IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 November 2025, accepted 3 December 2025, date of publication 8 December 2025,
date of current version 12 December 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3641299

RESEARCH ARTICLE

From Sensing to Segmentation: Transformer-Based Worker Activity Recognition for Industrial Assembly

ANDREA SBARAGLI[1] AND MATTEO NARDELLO[2], (Member, IEEE)
[1]Department of Engineering, Newcastle University, NE1 7RU Newcastle, U.K.
[2]Department of Industrial Engineering, University of Trento, 38123 Trento, Italy

Corresponding author: Andrea Sbaragli (andrea.sbaragli@ncl.ac.uk)

This work was supported by MUR PNRR through the Project Interconnected Nord-Est Innovation Ecosystem (iNEST) funded by European Union under NextGenerationEU under Grant ECS00000043.

ABSTRACT Human operators still provide strategic, value-added contributions to modern manufacturing systems, despite the widespread adoption of automation. This paper presents a cyber-physical system designed to perform fine-grained Worker Activity Recognition (WAR) and segment manual operations in human-centric assembly environments. The proposed system includes a modular and low-intrusion IoT acquisition layer that digitizes worker movements using Inertial Measurement Units (IMUs) and a radio frequency-based smart glove. While the IMUs track hand and back motion, the glove captures process interactions such as picking and depositing components. These multimodal data streams are processed by a cyber layer that performs time- and frequency-domain feature extraction and feeds the data into a Transformer-based neural network to classify five common assembly tasks. The results are post-processed by a Decision Support System to generate interpretable task segments and provide key performance indicators for industrial supervisors. The system is validated through a real-world deployment involving three workers assembling a vertical centrifugal pump, resulting in over 7 hours of annotated data. The Transformer model achieves an F1-score of 82% on this dataset. Thanks to its unobtrusive sensors, low deployment cost, and robustness to occlusions, the proposed solution offers a practical and scalable approach to enhancing visibility and traceability in manual industrial operations.

INDEX TERMS Human-centric manufacturing, industry 5.0, IoT, CPS, machine learning, worker activity recognition.

I. INTRODUCTION

The rise of wearable Internet of Things (IoT) technologies has enabled low-cost and distributed monitoring of human-centric manufacturing systems, offering new opportunities to enhance productivity, traceability, and operational insight [1]. The paradigm of Operator 4.0 emphasizes the strategic value of capturing and analyzing human activities on the shop floor, where detailed data on manual operations can support decision-making and process optimization [2], [3]. In this context, cyber-physical systems (CPS) integrate wearable sensors with intelligent processing layers, forming the backbone of human-centered digital manufacturing environments [4], aiding plant supervisors in decision-making [3].

However, recent global challenges — such as the COVID-19 pandemic, resource shortages, and the drive for sustainability — have accelerated the evolution towards Industry 5.0. This new paradigm focuses on enhancing sustainability and prioritizing worker well-being alongside automation [5]. A key component for achieving these objectives is accurate Worker Activity Recognition (WAR), which contextualizes manual tasks in dynamic and batch-dependent manufacturing systems. WAR techniques are typically classified into two

The associate editor coordinating the review of this manuscript and approving it for publication was Zhiwu Li.

© 2025 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

categories: Gross Movement Recognition and Fine-grained WAR approaches. Gross Movement Recognition targets large-scale motions (e.g., walking, carrying), often using motion capture (MOCAP) or real-time locating systems (RTLS) [6], [7], later on processed by machine learning algorithms to monitor logistic activities and resource utilization [8], [9]. Fine-grained WAR approaches, on the other hand, focus on detailed actions like tool usage or picking activities and involves two main computational steps [10]. First, IoT-based measurements are processed by a data engineering unit, which ensures a proper representation of the data and performs feature extraction both in time and frequency domains [11], [12]. Second, the processed data are classified using multi-class machine learning models, ranging from Support Vector Machines (SVMs) and Random Forests (RF) to advanced neural networks like CNNs and Long Short Term Memory (LSTM) networks [11], [13].

However, existing approaches suffer from two main limitations. First, many rely on vision-based systems like MOCAP, which are costly to deploy and sensitive to occlusions and spatial constraints. Second, training datasets are often recorded as isolated tasks (e.g., screwing or hammering), neglecting the complex transitions and variability of full assembly sequences [10], [14]. This makes it difficult to model the uncertainty inherent in real-world worker behavior, limiting the generalizability and deployability of such systems [10], [11], [13], [14].// To address these limitations, this paper proposes a novel cyber-physical system for fine-grained worker activity recognition in human-centric manufacturing environments. The system comprises:

- **A wearable IoT acquisition layer**, using two IMUs and an RFID-based smart glove to digitize both motion patterns and process interactions throughout full assembly sequences; The operator's dominant hand movements are tracked via acceleration profiles and the smart glove, while back movements are monitored through quaternion data streams
- **A Transformer-based cyber layer**, which processes time- and frequency-domain features to classify five task states, including multiple tool usage scenarios;
- **A Decision Support System**, which segments the classified tasks and generates interpretable outputs and key performance indicators (KPIs) for production oversight.

The system is validated through a real-world deployment involving three workers assembling a multistage centrifugal pump across 50 repetitions. The dataset contains over seven hours of annotated data and more than one million data points. The Transformer-based model achieves an F1-score of 82% on this dataset and demonstrates high accuracy on four public activity recognition benchmarks.// The main contributions of this work are:

- An operator-centric IoT acquisition layer that captures workers' motion patterns and process interactions during assembly activities through IMU wearables and an RFID-based smart glove. This layer offers high wearability and modularity, allowing flexible measurement configurations at reduced costs compared to MOCAP cameras.
- A Transformer-based neural network that learns patterns from the extracted features to perform fine WAR, classifying five distinct task states, including three related to screwing operations with different tools.
- A Decision Support System that integrates ML-based outputs with process interactions to segment workers' activities and provide key performance indicators (KPIs) for assembly processes.
- An extensive experimental campaign involving 50 repetitions of industrial pump assembly with three subjects, resulting in over seven hours of dataset collection.
- Validation of the Transformer-based neural network across the recorded dataset and four public activity recognition datasets.

The remainder of the article is organized as follows. Section II discusses the adoption of CPS to perform WAR in human-centric manufacturing systems with different enabling IoT technologies. Section III describes the developed CPS by highlighting the system's main parts, namely the *IoT acquisition layer*, the *cyber layer*, and the *Decision Support System*. Sections IV and V validate this digital solution by describing the experimental campaign carried out and the results achieved. Finally, Section VI ends this work by outlining the conclusions and further research opportunities.

II. RELATED WORK

Activity recognition is an increasingly significant research area with diverse applications, ranging from physical activity monitoring to industrial manufacturing processes [15]. In this context, machine learning-based CPS are favored for their ability to extract and classify human-relevant features from multimodal data, including video and wearable signals [16], [17]. In manufacturing environments, these monitoring approaches are typically referred to as Worker Activity Recognition (WAR), as they build a digital representation of workers' behaviors to support a range of operational goals. Beyond basic motion classification, WAR systems enable three main applications: improving worker safety through posture detection and ergonomic indices; facilitating human–automation collaboration; and segmenting manual assembly activities across production cycles. For example, [12] employed full-body IMUs and Hidden Markov Models to analyze postures and identify musculoskeletal risks. In collaborative scenarios, surface Electromyography and IMUs have been used to recognize operator gestures [18].

Despite their value, many of these approaches struggle to segment complete sequences of assembly tasks, which is essential for deriving accurate key performance indicators (KPIs) and managing process efficiency [13], [19]. This

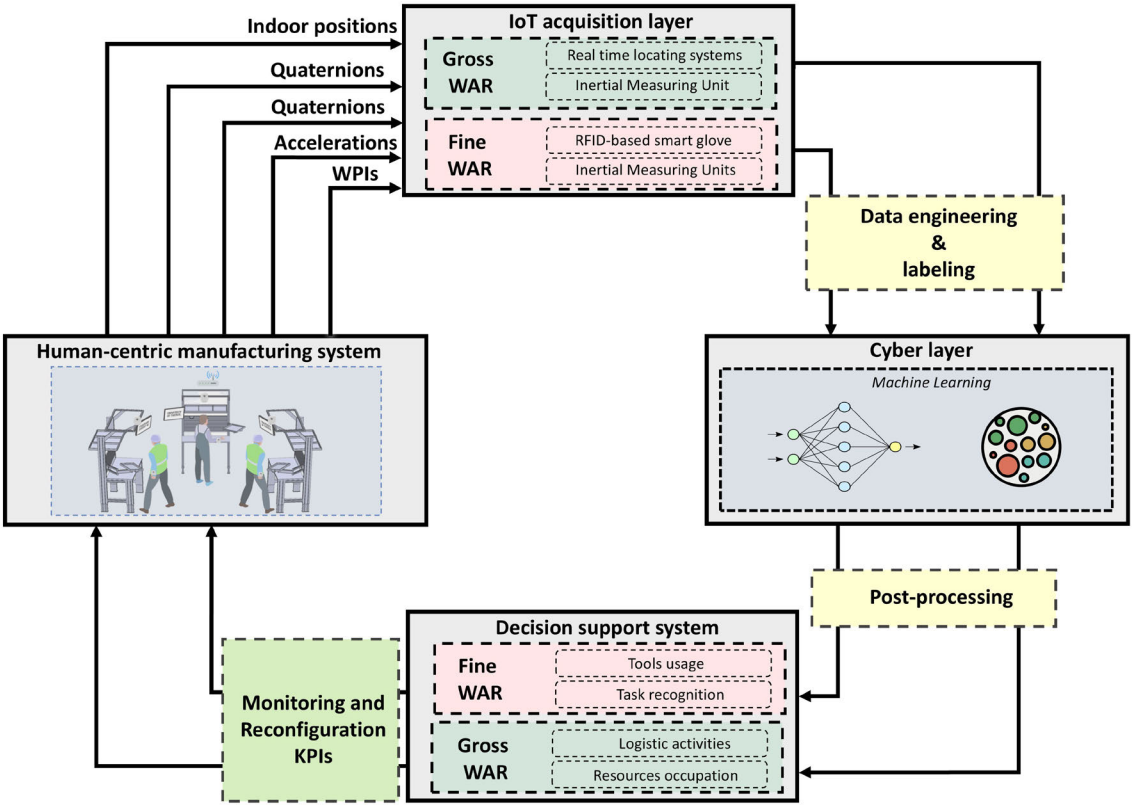


FIGURE 1. Overview of the proposed cyber-physical system for worker activity recognition. The architecture is composed of three main layers: (i) the IoT acquisition layer, which gathers multimodal data (e.g., indoor positions, accelerations, quaternions, and process interactions) using wearable IMUs, a smart RFID glove, and RTLS; (ii) the cyber layer, which performs data engineering, feature extraction, and task classification through a Transformer-based neural network; and (iii) the Decision Support System (DSS), which segments activities, integrates gross and fine WAR outputs, and computes key performance indicators (KPIs) for real-time monitoring and process optimization.

gap is particularly relevant in dynamic production settings characteristic of Industry 4.0 and 5.0 paradigms.

Various research contributions have focused on CPS for fine WAR. For example, [10] introduced a multi-modal WAR system that combines an IMU-based armband and cameras to digitize human actions and the industrial environment. This system uses a Convolutional Neural Network (CNN) with multiple layers to classify six basic assembly activities, such as tool usage and resting scenarios, achieving over 90% accuracy. However, practical deployment in industrial settings is hindered by several limitations, including potential occlusions in video-based acquisitions, high costs due to the dynamic nature of workers' movements across various locations, and a lack of dataset representativeness since tasks are recorded individually rather than within full assembly sequences. This acquisition approach does not model workers' variability in transitioning to the next task, limiting its applicability in real use cases. Similar limitations emerge by leveraging only camera-based CPS acquisition layers. For instance, [14] aims at segmenting the workstations' disassembly process of hard disk drives by using video acquisitions. A cyber layer employing a sequential variational autoencoder with CNN and Recurrent Neural Networks (RNN) was used to analyze spatial and temporal patterns in the images, classifying activities such as tool grabbing and manual operations with accuracies above 90%. Despite these successes, the approach faces structural challenges, such as product-dependent architectures that require fine-tuning for different production batches, leading to delayed deployment in industries with diverse product portfolios. Additionally, it fails to distinguish between manual and tool-assisted tasks, reducing its effectiveness in providing comprehensive visibility into assembly operations. As overlooking task transitions strongly limits applicability in real environments, [20] employs a single Kinect depth camera to detect and classify workers' activity changes in a manufacturing cell. Infra-red light is used to evaluate depth data and track human movement with 20 upper body joints. The acquired dataset involving 7 hours of data with two workers has two major limitations. First, this measuring unit suffers from occlusions as discussed before. Second, manual data removal is required due to the light reflection of industrial equipment. This provides repeated false skeleton detections, placing the worker under the conveyor or even the floor. Following noise removal, several features are extracted from body joints (e.g., the distance between hands), and

the data are labeled according to process tasks such as fetch, paint, and dry. An additional ''unknown'' class is introduced to account for uncertainties in the manufacturing cycle, particularly during worker-to-worker coordination and machine maintenance. Different configurations of Hidden Markov Models have been tested to segment task executions, leading to under-performing test set accuracies ranging from 65.2% to 67%. This investigation reiterates how camera measurements are challenged by the intrinsic variability of the human factor during manufacturing cycles.

Apart from camera-based methods, wearable-based CPS have also been explored for segmenting task executions. For example, [11] leveraged a smartphone to capture data on hand motions, including acceleration and orientation, along with environmental factors like brightness and noise levels. This data was processed using a variational autoencoder with an LSTM unit for classifying five activities, achieving perfect accuracy with short recordings. However, this approach shares similar data limitations as other methods, as short recordings do not accurately reflect real-world manufacturing sequences and overlook the transitions between tasks. When longer recordings were tested to model workers' motion pattern variability between tasks, the model's accuracy dropped to 76%. Additionally, the invasiveness of smartphone-based data acquisition poses challenges for widespread adoption in industrial settings, especially when more acquisition units are required. Reference [13] fills two consistent limitations of this approach. First, wrist-worn IMUs replace the smartphone acquisition system to improve workers' comfort and wearability during task executions. Second, the analysis targets the recognition of high-entropy motion patterns such as task transitioning. Acceleration and gyroscope data are fed into an SVM to classify various activities, from idle states to separating meat. However, classification accuracies were relatively low, peaking at 70% for a single worker and 63% for four combined workers. These lower accuracies impact the consistency of activity segmentation and usage in industrial environments. Reference [21] presents an additional investigation into segmenting assembly tasks by modeling the intrinsic uncertainty of human behavior in factory settings using five IMUs. Four inertial sensors are employed to monitor the motion patterns of both hands and one ankle, while a fifth sensor attached to the tool facilitates the detection of industrial equipment. The experimental campaign consisted of several sessions in which four screwing operations were individually recorded, totaling 10 hours of data collected from three participants. A null class was introduced to account for improperly executed assembly actions, as these signals were less descriptive or deviated from those of the four primary tasks. Different representations of the normalized motion patterns were used with CNN and LSTM models, incorporating oversampling and binary indicators to specify the use of the electric screwdriver. Despite these efforts, the balanced classification accuracy ranged from 35.01% to 58.81% across five datasets and only exceeded 80% when the null class was excluded. These results once again reiterate the limitations reported in the literature for deploying such solutions in industrial environments.

To conclude, while existing CPS have made strides in fine WAR, their applicability in real operating environments remains limited due to two main issues:

- **Cost and scalability:** MOCAP-based systems are sensitive to occlusions and require multiple deployed cameras to track workers across workstations. They also demand higher computational resources for video processing and ML inference.
- **Dataset realism:** most systems rely on isolated, short task recordings that fail to capture the full sequence of activities, and do not model the inherent variability in how workers transition between tasks.
- **Performances: the vast majority of approaches rarely exceed a classification accuracy of 70% when attempting to model the intrinsic uncertainty of human behavior in industrial assembly processes**

The CPS introduced in the following section addresses these limitations by enabling full-sequence segmentation using a lightweight, wearable sensor infrastructure, designed for effective integration in dynamic manufacturing environments.

III. CYBER-PHYSICAL SYSTEM

The architecture of the proposed cyber-physical system (CPS) is designed to perform Worker Activity Recognition (WAR) in human-centric manufacturing environments by combining wearable sensing, machine learning, and decision support. The level of WAR investigation determines the sensing granularity and the corresponding system components.

Specifically, **gross WAR** focuses on high-level motion and spatial behaviors, and is enabled through a UWB-based Real-Time Locating System (RTLS) and an IMU to track workers' indoor positions and orientations. In contrast, **fine WAR** targets detailed manual actions, such as tool usage and object manipulation, and relies on two IMUs and an RFID-enabled smart glove to monitor motion profiles and process interactions (WPIs in Figure 1). Accelerometers and gyroscopes capture 3D motion from the worker's dominant hand and back, generating rich temporal data over entire production cycles.

These multimodal signals are processed through a data engineering and labeling stage and fed into a machine learning-based cyber layer. Depending on the WAR modality, different models are applied: gross motion is analyzed via time-series and clustering techniques [8], [9], while fine motion is classified using a Transformer-based neural network trained on acceleration and quaternion features.

The resulting task classifications are post-processed and integrated with RFID-based events within a Decision Support System (DSS), which segments activities and extracts key performance indicators (KPIs) such as tool usage duration and task timing.

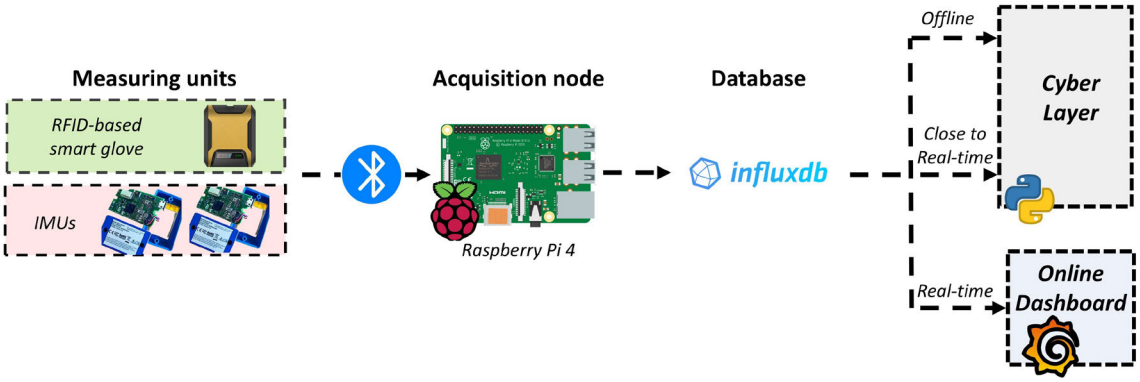


FIGURE 2. Proposed IoT system architecture. Wearable sensors (RFID glove and IMUs) transmit data via Bluetooth to a Raspberry Pi acquisition node, which stores measurements in an InfluxDB database. The data can then be processed offline, near real-time, or visualized in real time through an online dashboard.

It is important to highlight that both WAR modalities—fine and gross—are complementary. While gross WAR identifies logistics-related behaviors (e.g., manual material handling, zone transitions), fine WAR enables detailed segmentation of task execution and tool interactions [8]. This work primarily validates the fine WAR layer, as highlighted in red in Figure 1, and the following subsections describe its components in detail.

Figure 1 provides an overview of the system architecture. The proposed CPS is structured into three conceptual layers:

1) the **IoT acquisition layer**, responsible for collecting task-relevant motion and interaction data;
2) the **cyber layer**, which processes the sensor data through feature extraction and a Transformer-based classifier;
3) the **Decision Support System**, which segments tasks, associates them with process interactions, and computes KPIs for operational insight.

This modular architecture is designed to be deployable across a variety of assembly use cases where operators interact with different resources depending on production batches.

A. IoT ACQUISITION LAYER

The IoT acquisition layer relies on three wearable components to perform fine-grained activity recognition as shown in Fig. 2. Data acquisition is managed by a worker-specific Raspberry Pi 4 running Python scripts and handling concurrent Bluetooth connections. The architecture is modular and low-cost, allowing multiple data sources to be connected to the same node. This setup offers a clear advantage over MOCAP-based systems, which require multiple cameras, are susceptible to occlusions, and are harder to scale across the shop floor.To ensure data traceability and privacy, each measurement is tagged with a unique operator ID. The Raspberry Pi acts as a client node and transmits data to an InfluxDB instance hosted on a local server within the same network. The system supports both offline and near real-time operation modes: offline mode enables KPI generation after task completion, while Grafana dashboards allow for real-time visualization of raw data. The wearable units include: (i) a smart glove equipped with a battery-powered RFID reader to acquire process interactions by scanning passive tags deployed in storage locations (e.g., supermarket bins); and (ii) two IMUs that sample 3D acceleration and quaternion to track workers' motion profiles.

The smart glove is worn on the worker's dominant hand, with the electronics positioned on the back of the hand and the antenna integrated into the palm side to minimize intrusiveness during manual tasks. The glove incorporates a battery-powered RFID reader featuring a Cortex-M3 microcontroller and a linearly polarized antenna. The device allows output power modulation between 1 and 30 dBm, achieving a maximum reading range of 80 cm. It operates at a sampling frequency of 10 Hz and collects the unique tag ID, the received signal strength indicator (RSSI), and the timestamp for each scanned passive RFID tag within range. The complete unit weighs 64 g and is designed for continuous use in industrial environments without hindering the operator's movements.

Workers' motion profiles are captured using two wearable Inertial Measurement Units (IMUs), each integrating a low-power STM32L4 microcontroller that handles sensor configuration and data processing. The IMUs embed accelerometers, gyroscopes, and magnetometers, and stream both 3D acceleration and quaternion data at a sampling rate of 52 Hz. One IMU is positioned on the worker's dominant hand to capture detailed hand movements during task execution, while the second is placed on the upper back to track postural dynamics through quaternion orientation. Notably, the back-mounted IMU also supports gross WAR analysis, enabling broader motion tracking with minimal additional hardware. This dual-placement approach ensures comprehensive yet non-intrusive monitoring of fine and gross activities using a limited number of measuring units.

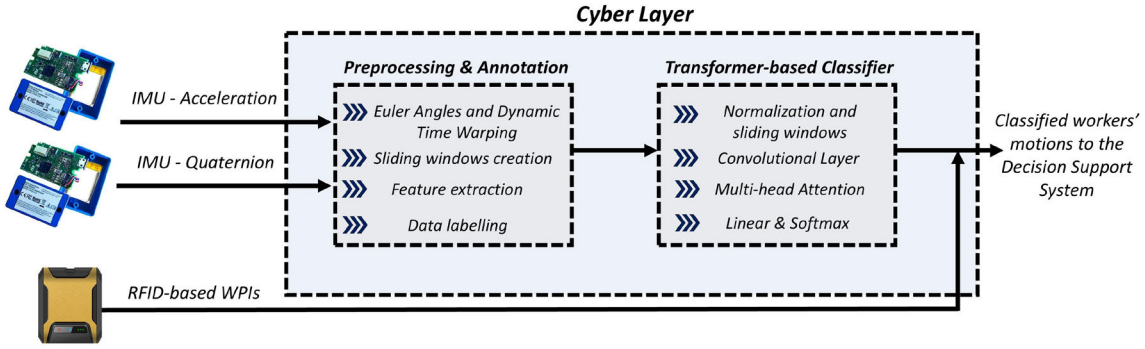


FIGURE 3. Data processing flow in the cyber layer. IMU-based acceleration and quaternion data undergo feature extraction and labeling before being processed by a Transformer-based neural network. Classified tasks are then combined with RFID-based worker–process interactions and passed to the Decision Support System for activity segmentation and KPI computation.

B. CYBER LAYER

This subsection describes the machine learning pipeline used for fine-grained Worker Activity Recognition (WAR) in low-standardized assembly environments. As shown in Figure 3, the IoT-acquired data streams follow two parallel paths. First, IMU signals capturing acceleration and orientation are processed in the cyber layer through feature extraction in both time and frequency domains (see Subsubsection III-B1), and then classified using a Transformer-based neural network (see Subsubsection III-B2). Second, the resulting activity classifications are combined with RFID-based process interaction data and forwarded to the Decision Support System (see Subsection III-C). This final layer post-processes the information to segment assembly tasks and extract relevant key performance indicators (KPIs) in support of human-centric manufacturing analysis.

1) PREPROCESSING & ANNOTATION

This preprocessing stage prepares the raw sensor data for the machine learning pipeline within the cyber layer. The first step involves converting the quaternion signals—captured from the IMU placed on the worker's back into Euler angles. This conversion is performed with respect to a reference acquisition in which the worker is standing still, allowing the computation of relative yaw, pitch, and roll values.

Next, the 3D acceleration data (from the hand-mounted IMU) and the Euler angle sequences are temporally aligned using the Dynamic Time Warping (DTW) algorithm to account for sampling delays and asynchronies.

The synchronized data streams are then segmented using a sliding window of 128 samples with 50% overlap between consecutive windows. For each window and each signal axis, statistical features are extracted in both the time and frequency domains. Time-domain features include mean, variance, and root mean square (RMS), while frequency-domain features include mean frequency and median frequency. Frequency features are averaged across axes, resulting in a final feature vector with a dimensions of 22 per window.

Finally, each windowed segment is manually labeled based on a video-based ground truth recorded during the experimental campaign. These labeled feature vectors form the input to the neural network classifier in the cyber layer.

2) TRANSFORMER-BASED CLASSIFIER

The Transformer-based model architecture is based on the encoder-only part proposed in [22], which has demonstrated superior performance in multivariate time-series classification compared to conventional state-of-the-art methods. The encoder-only design results in reduced model complexity and faster training times.

Prior to training, dataset samples are normalized and processed using a sliding window approach with a window size of 50 samples and a stride of 1. The preprocessed data are then arranged in batches of size 64 and split into training, validation, and test sets. Each sample consists of a temporal sequence of extracted features and is passed through the network, as illustrated in Figure 4.

To enhance temporal resolution, the input is first processed by a one-dimensional convolutional layer with 100 output channels, a kernel size of 3, and a stride of 1. Since the Transformer is a feed-forward architecture without recurrence or convolutional memory, a positional encoding mechanism is applied to inject information about the order of the sequence. This encoding is based on sine and cosine functions at varying frequencies, following the method introduced in [23].

The encoded feature sequences are then input into the Transformer encoder, which uses a Multi-Head Attention (MHA) mechanism with two heads. The attention mechanism relies on learnable Query, Key, and Value vectors and computes scaled dot-product attention followed by a softmax operation. Using multiple heads allows the model to learn different representations in parallel, which are then concatenated and projected to match the original dimension.

The core Transformer block also includes two Add & Norm layers with residual connections and batch normalization to address vanishing gradients. These are followed by a fully connected feed-forward network with two linear layers

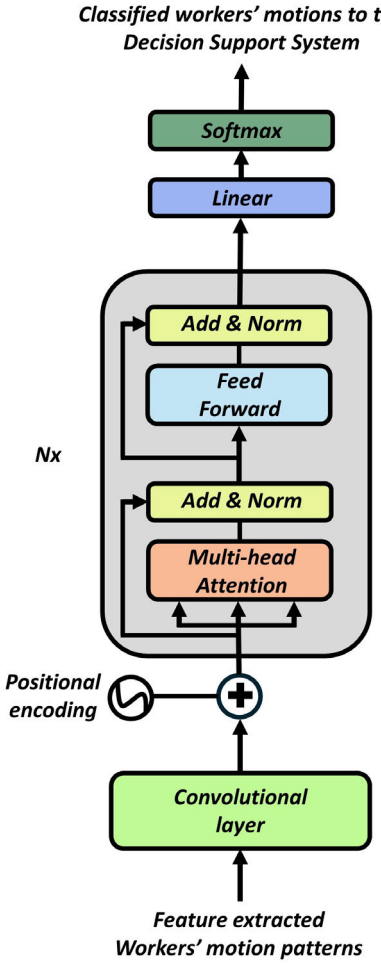


FIGURE 4. Transformer-based neural network architecture. The model processes feature-extracted motion patterns from IMU sensors through a convolutional layer followed by positional encoding. A stack of encoder blocks, including multi-head attention, feed-forward layers, and normalization, learns temporal dependencies. The output is passed through linear and softmax layers to classify workers' activities, which are then forwarded to the Decision Support System.

of inner dimension 1028 and a ReLU activation in between, introducing non-linearity into the architecture.

The final classification is performed by a linear layer followed by a softmax activation that outputs class probabilities. The resulting task predictions are forwarded to the Decision Support System (see Figure 3), where they are fused with RFID-based process interaction data to enable segmentation of manual task executions during the assembly workflow.

The neural network is implemented in PyTorch and trained on a Dell PowerEdge R740 equipped with an Intel Xeon Platinum 8468H processor and an NVIDIA A16 GPU.

C. DECISION SUPPORT SYSTEM

The Decision Support System (DSS) constitutes the final layer of the proposed CPS and is responsible for segmenting worker-specific manual activities in human-centric manufacturing environments. This layer operates by combining the classified motion patterns with RFID-derived process interactions (see Fig. 3) and performs two key functions.

First, a post-processing algorithm organizes the classified tasks (see Table 1) into temporal windows that group consecutive homogeneous activities. Each window is characterized by start and end timestamps, enabling the calculation of task duration. The DSS then queries the InfluxDB database to retrieve RFID-based process interactions. These interactions provide essential contextual information, such as the specific storage location (e.g., supermarket bin) involved in picking and depositing (P/D) operations.

To associate P/D events with specific storage locations, a local search algorithm analyzes the scanned RFID tags within each window. In cases where multiple tags are detected, the location with the highest number of reads is assigned. This procedure achieves an assignment accuracy of 98%. These fine-grained annotations help identify the transition points between assembly activities, since interactions with storage locations typically indicate the start of a new sequence of tasks. Special care is taken in scenarios where multiple pickings occur within a single assembly phase.

Figure 5 shows the activity segmentation results obtained during the assembly of an industrial pump. Each assembly activity segment (e.g., A1, A2, A3) comprises multiple classified tasks, including both manual operations—such as tool grabbing and screw preparation—and tool-assisted actions (e.g., using hexagonal, jack, or Phillips screwdrivers). The segmentation also highlights picking and depositing (P/D) windows, enriched with RFID-based annotations that identify the specific source locations involved, such as floor-level storage or designated supermarket bins. For instance, the third P/D event corresponds to items picked from both the ground and Bin 02.

In cases where multiple pickings occur within a single assembly activity, a rule-based approach is used to disambiguate events and to ensure correct temporal grouping and labeling. Blank intervals between P/D events or between manual tasks reflect transitions across distinct physical areas (e.g., between storage and workstations). These spatial gaps underscore the relevance of integrating gross activity recognition techniques, such as those proposed in [8] and [9], to fully capture logistical flows and resource utilization on the shop floor. Second, the DSS leverages the segmented tasks and activities to compute the following strategic key performance indicators (KPIs):

- **Activity duration**: exploiting historical data may signal potential bottlenecks and enable early alerts for implementing short-term corrective actions and long-term process reconfiguration. Indeed, the identification of tasks within each activity accelerates this process. This information represents a valuable input also for implementing heuristic approaches to optimize production scheduling.
- **Throughput time**: represents the sum of all activity durations for completing a product. This data is pivotal

for lean manufacturing principles to monitor operational efficiency.

- **Tool usage:** represents a valuable indicator to automate the availability and the performance in Overall Equipment Effectiveness Investigations (OEE). While the availability is given by the ratio of run time and planned production time, the performance takes into account the desired throughput time and produced units over the actual run time. This time-driven parameter may be leveraged by heuristics to optimize the availability while minimizing concurrent operations across different activities.

Although the experimental evaluation and Transformer-based detection performance are discussed in subsequent sections, this segmentation output already enhances process transparency by enabling fine-grained visibility into human-operated assembly workflows.

IV. INDUSTRIAL EVALUATION SETUP

The pilot industrial environment involves the assembly of a multistage vertical cast iron centrifugal pump. The assembly process takes place in an industrial workstation equipped with three manual screwdrivers and in a supermarket that stocks components, fasteners, and work-in-progress products (see Fig. 6 (a)). Based on operator experience, each pump assembly cycle lasts between 5 and 7 minutes, combining manual handling and tool-assisted operations. The assembly sequence is structured into three main activities (A1, A2, and A3), each comprising specific subtasks as detailed below:

- **A1 - Bush Mounting:** Operators begin by picking the pump shaft from the ground and retrieving the bush and its fasteners from a designated bin in the supermarket (Fig.6 (b)). The bush is manually fitted into the pump housing, followed by a preliminary screwing operation using the hexagonal screwdriver (Fig.6 (c)). The screws are not fully tightened at this stage, as the motor will be installed over the bushing in the next activity. The average duration of this task is 1.45 minutes, with a standard deviation of 0.41 minutes.
- **A2 - Motor Mounting:** Leaving the WIP pump in place, workers collect two additional components and proceed to install the motor onto the base structure. The bush is then permanently fixed using the hexagonal screwdriver. Subsequently, four bolts are manually positioned at the motor base and tightened using the jack screwdriver (Fig. 6 (d)). This is the most time-consuming activity in the sequence, with an average duration of 2.52 minutes and a standard deviation of 0.91 minutes.
- **A3 - Plates Mounting:** In the final step, two aluminum plates and four screws are retrieved from the supermarket. The plates are manually aligned and secured on each side of the pump using the Phillips screwdriver (Fig. 6 (e)). Once completed, the finished product is deposited back in the supermarket area, and the next assembly cycle begins. This task has an average duration of 1.23 minutes and a standard deviation of 0.37 minutes.

TABLE 1. Dataset classes.

Class	Task	Distribution [%]
P/D	Picking or depositing of assembly products, fasteners, and components from supermarket	14.3
Manual	Any manual operation in A1, A2, and A3	26.2
Hexagonal	Use of the Hexagonal screwdriver in T1 and T2	19.9
Jack	Use of the Jack screwdriver in A2	19.6
Phillips	Use of the Phillips screwdriver in A3	20.0

To validate the proposed methodology, three male subjects participated in the experimental campaign, each equipped with the described IoT devices as shown in Fig.6(f). Based on the defined assembly sequence, their motion profiles were annotated according to the activity classes summarized in Table1. While picking/depositing (P/D) operations and tool-specific actions are clearly represented in Fig.6, the Manual class aggregates a diverse set of intermediate actions, including, bolt alignment, screw handling, and transitions between tool usages. Although this class poses a challenge to the learning capabilities of the Transformer-based model due to the high variability and entropy of its motion patterns (see sectionIII-B2), capturing these transitional behaviors is essential for deploying CPS in real-world industrial environments. This inclusion marks a significant advancement over the reviewed approaches discussed in Section II.

Three male participants performed multiple repetitions of the entire assembly process, resulting in approximately 7 hours of recordings and over one million raw sensor samples. The participants' physical traits are as follows: Participant 1 — height (H) 178 cm, weight (W) 76 kg; Participant 2 — H 185 cm, W 80 kg; Participant 3 — H 192 cm, W 90 kg. In particular, the first two subjects completed 25 full assembly cycles each, while the third subject performed a focused 45-minute session alternating P/D, Hexagonal, Jack, and Philipps operations to address class imbalance. The recorded dataset was processed using the feature extraction pipeline described in section III-B1, yielding 19,278 labeled input samples. The final class distribution is reported in Table 1.

V. EXPERIMENTAL RESULTS AND ANALYSIS

This section presents the experimental results related to the proposed cyber layer (see Subsubsection III-B2). The evaluation focuses on two main aspects. First, the performance of the Transformer-based neural network in recognizing assembly tasks is benchmarked against two widely used neural architectures in WAR applications:

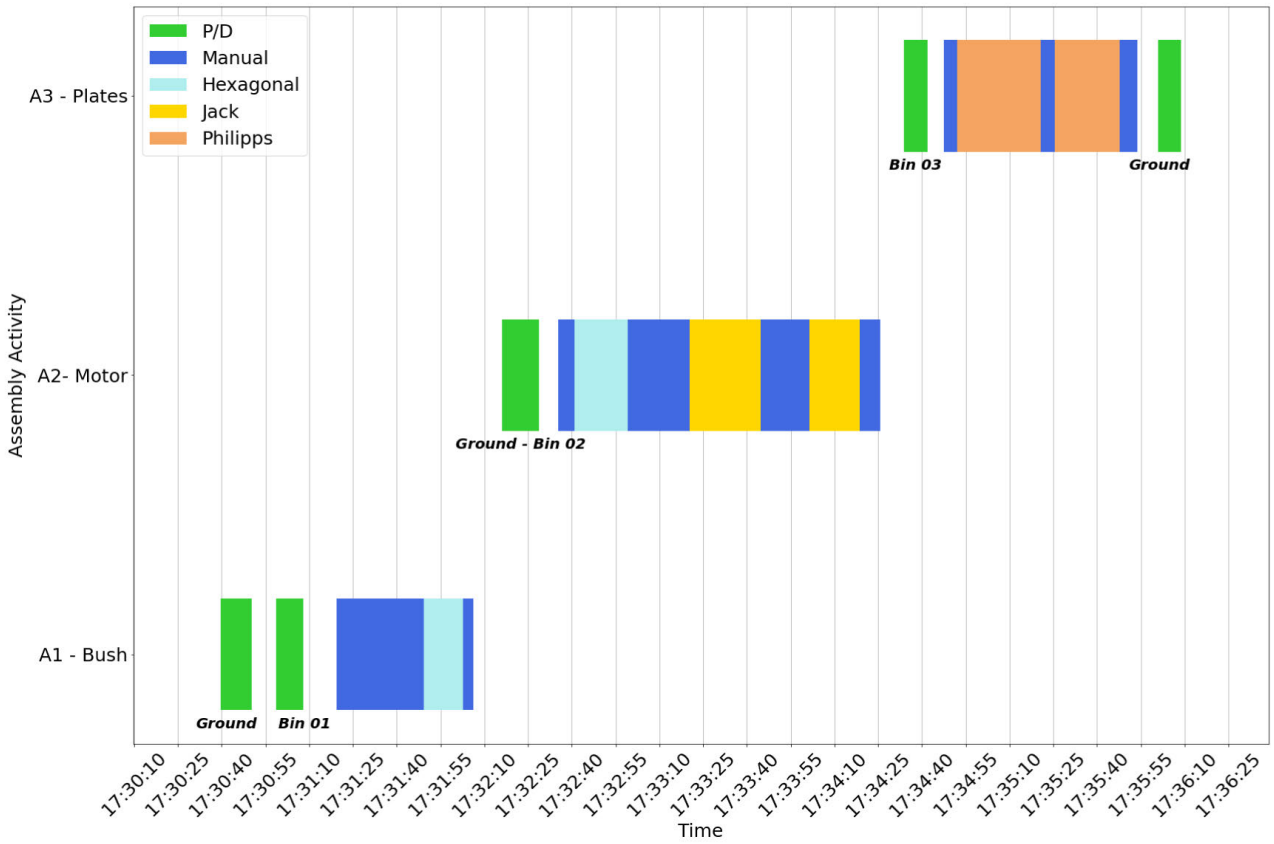


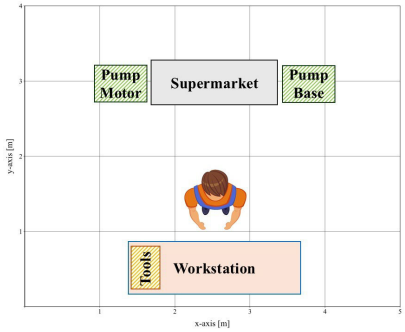
FIGURE 5. Segmentation of assembly activities based on task classification and process interactions. The chart shows three sequential assembly activities (A1 – Bush, A2 – Motor, A3 – Plates), each composed of multiple tasks including manual operations (blue), tool-assisted actions (light blue, yellow, orange), and picking/depositing (P/D) events (green). RFID annotations identify the source locations (e.g., Bin 01, Bin 02, Ground), while blank intervals indicate transitions between spatially separated work zones.

CNN and LSTM. For this evaluation, the cyber-layer data flow illustrated in Fig. 3 utilizes the same human-centric datastreams (e.g., IMUs and RFID) and the initial module (e.g., Preprocessing and Automation). However, the Transformer-based classifier is replaced by the CNN–LSTM implementation validated in [9] and [24]. This comparison considers three different input data, ranging from raw data to feature-extracted acceleration and quaternion profiles, as described in Subsubsection III-B1. Second, the generalizability of the Transformer model is validated using four publicly available activity recognition datasets. Unlike the previous performance investigation, the cyber layer utilizes the datastreams, feature extraction methods, and data partitions of public datasets together with the Transformer-based classifier shown in Fig. 4. Confusion matrices are employed in both benchmarking evaluations to compute standard classification metrics, including accuracy, precision, recall, and F1-score.

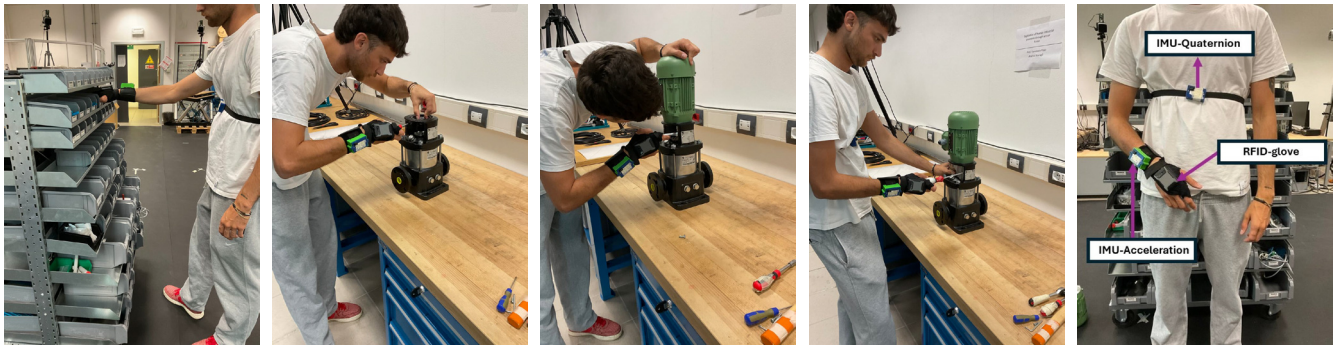
A. PERFORMANCE COMPARISON FOR FINE WAR

The dataset collected during the experimental campaign was used to test different combinations of neural networks and input data representations. For each of the three datasets, the data is split across the various sets using the same ratio and seed. Tab. 2 compares the classification performance of the Transformer, CNN, and LSTM architectures. The CNN and LSTM models follow the configurations proposed in [9] and [24], with the LSTM modified to use a single unit instead of two. The benchmark considers three input variants: raw IMU signals, feature-extracted accelerometer data (Acc.), and combined feature sets from accelerometers and quaternions (Acc. & Quat).

Overall, the use of raw data results in the lowest performance across all models, coupled with the longest training times—up to 3 minutes per epoch for the LSTM and approximately 1 minute per epoch for the CNN and Transformer. CNN performance remains weak even with feature-extracted datasets, yielding F1-scores below 60%. In contrast, both the LSTM and Transformer show significant performance improvements when feature extraction is applied. Notably, the Transformer consistently outperforms the LSTM, which had been previously validated for gross WAR tasks [9]. Interestingly, while the LSTM performs slightly better using only acceleration-based features, the Transformer benefits from the inclusion of quaternion data,



(a) Set-up



(b) Picking operation (c) Hexagonal screwing (d) Jack screwing (e) Philipps screwing (f) Subject wearing IoT sensors

FIGURE 6. Overview of the industrial-pilot pump assembly process. (a) Industrial set-up of the experimental campaign, (b) Picking components from supermarket bins, (c) screwing with a hexagonal tool, (d) screwing with a jack tool, (e) screwing with a Philipps tool, and (f) operator equipped with wearable IoT devices, including IMUs (for acceleration and orientation) and an RFID-based smart glove.

achieving a consistent 2% performance boost. Despite the increased input dimensionality, training times remain efficient, never exceeding 3 minutes per epoch. Figure 7 presents the training and inference performance of the best-performing Transformer configuration. Although the Cross Entropy loss curve (Fig.7 (a)) suggests potential overfitting, F1-scores stabilize after convergence, demonstrating model robustness over 300 training epochs. The confusion matrix in Fig.7 (c) refers to epoch 117, where the model achieved its minimum validation loss of 0.61 and a test-set F1-score of 82%. The vast majority of classification errors are associated with the Manual class, which includes a broad range of high-entropy transitions between task-specific operations (see Table 1). These operations range from tool grabbing to manual screw fitting and are often distinguished by limited duration and hardly replicable movements among workers. Despite this, the Manual class reaches an F1-score of 66%, while the remaining classes achieve scores ranging from 85% to 89%, highlighting the CPS's strong capabilities for activity segmentation and task-level monitoring in human-centric industrial settings.

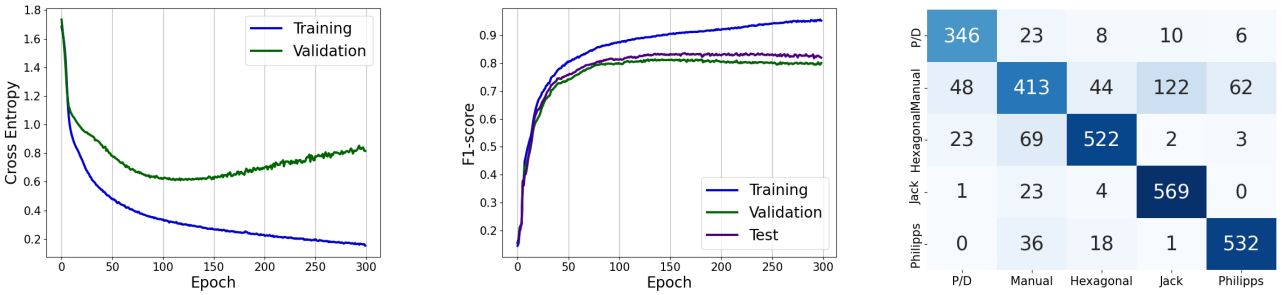
B. PERFORMANCE COMPARISON OF PUBLIC DATASETS

This second analysis evaluates the generalization capability of the proposed CPS using four publicly available

TABLE 2. Ablation study comparing the performance of three neural network architectures (CNN, LSTM, and Transformer) on different input representations: raw IMU data, feature-extracted accelerometer data (Acc.), and combined acceleration and quaternion features (Acc. & Quat). Results are reported in terms of Recall (Rec.), Precision (Pr.), F1-score (F1), and Accuracy (Acc.)

Input Data	Model	Rec.	Pr.	F1	Acc.
Raw	**CNN** [24]	20.0	26.3	15.0	31.2
Acc.		61.1	59.0	58.0	59.4
Acc. & Quat		55.0	57.2	51.3	53.1
Raw	**LSTM** [9]	31.1	15.1	19.3	34.2
Acc.		79.0	79.0	78.2	78.3
Acc. & Quat		79.3	78.0	78.0	78.1
Raw	**Transformer (ours)**	62.4	65.5	60.7	60.9
Acc.		80.4	80.1	80.0	80.5
Acc. & Quat.		**82.3**	**82.0**	**82.0**	**82.7**

Human Activity Recognition (HAR) datasets. The Transformer classifier is validated on the feature-extracted data corresponding to each dataset. These datasets are typically collected under controlled conditions with isolated activity windows, which, while simplifying model training, limit robustness and hinder practical deployment in real-world scenarios. Moreover, the literature presents a very limited amount of datasets focusing on activity



(a) Cross Entropy losses (b) F1-score over epochs (c) Confusion Matrix for the 117th model epoch

FIGURE 7. Training and inference metrics of the Transformer-based neural network with the Acc. & Quat. dataset.

recognition in industrial environments [25]. The vast majority of investigations target HAR of gross movements such as walking, lying, and sitting, which further facilitates pattern recognition of deep learning classifiers. Tab. 3 summarizes the test results obtained by training the proposed Transformer-based neural network on each dataset. On average, training required 15 seconds per epoch, with full convergence achieved in less than 10 minutes per dataset.

UCI HAR Dataset – The first evaluation uses the UCI HAR dataset [26], which includes triaxial linear accelerations and angular velocities acquired by a waist-mounted Samsung Galaxy S II during six physical activities: walking, walking upstairs/downstairs, sitting, standing, and lying. The dataset features 561 time and frequency-domain descriptors computed over 128-sample windows with 50% overlap. The Transformer achieves an F1-score of 98.3%, validating its effectiveness with high-dimensional input and heterogeneous motion profiles.

Skoda Dataset – The second benchmark involves the Skoda dataset [25], where a subject performs 70 repetitions of 10 automotive quality inspection activities (e.g., door closing and opening) while wearing 20 IMUs on the upper limbs. Despite its richness and inclusion of transition segments via a null class, the dataset is limited by single-subject acquisition and low wearability due to the large number of sensors. The best configuration used the mean and standard deviation over 64-sample windows with 50% overlap. Although the F1-score is 4% lower than the CNN-based results reported in [25], the Transformer converges in 7 minutes and achieves performance suitable for industrial deployment. Six classes exceed 95% F1-score, while the remaining four range between 78% and 85%, with the lowest score in the gasp checking class.

MHealth Dataset – This dataset includes 12 physical activities recorded from 10 subjects using IMUs placed on the chest, right wrist, and left ankle. The activity set broadly targets static positions (e.g., standing and lying) and dynamic ones, including jumping and running. The data consist of triaxial accelerometer and gyroscope signals processed using the feature extraction pipeline described in Subsubsection III-B1. The Transformer outperforms previous state-of-the-art methods [17], [27], reaching an F1-score of 99.9%.

TABLE 3. Performance of the proposed Transformer-based neural network on four publicly available human activity recognition (HAR) datasets. Results are reported in terms of Recall (Rec.), Precision (Pr.), F1-score (F1), and overall Accuracy (Acc.)

Dataset	Rec.	Pr.	F1	Acc.
Galaxy S II [26]	98.2	98.3	98.3	98.3
Skoda [25]	91.0	91.4	91.2	91.2
MHealth [28]	99.8	99.8	99.9	99.9
PAMAP2 [29]	87.1	88.5	90.2	89.2

PAMAP2 Dataset – Finally, the PAMAP2 dataset includes 12 activity classes recorded from 9 subjects using three IMUs located similarly to MHealth. Many activities overlap with the previous approach (e.g., lying and sitting), whereas many others, such as computer work and ironing, are dataset specific. The Transformer is trained on features extracted from accelerometer (±16g) and gyroscope signals. The model achieves an F1-score of 90.2%.

C. SYSTEM LATENCY EVALUATION

To assess the responsiveness of the IoT infrastructure, system latency was measured over a continuous 48-hour period by recording the time required to store acquired sensor data into the database. The two IMU devices exhibited an average latency of 61 *ms* with a standard deviation of 330 *ms*. In contrast, the RFID-enabled smart glove showed significantly lower latency, averaging 1 μs with a standard deviation of 22 μs.

VI. CONCLUSION

Worker Activity Recognition is a key enabler for improving visibility in human-centric manufacturing by identifying sub-tasks and analyzing manual operations. However, many existing approaches are evaluated in controlled settings using

isolated task recordings, which limits their applicability in real-world, dynamic assembly workflows.To address this gap, this work presented and validated a CPS designed to segment worker activities within industrial assembly lines using a three-layer architecture. The system includes: (i) an IoT acquisition layer that captures motion patterns and process interactions through two IMUs and an RFID-enabled smart glove; (ii) a cyber layer that performs feature extraction and activity classification using a Transformer-based neural network; and (iii) a Decision Support System that post-processes the output to segment assembly activities and compute key performance indicators. The proposed CPS was evaluated in a real industrial scenario involving three workers assembling a multistage centrifugal pump. The resulting dataset includes five activity classes and over one million labeled samples. Experimental results show that the Transformer-based model outperforms CNN and LSTM baselines, achieving an F1-score of 82%. Additional validation on four public activity recognition datasets further confirmed its robustness, with F1-scores exceeding 91%. Future research will focus on expanding the number of recognized activity classes to capture more nuanced manual operations and on enhancing the system's real-time capabilities to support proactive decision-making in manufacturing environments.

REFERENCES

[1] A. Albanese, M. Nardello, G. Fiacco, and D. Brunelli, "Tiny machine learning for high accuracy product quality inspection," *IEEE Sensors J.*, vol. 23, no. 2, pp. 1575–1583, Jan. 2023.

[2] D. R. Díaz, J. Stahre, T. Wuest, O. Noran, P. Bernus, Å. Fast-Berglund, and D. Gorecky, "Towards an operator 4.0 typology: A human-centric perspective on the fourth industrial revolution technologies," in *Proc. Int. Conf. Comput. Ind. Eng. (CIE)*, Tianjin, China, 2016, pp. 29–31.

[3] T. Ruppert, S. Jaskó, T. Holczinger, and J. Abonyi, "Enabling technologies for operator 4.0: A survey," *Appl. Sci.*, vol. 8, no. 9, p. 1650, Sep. 2018.

[4] S. Thiede, M. Juraschek, and C. Herrmann, "Implementing cyber-physical production systems in learning factories," *Proc. CIRP*, vol. 54, pp. 7–12, Jan. 2016.

[5] B. Gladysz, T.-A. Tran, D. Romero, T. van Erp, J. Abonyi, and T. Ruppert, "Current development on the operator 4.0 and transition towards the operator 5.0: A systematic literature review in light of industry 5.0," *J. Manuf. Syst.*, vol. 70, pp. 160–185, Oct. 2023.

[6] A. Luchetti, A. Carollo, L. Santoro, M. Nardello, D. Brunelli, P. Bosetti, and M. De Cecco, "Human identification and tracking using ultra-wideband-vision data fusion in unstructured environments," *ACTA IMEKO*, vol. 10, no. 4, pp. 124–131, Dec. 2021.

[7] T.-A. Tran, T. Ruppert, G. Eigner, and J. Abonyi, "Assessing human worker performance by pattern mining of Kinect sensor skeleton data," *J. Manuf. Syst.*, vol. 70, pp. 538–556, Oct. 2023.

[8] F. Pilati and A. Sbaragli, "Learning human-process interaction in manual manufacturing job shops through indoor positioning systems," *Comput. Ind.*, vol. 151, Oct. 2023, Art. no. 103984.

[9] F. Pilati, A. Sbaragli, G. P. R. Papini, and P. Capuccini, "An artificial neural network architecture to classify workers' operations in manual production processes," in *Proc. Int. Conf. Flexible Autom. Intell. Manuf.*, Cham, Switzerland. Springer, 2023, pp. 805–812.

[10] W. Tao, M. C. Leu, and Z. Yin, "Multi-modal recognition of worker activity for human-centered intelligent manufacturing," *Eng. Appl. Artif. Intell.*, vol. 95, Oct. 2020, Art. no. 103868.

[11] A. Mastakouris, G. Andriosopoulou, D. Masouros, P. Benardos, G.-C. Vosniakos, and D. Soudris, "Human worker activity recognition in a production floor environment through deep learning," *J. Manuf. Syst.*, vol. 71, pp. 115–130, Dec. 2023.

[12] A. Malaise, P. Maurice, F. Colas, and S. Ivaldi, "Activity recognition for ergonomics assessment of industrial tasks with automatic feature selection," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 1132–1139, Apr. 2019.

[13] A. R. M. Forkan, F. Montori, D. Georgakopoulos, P. P. Jayaraman, A. Yavari, and A. Morshed, "An industrial IoT solution for evaluating Workers' performance via activity recognition," in *Proc. IEEE 39th Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Jul. 2019, pp. 1393–1403.

[14] X. Zhang, D. Yi, S. Behdad, and S. Saxena, "Unsupervised human activity recognition learning for disassembly tasks," *IEEE Trans. Ind. Informat.*, vol. 20, no. 1, pp. 785–794, Jan. 2024.

[15] A. Y. Shdefat, N. Mostafa, Z. Al-Arnaout, Y. Kotb, and S. Alabed, "Optimizing HAR systems: Comparative analysis of enhanced SVM and k-NN classifiers," *Int. J. Comput. Intell. Syst.*, vol. 17, no. 1, p. 150, Jun. 2024.

[16] C. Gianoglio, A. Mohanna, A. Rizik, L. Moroney, and M. Valle, "On edge human action recognition using radar-based sensing and deep learning," *IEEE Trans. Ind. Informat.*, vol. 20, no. 3, pp. 4160–4172, Mar. 2024.

[17] N. A. Choudhury and B. Soni, "An adaptive batch size-based-CNN-LSTM framework for human activity recognition in uncontrolled environment," *IEEE Trans. Ind. Informat.*, vol. 19, no. 10, pp. 10379–10387, Oct. 2023.

[18] A. Kubota, T. Iqbal, J. A. Shah, and L. D. Riek, "Activity recognition in manufacturing: The roles of motion capture and sEMG+inertial wearables in detecting fine vs. Gross motion," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 6533–6539.

[19] A. Tóth, L. Nagy, R. Kennedy, B. Bohuš, J. Abonyi, and T. Ruppert, "The human-centric Industry 5.0 collaboration architecture," *MethodsX*, vol. 11, Dec. 2023, Art. no. 102260.

[20] D. J. Rude, S. Adams, and P. A. Beling, "Task recognition from joint tracking data in an operational manufacturing cell," *J. Intell. Manuf.*, vol. 29, no. 6, pp. 1203–1217, Aug. 2018.

[21] G. Sopidis, M. Haslgrübler, B. Azadi, B. Anzengruber-Tánase, A. Ahmad, A. Ferscha, and M. Baresch, "Micro-activity recognition in industrial assembly process with IMU data and deep learning," in *Proc. 15th Int. Conf. Pervasive Technol. Rel. Assistive Environments*, Jun. 2022, pp. 103–112.

[22] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff, "A transformer-based framework for multivariate time series representation learning," in *Proc. 27th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2021, pp. 2114–2124.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2025, pp. 5998–6008.

[24] Y. Kaya and E. K. Topuz, "Human activity recognition from multiple sensors data using deep CNNs," *Multimedia Tools Appl.*, vol. 83, no. 4, pp. 10815–10838, Jan. 2024.

[25] F. Ordóñez and D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, Jan. 2016.

[26] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones," *Esann*, vol. 3, pp. 437–442, Jan. 2013.

[27] K. D. Garcia, C. R. de Sá, M. Poel, T. Carvalho, J. Mendes-Moreira, J. M. P. Cardoso, A. C. P. L. F. de Carvalho, and J. N. Kok, "An ensemble of autonomous auto-encoders for human activity recognition," *Neurocomputing*, vol. 439, pp. 271–280, Jun. 2021.

[28] O. Banos, C. Villalonga, R. Garcia, A. Saez, M. Damas, J. A. Holgado-Terriza, S. Lee, H. Pomares, and I. Rojas, "Design, implementation and validation of a novel open framework for agile development of mobile health applications," *Biomed. Eng. OnLine*, vol. 14, no. Suppl 2, p. S6, 2015.

[29] A. Reiss and D. Stricker, "Introducing a new benchmarked dataset for activity monitoring," in *Proc. 16th Int. Symp. Wearable Comput.*, Jun. 2012, pp. 108–109.

ANDREA SBARAGLI received the B.S. and M.S. degrees in management engineering from the University of Bologna, Italy, in 2018 and 2021, respectively, and the Ph.D. degree in materials, mechatronics, and systems engineering from the University of Trento, Italy, in 2025, with a focus on cyber-physical systems for monitoring the efficiency and sustainability of human-centric manufacturing systems.

His doctoral research combined the Internet of Things (IoT) technologies with machine learning algorithms to develop innovative solutions for intelligent and adaptive industrial processes. During his Ph.D. studies, he also spent a visiting semester at the Department of Engineering Technology, University of Twente, the Netherlands. He is currently a KTP Associate with Newcastle University, as part of the Merit Program, where he leverages machine learning-based cyber-physical systems to enhance the monitoring, planning, and scheduling of human-centric manufacturing systems, driving digital transformation in construction manufacturing.

MATTEO NARDELLO (Member, IEEE) received the M.S. degree in electronics and telecommunications engineering and the Ph.D. degree in materials, mechatronics, and systems engineering from the University of Trento, Trento, Italy, in 2016 and 2020, respectively.

He is currently a Research Fellow with the Department of Industrial Engineering, University of Trento. His research interests include the investigation of machine learning techniques applied to resource-constrained embedded platforms (TinyML), with a special focus on the autonomous smart Internet of Things (IoT) devices and the study of new UWB-based architectures for indoor localization services. His current research interests include the modeling and hardware–software codesign of innovative solutions to reduce the power requirements of a distributed wireless sensor network for environmental sensing.