information

MDPI

*Article*

# Explainable Early Activity Recognition via Wearable Inertial Sensors for Human–Robot Collaboration in Agriculture

**Lefteris Benos [1,*], Erotokritos Skordilis [2], Remigio Berruto [3] and Dionysis Bochtis [1,4]**

1 Institute for Bio-Economy and Agri-Technology (IBO), Centre of Research and Technology-Hellas (CERTH), 57001 Thessaloniki, Greece; d.bochtis@certh.gr
2 Department of Business Technology, University of Miami, Coral Gables, FL 33124-6552, USA; sge12@miami.edu
3 Interuniversity Department of Regional and Urban Studies and Planning (DIST), University of Turin, Viale Mattioli 39, 10125 Torino, Italy; remigio.berruto@unito.it
4 FarmB Digital Agriculture S.A., Capital Trade Center, Building B, Laertou 22, 55535 Thessaloniki, Greece
* Correspondence: e.benos@certh.gr

**Abstract**

In open-field agriculture, timely human activity recognition (HAR) is critical for anticipating worker actions and enabling proactive human–robot collaboration. This study developed an offline early HAR framework based on pre-segmented activity sequences. Long Short-Term Memory (LSTM) networks were used in conjunction with wearable inertial sensors mounted on the chest, cervical region, lumbar region, and right and left wrists. Accelerometer, gyroscope, and magnetometer signals were collected from 20 participants during an outdoor agricultural material-handling task, with a mobile ground robot serving as the receiving platform for the crate. Each activity was represented by cumulative prefixes from 30% to 100% of its duration. At the 30% observation ratio, the model achieved a macro-F1 score of 0.9314, compared with 0.9481 for the full sequence, corresponding to an absolute difference of 0.0167. Shapley Additive Explanations (SHAP) were also used to identify the body locations, sensor modalities, and signal channels that contributed most. The early decisions were mainly supported by sensors placed on the trunk, with wrist sensors providing complementary information, particularly for standing classification. Multimodal inertial information was also important. The most influential inputs were mainly gyroscope and magnetometer channels from the chest, cervical region, and lumbar region. In conclusion, the high performance at early observation ratios highlights the potential to support more adaptive and better-coordinated robot-assistance strategies.

**Keywords:** Long Short-Term Memory (LSTM); Shapley Additive Explanations (SHAP); prefix-based classification; multi-location inertial sensors; signal-channel attribution; agricultural robotics



Academic Editor: Francesco Camastra

Received: 7 August 2026
Revised: 3 September 2026
Accepted: 4 September 2026
Published: 7 September 2026



## 1. Introduction

Agricultural robotics has enabled the task automation of several labor-intensive and repetitive tasks, including crop monitoring, harvesting, seeding, weeding, precision spraying, in-field transport, and material handling [1,2]. However, many agricultural tasks remain difficult to fully automate. Arguably, most of the challenges are found in open-field agricultural environments, which constitute dynamic ecosystems that are particularly susceptible to heterogeneity and unforeseen operational constraints [3]. These challenges are intensified by the variable lighting conditions, shifting weather patterns, uneven terrain, and biological diversity of crops [4,5]. The requirement for real-time safety is also of major importance [6]. This is because human workers and robotic systems frequently operate within the same shared workspace [7,8]. To mitigate the complexities inherent to these environments, synergy between humans and robots represents a promising approach [9,10]. In these systems, human cognitive strengths are combined with the precision and strength of robotic platforms to achieve common operational goals.

Central to this synergy is human–robot interaction (HRI), a multidisciplinary field that brings together robotics, artificial intelligence, sensor technologies, control engineering, and human factors. Human activity recognition (HAR) provides a necessary perception layer for HRI, since it enables robots to translate sensor-based observations of human movement into meaningful activity labels. This allows them to understand task progression and coordinate their actions with human workers [11].

### *1.1. Related Work on HAR in Agricultural HRI*

Driven by the growing interest in integrating human–robot collaboration in agriculture and the significant advancements in both sensor engineering and deep learning [12], two main methodological pathways have emerged in the relevant literature: vision-based and wearable-sensor-based HAR.

Indicatively, Morio et al. [13,14] investigated hidden Markov models for pea-production behavior recognition and Random Forests for fundamental worker-posture recognition. Their approach relied on a pan-tilt-zoom camera, color markers, silhouettes, and posture representations. More recent work has shifted toward video-based learning approaches. Pal et al. [15] utilized a combination of Mask Region Convolutional Neural Network (Mask R-CNN) and optical-flow descriptors with the intention of classifying human picker activities in strawberry polytunnels and apple orchards. Moreover, Moysiadis et al. [16] used depth cameras, skeleton extraction and machine-learning classifiers for the recognition of static hand gestures. In [17], this vision-based HRI approach was extended to dynamic human-movement recognition using pose extraction and deep-learning classification. In both studies, Robot Operating System (ROS) was also applied to translate human actions into unmanned ground vehicle (UGV) commands. Although the above systems supported HRI, some limitations were reported, associated principally with motion blur, lighting variation, and visual occlusions.

HAR based on wearable sensors, such as accelerometers, gyroscopes, and magnetometers embedded in inertial measurement units (IMUs), gloves, or smartphones, has also gained attention as it can overcome the aforementioned constraints. In short, a waist-mounted inertial sensing system with two accelerometer channels and one gyroscope channel was utilized in [18]. This system was used to evaluate k-Nearest Neighbors (kNN), support vector machines (SVM), and Artificial Neural Networks (ANN) for fall/non-fall classification and classification of certain agricultural activities. Gokul et al. [19] developed a wireless gesture-controlled robotic arm for weeding operations, where the hand gestures were captured via a sensor-embedded glove. Furthermore, Sharma et al. [20] proposed an approach based on neural networks for classifying usual agricultural activities, such as harvesting, transplantation, and walking, leveraging accelerometer data from a smartphone carried by a farmer. Martins et al. [21] investigated inertial-based posture recognition for agricultural tasks using wearable IMUs and five deep-learning architectures. They also applied the Gradient-weighted Class Activation Mapping (Grad-CAM) to improve the explainability of posture-classification decisions.

In the same vein, a situation-awareness layer via HAR was provided in [22], allowing a UGV to recognize specific worker activities so that human and robot actions can be synchronized during a collaborative harvesting scenario. Data were gathered during outdoor farm experiments from participants wearing five IMUs on strategically selected body locations. A Long Short-Term Memory (LSTM)-assisted HAR framework was implemented to investigate optimal sensor placement and multimodal sensor fusion. The same publicly available dataset was reused by Mekruksavanich and Jitpattanakul [23]. In this study, a lightweight one-dimensional ResNeXt architecture was proposed while different temporal-window lengths were evaluated. The methodological focus changed in [24], which used the same dataset. Specifically, Shapley Additive Explanations (SHAP)-based interpretation was used to quantify feature-level contributions associated with body location, sensing modality, and signal axis to the model predictions, while XGBoost was used to classify the activities.

A concise comparison of the reviewed studies and the proposed method is provided in Table 1. The comparison focuses on the agricultural context, sensor type, methodological approach, early-recognition capability, and explainability component.

**Table 1.** Comparison of related studies on HAR in agricultural HRI and the proposed method.

| Study | Agricultural Context/Task | Sensor Type | Method | Early Recognition | Explainability |
|---|---|---|---|---|---|
| Morio et al. [13] | Worker-posture recognition for understanding agricultural worker behaviors | Camera-based silhouettes and color-marker posture representations | Lookup-table-based and Random-Forest-based posture recognition | No | No |
| Morio et al. [14] | Agricultural worker behavior recognition for intelligent worker assistance in pea production | Pan-tilt-zoom camera and color-marker-based posture information | Hidden Markov models | No | No |
| Pal et al. [15] | Human picker activity classification in strawberry polytunnels and apple orchards | RGB-D/depth video | Mask R-CNN, optical flow, KDE, and K-means-based classification | No | No |
| Moysiadis et al. [16] | Static hand-gesture recognition for agricultural HRI | Depth camera and skeleton-based hand landmarks | ML classifiers and ROS-based robot-command execution | No | No |
| Moysiadis et al. [17] | Dynamic human-movement recognition for agricultural HRI | Vision-based human-movement data | LSTM-based movement classification and ROS-based UGV actions | No | No |
| Son et al. [18] | Fall and activity classification in agricultural workers | Two accelerometers and one gyroscope | kNN, SVM, and ANN | No | No |
| Gokul et al. [19] | Gesture-controlled agricultural weeding robot | Sensor-embedded glove | Gesture-based wireless robotic-arm control | No | No |

**Table 1.** *Cont.*

| Study | Agricultural Context/Task | Sensor Type | Method | Early Recognition | Explainability |
|---|---|---|---|---|---|
| Sharma et al. [20] | Agricultural activity detection, including harvesting, bed-making, transplantation, walking, and standing | Mobile-phone accelerometer | Neural networks | No | No |
| Martins et al. [21] | Inertial-based posture recognition for agriculture- and construction-related tasks | Wearable inertial sensors | CNN, LSTM, Transformer, CNN-LSTM, and CNN-Transformer models | No | Yes; Grad-CAM-based interpretation |
| Benos et al. [22] | Agricultural harvesting scenario involving human–robot collaboration | Wearable IMUs | LSTM-based HAR; sensor-placement and multimodal-fusion analysis | No | No |
| Mekruksavanich and Jitpattanakul [23] | Wearable-sensor-based activity recognition for agricultural HRI | Wearable IMUs | Lightweight 1D-ResNeXt | No; short-window analysis but no prefix-based early recognition | No |
| Benos et al. [24] | Explainable HAR for human–robot collaboration in agriculture | Wearable IMUs | XGBoost and SHAP | No | Yes; SHAP after complete-activity classification |
| Present study | Outdoor agricultural HRI crate-handling task | Wearable IMUs | LSTM and SHAP | Yes; cumulative activity prefixes from 30% to 100% | Yes; observation-ratio-specific SHAP analysis at body-location, sensor-modality, and signal-channel levels |

### 1.2. Research Gap, Novelty, and Contributions

As summarized in Table 1, existing studies have mainly focused on HAR where agriculture-related worker activities are classified using complete activity recordings. This retrospective approach, however, is less suitable for real-time agricultural HRI, as collaborative robots (also known as "cobots") often need to react before a worker has completed an action [25,26]. In time-series research, early classification specifically addresses this requirement by predicting the class of a partially observed time series, where earliness and accuracy usually represent two competing objectives [27]. Successful applications have been reported in several time-critical domains, such as medical diagnostics, quality monitoring in manufacturing, and crop type mapping [28,29]. In collaborative robotics, the ability to infer a worker's intended action before the action is completed is commonly discussed in terms of human intention recognition, action anticipation, and proactive human–robot collaboration [30,31].

An important methodological gap remains between early time-series classification and wearable-sensor-based agricultural HAR. Studies on wearable-sensor-based HAR in agriculture have shown that IMU data can support activity classification. Nevertheless, they have not quantified the amount of partial movement information required for reliable activity recognition. This is particularly important in collaborative field operations, where the practical value of HAR depends on both classification accuracy and whether the recognition is available early enough to support timely robot response. In practical terms, early predictions should be accompanied by interpretable information that can guide the optimization of sensor and body-location selection. This has the potential to reduce system cost, hardware complexity, and deployment burden in future agricultural HRI systems. Interpretable results can also enhance trust in future agricultural HRI systems by making model decisions more transparent. To the best of our knowledge, this prefix-specific explainability perspective has not yet been examined in agricultural HAR based on wearable sensors. In previous agriculture-related studies, model interpretation has been applied exclusively after complete-activity classification [21,24].

To address this gap, the present study introduces an explainable early-recognition framework for wearable IMU-based agricultural HRI. Each worker activity is treated as a multivariate temporal sequence. Progressively observed prefixes from 30% to 100% of the activity duration are evaluated using an LSTM-based classifier, ranging from early activity recognition, based on limited movement information, to full-sequence classification. In addition to model performance, explainability is assessed separately at each observation ratio to determine how the importance of body location, sensor type, and signal channel changes as more of the activity sequence becomes available.

Compared with previous agricultural HAR studies based on wearable IMUs, the present work extends the field in two ways:

- It evaluates HAR from cumulative prefixes of each movement sequence, instead of complete activity recordings. This links HAR performance to the amount of motion information available before the human action is completed.
- SHAP-based explainability is applied separately at each observation ratio. This allows the contribution of body locations, sensor modalities, and individual signal channels to be examined from early recognition to full-sequence classification.

Importantly, the study reports a new outdoor wearable-IMU dataset, where the human worker performed the material-handling task and the ground robot served as the receiving platform for the crate, a scenario inspired by prior agricultural studies [17,32–34]. Given the scarcity of publicly available agricultural HRI datasets, the dataset is made publicly available to support reproducibility and future benchmarking.

In summary, the results show that the proposed framework maintained high recognition performance even when only the early part of each activity was observed. The explainability analysis showed that early predictions were mainly supported by trunk-related sensors, namely the chest, cervical region, and lumbar region. At the modality and channel levels, the analysis showed that the model used information from all three IMU modalities, while gyroscope and magnetometer channels from central body locations were among the most influential inputs.

## 2. Materials and Methods

The proposed explainable early activity-recognition framework is summarized as an analysis pipeline in Figure 1. The workflow begins with data acquisition from both wearable IMUs and synchronized video recordings during an outdoor agricultural HRI task. Afterward, video-assisted annotation is conducted to define the six activity classes and assign the corresponding activity boundaries to the synchronized IMU time series. The synchronized recordings are subsequently checked, assembled into multichannel IMU inputs, and then transformed into cumulative activity prefixes, which represent different observation ratios. These prefixes are used to train and evaluate LSTM models under repeated participant-wise train/validation/test splits. Finally, explainability analysis is applied to identify the main signal factors contributing to the model predictions at each observation ratio. The methodological details of the pipeline are provided in the following subsections.

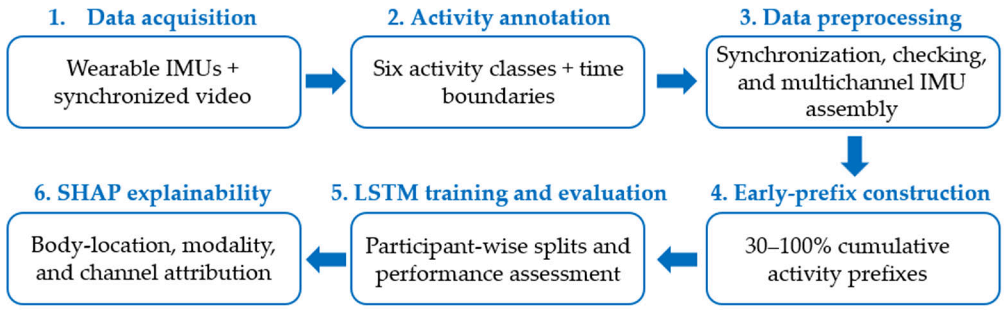



**Figure 1.** Analysis pipeline of the proposed explainable early activity-recognition framework.

### 2.1. Data Acquisition

The dataset was collected during outdoor experimental sessions carried out on a farm in central Greece, representing a collaborative agricultural handling scenario in which a worker and a UGV share the same workspace during crate transport. In particular, we followed the experimental protocol established in [22], namely the same participant group, task sequence, sensor configuration, and recording procedure. However, the dataset analyzed in this research consists entirely of new recordings acquired in a separate experimental session: no IMU recordings or trials from [22] were reused.

In the present study, the trials were performed with the Husky (Clearpath Robotics, Kitchener, ON, Canada) UGV, which had a crate-deposit height of 40 cm. This UGV was selected as a rugged, medium-sized, all-terrain robotic platform that is widely used in agricultural robotics research. The modular architecture of this robotic platform, combined with its ROS compatibility, provides the flexibility needed to integrate different sensors, payloads, and control modules for agricultural experiments [35,36]. This differs from the configuration used in [22], which included a Thorvald platform (SAGA Robotics SA, Oslo, Norway). In that case, the crate was deposited at a height of 80 cm. The lower deposition height used here may introduce variation in the recorded movement signatures, mainly during the final crate-placing phase, since participants had to adapt their trunk and wrist movements to the specific robot configuration. An overview of the IMU placement, six-class task protocol, and video-based annotation of synchronized IMU time series is illustrated in Figure 2.

In summary, twenty healthy volunteers participated in the experiments, consisting of 10 males and 10 females. Their mean age was 30.13 years, with a standard deviation of approximately 4.13 years. Moreover, their mean height and body mass were 1.71 m (SD = 0.10 m) and 70.20 kg (SD = 16.1 kg), respectively. Participants could participate in the experimental sessions only if they had no musculoskeletal injury or surgical intervention during the previous year that could affect their ability to perform the experimental task. Before the beginning of the trials, all participants provided informed consent. The synchronized video recordings were used only for offline activity annotation and are not published or shared. Additionally, all data used in this analysis were de-identified. A five-minute warm-up was also performed before testing to reduce the risk of injury during the lifting and carrying task [37].

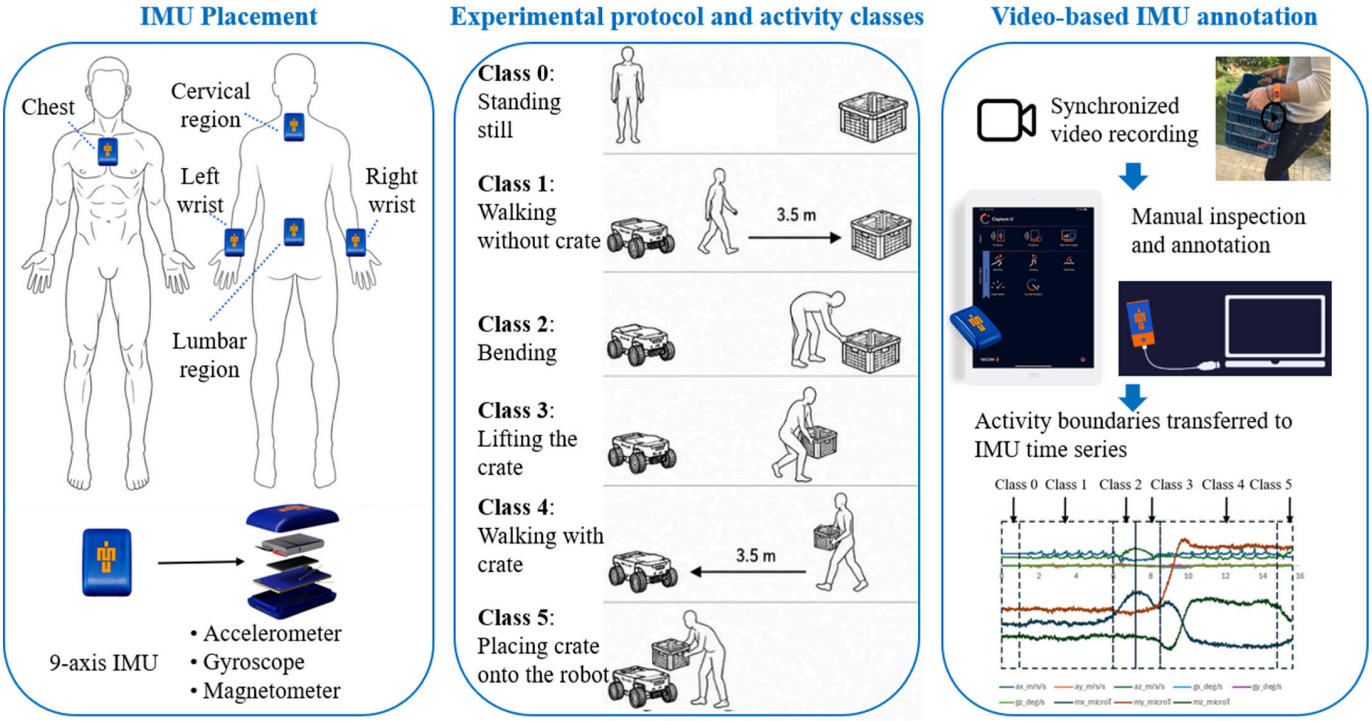



**Figure 2.** Overview of the IMU placement, experimental protocol, and video-based annotation workflow.

Each trial reproduced a simple crate-handling operation within a collaborative harvesting context. As can be seen in the middle section of Figure 2, the participants initially remained motionless until the starting signal was given. They then walked 3.5 m in a straight line without carrying a crate, bent down to approach the crate, lifted it from the ground to an upright standing posture, walked back 3.5 m while carrying the crate, and finally placed it onto the Husky UGV. Two crate-load conditions were included: an empty crate with a tare weight of 1.5 kg and a loaded crate with additional weight plates corresponding to approximately 20% of each participant's body mass. The additional mass was adjusted using 1 kg and 2.5 kg plates. Each participant completed both load conditions three times in randomized order and was allowed to perform the sequence at a self-selected pace.

Human motion was recorded using five Blue Trident inertial measurement units (IMUs; Vicon, Nexus, Oxford, UK). The IMUs were positioned on the chest, cervical region, lumbar region, right wrist, and left wrist. These placements were selected for the purpose of capturing the adjustments in body posture related to the trunk as well as the movements of the upper limbs involved in grasping, lifting, carrying, and placing the crate. The wrist sensors were fixed using Velcro straps. In contrast, the sensors on the chest, cervical region, and lumbar region were attached using double-sided tape. Each IMU included a tri-axial accelerometer, tri-axial gyroscope, and tri-axial magnetometer. This made it possible to record motion data across multiple axes. Data were collected at a sampling frequency of 50 Hz. This frequency was sufficient for recognizing the human activities examined in this study [23,38].

The recordings were stored directly on the IMUs in CSV format, with each file corresponding to one body location. Sensor synchronization was performed using Capture.U software (version 1.3; Vicon Motion Systems Ltd., Oxford, UK), while simultaneous video recording was used to support manual annotation of the continuous task sequence. The video data allowed the start and end of each activity segment to be identified consistently. Activity boundaries were manually annotated from synchronized videos and cross-checked with the corresponding inertial-sensor time series, with ambiguous transitions reviewed before final labeling. For the present study, the activity labels were defined using a six-class scheme. Table 2 summarizes the operational definition of the six activity classes used for manual annotation. Although both empty- and loaded-crate trials were included in the experimental protocol, they were not treated as separate activity classes in this study. Instead, they were retained within the same six activity labels to increase movement variability, following the rationale of [22].

**Table 2.** Operational definitions of the six activity classes used for manual annotation.

| Class | Activity | Operational Definition |
|---|---|---|
| 0 | Standing | The participant remains upright and stationary before initiating the task sequence. |
| 1 | Walking without crate | Activity begins with the onset of gait, when one foot contacts the ground to support body weight while the other prepares to lift for the next step. |
| 2 | Bending | This activity starts when the participant begins lowering the body toward the crate through trunk flexion, knee flexion, or both. |
| 3 | Lifting crate | This activity corresponds to the transition from a bent posture to an upright standing position while lifting the crate. |
| 4 | Walking with crate | This activity corresponds to gait while carrying the crate and begins once the participant initiates walking after lifting. |
| 5 | Placing crate onto the robot | This activity begins when the participant starts lowering the crate toward the robotic platform, typically through trunk and/or knee flexion, and ends when the crate is placed on the robot. |

### *2.2. Data Preprocessing and Prefix Construction*

#### 2.2.1. Data Quality Checks and Synchronized Signal Assembly

The files belonging to each experimental case were first grouped according to participant, trial, load condition, and body location. Consequently, each complete case consisted of five body-location recordings, corresponding to the chest, cervical region, lumbar region, right wrist, and left wrist. Data integrity was confirmed by verifying the synchronization, temporal alignment, and label consistency across all five IMU sensors. A total of 120 experimental cases were expected from the experimental design (20 participants $\times$ 2 load conditions $\times$ 3 repetitions). Cases were excluded when a technical recording problem, such as an interrupted or incomplete IMU recording, prevented complete five-sensor data from being obtained. Following this procedure, 88 complete cases were retained for subsequent analysis. The retained cases were nearly balanced across load conditions, with 45 empty-crate and 43 loaded-crate cases. All 20 participants remained represented in the final dataset, although the number of retained cases per participant ranged from one to six. This ensured that each retained activity instance included all five body-location recordings required for subsequent multichannel input construction.

Thus, a multichannel IMU signal array was constructed for each segmented activity instance by combining the synchronized recordings from the five body locations. From each IMU, twelve signal streams were retained: the three accelerometer axes ($a_x$, $a_y$, $a_z$), the three gyroscope axes ($g_x$, $g_y$, $g_z$), the three magnetometer axes ($m_x$, $m_y$, $m_z$), and the corresponding acceleration, angular-velocity, and magnetic-field magnitudes. The corresponding magnitudes, denoted as $a$, $g$, $m$, were computed as Euclidean norms of the three respective axes. Therefore, each time step was represented by 60 input channels

(5 body locations × 12 signal streams per IMU). No handcrafted statistical features, such as mean, standard deviation, or root mean square, were extracted.

### 2.2.2. Signal-Preservation Strategy

For the sake of retaining the transient features essential for early HAR, the IMU signals were kept unsmoothed prior to prefix extraction. Common noise-reduction steps, including median and low-pass filtering, were omitted to preserve the abrupt accelerations and rotations characterizing activity onset. Previous IMU-based HAR research on manual material-handling activities has shown that window length and temporal information are important for classification performance [39]. This choice was particularly important for short activity segments. For instance, in our analysis, the minimum full-segment length was 15 samples for both "bending" and "placing crate onto the robot". At the earliest evaluated observation ratio of 30%, these activities could therefore be represented by prefixes of only 5 samples, corresponding to approximately 0.10 s at the 50 Hz sampling rate. Similarly, "lifting crate" had a minimum full-segment length of 34 samples, giving a 30% prefix of 11 samples. Under these conditions, even short smoothing or median-filtering windows could affect a substantial part of the available early signal and potentially attenuate transient movement signatures relevant for classification. Therefore, the verified IMU trajectories were retained without additional smoothing or filtering prior to prefix extraction.

### 2.2.3. Activity-Prefix Extraction and Earliness Definition

No sliding-window or overlapping-window segmentation was applied in this analysis. Instead, each activity instance was represented by cumulative prefixes extracted from the manually annotated onset of the activity. This design was adopted because the objective of the study was to determine how much of an activity sequence must be observed before reliable classification is possible and not to classify arbitrary fixed-length windows from a continuous signal stream.

For a complete activity sequence of length $T$, each prefix was generated by retaining only the first $t$ samples, where:

$$t = \lceil rT \rceil, \tag{1}$$

with $r$ denoting the observation ratio. Observation ratios from 30% to 100% were evaluated in 10% increments, corresponding to 30%, 40%, 50%, 60%, 70%, 80%, 90%, and 100% of the complete activity duration. Lower observation ratios represented early recognition from partial movement information, whereas the 100% condition represented full-sequence classification.

For each observation ratio, earliness was defined as the unobserved proportion of the activity sequence at the prediction time following [27]:

$$E = \frac{T - t}{T} \times 100. \tag{2}$$

In the above equation, $t$ stands for the number of observed samples used for prediction. Hence, the nominal observation ratios of 30–100% correspond to earliness levels of 70–0%, respectively. The 10% and 20% observation ratios were excluded because the shortest activity segments would have produced prefixes with very few samples and potentially unstable sequence representations. The 30% observation ratio was therefore selected as a pragmatic lower bound for the early-recognition analysis rather than as a theoretically defined minimum for representing temporal structure. Although some of the shortest 30% prefixes still contained only a limited number of measured samples, the robustness of this earliest observation condition was further examined through the sensitivity analysis described in Section 2.2.4.

Across the dataset, each activity was represented by 88 segmented instances, corresponding to one segment per complete experimental case. To complement the percentage-based observation ratios with absolute timing information, Table 3 summarizes the activity durations and the corresponding observation times represented by the 30% prefixes. The 30% prefix observation time was calculated using the same prefix-definition rule, $t = \lceil 0.30 \cdot T \rceil$, and converted to seconds using the 50 Hz sampling rate. This provides the absolute time represented by the earliest evaluated observation ratio for each activity.

**Table 3.** Median activity duration and corresponding observation time represented by the 30% prefix.

| Activity | Median Full Duration (s) | Median 30% Observation Time (s) |
|---|---|---|
| Standing | 1.00 | 0.30 |
| Walking without crate | 3.24 | 0.98 |
| Bending | 1.06 | 0.32 |
| Lifting crate | 1.24 | 0.38 |
| Walking with crate | 3.15 | 0.96 |
| Placing crate onto the robot | 1.05 | 0.32 |

It should be noted that the above prefix-generation procedure was performed offline using the manually annotated activity boundaries. Therefore, the activity onset and complete activity duration, $T$, were known during dataset preparation, and the observation ratios were calculated as proportions of the complete activity segment. Accordingly, the present protocol evaluates early classification of pre-segmented activity sequences, rather than complete online early-activity recognition from a continuous IMU stream.

2.2.4. Fixed-Length Resampling and Tensor Construction

To provide a consistent temporal input dimension for the main LSTM training pipeline, all extracted prefixes were transformed to a fixed length before model training. This step was performed after true-prefix extraction, ensuring that the model received only the information available up to the selected observation ratio. Resampling was performed independently for each channel using linear interpolation over normalized prefix time. This operation standardized tensor length but did not introduce future information beyond the observed prefix. Importantly, this step was used to standardize the temporal length of the already extracted raw-channel prefixes and should not be interpreted as filtering or hand-crafted feature extraction. However, because linear interpolation constructs intermediate values between measured samples, it may influence the apparent temporal morphology of very short prefixes. Therefore, the effect of this representation choice was further examined through a sensitivity analysis at the 30% observation ratio, which represents the earliest and shortest-prefix condition.

The resampling procedure preserved the temporal order of the observed prefix while standardizing its length across activities, participants, and trials. In the present workflow, each prefix was resampled to 100 time steps. After resampling, each prefix was represented as a three-dimensional input tensor with dimensions corresponding to samples, time steps, and input channels. The input-channel dimension consisted of the 60 IMU-derived streams obtained from the five body locations. The corresponding target variable was encoded using integer labels for the six activity classes summarized in Table 2.

In the sensitivity analysis, the original native-length 30% prefixes were used directly, without fixed-length interpolation. The same participant-wise train/validation/test splits as in the main 30% experiment were retained. For each split, channel-wise standardization was fitted only on the training prefixes and then applied to the validation and test prefixes. The standardized variable-length prefixes were then zero-padded to 100 time steps and processed using the same lightweight LSTM configuration described in Section 2.3, with the addition of a masking layer before the LSTM so that padded timesteps were ignored.

2.2.5. Participant-Wise Splitting and Leakage-Safe Standardization

To evaluate generalization to unseen operators, a repeated participant-wise train/validation/test splitting protocol was used. Thus, splitting was not performed at the prefix or segment level. Instead, it was performed at the participant level to ensure that data from the same participant were never shared across training, validation, and test subsets within the same repetition. This prevented subject-specific movement patterns from leaking into the evaluation process. A total of 20 repeated splits were generated from the 20 participants. In other words, in each split, 16 participants were assigned to training, 2 to validation, and 2 to testing. The split design was balanced so that, across the 20 repetitions, each participant appeared 16 times in the training set, 2 times in the validation set, and 2 times in the test set.

The same participant-wise splits were used for all observation ratios from 30% to 100%. The repeated splits are interpreted as overlapping participant-wise split configurations rather than as 20 statistically independent experimental repetitions. Therefore, the mean and standard deviation values across splits describe performance variability across repeated unseen-participant evaluations.

Feature standardization was performed in a leakage-safe manner during LSTM tensor preparation. For each split and observation ratio, channel-wise z-score scaling was fitted using only the training subset. The mean and standard deviation were computed across the sample and time-step dimensions of the training tensor for each of the 60 input channels. The same training-derived scaling parameters were then applied to the corresponding training, validation, and test tensors. This ensured that validation and test data did not influence the scaling procedure before LSTM training, validation, and final testing.

*2.3. LSTM-Based Early Activity Recognition*

An LSTM classifier was used to perform early HAR from the multichannel IMU prefixes. The main analysis used a fixed lightweight LSTM architecture. Specifically, the model consisted of one LSTM layer with 64 units, followed by a dropout layer with a dropout rate of 0.25, a fully connected layer with 32 rectified linear unit (ReLU) neurons, a second dropout layer with a dropout rate of 0.20, and a final softmax layer with six output neurons corresponding to the six activity classes. The model was trained using the Adam optimizer with a learning rate of 0.001 and sparse categorical cross-entropy loss. Training was performed with a batch size of 32 for a maximum of 60 epochs. Early stopping was applied based on validation loss, with a patience of 8 epochs, and the model weights from the best validation-loss epoch were retained for final testing. In addition, the learning rate was reduced during training when the validation loss plateaued. This lightweight configuration was selected to provide sufficient recurrent capacity for modelling the 100-step multichannel sequences while limiting model complexity relative to the available participant-level dataset size. Early stopping, dropout regularization, and learning-rate reduction were used to reduce overfitting risk and improve training stability.

No exhaustive hyperparameter optimization was performed, because the primary objective of the study was not to maximize performance through architecture search, but to evaluate how recognition performance changed as a function of the observed proportion of the activity sequence. Keeping the main architecture fixed across all observation ratios also helped isolate the effect of the amount of available movement information by avoiding observation-ratio-specific model tuning. To partially assess the effect of model capacity, a limited sensitivity analysis was performed using smaller and larger lightweight LSTM configurations. Specifically, two additional architectures were tested: a smaller model with

https://doi.org/10.3390/info17090865

32 LSTM units and 16 dense units and a larger model with 128 LSTM units and 64 dense units. The analysis was conducted using all 20 participant-wise split configurations at three representative observation ratios, namely 30%, 70%, and 100%. The quantitative results of this sensitivity analysis are reported in Section 3.1.

As an additional modality-level robustness analysis, the selected LSTM architecture was retrained and evaluated at the 30% observation ratio after removing all magnetometer-derived channels from the five IMUs. Performance was compared with that of the full multimodal model. The quantitative results of this sensitivity analysis are reported in Section 3.3.1.

Separate LSTM models were trained for each participant-wise split and observation ratio. The analysis included 20 repeated participant-wise splits and eight observation ratios from 30% to 100%. Thus, a total of 160 split-specific and observation-ratio-specific LSTM models were trained in the main analysis. For each model, the training subset was used to fit the LSTM model parameters. The validation subset was used to monitor training, apply early stopping based on validation loss, and retain the best-performing training epoch.

Final evaluation was performed only after training, using the saved best LSTM model for each split and observation ratio. Each model was evaluated once on the corresponding untouched test set, which consisted of participants not used during training or validation. Performance was assessed using macro-F1 as the primary overall evaluation metric, with accuracy providing a complementary measure of classification performance. Class-specific F1-scores and confusion matrices were used for activity-level analysis.

In particular, accuracy was defined as the proportion of correctly classified test samples among all test samples:

$$\text{Accuracy} = \frac{N_{\text{correct}}}{N}. \tag{3}$$

In Equation (3), $N_{\text{correct}}$ is the number of correctly classified test samples, while $N$ is the total number of test samples. For each activity class $k$, precision, recall, and F1-score were computed from the corresponding true-positive $(\text{TP}_k)$, false-positive $(\text{FP}_k)$, and false-negative $(\text{FN}_k)$ counts:

$$\text{Precision}_k = \frac{\text{TP}_k}{\text{TP}_k + \text{FP}_k}, \tag{4}$$

$$\text{Recall}_k = \frac{\text{TP}_k}{\text{TP}_k + \text{FN}_k}, \tag{5}$$

$$\text{F1}_k = \frac{2 \cdot \text{Precision}_k \cdot \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k}. \tag{6}$$

Macro-F1 was computed as the unweighted mean of the class-specific F1-scores:

$$\text{macro-F1} = \frac{1}{K}\sum_{k=1}^{K} \text{F1}_k, \tag{7}$$

where $K$ is the number of activity classes. In our case, $K = 6$.

In the present analysis, macro-F1 was treated as the primary performance indicator, since it gives equal weight to all activity classes. Therefore, it is appropriate for evaluating recognition performance without allowing classes with higher support to dominate the overall estimate. Confusion matrices were also generated to examine activity-specific misclassification patterns. For each observation ratio, final performance was summarized across the 20 repeated participant-wise test evaluations as the mean and standard deviation. This allowed the evolution of recognition performance to be examined from early prefixes to full-sequence classification. To account for the repeated-participant structure, for each participant and observation ratio, macro-F1 was calculated from the held-out test predictions and averaged across the two split configurations in which that participant appeared in the test set. The resulting participant-level macro-F1 values were compared across the eight observation ratios using the non-parametric Friedman test. Because the same split IDs were retained across observation ratios, paired split-level differences between the 100% observation ratio and each earlier observation ratio were also calculated for macro-F1 and accuracy and are reported descriptively using mean differences and 95% confidence intervals. The corresponding results are presented in Section 3.1.1.

### 2.4. Explainability Analysis

Explainable Artificial Intelligence (XAI) constitutes a set of methodologies that make the predictions of otherwise complex "black-box" machine learning models more interpretable [40,41]. It aims to identify the input features, signals, or patterns that contribute most strongly to a given model decision. SHAP is a widely used tool within the XAI ecosystem. It has emerged as a leading method for enhancing model interpretability in several sectors, including agriculture [42].

In the present study, explainability analysis was conducted after model training and testing to identify which IMU signals contributed most to the LSTM predictions at each observation ratio. In particular, the SHAP GradientExplainer was applied to the trained LSTM models, as this method is suitable for differentiable neural-network architectures. Similar SHAP-based attribution analyses have recently been used to interpret deep sequence models in wearable-sensor locomotion recognition and to identify influential sensors and input signals [43]. In brief, GradientExplainer estimates expected gradient attributions, which extend the integrated gradients approach by using samples from a background dataset to approximate SHAP values relative to a reference distribution [44,45]. For each participant-wise split and observation ratio, the background samples were selected only from the corresponding training subset, so that validation and test data were not used to define the SHAP reference distribution.

SHAP values were computed for held-out test samples to interpret the behavior of the trained models on unseen participants. The test samples were used only for post hoc interpretation and not for model training, early stopping, or model selection. The main attribution analysis focused on correctly classified test samples. For each split and observation ratio, the SHAP background consisted of five training samples per activity class (30 samples in total), while up to five correctly classified test samples per class were selected for explanation. When fewer than five correctly classified samples were available for a given class, all available samples from that class were used. Because the correctly classified samples available for explanation could differ across observation ratios, the composition of the explained sample set could also vary.

The resulting SHAP summaries characterize the attribution patterns associated with successful model predictions. They do not capture the full behavior of the classifier, as misclassified samples were excluded from the main attribution analysis. For each explained sample, SHAP values were obtained for the LSTM input tensor consisting of 100 prefix time steps and 60 IMU channels. Because the model produced a six-class softmax output, class-specific attributions were considered. For correctly classified samples, the SHAP values corresponding to the predicted class were retained, which was equivalent to using the true activity class.

Feature importance was summarized using the mean absolute SHAP value. Specifically, the SHAP values associated with each input channel were summarized as a single mean absolute attribution value for each split and observation ratio before normalization.

For each participant-wise split $s$ and observation ratio $r$, the relative contribution of channel $j$ was calculated as:

$$RC_{j,s}^{(r)}(\%) = 100\frac{I_{j,s}^{(r)}}{\sum_{m=1}^{60} I_{m,s}^{(r)}}, \tag{8}$$

where $I_{j,s}^{(r)}$ denotes the mean absolute class-specific SHAP attribution of channel $j$. Thus, the relative contributions of the 60 input channels sum to 100% within each split and observation ratio. For body-location and sensor-type analyses, the relative contributions of the corresponding channels were summed. The same normalization was applied separately within each activity class, and the resulting split-specific contributions were averaged across the 20 participant-wise splits. This analysis enabled us to identify which body locations and types of sensor signals were most influential for the model predictions from early prefixes to full-sequence classification.

## 3. Results

### *3.1. Overall Early Activity-Recognition Performance and Robustness*

#### 3.1.1. Early-Recognition Performance

The overall test performance across observation ratios is summarized in Table 4. Each row corresponds to one observation ratio, from 30% to 100% of the complete activity sequence, together with the corresponding earliness level. As mentioned above, macro-F1 was used as the primary performance indicator, with accuracy providing a complementary measure of overall classification correctness. For each observation ratio, the result is provided as a mean ± standard deviation (SD) across the 20 participant-wise test splits. Hence, the mean values show the average performance, while SD expresses how much the performance varied when the unseen test participants changed. Notably, even at 30% observation, or 70% earliness, the model reached a macro-F1 score of 0.9314 ± 0.0702. The corresponding accuracy was equal to 0.9327 ± 0.0683. This illustrates that reliable HAR was possible from the early part of the movement sequence. Increasing the observation ratio did not result in a monotonic improvement, since both macro-F1 and accuracy remained within a narrow range across all ratios. The highest macro-F1 was observed at 90% observation (0.9500 ± 0.0797), with the corresponding highest accuracy at this ratio (0.9522 ± 0.0739). The full-sequence condition, namely 100% observation, produced a macro-F1 of 0.9481 ± 0.0695 and an accuracy of 0.9499 ± 0.0653.

**Table 4.** Summary of overall LSTM test performance at different observation ratios over the 20 repeated participant-wise test splits.

| Observation Ratio (%) | Earliness (%) | Macro-F1 | Accuracy |
|---|---|---|---|
| 30 | 70 | 0.9314 ± 0.0702 | 0.9327 ± 0.0683 |
| 40 | 60 | 0.9446 ± 0.0808 | 0.9455 ± 0.0789 |
| 50 | 50 | 0.9410 ± 0.0825 | 0.9421 ± 0.0799 |
| 60 | 40 | 0.9389 ± 0.0894 | 0.9403 ± 0.0858 |
| 70 | 30 | 0.9498 ± 0.0733 | 0.9502 ± 0.0729 |
| 80 | 20 | 0.9400 ± 0.0854 | 0.9428 ± 0.0782 |
| 90 | 10 | 0.9500 ± 0.0797 | 0.9522 ± 0.0739 |
| 100 | 0 | 0.9481 ± 0.0695 | 0.9499 ± 0.0653 |

To determine whether these differences reflected a systematic effect of observation ratio, the participant-level repeated-measures analysis described in Section 2.3 showed no significant overall effect on macro-F1, $\chi^2(7) = 6.07$, $p = 0.532$. Thus, there was no evidence of a systematic change in participant-level macro-F1 across the eight evaluated observation ratios.

The paired split-level comparisons, performed as described in Section 2.3, are summarized in Table 5. Differences were calculated as the 100% observation ratio minus each earlier observation ratio. For the main comparison between 30% and 100% observation, the mean paired differences were 0.0167 for macro-F1 and 0.0173 for accuracy. The corresponding descriptive 95% confidence intervals were −0.0117 to 0.0451 and −0.0090 to 0.0436, respectively. Thus, the observed difference between the earliest evaluated observation ratio and the full-sequence condition was small and not clearly separated from zero. This supports the interpretation that 30% prefixes provided performance close to full-sequence classification. These results should not be interpreted as demonstrating formal statistical equivalence or non-inferiority between the 30% and 100% observation conditions.

**Table 5.** Paired split-level differences between 100% observation and earlier observation ratios, reported as mean differences with descriptive 95% confidence intervals.

| Observation Ratio Compared with 100% | Δ Macro-F1, Mean [95% CI] | Δ Accuracy, Mean [95% CI] |
|---|---|---|
| 30% | 0.0167 [−0.0117, 0.0451] | 0.0173 [−0.0090, 0.0436] |
| 40% | 0.0035 [−0.0171, 0.0240] | 0.0044 [−0.0165, 0.0254] |
| 50% | 0.0071 [−0.0098, 0.0241] | 0.0078 [−0.0090, 0.0246] |
| 60% | 0.0093 [−0.0118, 0.0304] | 0.0096 [−0.0112, 0.0303] |
| 70% | −0.0016 [−0.0133, 0.0100] | −0.0002 [−0.0122, 0.0118] |
| 80% | 0.0081 [−0.0111, 0.0274] | 0.0072 [−0.0105, 0.0248] |
| 90% | −0.0018 [−0.0133, 0.0096] | −0.0023 [−0.0131, 0.0085] |

Figure 3 further illustrates the distribution of test macro-F1 scores for the 20 repeated participant-wise splits for each observation ratio. In contrast to Table 4, the boxplots show the internal distribution of the split-level results. Specifically, in each boxplot, the orange horizontal line signifies the median macro-F1, while the green triangle indicates the mean macro-F1. In addition, the box represents the interquartile range, corresponding to the middle 50% of split-level results. Circles show the outlier split-level results, while the whiskers show the non-outlier range. As shown in Figure 3, most splits achieved high macro-F1 values for all observation ratios. This supports the stability of the model. However, a small number of lower-performing outlier splits were also observed for all observation ratios. This indicated that some combinations of unseen test participants resulted in lower classification performance. In several observation ratios, the mean macro-F1 was lower than the median. This reflects the impact of the outliers. Overall, the boxplots indicate encouraging generalization to unseen participants within the present experimental cohort. Nevertheless, the lower-performing outlier splits indicate that recognition performance was not completely uniform across participants. This variability can be attributed to differences in how each participant performed the task, such as discrepancies in gait pattern, ergonomic approach taken during bending, lifting, crate handling, and placing the crate onto the robot. Therefore, the boxplots provide useful information beyond the mean values.

#### 3.1.2. Sensitivity to Input Representation and LSTM Model Capacity

In order to examine whether the earliest recognition result was affected by the fixed-length interpolation procedure, an additional sensitivity analysis was performed at the 30% observation ratio, as discussed in Section 2.2.4. This sensitivity analysis showed that the early-recognition result was not strongly dependent on the fixed-length interpolation procedure. Using the native variable-length 30% prefixes with zero padding and masking, the model achieved an accuracy of 0.9524 ± 0.0531 and a macro-F1 score of 0.9512 ± 0.0558 across the 20 participant-wise splits. These values were close to, and slightly higher than, the corresponding interpolated-prefix result at the 30% observation ratio, where accuracy was 0.9327 ± 0.0683 and macro-F1 was 0.9314 ± 0.0702. This indicates that the main early-recognition conclusion at the 30% observation ratio was preserved under an alternative variable-length representation using padding and masking.

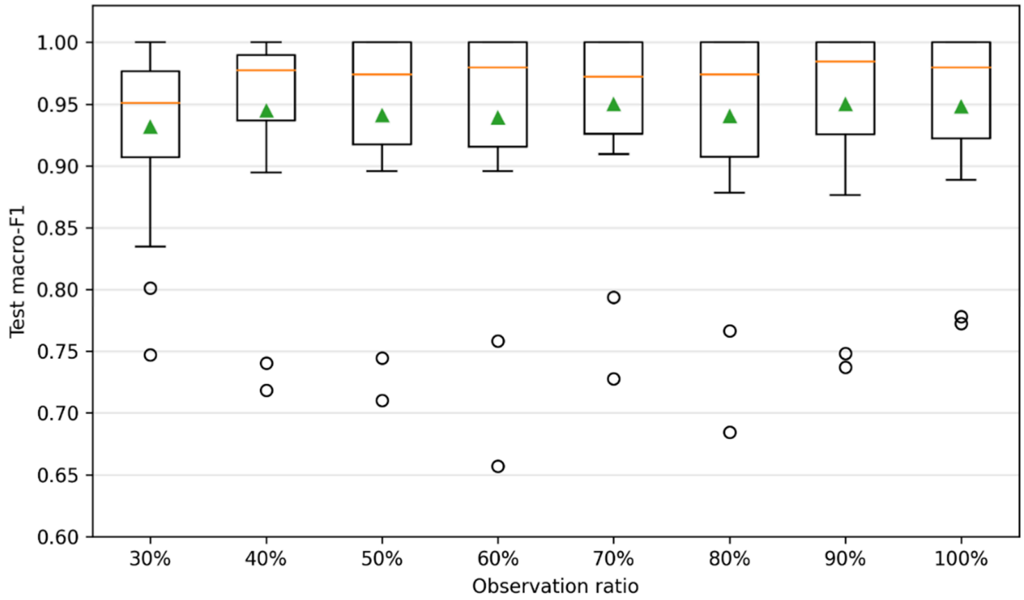



**Figure 3.** Boxplots of split-wise test macro-F1 for each observation ratio; orange lines indicate medians, green triangles indicate means, boxes indicate interquartile ranges, whiskers indicate non-outlier ranges, and circles indicate outlier split-level results.

To quantitatively assess the effect of LSTM model capacity, the limited architecture-sensitivity analysis described in Section 2.3 was evaluated at three representative observation ratios, namely 30%, 70%, and 100%. The results are reported in Table 6. Across the smaller (32 LSTM units), main (64 LSTM units), and larger LSTM configurations (128 LSTM units), macro-F1 values remained broadly comparable at all three observation ratios. These results indicate that the overall early-recognition performance trend was broadly preserved across the tested model capacities and was not strongly dependent on the selected LSTM size.

**Table 6.** LSTM architecture-sensitivity analysis at representative observation ratios. Values are reported as macro-F1 mean ± SD.

| LSTM Units | Dense Units | 30% Macro-F1 | 70% Macro-F1 | 100% Macro-F1 |
|---|---|---|---|---|
| 32 | 16 | 0.9398 ± 0.0623 | 0.9420 ± 0.0897 | 0.9265 ± 0.0883 |
| 64 * | 32 | 0.9314 ± 0.0702 | 0.9498 ± 0.0733 | 0.9481 ± 0.0695 |
| 128 | 64 | 0.9306 ± 0.0836 | 0.9519 ± 0.0761 | 0.9437 ± 0.0619 |

* Main architecture.

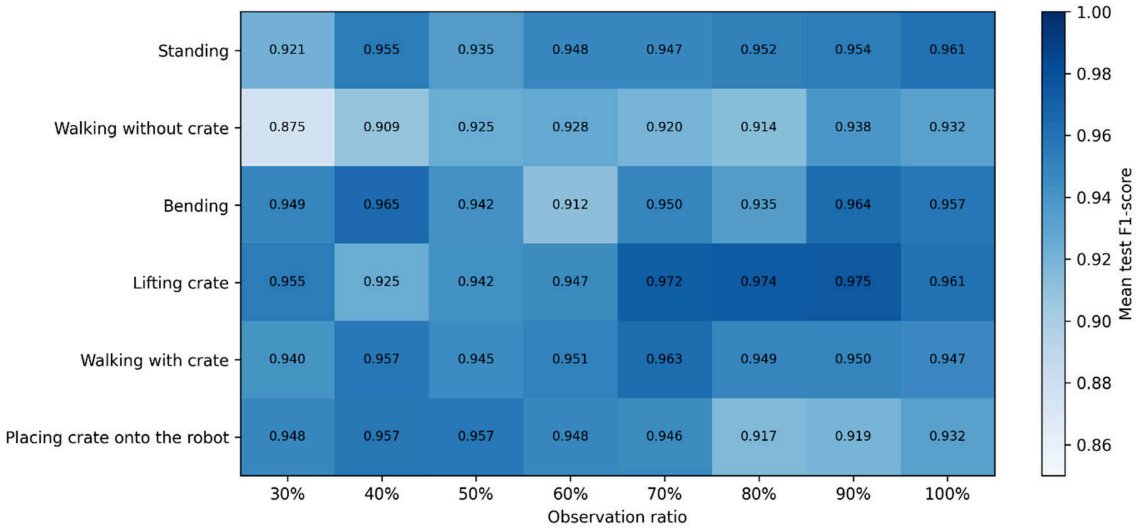
### 3.1.3. Computational Feasibility

The computational feasibility of the proposed framework was assessed using an offline inference benchmark of the selected lightweight LSTM architecture and the fixed-length 100 × 60 IMU-prefix tensors. The model contained 34,278 trainable parameters, corresponding to an approximate float32 parameter memory of 0.13 MB, while the mean saved model size was 0.423 MB. Under CPU-only TensorFlow execution, the mean single-prefix inference latency was 56.60 ± 4.77 ms. When test prefixes were processed in batches of 32, the mean processing time was 1.49 ± 0.31 ms per prefix. These results indicate that the selected LSTM architecture has a relatively low model-level computational cost. However, this benchmark reflects only offline model-level inference on preprocessed fixed-length IMU-prefix tensors. It does not include end-to-end deployment components such as online IMU streaming, synchronization, prefix construction, scaling/preprocessing, ROS communication, robot-control execution, or safety-control latency.

### *3.2. Activity-Level Results*

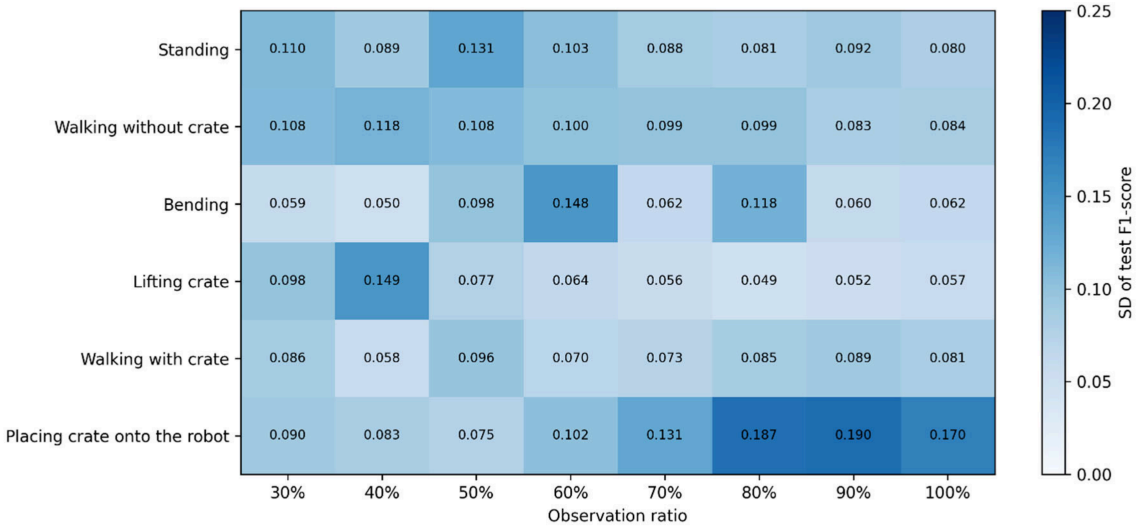
As far as the mean activity-specific test F1-scores are concerned, high scores were obtained for all six activities, as depicted in Figure 4a. This shows that the proposed LSTM-based framework was able to distinguish the main activity classes even from early activity prefixes. Nevertheless, the effect of increasing the observation ratio varied among activities and did not follow a uniform or monotonic pattern. In particular, "standing" displayed an overall upward tendency, as it increased from 0.921 at 30% observation to 0.961 at 100% observation. This improvement was accompanied by slight fluctuations in the interim observations. "Walking without crate" demonstrated the lowest performance at 30% observation (0.875), while it generally benefited from additional temporal information. "Bending", "lifting crate", and "walking with crate" remained consistently well recognized during all observation ratios, with only minor variations. In contrast, "placing crate onto the robot" achieved high performance at the early and intermediate observation ratios, but showed a moderate decrease at 80% and 90% observation. As a complementary heatmap, Figure 4b presents the variability in activity-specific test F1-score among the repeated participant-wise evaluations. In general, variability remained relatively limited for most activities, indicating stable model performance. The highest variability was observed for "placing crate onto the robot" at the later observation ratios, particularly from 80% to 100%, which may reflect variations in posture, wrist motion, and crate alignment.

It should also be noted that the same observation ratio does not correspond to the same absolute amount of temporal information for all activities, since the activity durations differed. Therefore, a 30% prefix may contain a larger number of samples for longer activities, such as walking with or without the crate, than for shorter transitional activities, such as bending. This may have also contributed to the activity-specific differences observed in Figure 4.




**(a)**

**Figure 4.** *Cont.*




(**b**)

**Figure 4.** Heatmaps showing the (**a**) mean activity-specific test F1-score and (**b**) variability in activity-specific test F1-score across repeated participant-wise evaluations for each activity and observation ratio.

With the aim of further investigating the relationship between activity-level performance and misclassification patterns, Figure 5 presents the aggregated row-normalized confusion matrix at an observation ratio of 30%. This ratio was selected because it represents the earliest evaluated recognition point and, thus, the most relevant condition for assessing early activity classification. For this analysis, the raw prediction counts from all 20 participant-wise test splits were first aggregated, and the resulting confusion matrix was then normalized by row.

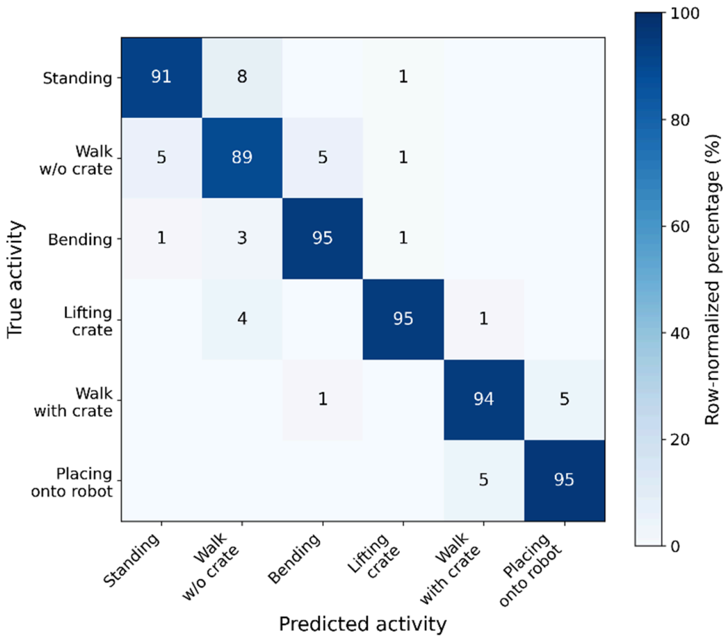



**Figure 5.** Aggregated row-normalized confusion matrix at 30% observation.

Overall, the confusion matrix of Figure 5 shows a strong diagonal structure, with correct classification rates ranging from approximately 89% to 95%. Hence, limited misclassification was observed at this early stage. The remaining errors were mainly noted between activities with temporally adjacent or biomechanically similar movement characteristics.

One confusion pattern was between “standing” and “walking without crate”. Specifically, 7.95% of standing samples were classified as “walking without crate”, while 5.11% of walking-without-crate samples were classified as “standing”. This may reflect the subtle inertial differences between stationary posture, postural adjustments, and early gait initiation before a clear walking pattern emerges. A second confusion pattern involved “walking without crate” and the adjacent crate-handling phases, particularly “bending” and, to a lesser extent, “lifting crate”. This may be explained by the fact that early prefixes captured transitional movements before the more distinctive bending or lifting patterns became fully expressed. A further relevant confusion occurred between the two loaded-task phases, namely “walking with crate” and “placing crate onto the robot”. This is biomechanically plausible because both activities involve load handling, constrained arm posture, trunk stabilization, and movement near the robot platform.

In addition to activity-level misclassifications, participant-wise performance variation was examined using the pooled test predictions across all observation ratios. Participant-wise accuracy ranged from 71.88% to 100%, while participant-wise macro-F1 ranged from 0.7117 to 1. This indicates that the model did not perform uniformly across all participants, despite the high overall performance. Moreover, most errors were concentrated in a small number of anonymized participants, with three participants accounting for approximately 67.2% of all misclassifications. This suggests that part of the remaining error may be related to inter-individual variability in movement execution rather than only to class-level ambiguity. Possible contributing factors include differences in walking speed, bending and lifting strategy, crate-handling posture, trunk stabilization, and small variations in wearable-sensor placement.

### *3.3. Interpretation of LSTM Predictions Using SHAP Values*

#### 3.3.1. SHAP Contribution Patterns over the Full Range of Observation Ratios

In this section, our focus is shifted to interpreting the contribution of the wearable IMU inputs to the LSTM predictions following the analysis detailed in Section 2.4. The following biomechanical interpretations should be considered plausible interpretations of the learned attribution patterns and not direct causal biomechanical evidence [46].

First, we examine the body locations that contributed most to the model’s predictions. Figure 6a shows the SHAP relative contribution of each body location for the observation ratios studied. Overall, the contribution pattern remained relatively stable from 30% to 100% observation. The chest sensor showed the highest contribution at all observation ratios, ranging from 23.4% to 24.5%. The cervical-region and lumbar-region sensors also contributed substantially. Their corresponding contributions ranged from 19.8% to 21.1% and from 20.4% to 21.8%, respectively. The right wrist showed a more moderate contribution, ranging from 18.1% to 19.5%, whereas the left wrist had the lowest contribution, ranging from 15.1% to 16.2%. From a biomechanical perspective, the higher contribution of the chest, cervical region, and lumbar-region sensors is meaningful because these body locations are directly relevant to the examined task sequence, as highlighted in previous studies [22–24].

The SHAP contributions were further grouped by sensor type. As shown in Figure 6b, the contributions at the modality level were relatively balanced for all observation ratios examined. Magnetometer signals showed the highest mean relative contribution, with an average value of 35.5%. The mean contribution of gyroscope signals followed closely with 34.7%, while accelerometer signals showed the lowest mean contribution of 29.8%. These results imply that the present LSTM model does not rely on a single signal modality. Instead, acceleration, angular velocity, and magnetic field signals provide useful information for capturing different aspects of movement execution.

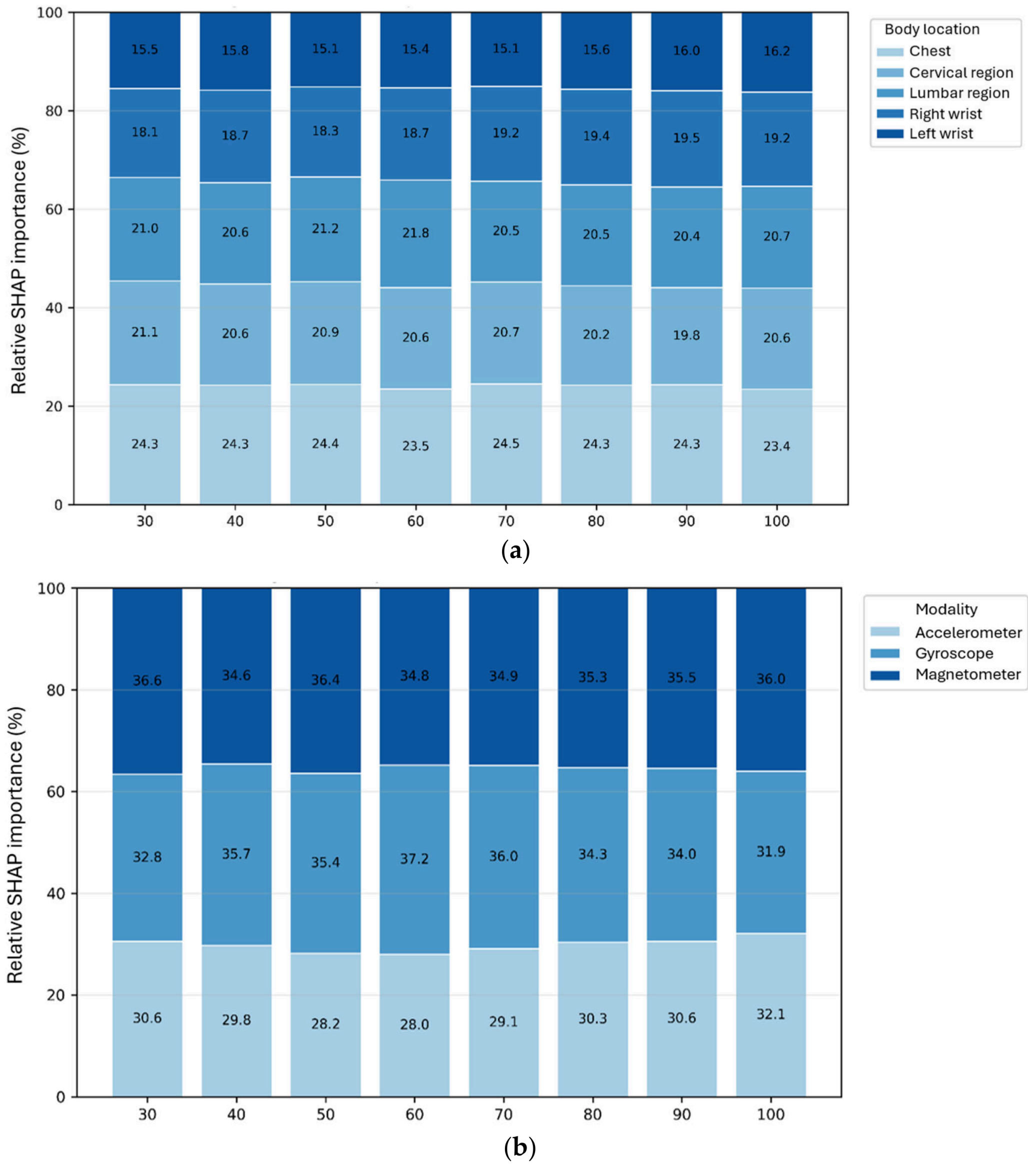



**Figure 6.** Relative SHAP contribution of each (**a**) body location and (**b**) sensor modality for each observation ratio.

As described in Section 2.3, the selected LSTM architecture was retrained and evaluated at the 30% observation ratio after removing all magnetometer-derived channels from the five IMUs. The resulting accelerometer–gyroscope model achieved a macro-F1 of $0.8765 \pm 0.1263$, compared with $0.9314 \pm 0.0702$ for the full multimodal model. Because "walking without crate" and "walking with crate" were performed in opposite directions along the experimental route, these two activities were examined specifically. Without magnetometer inputs, their F1 scores were $0.8571 \pm 0.1259$ and $0.9146 \pm 0.0997$, respectively. Mutual confusion between the two classes remained low, at 0.57% and 1.70%, respectively.

Focusing on the earliest evaluated recognition point at 30% observation, the body-location analysis (Figure 6a) showed that the chest sensor had the highest relative contribution, equal to 24.3%. This was followed by the cervical-region and lumbar-region sensors, with contributions of 21.1% and 21.0%, respectively. The right wrist accounted for 18.1%, while the left wrist had the lowest contribution, at 15.5%. At the modality level, Figure 6b showed that magnetometer signals had the highest contribution, equal to 36.6%, followed by gyroscope signals at 32.8% and accelerometer signals at 30.6%. Overall, the early LSTM predictions followed the same SHAP contribution pattern observed when considering the full set of observation ratios. This confirms the joint importance of trunk sensors and multimodal IMU information for early recognition.

### 3.3.2. Activity-Specific SHAP Contribution Patterns at the Earliest Observation Ratio

Examining SHAP contributions at 30% observation allows the identification of the body locations, sensor modalities, and individual IMU channels that support the model's earliest decisions. Figure 7a,b demonstrate that the relative importance of specific body locations and sensor modalities, respectively, varied by activity when observed at 30% observation.

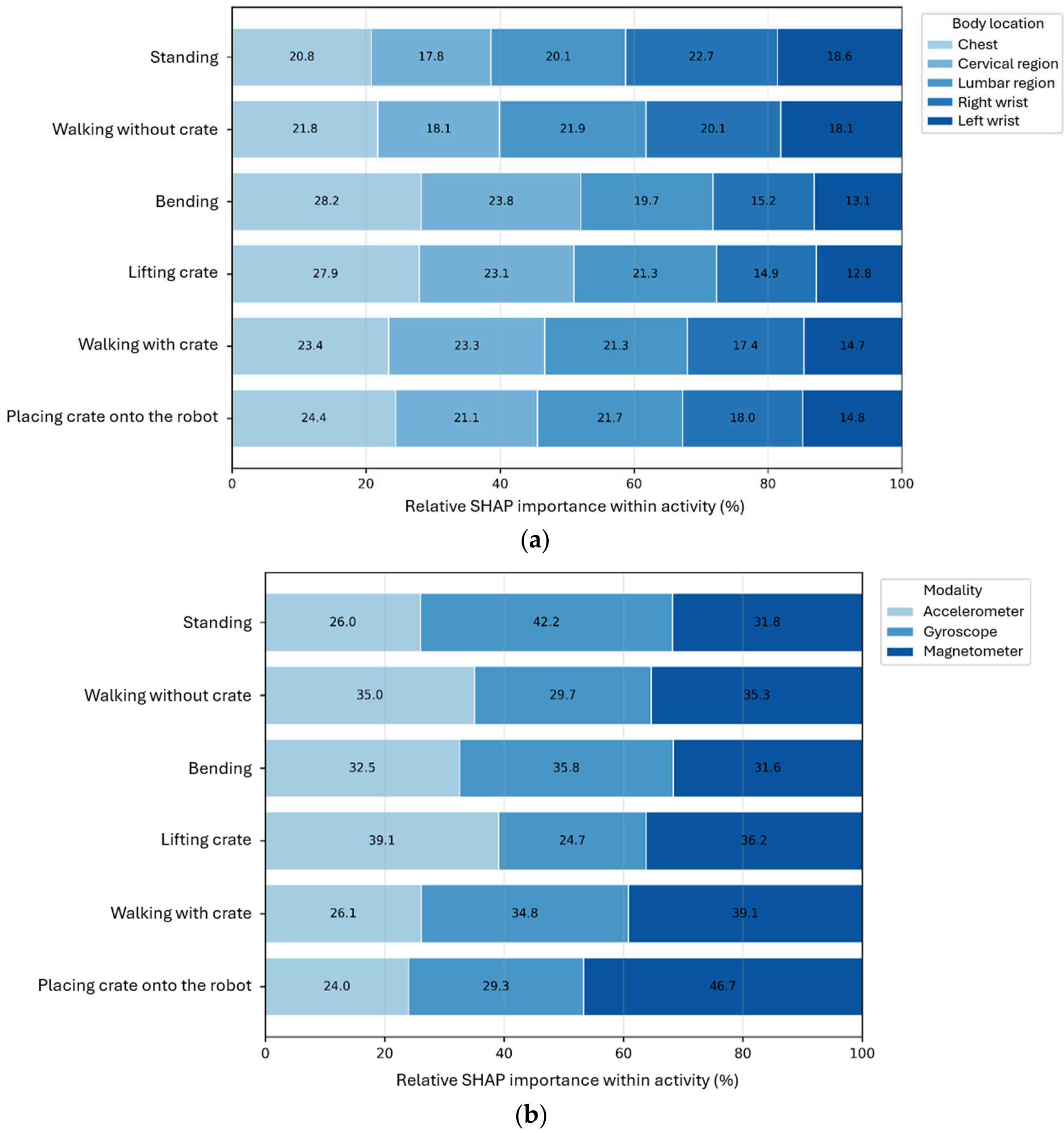



**Figure 7.** Activity-specific relative SHAP contribution patterns at 30% observation for each: (**a**) body location and (**b**) sensor modality within each activity class.

For "standing", the right wrist showed the highest SHAP contribution (22.7%), followed by the chest (20.8%), lumbar region (20.1%), left wrist (18.6%), and cervical region (17.8%). This difference may indicate minor variations, depending on the arm posture, balance adjustments, or preparatory movements of the upper limbs before gait initiation. It may also be associated with participant-specific movement habits. Overall, this difference should be interpreted cautiously and not as evidence of a general dominance of the right wrist over the left wrist. At the modality level, the signals were ranked by contribution as follows: gyroscope (42.2%), magnetometer (31.8%), and accelerometer (26.0%). This suggests that subtle angular-velocity information contributed to distinguishing the early standing phase from movement-related activities. However, the additional information from the other modalities was valuable.

For "walking without crate", the body-location contributions were relatively balanced, with the lumbar region (21.9%) and chest (21.8%) showing the highest values, followed by the right wrist (20.1%). The cervical region and left wrist contributed equally (18.1%). This pattern seems biomechanically plausible, as early walking involves coordinated whole-body motion, including trunk stabilization, lumbar-region movement, and natural arm swing. As shown in Figure 7b, the magnetometer accounted for 35.3%, the accelerometer for 35.0%, and the gyroscope for 29.7%.

For "bending", the chest showed the highest SHAP contribution (28.2%), followed by the cervical region (23.8%) and lumbar region (19.7%). In contrast, the right and left wrists showed lower contributions. This pattern appears biomechanically meaningful, since the early phase of bending is mainly characterized by trunk inclination and upper-body reorientation toward the crate. The high contribution of the chest and cervical-region sensors may therefore indicate changes in upper-trunk posture and spinal alignment. In turn, the lumbar-region contribution may capture lower-trunk motion during flexion. At the modality level, gyroscope signals showed the highest contribution (35.8%), followed by accelerometer (32.5%) and magnetometer signals (31.6%).

The "lifting crate" activity showed similar qualitative patterns to "bending". Specifically, the chest had the highest SHAP contribution (27.9%), followed by the cervical region (23.1%) and lumbar region (21.3%), while the wrists showed lower contributions. This finding is biomechanically reasonable, as the early lifting phase involves a transition from a flexed posture toward a more upright stance. The high contribution of the chest and cervical-region sensors may reflect upper-trunk reorientation and changes in spinal alignment. The lumbar-region contribution may capture lower-trunk involvement during the lifting transition. The accelerometer (39.1%) and magnetometer (36.2%) were the primary contributors to the initial lifting phase, highlighting the importance of movement intensity and body orientation. Rotational dynamics, driven by the gyroscope (24.7%), played a secondary role.

For "walking with crate", the patterns were similar to the above two activities: the chest and cervical region had the highest SHAP contributions, at 23.4% and 23.3%, respectively, while the lumbar region followed with 21.3%. The strong contribution of the chest, cervical region, and lumbar-region sensors may reflect the central role of trunk motion and body alignment during loaded gait. Magnetometer signals showed the highest contribution (39.1%), followed by gyroscope (34.8%) and accelerometer signals (26.1%). This indicates that, during this activity, the model mainly used information related to body orientation and rotational movement, probably because carrying the load reduced free arm motion and made trunk alignment more important.

For "placing crate onto the robot", the chest again showed the highest SHAP contribution at very early observation (24.4%), followed by the lumbar region (21.7%) and cervical region (21.1%). The high contribution of the chest, lumbar-region, and cervical-region sensors may therefore reflect the central role of trunk reorientation and spinal-region movement during the transition from carrying the crate to positioning it onto the robot. At the modality level, magnetometer signals showed the strongest contribution among all activities (46.7%), followed by gyroscope (29.3%) and accelerometer signals (24.0%). This suggests that orientation-related information was especially important during the early placement phase, where the participant had to align the body and crate toward the robot platform.

https://doi.org/10.3390/info17090865

Finally, a channel-level SHAP analysis was performed to interpret the earliest LSTM predictions. Each channel corresponds to a specific combination of body location, sensor modality, and signal axis or resultant magnitude. Since five body locations were used and each IMU provided twelve signals ($ax$, $ay$, $az$, $a$, $gx$, $gy$, $gz$, $g$, $mx$, $my$, $mz$, and $m$), a total of 60 input channels were examined. To facilitate our analysis, these variables are presented as "IMU signal_ body location". Figure 8 presents the top 20 individual IMU channels at 30% observation. In summary, most of the highly influential channels were related to trunk-mounted sensors, namely the chest, cervical region, and lumbar region. The highest-ranked channels were mainly gyroscope and magnetometer signals, including $gy\_chest$, $gy\_cervical\_region$, $gy\_lumbar\_region$, and $my\_cervical\_region$. These findings indicate that angular-velocity and orientation-related information from central body locations strongly supported early activity recognition. Several accelerometer channels, including $az_{chest}$, $az\_lumbar\_region$ and $ax\_cervical\_region$, also appeared as top-ranked inputs. This result shows that linear movement and movement-intensity information contributed to the model's decisions. Finally, channels related to the wrists, like $g\_right\_wrist$ and $g\_left\_wrist$, were also included in the top 20. Yet, they were less dominant compared to channels related to the trunk. This suggests an overall complementary role of wrist sensors.

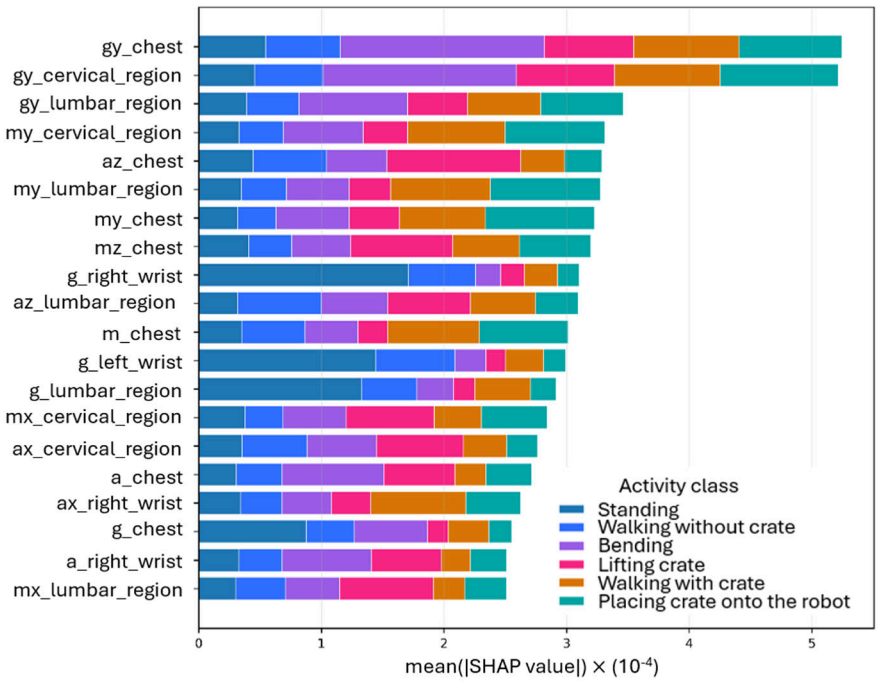



**Figure 8.** Top 20 IMU channels by SHAP importance at 30% observation, with stacked colors showing activity-specific contributions.

Focusing on the two most dominant channel contributions for each activity, the following patterns were observed:

- The recognition of "standing" was mainly supported by $g\_right\_wrist$ and $g\_left\_wrist$, indicating that wrist angular-velocity magnitude helped capture subtle upper-limb adjustments during the initial low-motion phase.
- For "walking without crate", the most important channels were $az\_lumbar\_region$ and $gy\_chest$, showing that lumbar-region acceleration and chest angular velocity contributed to the recognition of early gait-related movement.
- For "bending", the central channels were $gy\_chest$ and $gy\_cervical\_region$ which are consistent with the rotational nature of trunk flexion during the onset of bending.

- For “lifting crate”, *az_chest* and *gy_cervical_region* showed the strongest channel-level contributions, indicating that chest acceleration and cervical-region angular velocity provided important cues for recognizing the initial transition from bending toward lifting.
- For “walking with crate”, *gy_chest* and *gy_cervical_region* showed strong contributions, reflecting the importance of trunk and upper-body rotational motion during loaded gait.
- For “placing crate onto the robot”, *gy_cervical_region* and *my_lumbar_region* showed the strongest channel-level contributions, showing that upper-body angular motion and lumbar-region magnetic-field information supported the early recognition of crate placement.

It should be noted that additional channel inputs also supported the recognition of each class. However, only the two most influential channels were discussed here for brevity.

## 4. Discussion

The examined scenario was motivated by a collaborative harvesting operation, in which a human operator handles crop crates and a UGV may assist by transporting them to the desired location, similar to previous studies [17,32,33]. In this context, HAR based on wearable sensors can provide an important perception component for HRI. [9,47]. This study extends existing research involving wearable IMUs for human–robot collaboration in the agricultural sector. However, instead of classifying complete activities after the full movement sequence, it considers early partial movement information. This distinction can be decisive in agricultural HRI, where relying on retrospective HAR creates a lag that may prevent robots from responding effectively in real time. Conversely, early recognition could support more proactive robot behavior while human activity is still in progress [48]. This could provide additional time for decision-making, support smoother coordination with the human operator, and enable safer adaptation of robot behavior [6,49].

The strength of the proposed framework lies in the dual benefit it offers. On the one hand, the present LSTM model achieved high recognition performance even when only the first 30% of each activity was available. In absolute terms, the median 30% prefix corresponded to approximately 0.30–0.98 s across the examined activities, providing a practical indication of the temporal information available at the earliest evaluated observation point. These values represent the amount of activity observed in the present offline, pre-segmented analysis and should not be interpreted as end-to-end online decision or robot-response times. The high recognition performance at this early observation point may be related to distinctive movement signatures already being present during the onset of each activity. In particular, the initial phases of gait, trunk flexion, lifting, and load handling may already generate sufficiently distinct inertial patterns for the model to differentiate the examined activities before the complete movement is observed.

The additional robustness analyses further support this early-recognition finding. The paired split-level comparison showed only small differences between the 30% and 100% observation conditions, with the corresponding descriptive 95% confidence intervals spanning zero. Moreover, the early-recognition performance was preserved when native-length prefixes with padding and masking were used instead of fixed-length interpolation. The overall performance trend also remained broadly comparable across the examined smaller, main, and larger LSTM configurations. Nevertheless, performance was not completely uniform across unseen participants, indicating that inter-individual variability remains an important consideration.

The second benefit relates to explainability. The explainability analysis showed that “standing” was mainly supported by wrist angular-velocity magnitude signals, while "walking without crate" was supported by lumbar-region acceleration and chest angular velocity. "Bending" recognition was mainly supported by chest and cervical-region gyroscope signals, while "lifting" was supported by chest acceleration and cervical-region angular velocity. Finally, "walking with crate" was supported by chest and cervical-region gyroscope signals and "placing crate onto the robot" by cervical-region angular velocity and lumbar-region magnetic-field information.

These findings are consistent with previous studies, although the relevant literature covers full-activity classification. In particular, the higher contribution of the sensors on the chest, cervical region, and lumbar region agrees with the results of [22]. In that study, the combination of inertial sensors placed at these body locations yielded the highest F1 score compared to the other placement combinations for the three anatomical positions. Our results also mirror the trends identified by [24], where the SHAP analysis was also presented. That study also stressed the importance of trunk-related sensors, with wrist sensors providing complementary information. It is worth noting that the remarkable contribution of sensors worn on the chest, as shown in Figure 6a, was also reported in [22,23]. A quantitative comparison with closely related agricultural HAR studies provides additional context for the present results. Using all five IMUs, an F1-score of 0.908 was reported in [22], while the framework of [24] achieved a macro-F1 of 0.910. In comparison, the present analysis reached a macro-F1 of 0.948 at 100% observation. Although direct comparison should be interpreted cautiously because of differences in datasets, activity definitions, preprocessing, and evaluation protocols, these results indicate that the proposed framework achieves competitive full-sequence performance.

The aforementioned studies, including the present one, support the value of combining different sensor signals. This is because each modality captures different aspects of the activities: accelerometers capture linear movement, gyroscopes capture angular motion, and magnetometers provide orientation-related information. The magnetometer-removal analysis further supported the value of multimodal sensing, as performance decreased when magnetometer-derived channels were excluded. The low mutual confusion between the two walking classes also suggests that this effect was not solely due to their opposite walking directions. From a practical perspective, these attribution patterns may guide the design of more efficient wearable configurations. For example, the higher contribution of trunk-mounted sensors motivates future sensor-ablation studies of reduced wearable configurations. Such studies could determine whether fewer sensors can maintain classification performance while reducing hardware requirements, implementation costs, and wearable burden [50]. The relatively low model-level computational cost also supports the suitability of the selected lightweight LSTM architecture for further deployment-oriented evaluation.

At the channel level, the present findings show partial agreement with [24]. In that study, magnetometer-related channels were among the top-ranked features, stressing the role of magnetometers in predicting the examined activities. However, while that study emphasized mainly magnetometer-related channels, the present early-prefix analysis draws attention to both gyroscope and magnetometer channels from central body locations. This difference may be related to the different modelling approaches (LSTM instead of an XGBoost classifier) and the use of early prefixes. A different robot configuration was used in [24] with a crate-deposit height of 80 cm, whereas this analysis focused on a lower deposit height of 40 cm. In addition, the present study considered both axis-specific signals and resultant magnitudes, allowing a broader interpretation of the IMU information. The inclusion of resultant magnitudes is also consistent with previous HAR studies, such as [51,52], in which magnitude-derived features have been used as informative inputs for activity classification. In this context, it should be mentioned that, overall, SHAP values should be interpreted as model-attribution results and not as direct causal biomechanical

https://doi.org/10.3390/info17090865

evidence. They indicate how each feature contributed to the model predictions, while the associated biomechanical explanations should be considered plausible interpretations of the learned decision patterns [53].

It is important to note the limitations of this study. First of all, the dataset included 20 participants who were not professional agricultural workers. Consequently, the recorded movement patterns may not fully represent those of experienced workers. These workers may use different walking, bending, lifting, carrying, and crate-placement strategies due to experience, work habits, or repeated exposure to material-handling operations [54]. Therefore, the findings should be interpreted within the context of a controlled outdoor agricultural HRI scenario and not as a fully generalizable validation across broader agricultural populations or more diverse working conditions. Furthermore, our sensor configuration was limited to five IMUs at the specific body locations. This setup captured the main trunk and upper-limb movements of the examined task. However, sensors placed on the thighs, shanks, ankles, or feet could provide more direct information on step initiation, lower-limb coordination, as well as changes in gait caused by carrying a load [55]. A further limitation is that SHAP was applied to the interpolated prefixes. Although the padding-and-masking sensitivity analysis preserved the 30% classification performance, robustness of classification performance to the input representation does not necessarily imply identical SHAP attribution rankings. Therefore, some influence of interpolation on the SHAP attribution rankings, particularly for the shortest prefixes, cannot be excluded. Another limitation is that the study was based on offline analysis. The IMU data were collected, annotated, and then used to train and evaluate the LSTM models. No real-time HRI or online robot response was implemented, as this was outside the scope of the present study. Moreover, the present early-recognition protocol relied on pre-segmented activity sequences and did not include online activity-onset detection, transition detection, or prefix construction without prior knowledge of the final activity duration. As shown by Moysiadis et al. [17], real-time HRI requires additional system integration. This includes online data processing, communication with ROS nodes, translation of recognized human actions into robot commands, navigation control, obstacle avoidance, and safety validation under field conditions.

Based on the above limitations and the encouraging results of this work, future work should include professional agricultural workers, larger participant groups, and additional handling tasks performed under real field conditions. Such an evaluation could validate the ability of the early-recognition framework to maintain performance despite the inherent variability of real-world tasks. This variability includes the effects of worker experience, fatigue, environmental conditions, terrain variability, and task-specific movement strategies. Additional sensor configurations should also be investigated, such as lower-body IMU placements and other sensing sources like electromyography [56], which may provide complementary information. From a practical deployment perspective, future studies should also assess the stability, repeatability, and comfort of wearable-sensor placement during prolonged agricultural work. This is important because changes in attachment tightness, sensor orientation, or sensor displacement may affect signal consistency and recognition reliability [57]. Future studies should examine magnetometer robustness under different walking directions, route orientations, robot positions, and environmental conditions. This would help clarify whether the recorded magnetic information mainly reflects body orientation or experiment-specific heading and local magnetic-field effects.

Future research could also assess whether similar or improved performance can be achieved without increasing system complexity. Comparative evaluation could also include alternative sequence-learning architectures, such as Gated Recurrent Unit (GRU), one-dimensional Convolutional Neural Network (1D CNN), and Temporal Convolutional

Network (TCN) models. Conventional machine-learning baselines, such as XGBoost, could also be evaluated under the same participant-wise protocol. Finally, an important next step could be to integrate the proposed early-recognition model into an online HRI pipeline, where real-time early predictions could be translated into robot actions during field operation. Such integration should include end-to-end reliability assessment under field conditions, including failure detection, confidence-based decision thresholds, safe robot behavior under uncertain predictions, and the possibility of human override. Future SHAP-based analyses could also examine misclassified and low-confidence pre-dictions, as these cases are particularly relevant for assessing model behavior in safety-oriented HRI applications. For practical deployment, worker feedback and usability evaluation would be of major importance [58], as the collaborative system must be practical, inherently safe, and able to adapt to the unpredictable nature of agricultural environments.

## 5. Conclusions

In the present study, an explainable early activity-recognition framework was developed that uses wearable IMU sensors and LSTM models to recognize agricultural material-handling tasks. The main contribution of this work is that it moves from conventional full-activity HAR toward explainable early recognition, showing how accurately agricultural material-handling activities can be identified from partial movement information. It also identifies which body locations, IMU modalities, and individual signal channels support these early decisions. This can help move agricultural HAR from retrospective activity classification toward more proactive robot-assistance systems, where early activity predictions could support robot responses before the human activity is completed.

The main conclusions of the present study are summarized as follows:

- At 30% observation, corresponding to 70% earliness, the model reached a macro-F1 of $0.9314 \pm 0.0702$, which was only slightly below the full-sequence score of $0.9481 \pm 0.0695$. Notably, performance did not increase monotonically as the observation ratio increased. This indicates that the early part of the movement already contained substantial information for classification under the present experimental conditions.
- High activity-level performance was observed for all six activity classes, while limited misclassification was observed, mainly between activities with adjacent or similar movement characteristics.
- SHAP analysis showed that early LSTM predictions were mainly supported by trunk-related sensors, namely the chest, cervical region, and lumbar region. However, wrist sensors provided complementary information, particularly for the "standing" activity.
- The LSTM model did not rely on a single IMU modality. Instead, accelerometer, gyroscope, and magnetometer signals all contributed to early predictions. This stresses the value of combining linear-motion, angular-motion, and orientation-related information for predicting the activities examined in the present study.
- At the channel level, early recognition was mainly supported by gyroscope channels from central body locations, especially *gy_chest*, *gy_cervical_region* and *gy_lumbar_region*. Other important inputs included magnetometer and accelerometer channels, such as *my_cervical_region* and *az_chest*.

**Author Contributions:** Conceptualization, L.B. and D.B.; methodology, L.B., E.S. and R.B.; software, L.B. and E.S.; validation, L.B. and E.S.; formal analysis, L.B. and E.S.; investigation, L.B. and R.B.; data curation, L.B. and E.S.; writing—original draft preparation, L.B.; writing—review and editing, L.B., E.S., R.B. and D.B.; visualization, L.B.; supervision, D.B. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** This study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Ethical Committee under the identification code 1660 on 3 June 2020.

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

**Data Availability Statement:** The de-identified wearable-IMU dataset generated and used in this study is available through the iBO/CERTH Open Datasets platform: https://ibo.certh.gr/open-datasets/ (accessed on 3 September 2026). The synchronized video recordings used for activity annotation are not publicly shared due to participant privacy considerations.

**Conflicts of Interest:** Author Dionysis Bochtis is employed by the company farmB Digital Agriculture S.A. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

## References

1. Hernández, H.A.; Mondragón, I.F.; González, S.R.; Pedraza, L.F. Reconfigurable Agricultural Robotics: Control Strategies, Communication, and Applications. *Comput. Electron. Agric.* **2025**, *234*, 110161. [CrossRef]
2. Almtireen, N.; Ryalat, M. Robotics in Precision Agriculture: Task-, Platform-, and Evaluation-Oriented Review. *Robotics* **2026**, *15*, 81. [CrossRef]
3. Wang, R.F.; Xu, R.; Chee, P.W.; Wang, H.; Li, C. A Review of Visual Navigation for Agricultural Robots in Open Fields and Controlled Environments. *Comput. Electron. Agric.* **2026**, *248*, 111754. [CrossRef]
4. Zhang, Y.; Li, N.; Zhang, L.; Lin, J.; Gao, X.; Chen, G. A Review on the Recent Developments in Vision-Based Apple-Harvesting Robots for Recognizing Fruit and Picking Pose. *Comput. Electron. Agric.* **2025**, *231*, 109968. [CrossRef]
5. Bai, Y.; Zhang, B.; Xu, N.; Zhou, J.; Shi, J.; Diao, Z. Vision-Based Navigation and Guidance for Agricultural Autonomous Vehicles and Robots: A Review. *Comput. Electron. Agric.* **2023**, *205*, 107584. [CrossRef]
6. Benos, L.; Bechar, A.; Bochtis, D. Safety and Ergonomics in Human-Robot Interactive Agricultural Operations. *Biosyst. Eng.* **2020**, *200*, 55–72. [CrossRef]
7. Benos, L.; Moysiadis, V.; Kateris, D.; Tagarakis, A.C.; Busato, P.; Pearson, S.; Bochtis, D. Human–Robot Interaction in Agriculture: A Systematic Review. *Sensors* **2023**, *23*, 6776. [CrossRef] [PubMed]
8. Adamides, G.; Edan, Y. Human–Robot Collaboration Systems in Agricultural Tasks: A Review and Roadmap. *Comput. Electron. Agric.* **2023**, *204*, 107541. [CrossRef]
9. Lytridis, C.; Kaburlasos, V.G.; Pachidis, T.; Manios, M.; Vrochidou, E.; Kalampokas, T.; Chatzistamatis, S. An Overview of Cooperative Robotics in Agriculture. *Agronomy* **2021**, *11*, 1818. [CrossRef]
10. Anastasiou, E.; Ntakos, G.; Kanakari, E.; Bitsika, S.; Gemtou, M.; Katsaragakis, M.; Soudris, D.; Volioti, C.; Arvanitou, E.M.; Folina, M.T.; et al. Applications of Robotics and Extended Reality in Agriculture: A Review. *Smart Agric. Technol.* **2025**, *12*, 101521. [CrossRef]
11. Asif, S.; Callari, T.C.; Khan, F.; Eimontaite, I.; Hubbard, E.M.; Bahraini, M.S.; Webb, P.; Lohse, N. Exploring Tasks and Challenges in Human-Robot Collaborative Systems: A Review. *Robot. Comput. Integr. Manuf.* **2026**, *97*, 103102. [CrossRef]
12. Jalayer, R.; Jalayer, M.; Orsenigo, C.; Tomizuka, M. A Review on Deep Learning for Vision-Based Hand Detection, Hand Segmentation and Hand Gesture Recognition in Human–Robot Interaction. *Robot. Comput. Integr. Manuf.* **2026**, *97*, 103110. [CrossRef]
13. Morio, Y.; Inoue, T.; Tanaka, T.; Murakami, K. Worker Posture Recognition for Understanding Agricultural Worker Behaviors. *Eng. Agric. Environ. Food* **2017**, *10*, 208–222. [CrossRef]
14. Morio, Y.; Tanaka, T.; Murakami, K. Agricultural Worker Behavioral Recognition System for Intelligent Worker Assistance. *Eng. Agric. Environ. Food* **2017**, *10*, 48–62. [CrossRef]
15. Pal, A.; Leite, A.C.; Gjevestad, J.G.O.; From, P.J. A Video-Based Activity Classification of Human Pickers in Agriculture. *arXiv* **2023**, arXiv:2311.10885. [CrossRef]
16. Moysiadis, V.; Katikaridis, D.; Benos, L.; Busato, P.; Anagnostis, A.; Kateris, D.; Pearson, S.; Bochtis, D. An Integrated Real-Time Hand Gesture Recognition Framework for Human–Robot Interaction in Agriculture. *Appl. Sci.* **2022**, *12*, 8160. [CrossRef]
17. Moysiadis, V.; Benos, L.; Karras, G.; Kateris, D.; Peruzzi, A.; Berruto, R.; Papageorgiou, E.; Bochtis, D. Human–Robot Interaction through Dynamic Movement Recognition for Agricultural Environments. *AgriEngineering* **2024**, *6*, 2494–2512. [CrossRef]
18. Son, H.; Lim, J.W.; Park, S.; Park, B.; Han, J.; Kim, H.B.; Lee, M.C.; Jang, K.J.; Kim, G.; Chung, J.H. A Machine Learning Approach for the Classification of Falls and Activities of Daily Living in Agricultural Workers. *IEEE Access* **2022**, *10*, 77418–77431. [CrossRef]

19. Gokul, S.; Dhiksith, R.; Sundaresh, S.A.; Gopinath, M. Gesture Controlled Wireless Agricultural Weeding Robot. In Proceedings of the 2019 5th International Conference on Advanced Computing and Communication Systems, ICACCS, Coimbatore, India, 15–16 March 2019; pp. 926–929. [CrossRef]
20. Sharma, S.; Raval, J.; Jagyasi, B. Neural Network Based Agriculture Activity Detection Using Mobile Accelerometer Sensors. In Proceedings of the 2014 Annual IEEE India Conference (INDICON), Pune, India, 11–13 December 2014. [CrossRef]
21. Martins, D.R.; Cerqueira, S.M.; Santos, C.P. Seeking Optimal and Explainable Deep Learning Models for Inertial-Based Posture Recognition. *Knowl. Based Syst.* **2024**, *306*, 112700. [CrossRef]
22. Benos, L.; Tsaopoulos, D.; Tagarakis, A.C.; Kateris, D.; Bochtis, D. Optimal Sensor Placement and Multimodal Fusion for Human Activity Recognition in Agricultural Tasks. *Appl. Sci.* **2024**, *14*, 8520. [CrossRef]
23. Mekruksavanich, S.; Jitpattanakul, A. Efficient Wearable Sensor-Based Activity Recognition for Human–Robot Collaboration in Agricultural Environments. *Informatics* **2025**, *12*, 115. [CrossRef]
24. Benos, L.; Tsaopoulos, D.; Tagarakis, A.C.; Kateris, D.; Busato, P.; Bochtis, D. Explainable AI-Enhanced Human Activity Recognition for Human–Robot Collaboration in Agriculture. *Appl. Sci.* **2025**, *15*, 650. [CrossRef]
25. Hoffman, G.; Bhattacharjee, T.; Nikolaidis, S. Inferring Human Intent and Predicting Human Action in Human-Robot Collaboration. *Annu. Rev. Control Robot. Auton. Syst.* **2024**, *7*, 73–95. [CrossRef]
26. Yerebakan, M.O.; Hu, B. Human–Robot Collaboration in Modern Agriculture: A Review of the Current Research Landscape. *Adv. Intell. Syst.* **2024**, *6*, 2300823. [CrossRef]
27. Gupta, A.; Gupta, H.P.; Biswas, B.; Dutta, T. Approaches and Applications of Early Classification of Time Series: A Review. *IEEE Trans. Artif. Intell.* **2020**, *1*, 47–61. [CrossRef]
28. Rußwurm, M.; Courty, N.; Emonet, R.; Lefèvre, S.; Tuia, D.; Tavenard, R. End-to-End Learned Early Classification of Time Series for in-Season Crop Type Mapping. *ISPRS J. Photogramm. Remote Sens.* **2023**, *196*, 445–456. [CrossRef]
29. Ki, M.R.; Kim, D.H.; Abdelhamid, M.A.A.; Pack, S.P. Cancer and Aging Biomarkers: Classification, Early Detection Technologies and Emerging Research Trends. *Biosensors* **2025**, *15*, 737. [CrossRef] [PubMed]
30. Wang, T.; Liu, Z.; Wang, L.; Li, M.; Wang, X.V. Data-Efficient Multimodal Human Action Recognition for Proactive Human–Robot Collaborative Assembly: A Cross-Domain Few-Shot Learning Approach. *Robot. Comput. Integr. Manuf.* **2024**, *89*, 102785. [CrossRef]
31. Zhang, X.; Tian, S.; Liang, X.; Zheng, M.; Behdad, S. Early Prediction of Human Intention for Human–Robot Collaboration Using Transformer Network. *J. Comput. Inf. Sci. Eng.* **2024**, *24*, 051003. [CrossRef]
32. Peng, C.; Vougioukas, S.; Slaughter, D.; Fei, Z.; Arikapudi, R. A Strawberry Harvest-Aiding System with Crop-Transport Collaborative Robots: Design, Development, and Field Evaluation. *J. Field Robot.* **2022**, *39*, 1231–1257. [CrossRef]
33. Ye, Y.; Wang, Z.; Jones, D.; He, L.; Taylor, M.E.; Hollinger, G.A.; Zhang, Q. Bin-Dog: A Robotic Platform for Bin Management in Orchards. *Robotics* **2017**, *6*, 12. [CrossRef]
34. Vásconez, J.P.; Auat Cheein, F.A. Workload and Production Assessment in the Avocado Harvesting Process Using Human-Robot Collaborative Strategies. *Biosyst. Eng.* **2022**, *223*, 56–77. [CrossRef]
35. Thayananthan, T.; Zhang, X.; Huang, Y.; Chen, J.; Wijewardane, N.K.; Martins, V.S.; Chesser, G.D.; Goodin, C.T. CottonSim: A Vision-Guided Autonomous Robotic System for Cotton Harvesting in Gazebo Simulation. *Comput. Electron. Agric.* **2025**, *239*, 110963. [CrossRef]
36. Rapado-Rincon, D.; Kootstra, G. Tree-SLAM: Semantic Object SLAM for Efficient Mapping of Individual Trees in Orchards. *Smart Agric. Technol.* **2025**, *12*, 101439. [CrossRef]
37. Ding, L.; Luo, J.; Smith, D.M.; Mackey, M.; Fu, H.; Davis, M.; Hu, Y. Effectiveness of Warm-Up Intervention Programs to Prevent Sports Injuries among Children and Adolescents: A Systematic Review and Meta-Analysis. *Int. J. Environ. Res. Public Health* **2022**, *19*, 6336. [CrossRef] [PubMed]
38. Yamane, T.; Kimura, M.; Morita, M. Effects of Sampling Frequency on Human Activity Recognition with Machine Learning Aiming at Clinical Applications. *Sensors* **2025**, *25*, 3780. [CrossRef] [PubMed]
39. Sochopoulos, A.; Poliero, T.; Ahmad, J.; Caldwell, D.G.; Di Natali, C. Human Activity Recognition Algorithms for Manual Material Handling Activities. *Sci. Rep.* **2025**, *15*, 10954. [CrossRef] [PubMed]
40. Camilleri, M.A. Opening the Black Box: Operational Principles, Tools and Frameworks That Advance Explainable Artificial Intelligence (XAI) Models. *Technol. Forecast. Soc. Change* **2026**, *229*, 124710. [CrossRef]
41. Liu, X.; Huang, D.; Yao, J.; Dong, J.; Song, L.; Wang, H.; Yao, C.; Chu, W. From Black Box to Glass Box: A Practical Review of Explainable Artificial Intelligence (XAI). *AI* **2025**, *6*, 285. [CrossRef]
42. Rajbongshi, A.; Johora, F.T.; Hossain, A.; Sarker, M.S.; Rahman, M.H.; Rahman, M.W.; Alotaibi, F.T.; Moni, M.A. Leveraging Explainable AI for Sustainable Agriculture: A Comprehensive Review of Recent Advances. *Artif. Intell. Rev.* **2026**, *59*, 105. [CrossRef]
43. Coser, O.; Tamantini, C.; Tortora, M.; Furia, L.; Sicilia, R.; Zollo, L.; Soda, P. Deep Learning for Human Locomotion Analysis in Lower-Limb Exoskeletons: A Comparative Study. *Front. Comput. Sci.* **2025**, *7*, 1597143. [CrossRef]

44. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic Attribution for Deep Networks. In Proceedings of the 34th International Conference on Machine Learning (ICML), PMLR 70, Sydney, Australia, 6–11 August 2017; pp. 3319–3328.
45. Shap. GradientExplainer—SHAP Latest Documentation. Available online: https://shap.readthedocs.io/en/latest/generated/shap.GradientExplainer.html (accessed on 21 August 2026).
46. Neirz, P.; Allende, H.; Saavedra, C. Epistemic Frontiers and the Distinction between Causality, Information, and Predictability in Pattern Recognition. *Sci. Rep.* **2026**, *16*, 21803. [CrossRef] [PubMed]
47. Aiello, G.; Catania, P.; Vallone, M.; Venticinque, M. Worker Safety in Agriculture 4.0: A New Approach for Mapping Operator's Vibration Risk through Machine Learning Activity Recognition. *Comput. Electron. Agric.* **2022**, *193*, 106637. [CrossRef]
48. Eirale, A.; Martini, M.; Chiaberge, M. Human Following and Guidance by Autonomous Mobile Robots: A Comprehensive Review. *IEEE Access* **2025**, *13*, 42214–42253. [CrossRef]
49. Buerkle, A.; Eaton, W.; Lohse, N.; Bamber, T.; Ferreira, P. EEG Based Arm Movement Intention Recognition towards Enhanced Safety in Symbiotic Human-Robot Collaboration. *Robot. Comput. Integr. Manuf.* **2021**, *70*, 102137. [CrossRef]
50. Borella, E.; Cakmakci, U.B.; Gottardis, E.; Buratto, A.; Marchioro, T.; Badia, L. Effective Sensor Selection for Human Activity Recognition via Shapley Value. In Proceedings of the 2024 IEEE International Workshop on Metrology for Living Environment (MetroLivEnv), Chania, Greece, 12–14 June 2024; pp. 22–27. [CrossRef]
51. Pilataxi Piltaxi, J.I.; Trujillo Guerrero, M.F.; Benavides Laguapillo, V.C.; Rosales Acosta, J.A. Human Activity Recognition Using an Accelerometer Magnitude Value. In *Applied Technologies; Communications in Computer and Information Science*; 1194 CCIS; Springer: Cham, Switzerland, 2020; pp. 462–472. [CrossRef]
52. Bennasar, M.; Price, B.A.; Gooch, D.; Bandara, A.K.; Nuseibeh, B. Significant Features for Human Activity Recognition Using Tri-Axial Accelerometers. *Sensors* **2022**, *22*, 7482. [CrossRef] [PubMed]
53. Zheng, X.; Otten, E.; Reneman, M.F.; Lamoth, C.J. Explaining Deep Learning Models for Age-Related Gait Classification Based on Acceleration Time Series. *Comput. Biol. Med.* **2025**, *184*, 109338. [CrossRef] [PubMed]
54. Armstrong, D.P.; Beach, T.A.C.; Fischer, S.L. The Influence of Contextual and Theoretical Expertise on Generic and Occupation-Specific Lifting Strategy. *Hum. Factors* **2024**, *66*, 2590–2605. [CrossRef] [PubMed]
55. Martinez-Hernandez, U.; Awad, M.I.; Dehghani-Sanij, A.A. Learning Architecture for the Recognition of Walking and Prediction of Gait Period Using Wearable Sensors. *Neurocomputing* **2022**, *470*, 1–10. [CrossRef]
56. Moawad, A.; El-Khoreby, M.A.; Fawaz, S.I.; Issa, H.H.; Awad, M.I.; Abdellatif, A. New Framework for Human Activity Recognition for Wearable Gait Rehabilitation Systems. *Appl. Syst. Innov.* **2025**, *8*, 53. [CrossRef]
57. Gil-Martín, M.; López-Iniesta, J.; Fernández-Martínez, F.; San-Segundo, R. Reducing the Impact of Sensor Orientation Variability in Human Activity Recognition Using a Consistent Reference System. *Sensors* **2023**, *23*, 5845. [CrossRef] [PubMed]
58. Apraiz, A.; Lasa, G.; Mazmela, M.; Arana-Arexolaleiba, N.; Elguea, Í.; Escallada, O.; Osa, N.; Etxabe, A. The User Experience in Industrial Human-Robot Interaction: A Comparative Analysis of Unimodal and Multimodal Interfaces for Disassembly Tasks. *Robot. Comput. Integr. Manuf.* **2025**, *95*, 103045. [CrossRef]