machines

MDPI

Article

Neural-Network-Driven Intention Recognition for Enhanced Human–Robot Interaction: A Virtual-Reality-Driven Approach

Ali Kamali Mohammadzadeh, Elnaz Alinezhad and Sara Masoud *

Department of Industrial and Systems Engineering, Wayne State University, Detroit, MI 48202, USA; alikamali@wayne.edu (A.K.M.); ht6469@wayne.edu (E.A.)
* Correspondence: saramasoud@wayne.edu

Abstract: Intention recognition in Human–Robot Interaction (HRI) is critical for enabling robots to anticipate and respond to human actions effectively. This study explores the application of deep learning techniques for the classification of human intentions in HRI, utilizing data collected from Virtual Reality (VR) environments. By leveraging VR, a controlled and immersive space is created, where human behaviors can be closely monitored and recorded. Ensemble deep learning models, particularly Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM), and Transformers, are trained on this rich dataset to recognize and predict human intentions with high accuracy. While CNN and CNN-LSTM models yielded high accuracy rates, they encountered difficulties in accurately identifying certain intentions (e.g., standing and walking). In contrast, the CNN-Transformer model outshone its counterparts, achieving near-perfect precision, recall, and F1-scores. The proposed approach demonstrates the potential for enhancing HRI by providing robots with the ability to anticipate and act on human intentions in real time, leading to more intuitive and effective collaboration between humans and robots. Experimental results highlight the effectiveness of VR as a data collection tool and the promise of deep learning in advancing intention recognition in HRI.

Keywords: intention recognition; human–robot interaction; virtual reality; neural networks

Academic Editor: Kai Cheng

Received: 2 April 2025
Revised: 7 May 2025
Accepted: 13 May 2025
Published: 15 May 2025

Citation: Kamali Mohammadzadeh, A.; Alinezhad, E.; Masoud, S. Neural-Network-Driven Intention Recognition for Enhanced Human–Robot Interaction: A Virtual-Reality-Driven Approach. *Machines* **2025**, *13*, 414. https://doi.org/10.3390/machines13050414

1. Introduction

Human–Robot Interaction (HRI) refers to the ways in which humans and robots interact with each other and combines the knowledge and techniques from computer science, engineering, psychology, and design. HRI is crucial in manufacturing as it determines the efficiency, reliability, and safety of the human–robot system, and it has the potential to significantly improve the productivity and quality of manufacturing processes. HRI serves to align robot capabilities with human expectations. In addition, it can enhance the overall work environment by reducing the physical and mental workload of human workers and enabling them to focus on more challenging and rewarding tasks.

As depicted in Figure 1, HRI can be divided into three primary categories: Human–robot coexistence, which involves the separation of workspace between humans and robots without the need for synchronization of actions and intentions [1]; human–robot cooperation, where humans and robot work individually to achieve a common goal, sharing both time and space. Advanced technologies are employed to ensure collision detection and avoidance during this cooperation [2]; human–robot collaboration, which encompasses scenarios where humans and robots engage in complex tasks with direct interactions. It emphasizes explicit contact between humans and robots, highlighting a high level of cooperation and coordination [3].

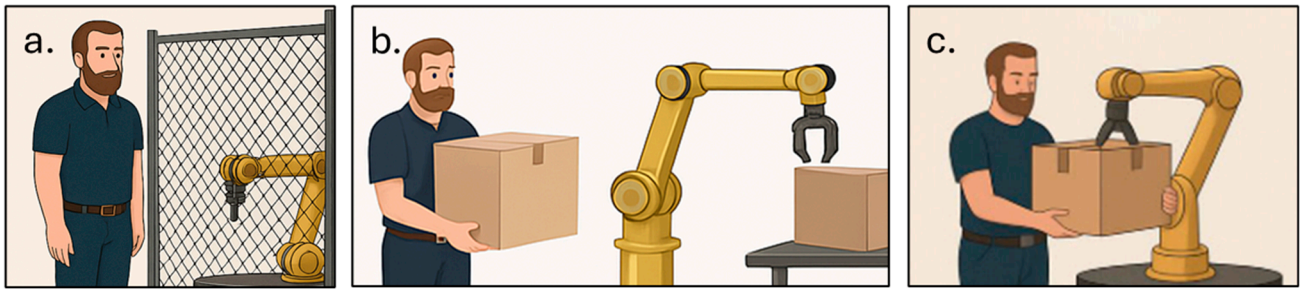

Figure 1. Three forms of human–robot interaction: (**a**) coexistence, (**b**) cooperation, and (**c**) collaboration.

Intention recognition is an essential aspect of collaborative HRI as it enables the robot to understand and respond to the actions of its human partners. In this study, we use the term intention to denote the underlying goal of the human, while activity refers to the observable motion sequence that accomplishes that goal. By recognizing human activities, robots can provide appropriate assistance and support, which can increase the efficiency and effectiveness of the collaboration between humans and robots. The recognition of activities can also lead to better safety by allowing the robot to anticipate and respond to potential hazards in real time [4].

An important and emerging context for applying intention recognition is within VR environments. VR is increasingly used in the manufacturing industry to enhance the design, testing, and training processes. VR provides a safe and controlled environment for manufacturing engineers to test and optimize the functionality and performance of their products and systems [5]. It enables engineers to simulate the production process and identify potential issues and solutions before they arise in the actual manufacturing process, reducing the cost and time associated with prototyping and testing. VR provides immersive training for manufacturing workers by offering hands-on virtual training that mimics real production. It enables workers to learn at their own pace and provides a safe environment to practice and make mistakes. Neural networks in VR environments can recognize and classify human activities using body posture, motion, and gestures. They help control and monitor worker actions for correct and safe performance and offer real-time feedback for improved performance and reduced injury risk. The use of neural networks for intention recognition in VR environments in manufacturing has the potential to improve safety, efficiency, and productivity [6].

The aim of this research is to assess the viability of using neural networks to recognize human intentions based on their activities in manufacturing environments. This research demonstrates the potential impact of deep learning on human–robot interaction. The remainder of this paper is structured as follows: Section 2 presents a review of the relevant literature, grounding our study in the broader context of intention recognition in manufacturing. Section 3 outlines the methodology of our research, including data collection and processing methods, and the specific neural network models employed. Section 4 provides the experimental design and a detailed analysis and discussion of the results. Finally, Section 5 concludes the paper and suggests directions for future research.

2. Literature Review

The contemporary manufacturing ecosystem is undergoing a transformative evolution, propelled by rapid technological advancements, and shifting paradigms of Industry 4.0 [7]. In this digitally driven landscape, human–robot collaboration emerges as a keystone to enhancing productivity, ensuring safety, and fostering innovation [2]. Now, Industry 5.0 introduces human-centric automation that prioritizes safety and personalization alongside productivity. However, for this collaboration to be seamless and efficient, a profound

understanding of human activities in real time is imperative. Intention recognition bridges this knowledge gap [3]. By enabling robots to interpret and respond to human movements and intentions dynamically, it significantly reduces operational errors and augments human capabilities. However, this is a challenging task arising from a multitude of factors including background clutter, partial occlusion, variations in scale and viewpoint, and changes in lighting, appearance, and frame resolution, among others [8].

HRI in manufacturing is a convergence of human ingenuity and robotic precision. As we stand on the precipice of the fourth industrial revolution, this symbiotic relationship between man and machine in manufacturing arenas is becoming increasingly paramount [9]. Historically, the manufacturing sector viewed robots as tools designed solely to enhance productivity through repetitive, mundane tasks, often operating in isolation behind safety barriers. These robots, while effective, had limited scopes of operation. They were utilized in processes deemed too perilous or monotonous for human involvement. However, the advent of advanced sensor technology, coupled with breakthroughs in artificial intelligence and machine learning, heralded a transformative shift in HRI over decades [10]. Today, the manufacturing landscape is dotted with 'collaborative robots' or 'cobots'. Unlike their predecessors, these machines are intricately designed to operate alongside human workers, not just as tools but as collaborative partners, sharing tasks and responsibilities [11].

The potential to enable seamless and intuitive collaboration between humans and robots has fueled a surge of interest in intention recognition for HRI in recent years [12]. Recognizing human intent is crucial for the development of robots that can proactively assist humans in a wide range of tasks, from manufacturing to personal care [13–15]. This has sparked a multidisciplinary interest among researchers in robotics, artificial intelligence, cognitive science, and ergonomics [16]. Combining human dexterity and judgment with the precision and endurance of robots offers numerous advantages. These include enhanced productivity, cost efficiency, improved quality, and greater safety and wellness in the workplace [11]. While this combination ensures tasks are completed faster and more efficiently, cobots can operate continuously, reducing labor costs and minimizing human errors [17]. In addition, cobots ensure consistent quality, while humans bring intricate detailing and adaptability to the table [18]. Finally, cobots can readily handle hazardous tasks, reducing human exposure to potential dangers. Furthermore, with the increasing modularity of these cobots, reprogramming them to adapt to a plethora of tasks is becoming hassle-free, offering unparalleled flexibility in manufacturing environments [19].

As the realm of HRI expands, ensuring human safety in this intertwined workspace is paramount. This concern is more than physical barriers; it is about real-time comprehension and reaction. Intention recognition, in this regard, becomes invaluable. By enabling robots with the capability to interpret and predict human actions and intentions in real time, potential mishaps can be proactively avoided [20,21].

In our research, we concentrate on 'Intention Recognition', a term that is sometimes equated with 'Activity Recognition'. This level of discernment is crucial for the development of context-aware and intelligent systems [22,23]. VR has been used for training, design, and collaborative tasks that involve both humans and robots [24]. Within this framework, intention recognition plays a crucial role. By identifying human actions and intentions in a virtual setting, VR systems not only provide a safe environment for data collection but also facilitate the rehearsal of collaborative tasks [8].

As displayed in Figure 2, a comprehensive literature search was conducted on reputable academic databases and research repositories using relevant keywords, including but not limited to "intention recognition", "activity recognition", "human–robot interaction", "virtual reality", and "extended reality", as well as their combinations. The search results were carefully filtered to remove duplicates, non-English language papers, unavailable

documents, and unrelated studies. Notably, the number of publications related to "virtual reality", "intention recognition", and "human–robot interaction" has more than doubled between 2018 and 2023, indicating a significant growth in these areas.

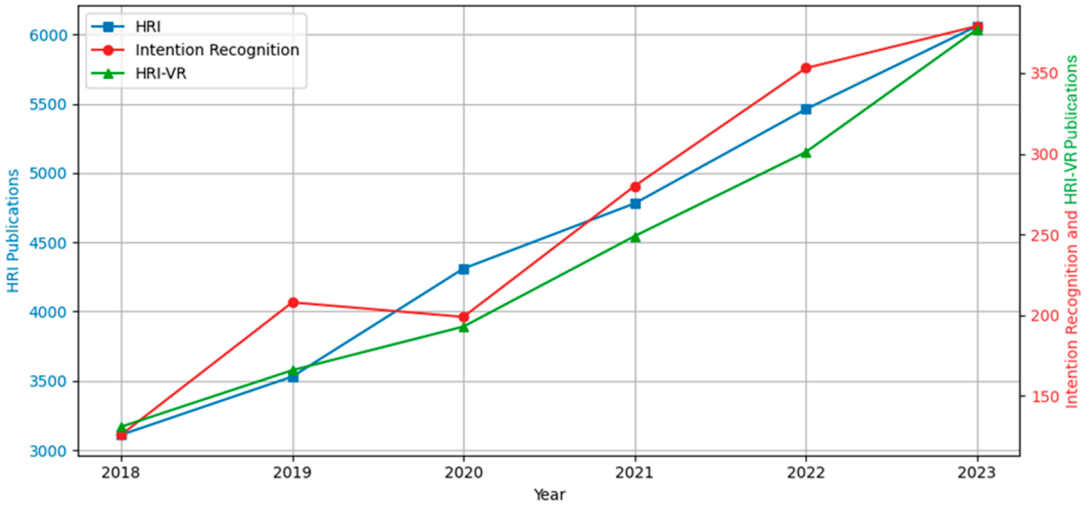

Figure 2. Publication trends on HRI, intention recognition, and HRI-VR.

Among the emerging trends, the integration of VR has gained significant traction as a powerful tool for developing and evaluating HRI, and in particular, human prediction including intention recognition in HRI systems [5,6,25,26]. Researchers have proposed various deep learning architectures, including CNNs and LSTM networks, to extract spatial and temporal features from this multimodal data [5,27–29]. To name some of the most recent notable studies in the field, Xia et al. proposed an XR system based on HoloLens 2 for a multimodal fusion intent recognition algorithm (MFIRA) algorithm that features the fusion of gesture and speech information through machine learning feature–layer fusion, as well as analysis of the conflicts between modality information [30]. Peng et al. proposed an intention recognition model and designed a communication architecture to assist intent sharing between production elements [31]. Xu et al. integrated virtual reality, computer vision, and human–robot interaction methodologies for remote robot control in a proxy data center environment, demonstrating the potential of VR-based human–robot interaction to enhance productivity and safety in industrial applications [32]. As manufacturing processes become more intricate and data-intensive, the role of neural networks in facilitating sophisticated intention recognition algorithms becomes paramount [33–35]. Recognizing human activities, particularly in VR settings, not only ensures safety and efficiency but also paves the way for the next generation of smart manufacturing units where human expertise and robotic precision harmoniously coexist. This study situates itself at the nexus of these developments, aiming to enhance the symbiotic relationship between humans and robots in the ever-evolving manufacturing sector.

Within the last few years, significant advancements have been made in the field of behavioral recognition and its impact on human–robot interactions. Awais and Henrich introduced a probabilistic-state-machines-based algorithm specifically designed for both explicit and implicit intention recognition in human–robot collaboration [36].

Stiefmeire et al. employed ultrasonic sensors and Hidden Markov Models for worker intention recognition [37], while Koskimaki et al. used a wrist-worn Inertial Measurement Unit (IMU) and a K-Nearest Neighbor model for classifying activities on industrial assembly lines [38]. Further, Maekawa et al. proposed an unsupervised approach for lead time estimation using smartwatch and IMU sensor data [39]. Zhu et al. tackled human intention recognition through a hidden Markov-based algorithm and deep convolution neural net-

works, focusing on hand gestures and gait periods [40]. Sun et al. utilized muscle electrical signals and K-Nearest Neighbor algorithms for gait motion recognition [41]. Masoud et al. introduced a task recognition framework for grafting operations using data gloves [33]. Buerkle et al. introduced a novel approach using a mobile electroencephalogram (EEG) to detect upper-limb movement intentions in advance by leveraging the human brain's ability to evaluate motor movements before execution [34]. Zhang et al. presented a human–object integrated approach for recognizing operator intention in human–robot collaborative assembly, leveraging spatial-temporal graph convolutional networks for action recognition and an improved YOLOX model for assembly part detection [35]. Zhang et al. developed a human–robot collaboration system using Electromyography (EMG) signals, enhancing motion intention accuracy through an optimized algorithm and deep reinforcement learning, resulting in reduced human effort in sawing tasks [42]. Li et al. introduced Proactive Human–Robot Collaboration, emphasizing a shift towards combining human intuition with robotic precision, enabling a more anticipatory and synergistic approach to manufacturing tasks [43]. Zhang et al. presented a predictive human–robot collaboration model for assembly tasks using a Convolution-LSTM-based approach. Tested in a vehicle seat assembly, the model enhanced efficiency and adaptability compared to non-predictive methods [44]. Finally, Sun et al. presented a digital twin framework for human–robot collaboration in assembly using an automobile generator case study, enhancing robot cognition and adaptability with intention recognition and task knowledge [45].

Although numerous studies exist on intention recognition at the intersection of HRI and VR, as highlighted in Figure 2, none have focused on delivering generalized tasks with high-resolution data on body movement trajectories. By capturing detailed data on human behavior and intention in virtual environments, VR facilitates the development of more accurate and adaptive intention recognition models, leading to safer and more efficient manufacturing processes. In this study, participants use wearable technology to immerse themselves in a virtual manufacturing environment where they must complete a series of tasks near a robot. These devices enhance immersion and continuously collect motion and gesture data. We propose a framework that leverages these continuous data streams to recognize the underlying tasks performed by the participants, which generate these data.

3. Materials and Methods

A breakdown of our proposed framework is illustrated in Figure 3. The proposed framework is built upon three main modules, which are data acquisition, data processing, and model training and evaluation.

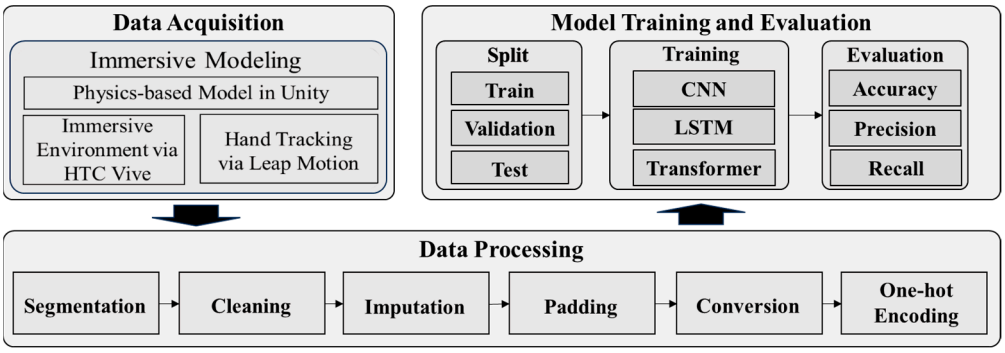

Figure 3. Our proposed intention recognition framework.

3.1. Data Acquisition

The Unity3D (https://unity.com/releases/editor/archive, accessed on 10 January 2020) game engine [46], the HTC Vive Pro Eye Arena system (HTC Corporation, Taoyuan

District, Taoyuan, Taiwan) [47], and Leap Motion (Walnut Creek, CA 94597, USA) [48] are used for creating immersive virtual environments and tracking user's spatiotemporal trajectories of body and hand, as displayed in Figure 4. Unity3D is a physics-based game development engine and platform that allows developers to create interactive 2D and 3D content, including simulations and VR experiences [46]. The HTC Vive Head-Mounted Display (HMD) serves a dual purpose of displaying the designed virtual environment to the users as well as tracking the trajectories of the users within this environment. Supplementing the HMD, three additional HTC-Vive trackers are employed, strategically attached to the participants' chest and elbows, to further track users' trajectories. These trackers further enhance our ability to capture the nuanced movements of the participants during various activities, thereby enriching the dataset for our study. Lastly, the Leap Motion sensor is used to model hands and track joint movements in the virtual environment and track users' hand joint movements. This technology enabled us to capture the fine-grained detail of hand movements, offering us an in-depth view of how these motions factor into different activities.

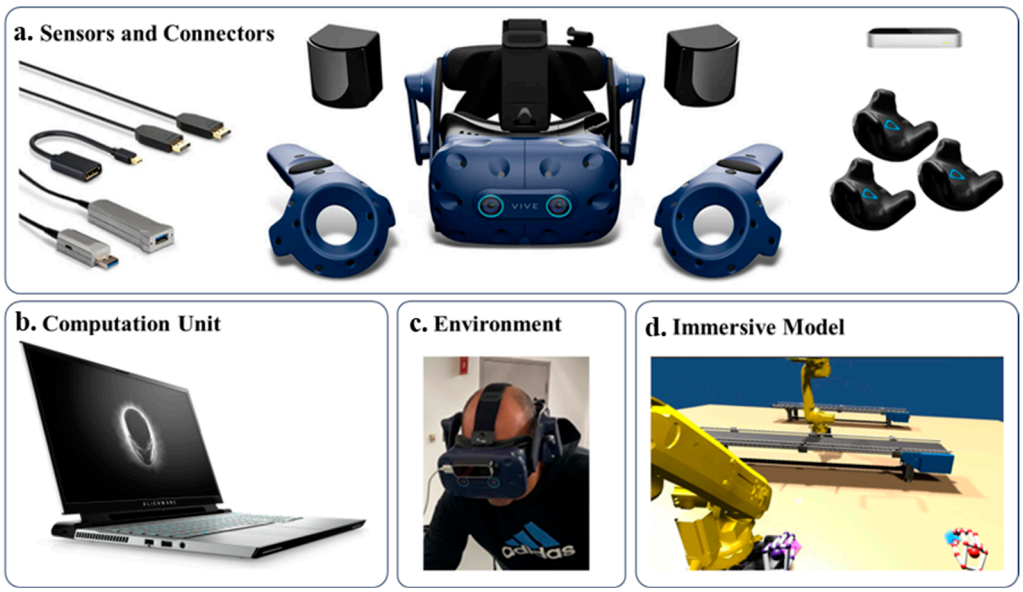

Figure 4. Overview of the experimental setup components. (**a**) Sensors and connectors, including the HTC Vive Pro headset, base stations, controllers, and trackers. (**b**) Computation unit used to run the VR system and data processing. (**c**) Experimental environment showing a participant interacting with the VR setup. (**d**) Immersive model of the virtual manufacturing environment used for intention recognition tasks.

Each participant is equipped with the proposed setup, recording the participants' movements during their engagement with the tasks in the virtual environment (as dictated orally by the authors) at 300 hz. The duration and nature of these activities (Table 1) are designed to capture a broad spectrum of actions and movements. The resulting movement data was automatically collected and stored in an Excel file for subsequent analysis. Table 1 outlines the list of activities performed by the participants during the study. These activities were carefully chosen based on their applicability to scenarios typically encountered within manufacturing or industrial contexts, thus offering a diverse spectrum of human motions for our investigation. The selection process for these activities not only leaned on their relevance to manufacturing settings but also considered the broader literature on the subject [49], in addition to the data available from established datasets such as UCI, WISDM, and GAMI [50].

Table 1. List of activities along with descriptions and context in manufacturing.

Activity Code	Activity Name	Activity Description	Context in Manufacturing
A_1	Idle	Standing in place	Serves as a baseline state, representing periods of inactivity
A_2	Bending	Bending down	Mimics actions such as interacting with lower machinery levels
A_3	Sitting	Sitting on the ground/chair	Reflects periods of rest or tasks performed while seated
A_4	Moving	Walking around	Indicates general mobility, transitioning tasks or locations
A_5	Relocating	Putting parts in a box/on table	Represents tasks involving object manipulation or repositioning
A_6	Grabbing with one hand	Grabbing a component in either right or left hand	Simulates handling of smaller or lightweight components, such as screws, bolts, or small tools, and allows for simultaneous engagement in another task, such as operating a machine control.
A_7	Grabbing with two hands	Grabbing a component with both hands	Represents handling of larger/heavier items that require secure grip and better control, such as larger assembled products, ensuring safety and minimizing risk of mishandling.

3.2. Data Processing

Data preprocessing follows the data acquisition phase. This step applies a series of steps on the combined observations, including data segmentation to break down the streams of time series, cleaning to remove any anomalies or inconsistencies, imputation to manage missing data, normalize the data, and reinforce the time series structure of the data. This preprocessing was crucial in preparing the data for accurate and reliable analysis in the later stages of the study. Additional steps such as padding, conversion, and one-hot encoding are performed to structure the data for the proposed deep neural networks.

Data segmentation via Change Point Analysis (CPA): CPA is a statistical technique employed to identify significant points or intervals in a time series where there is a notable change or shift in the underlying characteristics or behavior of the data. It is a valuable tool for detecting and understanding changes in trends, patterns, or distributions within a time series or any ordered sequence of data [51]. The primary objective of change point analysis is to pinpoint the locations of these changes and quantify their magnitude, timing, and potentially their causes. In the context of intention recognition in manufacturing, workers or employees typically perform sequences of actions one after another. Hence, it becomes essential to identify when activity changes occur to make more accurate predictions continually. To address this aspect, each of the observations collected during the data collection process comprises human movement data representing a sequence of activities rather than a single activity. The collected observations are subsequently broken down into single activities based on the results obtained from the change point analysis method.

To identify significant changes in the time series data and segment the observations accordingly, the Pelt algorithm is employed in conjunction with a Gaussian radial basis function (RBF) model [52]. The Pelt algorithm optimizes a cost function to efficiently detect change points, making it well-suited for handling large datasets and providing accurate outcomes. The cost function used in the Pelt algorithm is typically based on a specific statistical criterion, such as the sum of squared residuals, sum of absolute residuals, or Bayesian Information Criterion (BIC). The cost function evaluates the fit of each segment and penalizes

the introduction of additional change points. The optimal change points are determined by finding the points that minimize the cumulative cost over all possible segmentations. Denoting the time series data as $X = \{x_1, x_2, \ldots, x_T\}$, where x_t represents the data point at time index t, and T is the total number of time steps. The goal of the Pelt algorithm is to partition this time series into segments $\{(x_1, \ldots, x_{\tau_1}), (x_{\tau_1+1}, \ldots, x_{\tau_2}), \ldots, (x_{\tau_{k-1}+1}, \ldots, x_T)\}$, where τ_k is the change points. In conjunction with the Pelt algorithm, the Gaussian radial basis function, (1), is utilized.

$$f(x) = \sum_{i=1}^{N} w_i . \exp\left(\frac{(x - c_i)^2}{2\sigma^2}\right), \quad (1)$$

where $f(x)$ represents the function's output, x is the input, N is the total number of radial basis functions, w_i represents the weight associated with each basis function, c_i is the center of the i^{th} basis function, and σ is a parameter that controls the spread or width of the basic functions.

Data Cleaning: Following the collection of segments, the data undergoes several processing stages. The initial step is to apply a threshold of 30%, determining whether to retain or discard specific observations. This threshold is selected to maintain a high level of data quality.

Data Imputation: For the observations that pass the 30% threshold, imputation is employed using the Iterative Imputer method to fill in any missing data [53]. This imputation method models each feature with missing values as a function of other features and iteratively refines these estimates until a stable solution is reached. Subsequently, normalization is applied to standardize the data's scale, improving the model's ability to accurately interpret and learn from the data. Finally, the processed data is transformed into a time series format, providing a sequential structure essential for our analysis.

Data Padding: Given the uneven length of time series segments, padding is implemented to standardize data dimensions. Padding extends sequences with zeros until all match the length of the longest sequence.

Conversion to 3D Tensor: The standardized observations are transformed into a 3D tensor, shaped (3500, 126, 665). This tensor accounts for the number of observations, timestamps, and features, respectively.

One-Hot Encoding of Targets: For the final preprocessing step, the target labels are encoded using one-hot encoding. With seven activity classes, each class was represented as a binary vector in our dataset.

3.3. Model Training and Evaluation

Given the spatiotemporal structure of our dataset, CNN, CNN-LSTM, and CNN-Transformer are selected for their respective strengths and capabilities in managing complex spatial and temporal data structures. CNNs are well suited for extracting local spatial patterns and features from input data, making them effective for recognizing fine-grained motion and gesture details. LSTM networks are designed to model temporal sequences and are particularly useful for capturing the dynamic evolution of human actions over time, which is critical for understanding intentions in HRI. To further enhance the model's capacity to capture long-range dependencies and contextual information, we incorporated a CNN-Transformer architecture that combines the local feature extraction capabilities of CNNs with the global attention mechanism of Transformers. This hybrid approach enables the model to better integrate both short- and long-term temporal relationships, improving its ability to recognize complex and subtle human intentions in interactive scenarios.

3.3.1. CNN

The CNN model, renowned for its robustness and efficacy in handling multi-dimensional input, is used to extract local feature representations from the time series data. Its unique strength lies in recognizing spatial dependencies in the local frame, which makes it a viable choice for this study. CNNs have become the architecture of choice for a wide range of image processing and computer vision tasks due to their ability to learn hierarchical representations of input data. Originating from the field of deep learning, CNNs have been successfully employed in numerous applications such as image and video recognition, recommender systems, image generation, medical image analysis, and natural language processing, among others. Convolutional layers form the basic building block of a CNN and consist of a set of learnable filters (or kernels), which have a small receptive field but extend through the full depth of the input volume. As the filters slide over the input data, they perform a dot product operation, creating a feature map that represents the presence of specific features in the input. Given an input matrix I and a filter F, the convolution is defined as presented in (2), where * denotes the convolution operation. The resulting matrix is often called a feature map.

$$(I * F)(i, j) = \sum m \ \sum n \ I(m, \ n)F(i - m, \ j - n) \tag{2}$$

After each convolution operation, the feature maps are passed through a non-linear activation function, such as a Rectified Linear Unit (*ReLU*). This function (3) introduces non-linearity to the network, allowing it to learn more complex patterns.

$$ReLU(x) = \max(0, \ x) \tag{3}$$

Following one or more convolutional layers, pooling layers are used to reduce the spatial size of the representation, both to decrease the computational load and to help the model generalize better. The most common form of pooling is max pooling, where the maximum value is chosen from each cluster of neurons at the prior layer. Near the end of the network, fully connected layers are used to perform high-level reasoning. Neurons in a fully connected layer have connections to all activations in the previous layer, as seen in regular neural networks. Their purpose is to use these learned features to classify the input image. In a fully connected layer, the weights can be represented as a matrix W, and the biases as a vector b. Given input vector x, the output y of the fully connected layer is as displayed in (4).

$$y = Wx + b \tag{4}$$

The final layer in a CNN is typically a softmax or a sigmoid activation function for multi-class or binary classification tasks, respectively. These functions convert the output into a probability distribution over classes, providing a definitive prediction. The softmax function transforms an input vector z into a probability distribution over C classes as displayed in (5).

$$\text{Softmax}(z)_i = \frac{\exp(z_i)}{\sum_{j-1}^{C} \exp(z_i)}, \quad \text{For } i = 1, \ldots, C. \tag{5}$$

Through this structure, CNNs can leverage the spatial and temporal dependencies in input data through the application of relevant filters, providing an automatic, adaptive approach to feature extraction, thus enabling a more effective and efficient form of intention recognition.

3.3.2. Long Short-Term Memory Networks (LSTMs)

LSTMs are a type of recurrent neural network (RNN) architecture, explicitly designed to overcome the vanishing gradient problem associated with traditional RNNs. LSTMs are exceptionally suited for classifying, processing, and making predictions based on time series data, given their capacity for learning long-term dependencies. They have been widely used in a variety of applications, including speech recognition, language modeling, translation, and gesture recognition. The primary components of an LSTM unit are the cell state and three types of gates, namely the forget (6), input (7) and (8), and output (10) and (11), Given an input vector x_t at time t, the previous hidden state h_{t-1}, and the previous cell state c_{t-1} (9), the LSTM unit updates using sigmoid (σ) and tanh functions.

$$f_t = \sigma\Big(w_f.[h_{t-1}, x_t] + b_f\Big) \tag{6}$$

$$i_t = \sigma(w_i.[h_{t-1}, x_t] + b_i) \tag{7}$$

$$c_t = \tanh(w_c.[h_{t-1}, x_t] + b_c) \tag{8}$$

$$c_t = f_t \times c_{t-1} + i_t \times c_t \tag{9}$$

$$o_t = \sigma(W_o.[h_{t-1}, x_t] + b_o) \tag{10}$$

$$h_t = o_t \times \tanh(c_t) \tag{11}$$

where the forget gate (f_t), (6), determines which information from the cell state should be thrown away or kept. w_f and b_f are the weights and bias associated with the forget gate. The input gate updates the cell state with new information. It decides which values will be updated and creates a vector of new candidate values that could be added to the state. (7) uses sigmoid to decide which values should be updated in the cell state, while (8) uses tanh function to create a vector of new candidate values. w_i and b_i are the weights and bias associated with the forget gate. The cell state (9) runs along the entire chain of LSTM. It carries information from earlier time steps to later ones and can be updated or modified by the gates to forget certain values as dictated by the forget gate and adds new values as proposed by the input gate. Finally, the output gate decides what the next hidden state should be as displayed in (10). This hidden state, h_t, will be used in predictions at this time step, and will be transferred to the next LSTM unit as displayed in (11). Through this structure, LSTMs can handle long-term dependency problems. They can remember or forget information over a long period of time, making them highly effective for activity recognition tasks, particularly when the activities are of varying lengths or when the recognition system needs to account for temporal dependencies in the data.

3.3.3. Transformer

Transformers have emerged as dominant players in the machine learning arena, notably for sequence-related tasks. Originating in natural language processing, their applicability has now expanded across various domains, including intention and activity recognition. At the heart of the Transformer's prowess is its ability to parallelize sequence data processing. Unlike RNNs, which inherently work sequentially, Transformers process all data points of a sequence simultaneously. This parallel processing attribute, coupled with the model's innate capacity to determine the importance of different parts of a sequence, offers a significant computational advantage. The crux of a Transformer model is its attention mechanism, which assigns weights to different parts of the sequence depending

on their relevance. Given a sequence with data points represented as key (K), value (V), and query (Q) pairs, this mechanism can be broken down as displayed in (12):

$$\text{Attention}(Q,\ K,\ V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \times V \tag{12}$$

Here, d_k is the dimensionality of the key vectors. The result of this computation gives a weighted sum of values, where the weight assigned to each value depends on the query and key. Self-attention is a specialized form of attention mechanism where the model assigns weights by comparing each data point in a sequence to every other data point as shown in (13). This enables the model to evaluate dependencies without considering sequence order.

$$\text{Self} - \text{Attention}(X) = \text{Attention}(X, X, X) \tag{13}$$

Typically, a Transformer model contains an encoder that interprets the input data and a decoder that formulates the output. Each of these is made up of numerous identical layers, incorporating self-attention and feed-forward networks. In the context of intention recognition, the Transformer's capability to concurrently process all activity data offers an encompassing perspective of the entire sequence. Hence, when integrated with CNNs as a CNN-Transformer, the resultant model leverages the spatial feature extraction strength of CNNs with the sequence processing expertise of Transformers. This synergy allows for superior attention to crucial parts of the sequence, bolstering our expectations of exceptional performance in this work.

We train our models on the preprocessed data, tweaking parameters based on the performance on a validation set. The performance of the algorithms was then assessed using metrics such as precision, recall, F1-score, and support. Precision is the ability of the classifier not to label as positive a sample that is negative, whereas recall (or sensitivity) is the ability of the classifier to find all the positive samples. The F1-Score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0. Support is the number of actual occurrences of the class in the specified dataset. Precision, recall, and F1-score are metrics that evaluate the quality of the model's predictions, while support is the number of occurrences of each class in the dataset.

4. Experimental Setup

We enlisted the participation of a diverse group of six individuals, aged 19 to 39, including two females and four males to perform seven activities: 'Standing', 'Sitting', 'Bending', 'Walking', 'Pick up one hand', 'Pick up two hands', and 'Relocate'. The procedure for data collection began with a brief introduction, acquainting participants with the HTC-Vive system and Leap Motion and outlining the process and their role in it. Each observation included human movement data for performing a sequence of activities. The features collected over each observation include the position (x, y, z), rotation (x, y, z), velocity (x, y, z), and angular velocity (x, y, z) of each HTC-Vive tracker and head-mounted display (HMD), in addition to the HMD's forward vector, which is a normalized 3D vector (x, y, z) representing the direction of the headset's gaze. Furthermore, the coordinates of palm position and velocity; vectors for the palm normal and direction to the fingers; and lists of the attached fingers as demonstrated in Figure 5.

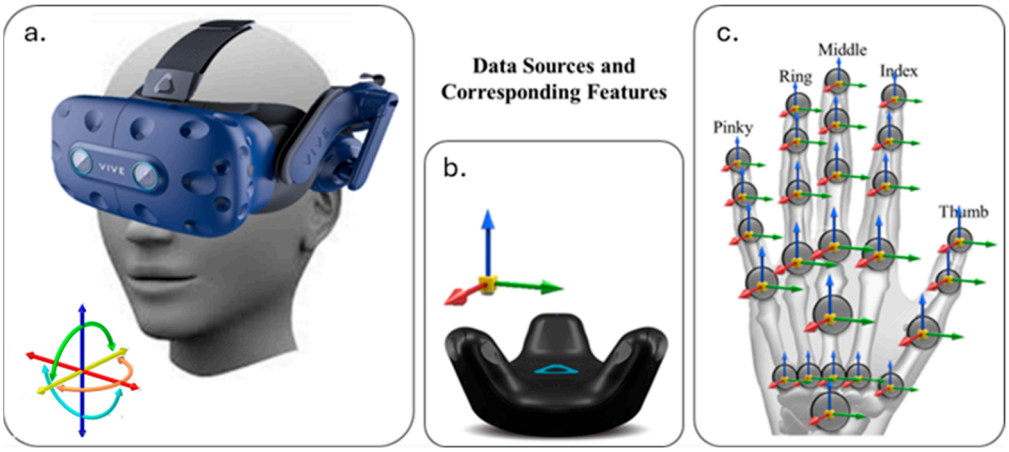

Figure 5. Data sources and corresponding features used in the system. (**a**) HTC Vive Pro headset providing head orientation and movement data; (**b**) HTC Vive Tracker capturing body or object motion with positional and rotational data; (**c**) Leap Motion sensor providing detailed hand and finger tracking, including joint positions and orientations for the thumb, index, middle, ring, and pinky fingers.

The HTC system is an "outside-in" tracking system that uses room-scale technology [47]. The utilized HTC Vive system uses external sensors, known as base stations, to track the position of the headset and trackers in a room. These base stations are placed in fixed positions around the coverage area (10 by 10 m^2) and emit infrared signals. The sensors on the headset and sensors detect these signals, allowing the system to precisely calculate their positions in 3D space [47]. Leap Motion is a motion-tracking technology designed to capture and interpret hand and finger movements with high precision, enabling users to interact with digital environments in a natural, touch-free manner [48]. The Leap Motion Controller is equipped with infrared cameras and LEDs that track the position and movement of a user's hands and fingers in 3D space within a close range [48]. This data is processed to allow users to manipulate objects, control applications, or navigate virtual environments through gestures alone.

After detecting the change points using the Pelt algorithm, we proceeded to segment the time series into distinct parts based on the identified changes. Each resulting segment represented a single-activity observation, and each segment was saved as a separate file. This process generated a total of 4400 single-activity observations, including 1200 instances of standing idle, 1000 instances of walking, 600 instances of pickup with one hand, and 400 instances of sitting, bending, pickup with both hands, and placing, each. Each of these single-activity observations was represented as a matrix with a shape of (X, 676), where 'X' denotes the number of timestamps collected during the observation, and 665 signifies the number of features captured in the data. To ensure uniformity, the observations were padded to achieve a consistent shape. By combining all the observations, we created a tensor shape (4400, 126, 676), allowing us to efficiently work with the data using the prediction models. Upon discarding data with missing values exceeding 30 percent for any given feature, a dataset of size 153,385,200 data points is amassed. This dataset translates into 3500 observations, each encapsulating 676 features with varying time stamps, ranging from 13 to 126, which are processed by going through various steps explained in the data processing section such as segmentation, cleaning, imputation, padding, data conversion, and one-hot encoding.

5. Results and Discussion

Following preprocessing, the observations are divided into training and validation sets using an 80/20 split, and the performance of our classification algorithms are evaluated using metrics like accuracy, precision, recall, and F-1 score. As illustrated in Figure 6, the CNN model boasts an overall precision, recall, and F-1 score of 0.95, indicating a high level of performance. It is important, however, to recognize the model's variance in performance across classes. Specifically, while it excelled in predicting both 'pick up' activities, its performance was slightly lower for 'walking' and 'standing' activities, with precision and recall scores of 0.95 and 0.86, and 0.82 and 0.93, respectively.

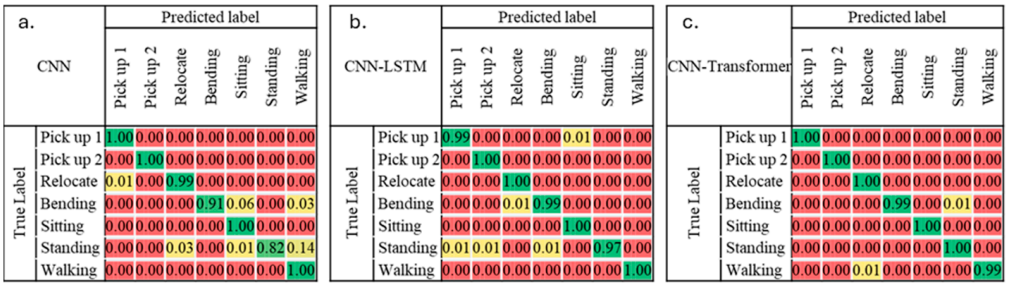

Figure 6. Confusion matrices showing the classification performance of (**a**) CNN, (**b**) CNN-LSTM, and (**c**) CNN-Transformer models across seven activity classes: Pick up 1, Pick up 2, Relocate, Bending, Sitting, Standing, and Walking. The color intensity represents the percentage of predicted versus true label for each of activities, with green indicating highest percentage, red indicating lowest percentage, and yellow in between. The CNN-Transformer shows the highest overall classification performance across activities.

Figure 6 also displays the performance of the ensemble CNN-LSTM network on each class. It can be observed that the model performed well in all categories, with precision, recall, and F1-score all above 0.97 for all classes. The model's overall accuracy on the test dataset stands at 0.99. However, it is noteworthy that the 'standing' category exhibits a slightly lower recall compared to the rest, indicating a few 'standing' instances are misclassified into other categories.

As outlined in Figure 7, the ensemble CNN-Transformer model displays exemplary performance, achieving an overall accuracy close to 1.00. Across all classes, the model exhibits high precision, recall, and F1-scores. Specific class metrics are as follows. 'Pick up one hand' and 'pick up two hands' have precision, recall, and F1-scores of 1.00. 'Relocate' and 'bending' achieve near-perfect performance with precision of 0.99 and 1.00, recall of 1.00 and 0.99, and F1-scores of 0.99 for both. 'Sitting' and 'standing' also achieve a scores of 1.00 across all measures. 'Walking' is close to perfect, with precision and F1-scores of 1.00, and a slightly lower recall of 0.99.

Given the performance of our models, it is evident that the CNN-Transformer model has surpassed the capabilities of both the CNN and the CNN-LSTM models (see Figure 7). The CNN-LSTM model encountered difficulties in distinguishing the 'standing' activity, a task that required accurate classification amidst the complexity of human postures and movements. Similarly, the standalone CNN model displayed some weaknesses, specifically with the 'walking' and 'standing' activities, where it yielded lower precision and recall. This implies that the model was less accurate and had more false positives and false negatives, indicating potential issues with inability to capture the complexities of human gait.

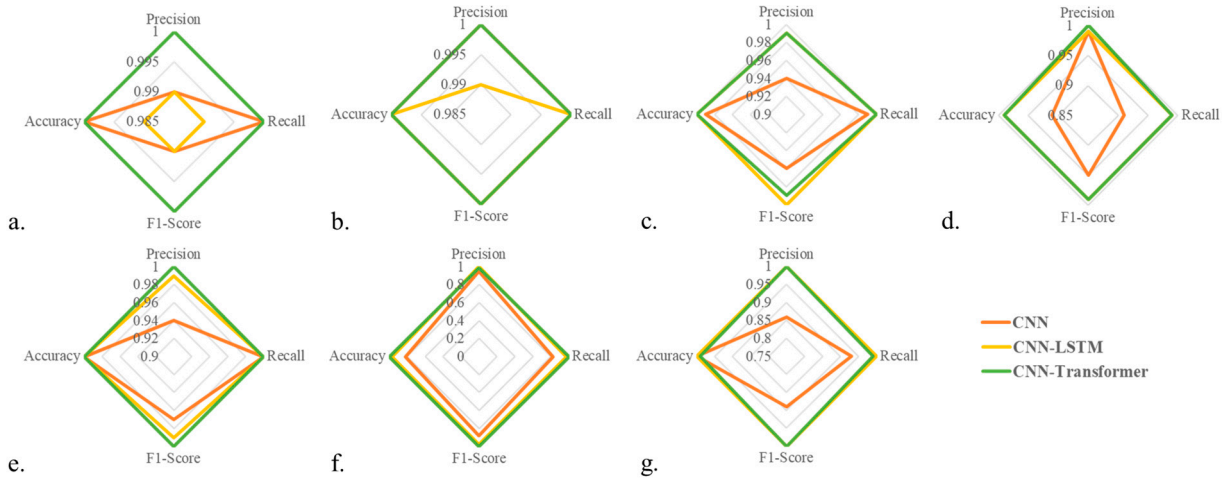

Figure 7. Radar plots comparing the performance of CNN, CNN-LSTM, and CNN-Transformer models across seven activities using four metrics: accuracy, precision, recall, and F1-score. Each subplot corresponds to a specific activity. (**a**) Grabbing with one hand; (**b**) Grabbing with two hands; (**c**) Relocating; (**d**) Bending; (**e**) Sitting; (**f**) Idle (Standing); (**g**) Walking.

Contrary to these struggles, the CNN-Transformer model stands out with its performance. Demonstrating near-perfect precision and recall across all activity classes, this model has shown marked improvement. It suggests that the Transformer architecture, renowned for its capability to handle temporal dependencies and its attention mechanism, has effectively enhanced the model's ability to capture the intricate patterns of human activities. Its performance is a clear indicator of significant advancements in its capacity to correctly identify, classify, and predict all activities. This progression signifies not just model enhancement but also strides towards the goal of robust and reliable human intention recognition.

Figure 8 presents the training and validation accuracy and loss curves over epochs for the CNN-Transformer model. The model exhibits rapid convergence, with both training and validation accuracy increasing steadily while the loss decreases, indicating effective learning. The close alignment between training and validation curves suggests that the model generalizes well without significant overfitting.

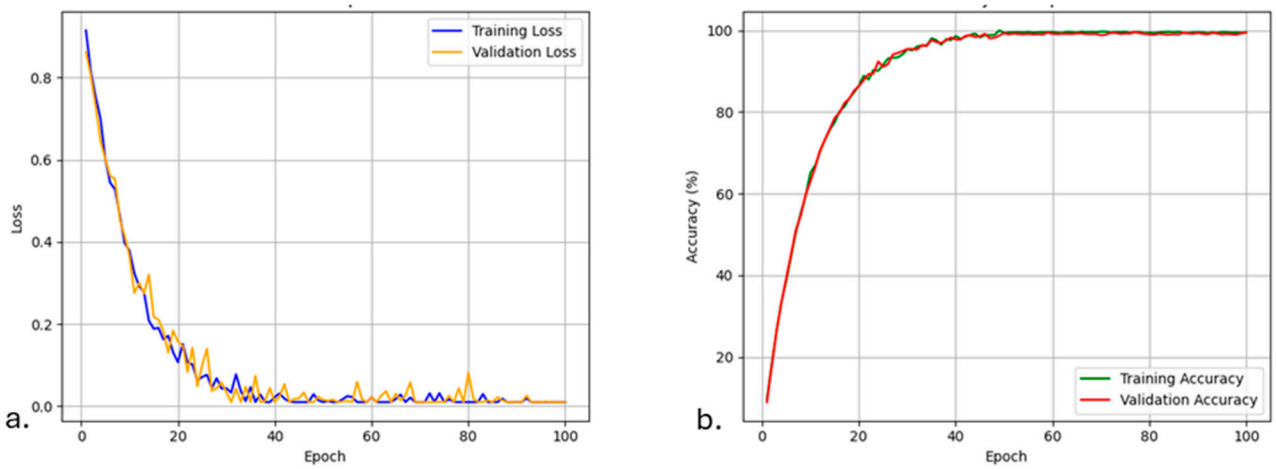

Figure 8. Loss (**a**) and accuracy (**b**) of the winning model (CNN-Transformer model) over epochs.

5.1. Sensitivity Analysis

As early intention recognition—the ability to recognize intentions before the completion of the corresponding movement—is gaining more popularity, the proposed intention recognition framework is tested against incomplete observations for all trained models. The incomplete observations are produced by truncating data at successive one-second intervals, assessing model accuracy using the F1-score as displayed in Figure 9.

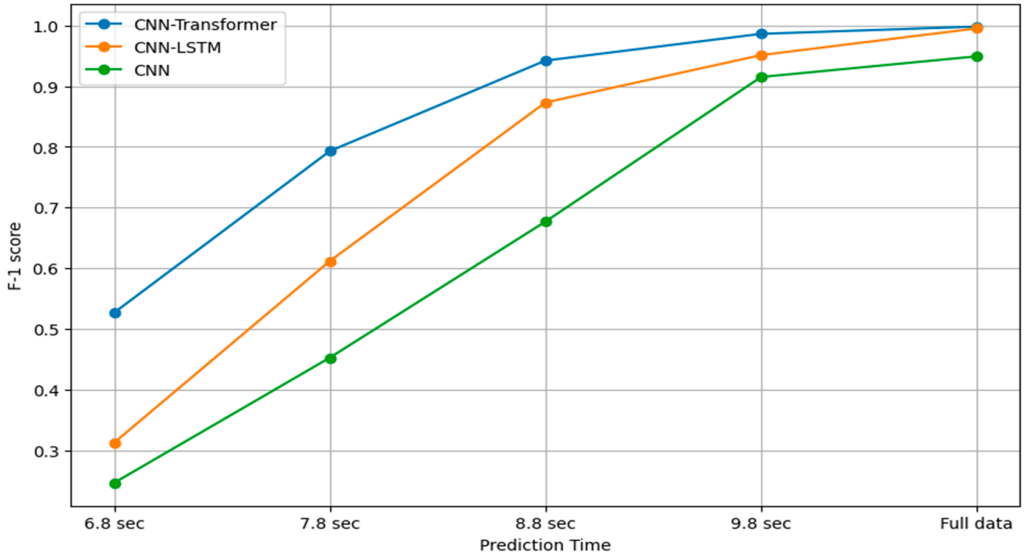

Figure 9. Performance of all trained models across different prediction horizons.

In Figure 9, the 'full data' condition represents the framework having unrestricted access to the continuous sensory data stream and autonomously performing segmentation using change point detection. Importantly, this does not mean the model uses long fixed time windows (e.g., ~10 s) for training or inference. Rather, the effective prediction horizon can be as short as one second, with the remaining sequence padded as needed. This flexibility allows the system to optimize segmentation and achieve the best performance. Figure 9 illustrates the performance of three different neural network architectures, CNN-Transformer, CNN-LSTM, and CNN, over varying prediction horizons. As seen in Figure 9, the CNN-Transformer consistently outperformed the others across all time intervals, especially as the prediction time decreases. All models' performances improve over longer prediction time, which is typical as more information provides better context and feature recognition for predictions. The CNN-Transformer outperforms the other models, suggesting that its architecture is most effective for the tasks involving both spatial and temporal data analysis. CNN-LSTM falls in the middle, offering a balance between temporal dynamics processing and feature extraction, while the CNN lacks the capabilities to excel on its own in tasks requiring understanding of temporal dynamics.

5.2. Comparison with Prior Work and State of the Art

Prior research on intention recognition in HRI has involved instances of using wearable technology (e.g., IMUs), video-based tracking, and physiological signals in real-world yet safe setup environments [34,36,38]. While these studies have provided valuable insights, they are often limited due to environmental noise, occlusion, and hardware constraints. Notably, the use of VR as a platform for intention recognition remains underexplored, and our study is among the first to leverage VR to collect large-scale, high-resolution, multimodal data on human activities in manufacturing-like settings. For example, [36] used probabilistic state machines for intention recognition in human–robot collaboration, while [38]

applied wrist-worn IMUs for behavior modeling in industrial assembly lines. More recently, [34] employed EEG signals to detect upper-limb movement intentions in advance, and [44] combined human–object interaction models with graph convolutional networks for assembly intention recognition. Although these approaches have demonstrated success, they often face challenges in generalizability, data sparsity, and real-time performance.

Our work advances the state of the art by introducing VR as a controlled and flexible experimental platform for studying intention recognition in human–robot interaction. This setup enables the simulation of diverse manufacturing tasks and the generation of large, temporally rich datasets with minimal environmental interference, ensuring high consistency and repeatability across experiments. Compared to prior sensor-based and vision-based systems, the VR-based approach provides precise control over task conditions and environmental variability, improving the consistency of model training and evaluation.

Importantly, while previous studies have applied deep learning methods such as LSTM networks and multimodal fusion models to intention recognition tasks [30], our study introduces a novel CNN-Transformer hybrid architecture that leverages both spatial and temporal dependencies, achieving superior performance in offline intention recognition. We explicitly frame our contribution as a feasibility and baseline study designed to explore the technical potential of integrating VR testbeds with advanced neural architectures, rather than presenting a fully deployable real-time system. While we recognize the limitations of our participant pool, which we explicitly acknowledge in the manuscript, we demonstrate the potential of combining dense, high-quality temporal data with innovative modeling approaches. This positions our work as a complementary and foundational contribution that bridges the gap between controlled experimental studies and the complex demands of real-world manufacturing environments, paving the way for future large-scale and real-time implementations.

5.3. Limitations and Real-World Implementation

While our CNN-Transformer model demonstrated strong precision and recall within the controlled VR environment, several important considerations and opportunities for future work remain.

First, although the study involved a modest participant group (n = 6), we collected a rich and detailed dataset comprising over 1.5 million fine-grained time-stamped data points from multiple sessions. This provided robust material for training and testing our models. That said, we recognize that the participant pool's size limits generalizability across broader populations, such as those varying in age, body type, or movement styles. We have transparently acknowledged this limitation and see it as an exciting direction for future work, where expanding to larger and more diverse participant pools will allow us to rigorously test the generalizability and scalability of the approach.

Second, while VR offers significant advantages as a controlled, flexible, and replicable experimental platform, we fully acknowledge that real-world manufacturing environments bring additional complexities—including unpredictable worker behaviors, occlusions, varying lighting, and operational disruptions—that were outside the scope of this study. We view our VR-based findings as an important feasibility step, providing foundational insights that will support future field deployments.

Third, we note that our framework currently relies on specialized VR hardware (e.g., HTC Vive, Leap Motion), which, while effective in a research setting, may present scalability challenges for widespread industrial use. Future extensions of this work will explore adaptation to industrial-grade sensors and more cost-effective solutions, such as wearable IMUs or camera-based systems, to enhance practical applicability.

Additionally, although the system is designed for real-time operation, this study focused on offline analysis to first establish and benchmark model performance. Moving forward, integrating the framework into live human–robot collaborative systems, evaluating real-time latency, and validating closed-loop control performance are essential next steps that we are eager to pursue.

Finally, while the current framework uses a bottom-up approach—predicting fine-grained actions (e.g., walking, standing) without requiring explicit high-level goal modeling—we see exciting opportunities for future work to incorporate hierarchical, top-down reasoning. Such integration would allow the system to connect low-level actions to overarching collaborative goals, improving both interpretability and scalability for more complex multi-agent scenarios.

6. Conclusions

In conclusion, our research proposes an intention recognition framework, while presenting a comprehensive exploration of utilizing neural networks, particularly within a manufacturing context. With an extensive dataset comprising over 150 million data points collected using wearable technologies and VR, a robust data preprocessing pipeline that involved cleaning the data, inputting missing values, padding sequences for consistency, converting data into a 3D tensor, and performing one-hot encoding on the target labels is deployed. CNN, CNN-LSTM, and CNN-Transformer models are trained and evaluated against a range of metrics including accuracy, precision, recall, and the F1-score. Our findings revealed that while all three models demonstrated high efficacy in intention recognition, the ensemble CNN-Transformer model outperforms in terms of precision, recall, and overall accuracy. As the CNN-Transformer model demonstrated superior performance on the original data, our findings open avenues for ongoing research aimed at early intention recognition tasks. This study brings us one step closer to achieving our goal: robust, reliable, and comprehensive human intention recognition, thereby paving the way for a new era of productivity and safety in the industry. Although the framework achieved high accuracy, its generalizability is constrained by three factors: (i) a small, demographically limited participant pool, (ii) data collected in a virtual-reality cell rather than on a physical factory floor, and (iii) evaluation without the robot closed-looped for end-to-end latency measurement. Future studies will include more diverse operators, real-world data capture, extending to top-down integration of goal-task hierarchies, and full robot-in-the-loop validation to address these gaps.

Author Contributions: Conceptualization, A.K.M. and S.M.; Data curation, A.K.M.; Formal analysis, A.K.M. and S.M.; Methodology, A.K.M., E.A. and S.M.; Resources, S.M.; Software, A.K.M. and E.A.; Supervision, S.M.; Validation, A.K.M., E.A. and S.M.; Visualization, S.M.; Writing—original draft, A.K.M., E.A. and S.M.; Writing—review & editing, A.K.M., E.A. and S.M. 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 project does not constitute human participant research according to the definition codified in the Common Rule at 45 CFR 46 and FDA regulations. This means that IRB review and oversight is not required for this project.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

Abbreviations

The following abbreviations are used in this manuscript:

HRI	Human–Robot Interaction
VR	Virtual Reality
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory
Cobot	Collaborative Robot
MFIRA	Multimodal Fusion Intent Recognition Algorithm
EEG	Electroencephalogram
EMG	Electromyography
UCI	University of California, Irvine
WSDM	Wireless Sensor Data Mining
GAMI	Gaze and Motion Information
CPA	Change Point Analysis
RBF	Radial Basis Function
ReLU	Rectified Linear Unit
BIC	Bayesian Information Criterion
RNN	Recurrent Neural Network
LED	Light-emitting Diode

References

1. Huang, Z.; Mun, Y.J.; Li, X.; Xie, Y.; Zhong, N.; Liang, W.; Geng, J.; Chen, T.; Driggs-Campbell, K. Seamless interaction design with coexistence and cooperation modes for robust human-robot collaboration. *arXiv* **2022**, arXiv:2206.01775.
2. Wang, N.; Zeng, Y.; Geng, J. A brief review on safety strategies of physical human-robot interaction. *ITM Web Conf. EDP Sci.* **2019**, *25*, 01015. [CrossRef]
3. Dhanda, M.; Rogers, B.A.; Hall, S.; Dekoninck, E.; Dhokia, V. Reviewing human-robot collaboration in manufacturing: Opportunities and challenges in the context of industry 5.0. *Robot Comput. Integr Manuf.* **2025**, *93*, 102937. [CrossRef]
4. Jahanmahin, R.; Masoud, S.; Rickli, J.; Djuric, A. Human-robot interactions in manufacturing: A survey of human behavior modeling. *Robot Comput. Manuf.* **2022**, *78*, 102404. [CrossRef]
5. Bottega, J.A.; Steinmetz, R.; Kolling, A.H.; Kich, V.A.; De Jesus, J.C.; Grando, R.B.; Gamarra, D.F.T. Virtual reality platform to develop and test applications on human-robot social interaction. In Proceedings of the IEEE 2022 Latin American Robotics Symposium (LARS), 2022 Brazilian Symposium on Robotics (SBR), and 2022 Workshop on Robotics in Education (WRE), São Bernardo do Campo, Brazil, 18–21 October 2022; pp. 1–6.
6. Walker, M.; Phung, T.; Chakraborti, T.; Williams, T.; Szafir, D. Virtual, augmented, and mixed reality for human-robot interaction: A survey and virtual design element taxonomy. *ACM Trans Hum Robot. Interact* **2023**, *12*, 1–39. [CrossRef]
7. Bai, C.; Dallasega, P.; Orzes, G.; Sarkis, J. Industry 4.0 technologies assessment: A sustainability perspective. *Int. J. Prod. Econ.* **2020**, *229*, 107776. [CrossRef]
8. Mohammadzadeh, A.K.; Allen, C.L.; Masoud, S. VR Driven Unsupervised Classification for Context Aware Human Robot Collaboration. In *International Conference on Flexible Automation and Intelligent Manufacturing*; Springer: Cham, Switzerland, 2023; pp. 3–11.
9. Kofer, D.; Bergner, C.; Deuerlein, C.; Schmidt-Vollus, R.; Heß, P. Human–robot-collaboration: Innovative processes, from research to series standard. *Procedia CIRP* **2021**, *97*, 98–103. [CrossRef]
10. Vrigkas, M.; Nikou, C.; Kakadiaris, I.A. A review of human activity recognition methods. *Front. Robot. AI* **2015**, *2*, 28. [CrossRef]
11. Vasconez, J.P.; Kantor, G.A.; Cheein, F.A.A. Human–robot interaction in agriculture: A survey and current challenges. *Biosyst. Eng.* **2019**, *179*, 35–48. [CrossRef]
12. Hu, Z.; Zhang, Y.; Li, Q.; Lv, C. Human–machine telecollaboration accelerates the safe deployment of large-scale autonomous robots during the COVID-19 pandemic. *Front. Robot. AI* **2022**, *9*, 853828. [CrossRef]
13. Liu, C.; Li, X.; Li, Q.; Xue, Y.; Liu, H.; Gao, Y. Robot recognizing humans intention and interacting with humans based on a multi-task model combining ST-GCN-LSTM model and YOLO model. *Neurocomputing* **2021**, *430*, 174–184. [CrossRef]
14. Liu, R.; Chen, R.; Abuduweili, A.; Liu, C. Proactive human-robot co-assembly: Leveraging human intention prediction and robust safe control. In Proceedings of the IEEE 2023 IEEE Conference on Control Technology and Applications (CCTA), Bridgetown, Barbados, 16–18 August 2023; pp. 339–345.

15. Robinson, N.; Tidd, B.; Campbell, D.; Kulić, D.; Corke, P. Robotic vision for human-robot interaction and collaboration: A survey and systematic review. *ACM Trans. Hum. Robot. Interact.* **2023**, *12*, 1–66. [CrossRef]
16. Kumar, P.; Chauhan, S.; Awasthi, L.K. Human activity recognition (har) using deep learning: Review, methodologies, progress and future research directions. *Arch. Comput. Methods Eng.* **2024**, *31*, 179–219. [CrossRef]
17. Katariya, M.; Mathur, A.; Singh, S. Human–Robot/Machine Interaction for Sustainable Manufacturing: Industry 5.0 Perspectives. In *Handbook of Intelligent and Sustainable Manufacturing*; CRC Press: Boca Raton, FL, USA, 2025; pp. 18–41.
18. Didwania, R.; Verma, R.; Dhanda, N. Application of Robotics in Manufacturing Industry. In *Machine Vision and Industrial Robotics in Manufacturing*; CRC Press: Boca Raton, FL, USA, 2024; pp. 57–84.
19. Othman, U.; Yang, E. Human–robot collaborations in smart manufacturing environments: Review and outlook. *Sensors* **2023**, *23*, 5663. [CrossRef]
20. Lasota, P.A.; Fong, T.; Shah, J.A. A survey of methods for safe human-robot interaction. *Found. Trends® Robot* **2017**, *5*, 261–349. [CrossRef]
21. AlShorman, O.; Alshorman, B.; Masadeh, M.S. A review of physical human activity recognition chain using sensors. *Indones. J. Electr. Eng. Inform.* **2020**, *8*, 560–573.
22. Rafferty, J.; Nugent, C.D.; Liu, J.; Chen, L. From activity recognition to intention recognition for assisted living within smart homes. *IEEE Trans. Hum. Mach. Syst.* **2017**, *47*, 368–379. [CrossRef]
23. Darafsh, S.; Ghidary, S.S.; Zamani, M.S. Real-time activity recognition and intention recognition using a vision-based embedded system. *arXiv* **2021**, arXiv:2107.12744.
24. Lei, Y.; Su, Z.; He, X.; Cheng, C. Immersive virtual reality application for intelligent manufacturing: Applications and art design. *Math. Biosci. Eng.* **2023**, *20*, 4353–4387. [CrossRef]
25. Inamura, T.; Mizuchi, Y. Sigverse: A cloud-based vr platform for research on multimodal human-robot interaction. *Front. Robot. AI* **2021**, *8*, 549360. [CrossRef]
26. Wonsick, M.; Padir, T. A systematic review of virtual reality interfaces for controlling and interacting with robots. *Appl. Sci.* **2020**, *10*, 9051. [CrossRef]
27. Hu, Y.; Zhang, X.-Q.; Xu, L.; He, F.X.; Tian, Z.; She, W.; Liu, W. Harmonic loss function for sensor-based human activity recognition based on LSTM recurrent neural networks. *IEEE Access* **2020**, *8*, 135617–135627. [CrossRef]
28. Huang, W.; Zhang, L.; Gao, W.; Min, F.; He, J. Shallow convolutional neural networks for human activity recognition using wearable sensors. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–11. [CrossRef]
29. Park, K.-B.; Choi, S.H.; Lee, J.Y.; Ghasemi, Y.; Mohammed, M.; Jeong, H. Hands-free human–robot interaction using multimodal gestures and deep learning in wearable mixed reality. *IEEE Access* **2021**, *9*, 55448–55464. [CrossRef]
30. Xia, Z.; Feng, Z.; Yang, X.; Kong, D.; Cui, H. MFIRA: Multimodal fusion intent recognition algorithm for AR chemistry experiments. *Appl. Sci.* **2023**, *13*, 8200. [CrossRef]
31. Peng, J.; Kimmig, A.; Wang, D.; Niu, Z.; Tao, X.; Ovtcharova, J. Intention recognition-based human–machine interaction for mixed flow assembly. *J. Manuf. Syst.* **2023**, *72*, 229–244. [CrossRef]
32. Xu, W.; Huff, T.; Ye, S.; Sanchez, J.R.; Rose, D.; Tung, H.; Tong, Y.; Hatcher, J.; Klein, M.; Morales, E.; et al. Virtual reality-based human-robot interaction for remote pick-and-place tasks. In Proceedings of the Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction, Boulder, CO, USA, 11–15 March 2024; pp. 1148–1152.
33. Masoud, S.; Chowdhury, B.; Son, Y.-J.; Kubota, C.; Tronstad, R. A dynamic modelling framework for human hand gesture task recognition. *arXiv* **2019**, arXiv:1911.03923.
34. 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. Manuf.* **2021**, *70*, 102137. [CrossRef]
35. Zhang, Y.; Ding, K.; Hui, J.; Lv, J.; Zhou, X.; Zheng, P. Human-object integrated assembly intention recognition for context-aware human-robot collaborative assembly. *Adv. Eng. Inform.* **2022**, *54*, 101792. [CrossRef]
36. Awais, M.; Henrich, D. Human-robot collaboration by intention recognition using probabilistic state machines. In Proceedings of the IEEE 19th International Workshop on Robotics in Alpe-Adria-Danube Region (RAAD 2010), Balatonfured, Hungary, 24–26 June 2010; pp. 75–80.
37. Stiefmeier, T.; Ogris, G.; Junker, H.; Lukowicz, P.; Troster, G. Combining motion sensors and ultrasonic hands tracking for continuous activity recognition in a maintenance scenario. In Proceedings of the 2006 10th IEEE International Symposium on Wearable Computers, Montreux, Switzerland, 11–14 October 2006; pp. 97–104.
38. Koskimäki, H.; Huikari, V.; Siirtola, P.; Röning, J. Behavior modeling in industrial assembly lines using a wrist-worn inertial measurement unit. *J. Ambient. Intell. Humaniz. Comput.* **2013**, *4*, 187–194. [CrossRef]
39. Maekawa, T.; Nakai, D.; Ohara, K.; Namioka, Y. Toward practical factory activity recognition: Unsupervised understanding of repetitive assembly work in a factory. In Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Heidelberg, Germany, 12–16 September 2016; pp. 1088–1099.

40. Zhu, L.; Wang, Z.; Ning, Z.; Zhang, Y.; Liu, Y.; Cao, W.; Wu, X.; Chen, C. A novel motion intention recognition approach for soft exoskeleton via IMU. *Electronics* **2020**, *9*, 2176. [CrossRef]
41. Sun, B.; Cheng, G.; Dai, Q.; Chen, T.; Liu, W.; Xu, X. Human motion intention recognition based on EMG signal and angle signal. *Cogn. Comput. Syst.* **2021**, *3*, 37–47. [CrossRef]
42. Zhang, T.; Sun, H.; Zou, Y. An electromyography signals-based human-robot collaboration system for human motion intention recognition and realization. *Robot Comput. Manuf.* **2022**, *77*, 102359. [CrossRef]
43. Li, S.; Zheng, P.; Liu, S.; Wang, Z.; Wang, X.V.; Zheng, L.; Wang, L. Proactive human–robot collaboration: Mutual-cognitive, predictable, and self-organising perspectives. *Robot Comput. Integr Manuf.* **2023**, *81*, 102510. [CrossRef]
44. Zhang, Z.; Peng, G.; Wang, W.; Chen, Y.; Jia, Y.; Liu, S. Prediction-based human-robot collaboration in assembly tasks using a learning from demonstration model. *Sensors* **2022**, *22*, 4279. [CrossRef]
45. Sun, X.; Zhang, R.; Liu, S.; Lv, Q.; Bao, J.; Li, J. A digital twin-driven human–robot collaborative assembly-commissioning method for complex products. *Int. J. Adv. Manuf. Technol.* **2022**, *118*, 3389–3402. [CrossRef]
46. Wang, S.; Mao, Z.; Zeng, C.; Gong, H.; Li, S.; Chen, B. A new method of virtual reality based on Unity3D. In Proceedings of the 2010 18th International Conference on Geoinformatics, Beijing, China, 18–20 June 2010; pp. 1–5. [CrossRef]
47. Buck, L.; Paris, R.; Bodenheimer, B. Distance Compression in the HTC Vive Pro: A Quick Revisitation of Resolution. *Front. Virtual Real.* **2021**, *2*, 728667. [CrossRef]
48. Weichert, F.; Bachmann, D.; Rudak, B.; Fisseler, D. Analysis of the accuracy and robustness of the leap motion controller. *Sensors* **2013**, *13*, 6380–6393. [CrossRef]
49. Jameer, S.; Syed, H. A DCNN-LSTM based human activity recognition by mobile and wearable sensor networks. *Alex. Eng. J.* **2023**, *80*, 542–552. [CrossRef]
50. Khare, S.; Sarkar, S.; Totaro, M. Comparison of sensor-based datasets for human activity recognition in wearable IoT. In Proceedings of the 2020 IEEE 6th World Forum on Internet of Things (WF-IoT), New Orleans, LA, USA, 5–9 April 2020; pp. 1–6.
51. Aminikhanghahi, S.; Cook, D.J. A survey of methods for time series change point detection. *Knowl. Inf. Syst.* **2017**, *51*, 339–367. [CrossRef]
52. Yu, Y.; Lee, M.; Lee, C.; Cheon, Y.; Baek, S.; Kim, Y.; Kim, K.; Jung, H.; Lim, D.; Byun, H.; et al. Estimating APC Model Parameters for Dynamic Intervals Determined Using Change-Point Detection in Continuous Processes in the Petrochemical Industry. *Processes* **2023**, *11*, 2229. [CrossRef]
53. Platias, C.; Petasis, G. A comparison of machine learning methods for data imputation. In Proceedings of the 11th Hellenic Conference on Artificial Intelligence, Athens, Greece, 2–4 September 2020; pp. 150–159.