

Privacy-Preserving 3D Lidar-Based Multi-Modal Activity Recognition in Human-Robot Interaction

Adel Baselizadeh^{1,2}, Md Zia Uddin³, Weria Khaksar⁴, Diana Saplacan Lindblom¹, Jim Torresen^{1,2}

¹Department of Informatics, University of Oslo, Oslo, Norway

²RITMO Centre for Interdisciplinary Studies in Rhythm, Time, and Motion, University of Oslo, Oslo, Norway

³Department of Sustainable Communication Technologies, SINTEF Digital, Oslo, Norway

⁴Faculty of Science and Technology, Norwegian University of Life Sciences, Ås, Norway
{adelb, dianasa, jimtoer}@ifi.uio.no, zia.uddin@sintef.no, weria.khaksar@nmbu.no

Abstract—Human activity recognition (HAR) involves using sensors to collect human data, which is then analyzed to identify their activities. HAR has numerous applications across various fields. One significant area is human-robot interaction (HRI). Recognizing user activities provides robots with valuable insights into the user's status, enhancing the efficiency of HRI. However, developing HAR models presents privacy challenges due to the collection of user data, especially in robots, which are often equipped with diverse sensors. This paper proposes privacy-preserving HAR methods within the context of HRI.

The paper investigates using privacy-preserving sensors, with a particular focus on 3D Lidar, to develop HAR models. It explores the integration of 3D Lidar with other sensors, including user-wearable sensors and robot-based sensors, including force/torque sensors, through multimodal deep learning (DL) approaches. Various DL-based sensor fusion methods, including data-level and feature-level fusion approaches, are thoroughly examined, and their accuracies for HAR are compared.

A novel dataset was collected to train the multimodal DL models, capturing various user activities during HRI. This dataset leverages 10 different sensors, including 9 privacy-preserving sensors, along with an RGB camera for reference. The paper considers nine distinct user activities, including physical interactions with a robot and commanding a robot to perform specific tasks. The results indicate that integrating the sensors' data at the feature-level achieves an 80.73% accuracy in recognizing various user activities during HRI.

Index Terms—Human activity recognition, HAR, human-robot interaction, HRI, 3D Lidar, multi-modal sensing, privacy, dataset.

I. INTRODUCTION

Human activity recognition (HAR) has garnered significant interest across various fields due to its ability to provide detailed information about user status and activities [1]. HAR is useful in a wide range of applications, including recognizing daily-life activities, detecting emergency situations, enhancing autonomous vehicle systems, and supporting ambient assisted living [2].

With the increasing presence of robots in various aspects of our lives, improving HRI has become paramount. One way to enhance HRI is to equip robots with HAR algorithms, making them able to recognize user activities [3]. This capability allows robots to understand users' status and activities better, empowering them to make more informed decisions in different HRI scenarios.

Robots serve as versatile platforms that can integrate various types of sensors, offering excellent opportunities for collecting

diverse user data. This data is crucial for developing robust HAR models. Additionally, due to their ability to be equipped with multiple sensors, robots are highly suitable for developing multimodal and HAR models, which have demonstrated superior performance in advanced HAR applications [4].

Several research studies have focused on developing HAR algorithms within the context of HRI. Some papers have proposed HAR methods specifically tailored for recognizing human activities during interactions with industrial robots [5]. These efforts aim to enhance the efficiency of utilizing robots in industrial settings, where they often work in collaboration with human operators. Similarly, there are research efforts addressing HAR in the context of agricultural robots [6].

Recently, significant attention has been paid to developing a new category of robots known as care robots and personal robot assistants. These robots are designed to be intimately involved in users' lives, operating close to individuals and offering support and assistance [7]. One crucial application of these robots is HAR to detect emergency situations, offering round-the-clock monitoring in home environments or health-care facilities like elderly care centers [8]. Thus, implementing practical HAR algorithms on these robots holds great importance. However, despite research addressing HAR in HRI across various robotic settings, there remains a noticeable gap in the field of care and personal robots.

Developing HAR models, especially those based on AI and DL, necessitates access to suitable datasets for training [9]. While various HAR datasets are available for different contexts, encompassing diverse sensors and human activities, there is a lack of a specific dataset tailored for HAR in HRI within the domain of care and personal assistant robots. These robots interact with users for different purposes, covering a broader range of activities than industrial or agricultural robots, which typically perform specific tasks. Consequently, datasets designed explicitly for HAR in HRI within the context of care/personal robots need to include this diversity in activities. Moreover, given that these robots are typically equipped with various sensors, these datasets can gather a wide range of user data. Such datasets offer the opportunity to develop HAR models based on diverse sensor data and facilitate the development of multimodal HAR models, which have shown notable performance improvements in HAR [4].

Additionally, user privacy is a crucial aspect to consider regarding HAR in the context of these robots [10]–[13].

Given their deployment in close proximity to users in their homes and care facilities, these robots have the potential to collect sensitive user data. Therefore, respecting user privacy in interactions with these robots is paramount. Consequently, developing HAR models for HRI in these robots should be approached with privacy concerns in mind [14]. This may entail exploring the utilization of privacy-preserving sensors that avoid collecting sensitive user data.

The 3D Lidar sensor has found widespread applications across various domains, including autonomous vehicles [15], drones [16], and robotics [17]. Its integration into robotics applications is on the rise, with many robots now being equipped with this sensor due to its capabilities in tasks such as robot navigation [18]. However, exploring this sensor's applicability for HAR has been limited, with only a few research efforts investigating its potential in this area [19]. It's worth noting that in the robotics literature, 3D Lidar is considered a privacy-preserving sensor [20]. Given the type of data it collects, 3D Lidar can provide sufficient information for HAR while gathering less detailed or sensitive user data, thereby respecting user privacy. Moreover, its ability to function independently of lighting and weather conditions makes it a suitable option for HAR in various contexts. Consequently, this sensor appears to be well-suited for HAR purposes while ensuring the preservation of user privacy. Additionally, given their prominent usage in robots, they can be integrated with other robot sensors, enabling multimodal sensing approaches.

This paper endeavors to develop privacy-preserving HAR algorithms for HRI within care and personal assistant robots. It focuses particularly on leveraging multimodal sensing algorithms, with a special emphasis on utilizing a 3D Lidar sensor. The contributions of this paper can be summarized as follows:

- 1) Developing DL-based HAR models, leveraging combinations of Convolutional Neural Networks (CNN) [21], Long Short-Term Memory (LSTM) [22], and Seq2Seq [23] and PointNet++ [24] architectures.
- 2) The integration of 3D Lidar sensors with other sensors, including user-worn (wrist) sensors and the robot's force/torque sensors, the first instance of such integration for HAR purposes.
- 3) Developing various multi-modal HAR algorithms for integrating data from the aforementioned sensors, including data-level and feature-level fusion techniques.

II. RELATED WORK

A. 3D Lidar-Based HAR

3D Lidar has emerged as an important sensor in robotics with diverse applications such as 3D mapping of the environment and navigation. However, its ability to collect detailed data, including human body gestures, makes it also potentially useful for HAR purposes. Despite its potential, the utility of 3D Lidar for HAR has not been as extensively studied as for other applications.

Notable research in the use of 3D Lidar for HAR includes the work of Roche et al., who utilized a combination of 3D Lidar and an RGB camera for activity recognition, specifically human daily life activities [19]. The RGB camera was

employed to detect user positions, followed by applying a CNN architecture to the 3D Lidar data to recognize user activities. In a separate study [22], Yu et al. utilized 3D Lidar alone for HAR, employing a density-based spatial clustering of applications with noise framework integrated with a DL network based on LSTM units to recognize human daily life activities. Another study by Dang et al. proposed a DL framework for recognizing four common types of human movements in daily life using 3D Lidar [25].

As previously mentioned, the use of 3D Lidar for HAR is still limited, with only a few studies exploring this area. A major limitation is the narrow focus on human daily life activities in the existing studies. Furthermore, the integration of 3D Lidar with other sensors is also limited, with the only integration being with an RGB camera in the work of Roche et al. [19]. Thus, there is still significant potential to address these limitations, which has been attempted in this work. Firstly, the focus has been expanded to a new category of activities, specifically HAR during HRI. This broadens the scope and applicability of 3D Lidar-based HAR from only daily life to other types of activities. Secondly, the integration of 3D Lidar with other privacy-preserving sensors, such as wearable and robot sensors, has been explored. This opens up new possibilities in multimodal sensing using 3D Lidar, which can result in more comprehensive and accurate HAR algorithms. Lastly, commonly used DL architectures for HAR, such as CNN and LSTM, along with established DL architectures, including the Seq2Seq mechanism and Pointnet++ framework, have been integrated and applied to developing HAR models. This paper aims to overcome the limitations identified in prior research and contribute to advancing 3D Lidar-based HAR.

B. Multi-modal DL approaches for HAR

Multimodal sensing in HAR has gained attention for its potential to improve activity recognition by combining data from multiple sensors [26]. Compared to using a single sensor, this approach offers greater robustness and accuracy. If one sensor fails to capture data, others can compensate, reducing the risk of data loss. Additionally, different sensors capture different activity features, allowing for a more comprehensive understanding. Overall, multimodal sensing enhances the reliability and effectiveness of HAR systems in real-world scenarios.

When it comes to multimodal DL for HAR, there are various approaches for integrating data from different sensors. Three commonly used approaches for sensor data integration in multimodal HAR are:

- *Data-level fusion*: combines raw data from multiple sensors before feeding it into DL models for activity recognition [27].
- *Feature-level fusion*: extracts features from each sensor using separate DL models, then merges them for further processing [28].
- *Decision-level fusion*: performs classification on each sensor's data independently, then combines the decisions via a high-level classifier [29].

This paper explores the integration of three different sensors, including 3D Lidar, user wrist-wearable sensors, and

force/torque sensors of a robot, using data-level and feature-level fusion approaches. A comparison is made of the accuracy of HAR achieved with each fusion method to determine the effectiveness of sensor integration for activity recognition.

III. THE PRIMA-CARE DATASET

PriMA-Care (Privacy-preserving Multimodal dataset for Human Activity Recognition in Care Robots) was collected to support HAR algorithm development in HRI contexts [30]. The dataset was gathered entirely through a robot-based system, with all sensors (except one wearable) integrated into the TIAGo robot [31], which is designed for assistive use in care environments. Due to technical limitations, data from the wearable sensor was collected separately.

The dataset includes 27 activities reflecting typical care scenarios, grouped into four categories: 1) Daily routines, 2) Physical HRI, 3) Robot control via hand gestures, and 4) User movements near the robot. These cover both user-robot interactions and user-only monitoring, aiming to enable more robust HAR models. The list of activities is provided in [30].

A. Chosen activities

As outlined earlier, the PriMA-Care dataset includes twenty-seven different activities, which involve interactions between a robot and a user or monitoring of the user by a robot in their home or care facility environments. These activities are categorized into four groups. Among these groups, two—physical HRI and user-commanded robot activities—are directly related to HRI. Since this paper focuses on studying HAR in the context of HRI, we have selected specific activities from these two categories in the PriMA-Care dataset. Accordingly, nine activities from these two groups have been selected to develop HAR models: "Giving an object to the robot," "Holding the robot's hand while walking," "Interacting with the robot dangerously," "Come towards me," "Get away from me," "Rotate," "Stop," "Bring me object 1," and "Bring me object 2." Fig. 3 displays the nine activities from PriMA-Care examined in this paper, captured by the 3D Lidar sensor.

B. Selected sensors

In this paper, we have focused on three specific sensors as a first step toward developing multimodal HAR models using the PriMA-Care dataset. One of the objectives is to investigate the potential applicability of 3D Lidar for HAR, making it a central component of our study. Additionally, since some of the activities involve physical HRI, force/torque sensors are crucial for recognizing these interactions, and thus, they were included in the study. Furthermore, activities involving user commands to the robot require users to move their hands in specific patterns, making wrist-worn wearable sensors beneficial for recognizing these activities. Therefore, a wrist-worn wearable sensor and force/torque sensors were chosen in addition to 3D Lidar to recognize the activities. All these sensors offer the advantage of being privacy-preserving. Their integration was considered to develop multimodal HAR models. The specifications of each sensor can be found in [30].

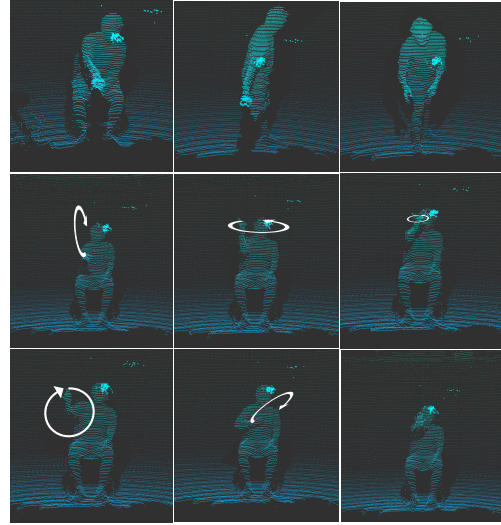


Fig. 1. The activities chosen from the PriMA-Care dataset for developing HAR models. From left to right: Top row: giving an object to the robot, holding the robot's hand while walking, interacting with the robot dangerously. Middle row: come towards me, rotate, bring me object 1. Bottom row: bring me object 2, get away from me, stop.



Fig. 2. ConVLSTM: the DL architecture used for processing wearable and force/torque sensor data.

IV. DL-BASED SENSOR DATA PROCESSING

The DL-based processing of sensor data for HAR involves two main steps. The first step entails processing the data from each sensor individually (for data-level fusion) or together (for feature-level fusion) to extract relevant features. The second step involves integrating these features using DL frameworks to recognize activities. This section will detail the DL architectures employed for processing data from each sensor.

The wearable sensor collects data from the accelerometer, gyroscope, and magnetometer in three directions at a sampling rate of 54 Hz, resulting in 9 data streams. The force/torque sensors measure forces and torques applied to the end-effector in three directions at 50 Hz, producing 6 data streams. The 3D Lidar sensor operates with different sampling rates and a number of data streams. For the PriMA-Care dataset collection, we used a 20 Hz sampling rate, with 128 data collection channels and a resolution of 1024. This results in 128×1024 data points at each timestamp, generating a point cloud with 131072 (128×1024) points.

To synchronize the data from all sensors, the wearable and force/torque sensors' data were downsampled to 20 Hz. This ensures that for each point cloud collected by the 3D Lidar at each timestamp, there are corresponding values from the wearable and force/torque sensors. This synchronization is crucial for developing multimodal HAR models by integrating data from different sensors.

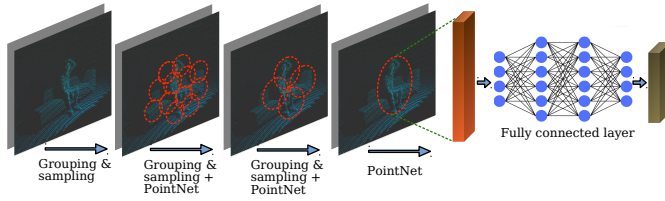


Fig. 3. A schematic of the PointNet++ architecture used for processing 3D Lidar point clouds. PointNet++ has a multi-layered structure that samples and groups the point cloud, applying the PointNet framework to each group iteratively, ultimately mapping the entire point cloud to a feature vector.

A. Wearable and force/torque sensors data processing

Due to the similar structures of the data collected by these sensors, i.e., data streams at 20 Hz after downsampling, similar DL architectures were employed for processing their data. These structures utilized a combination of CNN, LSTM, and FC layers. Fig. 4 illustrates the DL architecture designed for processing the sensor data (ConvLSTM).

As per existing literature, DL-based HAR on wearable sensor data or similar sources, which produce outputs in the form of time series, demonstrate higher accuracies when processing their data sequentially rather than at individual timestamps. Consequently, in the processing of wearable and force/torque sensors, we adopted a windowed approach, considering a sequence of 20 timestamps, equivalent to 1 second of data, fed into the DL models with an overlap of 10 timestamps, spanning 0.5 seconds, between these windows.

As shown in Fig. 4, the DL architecture for processing wearable and force/torque sensor data consists of 12 layers. The layers are color-coded as follows: blue for CNN, yellow for MaxPooling, green for Dropout, red for LSTM, and purple for the Fully Connected layer. The first two layers, indicated in blue, are 1-dimensional CNN layers with dimensions $(A, 200, 5)$. Here, A equals 6, 9, or 15, depending on whether the layer is applied to force/torque sensors, wearable sensors, or both, respectively. The kernel size for these CNN layers is 5. The third layer is a max-pooling layer with a size of 3. The fourth layer is a dropout layer with a dropout rate of 0.25. Layers 5 and 6 are also 1D CNN layers with dimensions $(200, 200, 2)$, where 2 is the kernel size. The seventh layer is another max-pooling layer, again with a size of 3. Layers 8 and 10 are LSTM layers with sizes $(200, 128)$ and $(128, 128)$, respectively. Layers 9 and 11 are dropout layers with a dropout rate of 0.25. The final layer is a fully connected layer that maps the 128 features to the number of classes, assuming the HAR task is applied to the sensor data without integration with 3D Lidar data.

B. 3D Lidar data processing

The 3D Lidar sensor collects data in the form of 3D point clouds, demanding specialized DL architectures for effective data processing. Standard DL models are often inadequate for this task. To address this, specific DL models are often utilized for 3D Lidar data processing. Among these, the PointNet++ architecture is frequently used in the literature for processing

3D Lidar point cloud data. PointNet++ excels at extracting meaningful features from this data for further analysis.

PointNet++ [24] is suitable for both classification and segmentation tasks on point clouds. Fig. 5 illustrates the PointNet++ architecture. Building upon the original PointNet architecture [32], PointNet++ employs a multi-layered approach to identify local features at various scales. It segments the input point cloud into overlapping regions through a systematic sampling and grouping process. Within each region, PointNet++ applies PointNet layers to analyze and learn local features. These local features are then progressively integrated to form a comprehensive global representation of the entire point cloud.

As mentioned earlier, each point cloud generated by the 3D Lidar comprises 131072 points obtained from 128 channels multiplied by 1024 points in each channel. However, during our training of DL models, we encountered a significant computational challenge when feeding all these data points to the models. The computational cost was prohibitively high, making it impossible for the GPU to handle all these points. In cases where the GPU could manage, the training time for a single step exceeded several hours, sometimes even extending beyond a full day.

To address this issue and make training feasible, we had to significantly reduce the number of point clouds. For this purpose, we extracted only those parts of the point clouds that captured the main movements of the subjects, while disregarding the other parts. As a result, we selected and considered 7300 points for training the models. Although this approach remained time-consuming, with each training step taking around 1 hour, it became possible to train the models based on point clouds within a manageable timeframe.

C. Data-level sensor fusion

In this study, we only fused the data from wearable and force/torque sensors at the data-level, excluding the integration of the 3D Lidar sensor data. This decision was made due to the complex and dissimilar data structure of the 3D Lidar sensor compared to the other sensors. Another rationale for not fusing 3D Lidar data with the other sensors at the data-level is the disparity in data volume. While wearable and force/torque sensors collect a similar amount of data (9 and 6 data streams, respectively), 3D Lidar sensors collect a substantially larger amount of data. Combining the raw data of the 3D Lidar sensor with the other sensors could introduce an imbalance in the input data, potentially biasing the results towards the 3D Lidar data. Therefore, for data-level fusion, we excluded the 3D Lidar sensor.

D. Feature-level sensor fusion

In the second approach of integrating the sensor data, we performed fusion at the feature-level. In this method, each sensor underwent individual processing based on the DL networks presented in Fig. 4 and Fig. 5. The features extracted from each sensor were then concatenated with each other. These concatenated feature tensors were subsequently fed into another DL architecture, typically including a set of fully

connected layers. Finally, these concatenated features were mapped to the activity classes.

Integrating these sensors after window-based processing with the 3D Lidar sensor presents a challenge due to the unequal number of data points between the 20-sample windows and the data from the 3D Lidar. Since the data points from the wearable and force/torque sensors are grouped into 20-sample sequences, there are now fewer of these groups compared to the number of point clouds from the 3D Lidar.

One potential solution to address this challenge would be to apply a similar strategy for processing the 3D Lidar data. This would involve considering a window comprising 20 point clouds from the 3D Lidar and processing this window, incorporating all these data points at each training step. However, implementing this approach is almost impossible.

Firstly, the PointNet++ framework utilized for 3D Lidar data processing cannot be applied to a sequence of 3D Lidar data. Secondly, modifying PointNet++ to accommodate this process, processing a sequence of 20 point clouds at each training step, poses significant computational challenges. Due to the substantial amount of data in the point clouds generated by the 3D Lidar, even powerful processors struggled to handle a combined set of point clouds.

The initial window-based method couldn't handle the unequal data sizes from 3D LiDAR, wearable, and force/torque sensors. To address this, we processed each 3D point cloud individually through PointNet++, extracted features, and saved them. A window of 20 such features was then created and combined with 20-sample windows from the other sensors, ensuring equal input sizes for training.

We used a Seq2Seq model to integrate the 3D LiDAR features. The 20 feature vectors were fed into the encoder, and the decoder produced a single output vector. This was fused with features from the wearable and force/torque sensors.

Figure 7 illustrates this Seq2Seq-based multimodal feature-level fusion approach.

V. RESULTS AND DISCUSSION

In this section, we present the results of the experiments conducted for training the HAR models. Approximately 80% of the data was used for training, while 20% was reserved for testing the trained models. Details of each sensor's data and the HAR models, along with the corresponding figures, are provided in Section IV. Five distinct models were developed based on the utilization of individual sensors and the integration of multiple sensors. These models are as follows:

- 1) Wearable sensor only
- 2) Wearable + force/torque data-level fusion
- 3) Wearable + force/torque feature-level fusion
- 4) 3D Lidar sensor only
- 5) 3D Lidar + wearable + force/torque feature-level fusion

We did not train a model solely on the force/torque sensor because this sensor only collected meaningful data during two specific activities involving physical interaction with the robot through its end-effector. In all other scenarios, the sensor data was largely meaningless and consisted mostly of noise.

All these models were trained on the same training set and tested on the same test set to ensure a fair comparison of model

accuracies. The performance metrics evaluated for each HAR model included accuracy, precision, recall, and F1-score.

Models were trained using an AMD Threadripper 1950X CPU and Nvidia GTX1080ti GPU, with Python and PyTorch. Training for the first three models (wearable and force/torque sensors) was fast, taking only a few minutes. The fourth model (3D LiDAR only) required 14 minutes per step, totaling 2 days for 200 epochs. The fifth, multimodal model (with Seq2Seq for LiDAR), took 1 hour per step, resulting in 8 days of training.

Table I presents the maximum and average scores of precision, recall, accuracy, and F1-score over 200 epochs for each of the trained models applied to the test set. Fig. 9 offers the confusion matrices for all five trained models.

Based on the results presented in Table I, focusing on accuracy, we observe that the model using only wearable sensors achieves the lowest average accuracy of 0.5515 and a maximum accuracy of 0.5900. This indicates that relying solely on wearable sensor data limits the model's ability to capture the complexity of human activities. In contrast, integrating force/torque sensor data with wearable sensors at the data-level results in a modest improvement, with an average accuracy of 0.5631 and a maximum of 0.6050. The additional data provides more context, but the fusion method at the data-level may not fully exploit the potential of the combined sensor information.

When feature-level fusion is employed, combining wearable and force/torque sensors, there is a notable enhancement in performance, with an average accuracy of 0.5791 and a maximum of 0.6403. feature-level fusion allows the model to more effectively leverage the strengths of both sensor types, leading to better classification accuracy. A significant improvement is observed when using only 3D Lidar data, which achieves an average accuracy of 0.6508 and a maximum accuracy of 0.7550. The 3D Lidar sensor provides rich spatial information, greatly aiding the model's ability to distinguish between different activities with higher precision and recall.

The integration of all sensors results in the highest performance, with an average accuracy of 0.6945 and a maximum accuracy of 0.8073. This comprehensive sensor integration combines the benefits of all individual sensor types, leading to superior accuracy and enhanced precision (average: 0.7012, max: 0.8169), recall (average: 0.6660, max: 0.7722), and F1-score (average: 0.6817, max: 0.7675). Multiple sensor data sources enable the HAR model to achieve the highest levels of performance across all metrics, demonstrating the value of multi-sensor integration in complex activity recognition tasks.

Overall, the results highlight the importance of sensor fusion in improving HAR model performance. While wearable sensors alone provide limited accuracy, integrating additional sensors, especially through feature-level fusion, significantly enhances the model's ability to accurately classify human activities. This comprehensive approach leads to improvements in precision, recall, and F1-score, ultimately providing a more robust and reliable HAR system.

Based on the confusion matrices depicted in Fig. 9, the findings elaborated in Table I and Fig. 8 are further clarified. From these matrices, it is clear that the model using only

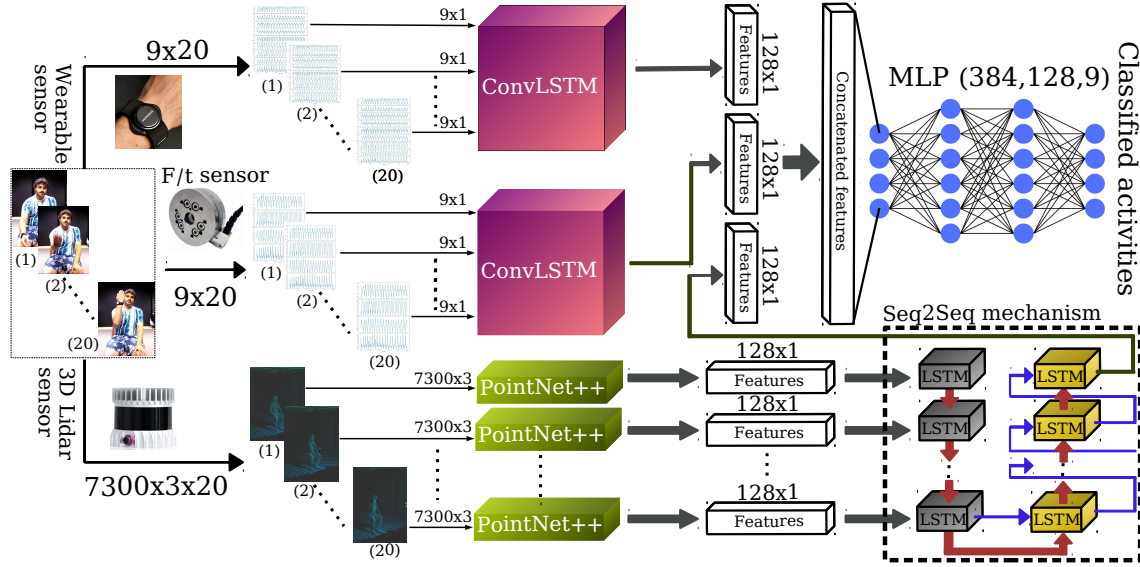


Fig. 4. The integration of wearable, force/torque, and 3D Lidar sensors at the feature-level. The data from wearable and force/torque sensors undergo separate processing using the ConvLSTM DL model illustrated in Fig. 4. Each model receives a sequence of 20 data points at every training step, producing a 128-feature vector as output. Meanwhile, the 3D Lidar data is processed using the PointNet++ architecture from Fig. 5. Each of the 20 point clouds is individually processed, containing 7300 points each, resulting in a 128-feature vector output for each cloud. For each set of 20 point clouds processed (with a 10-point overlap), the resulting 20×128 features are inputted into a Seq2Seq mechanism. Each 128-feature vector serves as an input to the Seq2Seq mechanism, organized in chronological order. Subsequently, the output of this mechanism, processing the 20×128 features, yields a single 128-feature vector. Finally, this output is concatenated with the 128-feature vectors obtained from the wearable and force/torque sensors. The resulting 384-feature vectors are then fed into a multilayer perceptron for the final classification. This MLP maps these features to a vector of 9 outputs, ultimately recognizing the activity.

TABLE I
AVERAGE AND MAXIMUM VALUES FOR PRECISION, RECALL, F1-SCORE, AND ACCURACY FOR DIFFERENT HAR MODELS.

HAR model	Precision		Recall		F1-Score		Accuracy	
	Avg	Max	Avg	Max	Avg	Max	Avg	Max
Only wearable sensor	0.6051	0.6396	0.5658	0.5970	0.5511	0.5857	0.5515	0.5900
Wearable & force/torque sensors data-level fusion	0.6217	0.6718	0.5787	0.6092	0.5669	0.6126	0.5631	0.6050
Wearable & force/torque sensors feature-level fusion	0.6373	0.7115	0.5944	0.6533	0.5939	0.6490	0.5791	0.6403
Only 3D Lidar	0.7230	0.7861	0.6675	0.7709	0.6317	0.7438	0.6508	0.7550
All sensors integrated	0.7012	0.8169	0.6660	0.7722	0.6817	0.7675	0.6945	0.8073

wearable sensors is not highly accurate in classifying different activities. Specifically, this model shows moderate accuracy for most activities except A5. This may be because A5 involves holding and moving the robot's arm, a task where the force/torque sensor is crucial, and wearable sensors alone cannot accurately recognize it. Notably, the model frequently misclassifies A5 as A6, which involves observing the robot dangerously. These two activities are unique in that they both involve standing and significant movement, unlike other activities, which mostly involve sitting.

The confusion matrix for the data-level fusion of wearable and force/torque sensors reflects a similar accuracy to the wearable-only model, as seen in Table I. However, a key difference is the significant improvement in classifying A5, highlighting the benefit of incorporating force/torque data.

The feature-level fusion of wearable and force/torque sensors shows a marked improvement in accuracy for A1 and A5 compared to both the wearable-only and data-level fusion models. These activities, which include physical interaction with the robot, benefit greatly from the combined sensor data, leading to significantly higher accuracy.

The confusion matrix for the 3D Lidar-only sensor indicates

that this sensor alone performs much better than the wearable-only and its integrations with force/torque sensors at both data and feature levels, except for activities A5 and A6. These two activities involve standing and extensive movement, which might challenge the 3D Lidar sensor due to changes in the user's position relative to the robot. Despite the 3D Lidar's capability to capture comprehensive user movements, the varied positions during these activities could lead to less accurate classification.

The final confusion matrix, for the model integrating all sensors, demonstrates the highest overall accuracy, as supported by Table I. This model leverages the strengths of the 3D Lidar for most activities while also incorporating the accuracy of the wearable and force/torque sensors for A5 and A6. For activities where the 3D Lidar excels, this integrated model shows superior performance, likely due to the combination of 3D Lidar with the other sensors. For A5 and A6, it outperforms the feature-level fusion of wearable and force/torque sensors, benefiting from the addition of the 3D Lidar.

In summary, although the highest overall accuracy achieved by the developed HAR models is not exceptionally high (80.37%), the integration of sensors significantly improves

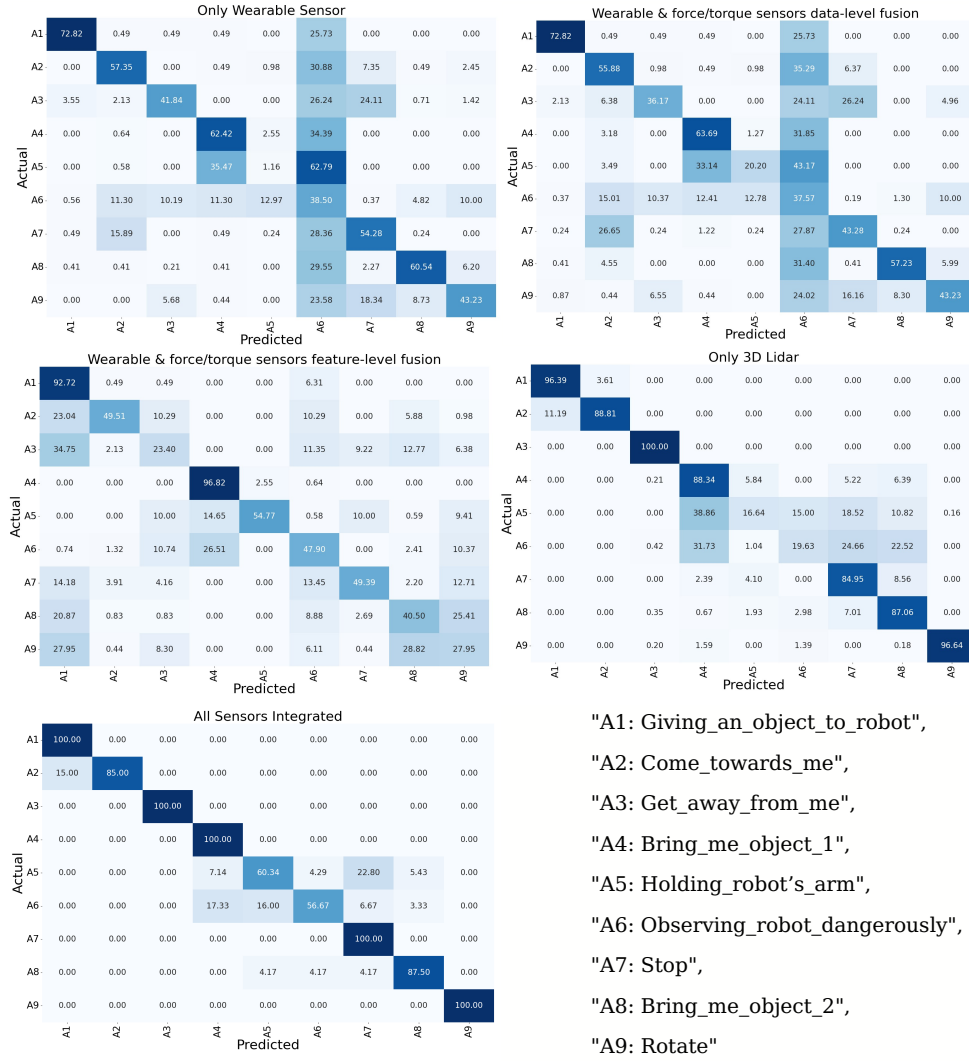


Fig. 5. Confusion matrices for all five trained HAR models, illustrating the classification performance across nine different activities and the accuracy of each HAR model in recognizing these activities.

the classification accuracy. This improvement is evident when adding force/torque sensor data to the wearable sensors and further enhanced by incorporating the 3D Lidar sensor.

VI. CONCLUSION

This paper proposes multimodal DL models for HAR in the context of HRI. The primary focus is on using privacy-preserving sensors integrated with each other, centered around the 3D Lidar sensor. A wearable sensor worn on the user's right wrist and the robot's force/torque sensors were also used and integrated with the 3D Lidar sensor. The aim was to cover various activities that a user may engage in when interacting with a robot. In this regard, nine distinct activities, including physical HRI and user commands to the robot, were studied. The activities were selected from a recently collected dataset named PriMA-Care, which includes approximately 27 user activities recorded using 10 different sensors.

Different methods of sensor data fusion using multimodal DL techniques were studied. These included data and feature-level fusion of the wearable and force/torque sensors and

feature-level fusion of all sensors. The results showed that integrating sensor data through these approaches improved the metrics for evaluating HAR. The feature-level integration of all sensors demonstrated the best performance, achieving the highest accuracy of 80.37%.

While sensor fusion effectively improved HAR accuracy, the best-obtained accuracy is not exceptionally high, which is a significant limitation of this work. However, the main goal of the paper was to compare how the integration of different sensors and various ways of sensor data integration would impact the performance. Thus, obtaining the best possible performance was less possible since we ran a number of different configurations rather than targeting only the most promising one. Another limitation is that decision-level sensor fusion, which the literature suggests can achieve higher HAR accuracies compared to feature or data-level fusion, was not explored in this study. Additionally, for 3D Lidar data processing, only the PointNet++ architecture was used, whereas more advanced DL frameworks could have been utilized. Another

constraint was reducing computational costs by cropping a part of the 3D Lidar point clouds that included user movements for processing. This approach may not be accurate or automated. More precise methods for extracting relevant parts from point clouds should be considered.

Future work will focus on various aspects. These include using decision-level sensor fusion and adding more activities to the study. Additionally, other sensors from PriMA-Care, particularly the thermal camera, will be integrated. This camera preserves privacy and aligns with the study's objectives. The work will also employ other benchmark DL models for point cloud processing. Furthermore, dimensionality reduction methods such as independent component analysis (ICA), principal component analysis (PCA), and linear discriminant analysis (LDA) will be used to extract relevant parts from the 3D Lidar data for further processing.

ACKNOWLEDGMENT

This work is partially supported by The Research Council of Norway as a part of the Vulnerability in the Robot Society (VIROS) project, under grant agreement 288285, Predictive and Intuitive Robot Companion (PIRC) project under grant agreement 312333 and through its Centers of Excellence scheme, RITMO with the project no. 262762.

REFERENCES

- [1] M. Vrigkas, C. Nikou, and I. A. Kakadiaris, "A review of human activity recognition methods," *Frontiers in Robotics and AI*, vol. 2, p. 28, 2015.
- [2] A. Taha, H. H. Zayed, M. E. Khalifa, and E.-S. M. El-Horbaty, "Human activity recognition for surveillance applications," in *Proceedings of the 7th International Conference on Information Technology*, 2015, pp. 577–586.
- [3] A. Roitberg, A. Perzylo, N. Somani, M. Giuliani, M. Rickert, and A. Knoll, "Human activity recognition in the context of industrial human-robot interaction," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*. IEEE, 2014, pp. 1–10.
- [4] S. Qiu, H. Zhao, N. Jiang, Z. Wang, L. Liu, Y. An, H. Zhao, X. Miao, R. Liu, and G. Fortino, "Multi-sensor information fusion based on machine learning for real applications in human activity recognition: State-of-the-art and research challenges," *Information Fusion*, vol. 80, pp. 241–265, 2022.
- [5] A. Roitberg, A. Perzylo, N. Somani, M. Giuliani, M. Rickert, and A. Knoll, "Human activity recognition in the context of industrial human-robot interaction," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*. IEEE, 2014, pp. 1–10.
- [6] A. Anagnostis, L. Benos, D. Tsaopoulos, A. Tagarakis, N. Tsolakis, and D. Bochtis, "Human activity recognition through recurrent neural networks for human-robot interaction in agriculture," *Applied Sciences*, vol. 11, no. 5, p. 2188, 2021.
- [7] E. Martinez-Martin and A. P. del Pobil, "Personal robot assistants for elderly care: an overview," *Personal assistants: Emerging computational technologies*, pp. 77–91, 2018.
- [8] K. T. Kadhim, A. M. Alsahlany, S. M. Wadi, and H. T. Kadhum, "An overview of patient's health status monitoring system based on internet of things (iot)," *Wireless Personal Communications*, vol. 114, no. 3, pp. 2235–2262, 2020.
- [9] A. D. Antar, M. Ahmed, and M. A. R. Ahad, "Challenges in sensor-based human activity recognition and a comparative analysis of benchmark datasets: A review," in *2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*. IEEE, 2019, pp. 134–139.
- [10] A. Baselizadeh, D. S. Lindblom, W. Khaksar, M. Z. Uddin, and J. Torresen, "Comparative analysis of vision-based sensors for human monitoring in care robots: Exploring the utility-privacy trade-off," in *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*, 2024, pp. 1794–1801.
- [11] A. Baselizadeh, "Advancing safety and privacy in human-robot interaction: Exploring learning-based control and multi-modal sensing approaches," 2024.
- [12] A. Baselizadeh, W. Khaksar, M. Z. Uddin, D. Saplacan, and J. Torresen, "Privacy-preserving user pose prediction for safe and efficient human-robot interaction," in *2023 IEEE 19th International Conference on Automation Science and Engineering (CASE)*, 2023, pp. 1–8.
- [13] J. Torresen, D. Saplacan, A. Baselizadeh, and T. Mahler, "Machine excellence tradeoffs to ethical and legal perspectives," in *2023 IEEE Conference on Artificial Intelligence (CAI)*, 2023, pp. 237–240.
- [14] M. Ryoo, B. Rothrock, C. Fleming, and H. J. Yang, "Privacy-preserving human activity recognition from extreme low resolution," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1, 2017.
- [15] J. Choi, S. Ulbrich, B. Lichte, and M. Maurer, "Multi-target tracking using a 3d-lidar sensor for autonomous vehicles," in *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*. IEEE, 2013, pp. 881–886.
- [16] O. Risbøl and L. Gustavsen, "Lidar from drones employed for mapping archaeology-potential, benefits and challenges," *Archaeological Prospection*, vol. 25, no. 4, pp. 329–338, 2018.
- [17] T. Yang, Y. Li, C. Zhao, D. Yao, G. Chen, L. Sun, T. Krajnik, and Z. Yan, "3d tof lidar in mobile robotics: A review," *arXiv preprint arXiv:2202.11025*, 2022.
- [18] A. Pfrunder, P. V. Borges, A. R. Romero, G. Catt, and A. Elfes, "Real-time autonomous ground vehicle navigation in heterogeneous environments using a 3d lidar," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 2601–2608.
- [19] J. Roche, V. De-Silva, J. Hook, M. Moencks, and A. Kondo, "A multimodal data processing system for lidar-based human activity recognition," *IEEE Transactions on Cybernetics*, vol. 52, no. 10, pp. 10027–10040, 2021.
- [20] M. Ohno, R. Ukyo, T. Amano, H. Rizk, and H. Yamaguchi, "Privacy-preserving pedestrian tracking using distributed 3d lidars," in *2023 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 2023, pp. 43–52.
- [21] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: analysis, applications, and prospects," *IEEE transactions on neural networks and learning systems*, vol. 33, no. 12, pp. 6999–7019, 2021.
- [22] Z. Yu, A. Taha, W. Taylor, A. Zahid, K. Rajab, H. Heidari, M. A. Imran, and Q. H. Abbasi, "A radar-based human activity recognition using a novel 3-d point cloud classifier," *IEEE Sensors Journal*, vol. 22, no. 19, pp. 18 218–18 227, 2022.
- [23] J. Gu, Z. Lu, H. Li, and V. O. Li, "Incorporating copying mechanism in sequence-to-sequence learning," *arXiv preprint arXiv:1603.06393*, 2016.
- [24] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *CoRR*, vol. abs/1706.02413, 2017. [Online]. Available: <http://arxiv.org/abs/1706.02413>
- [25] X. Dang, P. Jin, Z. Hao, W. Ke, H. Deng, and L. Wang, "Human movement recognition based on 3d point cloud spatiotemporal information from millimeter-wave radar," *Sensors*, vol. 23, no. 23, p. 9430, 2023.
- [26] M. A. Hanif, T. Akram, A. Shahzad, M. A. Khan, U. Tariq, J.-I. Choi, Y. Nam, and Z. Zulfikar, "Smart devices based multisensory approach for complex human activity recognition," *Comput. Mater. Contin.*, vol. 70, pp. 3221–3234, 2022.
- [27] K. Liu, N. Z. Gebräel, and J. Shi, "A data-level fusion model for developing composite health indices for degradation modeling and prognostic analysis," *IEEE Transactions on Automation Science and Engineering*, vol. 10, no. 3, pp. 652–664, 2013.
- [28] M. Ehatisham-Ul-Haq, A. Javed, M. A. Azam, H. M. A. Malik, A. Irtaza, I. H. Lee, and M. T. Mahmood, "Robust human activity recognition using multimodal feature-level fusion," *IEEE Access*, vol. 7, pp. 60 736–60 751, 2019.
- [29] N. Dawar, S. Ostadabbas, and N. Kehtarnavaz, "Data augmentation in deep learning-based fusion of depth and inertial sensing for action recognition," *IEEE Sensors Letters*, vol. 3, no. 1, pp. 1–4, 2019.
- [30] A. Baselizadeh, M. Z. Uddin, W. Khaksar, D. S. Lindblom, and J. Torresen, "Prima-care: Privacy-preserving multi-modal dataset for human activity recognition in care robots," pp. 233–237, 2024.
- [31] J. Pages, L. Marchionni, and F. Ferro, "Tiago: the modular robot that adapts to different research needs," in *International workshop on robot modularity, IROS*, vol. 290, 2016.
- [32] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," *CoRR*, vol. abs/1612.00593, 2016. [Online]. Available: <http://arxiv.org/abs/1612.00593>