2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Towards Enhanced Human Activity Recognition for Real-World Human-Robot Collaboration

Beril Yalcinkaya[1], Micael S. Couceiro[2], Lucas Pina[3], Salviano Soares[4], António Valente[5], Fabio Remondino[6]

Abstract— **This research contributes to the field of Human-Robot Collaboration (HRC) within dynamic and unstructured environments by extending the previously proposed Fuzzy State-Long Short-Term Memory (FS-LSTM) architecture to handle the uncertainty and irregularity inherent in real-world sensor data. Recognising the challenges posed by low-cost sensors, which are highly susceptible to environmental conditions and often fail to provide regular periodic readings, this paper introduces additional pre-processing blocks. These include two indirect Kalman filters and an additional LSTM network, which together enhance the input variables for the fuzzification process. The enhanced FS-LSTM approach is evaluated using real-world data, demonstrating its effectiveness in extracting meaningful information and accurately recognising human activities. This work underscores the potential of robotics in addressing global challenges, particularly in labour-intensive and hazardous tasks. By improving the integration of humans and robots in unstructured environments, this research contributes to the broader exploration of robotics in new societal applications, fostering connections and collaborations across diverse fields.**

I. INTRODUCTION

With the evolving landscape of robotics, the significance of Human-Robot Collaboration (HRC) has surged to the forefront [1]. This dynamic paradigm seeks to harness the unique strengths of both human intellect and machine precision, ushering in a new era of symbiotic productivity. The intricate interplay between human strategic thinking and the robust physical prowess of robots holds the promise of revolutionising various sectors. However, this harmonious collaboration encounters a significant hurdle: the inherent variability of human behaviour [2]. The intricacies of human actions, reactions, and decision-making defy rigid algorithms, posing a formidable challenge to the seamless integration of humans and robots. This unpredictability, while quintessentially human, necessitates a deeper understanding and prediction of behaviour to empower robots with the adaptability demanded by real-world, dynamic environments. In this context, the FEROX R&D EU project emerges as a beacon of innovation in HRC, joining human aspects with cutting-edge technologies. This project exemplifies innovation in HRC, aiming to transform the wild berry industry by introducing robotics solutions that aid pickers in harsh environments. By integrating robotics and artificial intelligence (AI), FEROX intends to enhance the efficiency and safety of berry pickers by providing detailed information about berry locations, abundance, and ripeness, where understanding the human behaviours is of the utmost importance.

A. Importance of human activity recognition

Human Activity Recognition (HAR) helps analysing sensor data to identify and detect human activities, enabling robots to have awareness of human actions [3]. HAR has found applications in robotics but also in various domains such as healthcare [4], smart home applications [5], and elderly assistance [6]. Traditional machine learning methods, such as Bayesian networks [7], random forest [8], and support vector machines [9], have been employed for HAR. Additionally, probabilistic methods, like Hidden Markov Models (HMM) [10] and other Markov-based approaches [11], have also been used to understand and predict human activities. Mixture models have been quite popular, as also presented in the authors work [12] on the probabilistic ensemble of classifiers, exploiting temporal information from previous time slices.

The recent deep learning popularity, with Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), has significantly improved the accuracy and robustness of HAR systems [13]. Similarly, the use of long short-term memory (LSTM) networks has gained prominence in HAR due to their ability to learn and remember over long sequences, making them suitable for time-series data inherent in human activities [14]. LSTM can capture temporal dependencies and intricate patterns in sequential data, as shown in our previous work on sports activities

*This work has been partly funded by the EU FEROX project (https://ferox.fbk.eu/) which received funding from the European Union's Horizon Europe Framework Programme under Grant Agreement No. 101070440. Views and opinions expressed are however those of the authors only and the European Commission is not responsible for any use that may be made of the information it contains. This work was also partially funded by FCT - Fundação para a Ciência e a Tecnologia (FCT) I.P., through national funds, within the scope of the UIDB/00127/2020 project (IEETA/UA, http://www.ieeta.pt/). This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020.DOI 10.54499/LA/P/0063/2020 https://doi.org/10.54499/LA/P/0063/2020/.

[1]Beril Yalcinkaya is with CORE R&D Department, Ingeniarius Lda, Alfena, Portugal. beril@ingeniarius.pt

[2]Micael Couceiro is with is with CORE R&D Department, Ingeniarius Lda, Alfena, Portugal. micael@ingeniarius.pt

[3]Lucas Pina is with is with CORE R&D Department, Ingeniarius Lda, Alfena, Portugal. lucas@ingeniarius.pt

[4]Salviano Soares is with is with School of Sciences and Technology-Engineering Department, University of Trás-os-Montes and Alto Douro (UTAD), Vila Real, Portugal and Institute of Electronics and Informatics Engineering of Aveiro (IEETA/LASI), University of Aveiro, Portugal salblues@utad.pt

[5]António Valente is with is with School of Sciences and Technology-Engineering Department, University of Trás-os-Montes and Alto Douro (UTAD), Vila Real, Portugal and NESC TEC—INESC Technology and Science, 4200-465 Porto, Portugal. avalente@utad.pt

[6]Fabio Remondino is with is with3D Optical Metrology (3DOM), Bruno Kessler Foundation (FBK), Trento, Italy. remondino@fbk.eu

recognition using wearable sensors [15] or gesture detection for robotics control [16]. More recently, transformer deep learning models, originally developed for natural language processing tasks, have also been adapted for other domains, including HAR [17]. The self-attention mechanism inherent in the transformer allows it to express individual dependencies between signal values within a time series, making it well-suited for HAR tasks. Lightweight transformers can achieve state-of-the-art results on HAR tasks using data from inertial measurement units (IMUs) embedded in mobile devices, while using fewer computing resources than traditional architectures [18].

B. Research question and objectives

Despite significant progress in the field, the challenge of managing the uncertainty inherent in human behavior persists, exacerbated by the high variability of actions across different contexts and the noise in real-world sensor data. This unpredictability jeopardizes the trust and safety crucial to Human-Robot Collaboration (HRC) systems, as inaccuracies in anticipating human intentions can lead to detrimental outcomes such as incorrect decisions, leading to accidents and injuries, thereby impacting trust and acceptability [19]. Various methodologies, from traditional machine learning and probabilistic models to advanced deep learning techniques, have been explored to mitigate these uncertainties. However, they struggle with real-time adaptability, capturing long-term dependencies and with challenges in computational efficiency and interpretability within the constrained environments of HRC systems. Our FS-LSTM framework emerges as a novel solution, intertwining the adaptability of fuzzy logic, the structured behavior representation of finite state machines, and the sequential data proficiency of Long Short-Term Memory (LSTM) networks. This integrated approach not only addresses the variability and noise challenges more effectively but also aligns with the computational realities of HRC systems, thereby enhancing the precision, adaptability, and safety of these interactions.

Building upon our prior work on the FS-LSTM (Fuzzy-State LSTM) HAR framework [20], the latest efforts encompass real-world testing employing Movesense sensory kits and smartphone sensors, alongside comparative assessments with LSTM and transformer models. Furthermore, the FS-LSTM architecture is expanded by incorporating additional sensors to capture both body and arm motion and orientation, thereby enhancing our understanding of human behaviors. To ensure precise estimation, Kalman fusion methods are utilized for body and arm orientation, while LSTM networks are harnessed to improve speed estimation.

II. METHODOLOGY

The FS-LSTM architecture, as depicted in Figure 1, is an innovative approach designed to address the uncertainty of human activity recognition and it involves five blocks labelled as A, B, C, D, E. In contrast to prior research [20], the present study refines the architecture to align with real-world scenarios. Specifically, the proposed extension integrates a smartphone affixed to the arm and a Movesense chest-worn sensor. This augmentation is motivated by the aspiration to capture comprehensive motion and tilt data from both the body.

Unlike our earlier work, this study employs an Indirect Kalman Filter and LSTM neural network to estimate speed and orientation values, as demonstrated in Block B. Subsequently, these values undergo a fuzzification process in Block C, yielding linguistic labels for BodyMotion, BodyTilt, ArmMotion and ArmTilt. These fuzzified features are then employed as inputs for Blocks D and E. Block D employs a state-based LSTM machine learning approach, utilising the fuzzified feature set as inputs. Distinct networks are trained for various runtime states, including a recovery state denoted as "Lost". Addressing uncertainty, Block E employs defuzzification. The fuzzified inputs play a pivotal role in a pre-established classification network. The resultant classification score undergoes fuzzification and subsequent defuzzification, facilitating the determination of state progression or retention.

A. Pre-processing blocks for real-world sensor data

In response to the complex and often erratic nature of real-world sensor data, this paper enhances FS-LSTM's handling of real-world sensor data with customized pre-processing. It employs Indirect Kalman Filters for orientation and an LSTM network to infer human body speed, bolstering the architecture's effectiveness beyond synthetic data tests.

1) Indirect Kalman filters for orientation estimation: For the purpose of orientation estimation, we have employed an attitude and heading reference systems (AHRS) filter [21]. This filter utilises a nine-axis Kalman filter structure, to estimate parameters including orientation, gyroscope offset, linear acceleration, and magnetic disturbance. Unlike direct methods for tracking orientation, the algorithm employs an indirect Kalman filter approach to model the error process. The central concept revolves around the continuous tracking of the error state vector x_k, which spans dimensions of 12-by-1, encompassing orientation errors, gyroscope bias, acceleration errors, and magnetic disturbance errors. This methodology enhances the robustness and accuracy of estimation by accounting for a diverse range of error sources inherent in the system.

The application of this indirect Kalman filter technique offers a potent avenue for estimating orientation, gyroscope bias, linear acceleration, and magnetic disturbance. By incorporating the error process and its iterative update mechanism, the AHRS filter substantially enhances the precision and stability of orientation estimation within systems focused on human activity recognition. The iterative process outlined below provides a comprehensive insight into how the algorithm progressively refines its estimates in a frame-by-frame manner.

$$x_k^- = 0$$

$$P_k^- = Q_k$$

$$y_k = z_k, \quad y_k \in R^n$$

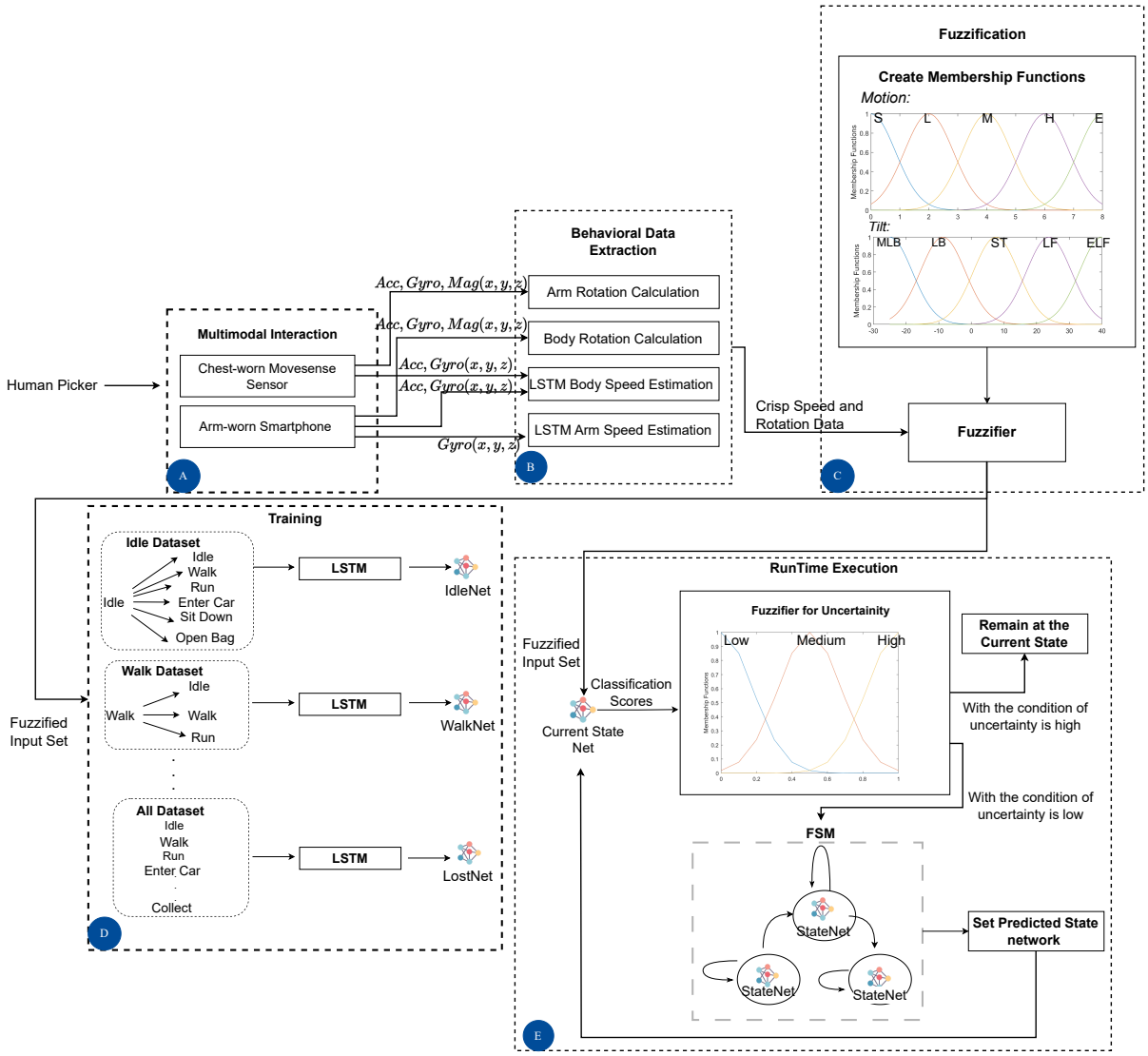

Fig. 1. The diagram of the proposed FS-LSTM architecture.

$$S_k = R_k + H_k P_k^- {H_k}^T$$

$$K_k = P_k^- H_k^T (S_k)^{-1}$$

$$x_k^+ = K_k y_k$$

$$P_{+k}^+ = P_k^- - K_k H_k P_k^-$$

In our analysis, x_k^- denotes the predicted (*a priori*) state estimate, capturing the system's anticipated state before incorporating new measurement data, reflecting the error dynamics of the process. The term P_k^- signifies the predicted (*a priori*)estimate covariance, quantifying the uncertainty of the *a priori* state estimate. y_k is a 9-dimensional vector, representing the discrepancy between actual observations and predictions. This vector encompasses three measurements each from the gyroscope, accelerometer, and magnetometer (x, y, z axes for each), serving as a critical feedback mechanism for updating the system's state. S_k, or the innovation covariance, assesses the innovation's uncertainty, crucial for scaling the Kalman gain, K_k, which optimizes the weight given to the innovation in the state update process. Upon integrating new measurements, x_k^+ emerges as the updated (*a posteriori*) state estimate, refined with the latest observational data, while P_k^+ updates the estimate covariance to reflect the reduced uncertainty post-measurement assimilation. The subscript k indicates the iteration step, with superscripts $+$ and $-$ distinguishing between *a posteriori* and *a priori* estimates, respectively. Utilizing data from linear acceleration, gyroscope, and magnetometer sensors across three axes (x, y, z) from both Movesense sensors and smartphones, we accurately estimate chest (BodyPitch) and arm (ArmPitch) orientations.

2) LSTM network for speed estimation: This section introduces the application of LSTM neural networks for speed estimation, more specifically BodySpeed and ArmSpeed. LSTM is chosen for body speed estimation due to the unreliability of Global Navigation Satellite Systems (GNSS) data when compared to the synthetic data considered in our

previous work [20]. The forest setting led to a substantial impairment of the GNSS signal captured by the smartphone due to tree interference. Therefore, a simple LSTM network has been adopted to uncover the intricate patterns within the lengthy sequential data from both Movesense and smartphone, while mitigating challenges, such as gradient explosions or vanishing gradients [22]. This choice underscores the LSTM's capability to effectively learn from and predict based on the complex temporal patterns inherent in sensor data, compensating for GNSS inaccuracies. Its architecture is particularly adept at navigating the challenges of sequence-based data analysis, making it an ideal choice for maintaining consistent performance in environments where GNSS reliability is compromised.

For the herein proposed pre-processing approach, the LSTM classes were established according to the average speed foreseen under different activities as follows:

$$\{\text{Walk}, \text{Run}, \text{Drive}, \text{All other activities}\} = \{0.5, 1.0, 1.5, 0.0\}$$

Subsequently, the LSTM network was trained utilising linear acceleration inputs (Acc_x, Acc_y, Acc_z) and gyroscope inputs (Gyr_x, Gyr_y, Gyr_z) obtained from both the Movesense kit and the smartphone. The aim was to align these inputs with the four designated outputs to estimate the BodySpeed. It is noteworthy that the ArmSpeed was directly defined by employing the smartphone's gyroscope reading along the y-axis.

B. Fuzzification of sensor data

Fuzzy logic enables computer systems to emulate human-like thinking and decision-making in uncertain and imprecise scenarios. It proves valuable when exact, reliable information is absent. For instance, the seemingly subjective phrase "the food is good" is enough for a person to determine an appropriate tip. This handling of uncertainty becomes crucial when precise data is unavailable.

The calculated speed and orientation in pitch serve as inputs for the fuzzification process. Specifically, speed is mapped to 5 linguistic labels for Motion of Body and Motion of Arm, while pitch is translated to 5 linguistic labels for Tilt of the body and arm. This approach optimizes the balance between detail and computational efficiency, allowing the FS-LSTM to effectively process and categorize human movements. The numerical data is mapped to Motion and Tilt into five linguistic variables using Gaussian membership functions. This is illustrated as:

$$S_U(t) = \begin{cases} BodyMotion \rightarrow \{S, L, M, H, E\} \\ BodyTilt \rightarrow \{MLB, LB, ST, LF, ELF\} \\ ArmMotion \rightarrow \{S, L, M, H, E\} \\ ArmTilt \rightarrow \{MLB, LB, ST, LF, ELF\} \end{cases}$$

Here, S, L, M, H, and E represent linguistic variables corresponding to Stopped, Low, Medium, High, and Extreme for Motion, respectively. Similarly, MLB, LB, ST, LF, and ELF correspond to Medium Lean Back, Lean Back, Straight, Lean Front, and Extremely Lean Front for Tilt. The resulting feature vector, denoted as $S_U(t)$, is subsequently utilised as input for the FS-LSTM method.

C. Activity recognition and sequence modelling

This section addresses the modelling of human activities by analysing transitions between activity states, encompassing both activity sequences and individual actions. The state diagram in Block E of Figure 1 visualises the activities of a human picker in a wild forest. This scenario depicts a sequence of locomotive movements ('Idle', 'Walk', 'Run'), berry collection/loading ('Collect'/'Load'), and vehicle driving ('Drive'), resulting in a configuration of 14 states. Among these states, 'Lost' serves as a recovery state. The states include both sequential elements like sitting and driving, as well as transitions like 'Sit Down' to 'Sit', 'Sit' to 'Stand Up', or 'Enter Car' to 'Drive'. Furthermore, the states present multiple potential trajectories; for instance, the 'Idle' state can transition to 'Open Bag', 'Walk', 'Run', 'Enter Car', or 'Sit Down'.

To model state transitions effectively, LSTM networks are employed. Within the proposed Human Activity Recognition (HAR) framework, each LSTM network operates on sequential data (fuzzy features discussed in the preceding section) and predicts viable transition states. This strategy precludes infeasible transitions, such as 'Sit Down' to 'Stand Up' or 'Collect' to 'Load', which deviate from the meticulously designed expert flow represented as a Finite State Machine (FSM). This methodology guarantees that the projected subsequent state aligns with feasible options, thereby facilitating decision-making in Human-Robot Collaboration (HRC) systems. Notably, a higher number of classes amplifies the model's complexity, demanding increased computational resources and extending runtime [23]. In comparison, a streamlined LSTM network for 'Sit Down', encompassing only realistic choices ('Sit Down' or 'Sit'), proves more efficient than one trained on the entirety of possible activities, especially when numerous outputs are improbable.

For prognosticating forthcoming activity sequences, a sequence-to-sequence classification approach is adopted. Each state-specific LSTM network utilises uniform sequential input ($S_U(t)$), embedding 20 features including the fuzzified BodyMotion, BodyTilt, ArmMotion and ArmTilt values derived from raw sensor data. Significantly, the outputs of each state-based LSTM network differ, tailored to specific states and potential transitions. States that lack feasibility are designated as 'Lost'. Notably, the Lost network operates on identical sequential input while employing an output dataset encompassing all possible states, functioning as a recuperative mechanism triggered by the preceding LSTM network.

Trained network models, detailed earlier, form a iterative feedback architecture to predict future states (Block E, Figure 1). Fuzzified outputs from Block C serve as inputs for the pertinent trained network model linked to the current state—initially known as the Lost network—iteratively adapting with architecture progression. This architecture

employs network model classification scores as posterior probabilities grounded in fuzzy input sets, computed via Bayes' Theorem. FS-LSTM operates on the assumption that if top scores are low or similar across classes, state network models might struggle to make confident predictions, causing uncertainty. Despite this uncertainty, FS-LSTM leverages the highest-scored class for predictions. Evaluating confidence and interpreting scores before decisions is crucial. Fuzzy logic assesses score uncertainty.

Similar to the Motion and Tilt fuzzy variables, classification scores are used to fuzzify and generate fuzzy linguistic labels for Uncertainty—Low, Medium, and High. Triangular fuzzifiers gauge membership degrees. Unlike Motion and Tilt variables, inference employs state-specific rules. An example of fuzzy rules for the 'Close Bag' state assesses uncertainty in classification scores for possible transitions from 'Idle', 'Close Bag', and 'Lost' recovery state.

In fuzzy inference, rules apply to fuzzified inputs, calculating rule fulfilment through aggregation. Defuzzification follows, transforming fuzzy outputs into crisp outputs via fuzzy sets and corresponding membership degrees. Aggregation results convert into crisp values via the centroid method, expressing uncertainty as a crisp value between 0 and 1. This crisp value gauges confidence in each network model before state determination.

Comparing the crisp uncertainty value from defuzzification to a predefined threshold assesses model confidence. If the value is below the threshold, the model confidently predicts based on the highest score the next state identified. The selected network model corresponds to the predicted state for the next iteration. If uncertainty exceeds the threshold, the system retains the state. Subsequent iterations classify using the current network model. This iterative process persists until the model achieves a suitable prediction confidence.

III. EXPERIMENTAL RESULTS

A. Setup and sensor data collection

The experimental setup utilises a Movesense sensor[1] attached to the chest and a smartphone affixed to the arm, as illustrated in Figure 2. The ground truth activities were acquired by merging data from static and mobile cameras, capturing the picker's actions. These videos underwent careful manual annotation to accurately label each activity, ensuring precise alignment with the picker's actions for reliable analysis and evaluation. The data acquisition process involves a specially designed app created using the Unity game engine[2], leveraging various community-contributed assets, such as the Movesense Sensor Plugin[3] and the Graph Maker[4].

The collected data, designated as T_U, is detailed below, with 's' denoting smartphone readings and 'm' representing Movesense sensor measurements. 45,534 samples collected

[1]https://www.movesense.com/movesense-active/

[2]https://unity.com/

[3]https://assetstore.unity.com/packages/tools/integration/movesense-sensor-plugin-118242

[4]https://assetstore.unity.com/packages/tools/gui/graph-maker-11782

Fig. 2. A human picker performing the "Collect" activity.

at a frequency of 30 Hz over a span of 45 minutes. A straightforward transformation is applied to ensure uniformity and coherence across various sensor outputs. For model development and validation, the dataset is split into training (70%), and testing (30%) subsets.

$$T_U(t) = \big[\, Gyr_x^s\ Gyr_y^s\ Gyr_z^s\ Gyr_x^m\ Gyr_y^m\ Gyr_z^m\ Mag_x^s\ Mag_y^s\ Mag_z^s\ Mag_x^m\ Mag_y^m\ Mag_z^m\ Acc_x^s\ Acc_y^s\ Acc_z^s\ Acc_x^m\ Acc_y^m\ Acc_z^m\ \big] \quad (1)$$

B. Methods and evaluation metrics

In the methods section, we detail the experimental approaches employed to investigate the intricate interplay between body and arm-related data in predictive modelling. Our investigation unfolds through four distinct and carefully designed experiments, each focusing on elucidating the nuances of data fuzzification and the intricate relationships between bodily and arm-associated variables.

Experiment 1. FS-LSTM with Body and Arm, involves the meticulous fuzzification of calculated variables, including BodySpeed, BodyPitch, ArmSpeed, and ArmPitch, all derived from sensor data as illustrated in Block B of Figure 1. The ensuing fuzzified data serves as the foundation for training multiple LSTM models, as depicted in Block D. This experiment delves into the implications of amalgamating body and arm-related data on predictive accuracy, effectively mitigating the influence of uncertainty.

Experiment 2. FS-LSTM with Body, hones in exclusively on body-related data, with a focal point on BodySpeed and BodyPitch. These parameters undergo fuzzification within the framework of a Fuzzy-LSTM architecture. The outcomes of this experiment provide insights into the significance of data fuzzification when considering body movements alone.

Experiment 3. LSTM with Body and Arm, on the incorporation of both BodySpeed and BodyPitch alongside ArmSpeed and ArmPitch into an LSTM model, without the introduction of fuzzification. This approach seeks to discern whether the omission of fuzzification and the exploration of the combined neural network sets can yield improvements in comparison to Experiment 1.

Experiment 4. Transformer with Body and Arm, concentrates on the realm of both body- and arm-related data devoid of fuzzification, as with Experiment 3, though this time using the Transformer model. This experiment aims to gauge our model's aptitude for pattern recognition and prediction when compared to the state-of-the-art approach for time series classification.

C. Benchmark

TABLE I

	FS-LSTM Body and Arm	FS-LSTM Body	LSTM Body and Arm	Transformer Body and Arm
Accuracy	**86.6%**	72.3%	74.4%	76.9%
Sensitivity	**0.94**	**0.94**	0.85	0.14
Specificity	0.94	0.95	0.93	**0.98**
Convergence Time Median	**0.03s**	**0.03s**	0.07s	0.07s
Convergence Time Mean	**0.42s**	1.44s	0.61s	0.51s
Convergence Time Std	2.67s	3.33s	2.01s	**1.11s**

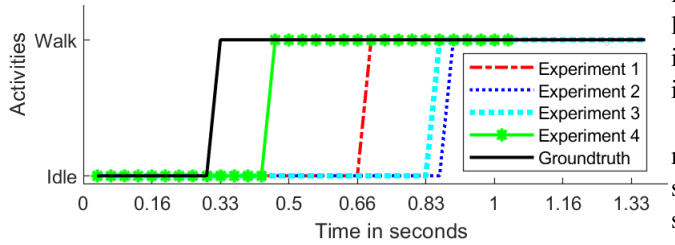

Fig. 3. Convergence time plotting along an activity sequence.

The efficiency of the FS-LSTM architecture in HAR can be gleaned through the emergent patterns from the experiments. Topping the list is Experiment 1, which employs both body and arm-related data, achieving an accuracy of 86.6%. By channelling the data through fuzzified variables and the FSM integration, intricate activity patterns become decipherable, representing the architecture's optimal performance. The convergence times present another crucial dimension. Both FS-LSTM iterations (Experiments 1 and 2) display strikingly similar median convergence times, significantly outpacing their traditional LSTM counterparts. This expeditious convergence, especially pronounced in Experiment 1, underscores FS-LSTM's computational prowess when grappling with expansive datasets. Nevertheless, Experiment 2 experiences a wider spread of convergence times, hinted by its elevated mean and standard deviation. This unpredictability in convergence aligns with the challenges faced by the FSM within the FS-LSTM framework when limited to just body-related data. The FSM, while reducing computational complexity by honing in on a narrower class range, can, with limited features, converge prematurely to incorrect classes. Such missteps not only influence immediate results but, given the architecture's sequential nature, can have cascading effects on subsequent iterations. Evaluating sensitivity and specificity offers more depth. Across the board, high sensitivity indicates a consistent ability to correctly pinpoint specific activities. Remarkably, despite its reduced accuracy, Experiment 2's FS-LSTM model showcases a specificity comparable to the best performer (Experiment 1). This suggests that, while there might be challenges in classifying certain activities accurately, the model is resilient in not falsely categorising negative cases. Experiment 4 introduces a novel perspective with the use of a transformer model trained with body and arm data. While demonstrating competitive accuracy at 76.9%, its sensitivity score of 0.14 raises concerns about its ability to accurately identify activities. This disparity highlights the intricate trade-offs between different architectures and emphasises the challenges of optimising accuracy while maintaining a balanced sensitivity-specificity interplay. The transformer, although a powerful architecture, could face challenges in handling the complexities of the data and maintaining accuracy with limited data where local temporal dependencies are more crucial. The convergence time metrics, nevertheless, indicate that the transformer model can promptly identify transitions, holding potential for expedited activity recognition. Despite having a generally larger convergence time compared to FS-LSTM due to its lower accuracy, the transformer model demonstrates rapid identification of the correct class when it occurs, as shown in Figure 3.

In summation,the FS-LSTM while the demonstrates optimal performance and efficiency with the dataset used in our study, demonstrating rapid convergence and a harmonious sensitivity-specificity dynamic, its nuanced behaviour with pared-down inputs highlights both its architectural sophistication and the associated challenges. As further refinements to this architecture beckon, understanding and navigating these intricacies will be instrumental in harnessing its full potential.

IV. CONCLUSIONS AND FUTURE WORKS

The paper provided valuable insights into the effectiveness of the proposed FS-LSTM architecture for HAR. Experiments have revealed distinct performance patterns that shed light on the model's capabilities and limitations. Notably, while the FS-LSTM architecture demonstrates clear superiority compared to the traditional LSTM and the transformer model, its performance with limited datasets raises intriguing questions about its adaptability and the potential influence of its inherent complexities. Further investigations could focus on refining the model's architecture or its training approach to optimise performance across varied input data, ensuring that the system's adaptability is not compromised by its sophistication. Moreover, the LSTM networks within the FS-LSTM architecture could be replaced with transformer models. Indeed, its ability to handle long-term dependencies without the need for recurrent layers could offer improved performance, especially when dealing with activities that are characterised by extended sequences of intricate movements. It would be particularly interesting to see how the fuzzy logic and FSM components interact with the transformer's attention mechanisms, possibly creating a more robust and adaptable model. Future studies incorporating multi-user evaluations will be key to refining and validating the FS-LSTM architecture for widespread HRC applications.

REFERENCES

[1] A. Bauer, D. Wollherr, and M. Buss, "Human–robot collaboration: a survey," *Int. J. Humanoid Robotics*, vol. 5, no. 01, pp. 47–66, 2008.

[2] E. Schmerling, K. Leung, W. Vollprecht, and M. Pavone, "Multimodal probabilistic model-based planning for human-robot interaction," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3399–3406, IEEE, 2018.

[3] L. E. Parker, "The effect of action recognition and robot awareness in cooperative robotic teams," in *Proceedings 1995 IEEE/RSJ International Conference on Intelligent Robots and Systems. Human Robot Interaction and Cooperative Robots*, vol. 1, pp. 212–219, IEEE, 1995.

[4] Y. Wang, S. Cang, and H. Yu, "A survey on wearable sensor modality centred human activity recognition in health care," *Expert Systems with Applications*, vol. 137, pp. 167–190, 2019.

[5] S. Mekruksavanich and A. Jitpattanakul, "Lstm networks using smartphone data for sensor-based human activity recognition in smart homes," *Sensors*, vol. 21, no. 5, p. 1636, 2021.

[6] A. Jalal, S. Kamal, and D. Kim, "A depth video sensor-based lifelogging human activity recognition system for elderly care in smart indoor environments," *Sensors*, vol. 14, no. 7, pp. 11735–11759, 2014.

[7] L. Liu, S. Wang, B. Hu, Q. Qiong, J. Wen, and D. S. Rosenblum, "Learning structures of interval-based bayesian networks in probabilistic generative model for human complex activity recognition," *Pattern Recognition*, vol. 81, pp. 545–561, 2018.

[8] S. Balli, E. A. Sağbaş, and M. Peker, "Human activity recognition from smart watch sensor data using a hybrid of principal component analysis and random forest algorithm," *Measurement and Control*, vol. 52, no. 1-2, pp. 37–45, 2019.

[9] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine," in *Ambient Assisted Living and Home Care: 4th Int. Workshop, Vitoria-Gasteiz, Spain, December 3-5, 2012. Proceedings 4*, pp. 216–223, Springer, 2012.

[10] D. Trabelsi, S. Mohammed, F. Chamroukhi, L. Oukhellou, and Y. Amirat, "An unsupervised approach for automatic activity recognition based on hidden markov model regression," *IEEE Transactions on automation science and engineering*, vol. 10, no. 3, pp. 829–835, 2013.

[11] J. P. Vital, D. R. Faria, G. Dias, M. S. Couceiro, F. Coutinho, and N. M. Ferreira, "Combining discriminative spatiotemporal features for daily life activity recognition using wearable motion sensing suit," *Pattern Analysis and Applications*, vol. 20, pp. 1179–1194, 2017.

[12] A. C. Nunes Rodrigues, A. Santos Pereira, R. M. Sousa Mendes, A. G. Araújo, M. Santos Couceiro, and A. J. Figueiredo, "Using artificial intelligence for pattern recognition in a sports context," *Sensors*, vol. 20, no. 11, p. 3040, 2020.

[13] K. Chen, D. Zhang, L. Yao, B. Guo, Z. Yu, and Y. Liu, "Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities," *ACM Computing Surveys (CSUR)*, vol. 54, no. 4, pp. 1–40, 2021.

[14] K. Xia, J. Huang, and H. Wang, "Lstm-cnn architecture for human activity recognition," *IEEE Access*, vol. 8, pp. 56855–56866, 2020.

[15] D. Araújo, M. Couceiro, L. Seifert, H. Sarmento, and K. Davids, *Artificial intelligence in sport performance analysis*. Routledge, 2021.

[16] K. Tatarian, M. S. Couceiro, E. P. Ribeiro, and D. R. Faria, "Stepping-stones to transhumanism: An emg-controlled low-cost prosthetic hand for academia," in *2018 International Conference on Intelligent Systems (IS)*, pp. 807–812, IEEE, 2018.

[17] I. Dirgová Luptáková, M. Kubovčík, and J. Pospíchal, "Wearable sensor-based human activity recognition with transformer model," *Sensors*, vol. 22, no. 5, p. 1911, 2022.

[18] S. EK, F. Portet, and P. Lalanda, "Lightweight transformers for human activity recognition on mobile devices," *arXiv preprint arXiv:2209.11750*, 2022.

[19] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "Would you trust a (faulty) robot? effects of error, task type and personality on human-robot cooperation and trust," in *Proc. 10th ACM/IEEE Human-Robot Interaction Conf.*, pp. 141–148, 2015.

[20] B. Yalçinkaya, M. S. Couceiro, S. P. Soares, and A. Valente, "Human-aware collaborative robots in the wild: Coping with uncertainty in activity recognition," *Sensors*, vol. 23, no. 7, p. 3388, 2023.

[21] R. Munguia and A. Grau, "An attitude and heading reference system (ahrs) based in a dual filter," in *ETFA2011*, pp. 1–8, IEEE, 2011.

[22] I. U. Haq, A. Ullah, S. U. Khan, N. Khan, M. Y. Lee, S. Rho, and S. W. Baik, "Sequential learning-based energy consumption prediction model for residential and commercial sectors," *Mathematics*, vol. 9, no. 6, p. 605, 2021.

[23] S. Han, J. Kang, H. Mao, Y. Hu, X. Li, Y. Li, D. Xie, H. Luo, S. Yao, Y. Wang, *et al.*, "Ese: Efficient speech recognition engine with sparse lstm on fpga," in *Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, pp. 75–84, 2017.