Optimizing Deep LSTM Model through Hyperparameter Tuning for Sensor-Based Human Activity Recognition in Smart Home

Mariam El Ghazi, Noura Aknin
Information Technology and Modeling, Systems Research Unit, Abdelmalek Essaadi University, Tetouan, Morocco
E-mail: mariam.elghazi@etu.uae.ac.ma , noura.aknin@uae.ac.ma

Keywords: long short-term memory (LSTM), hyperparameter tuning, batch normalization, deep learning, wearable sensors, human activity recognition (HAR)

Received: October 10, 2023

Human Activity Recognition (HAR) holds significant potential in healthcare, smart homes, sports, and security, mainly benefiting the well-being of elderly individuals and dependents. This research introduces an innovative deep learning-based approach to HAR, using wearable sensors in smart home environments. In this paper, we conduct a comprehensive review of the state of the art, offering insights into existing methods, classification techniques, their performances, hyperparameter tuning strategies, findings, limitations, and future directions. We propose an LSTM-based deep model enriched with batch normalization and perform a hyperparameter tuning using Bayesian Optimization; then, we evaluate the model on the PAMAP2 public dataset. The model outperforms previous studies, achieving remarkable performance metrics, including accuracy at 97.71%, F1 score, precision, and recall, approaching 96.66%, 96.85%, and 96.55%, respectively. We plan to assess the model's generalization capabilities for future work by training it on diverse datasets such as Opportunity and WISDM. Furthermore, we aim to enhance the model by exploring hybrid deep model architectures and alternative hyperparameter tuning approaches. These efforts maximize the model's efficiency and adaptability in real-world scenarios.

Povzetek:

1 Introduction

HAR has emerged as a crucial research area with wide-ranging applications in healthcare, smart homes, sports, and security [1]. Automatically recognizing and categorizing human activities can significantly enhance the well-being and independence of elderly individuals and those needing care [2]. In smart home settings, Human Activity Recognition (HAR) systems are essential for delivering context-aware services, monitoring residents' activities, and promptly notifying caregivers in the event of unusual situations [3].

While video-based approaches can achieve HAR, they often raise privacy concerns due to continuous surveillance requirements [4]. Sensor-based HAR using wearable devices has gained popularity to address these privacy issues. The data can be discreetly collected without compromising individuals' privacy using wearable sensors, such as accelerometers, gyroscopes, and temperature sensors [5].

This study focuses on sensor-based HAR using deep learning models, specifically LSTM. This network is suitable for handling time series data, which is essential for HAR tasks as activities are often characterized by sequential patterns over time. LSTM's ability to capture long-term dependencies and handle variable-length input sequences makes it an ideal choice for this time-sensitive problem.

The main contributions of this paper are as follows.

1) We conduct an in-depth review of the state of the art in sensor-based HAR using deep learning. This review provides valuable insights for readers, offering a thorough understanding of existing methods, classification techniques, hyperparameter tuning approaches, key findings, limitations, and future research directions. This comprehensive overview serves as a benchmark for comparing advancements in this domain.

2) We systematically extract the performance metrics achieved by models in previous studies, including accuracy, F1 score, precision, and Recall. Additionally, we assess whether these studies employed validation methods such as k-fold cross-validation.

3) We propose a novel LSTM-based model featuring batch normalization. To enhance its performance, we conduct hyperparameter tuning using Bayesian optimization.

4) We evaluate the efficacy of our proposed LSTM-based model on the publicly available wearable sensor dataset PAMAP2. We demonstrate the model's effectiveness through a rigorous assessment using accuracy, F1 score, precision, and Recall metrics. Furthermore, we ensure the reliability and generalizability of our model by performing a 10-fold cross-validation.

5) In addition to showcasing our experimental results, we compare them with those reported in the state-of-the-art. This comparative analysis positions our proposed method within the broader context of existing research, highlighting its strengths and contributions.

The paper is organized into distinct sections as follows: Section 2 provides a thorough review of prior studies related to sensor-based Human Activity Recognition (HAR). In Section 3, the proposed materials and methods are presented. Section 4 is dedicated to detailing the experiments and presenting the results. Finally, Section 5 concludes the paper. This structured approach ensures clarity and coherence throughout the document.

Table 1: State-of-the-art of sensor-based HAR using deep learning

Study	Year	Classification Method	Datasets	Hyperparameter (HP)	Findings	Limitations/ Future Works
Hammerla et al[6]	2016	CNN LSTM	-Opportunity -PAMAP2 -Daphnet Gait dataset	-fANOVA to investigate the impact of HP on model performance	-The best-performing model is CNN -DL guidelines for practitioners	-Explore more hp to optimize the models -explore more datasets and more complex models
Ma et al.[7]	2019	AttnSense (CNN, GRU, Attention Mechanism)	Heterogeneous, UniMiB-SHAR, and PAMAP2	The study explores HP impact, including CNN structure and sliding window width.	The results confirm the model's effectiveness in capturing dependencies in sensing signals' spatial and temporal domains.	Not mentioned
Xu et al.[8]	2019	Inno HAR (inception NN+RNN+GRU)	Opportunity PAMAP2 Smartphone Dataset	No mention of HP Tuning.	The proposed model exhibits superior performance and demonstrates strong generalization. Also, it has significant potential for real-time applications.	- Consider adjusting the network structure, including kernel sizes and connection methods. -Address the problem of class imbalance for HAR.
Wan et al[9]	2020	CNN LSTM Bi-LSTM	UCI PAMAP2.	YES, but the HP tuning techniques and ranges are not clearly explained.	The results showed that the CNN model outperforms other models	Explore new sensors for HAR. Investigate transfer learning's impact. If explore diverse HP impacts and identify optimal settings for varied datasets and applications.
Gao et al.[10]	2020	DanHAR (CNN and Attention Mechanism)	WISDM PAMAP2 UNIMIB SHAR OPPORTUNITY	No mention of HP tuning	The proposed model provided good results	Explore the impact of different HP on the performance of danhar and investigate the effectiveness on other datasets.
Xu et al.[11]	2022	Inception-LSTM with Attention Mechanism	Self-built dataset PAMAP2	No HP tuning, the HP are set by the authors and kept consistent.	The proposed model outperforms traditional algorithms in terms of accuracy and convergence speed	**Limitation:** Requires substantial training data, posing potential overfitting risks. **Future Directions:** Explore alternative models and regularization techniques. Extend model application to diverse contexts.
Thakur et al.[12]	2022	ConvAE-LSTM Model (CNN, LSTM, Auto Encoder)	WISDM UCI PAMAP2 Opportunity	The authors examine the impact of hyperparameters, such as the type of Optimizer used, the number of epochs, and the batch size, which affect the model performance.	The proposed Conv-AE-LSTM model provided good performance.	-The proposed method should be analyzed for its applicability in real-life applications -To compare the proposed method with other DL models, and examine other datasets. Also, further HP tuning could be performed to optimize the model's performance.
Tehrani et al.[13]	2023	Bi-LSTM	AreM Mhealth PAMAP2	The study investigates the Optimal window size and percentage of votes.	The proposed method attained a higher performance.	Improve the network's performance by optimizing the hyperparameters and exploring other types of neural networks.
Challa et al.[14]	2023	CNN+biLSTM+ HP Tuning	PAMAP2 UCI-HAR MHEALTH	The study used the Rao-3 metaheuristic optimization to search for optimal HP.	The authors searched for optimal HP values for DL models because they are crucial for the best performance. The proposed model achieved good results	Full text not open access
Kumar et al. [15]	2023	GRU	WISDM PAMAP2 KU-HAR	Full text not open access	The model achieved good results	The experimental outcomes offer an understanding of the practicality of the proposed model and suggest potential avenues for future research.

2 Related works

2.1 Human activity recognition overview: Human Activity Recognition (HAR) has become a pivotal research area with widespread applications in healthcare, smart homes, sports, and security [1]. The automatic detection and classification of human activities are crucial for enhancing the quality of life for elderly individuals and dependents, especially in smart home environments [2]. As the introduction mentions, this study will focus on sensor-based HAR to preserve the resident's privacy in a smart home.

2.2 Current trends of the state-of-the-art (SOTA): In the realm of deep learning models, various architectures, including Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU), have been employed for HAR. These models have demonstrated promising results in recognizing human activities based on sensor data. Table 1 summarizes the state of the art of sensor-based HAR using deep learning. the noticeable trends in deep learning for HAR

Trends in deep learning models for HAR: CNN emerged as the most widely used deep learning model across studies (Hammerla et al. [6], Ma et al. [7], Wan et al. [9], Gao et al. [10], Challa et al. [14]). CNNs are preferred for their ability to capture spatial features, making them suitable for sensor-based activity recognition. Besides, LSTM networks are also ubiquitous (Wan et al. [9], Xu et al. [11], Tehrani et al. [13]). Bi-directional LSTMs, in particular, are explored for their effectiveness in capturing temporal dependencies [13]. Some Studies integrated Attention mechanisms into models (Ma et al. [7], Gao et al. [10], Xu et al. [11]) to enhance the focus on specific segments of input sequences, contributing to improved performance. Another study conducted by Kalabakov et al.[16] uses DeepConvLSTM to transfer knowledge between two datasets, revealing that transferring the weights of fewer convolutional layers is more effective. Some practical implications of sensor-based Human Activity Recognition (HAR) in areas such as sports[17], surveillance [18], and fall detection[19] aim to ensure a healthier lifestyle for older people.

Hyperparameter tuning: Several studies (Hammerla et al. [6] and Thakur et al. [12]) explicitly explore hyperparameters, emphasizing their impact on model performance. However, there is a lack of consistency across studies regarding hyperparameter tuning.

Dataset diversity: In the field of human activity recognition, researchers have used publicly available datasets such as Opportunity[20], WISDM V1.1[21], and PAMAP2[22]. These datasets have allowed them to develop and test activity recognition methods using motion sensor data, reflecting an effort to generalize models across different contexts. It is noticeable that the PAMAP2 dataset is among the most used for HAR due to the size of the dataset, the Variety of the performed activities, and multiple subjects.

Evaluation metrics and model performance: Table 2 presents a variety of evaluation metrics across different studies, reflecting a lack of standardized reporting practices. While some studies provide accuracy, F1 score, precision, and Recall (Wan et al. [9], Xu et al. [11]), others have incomplete metrics (Hammerla et al. [6], Gao et al. [10]). This inconsistency makes direct comparisons challenging and emphasizes the need for standardized evaluation practices in sensor-based HAR research. Two standout models in sensor-based Human Activity Recognition are the Inception-LSTM proposed by Xu et al. (2022) with an accuracy of 95.04% [11]. Additionally, the hybrid CNN and bi-LSTM model with hyperparameter tuning introduced by Challa et al. (2023) achieved a slightly lower accuracy of 94.91% [14]. Challa et al. underscores the critical role of hyperparameter tuning for optimizing the performance of their model in activity recognition tasks.

Table 2: Related works model performance on PAMAP2 dataset

Study	Year	Classification Model	Accuracy	F1 score	Precision	Recall	K-fold cross Validation
Hammerla et al[6]	2016	CNN	-	93,70%	-	-	No
Hammerla et al[6]	2016	LSTM	-	92,90%	-	-	No
Ma et al.[7]	2019	AttnSense	-	89,30%	-	-	Yes (4 folds)
Xu et al.[8]	2019	Inno HAR	-	93,50%	-	-	No
Wan et al[9]	2020	CNN	91,00%	91,16%	91,66%	90,85%	No
Wan et al[9]	2020	LSTM	85,86%	85,34%	86,51%	84,67%	No
Wan et al[9]	2020	Bi-LSTM	89,52%	89,40%	90,19%	89,02%	No
Gao et al.[10]	2020	DanHAR	93,16%	-	-	-	No
Xu et al.[11]	2022	Inception-LSTM	95,04%	95,13%	95,06%	95,21%	No
Thakur et al.[12]	2022	ConvAE-LSTM	94,33%	94,46%	-	-	Yes (5 folds)
Tehrani et al.[13]	2023	Bi-LSTM	-	93.41%	93.41%	93.47 %	No
Tehrani et al.[13]	2023	Bi-LSTM	-	93,41%	93,41%	93,47%	No
Challa et al.[14]	2023	CNN+biLSTM+HP Tuning	94,91%	-	-	-	No
Kumar et al. [15]	2023	GRU	94,77%	-	-	-	No

Most sensor-based Human Activity Recognition studies lack the essential practice of employing K-fold cross-validation to evaluate the reliability and generalizability of deep learning models. Notably, only Ma et al. [7] (2019) and Thakur et al. [12] (2022) have incorporated 4-fold and 5-fold cross-validation, respectively, highlighting the need for a more standardized evaluation methodology. The absence of K-fold cross-validation across studies underscores the importance of a consistent approach for reliable comparisons.

2.3 Identified gaps in the literature: Despite the progress in HAR using deep learning models, a critical analysis reveals several gaps in the existing literature:

- **Limited hyperparameter tuning:** One notable gap is the lack of emphasis on hyperparameter tuning in several
  studies (e.g., Ma et al. [7], Gao et al. [10]) highlight a potential gap in the exploration of optimal model configurations, potentially impacting the models' entire performance.
- **Inconsistent evaluation metrics:** Another identified gap is the inconsistency in the choice of evaluation metrics across studies. While some focus on accuracy, others may neglect other essential metrics like F1 score, precision, and Recall, leading to an incomplete assessment of model performance.
- **Sparse adoption of K-Fold cross validation:** Few studies employ K-fold cross-validation to validate their models rigorously. This approach provides a more robust understanding of a model's generalizability, yet it remains underutilized in the current literature.

2.4. Future directions: In our state-of-the-art analysis, numerous future directions proposed by previous studies provide valuable insights into the evolving landscape of research and innovation.

- **Standardized practices:** There is a need for standardized practices, including consistent hyperparameter tuning and reporting guidelines, to facilitate reproducibility and comparison across studies.
- **Transfer learning exploration:** Future research could further explore the potential of transfer learning in sensor-based HAR, leveraging knowledge from pre-trained models to improve generalization.
- **Handling class imbalance:** Strategies to address class imbalance should be a focus of future work to enhance the robustness and applicability of models in real-world scenarios.

2.5 Addressing gaps in our proposed model: In light of the identified gaps, our work makes significant strides in advancing the field:

- **In-depth hyperparameter tuning:** Our proposed model incorporates hyperparameter tuning using Bayesian optimization. This deliberate approach enhances our model's adaptability and performance, addressing the previously observed gap.
- **Comprehensive evaluation metrics:** To overcome the inconsistency in evaluation metrics, we conduct a comprehensive assessment, including accuracy, F1 score, precision, and Recall. This ensures a thorough understanding of our model's performance across various dimensions.
- **Rigorous K-Fold cross validation:** Recognizing the importance of model validation, we implement a rigorous 10-fold cross-validation methodology. This validation strategy ensures the reliability and generalizability of our model's performance, addressing the underutilization of K-fold cross-validation in previous studies.

In summary, our work contributes to the evolution of sensor-based HAR by introducing an optimized LSTM model, specifically addressing gaps related to hyperparameter tuning, evaluation metrics, and model validation. Through these advancements, we offer a refined and optimized deep-learning model tailored to the intricacies of wearable sensor data in smart home environments.

3 Material and methods

The study introduces an LSTM-based Human Activity Recognition framework (as shown in Figure 1) that utilizes wearable sensor data from the PAMAP2 dataset. The sensors are placed on the chest, ankle, and hand to collect data related to 12 activities performed by individuals. The methodology involves three main stages: data preprocessing and segmentation in the first stage, data splitting into training and testing sets in the second stage, and training and hyperparameter tuning, followed by the model evaluation in the final stage. During the model training and tuning phase, the data is split into 70% for the training set and 30% for the testing set. Our proposed LSTM model is tested using the validation data, and hyperparameters are optimized using the Bayesian optimization approach. Subsequently, the hyperparameter-tuned models are evaluated using the test data to measure their recognition performance and to compare their effectiveness.

3.1 PAMAP2 dataset*:*

The dataset used in this study is The PAMAP2 dataset, which is widely employed in HAR research due to its extensive usage and relevance in this field. It contains sensor data from wearable devices, including inertial measurement units (IMUs) and physiological sensors. The dataset comprises recordings of various physical activities performed by participants, covering a wide range of movements and intensities. Multiple participants were involved, allowing for the study of individual variations in activity recognition. Each activity recording is labeled, providing ground truth data for Training and evaluating HAR models. The dataset's structured format includes separate files for different sensor modalities,

facilitating analysis and combining data for activity recognition. The PAMAP2 dataset is a valuable resource for advancing sensor-based activity recognition research. Table 3 presents a comprehensive overview of the PAMAP2 dataset, incorporating information from the provided documentation and our conducted experiment [22].

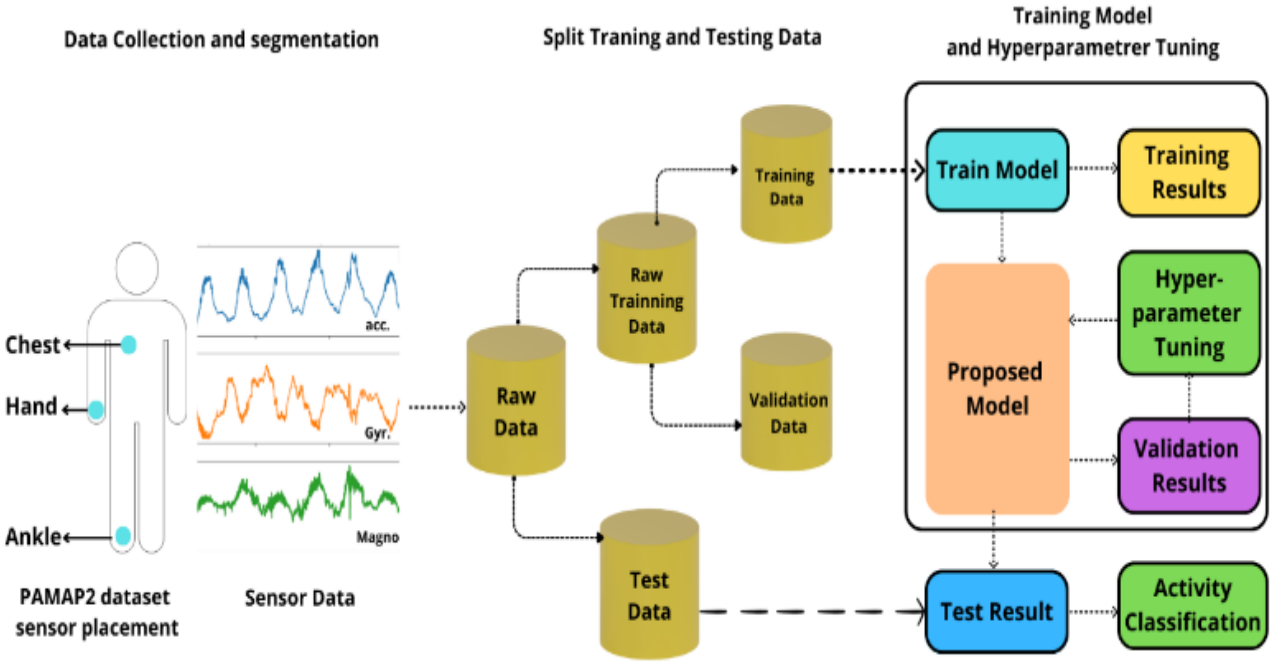

Figure 1: The schematic diagram of the proposed LSTM-Based model for sensor-based HAR

Table 3: PAMAP2 dataset description

Dataset	**Labels**	**Sampling rate**	**Windows size**	**Overlap**	**Features vector**	**Total of Segments** (20588)	
						Training 70%	***Testing 30%***
PAMAP2	12	100 Hz	1s	50%	(20588, 42)	(14412,100,42)	(6176,100,4)

Table 4 and Figure 2 show the number of instances per activity in the PAMAP2 dataset. Despite the imbalanced distribution, both the training and testing sets contain instances for all activities, ensuring that the model can be evaluated on the entire range of activities present in the dataset.

Table 4: PAMAP2 dataset instances per activity data distribution

Activity No	Class id	Activity label	#Instances
1	0	Lying	142931
2	1	Sitting	83738
3	2	Standing	99973
4	3	Walking	122906
5	4	Running	43050
6	5	Cycling	91340
7	6	Nordic Walking	111832
12	7	Ascending stairs	59314
13	8	Descending stairs	46830
16	9	Vacuum cleaning	86959
17	10	Ironing	125228
24	11	Rope jumping	15453

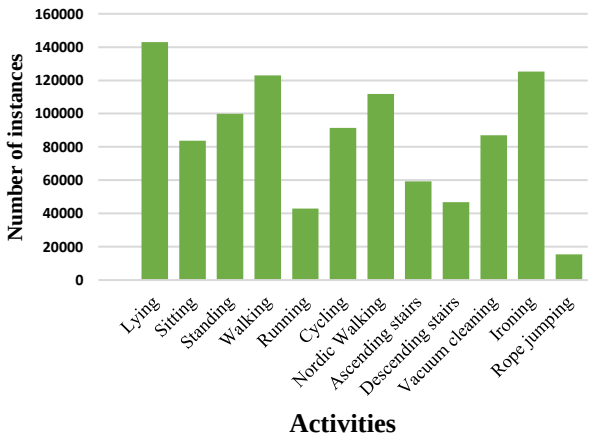

Figure 2: Instances per activity data distribution in PAMAP2 dataset

3.2 Long short-term memory (LSTM)

LSTM networks are a subset of recurrent neural networks (RNNs) and play a crucial role in time series applications, especially in Human Activity Recognition (HAR). HAR categorizes activities based on sensor data like accelerometer and smartphone gyroscope readings. The effectiveness of LSTM networks in HAR stems from their ability to effectively capture and represent long-term dependencies inherent in sensor data [23].

Figure 3 shows the internal structure of LSTM consists of several components, including:

1)Input gate: (it) This gate manages the flow of information from the input to the memory cell. It consists of a sigmoid activation function that produces an output value ranging from 0 to 1, determining the extent to which the input should be allowed to pass through.

$$i_t = \sigma\,(W_i\, xt + U_i\, ht_{-1} + b_i)\ \textit{(Equation 1)}$$

2)Forget gate: (ft) This gate manages the flow of information from the previous memory cell to the current memory cell. Additionally, it consists of a sigmoid activation function that produces an output value between 0 and 1, denoting the degree to which the previous memory cell should be forgotten.

$$f_t = \sigma(W_f * x_t + U_f * h_{t-1} + b_f)\ \textit{(Equation 2)}$$

3) Output gate (Ot): This gate manages the flow of information from the memory cell to the output. It consists of a sigmoid activation function that generates an output value between 0 and 1, signifying the proportion of the memory cell to be output.

$$o_t = \sigma\,(W_o\, xt + U_o\, ht_{-1} + b_o)\ \textit{(Equation 3)}$$

4) Temporary Cell Content (˜Ct): A candidate vector of new cell content that can be added to the cell state.

$$\tilde{C}_t = tanh\,(W_c\, xt + U_{c\ ht-1} + b_c)\ \textit{(Equation 4)}$$

5)Cell state(c_t): This is the internal state of the memory cell that is updated based on the input gate, the forget gate, and the memory cell.

$$c_t = f_t\, ct_{-1} + i_t\, \tilde{C}_t\ \textit{(Equation 5)}$$

6)Hidden State (h_t): The output of the LSTM cell is a filtered version of the cell state.

$$h_t = \sigma_t * tanh(c_t)\ \textit{(Equation 6)}$$

In the above equations:

- x_t is the input at time step t.
- h*t-1* is the previous hidden state (output) of the LSTM at time step t-1.
- σ represents the sigmoid activation function.
- W_f, W_i, W_o, W_c are weight matrices for the input.
- U_f, U_i, U_o, U_c are weight matrices for the previous hidden state.
- b_f, b_i, b_o, b_c are biased terms.
- tanh represents the hyperbolic tangent activation function.

By employing these gate operations along with the memory cell, LSTM can adeptly capture long-range

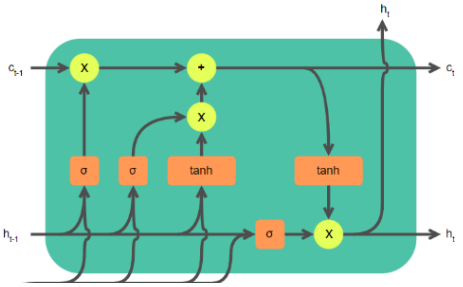

Figure.3. The Internal Structure of The LSTM Cell[31]

dependencies in sequential data and make precise predictions for future time steps.

LSTM Process for HAR:

1) Data Preparation: The initial step involves preprocessing the raw sensor data to extract essential features like acceleration, velocity, and orientation. Subsequently, the data is partitioned into fixed-length sequences, each representing a distinct activity.

2)Input Encoding: To facilitate input into the LSTM network, sensor data sequences undergo encoding. This typically entails transforming the data into a three-dimensional tensor with dimensions (samples, time steps, features).

3)Model Training: Next, the LSTM network is trained on the encoded sensor data to discern patterns and correlations between input sequences and their respective activity labels. During Training, the LSTM network's internal state is continually updated based on the input sequence, enabling predictions about the activity label.

4) Model Prediction: Once the LSTM network completes its Training, it becomes equipped to predict activity labels for new sensor data sequences. These sequences are fed into the LSTM network, which dynamically updates its internal state according to the input and produces the anticipated activity label.

3.3 Hyperparameter tuning with bayesian optimization

Hyperparameters are critical parameters in deep learning approaches as they directly influence the behavior of training algorithms and substantially impact the performance of deep learning models. Bayesian optimization emerges as a practical and efficient method for solving function optimization problems prevalent in computing, especially when seeking optimal model configuration. This approach is particularly suited for tackling related-function problems characterized by the absence of a closed analytical form. Bayesian optimization proves applicable to addressing various related function challenges, including computationally demanding tasks, intricate derivative evaluations, and non-convex functions [24] [25].

To use Bayesian optimization for time series problems and sensor HAR LSTM, the following steps can be followed:

1. *Define the search space:* Specify the hyperparameters to be optimized, such as the number of LSTM layers, the number of hidden units, the learning rate, the dropout rate, etc.
2. *Define the objective function:* This function evaluates the performance of the LSTM model using the given hyperparameters.
3. *Initialize the Bayesian optimization algorithm:* Set the initial hyperparameter values and corresponding objective function values.
4. *Iterate the optimization process:* The Bayesian optimization algorithm leverages a probabilistic model and an acquisition function to determine the subsequent set of hyperparameters for evaluation. Subsequently, the objective function is assessed with these hyperparameters, and the outcomes are utilized to update the probabilistic model.
5. *Repeat step four until convergence:* The optimization procedure persists until a specified stopping condition is satisfied, such as completing a predetermined number of iterations or achieving a targeted level of performance.

3.4 Batch normalization

Batch normalization is a technique utilized in deep learning, including LSTM networks, to address the internal covariate shift problem during Training. It normalizes each layer's inputs in a mini-batch, making the data more centered around zero with unit variance. This leads to improved training stability, faster convergence, and reduced sensitivity to weight initialization. Batch normalization is a crucial tool for enhancing the efficiency and accuracy of neural network models, including those used in HAR and time-series classification tasks[26].

3.5 Validation protocol

The K-fold cross-validation protocol is a widely employed technique in machine learning for assessing model performance. It entails partitioning the dataset into subsets or folds [27]. In this study, we use the 10-fold cross-validation. The model is trained on nine folds in each iteration and then validated on the remaining fold. This process is repeated multiple times to ensure robust evaluation. This process is repeated ten times, and performance metrics are averaged across iterations to obtain an overall performance estimate. In this study, we use this validation protocol to provide a more robust and less biased assessment of the model's ability to generalize to new data, making efficient use of available data for evaluation.

3.6 Evaluation metrics

The experiment used various evaluation metrics to assess the HAR model's performance. These metrics included accuracy, F1 score, precision, Recall, and the confusion matrix. Accuracy measures the general correctness of the model's predictions. The F1 score balances precision and Recall. Precision assesses the accuracy of positive predictions made by the model, whereas Recall evaluates the model's capability to identify positive instances correctly. The confusion matrix offers a comprehensive view of the model's performance across various classes, providing insights into true positive, true negative, false positive, and false negative classifications. Together, these metrics comprehensively evaluate the HAR model's accuracy, reliability, and predictive capabilities for various human activities. Table 5 summarizes these evaluation metrics, including accuracy, precision, Recall, and F-measure. Understanding these performance metrics requires knowledge of four fundamental terms used in their measurement: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). [28]

Table 5: Evaluation metrics [28]

Metric	Formula	Definition
Accuracy	$\frac{tp+tn}{tp+tn+fp+fn}$	the ratio of correct predictions and overall predictions
Precision	$\frac{tp}{tp+fp}$	The ratio of correct predictions to the total predicted
Recall of sensitivity	$\frac{tp}{tp+fn}$	the ratio of correct predictions to the samples in the actual class
F1 score / F-measure	$\frac{2(recall*precision)}{recall+precision}$	The weighted average of precision and Recall if the data is imbalanced

4 Experiments and results

4.1 Experimental design

The experimental design of this study is structured to systematically investigate the effectiveness of optimizing the LSTM-based proposed model for sensor-based Human Activity Recognition (HAR) in smart homes through hyperparameter tuning. The research questions (RQ) and corresponding hypotheses(H) guide the study's objectives and validate the proposed model's performance.

RQ1: *How does hyperparameter tuning impact the performance of LSTM models in sensor-based Human Activity Recognition?*

H1: Systematic hyperparameter tuning significantly enhances the accuracy and robustness of LSTM models compared to default configurations

RQ2: *How does the proposed optimized LSTM model perform compared to previous Studies?*

H2: The optimized LSTM model will outperform other models in terms of accuracy, precision, Recall, and F1 score.

RQ3: *How does the inclusion of batch normalization in the LSTM model affect its convergence speed and overall performance in Human Activity Recognition tasks?*
H3: The addition of batch normalization will contribute to faster convergence and improved model performance by mitigating internal covariate shifts.

RQ4: *How applicable are the findings of our optimized LSTM model to real-life smart home environments, considering practical challenges and variations in user behaviors?*
H4: The model's performance will remain robust in real-world scenarios, offering practical implications for smart home applications.

4.2 Experimental environment and hyperparameter optimization setup

This section presents the results obtained from the experiments performed on an NVIDIA GPU T4 using the Google Colab platform. The LSTM network's hyperparameters were optimized through Bayesian Hyperparameter Optimization, utilizing the Keras Tuner library[29]. The experiment setup is detailed in Table 6.

Table 6: Experiment environment setup

Platform	Google Colab
GPU	NVIDIA GPU T4
RAM	15 GB
Tenserflow version	2.12.0
Keras Version	2.12.0
Keras Tuner Version	1.3.5

4.3 The proposed model

Our proposed model is an LSTM-based HAR (as shown in Figure 4). The model consists of three LSTM layers, which are particularly effective for capturing sequential patterns in time-series data.
Dropout layers are incorporated between the LSTM layers to prevent overfitting and enhance generalization. Dropout randomly deactivates certain neurons during Training, effectively reducing the model's reliance on specific features and encouraging more robust learning. Additionally, batch normalization is applied to stabilize the training process by normalizing the input to each layer. This ensures a more consistent and faster convergence during Training. The combination of dropout and batch normalization contributes to the model's ability to handle the complexity of sensor data, leading to improved performance in classifying human activities accurately.

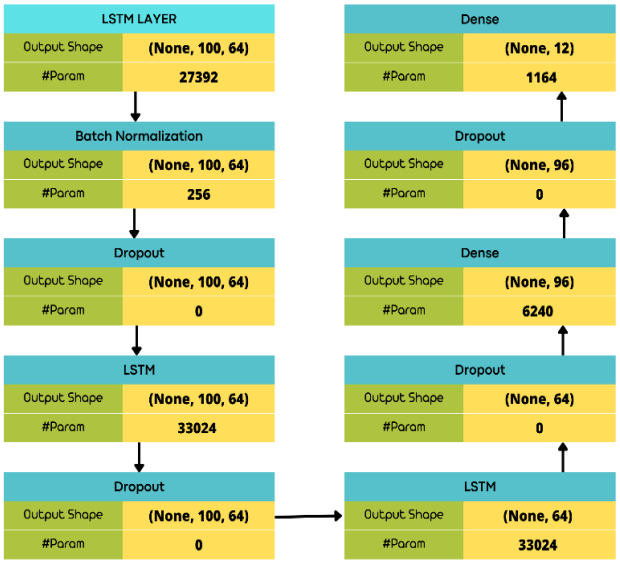

Figure 4: Structure of the proposed model

4.4 Experiments process

The experiment is conducted in three distinct stages: data preprocessing, Training, and hyperparameter tuning, and model Evaluation.

4.4.1. Data preprocessing

In the experiment's data preprocessing stage, the raw sensor data obtained from wearable devices is prepared for the proposed model. The data is first cleaned by dropping irrelevant orientation columns and removing transient activity rows. Non-numeric data is converted to

numeric, and missing values are interpolated to ensure data completeness. The data is then scaled to normalize the input features, enabling uniformity in the data distribution.

Next, the labels are encoded and converted into categorical variables. This step is crucial for classifying the activities during model training. The data is then split into training and testing sets, with 70% of the data used for Training and 30% for testing. The data is segmented into overlapping windows to facilitate the LSTM model's input format. The window size is set to 1s, and overlapping is 50 %.

The data segmentation stage creates segments and corresponding labels for Training and testing. The segments and labels are reshaped to ensure compatibility with the LSTM model's input format. Finally, the experiment confirms the shape of the training and testing segments before proceeding to the model training and evaluation stages (demonstrated in Table 3).

4.4.2. Training and hyperparameter tuning

This stage aims to optimize the model's performance. Hyperparameter tuning is conducted using Bayesian optimization, intelligently training the model while searching for the best combination of hyperparameters. This fine-tuning process enhances accuracy and generalization in classifying human activities.

Table 7: Hyperparameter ranges for model optimization

Hyperparameter	Range
LSTM Units	Integer from 64 to 256 with a step of 32
Dense Units	Integer from 32 to 128 with a step of 32
Dropout Rate	Float from 0.1 to 0.5 with a step of 0.1
Optimizer	Choice of 'ADAM,' 'RMSprop', or 'SGD'
Learning Rate	Choice of 1e-3, 1e-4, or 1e-5
Batch Size	Choice of 32, 64, or 128

Table 7 summarizes the hyperparameter Ranges. These ranges represent the search space for hyperparameters during the Bayesian Hyperparameter Optimization process using the Keras Tuner. The models are fine-tuned by adjusting several critical hyperparameters: the LSTM Units, Dense Units, Dropout Rate, Optimizer, Learning Rate, and Batch Size. Through exploration and Tuning of the proposed models, this eventually results in improved accuracy and robust performance.

In this study, the Bayesian Optimization process using the Keras Tuner library performs 10 trials to intelligently explore the hyperparameter space and identify the most optimal configurations for the proposed LSTM-based HAR model. Each trial involves tuning the hyperparameters while training the model for 50 epochs to ensure comprehensive learning and convergence. The combination of 10 trials and 50 epochs contributes to a thorough search and fine-tuning of the model, leading to improved accuracy and robust performance in HAR tasks.

Table 8 summarizes the hyperparameters found by the Keras Tuner library using Bayesian Optimization for the best LSTM-based HAR model. These hyperparameters include 64 LSTM units, 96 dense units, a dropout rate of 0.1, a batch size of 128, a learning rate of 0.001 with the ADAM optimizer, and 50 epochs for Training. These optimized hyperparameters lead to a well-balanced and efficient model that effectively classifies human activities with improved precision and performance.

Table 8: The Summarized hyperparameters of the proposed model found by keras tuner

Structure Hyperparameters	
LSTM Units	64
Dense Units	96
Dropout rate	0.1
Training Hyperparameters	
Batch Size	128
Learning rate	0.001
Optimizer	Adam
Epochs	50
Loss Function	Cross-entropy

4.4.3. Model evaluation

To evaluate the proposed model's performance on the PAMPA2 dataset, a 10-fold cross-validation approach was employed. The evaluation metrics used include accuracy, precision, Recall, and F1-score. These metrics were compared against those reported in previous literature studies conducted on the same dataset, enabling a comprehensive assessment of the proposed model's effectiveness and advancements in HAR.

4.5 Experimental results

In this section, we discuss the experimental results of the proposed method in terms of accuracy, F1 score, precision, and Recall. To prove the capability of our proposed model, we also compare its results with other approaches from numerous previous studies, as demonstrated in section 2.

The performance of our proposed LSTM-based model was assessed using 10-fold cross-validation, as shown in Table 9. The results showed consistent and reliable performance across all folds, with a mean cross-validation score of 97.71% and a small standard deviation of +/- 0.4. The model achieved an average F1 score of 0.96660, accurately classifying positive and negative instances. The precision score averaged at 0.96855, reflecting the model's ability to correctly predict positive instances, while the average recall score was 0.96549, showing its capability to identify positive instances out of all actual positives. The model's accuracy ranged from 97.16% to 98.54% across folds, with an average accuracy of 97.71%. The consistent high performance and minor variation in

these metrics indicate that the proposed LSTM-based HAR model effectively classifies human activities.

Table 9: The model performance on 10-fold cross-validation.

Fold	Cross-validation score	F1 score	Precision	Recall	Accuracy
Fold1	0.97365	0.96233	0.96435	0.96093	97.36%
Fold2	0.97157	0.96127	0.96107	0.96171	97.16%
Fold3	0.97641	0.96925	0.97149	0.96914	97.64%
Fold4	0.97363	0.96312	0.96830	0.95911	97.36%
Fold5	0.97363	0.96323	0.96558	0.96122	97.36%
Fold6	0.98126	0.97075	0.97248	0.96923	98.13%
Fold7	0.97918	0.96522	0.96495	0.96626	97.92%
Fold8	0.98543	0.97364	0.97676	0.97135	98.54%
Fold9	0.97918	0.96447	0.96671	0.96292	97.92%
Fold10	0.97710	0.97267	0.97382	0.97302	97.71%
Mean	**0.97.71 +/- 0.4**	**0.96660**	**0.96855**	**0.96549**	**97.71%**

The classification report Table 10 presents the evaluation metrics for a multi-class classification model. The report shows precision, Recall, and F1-score for each activity. Overall, the model demonstrates strong performance with an accuracy of 97%, indicating a high level of correct predictions across all classes.

Table 10: Classification report of the proposed model

Activity	Precision	recall	F1-score
Lying	0.99	1.00	1.00
Sitting	0.99	0.99	0.97
Standing	0.96	0.89	0.93
Walking	1.00	0.99	1.00
Running	0.98	0.95	0.97
Cycling	1.00	0.99	1.00
Nordic Walking	0.98	1.00	0.99
Ascending stairs	0.91	0.87	0.89
Descending stairs	0.85	0.89	0.87
Vacuum cleaning	0.98	0.98	0.98
Ironing	0.99	0.98	0.98
Rope jumping	0.89	0.95	0.92

The plot of accuracy and loss demonstrates the model's performance during Training, showcasing the increase in accuracy and decrease in loss over epochs, as shown in Figure 5. The confusion matrix of the proposed model is illustrated in Figure 6.

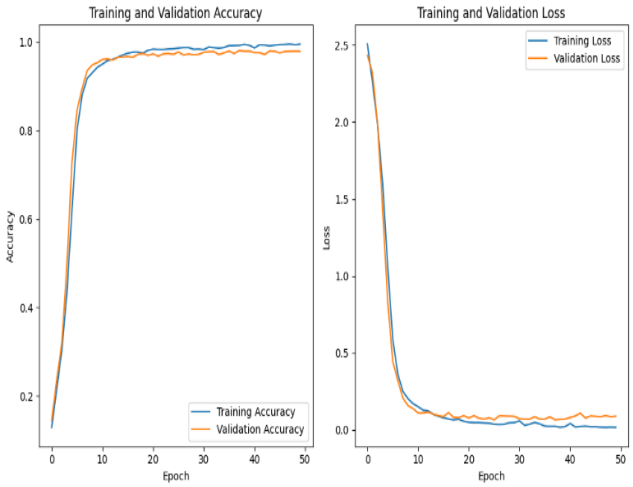

Figure 5: Accuracy and loss training performance of our proposed model

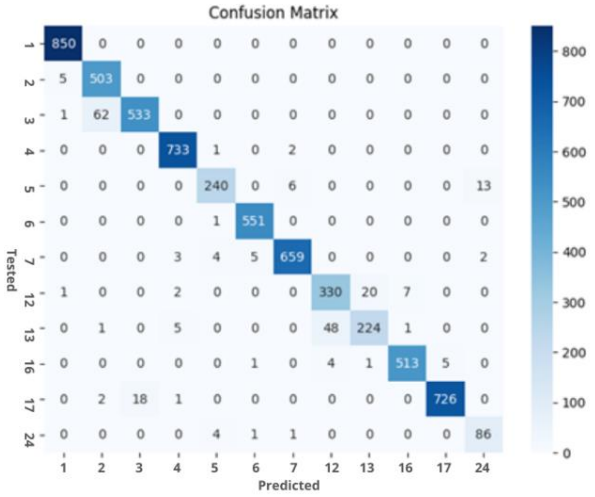

Figure 6: Confusion matrix of the proposed model.

4.6 Comparative results analysis

In this comparative results analysis, we evaluate the performance of our proposed LSTM-based model with hyperparameter tuning and batch normalization against a selection of previous studies in Human Activity Recognition (HAR), examining key metrics including accuracy, F1 score, precision, and recall as demonstrated in Table 11.

Table 11: Comparison with previous works

Study	Year	Classification method	Accuracy	F1 score	Precision	Recall
Hammerla et al[6]	2016	CNN	-	93,70%	-	-
Hammerla et al[6]	2016	LSTM	-	92,90%	-	-
Ma et al.[7]	2019	AttnSense	-	89,30%	-	-
Xu et al.[8]	2019	Inno HAR	-	93,50%	-	-
Wan et al[9]	2020	CNN	91,00%	91,16%	91,66%	90,85%
Wan et al[9]	2020	LSTM	85,86%	85,34%	86,51%	84,67%
Wan et al[9]	2020	Bi-LSTM	89,52%	89,40%	90,19%	89,02%
Gao et al.[10]	2020	DanHAR	93,16%	-	-	-
Xu et al.[11]	2022	Inception-LSTM	95,04%	95,13%	95,06%	95,21%
Thakur et al.[12]	2022	ConvAE-LSTM	94,33%	94,46%	-	-
Tehrani et al.[13]	2023	Bi-LSTM	-	93,41%	93,41%	93,47%
Challa et al.[14]	2023	CNN+biLSTM+HP Tuning	94,91%	-	-	-
Kumar et al. [15]	2023	GRU	94,77%	-	-	-
Our proposed model		**An LSTM-based model with HP Tuning**	**97,71%**	**96,66%**	**96,85%**	**96,55%**

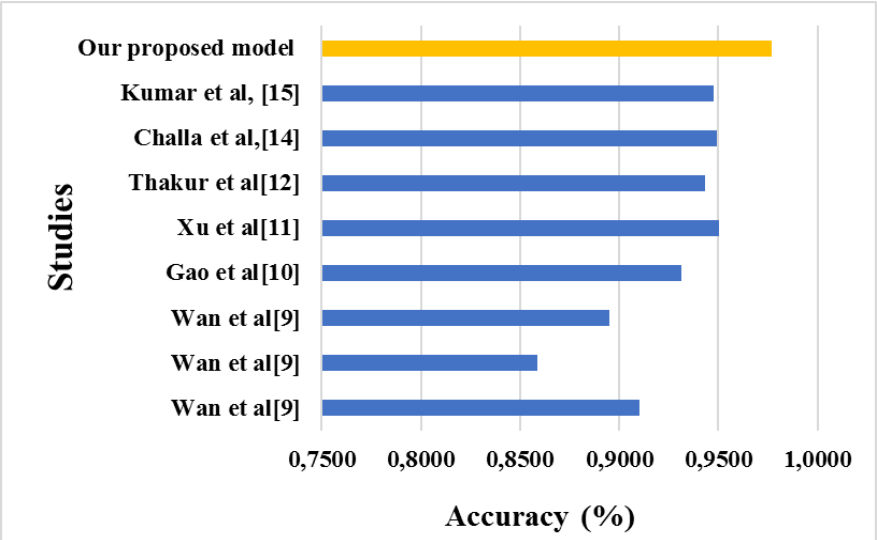

Figure 7: Accuracy comparison of our proposed model against previous studies

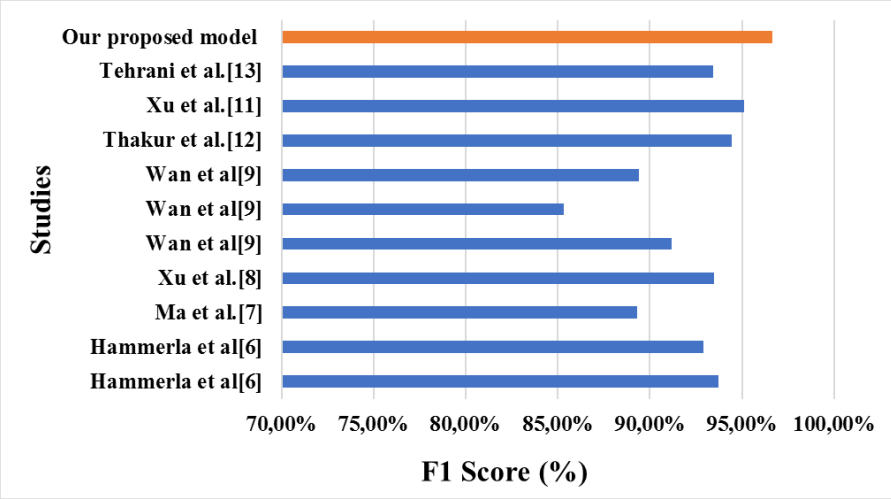

Figure 8: F1 score comparison of our proposed model against previous studies

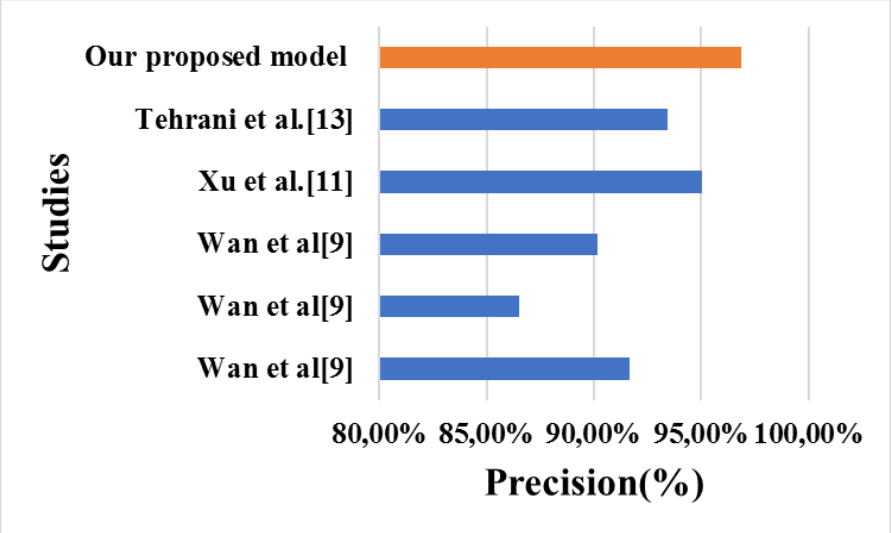

Figure 9: Precision comparison of our proposed model against previous studies

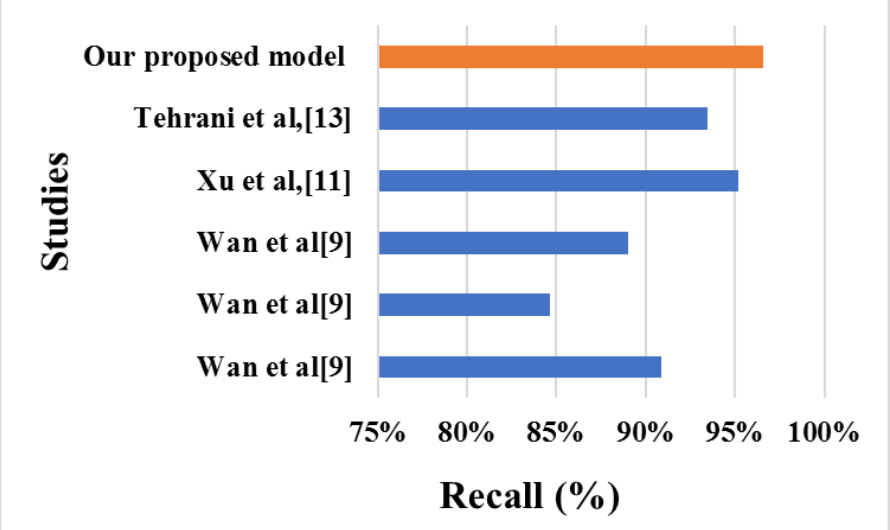

Figure 10: Recall comparison of our proposed model against previous studies

Table 11 compares the performance of our proposed LSTM-based model and previous studies on the PAMAP2 dataset. The results demonstrate the remarkable efficacy of our model, achieving an accuracy of 97.71%, an F1 score of 96.66%, a precision of 96.855%, and a recall of 96.549%. These metrics collectively outperform all previously detailed state-of-the-art studies listed in Table 11. Particularly noteworthy is the improvement over the closest competitor, the Inception-LSTM by Xu et al. (2022) [11], with enhancements of approximately 2.67% in accuracy, 1.53% in F1 score, 1.795% in precision, and 1.329% in Recall. Figures 7 to 10 present visualization charts comparing our model to previous studies regarding accuracy, F1 score, precision, and Recall, respectively.

This considerable advancement underscores the effectiveness of our model's architecture, hyperparameter tuning, and batch normalization in pushing the boundaries of sensor-based Human Activity Recognition. The substantial improvement in accuracy holds significant implications for real-world applications, emphasizing the practical relevance of our contributions.

4.7 Discussion

In addressing the complex challenges of sensor-based Human Activity Recognition (HAR) for smart home environments, our study aimed to optimize Long Short-Term Memory (LSTM) models through hyperparameter tuning, batch normalization, and rigorous evaluation and validation. We delve into the results, providing a detailed analysis to answer our research questions and validate our hypotheses.

4.7.1. Impact of hyperparameter tuning

Our investigation into the impact of hyperparameter tuning on our proposed LSTM model performance unveils nuanced insights into specific model parameters. The sensitivity of deep learning models, such as LSTM, to hyperparameter changes makes the manual search for optimal configurations challenging.

Our study employed tools like Keras Tuner, leveraging Bayesian optimization to systematically fine-tune critical hyperparameters, including LSTM Units, Dense Units, Dropout Rate, Optimizer, Learning Rate, and Batch Size.

The adjustments made in LSTM Units demonstrated a notable effect on the model's ability to capture long-term dependencies, contributing significantly to enhanced accuracy. Similarly, fine-tuning Dense Units allowed for a more nuanced representation of complex patterns within the sensor data, further improving the model's robustness.

Our exploration of the Dropout Rate emphasized its impact on regularization, mitigating overfitting risks and promoting model generalization. The choice of Optimizer played a pivotal role in optimizing convergence speed and overall model performance, with the model showcasing superior results with the selected optimization strategy.

Additionally, adjustments in Learning Rate and Batch Size showcased their influence on the model's learning dynamics and computational efficiency. Bayesian optimization through Keras Tuner facilitated an efficient search for the best configuration, considering the intricate interplay of these hyperparameters.

The quantitative improvements observed across these individual parameters collectively underscore the importance of meticulous hyperparameter tuning. This validated our hypothesis and provided a granular understanding of how each parameter contributes to the overall robustness and accuracy of our LSTM-based model.

4.7.2. The effect of batch normalization

Our exploration of batch normalization's effects on the LSTM model reveals notable improvements. Batch normalization contributes to faster convergence and enhances overall model performance, mitigating internal covariates [26]. This technique normalizes the inputs of a layer during Training, leading to faster and more stable Training of deep neural networks. This aligns with our hypothesis and underscores the significance of normalization techniques in optimizing deep learning models for HAR tasks.

4.7.2. Comparison with previous studies

Quantitatively comparing our optimized LSTM model with state-of-the-art studies on the PAMAP2 dataset demonstrates its superior performance. The model excels in accuracy, precision, Recall, and F1 score, outperforming previous models. The model's mean accuracy across all folds was approximately 97.71%, with a small standard deviation of 0.4%. The mean F1-score was approximately 96.66%, indicating a good balance between precision and Recall. Compared to previous literature, the proposed model outperformed most other approaches.

The practical implications of our optimized LSTM model indicate that the model's robustness extends to real-world smart home scenarios, including applications in healthcare, fitness tracking, and human-computer interaction. When discussing potential applications, risks, and ethical implications, we recognize the need for ongoing ethical discourse in the rapidly evolving landscape of technology.

4.7.4. Statistical analysis

In this analysis, we opted for the Wilcoxon signed-rank test, a robust non-parametric method tailored for evaluating a single algorithm's performance across diverse studies [30]. This approach aligns with our objective of comparing the efficacy of your proposed model against other studies in a scenario marked by a single algorithm being tested across multiple contexts. The Wilcoxon test, being non-parametric, is particularly advantageous for these analyses, as it eliminates the necessity for strict assumptions regarding the distribution of the data. The p-values obtained from the Wilcoxon test are instrumental in providing valuable indications about the statistical significance of the

observed differences [30]. With a conventional significance level of 0.05, the obtained p-value of 0.0079 for accuracy signifies a statistically significant distinction. This result underlines a significant divergence in the accuracy of your model compared to other studies. Similarly, the F1 score exhibits a p-value of 0.00195, emphasizing a substantial difference in your model's F1 score relative to the studies. Precision and Recall, while yielding p-values of 0.0625 each, fall just short of conventional significance levels.

The statistical analysis underscores the significant outperformance of your proposed model in terms of accuracy and F1 score. While precision and recall differences may not be statistically significant at the 0.05 threshold, they signal intriguing nuances deserving further exploration, contributing to a comprehensive understanding of your model's comparative performance.

In Summary, Our Proposed framework, deep multi-layer LSTM with Bayesian Optimization and batch normalization, achieved outstanding results in accurately classifying human activities using data from various wearable sensors. The three-stage experimental setup, including data preprocessing, hyperparameter tuning using Bayesian optimization, and model validation with 10-fold cross-validation, contributed to the model's robustness and effectiveness. The optimized hyperparameters, including LSTM units of 64, dense units of 96, and a dropout rate of 0.1, were identified through Bayesian optimization. The model's mean accuracy across all folds was approximately 97.71%, with a small standard deviation of 0.40%. The mean F1-score was approximately 0.9666, indicating a good balance between precision and Recall. Compared to previous literature, the proposed model outperformed most other approaches. The model's capability to handle multi-sensor data and its successful hyperparameter tuning make it highly applicable in real-world scenarios. However, further testing on diverse datasets and real-world conditions is warranted to validate its generalizability. Overall, our proposed model showcases remarkable performance and potential for practical applications in various domains.

5 Conclusion

In conclusion, our research significantly advances the Human Activity Recognition (HAR) field using wearable sensor data, with a particular focus on smart home environments. Through an exhaustive review of the state of the art, we have presented a comprehensive understanding of existing methods, classification techniques, hyperparameter tuning approaches, findings, limitations, and future directions.

Our proposed LSTM-based deep model, enhanced by batch normalization and hyperparameter tuning using Bayesian optimization, has demonstrated exceptional performance. Achieving an accuracy of 97.71% and impressive values for F1 score, precision, and Recall (approximately 96.66%, 96.85%, and 96.55%, respectively), our model outperforms previous studies, underscoring the crucial role of hyperparameter optimization in activity classification. Looking ahead, we aim to evaluate our model further on diverse datasets such as OPPORTUNITY and WISDM to enhance its generalization capabilities. Our commitment to ongoing optimization involves exploring more complex deep model architectures and alternative hyperparameter tuning approaches. This pursuit aligns with our goal of maximizing efficiency and adaptability in real-world scenarios.

References

[1] S. S. Zhang *et al.*, “Deep Learning in Human Activity Recognition withWearable Sensors: A Review on Advances,” *Sensors*, vol. 22, no. 4, p. 1476, Feb. 2022. https://doi.org/10.3390/s22041476

[2] M. Gochoo, F. Alnajjar, T. H. Tan, and S. Khalid, “Towards privacy-preserved aging in place: A systematic review,” *Sensors*, vol. 21, no. 9, p. 3082, Apr. 2021. https://doi.org/10.3390/s21093082

[3] A. Kristoffersson and M. Lindén, “A systematic review on the use of wearable body sensors for health monitoring: A qualitative synthesis,” *Sensors (Switzerland)*, vol. 20, no. 5, p. 1502, Mar. 2020. https://doi.org/10.3390/s20051502

[4] M. M. Islam, S. Nooruddin, F. Karray, and G. Muhammad, “Human activity recognition using tools of convolutional neural networks: A state of the art review, data sets, challenges, and future prospects,” *Comput. Biol. Med.*, vol. 149, no. August, p. 106060, Oct.2022. https://doi.org/10.1016/j.compbiomed.2022.106060

[5] Y. Zhang, I. D’haeseleer, J. Coelho, V. Vanden Abeele, and B. Vanrumste, “Recognition of bathroom activities in older adults using wearable sensors: A systematic review and recommendations,” *Sensors*, vol. 21, no. 6, pp. 1–23, Mar. 2021. https://doi.org/10.3390/s21062176

[6] N. Y. Hammerla, S. Halloran, and T. Plötz, “Deep, convolutional, and recurrent models for human activity recognition using wearables,” *IJCAI Int. Jt. Conf. Artif. Intell.*, vol. 2016-Janua, pp. 1533–1540, 2016. https://doi.org/10.48550/arXiv.1604.08880

[7] H. Ma, W. Li, X. Zhang, S. Gao, and S. Lu, “Attnsense: Multi-level attention mechanism for multimodal human activity recognition,” *IJCAI Int. Jt. Conf. Artif. Intell.*, vol. 2019-Augus, pp. 3109–3115, 2019. https://doi.org/10.24963/ijcai.2019/431

[8] C. Xu, D. Chai, J. He, X. Zhang, and S. Duan, “InnoHAR: A deep neural network for complex human activity recognition,” *IEEE Access*, vol. 7, no. c, pp. 9893–9902, 2019.

https://doi.org/10.1109/ACCESS.2018.2890675

[9] S. Wan, L. Qi, X. Xu, C. Tong, and Z. Gu, "Deep Learning Models for Real-time Human Activity Recognition with Smartphones," *Mob. Networks Appl.*, vol. 25, no. 2, pp. 743–755, Apr. 2020. https://doi.org/10.1007/s11036-019-01445-x

[10] W. Gao, L. Zhang, Q. Teng, J. He, and H. Wu, "DanHAR: Dual Attention Network for multimodal human activity recognition using wearable sensors," *Appl. Soft Comput.*, vol. 111, pp. 1–11, Jun. 2021. https://doi.org/10.1016/j.asoc.2021.107728

[11] Y. Xu and L. Zhao, "Inception-LSTM Human Motion Recognition with Channel Attention Mechanism," *Comput. Math. Methods Med.*, vol. 2022, 2022. https://doi.org/10.1155/2022/9173504

[12] D. Thakur, S. Biswas, E. S. L. L. Ho, and S. **Chat**topadhyay, "ConvAE-LSTM: Convolutional Autoencoder Long Short-Term Memory Network for Smartphone-Based Human Activity Recognition," *IEEE Access*, vol. 10, pp. 4137–4156, Jun. 2022. https://doi.org/10.1109/ACCESS.2022.3140373

[13] A. Tehrani, M. Yadollahzadeh-Tabari, A. Zehtab-Salmasi, and R. Enayatifar, "Wearable Sensor-Based Human Activity Recognition System Employing Bi-LSTM Algorithm," *Comput. J.*, no. April, 2023. https://doi.org/10.1093/comjnl/bxad035

[14] S. K. Challa, A. Kumar, V. B. Semwal, and N. Dua, "An optimized deep learning model for human activity recognition using inertial measurement units," *Expert Syst.*, vol. 40, no. 10, p. e13457, Dec. 2023. https://doi.org/10.1111/exsy.13457

[15] P. Kumar and S. Suresh, "RecurrentHAR: A Novel Transfer Learning-Based Deep Learning Model for Sequential, Complex, Concurrent, Interleaved, and Heterogeneous Type Human Activity Recognition," *IETE Tech. Rev.*, vol. 40, no. 3, pp. 312–333, May 2023.https://doi.org/10.1080/02564602.2022.2101557

[16] S. Kalabakov, M. Gjoreski, H. Gjoreski, and M. Gams, "Analysis of deep transfer learning using deepconvlstm for human activity recognition fromwearable sensors," *Informatica*, vol. 45, no. 2, pp. 289–296, 2021. https://doi.org/10.31449/inf.v45i2.3648

[17] H. A. Imran, "Khail-Net: A Shallow Convolutional Neural Network for Recognizing Sports Activities Using Wearable Inertial Sensors," *IEEE Sensors Lett.*, vol. 6, no. 9, 2022. https://doi.org/10.1109/LSENS.2022.3197396

[18] R. Piltaver, B. Cvetkovic, and B. Kaluža, "Denoising human-motion trajectories captured with ultra-wideband real-time location system," *Informatica*, vol. 39, no. 3, pp. 311–322, 2015.

[19] M. Luštrek and B. Kaluža, "Fall detection and activity recognition with machine learning," *Informatica*, vol. 33, no. 2, pp. 205–212, 2009.

[20] D. Roggen *et al.*, "Collecting complex activity datasets in highly rich networked sensor environments," *INSS 2010 - 7th Int. Conf. Networked Sens. Syst.*, pp. 233–240,2010. https://doi.org/10.1109/INSS.2010.5573462

[21] Kwapisz, J. R., Weiss, G. M., & Moore, S. A. (2011). Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter*, *12*(2), 74-82.

[22] "PAMAP2 Physical Activity Monitoring - UCI Machine Learning Repository." https://archive.ics.uci.edu/dataset/231/pamap2+physical+activity+monitoring (accessed Jul. 21, 2023).

[23] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780,Nov.1997. https://doi.org/10.1162/NECO.1997.9.8.1735

[24] J. Suto, "The effect of hyperparameter search on artificial neural network in human activity recognition," *Open Comput. Sci.*, vol. 11, no. 1, pp. 411–422, 2021. https://doi.org/10.1515/comp-2020-0227

[25] S. Raziani and M. Azimbagirad, "Deep CNN hyperparameter optimization algorithms for sensor-based human activity recognition," *Neurosci. Informatics*, vol. 2, no. 3, p. 100078, Sep. 2022. https://doi.org/10.1016/j.neuri.2022.100078

[26] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *32nd Int. Conf. Mach. Learn. ICML 2015*, vol. 1, pp. 448–456, 2015.

[27] X. Zhang and C. A. Liu, "Model averaging prediction by K-fold cross-validation," *J. Econom.*, vol. 235, no. 1, pp. 280–301, 2023. https://doi.org/10.1016/j.jeconom.2022.04.007

[28] M. El Ghazi and N. Aknin, "A Comparison of Sampling Methods for Dealing with Imbalanced Wearable Sensor Data in Human Activity Recognition using Deep Learning", vol. 14, no. 10, pp. 290–305, 2023. https://doi.org/10.14569/IJACSA.2023.0141032

[29] "KerasTuner." https://keras.io/keras_tuner/ (accessed Jul. 23, 2023).

[30] R. Woolson, "Wilcoxon Signed-Rank Test", *Wiley Encyclopedia of Clinical Trials*,2008. https://doi.org/10.1002/9780471462422.eoct979

[31] "File:LSTM cell.svg - Wikimedia Commons." https://commons.wikimedia.org/wiki/File:LSTM_cell.svg (accessed Jul. 23, 2023).