algorithms

MDPI

Article

Human Activity Recognition: A Comparative Study of Validation Methods and Impact of Feature Extraction in Wearable Sensors

Saeed Ur Rehman [1], Anwar Ali [2,*], Adil Mehmood Khan [1] and Cynthia Okpala [1]

[1] Faculty of Science and Engineering, University of Hull, Cottingham Rd., Hull HU6 7RX, UK; s.rehman2@hull.ac.uk (S.U.R.); a.m.khan@hull.ac.uk (A.M.K.)
[2] Department of Electronic and Electrical Engineering, Swansea University Bay Campus, Swansea SA1 8EN, UK
* Correspondence: anwar.ali@swansea.ac.uk

Abstract: With the increasing availability of wearable devices for data collection, studies in human activity recognition have gained significant popularity. These studies report high accuracies on k-fold cross validation, which is not reflective of their generalization performance but is a result of the inappropriate split of testing and training datasets, causing these models to evaluate the same subjects that they were trained on, making them subject-dependent. This study comparatively discusses this validation approach with a universal approach, Leave-One-Subject-Out (LOSO) cross-validation which is not subject-dependent and ensures that an entirely new subject is used for evaluation in each fold, validated on four different machine learning models trained on windowed data and select hand-crafted features. The random forest model, with the highest accuracy of 76% when evaluated on LOSO, achieved an accuracy of 89% on k-fold cross-validation, demonstrating data leakage. Additionally, this experiment underscores the significance of hand-crafted features by contrasting their accuracy with that of raw sensor models. The feature models demonstrate a remarkable 30% higher accuracy, underscoring the importance of feature engineering in enhancing the robustness and precision of HAR systems.

Keywords: machine learning; LOSO; human activity recognition

Citation: Rehman, S.U.; Ali, A.; Khan, A.M.; Okpala, C. Human Activity Recognition: A Comparative Study of Validation Methods and Impact of Feature Extraction in Wearable Sensors. *Algorithms* **2024**, *17*, 556. https://doi.org/10.3390/a17120556

Academic Editor: Antonio Della Cioppa

Received: 24 September 2024
Revised: 25 October 2024
Accepted: 27 October 2024
Published: 5 December 2024

1. Introduction and Background

Weiser's initial vision of ubiquitous computing foresaw computers seamlessly integrating into daily life, operating inconspicuously to enhance human experience without intrusion [1]. Today, this vision has materialized, with computers seamlessly incorporated into personal smart devices. Pervasive computing, a key aspect, achieves non-intrusive functionality by allowing computers to detect and respond to user actions implicitly based on their environment. The emergence of wearable sensor devices represents a significant development in the ongoing activity recognition research landscape, enabling individuals to monitor their physical activity seamlessly.

Smartphones, smart watches, or bands are the most common wearable sensors. They are embedded with motion sensors such as accelerometers, gyroscopes, and magnetometers. Accelerometers gauge the acceleration of an object for the rate of change in its velocity, measured in meters per second (m/s). Gyroscopes determine the orientation and angular velocity, measured in degrees per second (°/s), and magnetometers measure the magnetic field strength at a specific position, denoted in tesla units (T) [2].

The activity recognition chain [3] propounds the framework for activity classification applied to machine learning. It summarily involves data reading from sensors, preprocessing, segmentation, feature selection, and classification as can be seen in Figure 1. While most HAR studies rely on this protocol, its effectiveness can be stymied by validation methods, which are techniques used to estimate how well they will generalize to unseen data by evaluating them on different subsets of the available data [4].

Figure 1. The process of the activity recognition chain.

Studies in HAR have typically reported high accuracies [5,6]. They have all also been evaluated using questionable k-fold cross-validation. K-fold is a common validation approach in HAR systems, and it achieves high accuracies in generalization [7]. The approach randomly splits the data into training and testing subsets, repeating the process in the specified *k* times. For systems that are trained on participants' continuous data, this can cause a leak, where participant samples exist in both training and testing sets, leading to classification accuracy, as these data have already been seen [8,9]. Ensuring trained models are tested on unseen samples should be the true test of their predictive accuracy.

This work investigates the impact of validation methodologies on classification algorithms for HAR systems in wrist-worn wearables. It analyzes how different HAR models trained on feature vectors generalize to k-fold cross-validation and LOSO. The performance of models trained on raw data will also be examined.

In summary, this study examines the generalization performance of HAR systems trained on extracted features and validated on k-fold and LOSO cross-validation. We highlight the inappropriate splitting and data leakage in k-fold cross-validation leading to higher accuracies over LOSO and compare the testing performance of HAR systems trained on raw data to those trained on extracted features, validating the importance of feature vectors. We propose a combination of training models on extracted feature vectors as well as validating with LOSO for better-generalized HAR systems.

Background

Previous studies have highlighted the inappropriateness of the traditional train–test split and k-fold cross-validation. Ref. [10] emphasized the tendency for clinical studies that utilize data segmented from the same subject data and employ k-fold cross-validation to frequently overestimate the accuracy of predictions. They suggest the possibility that overlapping frames likely capture identical activity within a similar context; consequently, these adjacent segments are highly correlated. Ref. [11] corroborate this with their study demonstrating that assessing performance with a randomized training and test split indicates remarkably accurate identification of eight activities, attaining a 96% F1 score in the context of singular participant and stationary position data. However, when subjected to thorough backtesting, the F1 score declines, settling at 54%, indicating challenges with robustness and generalization.

Ref. [3] propose subject cross-validation in their educational approach to the common challenges in the ARC chain of HAR. Ref. [12] highlighted the impact of overlapping windows, a common segmentation technique, in the correlation challenges of k-fold cross-validation, recommending subject CV (further explained in Section 3) as a performance evaluator for HAR systems. A recent study by [13] further validates this conclusion on three datasets of varying modalities: the CASAS containing binary motion sensors, MHEALTH, and PAMAP2 on-body inertial sensors trained on Random Forest (RF) and Graph Neural Networks (GNNs). These showed that the traditional techniques' reported accuracy is highly overestimated, regardless of the data type.

Ref. [14] introduced a residual learning framework designed to simplify the training of much deeper networks than those traditionally used. By restructuring the layers to learn residual functions that relate to the layer inputs rather than learning functions without reference, ref. [14] made training more manageable. Comprehensive empirical results demonstrate that these residual networks are more straightforward to optimize and benefit from significantly increased depth, achieving higher accuracy. Ref. [15] propose an efficient approach to reduce DenseNet redundancy by replacing its bottleneck with their SMG

module, enhanced with local residuals. The SMG module uses a two-stage pipeline designed to integrate previous outputs. It gradually condenses redundant features with hierarchical convolutions, followed by multi-kernel depth-wise convolutions. This results in a compact output with richer, multi-scale features.

This work explores the effectiveness of LOSO based on training models on extracted features. These systems show a more valid and lower generalization accuracy of 76% when validated with LOSO and 89% on k-fold cross-validation.

2. Dataset and Methodology

2.1. Dataset

For this study, we explore PAMAP2 (Physical Activity Monitoring Dataset), an open-source dataset available at the UCI Machine Learning Repository [16]. It contains a variety of physical activities and postures recorded under laboratory conditions.

Collection

Nine subjects performed an extensive range of physical activities, including a combination of household and sport activities. The participants wore three wireless Inertial Measurement Units (IMUs) and a heart rate monitor and were made to follow the same protocol when performing 12 key activities. These sensors were placed on the chest, the wrist of the dominant arm, and the ankle. Each IMU contains a 3D magnetic sensor, two 3D accelerometers, and a 3D gyroscope. The IMU sampling frequency is 100Hz, and the sampling frequency of the heart rate monitor is 9Hz, which means that data are collected at 0.01 s and 0.111 s, respectively.

In Figure 2, the distribution of activities is represented, including activities like walking, which are more prevalent than rope-jumping.

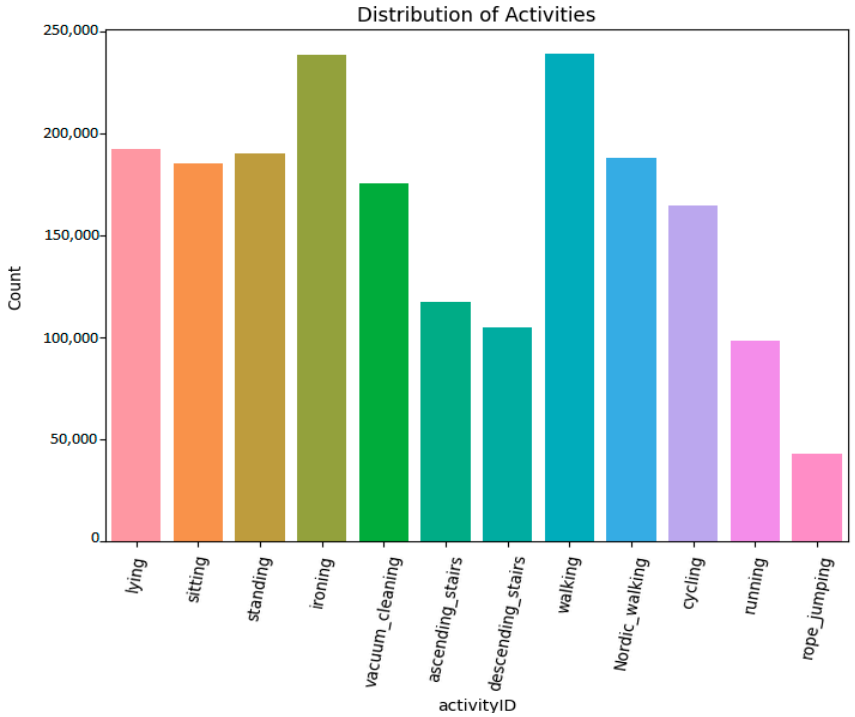

Figure 2. Distribution of Activities in PAMAP2 dataset.

2.2. Methodology

This section describes the proposed methodology. Adapted from the ARC chain framework, it comprises two modules. The first module involves HAR systems trained on extracted time and frequency domain features from hand sensors and validated on k-fold

and LOSO cross-validation. The second module involves training raw sensor data on fewer classifiers and validating them as described above. The impact of each module is compared, highlighting their benefits and drawbacks.

2.2.1. Data Preprocessing

The raw data contained noise due to how it was collected, making this step mandatory before performing classification.

Missing values, indicated with *NaN*, were found in all sensor features. They were handled by linear interpolation, which estimates values based on the surrounding data points [17].

Redundant features of time stamp, 6g acceleration, and orientation were removed. The 6g acceleration and orientation were established as invalid by the authors. Activity code 0 represents a break between activities and was dropped from the dataset. Subject 9 only performed 5 of the 18 activities; therefore, this subject's data were also dropped.

2.2.2. Feature Engineering and Exploratory Data Analysis

Exploratory Data Analysis (EDA) was carried out to visually present trends in the dataset using the Python Seaborn and Matplotlib libraries [18–24]. For intensity and visualization of the sensor data, the magnitude of the sensor features was derived. Box plots of the sensors, along with heart rate and temperature features, were plotted for outlier detection. While extreme outliers were detected, as seen in Figure 3, none were removed, as abnormal data readings are expected during vigorous activities.

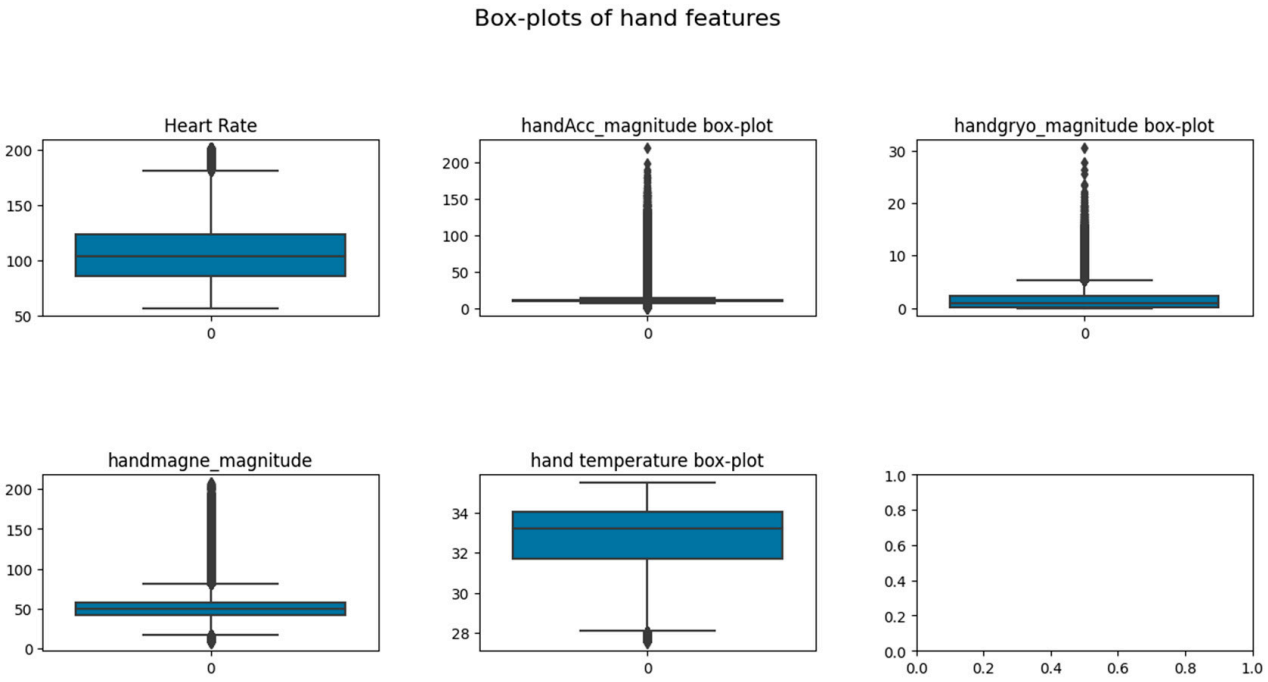

Figure 3. Box plot of hand features.

The relationships between the features and trends in the sensors were explored. Figure 4 shows the variations in the hand sensor reading across the activities, with the gyroscope capturing the most variation in activities compared to the other devices, especially during rapid movements.

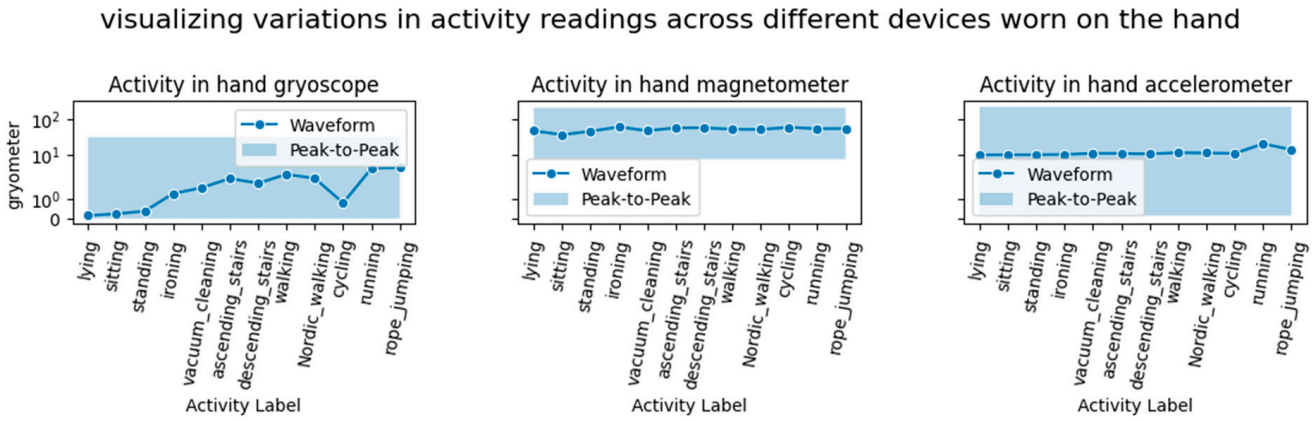

Figure 4. Variation in activities based on sensor readings.

2.2.3. Feature Segmentation and Extraction

Segmentation using windowing is important for extracting meaningful information from real-time data, as it facilitates the recognition of patterns. It also improves computational efficiency, as it provides a structured input format for ML models and processes smaller segments, improving computational efficiency [25]. For this study, a sliding window size of 200 was applied with an overlap of 100. This represents 2 s of sensor reading, as it was collected at 100 Hz.

Some experiments utilize only accelerometer readings [26]. Gyroscopic data were included in this study, as they clearly showed the most variation in activities (Figure 4). For each x, y, and z-axis, hand-crafted (manually engineered) time and frequency domain features were extracted. The time–domain features were derived by performing statistical calculations that offer simple and low computational complexity, while the frequency–domain features capture inherent patterns and trends by decomposing the data into real and imaginary values, which represent wave characteristics. This decomposition is achieved using methods such as Fourier or Wavelet transforms [27]. From this, 70 hand-crafted features were extracted. Table 1 provides a list of features extracted from the sensors.

Table 1. List of time and frequency domain features extracted from the sensors.

Feature	Domain
Mean	Time
Median	Time
Max/min	Time
Kurtosis	Time
Standard deviation	Time
Variance	Time
Covariance between axes	Frequency
Energy	Frequency
Entropy	Time

This step was bypassed in the second module of this experiment, as raw sensor data were input into the classifiers [28].

2.2.4. Feature Selection

To ensure that the models were trained on the most optimal features, Minimum Redundancy Maximum Relevance (mRMR), a method of feature selection, was used to select 50 features. This is a filter-based algorithm that combines the condition of minimum redundancy, achieved by avoiding highly correlated features, and maximum relevance, maximized by prioritizing features that are most important to the dependent variable, into a function.

2.2.5. Modeling and Classification

Subject data were set to an equal number of data points to ensure equal training data size in both validation techniques, resulting in 170,000 rows of data for all subjects.

Features for both modules were trained on four machine learning classifiers: random forest (RF), logistic regression (LR), support vector machines (SVMs), and k-nearest neighbors (KNN) and implemented on scikit learn [26]. For optimization, a randomized grid search was used to select some of the best hyperparameters for the RF classifier, and an elbow plot was used in selecting an appropriate number of neighbors (k) for the KNN model using cross-validation.

2.3. Evaluation

The validation methods can be classified into two main categories.

(i) K-fold cross-validation involves splitting the dataset into k parts. While one part is reserved for testing, the other will be used in training. This is repeated k times, and a different part is used each time. This method reports higher accuracy but struggles with generalization.

(ii) Leave-One-Subject-Out (LOSO) cross-validation is a variant of k-fold cross-validation, but one subject is left out of the entire fold for testing. The approach employs a testing subset of size p and a training subset of size $n - p$, where p retains all samples from a specific subject together. This process is iterated n times, ensuring that a given subject is not simultaneously present in both testing and training sets. This approach facilitates model generalization by incorporating data from diverse subjects.

All of the trained models are validated using k-fold and LOSO cross-validation. For k-fold, the k value was set to 8, meaning each time, 7 of 8 of the data points were used for training, and 1 of 8 was used for testing. This value was set to match the training data for LOSO for standardized comparison. For LOSO, each time, seven of the eight subjects' data were used for training, and the remaining one was used for testing.

Accuracy, precision, and recall were used to calculate the performance metrics of all the approaches. While accuracy evaluates correct classifications, precision and recall measure performance in the predicted classes.

3. Results

This section presents the results of the two experiment areas: the performance of the four models in k-fold and LOSO cross-validation. Secondarily, the performance of the raw sensor models is compared with HC models on RF and LR classifiers.

3.1. Performance of k-Fold and LOSO Methods Across Models

Figure 5 shows the training and testing accuracies of all four models validated on eight-fold and LOSO CV.

Both validation methods achieved relatively equal training accuracies across all four models, with random forest achieving the highest result of 97%. In testing, expectedly, LOSO achieved lower accuracies than k-fold in all models. This can be deduced to be caused by the variability in how the different subjects performed the activities, which the LOSO models were less effective in differentiating, especially when closely tied to individual users.

In analyzing the performance of k-fold and LOSO on the random forest, Table 2 provides a classification report for k-fold in each of the eight folds, and Table 3 is a classification report for LOSO in each subject-specific test instance. For subject 8, the model was trained on subjects 1–7 and evaluated on 8.

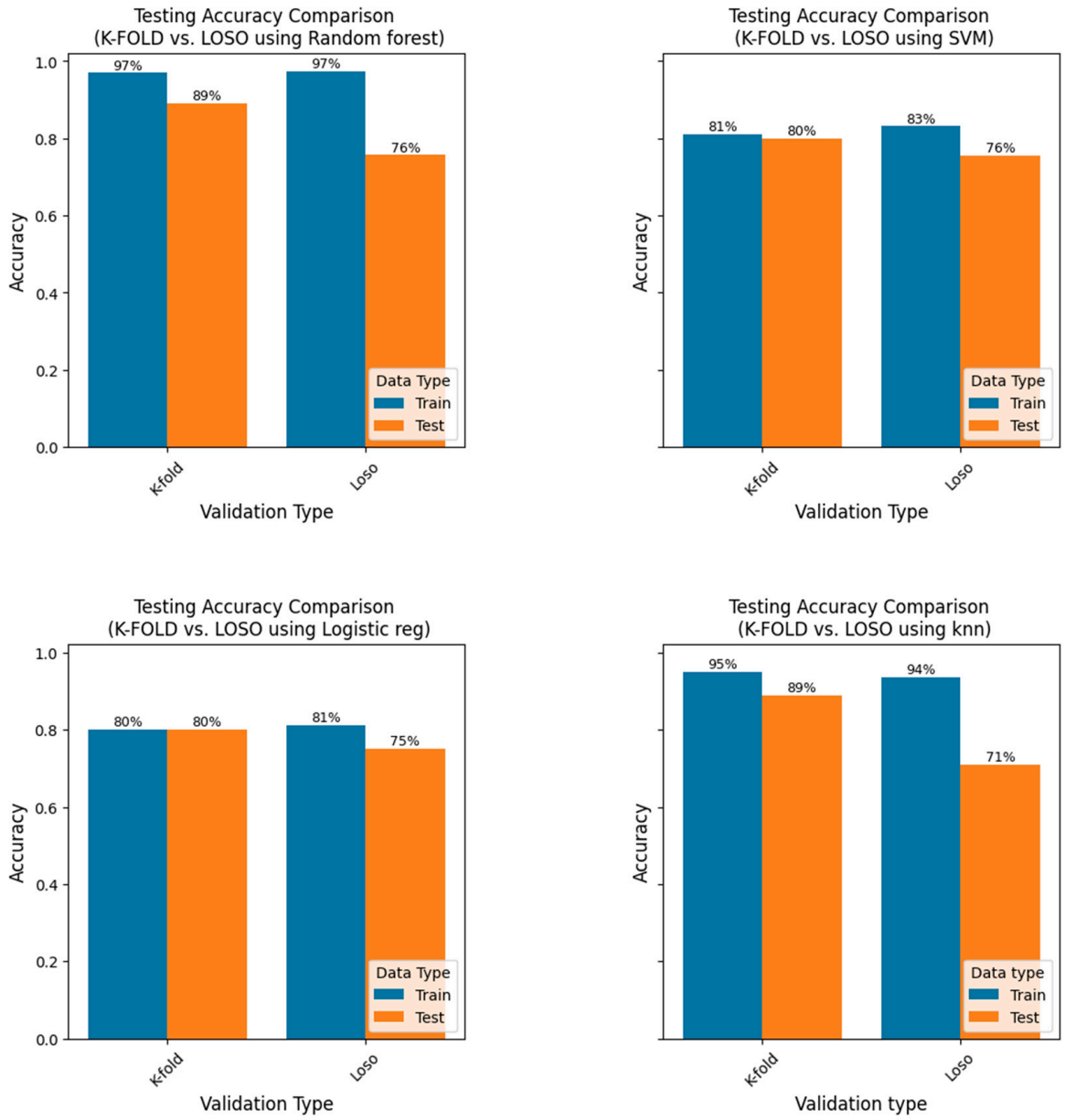

Figure 5. Plot of all models' training vs. testing accuracies for k-fold and LOSO.

Table 2. Classification report of random forest on each fold of k-fold.

	Accuracy	Precision	Recall	Model
0	0.890995	0.901599	0.893519	random forest
1	0.890338	0.911733	0.892361	random forest
2	0.895080	0.914623	0.896021	random forest
3	0.873740	0.899322	0.875867	random forest
4	0.882039	0.890860	0.886018	random forest
5	0.880261	0.897396	0.877654	random forest
6	0.889152	0.909028	0.891709	random forest
7	0.882632	0.907564	0.885401	random forest

Table 3. Classification report of random forest on each subject on LOSO.

Model	Subject	Accuracy	Precision	Recall	F1-Score
Rf_model	1	0.684866	0.693740	0.684866	0.667359
	2	0.812204	0.825516	0.812204	0.808839
	3	0.770006	0.785704	0.770006	0.759448
	4	0.755187	0.784565	0.755187	0.753585
	5	0.623815	0.579001	0.623815	0.583957
	6	0.772970	0.780498	0.772970	0.768426
	7	0.783175	0.810397	0.783175	0.784631
	8	0.852401	0.862090	0.852401	0.854501

As observed based on the performances, there is a huge variation in accuracy amongst the subjects in LOSO compared to k-fold, where the variation is negligible. This highlights the capacity of LOSO to capture variability among subjects.

Tables 4 and 5 are classification reports of k-fold and LOSO. k-fold generally achieved higher, precision which means it correctly classified the activities better than LOSO. This lower LOSO classification can be explained as differences among subjects in the training and testing datasets.

Table 4. Classification report of k-fold on random forest.

	Precision	Recall	F1 Score	Support
1	0.89	0.72	0.79	1893
2	0.69	0.76	0.73	1819
3	0.77	0.78	0.77	1868
4	0.89	0.73	0.80	1629
7	0.00	0.00	0.00	59
12	0.57	0.64	0.60	1138
13	0.75	0.56	0.64	1017
16	0.79	0.83	0.81	1721
17	0.73	0.90	0.80	2353
accuracy			0.76	13,497
macro avg	0.68	0.66	0.66	13,497
weighted avg	0.76	0.76	0.76	13,497

Table 5. Classification report of LOSO on random forest.

	Precision	Recall	F1 Score	Support
1	0.98	0.91	0.95	1893
2	0.91	0.87	0.89	1819
3	0.83	0.89	0.86	1868
4	0.96	0.93	0.95	1629
7	0.97	1.00	0.98	59
12	0.88	0.82	0.85	1138
13	0.92	0.76	0.83	1017
16	0.89	0.86	0.88	1721
17	0.80	0.94	0.86	2353
Accuracy			0.89	13,497
Macro avg	0.90	0.89	0.89	13,497
Weighted avg	0.89	0.89	0.89	13,497

In analyzing specific activities, LOSO achieved 0% across all metrics for class 7, Nordic walking. This means it could not predict a single correct instance. This was caused by the low support value of 59 for this class, meaning that a substantial number of instances for this class was not available and was not represented in each training instance, leading to poor model generalization. K-fold, however, achieved 98% accuracy for the same class

LOSO, and k-fold achieved the highest recall of 90% and 98% in the lying activity, class 1. This static activity is less characterized by a person, meaning that the slight accelerations involved in these activities are likely similar across various groups of individuals. The opposite is true for class 12, ascending stairs, which people with different fitness levels and body types potentially perform differently.

3.2. Performance of Raw Data Models and Extracted HC Feature Models

The testing accuracy of the models trained on raw sensor signals was compared with that of the results above for the random forest and logistic regression classifiers.

From Figure 6, in both validation techniques, the models performed better when trained on HC features than raw data, resulting in an absolute percentage difference of 30% and 25% for k-fold and LOSO, respectively, on the random forest. This is attributed to the interpretability of HC features, as they are designed with insight information regarding each activity. Table 1 presents the detailed characteristics of the extracted features, presenting a meaningful representation of the activities and leading to improved performance. Inputting raw data into the models creates a challenge in this area, as the models might not interpret the complex patterns that they are learning.

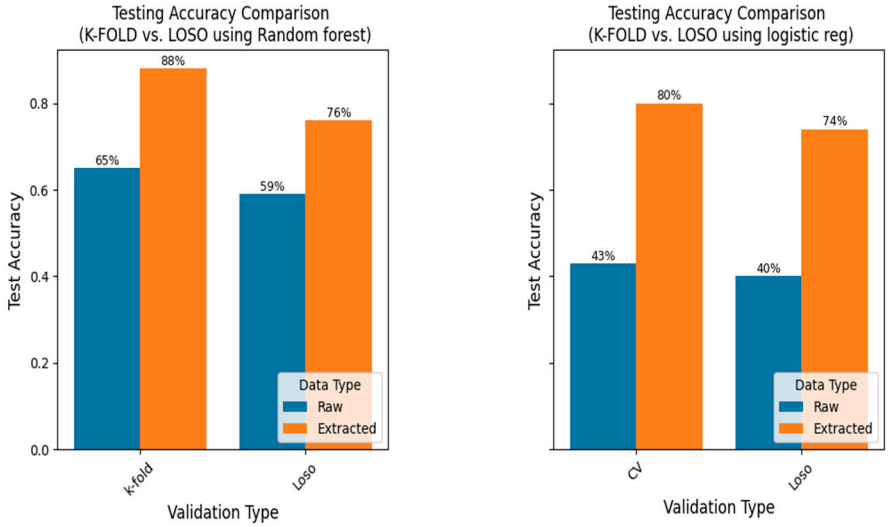

Figure 6. Testing accuracy comparison of raw sensor models and feature-trained models validated with k-fold and LOSO on random forest and logistic regression classifiers.

In both experiment modules, the random forest achieved the highest accuracy. This attributed to:

(i) Its ensemble learning property, which combines the predictions of multiple decision trees, allowing it to adapt well to different activity characteristics, even where they may exhibit complex and diverse patterns.
(ii) Its measure of feature importance, selecting the most relevant features that contribute more to the model's predictive performance. Where features are crucial to distinguishing between activities, it assigns higher importance to these features.
(iii) Its robustness in handling imbalanced classes, which is the case in our dataset, as seen in Tables 4 and 5.

4. Discussion

The core aim of this experiment was to emphasize the exaggerated accuracy derived from k-fold cross-validation in HAR models. From Section 3, LOSO achieved a lower accuracy compared to k-fold. This is due to subject-specific data leakage in the former. K-fold cross-validation randomly splits the data into subsets, with the possibility of subject-specific information presenting in multiple folds simultaneously, leading to unintentional

data leakage. This results in the model appearing more robust during k-fold validation, but less so when faced with entirely new subjects in LOSO validation.

Therefore, LOSO accuracy, which rigorously tests the model's ability to generalize to unseen subjects, provides a more accurate evaluation in the presence of subject-specific data and more realistic results in real-life deployment.

Considering that overlapping windows and a small window size of 2 s were used in this experiment, overlap between adjacent windows in subject data creates dependency, also causing over-estimation in k-fold generalization. As activities span multiple overlapping windows, predictions for one window are influenced by information in the neighboring windows, leading to dependencies between the training and testing fold and overestimation of the model's generalization performance [29–37]. With overlapping dependence occurring within the same subject, LOSO is not impacted by this, as it focuses on one subject at a time during testing.

For a personalized activity recognition system that is tailored to a specific user, k-fold would be better suited. Ref. [38] consider it a hybrid of the subject-independent models, which refers to the LOSO approach, and subject-dependent models, which are models built specifically with the data of the final user. While the subject-independent approach expectedly achieves the highest accuracy of all personalized models, it is not realistic in real-life use because the models cannot be built for every user, except in special cases. This makes k-fold a preferred alternative.

In considering performance improvements on LOSO to develop a model that generalizes better, training with a similar population to user characteristics should be considered. Evaluating the subject-level performance of LOSO in Table 6, the highest testing accuracy of 87% was achieved when the model was trained on subjects 1–7 and tested on subject 8, while the lowest accuracy was achieved when the model was trained on the other subjects and tested on subject 5. These results are caused by the semblance and dissimilarity between the validation subject and those included in the training set, respectively. This means that subject 8 has a high similarity with the trained subjects, while subject 5 has a low similarity. Figure 7 captures this trend. Training models on data similar to those of its final user increases its adaptability and performance. Previous studies [39–50] highlight this, where fitness level contributed to the characterization of sporting activities like running, and demographics like gender and body physique informed moderate activities like walking.

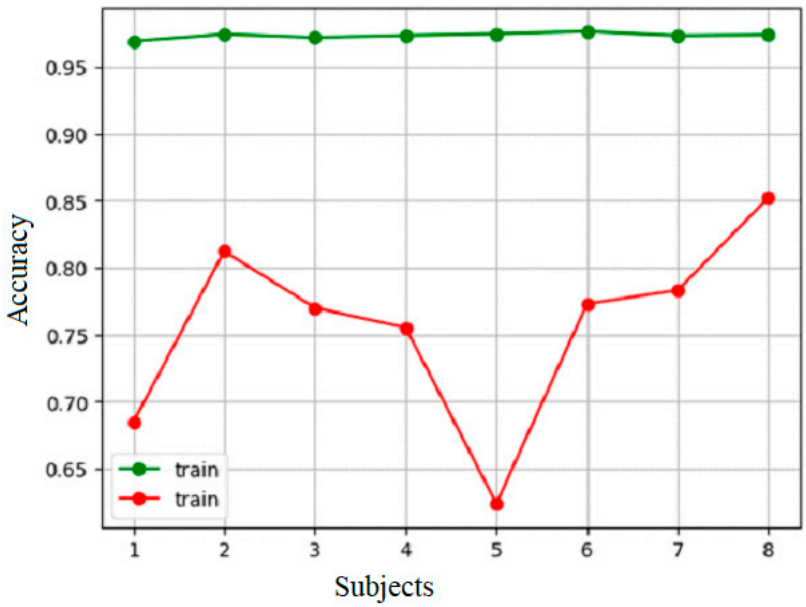

Figure 7. Training vs. testing accuracy in each subject in LOSO.

Table 6. Classification report of all models in each LOSO test subject instance.

Model	Subject	Accuracy	Precision	Recall	F1 Score
Rf_model	1	0.684866	0.693740	0.684866	0.667359
	2	0.812204	0.825516	0.812204	0.808839
	3	0.770006	0.785704	0.770006	0.759448
	4	0.755187	0.784565	0.755187	0.753585
	5	0.623815	0.579001	0.623815	0.583957
	6	0.772970	0.780498	0.772970	0.768426
	7	0.783175	0.810397	0.783175	0.784631
	8	0.852401	0.862090	0.852401	0.854501
knn_model	1	0.613650	0.589627	0.613650	0.594988
	2	0.732227	0.744817	0.732227	0.724324
	3	0.710136	0.714393	0.710136	0.698735
	4	0.768820	0.783423	0.768820	0.768041
	5	0.548578	0.520800	0.548578	0.520092
	6	0.724956	0.729998	0.724956	0.717539
	7	0.778436	0.785914	0.778436	0.778748
	8	0.804979	0.806358	0.804979	0.804501
lr_model	1	0.736499	0.741289	0.736499	0.727663
	2	0.803318	0.808627	0.803318	0.792767
	3	0.743331	0.766390	0.743331	0.732931
	4	0.800830	0.817979	0.800830	0.801867
	5	0.592417	0.545891	0.592417	0.561564
	6	0.772377	0.784693	0.772377	0.770848
	7	0.718009	0.673292	0.718009	0.677105
	8	0.836989	0.847843	0.836989	0.836116
svm_model	1	0.719288	0.705113	0.719288	0.700663
	2	0.816351	0.824913	0.816351	0.808968
	3	0.765857	0.796976	0.765857	0.758134
	4	0.794309	0.816359	0.794309	0.793946
	5	0.599526	0.563832	0.599526	0.569796
	6	0.789567	0.798942	0.789567	0.787912
	7	0.704976	0.657929	0.704976	0.664564
	8	0.859514	0.861714	0.859514	0.859766

The secondary focus of this study was on the importance of hand-crafted features in training HAR systems, which was achieved by training on raw sensors and comparing the results with the former. While this was established in the earlier section, segmentation and windowing also positively impacted the performance of the HC models. By segmenting into smaller windows, relevant patterns like the sequence of movements of activity over time were captured by the models within smaller time intervals. Their computational efficiency was also improved, as the models were processed in smaller chunks.

5. Conclusions and Future Work

In this paper, we targeted the generalization challenge associated with k-fold validation in a real-world scenario while also highlighting the relevance of feature engineering. Our results show a drop in performance when validated on LOSO over k-fold. We conclude that HAR systems should be trained on extracted features instead of raw sensor data, and that LOSO is best suited for universal recognition systems, while k-fold is suitable for personalized models. One potential method to enhance the performance of LOSO is to utilize features that exhibit greater commonality among subjects with diverse characteristics. For future work, this can be addressed by a combination of increasing the data subject size and applying deep learning techniques. With more subjects, more information can be captured from a diverse demography. Applying deep learning techniques, as well automatically extracting the most relevant features as opposed to manually designing these features, is heavily reliant on domain knowledge. We recognize that the performance based on the

time consumption and memory cost is also a key factor to evaluate, so we have a plan to incorporate it in our future work to provide a more comprehensive understanding of the computational efficiency of our methods.

Author Contributions: Conceptualization, S.U.R. and C.O.; Methodology, S.U.R. and C.O.; Software, S.U.R. and C.O.; Validation, S.U.R. and C.O.; Formal analysis, S.U.R., A.A., A.M.K. and C.O.; Investigation, S.U.R., A.A. and A.M.K.; Resources, S.U.R., A.A., A.M.K. and C.O.; Data curation, S.U.R. and C.O.; Writing—original draft preparation, S.U.R., A.A., A.M.K. and C.O.; Writing—review and editing, S.U.R., A.A. and A.M.K.; Visualization, S.U.R., A.A., A.M.K. and C.O.; Supervision, S.U.R.; Project administration, S.U.R. and A.A.; Funding acquisition, A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

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

References

1. Weiser, M. The computer for the 21st Century. *IEEE Pervasive Comput.* **2002**, *1*, 19–25. [CrossRef]
2. Jiang, S.; Shull, P.B.; Lv, B.; Sheng, X.; Zhang, C.; Wang, H. Development of a real-time hand gesture recognition wristband based on sEMG and IMU sensing. In Proceedings of the 16 IEEE International Conference on Robotics and Biomimetics (ROBIO), Qingdao, China, 3–7 December 2016. Available online: https://ieeexplore.ieee.org/document/7866498 (accessed on 8 December 2023).
3. Bulling, A.; Blanke, U.; Schiele, B. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Comput. Surv.* **2014**, *46*, 1–33. [CrossRef]
4. Bragança, H.; Colonna, J.G.; Oliveira, H.A.B.F.; Souto, E. How Validation Methodology Influences Human Activity Recognition Mobile Systems. *Sensors* **2022**, *22*, 2360. [CrossRef] [PubMed]
5. Wang, D.; Candinegara, E.; Hou, J.; Tan, A.-H.; Miao, C. Robust Human Activity Recognition Using Lesser Number of Wearable Sensors. In Proceedings of the 2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC), Shenzhen, China, 15–17 December 2017. Available online: https://ink.library.smu.edu.sg/sis_research/5468/ (accessed on 5 December 2023).
6. Xu, J.; Han, J.; Nie, F.; Li, X. Multi-view Scaling Support Vector Machines for Classification and Feature Selection. *IEEE Trans. Knowl. Data Eng.* **2019**, *32*, 1419–1430. [CrossRef]
7. Arlot, S.; Celisse, A. A survey of cross-validation procedures for model selection. *Stat. Surv.* **2010**, *4*, 40–79. [CrossRef]
8. Ferrari, A.; Micucci, D.; Mobilio, M.; Napoletano, P. Trends in human activity recognition using smartphones. *J. Reliab. Intell. Environ.* **2021**, *7*, 189–213. [CrossRef]
9. Little, M.A.; Varoquaux, G.; Saeb, S.; Lonini, L.; Jayaraman, A.; Mohr, D.C.; Kording, K.P. Using and understanding cross-validation strategies. Perspectives on Saeb et al. *GigaScience* **2017**, *6*, gix020. [CrossRef] [PubMed]
10. Hammerla, N.Y.; Plötz, T. Let's (not) stick together: Pairwise similarity biases cross-validation in activity recognition. In Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Osaka, Japan, 7–11 September 2015. [CrossRef]
11. Widhalm, P.; Leodolter, M.; Brändle, N. Top in the Lab, Flop in the Field? Evaluation of a sensor-based travel activity classifier with the SHL dataset. In Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, Singapore, 8–12 October 2018. [CrossRef]
12. Dehghani, A.; Glatard, T.; Shihab, E. Subject Cross Validation in Human Activity Recognition. *arXiv* **2019**. [CrossRef]
13. Tello, A.; Degeler, V.; Lazovik, A. Too Good to Be True: Performance overestimation in (re)current practices for Human Activity Recognition. *arXiv* **2023**. [CrossRef]
14. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
15. Yang, C.; An, Z.; Zhu, H.; Hu, X.; Zhang, K.; Xu, K.; Li, C.; Xu, Y. Gated convolutional networks with hybrid connectivity for image classification. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12581–12588.
16. Reiss, A. PAMAP2 Physical Activity Monitoring [Dataset]. *UCI Machine Learning Repository.* 2012. [CrossRef]
17. Caruso, C.; Quarta, F. Interpolation methods comparison. *Comput. Math. Appl.* **1998**, *35*, 109–126. [CrossRef]
18. Hunter, J.D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95. [CrossRef]
19. Gholamiangonabadi, D.; Kiselov, N.; Grolinger, K. Deep Neural Networks for Human Activity Recognition with Wearable Sensors: Leave-One-Subject-Out Cross-Validation for Model Selection. *IEEE Access* **2020**, *8*, 133982–133994. Available online: https://ieeexplore.ieee.org/abstract/document/9144538 (accessed on 5 December 2023). [CrossRef]
20. Ambati, L.S.; El-Gayar, O. Human Activity Recognition: A Comparison of Machine Learning Approaches. *J. Midwest Assoc. Inf. Syst. (JMWAIS)* **2021**, *2021*, 4. [CrossRef]

21. Andreu-Perez, J.; Leff, D.R.; Ip, H.M.D.; Yang, G.-Z. From Wearable Sensors to Smart Implants Toward Pervasive and Personalized Healthcare. *IEEE Trans. Biomed. Eng.* **2015**, *62*, 2750–2762. [CrossRef]
22. Asghari, P.; Nazerfard, E. Activity Recognition Using Hierarchical Hidden Markov Models on Streaming Sensor Data. In Proceedings of the 2018 9th International Symposium on Telecommunications (IST), Tehran, Iran, 17–19 December 2018. [CrossRef]
23. Baldominos, A.; Isasi, P.; Saez, Y. Feature selection for physical activity recognition using genetic algorithms. In Proceedings of the 2017 IEEE Congress on Evolutionary Computation (CEC), Donostia, Spain, 5–8 June 2017. Available online: https://ieeexplore.ieee.org/document/7969569 (accessed on 5 December 2023).
24. Waskom, M.; Botvinnik, O.; O'Kane, D.; Hobson, P.; Lukauskas, S.; Gemperline, D.C.; Augspurger, T.; Halchenko, Y.; Cole, J.B.; Warmenhoven, J.; et al. *Mwaskom/Seaborn*; v0.8.1. Zenodo: Brussel Belgium, 2017. Available online: https://doi.org/10.5281/zenodo.883859 (accessed on 5 December 2023).
25. Scikit-learn.org. 1. Supervised Learning—Scikit-Learn 0.21.3 Documentation. 2019. Available online: https://scikit-learn.org/stable/supervised_learning.html#supervised-learning (accessed on 5 December 2023).
26. Dargie, W. Analysis of Time and Frequency Domain Features of Accelerometer Measurements. In Proceedings of the 18th International Conference on Computer Communications and Networks, San Francisco, CA, USA, 3–6 August 2009. Available online: https://ieeexplore.ieee.org/document/5235366 (accessed on 5 December 2023).
27. Aguirre, P.L.; Torres, L.A.; Lemos, A.P. Autoregressive modeling of wrist attitude for feature enrichment in human activity recognition. In Proceedings of the Congresso Brasileiro de Inteligência Computacional, Niterói, RJ, Brazil, 30 October–1 November 2017.
28. Banos, O.; Galvez, J.-M.; Damas, M.; Pomares, H.; Rojas, I. Window Size Impact in Human Activity Recognition. *Sensors* **2014**, *14*, 6474–6499. [CrossRef]
29. Bock, M.; Hölzemann, A.; Moeller, M.; Laerhoven, K. Tutorial on Deep Learning for Human Activity Recognition. 2021. Available online: https://arxiv.org/pdf/2110.06663.pdf (accessed on 5 December 2023).
30. Module 4: Frequency Domain Signal Processing and Analysis. Available online: https://www.ccee.ncsu.edu/ccli-sensors/wp-content/uploads/sites/4/2015/09/Module-4.pdf (accessed on 5 December 2023).
31. Feng, Z.; Mo, L.; Li, M. A Random Forest-based Ensemble Method for Activity Recognition. In Proceedings of the 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milan, Italy, 25–29 August 2015. [CrossRef]
32. Lara, O.D.; Labrador, M.A. A Survey on Human Activity Recognition using Wearable Sensors. *IEEE Commun. Surv. Tutor.* **2013**, *15*, 1192–1209. [CrossRef]
33. Gupta, A.; Gupta, K.; Gupta, K.; Gupta, K. A Survey on Human Activity Recognition and Classification. In Proceedings of the 2020 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 28–30 July 2020. Available online: https://ieeexplore.ieee.org/document/9182416 (accessed on 5 December 2023).
34. Zdravevski, E.; Lameski, P.; Trajkovik, V.; Kulakov, A.; Chorbev, I.; Goleva, R.; Pombo, N.; Garcia, N. Improving Activity Recognition Accuracy in Ambient-Assisted Living Systems by Automated Feature Engineering. *IEEE Access* **2017**, *5*, 5262–5280. [CrossRef]
35. Jain, A.; Kanhangad, V. Human Activity Classification in Smartphones Using Accelerometer and Gyroscope Sensors. *IEEE Sens. J.* **2018**, *18*, 1169–1177. [CrossRef]
36. Li, K.; Habre, R.; Deng, H.; Urman, R.; Morrison, J.; Gilliland, F.D.; Ambite, J.L.; Stripelis, D.; Chiang, Y.Y.; Lin, Y. Applying multivariate segmentation methods to human activity recognition from wearable sensors' data. *JMIR Mhealth Uhealth* **2019**, *7*, e11201. [CrossRef]
37. Quigley, B.; Donnelly, M.; Moore, G.; Galway, L. A Comparative Analysis of Windowing Approaches in Dense Sensing Environments. *Proceedings* **2018**, *2*, 1245. [CrossRef]
38. Ferrari, A.; Micucci, D.; Mobilio, M.; Napoletano, P. On the Personalization of Classification Models for Human Activity Recognition. *IEEE Access* **2020**, *8*, 32066–32079. Available online: https://ieeexplore.ieee.org/abstract/document/8995531 (accessed on 5 December 2023). [CrossRef]
39. Sztyler, T.; Stuckenschmidt, H.; Petrich, W. Position-aware activity recognition with wearable devices. *Pervasive Mob. Comput.* **2017**, *38*, 281–295. [CrossRef]
40. Lane, N.D.; Xu, Y.; Lu, H.; Hu, S.; Choudhury, T.; Campbell, A.T.; Zhao, F. Enabling large-scale human activity inference on smartphones using community similarity networks (csn). In Proceedings of the 13th International Conference on Ubiquitous Computing, Beijing China, 17–21 September 2011. [CrossRef]
41. Mannini, A.; Sabatini, A.M. Machine Learning Methods for Classifying Human Physical Activity from On-Body Accelerometers. *Sensors* **2010**, *10*, 1154–1175. [CrossRef]
42. McKhann, G.; Drachman, D.; Folstein, M.; Katzman, R.; Price, D.; Stadlan, E.M. Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group* under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology* **2010**, *34*, 939. [CrossRef]
43. Micucci, D.; Mobilio, M.; Napoletano, P. UniMiB SHAR: A Dataset for Human Activity Recognition Using Acceleration Data from Smartphones. *Appl. Sci.* **2017**, *7*, 1101. [CrossRef]
44. Nweke, H.F.; Teh, Y.W.; Al-garadi, M.A.; Alo, U.R. Deep Learning Algorithms for Human Activity Recognition Using Mobile and Wearable Sensor networks: State of the Art and Research Challenges. *Expert Syst. Appl.* **2018**, *105*, 233–261. [CrossRef]

45. Lara, O.D.; Pérez, A.J.; Labrador, M.A.; Posada, J.D. Centinela: A human activity recognition system based on acceleration and vital sign data. *Pervasive Mob. Comput.* **2012**, *8*, 717–729. [CrossRef]
46. Perez, A.J.; Zeadally, S. Recent Advances in Wearable Sensing Technologies. *Sensors* **2021**, *21*, 6828. [CrossRef]
47. Reyes-Ortiz, J.L.; Anguita, D.; Ghio, A.; Parra, X. Human Activity Recognition Using Smartphones Dataset. 2012. Available online: https://archive.ics.uci.edu/dataset/240/human+activity+recognition+using+smartphones (accessed on 5 December 2023).
48. Reyes-Ortiz, J.L.; Oneto, L.; Samà, A.; Parra, X.; Anguita, D. Transition-Aware Human Activity Recognition Using Smartphones. *Neurocomputing* **2016**, *171*, 754–767. [CrossRef]
49. Saez, Y.; Baldominos, A.; Isasi, P. A Comparison Study of Classifier Algorithms for Cross- Person Physical Activity Recognition. *Sensors* **2016**, *17*, 66. [CrossRef] [PubMed]
50. Tapia, E.M.; Intille, S.S.; Haskell, W.; Larson, K.; Wright, J.; King, A.; Friedman, R. Real-Time Recognition of Physical Activities and Their Intensities Using Wireless Accelerometers and a Heart Rate Monitor. In Proceedings of the 2007 11th IEEE International Symposium on Wearable Computers, Boston, MA, USA, 11–13 October 2007. [CrossRef]