Enhancing HAR Novelty Detection with Activity Confusion Analysis and Clustering

Thaís B. de M. B. de Sousa[1], Lívia A. Cruz[1], Criston P. de Souza[1], Regis P. Magalhães[1], José A. F. de Macêdo[1]

[1]Universidade Federal do Ceará (UFC)

criston@ufc.br,

{thaisbezerra,jose.macedo,livia,regis}@insightlab.ufc.br

Abstract. *The field of medicine and healthcare, especially Human Activity Recognition (HAR), has been experiencing the benefits of technology for tracking people's habits. In this context, novelty detection aims to detect if an activity performed by a subject hasn't been seen before, in the past training data. However, the novelty may not be properly recognized if there is a similar action in the train set. Thus, this work proposes an analysis of activity pairs to determine which pairs are most confused when using three traditional models for novelty detection: Local Outlier Factor, One-class SVM, and Isolation Forest. After that, we employ agglomerative clustering to gather similar activities, and the clustered activities are used as input to a leave-one-activity-out novelty detection approach. We conclude that the Isolation Forest model achieves the best results in the activity pairs analysis, yielding an F1 score of 88.4%. Finally, the leave-one-activity-out methodology is evaluated with a sliding window technique, and the F1 score on the Local Outlier Factor model increases from 66.3% (without grouping similar activities) to 74.6% (grouping similar activities).*

1. Introduction

Human Activity Recognition (HAR) consists of understanding people's daily behavior by analyzing people themselves or their surrounding environment [Das Antar et al. 2019]. Many approaches for detecting human activity are based on different input sources, such as video, smart home sensors, wearables, and smartphones [Das Antar et al. 2019]. These technologies can monitor physical activities, assess a worker's health in the workplace, or recommend minor behavior interventions to help individuals maintain a healthy lifestyle.

HAR approaches typically focus on supervised classification based on predefined activities; however, it is essential to identify new patterns of unknown activities due to the diverse nature of human habits [Kim and Lee 2024]. In this context, novelty detection refers to determining whether a data point observed during the test differs significantly from the data seen during training [Pimentel et al. 2014]. The novelty class is not available during training; this problem can also be known as One-Class Classification, Outlier Detection, Anomaly Detection [Pimentel et al. 2014], or even Out-of-Distribution (OOD) Detection [Yang et al. 2024]. In particular, anomaly detection often focuses on identifying "abnormal" or "erroneous" data, while novelty detection does not establish a definition of "normal" but instead aims to detect new classes in a semantic context [Yang et al. 2024]. Additionally, the OOD problem is related to Open Set Classification,

where the goal is for machine learning models to identify classes or data that were not observed during training [Boyer et al. 2021]. This work focuses on novelty detection within a framework more closely related to Out-of-Distribution (OOD) Detection and Open-Set Classification.

In novelty detection, the training set may occasionally contain activities that are slightly similar to the novelty, leading to misclassification (that is, a real novelty is incorrectly designed as non-novelty). In fact, several studies have addressed this issue of confusing classes [Boyer et al. 2021], also referred to as interclass similarity [Jin et al. 2022]. This work proposes a preliminary analysis of activity pairs to assess which classes are most confused, using three classical novelty detection models (Local Outlier Factor, One-Class SVM, and Isolation Forest), in the context of wearable human activity data. After this, similar activities are aggregated using hierarchical clustering; then, we evaluate novelty detection using a "Leave-one-activity-out" approach, i.e., training with all activities except the novelty. In addition, we compare different preprocessing choices, varying the segment sizes, the set of features, and the segmentation strategy. Thus, in this work, we contribute with a methodology adaptable to various datasets and models, and the resulting activity groups can be leveraged to train not only novelty detection models but also HAR classification models.

The remainder of this paper is structured as follows: Section 2 reviews related works; Section 3 outlines data preprocessing, datasets, and methodology; Section 4 presents the experimental results; and Section 5 provides conclusions and future work.

2. Related work

Similarly to our work, other studies have proposed methods applied to data collected by wearable devices. In the HAR context, [Xu et al. 2019] utilize the Random Forest classifier for elderly home activities, with the use of wrist accelerometer data from the HMP dataset. They provide an activity similarity matrix based on the confusion matrix and propose a way to correct misclassification based on the location and time of the activity. [Staab et al. 2022] utilize smartwatch data to detect out-of-distribution (OOD) activities and perform human activity recognition (HAR). The study focuses on collecting data from four specific activities — writing, drinking, eating, and blowing the nose — and highlights their similarities, which make classification more challenging. Some pipelines were tested, including LSTMs and traditional machine learning algorithms such as Random Forest. Ultimately, the best configuration involved using only an LSTM model: first classifying data as OOD or not, and then assigning the correct label to recognized activities. However, the study did not explore any novelty detection models.

[Jin et al. 2022] utilize deep learning to solve three different problems: closed-set, pseudo-open-set, and completely open-set classification on human activity; three datasets are employed, including PAMAP2. They introduce a new loss function and a rejection score to discriminate the classes. On the open-set task, they employ an approach very similar to our "leave-one-activity-out"; they conclude that their deep learning algorithm outperforms classic models like One-Class SVM and Isolation Forest. [Munoz-Organero 2019] provide an algorithm based on Deep Recurrent Networks (DRNN) for outlier detection and posterior sub-activity classification, using RealWorld(HAR) and Opportunity datasets. The approach first utilizes outlier removal to

cleanse the training data, then evaluates classification to identify an activity inside another (such as "walking" segments inside a "climbing stairs" activity). Some different evaluation techniques are employed, including intra-subject and inter-subject (similar to leave-one-subject-out). However, the work focuses on identifying "walking" and could not have explored a more diverse set of activities.

[Boyer et al. 2021] compare traditional methods with image domain methods to address the challenge of out-of-distribution (OOD) detection in data from inertial sensors. They utilize MHEALTH and SPARS as data sources, and collect a new dataset as well. To mitigate the confusion among classes, they group them in a way that does not compromise OOD detection and arbitrarily select certain fixed classes to designate as OOD. They also experimented with a class removal strategy to determine whether eliminating or redesignating confusing classes could improve OOD performance. Their results indicate that traditional algorithms outperformed deep learning methods in terms of OOD detection.

Unlike [Munoz-Organero 2019], this work does not seek to refine the training classes by removing outliers and ambiguous instances. Rather, in line with the approach proposed by [Boyer et al. 2021], we explore whether clustering similar classes can improve the performance of traditional methods — which, as reported by [Jin et al. 2022], currently achieve suboptimal results.

3. Data and methods

In our approach, similar activities are grouped and evaluated together to improve the model's ability to detect novelty. The model evaluation consists of a leave-one-activity-out approach, where each activity is considered a novelty once, and the others are considered non-novelties. Since each subject has a peculiar way of performing activities, we evaluate the models for each subject separately.

This section outlines the following: first, input preprocessing and the datasets used; next, the methodology for grouping similar activities for each subject and the leave-one-activity-out evaluation; and finally, the novelty detection models evaluated in the experiments.

3.1. Input preprocessing

The problem input is a multivariate time series with data collected from Inertial Measurement Units (IMUs), such as smartwatches. They typically have the following sensors: accelerometer, gyroscope, magnetometer, and heart rate. Some datasets are collected with sensors on different body parts, such as the wrist, chest, or ankle. We focus on wrist sensors, as they are commonly used, but we also include one evaluation with all sensors for the best model.

The time series is divided into non-overlapping fixed-size segments, that must be classified as novelty or non-novelty. Each segment has a label indicating the user's activity during that segment's time interval. We also evaluate a sliding window approach for the best model and hyperparameters, on the leave-one-activity-out strategy. We assume the user performs a single activity at each segment, and therefore, the segments with more than one activity are removed from the input.

Code	Activity	Code	Activity
1	Walking	10	Eating Pasta
2	Jogging	11	Drinking from Cup
3	Stairs	12	Eating Sandwich
4	Sitting	13	Kicking (Soccer Ball)
5	Standing	14	Playing Catch...
6	Typing	15	Dribbling (Basketball)
7	Brushing Teeth	16	Writing
8	Eating Soup	17	Clapping
9	Eating Chips	18	Folding Clothes

(a) WISDM Activities code

Code	Activity	Code	Activity
1	Nordic walking	7	Rope jumping
2	Ascending stairs	8	Running
3	Cycling	9	Sitting
4	Descending stairs	10	Standing
5	Ironing	11	Vacuum cleaning
6	Lying	12	Walking

(b) PAMAP2 Activities code

Table 1. Activities codes.

We extract time series features from the segments with the Catch22 toolbox [Lubba et al. 2019], which provides a comprehensive set of interpretable and efficient time series descriptors. It generates, for each original feature, 22 attributes divided into five categories: distribution, autocorrelation, run length, value, and frequency domain. The time component is ignored after this process, so the data becomes tabular. We additionally fill the null values using the mean; this is necessary for PAMAP2 dataset, as will be further explained.

3.2. Datasets

The experiments utilize two public datasets, PAMAP2 [Reiss and Stricker 2012] and WISDM [Weiss 2019], which are widely adopted in the HAR context.

PAMAP2. This dataset is labeled for the problem of human activity detection, which we adapt here for the novelty detection problem. Data were collected from nine volunteers who had performed 12 kinds of activities, described in table 1. The dataset has as labels the activity the subject is executing at each point. Each subject used 3 IMUs: one on the dominant side's wrist, one on the chest, and one on the dominant side's ankle. Each IMU has a sampling rate of 9 Hz for the heart rate and 100 Hz for the other sensors. The imputation of the mean heart rate value in the preprocessing compensates for the lower heart rate sensor sample rate. To facilitate the comparison with the WISDM dataset, we consider only the accelerometer and gyroscope attached to the hand. Additionally, an evaluation using all sensors is also presented for the best model and hyperparameters.

WISDM. This dataset is useful for biometric models and activity recognition. 51 users performed 18 activities of daily living, present in table 1, with data recorded by accelerometer and gyroscope sensors from a smartwatch and a smartphone. The activities can be divided into non-hand-oriented and hand-oriented(general and eating). Both devices have a 20 Hz sample rate, and each row is labeled with the corresponding activity. In this work, we use only the smartwatch data to make a fair comparison with PAMAP2 hand's sensors. Furthermore, we remove time gaps longer than 10 seconds between activity executions.

3.3. Grouping of similar activities

This step aims to reduce model misclassification of confusing pairs of activities. The models are evaluated considering a pair of activities, one of them is the known activity, and

the other is the novelty one. Thus, one model is trained for each pair of activities and each distinct user. The segments of other activities and users were utilized for hyperparameter optimization through grid search, with the F1 score as the optimization metric. Finally, we average the results of each pair of activities, weighted by the number of segments in the test sets.

This process is detailed below in three steps: i) how a pair of activities of a subject is evaluated, ii) how the hyperparameters are optimized, and iii) how results are employed to cluster activities among all subjects.

i) Evaluation of activities pairs by subject. For each subject S and each pair of activities A and B of S, we evaluate a model considering some segments of A as the 'novelty' class and some segments of B as the 'non-novelty' class. Let $n = \min(|A|, |B|/3)$, where $|A|$ and $|B|$ are the number of segments of A and B, respectively. The test set comprises n segments drawn from A and n segments drawn from B. This way, the test set has the same number of segments of 'novelty' and 'non-novelty' classes. The train set is composed of the remaining segments of B, so it has at least the same number of segments as the test set. Indeed, if $n = |A|$, the test size is $2|A|$, and the train size is $|B| - |A|$, which is bigger, because we know that $|A| < |B|/3$. On the other case, the test and the train sizes are both equal to $2|B|/3$. As expected in novelty detection, the train set has only 'non-novelty' segments. The segments in the train set are considered the known 'non-novelties' and are used to construct the distribution of 'normal' activities. Based on this distribution, the segments in the test set are classified as 'novelty' or 'non-novelty,' and the macro F1 metric is calculated for the tuple (S, A, B).

ii) Hyperparameters optimization. The best combination of hyperparameters for subject S and activities A, B is determined considering all subjects except S and the average of the F1 weighted by the sizes of the test sets. The F1 of (S, A, B) is obtained using the best hyperparameters for (S, A, B). After performing the grid search, the final model is trained again using the best hyperparameters for each (S, A, B). For each pair of activities A and B, we calculate the average of the macro F1 across all the subjects using the final model, weighted by the number of segments evaluated.

iii) Activities clustering. We detect similar activities based on their overall performance of models on pair activities, assuming the most similar activities are those with the worst results. Thus, we cluster similar activities based on the F1 score. Note that the lower the F1, the "closer" the activities are, meaning that if one activity is wrongly classified as a non-novelty, it may be difficult to recognize it as a novelty, and vice versa. We use the agglomerative clustering algorithm [Lukasová 1979] with a linkage strategy that combines clusters based on the average distance between points in the clusters. The distance between two activities A and B is the harmonic mean of the F1 scores obtained from models that were trained with activity A as a non-novelty and B as a novelty, and vice versa.

3.4. Leave-one-activity-out evaluation

Now we evaluate the detection of novelty models on grouped activities. We retrain the models, employing a leave-one-activity-out approach. For each activity A and subject S, we create a test set in which the segments of A are labeled as 'novelty', while a sample of segments from the other activities of S is drawn and labeled as 'non-novelty'. The number

of segments drawn from the other activities is the same as the number of segments of A.

The training set is composed of the remaining segments of S. The segments in the train set are considered the known 'non-novelties' and are used to construct the distribution of 'normal' activities. Based on this distribution, the segments in the test set are classified as 'novelty' or 'non-novelty', and the metric F1 score is calculated. The F1 score for the subject S is the average of the F1 of its activities, weighted by the number of segments of each activity. For the hyperparameter optimization, we use a grid search. We evaluate the F1 of each subject and average the results, weighted by the number of segments of each subject.

3.5. Novelty detection models

The novelty detection models listed below are evaluated in this work. We focus on simple strategies, thus excluding those based on deep learning.

The trivial baseline is a theoretical reference for the average F1 score of a random classifier and is not implemented. All other models have an implementation available in the Scikit-learn library [Pedregosa et al. 2011]. The hyperparameters tested for each model are listed in Table 2.

Trivial baseline. Consider a random classifier with independent draws where each segment is classified as a 'novelty' with probability p, and is classified as 'non-novelty' with probability $1 - p$. Recall that the test set has the same number of 'novelty' and 'non-novelty' segments. In this case, the precision of the trivial baseline is 0.5 (assuming $0 < p < 1$) for both the 'novelty' and 'non-novelty' classes. On the other hand, the recall is p for the 'novelty' class and $1 - p$ for the 'non-novelty' class. Therefore, the F1 score is $p/(0.5+p)$ for the 'novelty' class and $(1-p)/(0.5+1-p)$ for the 'non-novelty' class. Thus, the average F1 score is $(p/(0.5 + p) + (1 - p)/(0.5 + 1 - p))/2$ with maximum value 0.5 when $p = 0.5$.

Local Outlier Factor. The Local Outlier Factor (LOF) [Breunig et al. 2000] is an unsupervised anomaly detection algorithm that computes the local density deviation of a data point with respect to its neighbors.

One-class Support Vector Machine The One-class Support Vector Machine (OC-SVM) [Schölkopf et al. 2001] is a method for novelty detection that learns a decision boundary around the data points. Given an underlying probability distribution P in the feature space, the method tries to estimate a region S such that the probability of a point being outside S is less than a user-defined threshold $0 \leq \texttt{nu} < 1$. This is performed by estimating a function f such that f is positive for points in S and negative for points outside S. The form of f is a kernel expansion of a small subset of the training data, and it is regularized by the length of the weight vector. To keep almost the same execution time for both datasets, the set of possible hyperparameter values is reduced for WISDM since it has more subjects. For WISDM, the values for `nu` range from 0.1 to 1.0 with a step of 0.1, the kernel options remain the same, and the polynomial degree consists of integer values ranging from 1 to 4.

Isolation Forest. The Isolation Forest (IF) [Liu et al. 2008] consists of selecting a random feature and a random split value between the maximum and minimum values of the feature. The process is repeated recursively until all points are isolated, producing a tree.

Model	Hyperparameters
LOF	n_neighbors={1,2,3,4,5} ; novelty = true
OC-SVM	nu = {0.01, 0.02, ... 1.0} ; kernel = {linear, poly, rbf, sigmoid} ; degree = {1,2,3,4,5}
IF	n_estimators = {100,200,300,400 } ; random_state = 0

Table 2. Hyperparameters optimization

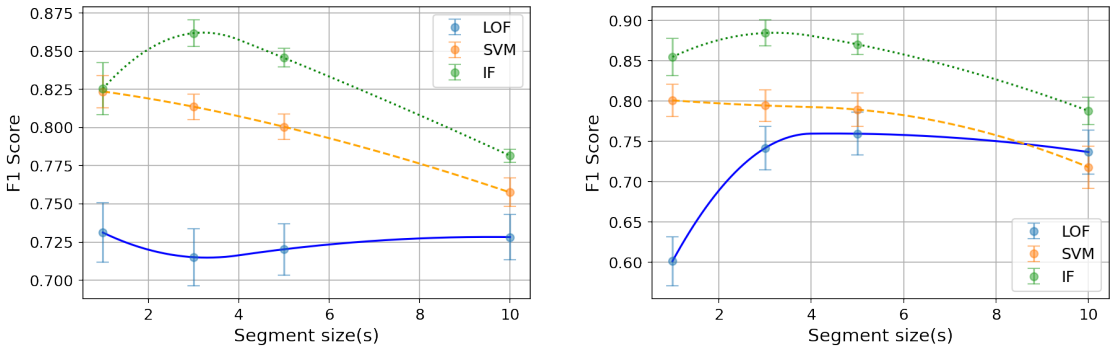
Many trees are built, and the average path length to isolate a point is used as a measure of the point's anomaly.

4. Experimental results

This section discusses the results of our experimental evaluation. Section 4.1 analyzes how segment size impacts performance, Section 4.2 shows results for grouping similar activities, and Section 4.3 compares models using the leave-one-activity-out method.

4.1. Segment size selection

We calculate the average F1 score for each model across different datasets and segment sizes (1s, 3s, 5s, and 10s). Figure 1a presents the results. The error bars indicate 95% confidence interval. Note that for both datasets, the IF model obtained the highest average F1 score for all tested segment sizes.



(a) WISDM dataset. **(b) PAMAP2 dataset.**

Figure 1. Average F1 score of each model, varying segment size.

For WISDM (Figure 1a), the segment size variation does not produce significant changes on average F1 for the LOF model. Nevertheless, for OC-SVM, we notice an average F1 decrease as we increase the segment size; while in IF the average F1 reaches a maximum at 3s segment size. Statistical tests were conducted to check significant differences, and all statistical tests in this work were performed assuming a 95% confidence level. A Friedman test was conducted to compare all 12 model and segment size combinations, indicating a statistically significant difference between at least two of the data groups. After this, we performed the Dunn test to check an average F1 score significant difference among segment sizes for IF, concluding that the only pair with no significant statistical difference is the (1s, 5s) pair (p-value of 0.527). Besides, since the 3s segment size is different from all others and presents the highest average F1, we conclude that the 3s segment size is the best for the IF model. We also performed the Dunn test to check an

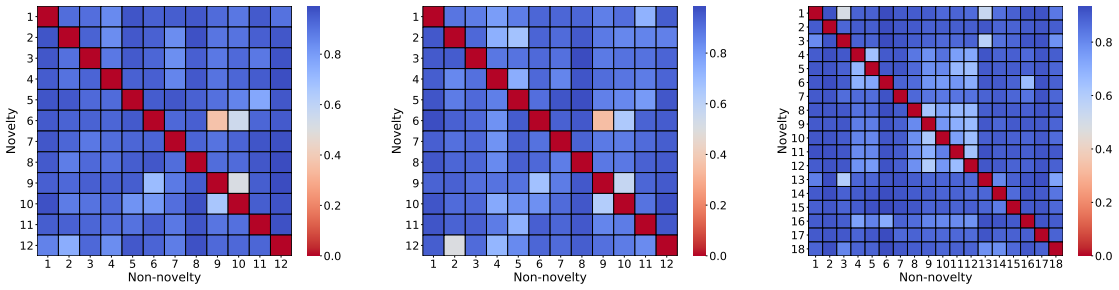


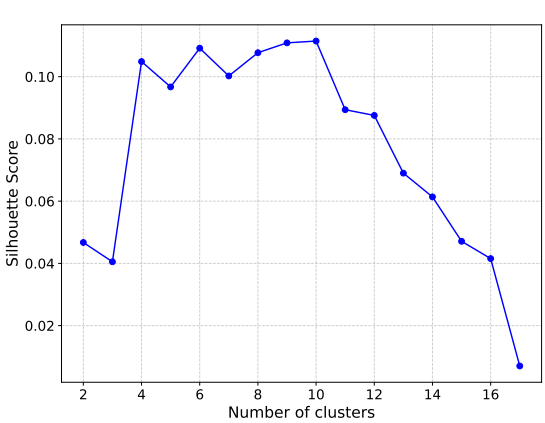
Figure 2. Activity pairs F1 heatmaps, IF, 3s. (a) PAMAP2-All features ; (b) PAMAP2-Acc/gyr; (c) WISDM-Acc/gyr.

average F1 score significant difference among the models for 3s segment size, concluding that all three models are different from each other. As a result, the IF model with a 3s segment size is the best configuration for the WISDM dataset.

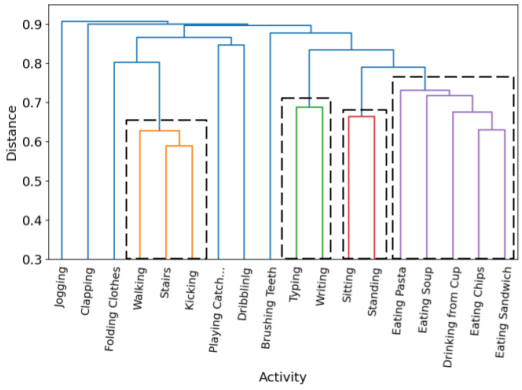
For PAMAP2 (Figure 1b), analyzing the segment size effect on each model, we conclude that LOF shows a peak in 5s segment size. Meanwhile, OC-SVM presents analogous values for 1s to 5s and a big decrease in 10s. The same pattern occurs for the IF model. A Friedman test revealed a statistically significant difference between at least two groups of data. For IF, the Dunn test showed that the 10s segment size is the only one with a p-value less than 0.05, but there is no significant distinction among the 1s, 3s, and 5s segment sizes. Applying the Dunn test to check the models' effect on the 3s segment size, we conclude that the average F1 score of IF is distinct from the two others, allowing us to conclude that the IF is the best model with statistical significance. Moreover, since the PAMAP2 dataset has a larger set of features than WISDM (not only the accelerometer and gyroscope of hand), we took the best model (IF) and segment size (3s) combination to analyze its results when using all available features. Using all features, we obtain an average F1 score of 91.2%; and using only the accelerometer and gyroscope of the hand, we obtain 88.4%, and this difference has statistical significance. Therefore, we conclude that employing a larger number of features improves the model's performance.

4.2. Grouping similar activities

The identification of similar activity pairs is based on the average F1 score obtained when considering one of the activities as a novelty and the other as a non-novelty, i.e., the lower the average F1 score, the more similar the activities are. Figure 2 shows the heatmaps of the average F1 score obtained by the IF with a segment size of 3s for (a) PAMAP2 using all features, (b) PAMAP2 using only the accelerometer and gyroscope of hand, and (c) WISDM using only the accelerometer and gyroscope of hand (i.e., all features available in WISDM). The IF model with 3s segments was chosen as it is the best configuration for both datasets, as discussed in Section 4.1. The bluer cells indicate a higher average F1 score for the activity pair, while the redder cells indicate a lower average F1 score. The activities represented in the heatmap are coded according to Table 1. For example, in Figure 2a, the pair of activities 6 and 9 (lying and sitting) present a low F1 score (0.36), indicating that these activities are similar. One possible explanation is that both are sedentary activities. In Figure 2c, there is a square of low F1 scores from activities 9 to 12, which are all eating and drinking activities; so it is harder to distinguish them.

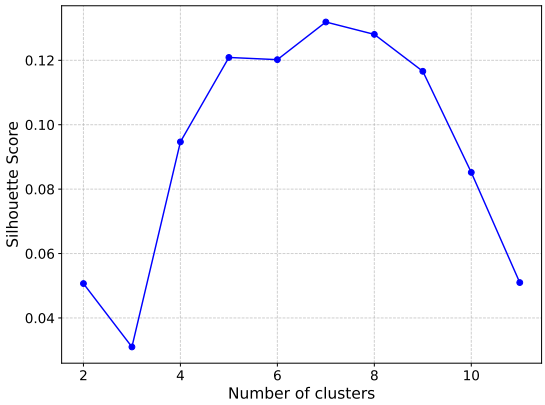


(a) Silhouette score vs number of clusters.

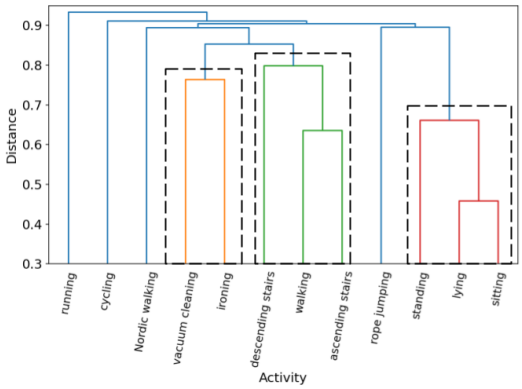


(b) Agglomerative clustering: dashed rectangles indicate activity grouping.

Figure 3. Grouping of similar activities in WISDM dataset (10 clusters).



(a) Silhouette score vs number of clusters.



(b) Agglomerative clustering: dashed rectangles indicate activity grouping.

Figure 4. Grouping of similar activities in PAMAP2 dataset (7 clusters).

Figure 3a shows the silhouette score as a function of the number of clusters for the WISDM dataset, achieving the maximum silhouette score with 10 clusters. Using the agglomerative clustering algorithm, we group the activities into 10 clusters, as shown in Figure 3b. The same analysis was performed for the PAMAP2 dataset, where the maximum silhouette score is achieved with 7 clusters, as shown in Figure 4a. Figure 4b shows the resulting activity groups for the PAMAP2 dataset. In both datasets, we observe that some light activities and walking-related ones were grouped, whereas the more intense ones were kept separated, indicating that it is easier to detect novelties in more intense physical activities.

4.3. Leave-one-activity-out evaluation

After grouping similar activities as described in Section 4.2, we evaluate the models using the leave-one-activity-out approach. Table 3 shows the average F1 score of each model and dataset, for both the original activities and after grouping similar activities using the optimal number of clusters. All models have an increase in the average F1 score after grouping similar activities. These differences have statistical significance according to

Dataset	Model	Original	Grouped
PAMAP2	IF	53.2	**66.6**
	LOF	62.0	63.8
	OC-SVM	52.5	58.0
WISDM	IF	46.0	55.3
	LOF	56.1	**61.8**
	OC-SVM	52.1	53.0

Table 3. Leave-one-activity-out F1 scores (%) for the original activities and after grouping similar activities.

the Wilcoxon test, except the LOF for the PAMAP2 dataset (p-value = 0.19).

For the PAMAP2 dataset with the grouping of similar activities, the Friedman test refused the hypothesis of equal performance among the models, and Dunn's test showed that only the OC-SVM model is different from the other models (lower average F1 score). Therefore, the IF and LOF models are the best models for the PAMAP2 dataset. By employing all features of the PAMAP2 dataset and grouping similar activities, the IF model achieved an average F1 score of 68.1%. On the other hand, by using only the accelerometer and gyroscope of the hand, the average F1 score was 66.6%, but this difference is not statistically significant (p-value = 0.098 on Wilcoxon's test).

For the WISDM dataset with the grouping of similar activities, the Friedman test refused the hypothesis of equal performance among the models, and Dunn's test showed that all models are different from each other. Therefore, the LOF model is the best model for the WISDM dataset.

In general, the leave-one-activity-out evaluation produced significantly worse results when compared to the results for activity pairs (Section 4.2). This suggests that even after grouping similar activities, the tested models struggled to effectively recognize novelties when all other activities were mixed.

As a final experiment, Tables 4a and 4b show the average F1 scores of the best configuration (LOF model and 3s segment size) for two time-series segmentation strategies (disjoint intervals and sliding window) and grouping or not grouping similar activities. The sliding window strategy produced better results for both the original and grouped activities. In fact, for the WISDM dataset, a Friedman test followed by a Dunn's test indicated that all configurations on Table 4b are different from each other, which means that grouping and applying the sliding window is the best choice. On PAMAP2, the same test revealed that on both segmentation types, there is no statistical difference in grouping or not grouping the activities.

Grouping \ Sliding window	No	Yes
No	62.0	73.4
Yes	63.8	73.7

(a) PAMAP2

Grouping \ Sliding window	No	Yes
No	56.1	66.3
Yes	61.8	74.6

(b) WISDM

Table 4. LOF model, F1 scores (%) for LOAO evaluation, varying time-series segmentation strategies and grouping or not similar activities.

5. Conclusions

The study evaluated three novelty detection models (LOF, OC-SVM, and IF) to recognize new activities in a dataset, including an activity pair analysis. The IF model showed the best performance on both datasets, achieving an F1 score of 88.4% with 3-second segments on PAMAP2 using only accelerometer and gyroscope data from the hand. When all features were used, the F1 score increased to 91.2%, indicating a significant difference. It is concluded that the IF model is effective for novelty detection when considering only two activities, one known and one novel.

Agglomerative clustering was conducted to group similar activities based on the F1 score obtained from the activity pair analysis. The results indicated that sedentary activities, such as standing and lying down, and meal-related activities form clusters. In contrast, more intense activities, like running and cycling, did not form clusters, indicating that they are more easily recognized as novel activities. Applying the leave-one-activity-out approach and grouping similar activities improved all models' performance. For PAMAP2, IF and LOF performed best, with F1 scores of 66.6% and 63.8%, respectively. For WISDM, LOF was the top model, achieving 61.8% F1. An experiment using sliding window segmentation further improved results, raising WISDM's LOF performance to 74.6%. Despite higher processing costs, the sliding window approach proved beneficial. The leave-one-activity-out approach resulted in significantly lower metrics than the activity pair analysis, as mixing multiple activities creates a more realistic yet challenging scenario. Even with grouped activities, some confusion likely remains among certain groups. This may be due to the clustering strategy, which did not account for individual differences when defining the clusters.

Future work could focus on selecting a subset of activities to be treated as novel, rather than rotating all available classes as novelty cases. Additionally, evaluating sliding windows with varying overlap degrees could further improve performance.

6. Acknowledgments

Part of the results presented in this work were obtained through the project "CENTER OF EXCELLENCE IN ARTIFICIAL INTELLIGENCE – AI4WELLNESS", funded by Samsung Eletronica da Amazônia Ltda., under the Information Technology Law No. 8.248/91.

References

Boyer, P., Burns, D., and Whyne, C. (2021). Out-of-distribution detection of human activity recognition with smartwatch inertial sensors. *Sensors*, 21(5):1669.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104.

Das Antar, A., Ahmed, M., and Ahad, M. A. R. (2019). Challenges in sensor-based human activity recognition and a comparative analysis of benchmark datasets: A review. In *2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, pages 134–139.

Jin, L., Wang, X., Chu, J., and He, M. (2022). Human activity recognition machine with an anchor-based loss function. *IEEE Sensors Journal*, 22(1):741–756.

Kim, H. and Lee, D. (2024). Clan: A contrastive learning based novelty detection framework for human activity recognition.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). Isolation forest. In *IEEE International Conference on Data Mining*, pages 413–422. IEEE.

Lubba, C. H., Sethi, S. S., Knaute, P., Schultz, S. R., Fulcher, B. D., and Jones, N. S. (2019). catch22: Canonical time-series characteristics: Selected through highly comparative time-series analysis. *Data Mining and Knowledge Discovery*, 33(6):1821–1852.

Lukasová, A. (1979). Hierarchical agglomerative clustering procedure. *Pattern Recognition*, 11(5):365–381.

Munoz-Organero, M. (2019). Outlier detection in wearable sensor data for human activity recognition (har) based on drnns. *IEEE Access*, 7:74422–74436.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pimentel, M. A., Clifton, D. A., Clifton, L., and Tarassenko, L. (2014). A review of novelty detection. *Signal Processing*, 99:215–249.

Reiss, A. and Stricker, D. (2012). Introducing a new benchmarked dataset for activity monitoring. *2012 16th International Symposium on Wearable Computers*, pages 108–109.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471.

Staab, S., Krissel, S., Luderschmidt, J., and Martin, L. (2022). Recognition models for distribution and out-of-distribution of human activities. In *2022 18th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, pages 6–12. IEEE.

Weiss, G. M. (2019). WISDM smartphone and smartwatch activity and biometrics dataset. *UCI Machine Learning Repository: WISDM Smartphone and Smartwatch Activity and Biometrics Dataset Data Set*, 7:133190–133202.

Xu, H., Pan, Y., Li, J., Nie, L., and Xu, X. (2019). Activity recognition method for home-based elderly care service based on random forest and activity similarity. *IEEE Access*, 7:16217–16225.

Yang, J., Zhou, K., Li, Y., and Liu, Z. (2024). Generalized out-of-distribution detection: A survey.