

From Motion to Notion: Teaching a Machine to Recognize Your Activities

Clara Wong
McMaster University
Hamilton, Canada
wongc148@mcmaster.ca

Abstract—This document presents a real-time human activity recognition (HAR) system executed on a Raspberry Pi, collecting accelerometer data. During preprocessing, nine statistical features were extracted from sliding windows, and a support vector machine (SVM) classifier was trained. The model was able to successfully classify sitting, walking, running, and turning with high accuracy. Visualization plots, such as PCA and feature distribution, corroborate the strong separability between the activity classes. Future work includes incorporating data from other sensors, collecting a larger and more diverse training dataset, and exploring advanced models to improve accuracy and generalizability.

I. INTRODUCTION

Cyber-Physical Systems (CPS) integrate sensing, computation, and actuation in a closed feedback loop to enable real-time interaction between the real-world and computational network. CPS technologies are crucial in many biomedical applications where continuous physiological sensing coupled with real-time decision-making is necessary, such as wearable health monitors, rehabilitation devices, and fall-protection systems [1]. Human Activity Recognition (HAR) is the process of using sensor data from devices and machine learning algorithms to detect and classify human activities [2]. This project explores the design, implementation, and validation of a comprehensive CPS for real-time HAR using a Raspberry Pi and the Sense HAT's onboard accelerometer. The result was an end-to-end HAR pipeline, starting with data acquisition, followed by model training and evaluation, and culminating in deployment on a Raspberry Pi with LED-based activity feedback. The challenges associated with real-time classification on a computationally limited device like the Raspberry Pi were also explored.

II. METHODOLOGY

A. Data Collection and Pre-Processing

Four activities were classified for this project: sitting, walking, turning clockwise, and running. To achieve this, accelerometer data was collected with the Raspberry Pi Sense HAT v2. It utilizes the STMicro LSM9DS1 chip, an inertial measurement unit (IMU) that includes an accelerometer. The IMU has a 16 g acceleration measurement range with a 16-bit resolution [3]. To collect data, the Raspberry Pi was placed at waist level; studies found that a waist-mounted accelerometer best represented human motion, and several studies obtained an accuracy score exceeding 90% [4]. Sampling frequency

is also an important consideration. Karantonis et al. reported that all measured body measurements are contained within frequency components below 20 Hz [5]. According to the Nyquist sampling theorem, this means a sampling frequency of at least 40 Hz would be sufficient to capture signal information and avoid aliasing. A sampling frequency of 50 Hz was chosen because a study by Jaen-Vargas et al. found that a window length of 1-2 seconds at 50 Hz was the optimal combination of window size and sampling rate for motion recognition, providing the best trade-off between recognition performance and speed [6]. Therefore, the data were segmented into window lengths of 2 seconds; since the sampling frequency is 50 Hz, the window length is set to 100 samples. The sliding window technique was employed, with an overlap of 75%, as a study by Nurwulan and Jiang found that this provided the highest accuracy with similar features (mean and standard deviation of acceleration) used [7]. Each window was classified based on a majority vote. For example, if 70 samples were labelled as walking and 30 samples were labelled as sitting, the overall window would be classified as walking. Data not labelled with an activity during data collection (containing 'NaN') were cleaned from the dataset.

B. Model Description

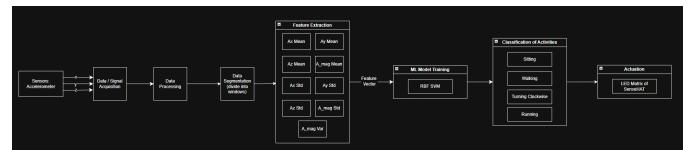


Fig. 1. System Architecture of the HAR system.

The HAR model pipeline consists of sensor data acquisition, data preprocessing, feature extraction, classification, and feedback. Data is collected from a waist-mounted Raspberry Pi SenseHAT. Data is then segmented into overlapping windows. Features are extracted for each window, and a radial basis function (RBF) support vector machine (SVM) classifies the sample into one of the following four activities: sitting, walking, turning, or running. The trained model is deployed on the Raspberry Pi for real-time classification, with predicted activities displayed on the SenseHAT's LED matrix. An overview of the system architecture is shown in Fig. 1.

More in-depth information about the model is shown in the sequence diagram (Fig. 10) and class diagram (Fig. 11) in the Supplemental Information section.

C. Training

A radial basis function (RBF) support vector machine (SVM) model was trained, specifically a one-vs-one classifier. A one-vs-one (OvO) classifier was chosen over a one-vs-rest (OvR) classifier for several reasons. First, the number of classes (four activities) is small. The OvO approach is preferred for small-class models since pairwise comparisons are done, resulting in higher accuracy for SVMs with non-linear kernels like RBF. Additionally, the training data contains an unequal number of samples for each activity. An OvO would not suffer as much from class imbalance as OvR does, which would lead to a biased model that favours the larger class. Finally, the training data is relatively small, so computational cost was not a limiting factor [8].

The RBF SVM in scikit-learn inherently uses a hinge-loss function, which penalizes misclassified samples and samples that are too close to the decision boundary. Since HAR signals are inherently noisy, the hinge-loss function is appropriate because it creates a large margin while still allowing margin violations when necessary. The C parameter in the SVM function controls the trade-off between achieving a wider margin and minimizing classification loss. A larger C parameter focuses more on correctly classifying a training sample, while a smaller C emphasizes a wider margin [9].

To determine the best hyperparameters, a grid search was performed for two kernel options ('rbf' and 'linear') and three C values (1, 10, 100). The best parameters were determined as *kernel* = 'rbf' and *C* = 10. This indicates that non-linear decision boundaries with a moderate C value provide the best trade-off between margin size and classification performance for the collected dataset.

D. Feature Engineering

The SVM model was trained on nine features:

- mean acceleration in the x-axis, y-axis, z-axis, and overall magnitude;
- standard deviation of acceleration in the x-axis, y-axis, z-axis, and overall magnitude;
- variance in the acceleration magnitude.

The mean and standard deviation were taken for all triaxial directions to ensure important directional information is maintained. The activities have defining directional patterns that would be lost if only the mean and standard deviation of the acceleration magnitude were considered. For example, turning and walking may produce similar magnitudes of acceleration, but they can be differentiated after breaking down each acceleration component. While walking may produce periodic, vertical oscillations in the z-axis, turning involves the body rotating around the vertical axis, which involves more lateral (x- and y-axis) acceleration changes. Only the variance of the magnitude of acceleration was considered for the model since

movement intensity is captured in the magnitude value. The features were then scaled before training the model on them.

E. Validation Model

A 10-fold cross-validation model was used to validate the model. A K-fold cross-validation model was chosen because it efficiently uses a small dataset by ensuring every data point is used for both testing and training. Moreover, K-fold cross-validation provides reliable estimates, reducing the likelihood of overfitting or underfitting. Furthermore, it works well on small datasets compared to other validation models. A larger K value was chosen (K = 10 rather than K = 5) because it offers the best trade-off between bias and variance. A large K value is also more effective on small datasets, since each fold uses 90% of the data for training, leading to a less biased estimate of the model's performance [10].

F. HAR Protocol

The procedure for designing and implementing this model is as follows:

- 1) *Data Acquisition*: acquiring data from the SenseHAT accelerometer sensors. The data is collected by placing the Raspberry Pi at waist level and recording the four activities (sitting, walking, running, turning) consecutively, with labelling achieved by using the push button. A sampling frequency of 50 Hz was used. The activities were repeated multiple times to get a wide range of data for each activity.
- 2) *Data Segmentation*: divide the data into two-second windows (each containing 100 samples at a sampling rate of 50 Hz). Assign to each window the label of the activity that appears most frequently within that interval. Use a step size of 25 samples to create sliding windows, resulting in a 75% overlap between consecutive windows.
- 3) *Feature Extraction*: within each window, compute the features the model is trained on, as mentioned in the Feature Engineering section. Then, scale the features before using them for training.
- 4) *Training*: tune the hyperparameters (kernel and C parameter) via grid search; train the SVM model with the best parameters.
- 5) *Validation*: validate the model using the k-fold cross-validation model. A 10-fold model was used.
- 6) *Evaluate on Test Data*: apply the final trained model on the test set and visualize the performance using metrics such as the confusion matrix, principal component analysis (PCA) plot, and classification report.
- 7) *Real-time Classification*: deploy the trained model and real-time data acquisition pipeline on the Raspberry Pi to achieve real-time activity [11].

III. RESULTS

The validation results from the 10-fold cross-validation showed high accuracy, as demonstrated by the results seen in Table I and Fig. 2 and Fig. 3. The mean accuracy was 98.9%.

The boxplot depicts the misclassified samples, with walking and running having a couple of inaccurate predictions.

TABLE I
10-FOLD CROSS-VALIDATION ACCURACY RESULTS

Fold Number	Accuracy
1	100%
2	100%
3	97.4%
4	100%
5	100%
6	97.4%
7	100%
8	100%
9	97.3%
10	97.3%
Mean Accuracy	98.9%

The test results also reflected the high accuracy seen in the validation model. In Table II, the precision, recall, and f1-scores for all activities were 100%, indicating that all windows in the test set were classified correctly.

TABLE II
CLASSIFICATION REPORT

	Precision	Recall	f1-score	Support
Run	1.00	1.00	1.00	18
Sit	1.00	1.00	1.00	25
Turn CW	1.00	1.00	1.00	35
Walk	1.00	1.00	1.00	17

The high accuracy is also reflected in the confusion matrix (Fig. 4), where all predicted classes match the actual outcome. The separability of the activities is also reflected in the PCA and distribution plots, as outlined in the Discussion section.

In addition to testing the model on the test data, a live demonstration of the real-time HAR classification was completed. The CPS was able to accurately detect sitting and running. However, the CPS showed some difficulty with differentiating between walking and turning clockwise.

IV. DISCUSSION

Fig. 5 depicts a PCA plot. In the PCA plot, the activity clusters are, for the most part, separated from each other, indicating a degree of difference between the classes. Sitting has two tight clusters that are well separated from the other dynamic activities, indicating low variance within the class. The low variance is expected for sitting, and the two clusters may correspond to different sitting postures that result in different acceleration values. Running also forms a well-separated cluster from the other activities. The cluster is not as tight, indicating greater variability in acceleration during running. Walking and turning display some overlap, which is expected since they are comparable in movement intensity. However, the clusters are still distinct enough for the SVM to find a separating boundary. The walking cluster is moderately spread out, indicating a higher variability than sitting and turning. Turning creates tight clusters, indicating low variance,

and the two clusters may again be due to differences in posture during data acquisition.

In addition to the PCA plot, the following distribution graphs for the acceleration magnitude for each activity were also created:

- 1) Distribution of acceleration magnitude (Fig. 6). The violin plot shows little spread in sitting acceleration, high spread in running acceleration, and moderate spread for walking and turning. The walking and turning plots show wider sections at approximately the same magnitude of acceleration. This suggests that other features are needed to differentiate between these two activities.
- 2) Distribution of mean acceleration magnitude (Fig. 7). Running had the highest mean acceleration, sitting had the lowest, and walking and turning had overlapping mean acceleration ranges, with turning being slightly higher. Mean acceleration shows the overall level of motion. The mean distribution plot demonstrates that running is high-intensity, turning and walking are moderate intensity, and sitting is low-intensity. It provides a baseline for activity intensity, clearly separating static and dynamic activities.
- 3) Distribution of acceleration magnitude standard deviation (Fig. 8). The plot shows that running had the highest standard deviation of acceleration, followed by walking, turning, and sitting with the lowest standard deviation. Walking and turning had overlapping standard deviation ranges, with walking having an overall slightly higher standard deviation. The standard deviation of acceleration indicates the spread of the acceleration values, with a higher standard deviation indicating large fluctuations (correlated with high-intensity activities) and a lower standard deviation indicating small fluctuations (correlated with low-intensity activities). These patterns are reflected in the violin plot.
- 4) Distribution of acceleration magnitude variance (Fig. 9). Similar to the standard deviation distribution, running had the highest variance in acceleration, followed by walking, turning, and sitting with the lowest variance. The variance of acceleration measures how much acceleration fluctuates around the mean. A high variance is expected for high-intensity activities and low variance for low-intensity activities; this is reflected in the violin plot. This reflects the complexity and variability of the activity, helping to differentiate between activities with similar mean accelerations.

Overall, activities were recognized with a high degree of accuracy, as demonstrated by the classification report results and plots of the confusion matrix, PCA, and distribution of acceleration magnitude for various features. The PCA and feature distribution plots provide insight into why walking and turning clockwise were occasionally misclassified during the real-time HAR demonstration. In both visualizations, the feature clusters for walking and turning displayed some overlap, reflecting their similar motion intensity, explaining

the classifier's difficulty in distinguishing between them. The limited dataset and minor inconsistencies in the positioning of the Raspberry Pi during data acquisition may also explain the marginal differences observed in the acceleration data. To mitigate this issue in the future, other sensors could be incorporated. For instance, a gyroscope captures angular velocity and would help differentiate rotational motion from walking. Additional features should also be explored. While mean, standard deviation, and variance performed well for this dataset, a broader range of features should be evaluated to identify the most discriminative ones. Potential features to consider include jerk (derivative of acceleration), signal energy, entropy, and frequency-domain features like spectral power. Expanding the dataset to include more trials and a variety of subjects would also further improve model generalization and improve the classification accuracy across the general population.

V. CONCLUSION

This project demonstrates the successful development of a CPS for HAR using a Raspberry Pi, SenseHat, and a machine-learning classification pipeline. The system achieved high accuracy on the test data, strong cross-validation accuracy, and clear separation between most activities in the PCA and distribution plots; it was able to reliably classify high-intensity (running) activities from moderate-intensity (walking and turning clockwise) and low-intensity (sitting) activities. The results demonstrate that a lightweight system like the Raspberry PI can perform accurate real-time activity recognition with the right features and preprocessing steps.

However, a closer inspection of the PCA and feature distribution plots, along with the live demonstration of the HAR, highlights some limitations. Walking and turning show overlapping feature distributions, which led to some misclassifications in the demonstration. The overlap is likely due to the comparable motion intensities of walking and turning, combined with the limited discriminative capability of features based solely on acceleration. Additionally, the dataset was collected from a single subject; therefore, the model is likely not generalizable to the broader population. This highlights the importance of building an HAR system that incorporates a variety of sensors, robust features, and is trained on larger datasets to develop a universally accurate system.

These issues can be addressed in the future by incorporating a gyroscope to better capture rotation motion, which could more effectively separate the activities, especially walking and turning. Additionally, expanding the feature set to determine the most discriminative features can also enhance the model's accuracy. Moreover, collecting data from multiple subjects in different environments would improve the model's robustness and reduce bias towards a specific movement pattern. Furthermore, other advanced models aside from SVM (eg. Convolutional Neural Networks, Bayesian Neural Networks, Random Forest) should be explored to help improve the existing model. Overall, these improvements would improve

the reliability and scalability of the system when used in real-world applications in the healthcare field.

REFERENCES

- [1] "Cyber-physical systems - an overview," <https://www.sciencedirect.com/topics/engineering/cyber-physical-systems>, accessed: Dec. 01, 2025.
- [2] H. F. Nweke, Y. W. Teh, M. A. Al-Garadi, and U. R. Alo, "Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges," *Expert Systems with Applications*, vol. 105, pp. 233–261, Sep. 2018.
- [3] "Raspberry pi sense hat v2," <https://www.pishop.ca/product/raspberry-pi-sense-hat-v2/>, 2025, accessed: Dec. 01, 2025.
- [4] F. Attal, S. Mohammed, M. Dedabirshvili, F. Chamroukhi, L. Oukhellou, and Y. Amirat, "Physical human activity recognition using wearable sensors," *Sensors*, vol. 15, no. 12, pp. 31 314–31 338, 2015.
- [5] "Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring," <https://ieeexplore.ieee.org/abstract/document/1573717>, accessed: Oct. 26, 2025.
- [6] M. Jaén-Vargas *et al.*, "Effects of sliding window variation in the performance of acceleration-based human activity recognition using deep learning models," *PeerJ Computer Science*, vol. 8, p. e1052, 2022.
- [7] N. R. Nurwulan and B. C. Jiang, "Window selection impact in human activity recognition," *International Journal of Innovative Technology and Interdisciplinary Sciences*, vol. 3, no. 1, pp. 381–394, 2020.
- [8] GeeksforGeeks, "Multiclass classification using support vector machines (svm)," <https://www.geeksforgeeks.org/machine-learning/multi-class-classification-using-support-vector-machines-svm/>, 2024.
- [9] ———, "Rbf svm parameters in scikit learn," GeeksforGeeks, 07 2023. [Online]. Available: <https://www.geeksforgeeks.org/python/rbf-svm-parameters-in-scikit-learn/>
- [10] ———, "K fold cross validation in machine learning," GeeksforGeeks, 04 2025. [Online]. Available: <https://www.geeksforgeeks.org/machine-learning/k-fold-cross-validation-in-machine-learning/>
- [11] A. Ferrari *et al.*, "Trends in human activity recognition using smartphones," *Journal of Reliable Intelligent Environments*, vol. 7, no. 3, pp. 189–213, 2021.

VI. SUPPLEMENTAL INFORMATION

A. Data Plots

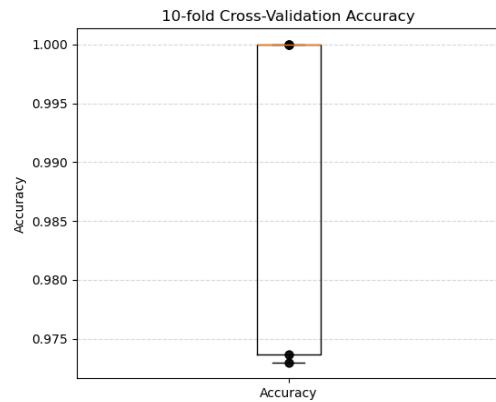


Fig. 2. Box plot of 10-fold cross-validation accuracy.

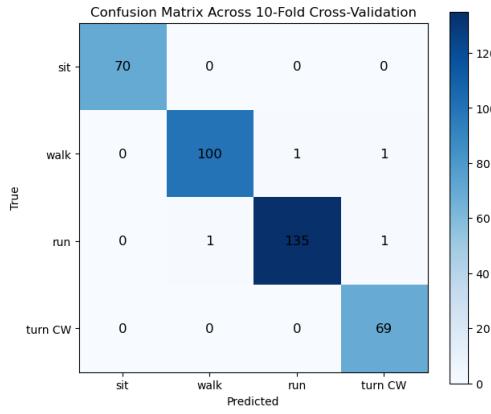


Fig. 3. Confusion matrix for 10-fold cross-validation.

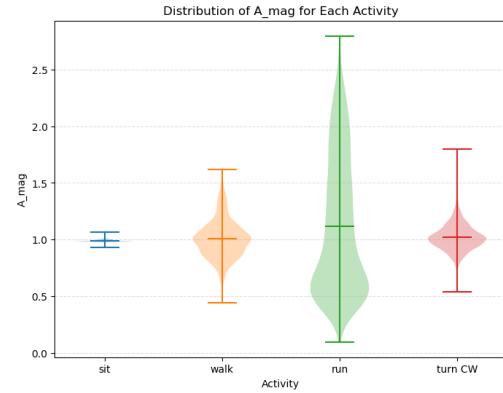


Fig. 6. Distribution of A_Mag for each activity.

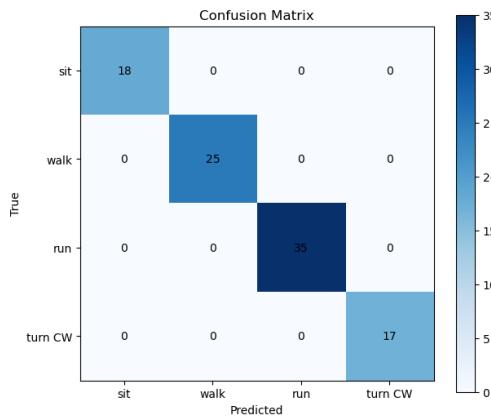


Fig. 4. Confusion matrix of test data.

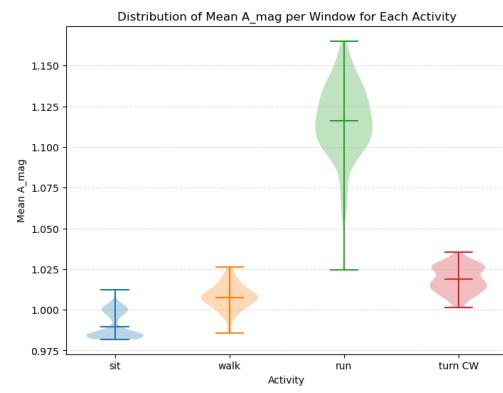


Fig. 7. Distribution of mean A_mag for each activity.

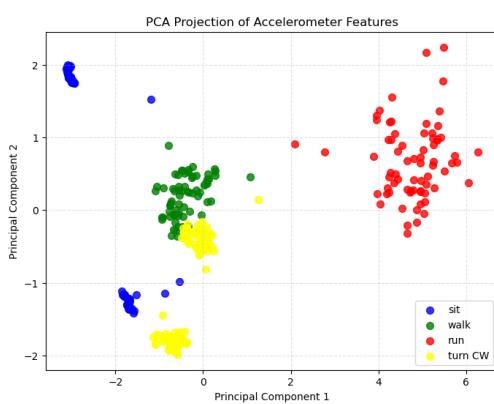


Fig. 5. PCA projection of accelerometer features.

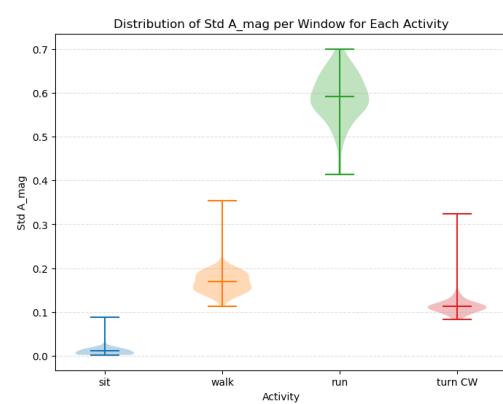


Fig. 8. Distribution of A_Mag standard deviation for each activity.

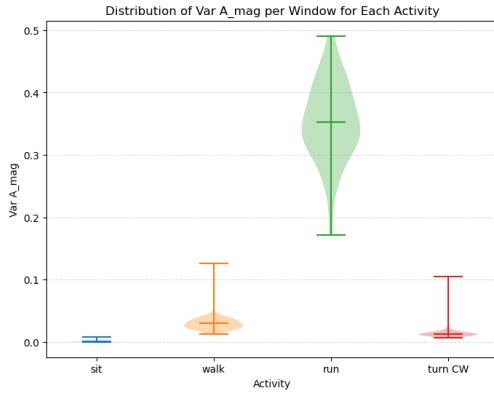


Fig. 9. Distribution of A_Mag variance for each activity.

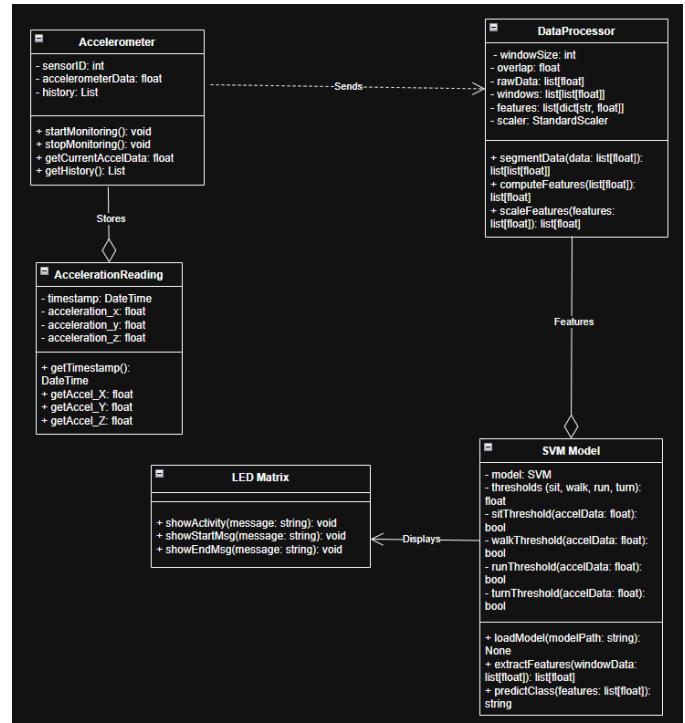


Fig. 11. Class Diagram of the HAR system.

B. UML of CPS

C. Source Code

Python Script for Training and Validating SVM Model

1) Setup:

```

1 import pandas as pd
2 from sklearn.preprocessing import StandardScaler
3 from sklearn.svm import SVC
4 from sklearn.metrics import classification_report, confusion_matrix
5 from sklearn.multiclass import OneVsOneClassifier
6 from sklearn.model_selection import cross_val_score, GridSearchCV,
7     train_test_split, cross_val_predict
8 from sklearn.decomposition import PCA
9 import matplotlib.pyplot as plt
10 import numpy as np
11 import joblib
  
```

2) Load Data and Window Preparation:

```

data = pd.read_csv('training_data.csv')

# Filter for the activities
data = data[data['activity'].isin(['sit', 'walk', 'run', 'turn CW'])]

# Prepare features (windowing)
WINDOW_SIZE = 100 # 2 seconds at 50Hz
def compute_window_features(ax, ay, az):
    ax = np.array(ax)
    ay = np.array(ay)
    az = np.array(az)
    a_mag = np.sqrt(ax**2 + ay**2 + az**2)
    return [np.mean(ax), np.mean(ay),
            np.mean(az), np.mean(a_mag),
            ...]
  
```

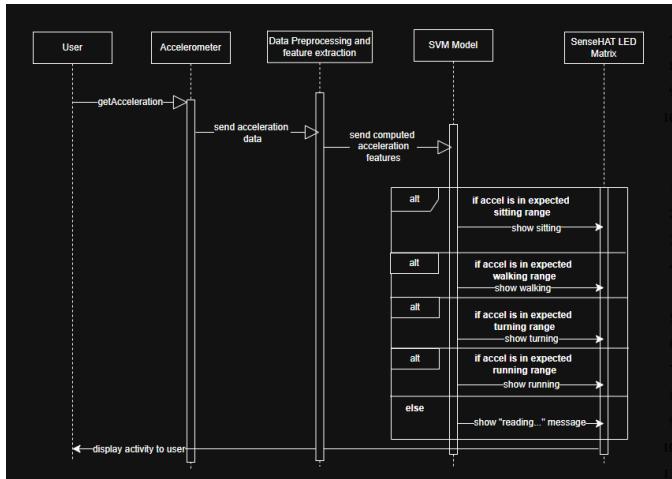


Fig. 10. Sequence Diagram of the HAR system.

```

14         np.std(ax), np.std(ay),
15             ↪ np.std(az), np.std(a_mag),
16             np.var(a_mag)]
17 STEP_SIZE = WINDOW_SIZE // 4 # 75% overlap
18
19 windows_features = []
20 windows_labels = []
21 windows_activities = []
22
23 # Loop with step of 25 steps
24 for start in range(0, len(data) - WINDOW_SIZE,
25     ↪ STEP_SIZE):
26     window = data.iloc[start : start +
27         ↪ WINDOW_SIZE]
28
29     ax = window['Ax'].values
30     ay = window['Ay'].values
31     az = window['Az'].values
32     acts = window['activity'].values
33
34     # Compute features for the window
35     features = compute_window_features(ax, ay,
36         ↪ az)
37     windows_features.append(features)
38
39     # Label by most common activity in the
40     ↪ window
41     most_common = window['activity'].mode()[0]
42     windows_labels.append(most_common)
43
44     windows_activities.append(list(acts))
45
46 X = np.array(windows_features)
47 y = np.array(windows_labels)
48
49 3) Data Preparation, Tune Hyperparameters:
50 X_train, X_test, y_train, y_test =
51     ↪ train_test_split(
52         X, y, test_size=0.2, random_state=42,
53             ↪ stratify=y
54 )
55
56 scaler = StandardScaler()
57 X_train_scaled = scaler.fit_transform(X_train)
58 X_test_scaled = scaler.transform(X_test)
59
60 # Hyperparameter tuning with GridSearchCV
61 param_grid = {
62     'estimator_C': [1, 10, 100],
63     'estimator_kernel': ['linear', 'rbf'],
64     'estimator_probability': [True]
65 }
66
67 grid = GridSearchCV(
68     OneVsOneClassifier(SVC(probability=True)),
69     param_grid,
70     cv=5,
71     scoring='accuracy'
72 )
73
74 grid.fit(X_train_scaled, y_train)
75
76 print("Best params:", grid.best_params_)
77 print("Best CV score:", grid.best_score_)
78
79 4) Train and Validate Model:
80
81 # train SVM with best params, one-vs-one
82     ↪ classification
83
84 model = OneVsOneClassifier(SVC(kernel=grid.be_j
85     ↪ st_params_['estimator_kernel'],
86     ↪ C=grid.best_params_['estimator_C'],
87     ↪ probability=True))
88 model.fit(X_train_scaled, y_train)
89
90
91 # SVM Model Validation
92 # Validate with 10-fold cross-validation
93 # use k = 10 since smaller dataset
94 scores = cross_val_score(
95     OneVsOneClassifier(SVC(kernel=grid.best_p_j
96         ↪ arams_['estimator_kernel'],
97         ↪ C=grid.best_params_['estimator_C'],
98         ↪ probability=True)),
99     X_train_scaled,
100    y_train,
101    cv=10,
102    scoring='accuracy'
103 )
104
105 y_pred_cv = cross_val_predict(
106     OneVsOneClassifier(SVC(kernel=grid.best_p_j
107         ↪ arams_['estimator_kernel'],
108         ↪ C=grid.best_params_['estimator_C'],
109         ↪ probability=True)),
110     X_train_scaled,
111     y_train,
112     cv=10
113 )
114
115 print("Cross-validation accuracy scores:",
116     ↪ scores)
117 print("Mean CV accuracy:", scores.mean())
118
119 # Plot cross-validation scores
120 plt.figure(figsize=(6, 5))
121 plt.boxplot(scores, vert=True)
122
123 # Also plot individual points
124 plt.scatter(
125     [1]*len(scores), scores,
126     color='black', s=40, alpha=0.7
127 )
128
129 plt.title("10-fold Cross-Validation Accuracy")
130 plt.ylabel("Accuracy")
131 plt.grid(True, axis='y', linestyle='--',
132     ↪ alpha=0.5)
133 plt.xticks([1], ['Accuracy'])
134 plt.show()
135
136 # Confusion matrix for CV predictions
137 cm_cv = confusion_matrix(y_train, y_pred_cv)
138 classes = ["sit", "walk", "run", "turn CW"]
139
140 plt.figure(figsize=(7, 6))
141 plt.imshow(cm_cv, cmap="Blues")
142 plt.colorbar()
143
144 plt.xticks(np.arange(len(classes)), classes)
145 plt.yticks(np.arange(len(classes)), classes)
146
147 for i in range(len(classes)):
148     for j in range(len(classes)):
149         plt.text(j, i, cm_cv[i,j],
150             ↪ ha='center', va='center',
151             ↪ color='black', fontsize=12)
152

```

```

57 plt.xlabel("Predicted")
58 plt.ylabel("True")
59 plt.title("Confusion Matrix Across 10-Fold
   ↳ Cross-Validation")
60 plt.show()
61 5) Evaluate Test Set:
62 # evauluate test set
63 y_pred = model.predict(X_test_scaled)
64 print("Classification Report:\n",
   ↳ classification_report(y_test, y_pred))
65
66 # Visualize on test data
67 # a) confusion matrix
68 cm = confusion_matrix(y_test, y_pred)
69 print("Confusion Matrix:\n", cm)
70
71 # Visualize Confusion Matrix
72 plt.figure(figsize=(7, 6))
73 plt.imshow(cm, cmap="Blues")
74 plt.colorbar()
75
76 classes = ["sit", "walk", "run", "turn CW"]
77
78 # Axis labels
79 plt.xticks(np.arange(4), classes)
80 plt.yticks(np.arange(4), classes)
81 # label squares
82 for i in range(4):
83     for j in range(4):
84         plt.text(j, i, cm[i, j], ha="center",
   ↳ va="center", color="black")
85
86 plt.xlabel("Predicted")
87 plt.ylabel("True")
88 plt.title("Confusion Matrix")
89 plt.show()
90
91 # b) PCA Visualization of Feature Space
92 pca = PCA(n_components=2)
93 X_pca = pca.fit_transform(X_train_scaled)
94
95 plt.figure(figsize=(8, 6))
96 # Use a fixed activity order and explicit
   ↳ color mapping so plots are consistent
97 activity_list = ['sit', 'walk', 'run', 'turn
   ↳ CW']
98 color_map = {'sit': 'blue', 'walk': 'green',
   ↳ 'run': 'red', 'turn CW': 'yellow'}
99
100 for act in activity_list:
101     col = color_map.get(act, 'gray')
102     idx = (y_train == act)
103     plt.scatter(
104         X_pca[idx, 0],
105         X_pca[idx, 1],
106         label=act,
107         color=col,
108         s=50,
109         alpha=0.8
110     )
111
112 plt.title("PCA Projection of Accelerometer
   ↳ Features")
113 plt.xlabel("Principal Component 1")
114 plt.ylabel("Principal Component 2")
115 plt.legend()
116 plt.grid(True, linestyle='--', alpha=0.4)
117 plt.show()
118
119 # c) feature distribution per activity using
   ↳ A_Mag only
120 plt.figure(figsize=(8, 6))
121
122 for act in activity_list:
123     values = data[data['activity'] ==
   ↳ act]['A_mag']
124     plt.violinplot(values,
   ↳ positions=[activity_list.index(act)],
   ↳ showmeans=True)
125
126 plt.title("Distribution of A_mag for Each
   ↳ Activity")
127 plt.xlabel("Activity")
128 plt.ylabel("A_mag")
129 plt.xticks(range(len(activity_list)),
   ↳ activity_list)
130 plt.grid(True, axis='y', linestyle='--',
   ↳ alpha=0.4)
131 plt.show()
132
133 # d) feature distribution using mean A_mag per
   ↳ window
134 plt.figure(figsize=(8, 6))
135 mean_a_mag_per_window = []
136
137 for features in windows_features:
138     mean_a_mag_per_window.append(features[3])
139     # mean A_mag is the 4th feature
140 mean_a_mag_per_window =
141     np.array(mean_a_mag_per_window)
142 for act in activity_list:
143     values = mean_a_mag_per_window[np.array(wi
   ↳ ndows_labels) == act]
144     plt.violinplot(values,
   ↳ positions=[activity_list.index(act)],
   ↳ showmeans=True)
145
146 plt.title("Distribution of Mean A_mag per
   ↳ Window for Each Activity")
147 plt.xlabel("Activity")
148 plt.ylabel("Mean A_mag")
149 plt.xticks(range(len(activity_list)),
   ↳ activity_list)
150 plt.grid(True, axis='y', linestyle='--',
   ↳ alpha=0.4)
151 plt.show()
152
153 # e) feature distribution using std of A_mag
   ↳ per window
154 plt.figure(figsize=(8, 6))
155 std_a_mag_per_window = []
156
157 for features in windows_features:
158     std_a_mag_per_window.append(features[7])
159     # std A_mag is the 8th feature
160 std_a_mag_per_window =
161     np.array(std_a_mag_per_window)
162 for act in activity_list:
163     values = std_a_mag_per_window[np.array(wi
   ↳ ndows_labels) == act]
164     plt.violinplot(values,
   ↳ positions=[activity_list.index(act)],
   ↳ showmeans=True)
165
166 plt.title("Distribution of Std A_mag per
   ↳ Window for Each Activity")

```

```

102 plt.xlabel("Activity")                                26
103 plt.ylabel("Std A_mag")                             27
104 plt.xticks(range(len(activity_list)),               28
105    ↳ activity_list)
106 plt.grid(True, axis='y', linestyle='--',           29
107    ↳ alpha=0.4)
108 plt.show()
109
110 # f) feature distribution using var A_mag per
111    ↳ window
112 plt.figure(figsize=(8, 6))
113 var_a_mag_per_window = []
114
115 for features in windows_features:
116     var_a_mag_per_window.append(features[8])
117     ↳ # var A_mag is the 9th feature
118 var_a_mag_per_window =
119     ↳ np.array(var_a_mag_per_window)
120 for act in activity_list:
121     values = var_a_mag_per_window[np.array(wi_30
122     ↳ ndows_labels) == act]
123     plt.violinplot(values,
124     ↳ positions=[activity_list.index(act)],
125     ↳ showmeans=True)
126
127 plt.title("Distribution of Var A_mag per          32
128    ↳ Window for Each Activity")
129 plt.xlabel("Activity")
130 plt.ylabel("Var A_mag")
131 plt.xticks(range(len(activity_list)),             33
132    ↳ activity_list)
133 plt.grid(True, axis='y', linestyle='--',           34
134    ↳ alpha=0.4)
135 plt.show()

```

Python Script for Data Acquisition

```

1 from sense_hat import SenseHat
2 from datetime import datetime
3 import joblib
4 import numpy as np
5 import csv, time, math
6
7 sense = SenseHat()
8 sense.clear()
9 #filename = "deliverable1_log_" +
10    ↳ datetime.now().strftime("%H%M%S") + ".csv"
11 filename = "final_training_data.csv"
12 button_labels = {"up": "walk", "down": "sit",
13    ↳ "left": "run", "right": "turn CW"}           37
14 activity = None
15 active = False
16 msg = ""
17 color_map = { "sit": [0, 0, 255],
18    ↳ "walk": [0, 255, 0],                         39
19    ↳ "run": [255, 0, 0],
20    ↳ "turn CW": [255,           ↳ 255, 0]
21    ↳ }
22 color = [255, 255, 255]
23
24 with open(filename, 'w', newline='') as f:
25     writer = csv.writer(f)
26     writer.writerow(["time", "Ax", "Ay",           40
27        ↳ "Az", "A_mag", "active",
28        ↳ "activity"])
29     sense.show_message("START",                   42
30        ↳ text_colour=[255, 255, 255],
31        ↳ scroll_speed=0.1)                         43

```

```

while True:
    for e in
        ↳ sense.stick.get_events():
            if e.direction == "up"
                ↳ and e.action ==
                ↳ "pressed" or
                ↳ e.direction ==
                ↳ "down" and
                ↳ e.action ==
                ↳ "pressed" or
                ↳ e.direction ==
                ↳ "left" and
                ↳ e.action ==
                ↳ "pressed" or
                ↳ e.direction ==
                ↳ "right" and
                ↳ e.action ==
                ↳ "pressed":
                    active = True
                    activity =
                        ↳ button_la_
                        ↳ bels [
                        ↳ e.directi_
                        ↳ on ]
#msg = "sit"
                    if
                        ↳ activity
                        ↳ == "sit"
                    else
                        ↳ "walk"
                    msg = activity
                    color = color_j
                        ↳ _map.get(j
                        ↳ activity,
                        ↳ [255, 255,
                        ↳ 255])
                    print("activi_
                        ↳ ty:",
                        ↳ activity)
                    sense.show_me_
                        ↳ ssage (
                        ↳ msg , tex_
                        ↳ t_colour =
                        ↳ color ,
                        ↳ scroll_sp_
                        ↳ eed=0.05)
                elif e.direction ==
                    ↳ "middle" and
                    ↳ e.action ==
                    ↳ "pressed":
                        sense.show_me_
                            ↳ ssage("EN_
                            ↳ D",
                            ↳ text_colou_
                            ↳ ur=[255,
                            ↳ 255, 255],
                            ↳ scroll_sp_
                            ↳ eed=0.05)
                        active = False
                        print("record_
                            ↳ ing
                            ↳ stopped")
                        sense.clear()
                        f.close()
                        exit()

```

