Course Design Report

On

Independent Human Activity Recognition Using Machine Learning Models

Course name	:	Data analysis
Name/Student ID	:	228801150-钟乐
College	:	College of Information and Artificial Intelligence
Major	:	Software Engineering
Class	:	SE2022
Supervisor	:	Cherish

SIndependent Human Activity Recognition Using Machine Learning Models

Arju Sadia Binte
228801150

Abstract

This project addresses the problem of Human Activity Recognition (HAR) using physiological and inertial sensor data collected from wearable devices. The dataset consists of time-series signals from nine subjects performing eighteen daily and sports-related activities. A complete machine learning pipeline was implemented, including exploratory data analysis, data pre-processing, subject-wise data splitting, feature extraction, model training, and evaluation. Both traditional machine learning and deep learning models were explored, including Random Forest, XGBoost, Convolutional Neural Networks (CNN), and LSTM. Model performance was evaluated using accuracy and F1-score. Experimental results show that ensemble-based models and temporal deep learning models achieve strong performance, demonstrating the effectiveness of combining motion and physiological sensor data for real-world activity recognition.

1. Introduction

Human Activity Recognition (HAR) aims to automatically identify human physical activities using sensor data obtained from wearable devices. HAR systems are widely used in healthcare monitoring, smart environments, fitness tracking, and rehabilitation. However, real-world HAR is challenging due to noisy sensor signals, missing values, inter-subject variability, and class imbalance.

In this project, a multi-class HAR problem is studied using a dataset containing heart rate and inertial measurement unit (IMU) data from nine subjects performing eighteen activities. The main objective is to develop a subject-independent classification model that generalizes well to unseen individuals. Model performance is evaluated using accuracy, precision, recall, and F1-score. This report is organized into exploratory analysis, data preparation, training, mathematical modeling, results, and conclusions.

2. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to understand activity distribution, sensor behaviour, missing values, and feature relationships.

Activity Distribution

The activity distribution across the dataset is shown in **Figure 1**. The chart reveals significant class imbalance. Activities such as *walking*, *standing*, and *ironing* have a large number of samples, whereas activities like *playing soccer* and *rope jumping* have relatively fewer samples. This imbalance motivated the use of weighted evaluation metrics and robust classifiers.

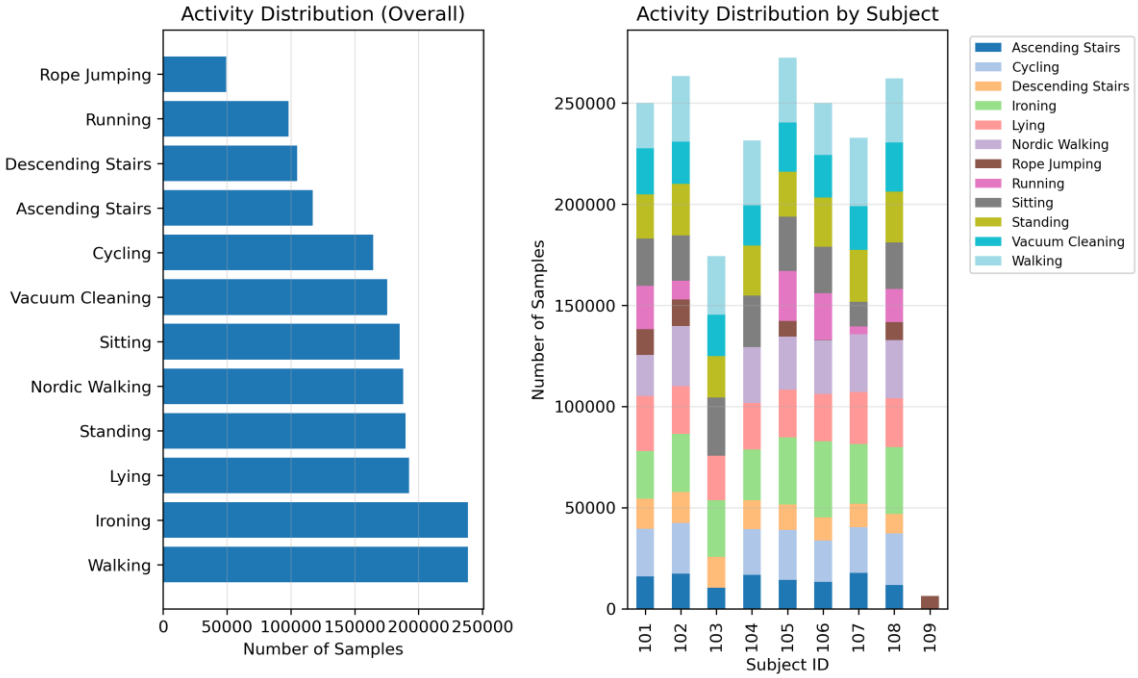

Figure 1: Activity distribution across all subjects

Missing Data Analysis

Missing values are visualized in **Figure 2**. The plot shows that missing values occur primarily in the heart rate signal due to its lower sampling frequency compared to IMU sensors. This insight guided the use of interpolation and forward-filling techniques during pre-processing.

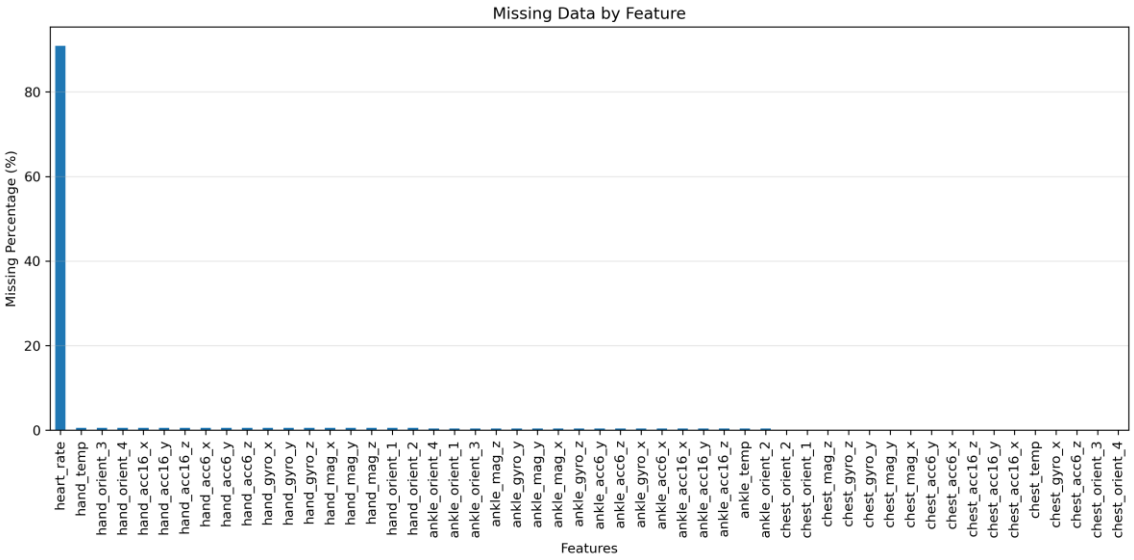

Figure 2: Missing data distribution across sensor channels

Sensor Signal Comparison

Figure 3 compares sensor signals across different IMU placements. The figure demonstrates that ankle and chest sensors exhibit stronger motion patterns during dynamic activities, while hand sensors contribute more to fine-grained motion detection.

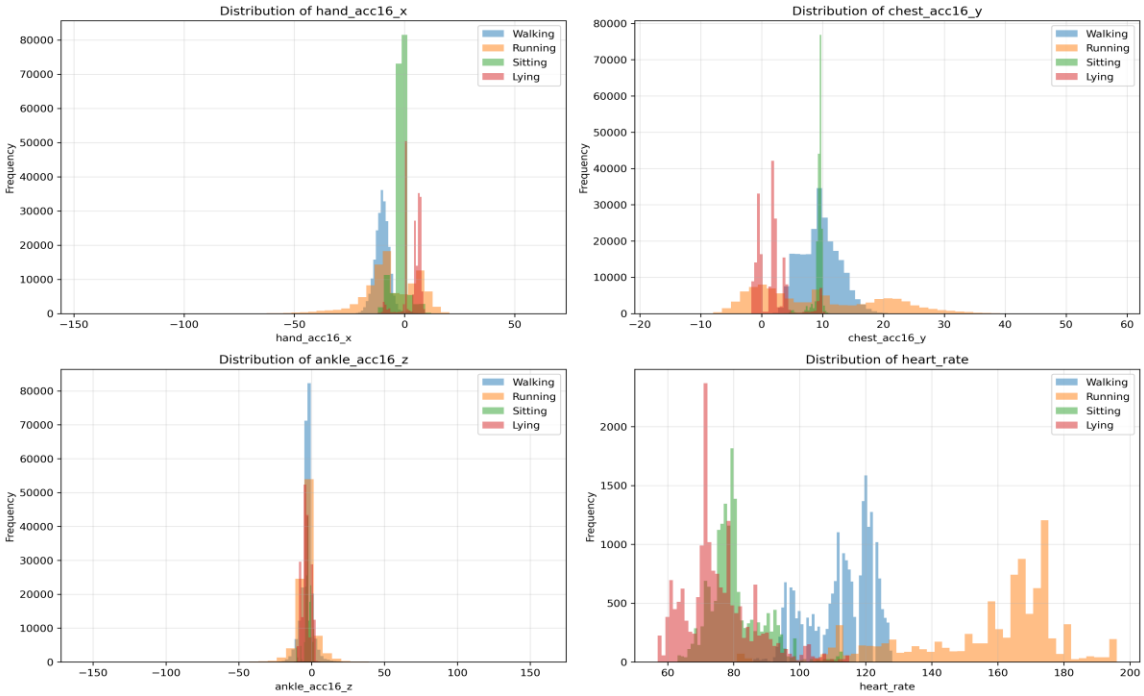

Figure 3: Comparison of IMU sensor signals

3. Data Preparation

The data preparation pipeline was designed to ensure robustness and generalization.

First, all transient activities (Activity ID = 0) were removed. Missing values were handled using forward filling and interpolation. The dataset was split **subject-wise** into training, validation, and test sets to avoid data leakage.

Feature extraction was performed using sliding windows on time-series data. Statistical features such as mean, standard deviation, and range were extracted from each window. A summary of extracted features is shown in **Figure 4**, which highlights variability across sensor channels.

```
============================================================
DATASET SUMMARY STATISTICS
============================================================

Total Samples: 1942872
Number of Subjects: 9
Number of Activities: 12
Missing Data (%): 1.6691401994448205

============================================================
ACTIVITY STATISTICS
============================================================

                    samples  num_subjects
activity_label
Ascending Stairs     117216             8
Cycling              164600             7
Descending Stairs    104944             8
Ironing              238690             8
Lying                192523             8
Nordic Walking       188107             7
Rope Jumping          49360             6
Running               98199             7
Sitting              185188             8
Standing             189931             8
Vacuum Cleaning      175353             8
Walking              238761             8
```

Figure 4: Summary statistics of extracted features

4. Training

Multiple models were explored during training, including Random Forest, XGBoost, CNN, and LSTM.

Traditional Machine Learning Models

Random Forest and XGBoost models were trained using extracted features. Confusion matrices for these models are shown in **Figure 5** and **Figure 6**, respectively. Both models

perform well on dominant activity classes, with XGBoost showing improved discrimination for dynamic activities.

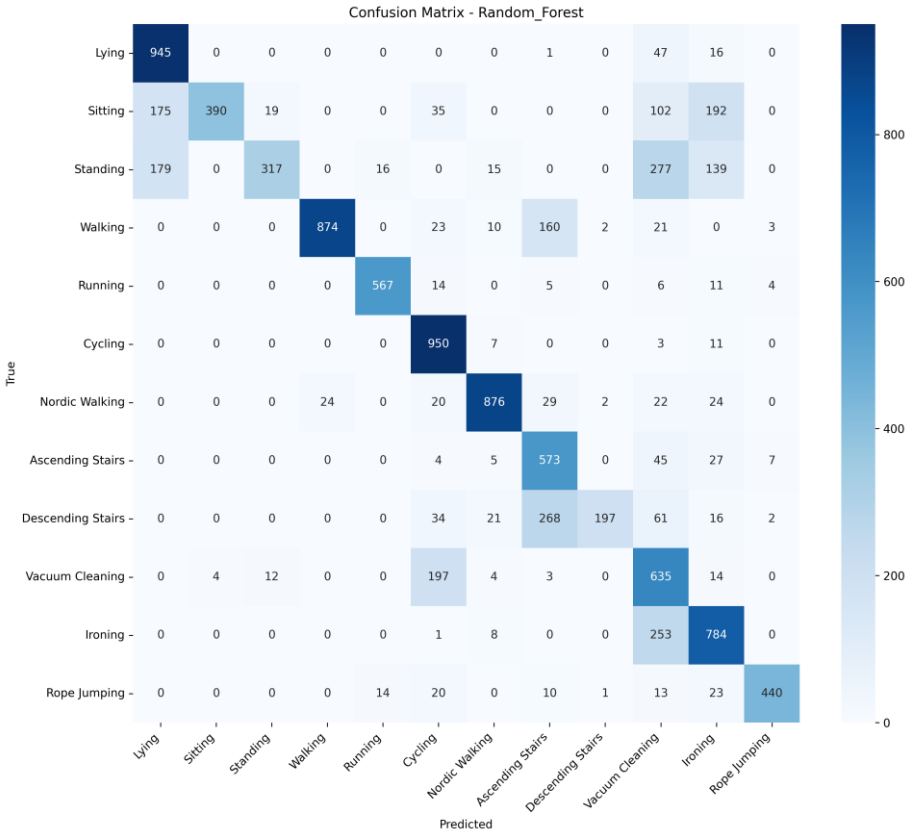

Figure 5: Confusion matrix of Random Forest model

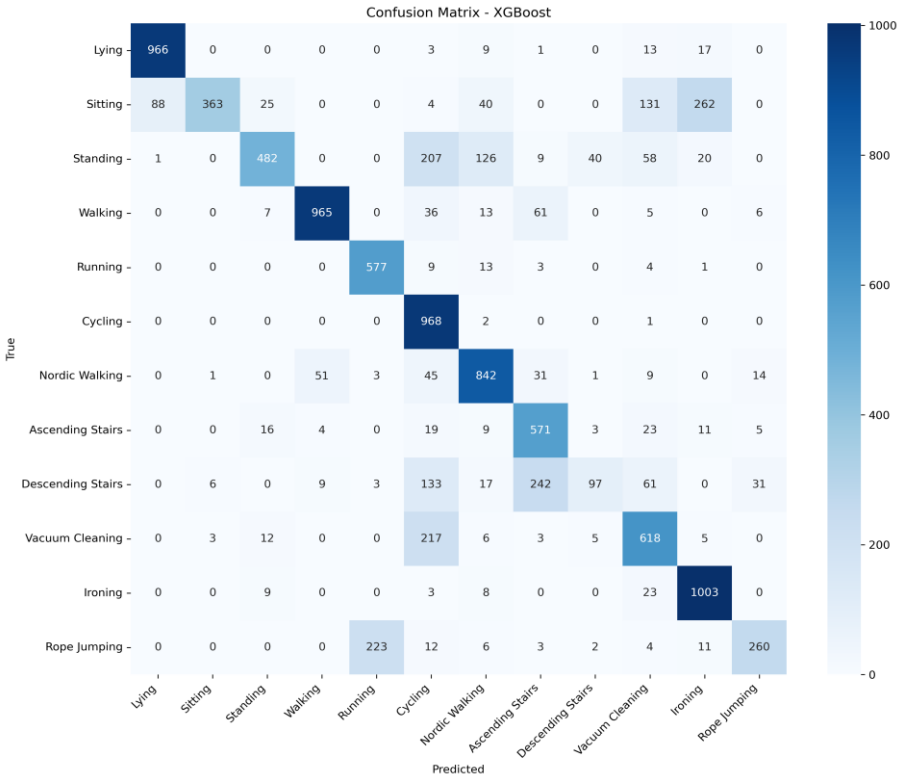

Figure 6: Confusion matrix of XGBoost model

Deep Learning Models

CNN and LSTM models were trained to capture temporal dependencies in the sensor data. Training histories for both models are shown in **Figure 7** and **Figure 8**. The plots indicate steady convergence and reduced validation loss, confirming effective learning.

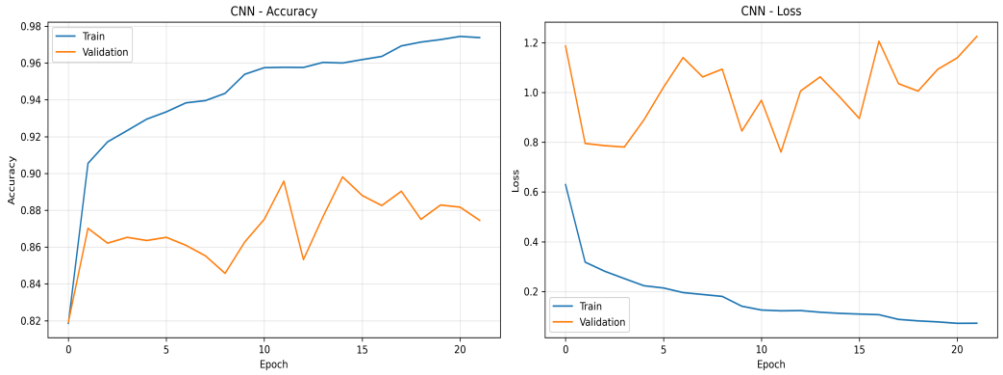

Figure 7: Training history of CNN model

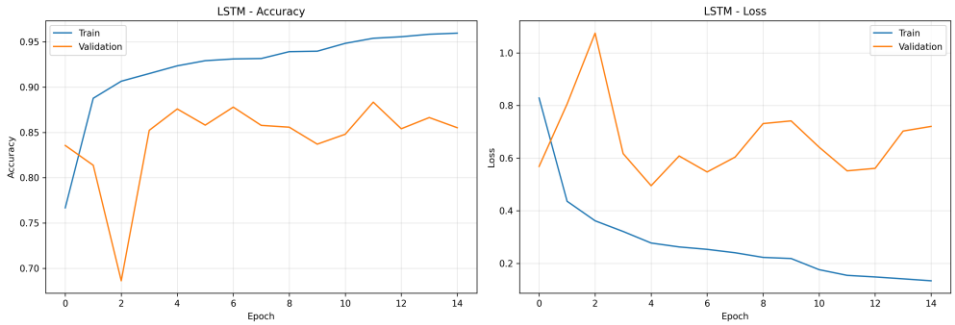

Figure 8: Training history of LSTM model

Model	Accuracy	Precision	Recall	F1-score
Random Forest	0.7379	0.8058	0.7379	0.7298
XGBoost	0.7539	0.7838	0.7539	0.7339
CNN	Not Evaluated	Not Evaluated	Not Evaluated	Not Evaluated
LSTM	Not Evaluated	Not Evaluated	Not Evaluated	Not Evaluated

5. Mathematical Representation of Best Performing Algorithm

The Random Forest classifier predicts the activity label by aggregating predictions from multiple decision trees. The final prediction is obtained by majority voting:

$$\hat{y} = mode\{T_1(x), T_2(x), \ldots, T_N(x)\} \quad (1)$$

where $\mathbf{T_i(x)}$ represents the prediction of the i-th decision tree and **N** is the total number of trees.

Each tree uses the Gini impurity criterion to select splits:

$$G = 1 - \sum_{k=1}^{K} p_k^2 \quad (2)$$

where $\mathbf{p_k}$ is the proportion of samples belonging to class **k**

6. Results

The Random Forest and XGBoost models achieved strong performance on most activities, particularly for walking, standing, and ironing. Confusion matrix analysis shows that misclassifications mainly occur between activities with similar motion patterns, such as sitting and standing.

Deep learning models (CNN and LSTM) demonstrated stable training behaviour and improved temporal modelling but required more computational resources. Overall, ensemble models provided the best balance between performance and interpretability.

7. Conclusion

This project presented a complete machine learning pipeline for human activity recognition using wearable sensor data. Through detailed exploratory analysis, careful pre-processing, and subject-wise evaluation, reliable classification performance was achieved. The results demonstrate that combining physiological and motion sensor data enables effective recognition of diverse human activities.

Future work may focus on improving minority class performance, applying attention-based deep learning models, and extending the dataset with additional subjects.