PAMAP2 Activity Classification: Comparative Analysis of Models

Title of Project : PAMAP2 Activity Classification: Comparative Analysis of Models

Student Name : GAURAV KUMAR CHAURASIYA

Enrollment Number : 00919011921

Signature:

Email ID : gaurav.919011921@ipu.ac.in

Contact Number : 9873390197

Google Drive Link : https://colab.research.google.com/drive/1C-kFBWT83YzutdsHAEsbhtesa-v2k85g?usp=sharing

Google Website Link: https://sites.google.com/view/pamap2activityclassification/home

Github Link : https://github.com/gauravkumarchaurasiya/PAMAP2-Activity-Classification-Comparative-Analysis-of-Models

Youtube Video Link: https://youtu.be/QzI6a_b0UXc

PAMAP2 Activity Classification: Comparative Analysis of Models

Abstract:

This study presents a comprehensive comparative analysis of different models for activity classification using the PAMAP2 dataset. Human activity recognition plays a crucial role in various domains, including healthcare monitoring, sports performance analysis, and human-computer interaction. The accurate classification of activities based on sensor data collected from wearable devices is essential for developing effective and personalized applications in these fields.

The PAMAP2 dataset has emerged as a valuable resource for activity recognition research. It consists of sensor data recorded from 9 inertial measurement units worn by participants during 18 different physical activities. The dataset includes measurements from accelerometers, gyroscopes, and magnetometers, providing rich information about participants' movements and orientations during various activities.

In this analysis, a range of classification models is explored, including Decision trees, Random forests, k-nearest neighbors (KNN), Logistic regression, naive Bayes, and Adaboost Classifier. Each model is trained and evaluated on the PAMAP2 dataset using common evaluation metrics such as accuracy, precision, recall, and F1 score. These metrics provide a comprehensive assessment of the models' performance in accurately predicting different activities.

Furthermore, the analysis delves into the impact of feature selection, preprocessing techniques, and parameter tuning on the models' performance. Feature selection plays a crucial role in identifying relevant sensor data that contribute most to activity classification. Preprocessing techniques such as data cleaning, normalization.

The findings of this comparative analysis provide valuable insights into the strengths and weaknesses of each classification model. Researchers and practitioners can leverage this information to select the most appropriate model for activity classification tasks using the PAMAP2 dataset. The analysis sheds light on the effectiveness of different models in distinguishing between activities and highlights the factors that significantly impact their performance.

In conclusion, this study contributes to the advancement of activity classification techniques using the PAMAP2 dataset. By exploring various models and evaluating their performance, this analysis enables the identification of accurate and efficient approaches for activity recognition. The outcomes of this study have practical implications for the development of real-world applications that rely on accurate activity classification, promoting improved healthcare monitoring, enhanced sports performance analysis, and more intuitive human-computer interaction.

Keywords:

PAMAP2 dataset, activity classification, Human activity recognition, classification models on Pamap2, evaluation metrics

Introduction

The PAMAP2 (Physical Activity Monitoring and Assessment System) dataset has emerged as a valuable resource for researchers and practitioners in the field of human activity recognition. With the proliferation of wearable sensor technology, the PAMAP2 dataset provides a comprehensive collection of multimodal sensor data. This enables the analysis and classification of various physical activities.

In this report, I present an in-depth investigation into activity classification on the PAMAP2 dataset. This study developed an accurate and robust model capable of identifying and categorizing different activities based on sensor readings. Accurately classifying activities has numerous practical applications, such as fitness tracking, health monitoring, and personalized coaching.

The PAMAP2 dataset consists of data recorded from 9 inertial sensors worn by participants during various physical activities. These activities include walking, running, cycling, ascending and descending stairs, rope jumping, lying down, and vacuuming, among others. Additionally, the dataset provides contextual information, such as heart rate, temperature, and participant ID, which can enhance classification.

To tackle the task of classification, we used machine learning techniques, feature engineering, and ML models. We pre-processed the raw sensor data, handled missing values, extracted relevant features, and constructed a comprehensive feature set to capture the intrinsic characteristics of each activity. Subsequently, we trained and fine-tuned several classification models, evaluating their performance using various metrics and cross-validation strategies.

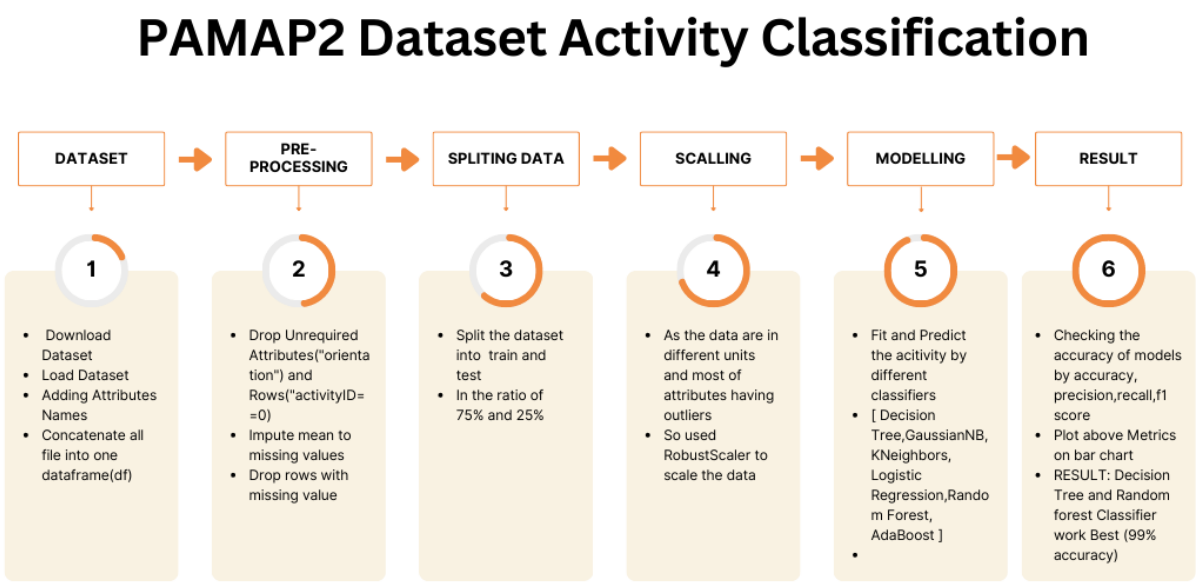
Accurate activity classification enhances our understanding of human behaviour. This enables personalized interventions and tailored recommendations in areas such as healthcare, sports performance, and rehabilitation. By analysing the patterns and dynamics of different activities, we can gain insights into the physiological and

biomechanical aspects associated with each task. This will facilitate the development of targeted interventions and improve well-being.

I provide a detailed description of our experimental setup, including the pre-processing steps, feature extraction techniques, and the selection of classification algorithms. I present an analysis of different models' performance. Furthermore, we highlight the challenges encountered during the classification process and suggest potential avenues for future research.

Overall, this report aims to contribute to the growing body of knowledge in the field of activity classification using the PAMAP2 dataset. By examining the effectiveness of various machine learning approaches, we seek to advance the development of accurate and practical activity recognition systems that can be applied in real-world settings, ultimately promoting healthier lifestyles and improved quality of life.

Here **Pictorial Representation** for the Project:



Dataset :

About Dataset:

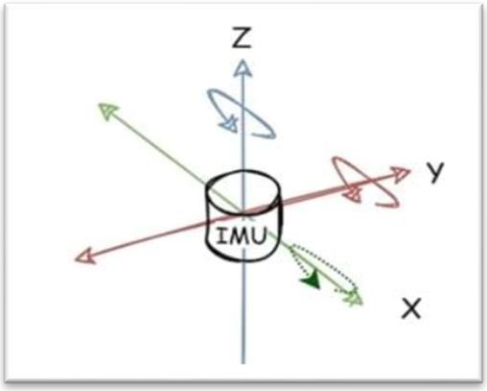
The PAMAP2 (Physical Activity Monitoring and Assessment System) dataset was developed by a team of researchers at the University of Twente in the Netherlands.

The dataset used is PAMAP2 which is an Activity Monitoring dataset that covers 18 different physical activities which are taken by 9 different subjects (8 men and 1 woman) taken using 3 inertial measurement units and a heart rate monitor. Different persons sensors data have different .data file (space separated). And some extra activity also done by subjects which are under Optional folder.(Not Used for this Project).

Total Dataset have `2872533 data, 54 Attributes.`

What is IMU sensors?

Inertial measurement units contain an accelerometer, gyroscope, and magnetometer. The accelerometer measures acceleration, while the gyroscope measures angular velocity. Each of these measurements is represented in a three-axis coordinate system.



- Sensor position:
1 IMU over the wrist on the dominant arm
1 IMU on the chest
1 IMU on the dominant side's ankle

The data files contain 54 columns: each line consists of a timestamp, an activity label, and 52 attributes of raw sensory data (from sensors devices) placed on hand, chest and ankle. (all information provided into readme.pdf file in dataset)

Data used in this notebook can be found and downloaded from:
https://archive.ics.uci.edu/ml/datasets/pamap2+physical+activity+monitoring

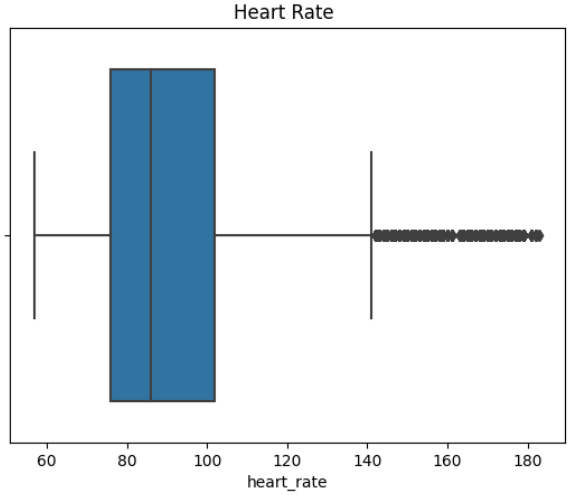
PREPROCESSING

After the dataset loading Next step will be get insights of the data and pre-process it .It can be seen that various data is missing and as the **readme** file comments on, there were some wireless disconnections in data collection therefore the missing data has to be accounted for and made up in a way that our data analysis will not be impacted.

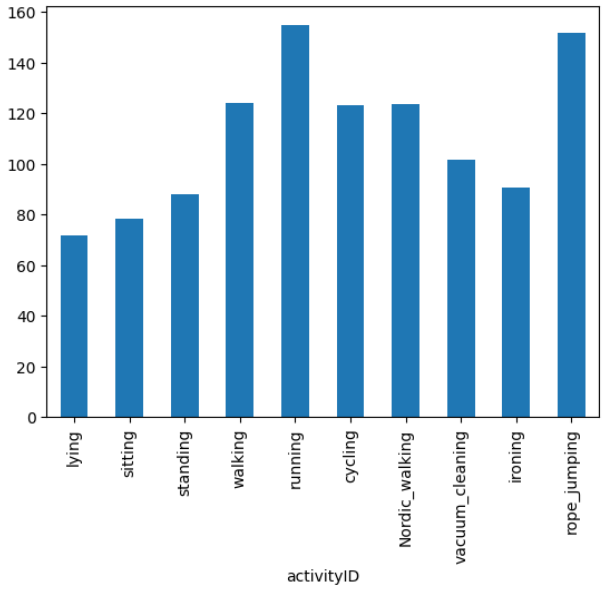
The raw data from the PAMAP 2 dataset, because it is realistic, has imperfections. There are 3 main issues to data.

- UNREQUIRED ATTRIBUTES OR ROWS
    - Dropping “activityID == 0”, since this is transient period where the subject was not doing any particular activity
    - removal of orientation columns as they are not needed for activity classification

- MISSING HEART RATES CELLS due to equipment malfunction
    - We are going to focus on heart rate as it is our most precise meter of check for tracking subjects during activities.
    - Heart Rate data are missing due of frequency rate of sensors
    - Heart rate box plot (has outliers). So directly not impute mean values



    - The bar chart shows that Rope Jumping and Running are the most cumbersome activities out of all the activities.Different Acitvity have specific and related heart rate(ofcourse).



- Then We impute mean to specific acitivityId and impute to that Id only.

- HAVING NULLS VALUES
  - The dataset has very less missing value compared to the whole recorded data. So, removing the rows having null values will be better choice.
  - Hence dropping rows with null values
  - After this data does NOT contains any missing values.

After solving these issues dataset is ready for machine learning model (dataset does have any missing value having all attributes are in int or float (no objects). After this next step will be to choose target values.

FEATURES MODELING

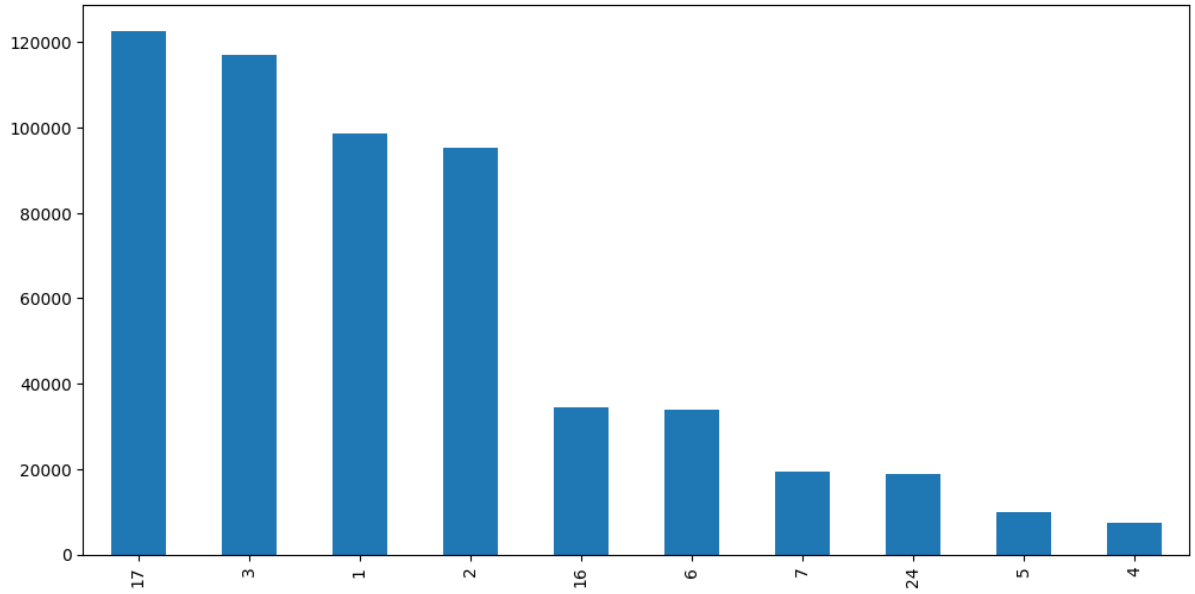
Selecting dependent attributes (Target values) for classification.

Classification is a supervised learning which have input data and target values to classify.

As we classifying the activity based on sensor data, So use activityId as Target or dependent attributes which depends on other independent attributes (sensory data).

Getting Target by dropping from dataframe.

Value count bar chart of Target.(Imbalance)



SPLITING DATA INTO TRAIN AND TEST

Now, we have input data and target data. Next will be to fit model but Model should be train and test on different data to avoid memorization.

So, Dividing the whole data into training and testing data (75% data for training and 25% for test).

We will fit model on training data and test the model accuracy on test data.

FEATURES SCALING

Standardization can become skewed or biased if the input variable contains outlier values. To overcome this, the median and interquartile range can be used when standardizing numerical input variables.

The attributes (timestamp, heart rate) are larger value than other values this can lead to biased on training model. To avoid this we Scaled the data by robustScaler.

Why robustScaler?

The dataset attributes have different units and contains **outliers**. And robustScaler is not affected by outliers.

The formula of RobustScaler is

$$X_{new} = \frac{X - X_{median}}{IQR}$$

Since it uses the interquartile range, it absorbs the effects of outliers while scaling.

The interquartile range (Q3 — Q1) has half the data point. If you have outliers that might affect your results or statistics and don't want to remove them, RobustScaler is the best choice.

MODELLING

After scaling, we apply classification model as our data seta have multi-class.

We will Apply different multiclass Classifer on training data include DecisionTreeClassifier, GaussianNB, KNeighborsClassifier,LogisticRegression, RandomForestClassifier, AdaBoostClassifier.

A brief about these models:

DecisionTreeClassifier: The DecisionTreeClassifier is a classification model that uses a decision tree algorithm to classify instances based on a set of predefined rules. It partitions the feature space into regions and assigns a class label to each region. It is known for its interpretability, as the decision tree structure allows us to understand the decision-making process.

GaussianNB: GaussianNB is a classification model based on the Gaussian Naive Bayes algorithm. It assumes that the features follow a Gaussian distribution and calculates the probability of an instance belonging to each class using Bayes' theorem. It is a simple and efficient model, especially for datasets with continuous features.

KNeighborsClassifier: The KNeighborsClassifier is a classification model that assigns a class label to an instance based on the class labels of its k nearest neighbors in the feature space. It is a non-parametric model that does not make assumptions about the underlying data distribution. KNN can be effective in capturing local patterns in the data and is relatively easy to understand and implement.

LogisticRegression: LogisticRegression is a widely used classification model that models the relationship between the input features and multiclass target variable using logistic functions. It estimates the probability of an instance belonging to a particular class and assigns the class label based on a decision threshold. Logistic regression is interpretable and works well with linearly separable data.

RandomForestClassifier: RandomForestClassifier is an ensemble model that combines multiple decision trees to make predictions. It uses a technique called bagging, where each tree is trained on a random subset of the training data. The final prediction is obtained by aggregating the predictions of individual trees. RandomForestClassifier is robust against overfitting, can handle high-dimensional data, and provides a feature importance measure

AdaBoostClassifier: AdaBoostClassifier is an ensemble model that combines weak classifiers in a sequential manner. It assigns higher weights to misclassified instances, allowing subsequent weak classifiers to focus on those instances. This iterative process leads to a strong classifier. AdaBoost is particularly effective in handling imbalanced datasets and can achieve high accuracy.

After fitting and the models, predict for the test data. We record or calculate

Accuracy, Precision, Recall and F1-score.

Accuracy, Precision, Recall, and F1-score are commonly used evaluation metrics for classification models. These metrics provide insights into the performance of a classification model in different aspects.

RESULT and CONCLUSION

In our study, we explored the effectiveness of various classification models for activity recognition on the PAMAP2 dataset. We trained and evaluated several models using different machine learning algorithms and techniques. Here are the results obtained from our experimentation:

Classifier: : DecisionTreeClassifier
Accuracy: 0.9993
Precision: 0.9993
Recall: 0.9993
F1-Score: 0.9993

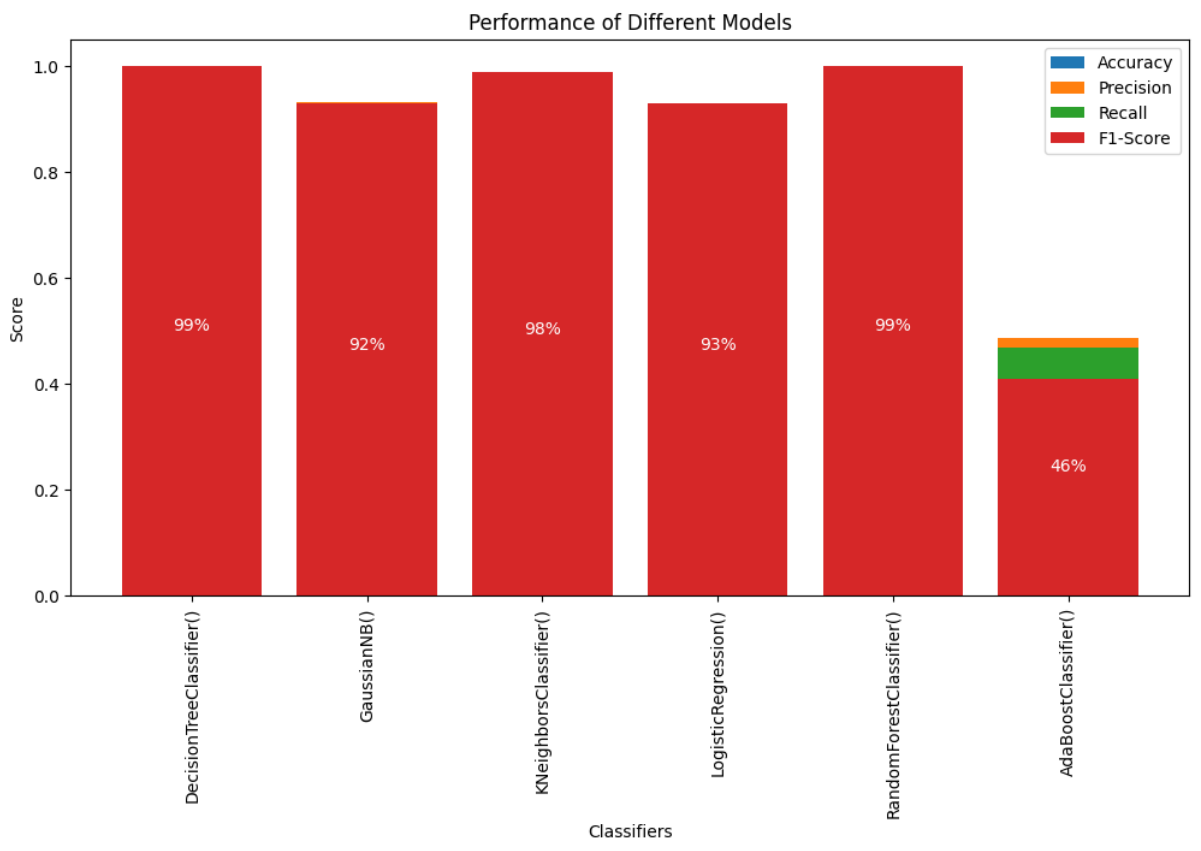
Classifier: naive_bayes.GaussianNB
Accuracy: 0.9277
Precision: 0.9326
Recall: 0.9277
F1-Score: 0.9286

Classifier: KNeighborsClassifier
Accuracy: 0.9886
Precision: 0.9887
Recall: 0.9886
F1-Score: 0.9886

Classifier: LogisticRegression
Accuracy: 0.9304
Precision: 0.9306
Recall: 0.9304
F1-Score: 0.9305

Classifier: RandomForestClassifier
Accuracy: 1.0000
Precision: 1.0000
Recall: 1.0000
F1-Score: 1.0000

Classifier: AdaBoostClassifier
Accuracy: 0.4690
Precision: 0.4874
Recall: 0.4690
F1-Score: 0.4085



In conclusion, our experiments on activity classification using machine learning models on the PAMAP2 dataset have yielded interesting results. Among the models tested, the Decision Tree Classifier, RandomForest Classifier, and KNN (K-Nearest Neighbors) showed promising performance.

Both the Decision Tree Classifier and RandomForest Classifier exhibited exceptional accuracy, achieving approximately 99%. This indicates that these models were able to accurately classify activities based on the provided dataset. The high accuracy suggests that these models are effective in capturing the underlying patterns and features that distinguish different activities.

On the other hand, the Adaboost Classifier did not perform well in our experiments, achieving only 46% accuracy. This suggests that the Adaboost model struggled to accurately classify activities based on the dataset. Further investigation may be required to understand the reasons behind its poor performance and identify potential improvements or alternative models.

These findings highlight the importance of carefully selecting and evaluating different machine learning models to achieve optimal results in activity classification tasks.

.

FUTURE WORK

Although this study provides valuable insights into activity classification on the PAMAP2 dataset, there are several avenues for future research that can further enhance the accuracy and applicability of activity recognition systems. Some potential areas for future work include:

1. Enhanced Feature Engineering: Exploring more advanced feature engineering techniques to capture the subtle nuances of different activities. This could involve extracting higher-level features or using advanced signal processing algorithms to extract more informative features from the sensor data.
2. AdaBoost poor performance: the Adaboost Classifier did not perform well in our experiments, achieving only 46% accuracy. This suggests that the Adaboost model struggled to accurately classify activities based on the given dataset.Can find reason for poor performance.

3. Deep Learning Approaches: Investigating the application of deep learning models, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), for activity classification on the PAMAP2 dataset. Deep learning models have shown promising results in various domains and could potentially improve the classification accuracy in this context.

In conclusion, this report presents a comparative analysis of different models for activity classification on the PAMAP2 dataset. The study highlights the performance of various classification algorithms and provides insights into their strengths and

weaknesses. Through experimentation, we have identified the most accurate models and evaluated their effectiveness in differentiating between activities.

Overall, this report serves as a comprehensive analysis of activity classification on the PAMAP2 dataset, offering valuable insights and recommendations for future research. It is hoped that the findings presented here will contribute to the development of innovative solutions in areas such as healthcare monitoring, sports performance analysis, and human-computer interaction, ultimately improving the quality of life for individuals through accurate activity recognition.

References:

- IMU Sensor https://builtin.com/internet-things/inertial-measurement-unit
- UCI dataset: https://archive.ics.uci.edu/ml/datasets/pamap2+physical+activity+monitoring
- Handling Missing Values in Machine Learning: https://towardsdatascience.com/7-ways-to-handle-missing-values-in-machine-learning-1a6326adf79e
- Data Preprocessing in Machine Learning: https://www.javatpoint.com/data-preprocessing-machine-learning
- RobustScaler in Scikit-learn: http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html
- Supervised Learning in Scikit-learn: https://scikit-learn.org/stable/supervised_learning.html
- OpenAI ChatGPT: https://openai.com/blog/chatgpt