Name = MD ASHRAF E-mail = mdashraf9723@gmail.com Intern ID = 21919235

TASK-1, SUMMARY

Detailed Approach and Observations

1. Introduction

The goal of this project was to classify human activities using raw accelerometer signals from the UCI HAR Dataset. We employed Traditional Machine Learning (ML) approaches by extracting features using TSFEL (Time Series Feature Extraction Library) and training three classifiers:

- Random Forest (RF)
- Support Vector Machine (SVM)
- Logistic Regression (LR)

The primary objective was to assess how well statistical features extracted from inertial sensor data contribute to activity recognition.

---

2. Methodology

2.1 Data Preprocessing

- The dataset consists of raw accelerometer readings (X, Y, Z) collected from a smartphone worn by participants.
- We specifically used the body acceleration signals from the Inertial Signals folder.
- The training and testing datasets were separately loaded, ensuring no data leakage.
- Labels (activity classes) were loaded and converted into categorical format for multi-class classification.

2.2 Feature Extraction Using TSFEL

Instead of feeding raw signals directly into ML models, we extracted statistical features using TSFEL, which provides an automated framework for time-series feature extraction. The process involved:

- Selecting statistical domain features, which analyze the signal's numerical distribution.
- Extracting features independently from each sample's time-series data.
- Filtering out samples shorter than a defined minimum length to ensure valid feature computation.

2.3 Handling Missing and Infinite Values

- Since some features might produce NaN (Not a Number) or infinite values, we applied np.nan_to_num() to replace them with finite numbers.

2.4 Feature Standardization

- Standardization was applied using StandardScaler() to normalize the extracted features, ensuring that models handle them uniformly.

---

3. Model Training & Evaluation

We experimented with three classical ML models, training them on the extracted feature set and evaluating performance using accuracy.

3.1 Random Forest (RF)

- **Accuracy: 59.28%**
- RF is an ensemble learning method that combines multiple decision trees to improve generalization.
- Performed the best, likely due to its ability to handle non-linearity and interactions between extracted features.

3.2 Support Vector Machine (SVM)

- **Accuracy: 53.58%**
- SVM attempts to find an optimal hyperplane to separate classes.
- Performed the worst, indicating that the extracted features might not be well-suited for linear separation.
- Feature selection or kernel tuning might be necessary to improve performance.

3.3 Logistic Regression (LR)

- **Accuracy: 57.28%**
- Logistic Regression, a linear classifier, performed better than SVM but worse than RF.
- Indicates that some linear relationships exist in the extracted features, but they alone are not sufficient for high accuracy.

---

4. Key Observations

1. Random Forest performed the best **(59.28%),** suggesting that tree-based models are better suited for handling feature interactions in this dataset.
2. Logistic Regression (**57.28%)** performed slightly better than SVM, showing that linear models capture some patterns but not all.
3. SVM had the lowest accuracy **(53.58%),** likely because the dataset features are not easily separable using a linear or even RBF kernel.
4. Feature extraction played a crucial role in model performance, and further improvements might be achieved by selecting more relevant features or using deep learning approaches.

---

5. Possible Improvements

To improve classification accuracy, the following strategies could be considered:

5.1 Feature Engineering

- Experimenting with other TSFEL domains such as temporal or spectral features.
- Using Principal Component Analysis (PCA) or feature selection techniques to reduce redundant features.

5.2 Model Optimization

- Tuning hyperparameters for RF, SVM, and LR using GridSearchCV or Bayesian Optimization.
- Trying Gradient Boosting methods like XGBoost or LightGBM for better feature handling.

5.3 Deep Learning Approaches

- Using a CNN+LSTM hybrid model, which has been successful in time-series-based classification.
- Training an end-to-end LSTM or Transformer-based architecture on raw accelerometer data instead of relying on feature extraction.

---

6. Conclusion

This study demonstrates that traditional ML models can be effective for activity recognition when combined with feature extraction.

- Random Forest was the most effective model, suggesting that ensemble methods can handle feature interactions better than linear models.
- The extracted statistical features provided moderate classification accuracy, but additional feature engineering could further improve results.
- Future work should explore deep learning methods, advanced feature selection techniques, and sensor fusion to enhance accuracy beyond 60% .