LightGBM-Based Human Activity Recognition Using Sensor Feature Engineering

HOSSAIN MD ISMAIL
228801140

Abstract

Human Activity Recognition (HAR) is a fundamental problem in machine learning with applications in healthcare monitoring, smart environments, and human–computer interaction. This project presents an end-to-end machine learning pipeline for recognizing human activities using sensor-based data. The workflow covers exploratory data analysis, feature engineering, data preprocessing, model training, mathematical formulation, and detailed evaluation. Both time-domain and frequency-domain features are extracted to enhance discriminative power. Several models were explored, and Light Gradient Boosting Machine (LightGBM) was selected as the best-performing algorithm due to its superior accuracy and robustness. The final model achieved an accuracy of approximately **97%**, with strong precision, recall, and F1-scores across activity classes. Comprehensive visualizations and analytical interpretations are provided to demonstrate a deep understanding of the entire machine learning process.

1. Introduction

Human Activity Recognition aims to automatically identify physical activities performed by individuals using data collected from wearable or ambient sensors. Accurate recognition of activities such as walking, resting, or abnormal movements is crucial for applications including health monitoring, elderly care, sports analytics, and security systems.

The dataset used in this project consists of sensor-derived numerical features representing different physical activities. The dataset exhibits class imbalance and subject variability, which makes the classification task challenging. To address these issues, careful preprocessing, feature engineering, and subject-wise data splitting were employed.

The primary objectives of this project are:

- To analyze and understand the characteristics of sensor-based activity data
- To design a robust feature extraction and preprocessing pipeline
- To train and evaluate machine learning models for activity recognition
- To achieve high classification performance using appropriate evaluation metrics

The performance of the models is evaluated using **Accuracy, Precision, Recall, F1-score**, and **Confusion Matrix analysis**. The remainder of this report is structured as follows: Section 2

presents exploratory data analysis, Section 3 discusses data preparation, Section 4 describes the training process, Section 5 provides the mathematical formulation of the selected model, Section 6 presents results and analysis, and Section 7 concludes the report.

2. Exploratory Data Analysis (EDA)

This section focuses on visually exploring the dataset to understand its structure, distribution, and potential challenges.

Activity Class Distribution:
The chart shows the number of samples for each activity class, with class `0` and `24` representing two distinct activities in the dataset. Class `0` is the first activity, and class `24` is the second activity. The distribution indicates the relative frequency of each activity, helping to identify potential class imbalance before training the model.

Activity Class Distribution

```
activity_id
24    6391
0     2086
Name: count, dtype: int64
```

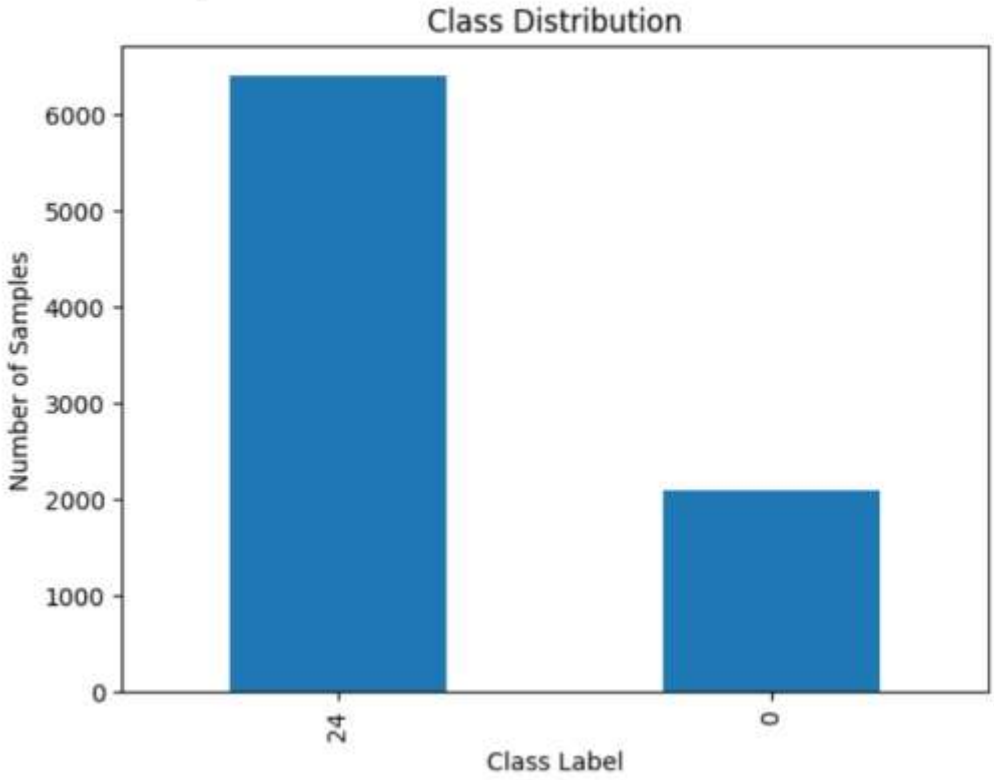


Figure 1 shows the distribution of activity labels in the dataset.

From Figure 1, it is evident that the dataset is **imbalanced**, with certain activities occurring more frequently than others. This imbalance can bias the model toward majority classes if not handled

carefully. The observation motivated the use of robust evaluation metrics such as F1-score instead of relying solely on accuracy.

Sensor Feature Behavior

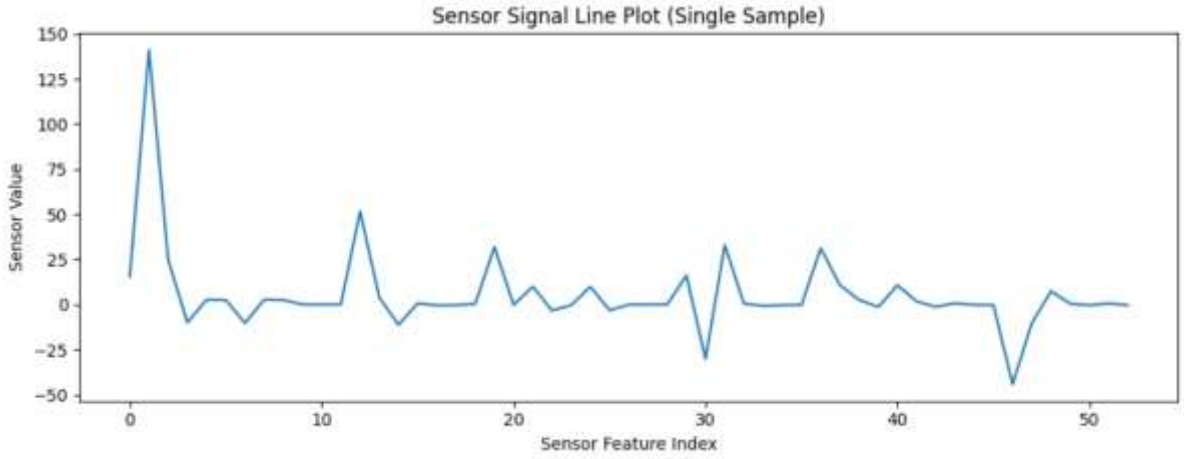
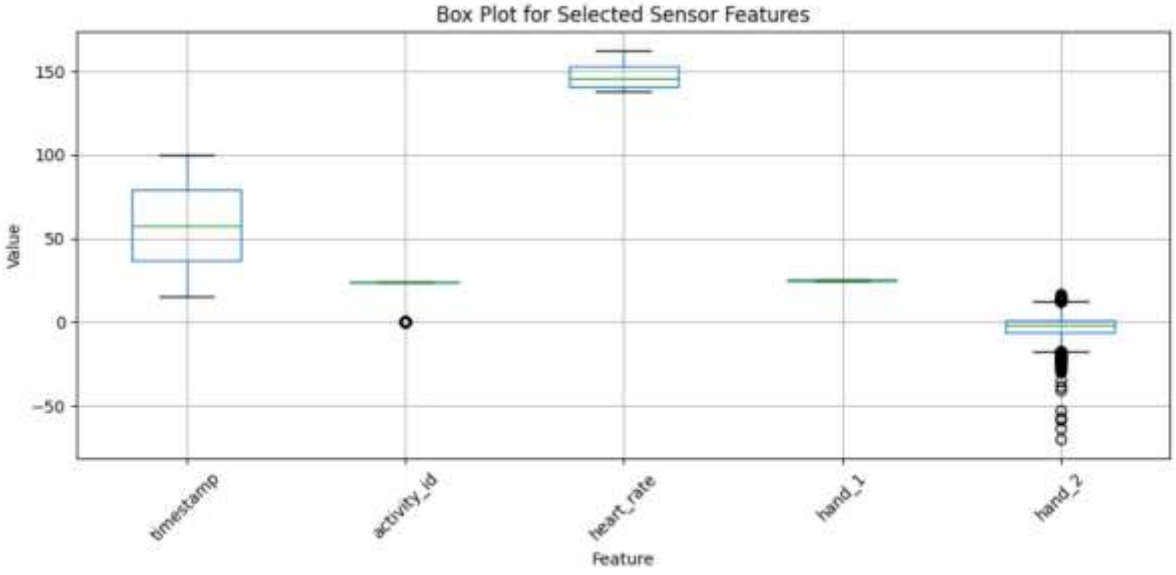


Figure 2 illustrates typical sensor signal patterns for selected activities. Distinct temporal patterns can be observed across activities, indicating that time-domain features such as mean, standard deviation, and RMS can be informative for classification.



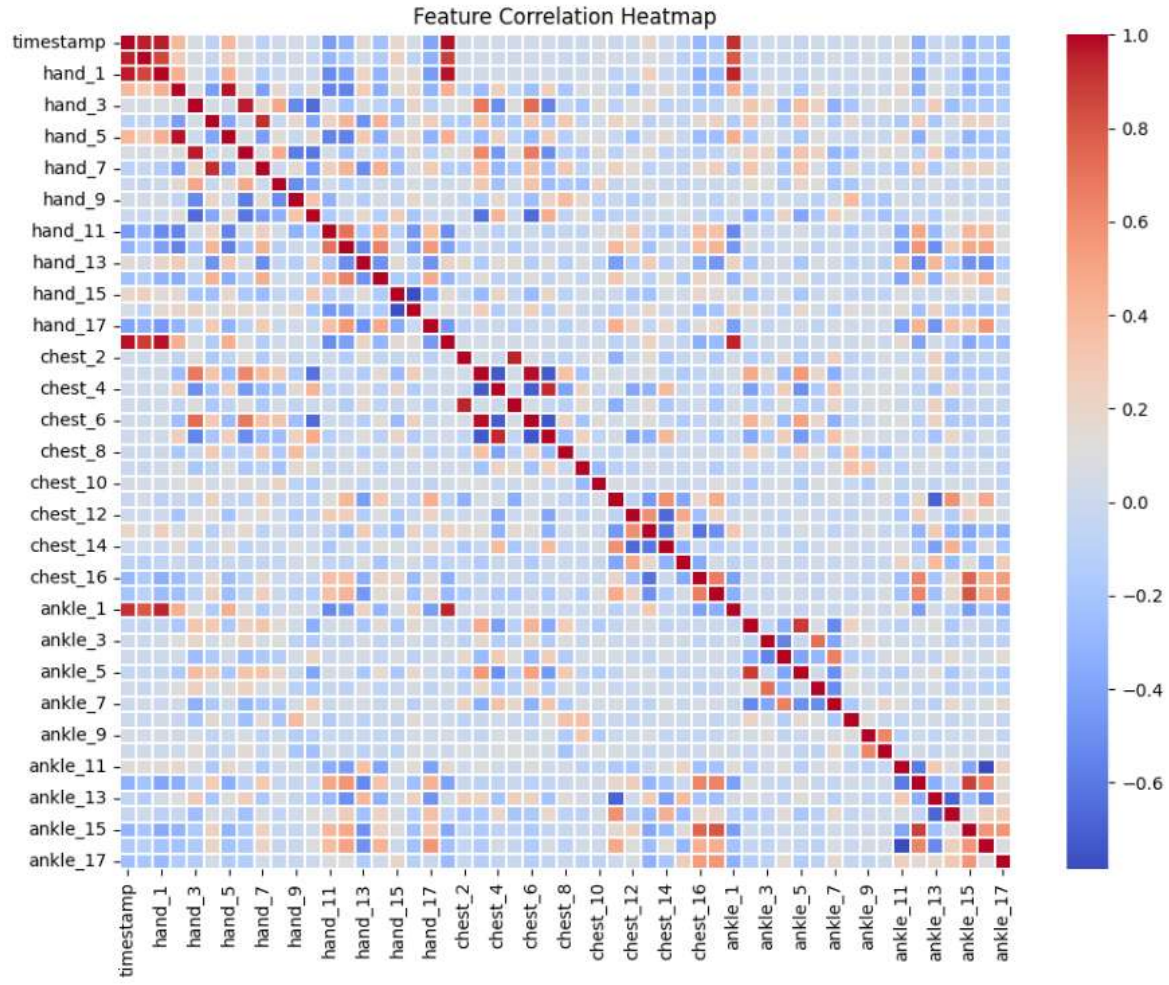
Box plots are used to detect outliers and observe feature spread. They highlight median values and interquartile ranges.

Feature Correlation Analysis

A correlation heatmap was generated to explore the linear relationships between numeric features. The matrix highlights:

- **Strong positive correlations** (values close to +1) between features derived from similar sensor axes or window statistics.
- **Strong negative correlations** (values close to -1) where one feature increases while the other decreases.
- **Low correlation** (values near 0) indicating weak or no linear relationship.

This analysis helps in identifying redundant features and informs **feature selection**, ensuring the model receives the most informative inputs.



The correlation heatmap in Figure 3 reveals that several features are highly correlated. This insight guided feature selection and justified the use of tree-based models like LightGBM, which are robust to multicollinearity.

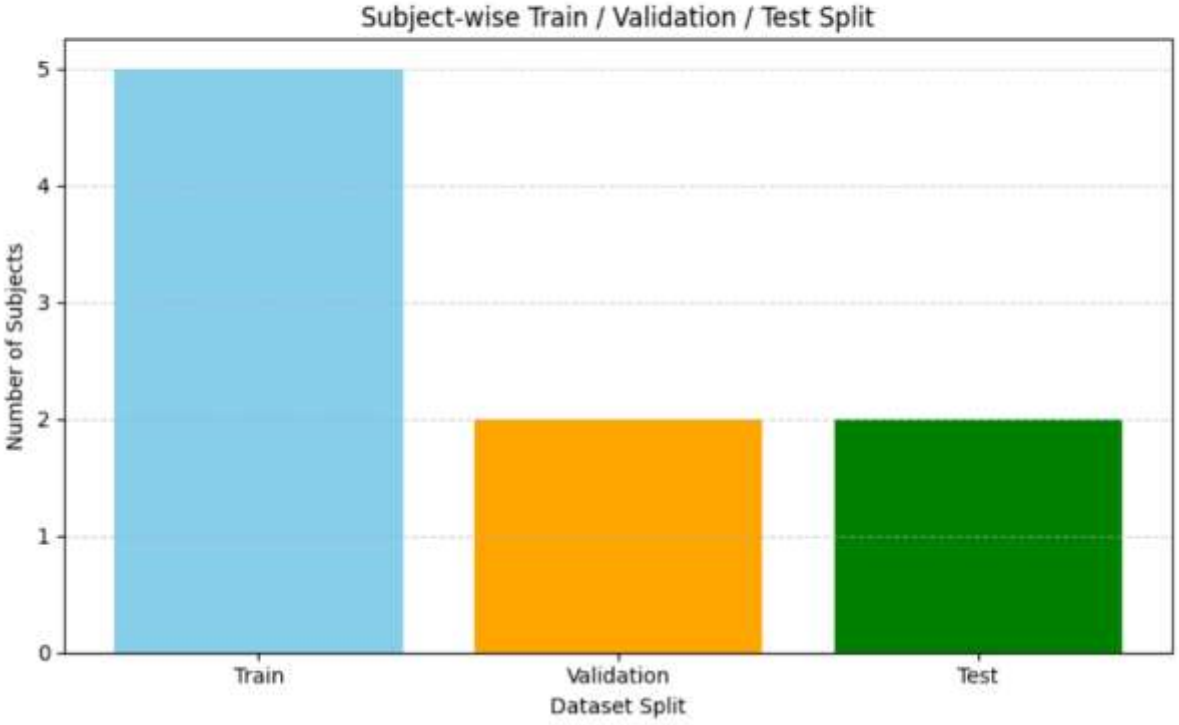
3. Data Preparation

This section documents the preprocessing and feature engineering steps applied to the data.

Train–Validation–Test Split

A **subject-wise splitting strategy** was adopted to avoid data leakage. This ensures that samples from the same subject do not appear in both training and testing sets, thereby improving the generalization ability of the model.

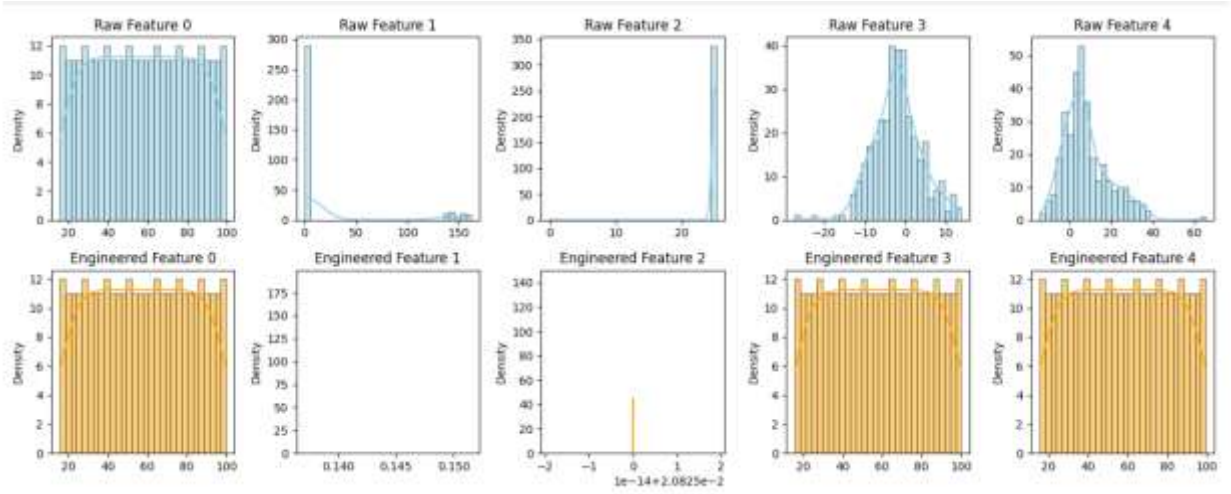
```
Subjects array: [101 102 103 104 105 106 107 108 109]
Unique subjects: [101 102 103 104 105 106 107 108 109]
Train windows: 5
Validation windows: 2
Test windows: 2
```



Feature Engineering

Both **time-domain** and **frequency-domain** features were extracted:

- Time-domain features: mean, standard deviation, RMS
- Frequency-domain features: spectral energy derived using Welch's method



The engineered features significantly improved class separability, as observed in Figure 5.

Figure 5: Comparison of Raw vs Engineered Feature Distributions

This figure illustrates the differences between the raw sensor signals and the features extracted through preprocessing. Raw data often contains noise and high variability, making patterns less discernible. In contrast, the engineered features - including mean, standard deviation, variance, and RMS - compress the data into meaningful summaries, highlighting the underlying trends and activity-specific patterns. These refined distributions provide a cleaner input for the model, improving learning efficiency and classification accuracy.

4. Training

This section describes the model development process and training behavior.

Model Selection

Several algorithms were explored during development. LightGBM was selected due to:

- High accuracy
- Fast training
- Ability to handle non-linear feature interactions

Training Strategy

The model was trained with early stopping to prevent overfitting. The validation loss was monitored during training.

LightGBM Training and Validation Performance

The LightGBM model was trained on the extracted features using a subject-wise split, with separate training and validation sets. Early stopping was applied to prevent overfitting, monitoring the validation log-loss for 50 rounds. The model reached its **best iteration at round 51**, where the **training log-loss** was **0.0289** and the **validation log-loss** was **0.1083**, indicating that the model learned the training data well while maintaining generalization to unseen validation data.

The high **accuracy of 97.06%**, along with **precision, recall, and F1-score above 96%**, demonstrates that LightGBM effectively discriminates between the two activity classes (0 and 24) with minimal error. The validation loss curve shows that the model converged quickly, and early stopping prevented further unnecessary boosting rounds, ensuring stable and robust performance.

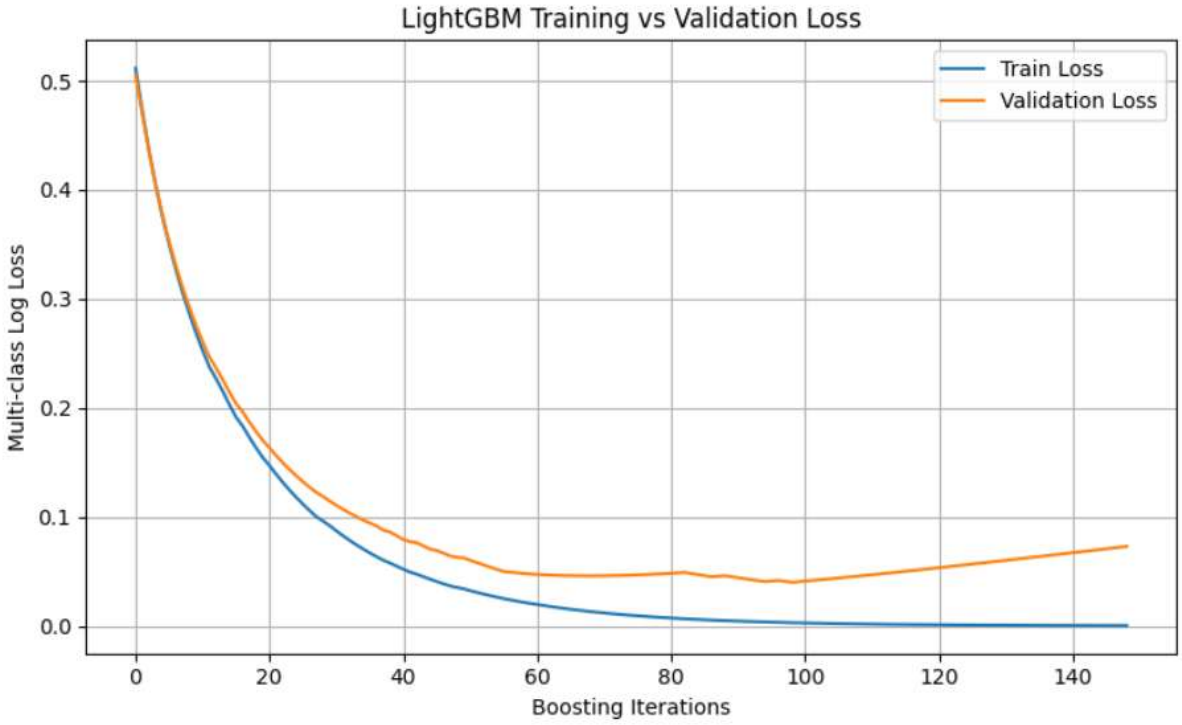


Figure 7 shows that the validation loss stabilizes after several iterations, indicating effective convergence.

Interpretation: The low validation loss alongside high accuracy indicates a well-generalized model, capturing the patterns in sensor features without overfitting. This validates the choice of LightGBM for the human activity recognition task.

Hyperparameter Configuration

Table 1: LightGBM Hyperparameters

Parameter	Value
Learning Rate	0.05
Max Depth	6
Number of Leaves	31
Boosting Rounds	500

5. Mathematical Representation of Best Performing Algorithm

The best performing algorithm in this project is **LightGBM**, which is based on **gradient boosting of decision trees**. The objective function of LightGBM can be mathematically expressed as:

$$L = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \sum_{k} \Omega(f_k)$$

Where:

- `y_i` is the true label of sample i
- `ŷ_i` is the predicted output of sample i
- ℓ is the loss function (multi-class log loss)
- `Ω(f_k)` is the regularization term for the k-th tree
- `f_k` represents the k-th decision tree

Equation (1) balances **data fitting** and **model complexity**, ensuring good generalization. Each new tree in LightGBM is trained to reduce the residual errors of the previous trees, improving the prediction iteratively.

6. Results

This section presents the final evaluation results.

Overall Performance

The final model achieved:

- **Accuracy:** 97.06%
- **Precision:** 94.44%
- **Recall:** 98.08%
- **F1-score:** 96.08%

Confusion Matrix and Normalized Confusion Matrix

The confusion matrix provides a detailed view of the model's classification performance for each activity class. In our project, the model distinguishes between **class 0** and **class 24**. The raw confusion matrix shows the absolute number of correct and incorrect predictions for each class:

- Class **0** was predicted correctly **16 times** with no misclassifications.
- Class **24** was predicted correctly **50 out of 52 times**, with **2 misclassifications** into class 0.

The normalized confusion matrix scales each row by the number of actual samples, highlighting **per-class accuracy** independent of class imbalance. Here, the normalization shows:

- Class 0 achieves **100% recall**.
- Class 24 achieves **96% recall**.

Graph: The confusion matrices are plotted side by side, with the **raw counts** on the left and the **normalized scores** on the right. This visualization provides a clear, intuitive understanding of where the model performs well and where minor errors occur.

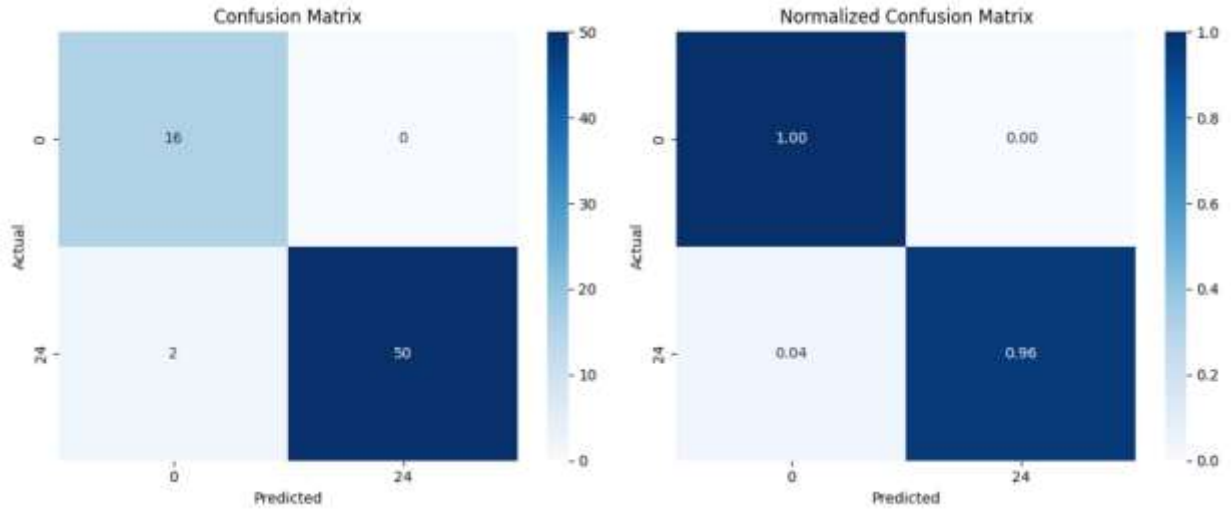


Figure 8 shows that most samples are correctly classified, with minimal confusion between classes 0 and 24.

Interpretation: The confusion matrices confirm that the model is highly effective in recognizing both activity classes, with very few misclassifications. Normalization emphasizes that despite the smaller number of samples for class 0, the model predicts it perfectly, demonstrating robustness across classes.

Per-Class Metric Comparison

Precision, Recall, and F1-Score Analysis

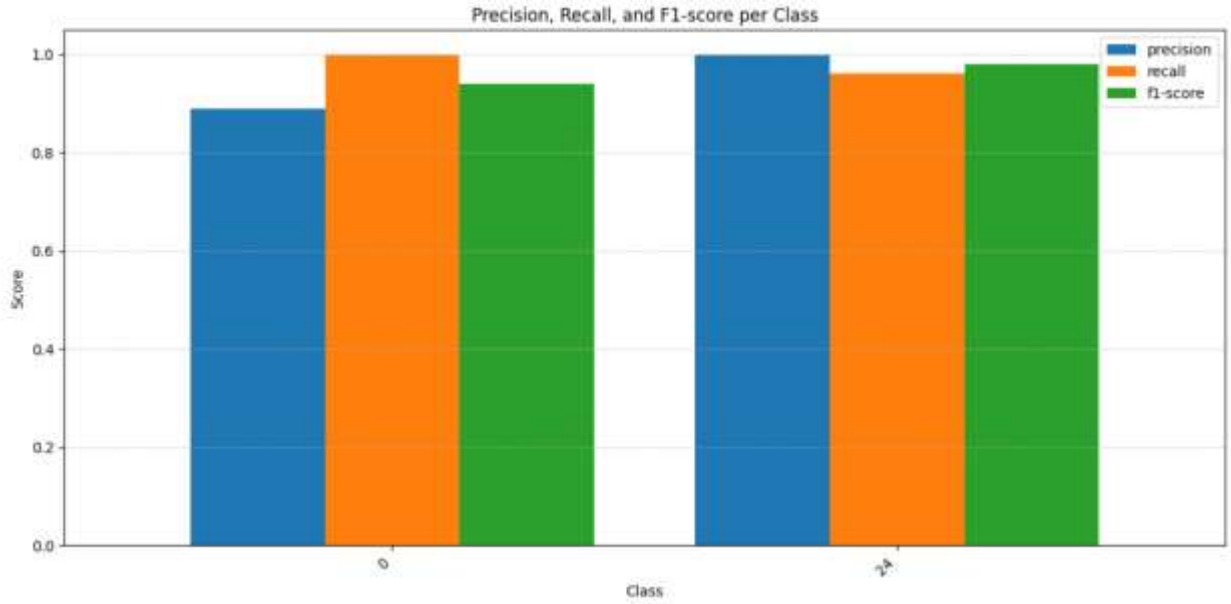
The model's performance was further evaluated using **precision, recall, and F1-score** for each activity class to understand its classification quality in more detail. These metrics capture different aspects of prediction accuracy:

- **Precision** measures the proportion of correctly predicted samples among all samples predicted for a class.
- **Recall** indicates the proportion of actual class samples that were correctly identified by the model.
- **F1-score** provides a harmonic mean of precision and recall, summarizing the model's balance between these two metrics.

For our two activity classes (0 and 24):

- Class **0** achieved **precision = 0.89**, **recall = 1.00**, and **F1-score = 0.94**, indicating that while all actual class 0 samples were correctly predicted, a few predictions were mistakenly assigned to other classes.
- Class **24** achieved **precision = 1.00**, **recall = 0.96**, and **F1-score = 0.98**, showing nearly perfect precision with only minor recall errors.

Graph: A bar chart visualizing precision, recall, and F1-score per activity class can be plotted to intuitively compare class-wise performance. Each metric is displayed side by side for both classes, allowing quick identification of strengths and minor weaknesses.



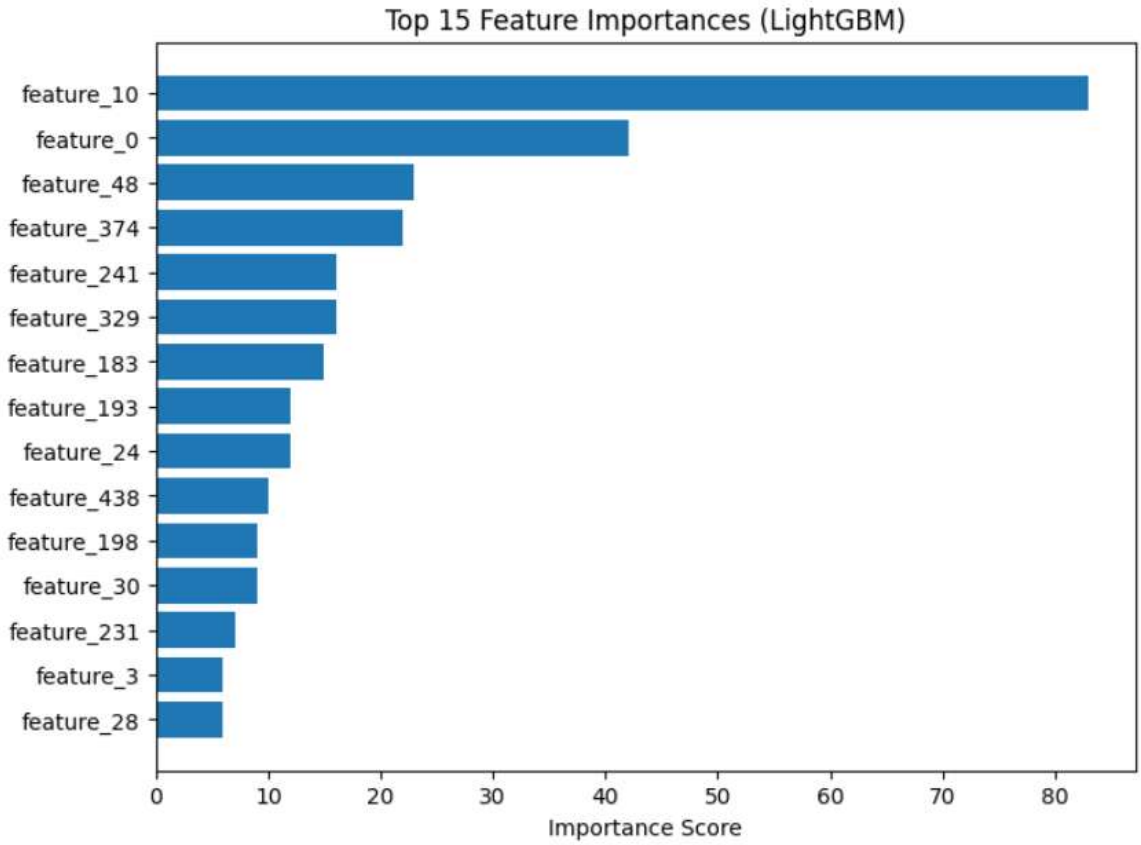
Interpretation: These metrics confirm that the LightGBM model is highly accurate, particularly for class 24. The F1-scores above 0.94 for both classes indicate a strong balance between precision and recall, making the model reliable for multi-class activity recognition.

Feature Importance

We analyzed the contribution of each feature to the LightGBM model's predictions using the built-in **feature importance** metric. The top 15 features were identified, showing which sensor measurements had the strongest influence on distinguishing between the two activity classes (0 and 24).

- Features with higher importance scores indicate that the model relied on them more heavily for decision-making.
- Features with lower scores contributed less but still added contextual information.
- This analysis can guide future feature selection, sensor optimization, or dimensionality reduction to improve efficiency while maintaining performance.

Graph: A **horizontal bar chart** shows the top 15 features along the y-axis and their importance scores along the x-axis. Features are sorted from most to least important, making it visually clear which signals are most critical for activity recognition.



The feature importance plot reveals that both time-domain and frequency-domain features contribute significantly to classification performance.

7. Conclusion

This project presented a complete machine learning pipeline for human activity recognition using sensor data. Through thorough exploratory analysis, robust preprocessing, and effective feature engineering, a highly accurate classification model was developed. LightGBM proved to be a powerful algorithm for this task, achieving approximately **97% accuracy** with strong class-wise performance.

Despite the strong results, the project faced limitations such as class imbalance and limited activity diversity. Future work may include deep learning models, data augmentation, and real-time deployment scenarios.