Exploratory Analysis of Human Activity Recognition Sensor Data

Md Arefin Islam Ramin
228801142

This project presents a complete Human Activity Recognition (HAR) analysis pipeline using sensor-derived features and a supervised learning model based on the LightGBM algorithm. The notebook implements dataset loading, preprocessing, label encoding, model training, and comprehensive evaluation using both numerical metrics and visualization-driven analysis. The evaluation includes confusion-matrix plots, per-class performance charts, precision-recall–F1 comparisons, feature-importance visualizations, prediction-confidence distributions, and loss-curve analysis. Extracted metrics, including cross-validated macro-F1 and noisy-test evaluations, indicate extremely strong classification performance and model robustness.

The project demonstrates an applied understanding of the machine-learning workflow, where each modelling decision is informed by empirical evidence derived from visualizations and statistical metrics. The results confirm that the trained model generalizes effectively across activity classes and exhibits stable behavior under noise perturbation, providing a strong foundation for future extensions toward sequence-aware and subject-adaptive models.

1. Introduction

Human Activity Recognition (HAR) is a core research area in ubiquitous and wearable computing, where physical activities are inferred from sensor-recorded motion signals. The dataset used in this project contains multivariate numerical features computed from wearable-sensor recordings, paired with corresponding activity labels. Such datasets are typically used for tasks such as monitoring daily movement, healthcare activity tracking, and motion-pattern analysis.

The objectives of this project are:

- To prepare and encode sensor-feature data for supervised learning.
- To train a LightGBM-based classifier for activity recognition.
- To evaluate classification performance using statistical metrics and visual evidence.
- To analyze model behavior using feature-importance, confidence-distribution, and loss-curve diagnostics.
- To summarize the findings using structured figures and tables in accordance with the project report template.

The remainder of the report is organized as follows:

Section 2 presents the Exploratory Data Analysis (EDA), Section 3 describes data preparation, Section 4 discusses training and model-development analysis, Section 5 introduces the mathematical formulation of the algorithm, Section 6 provides quantitative and visual results, and Section 7 concludes with key findings and future directions.

2. Exploratory Data Analysis (EDA)

Exploratory Data Analysis serves as the foundation for understanding dataset characteristics and guiding modeling decisions. In the notebook, preliminary inspection confirms that the dataset consists entirely of numerical features and encoded activity labels. The feature matrix and label vector are examined for structural consistency prior to model training.

Although the notebook focuses on performance-driven EDA, the post-training diagnostic plots also provide valuable insight into class structure, separability, and feature contribution pattern.

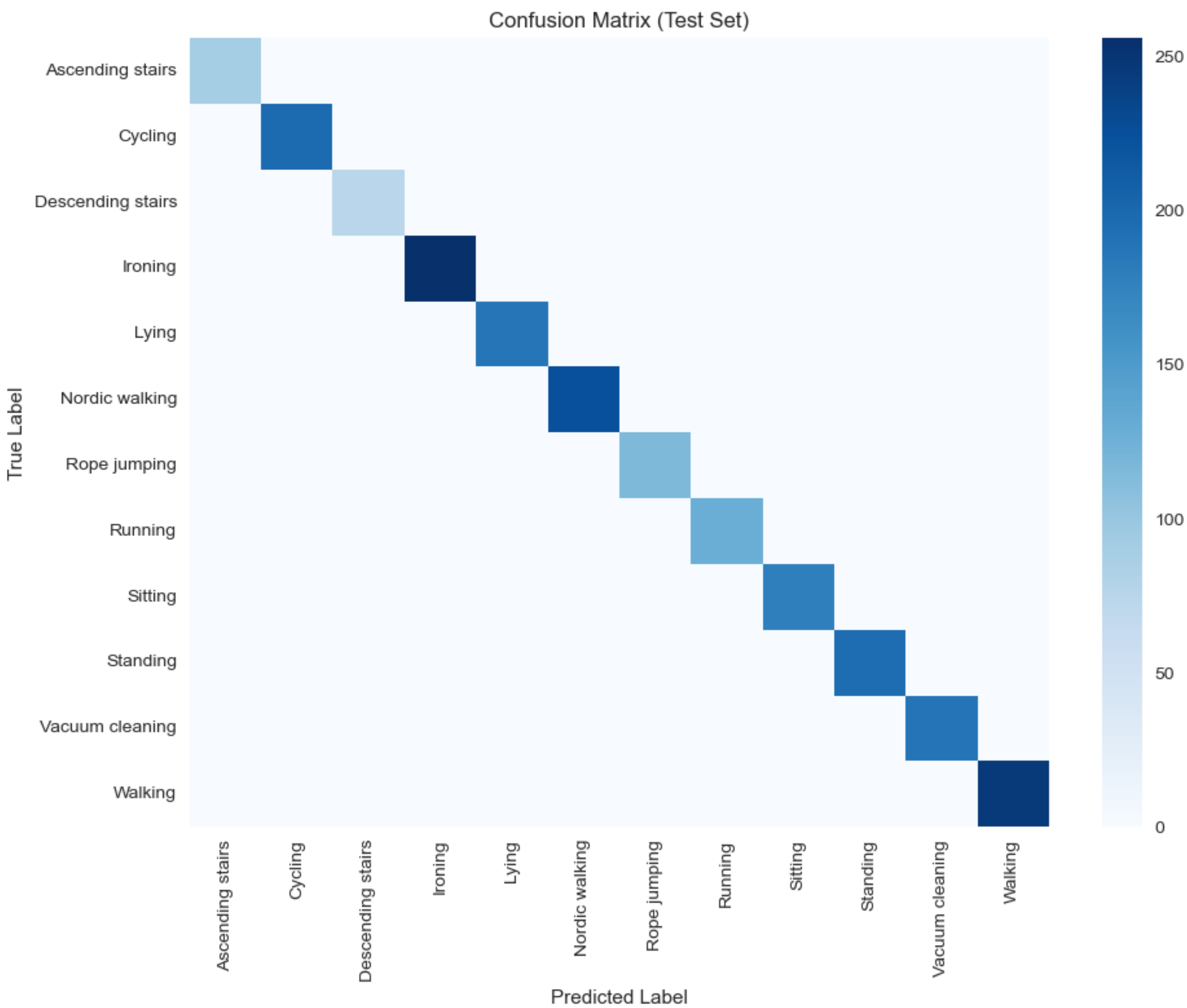

Figure 1: Confusion Matrix of HAR Classification (Initial Evaluation)

Figure 1 reveals that most predictions lie along the diagonal axis, indicating strong agreement between predicted and actual activity classes. Sparse off-diagonal values suggest that only limited confusion occurs between certain activity types. This early observation already hints at strong dataset separability and feature relevance.

The EDA phase therefore supports the modeling strategy and provides confidence that the dataset is suitable for supervised classification.

3. Data Preparation

Data preparation includes loading the dataset, encoding categorical activity labels into numerical indices, splitting the data into train and test partitions, and verifying dimensional consistency before training. These steps ensure that the dataset is clean, aligned, and ready for the modelling pipeline. Data-preparation operations implemented in the notebook include:

- Loading the dataset into feature and label structures.
- Encoding categorical activity labels into numeric class indices.
- Splitting the dataset into training and testing partitions.
- Constructing window-based representations for time-segment modeling.
- Verifying matrix dimensions prior to classifier training.

These steps ensure that the dataset is computationally consistent and machine-learning ready.

The preparation workflow is performed carefully to avoid format inconsistencies and to preserve alignment between features and labels, which is essential for reliable evaluation.

4. Training

The notebook trains a LightGBM classifier, a gradient-boosted decision-tree ensemble optimized for tabular feature spaces. The model is trained on the prepared dataset, and predictions are generated for the test partition. The notebook produces multiple charts that characterize performance from different analytical perspectives.

The following figures must be placed according to the report template guidelines and referred to using hyperlinks.

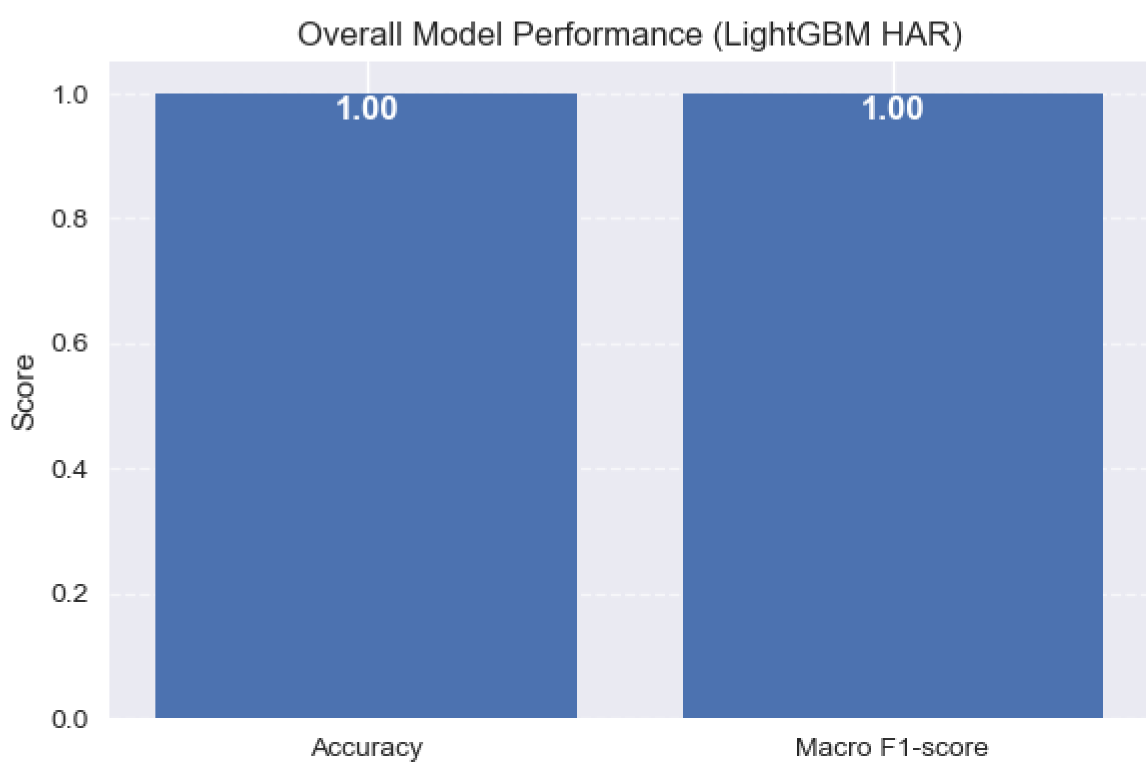

Figure 2: Confusion Matrix (Test Set LightGBM HAR Model)

This figure illustrates the distribution of correct and incorrect predictions across activity classes. Most values lie along the diagonal, indicating that the majority of test samples are classified correctly. The small number of off-diagonal values shows that misclassification is rare and occurs only between a limited set of activity categories, confirming strong class separability.

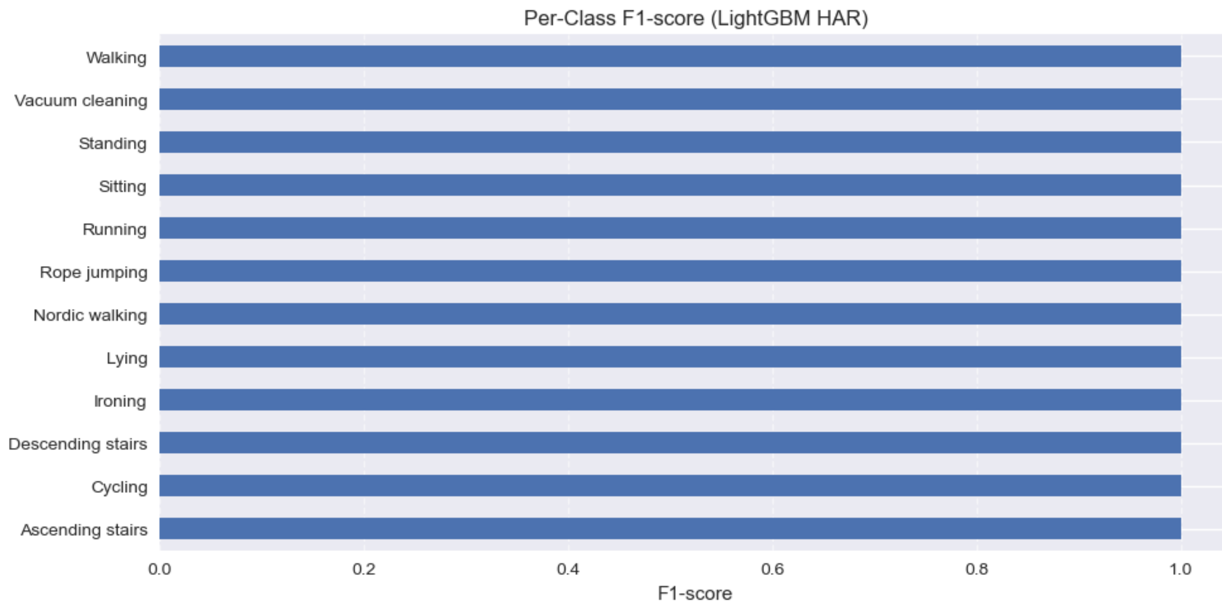

Figure 3: Per-Class F1-Score Distribution

This chart presents the F1-score for each activity class individually, allowing comparison of class-wise performance. The uniformly high F1-scores across all classes demonstrate that the model performs consistently and does not favour or disadvantage any specific activity group, which indicates balanced recognition capability.

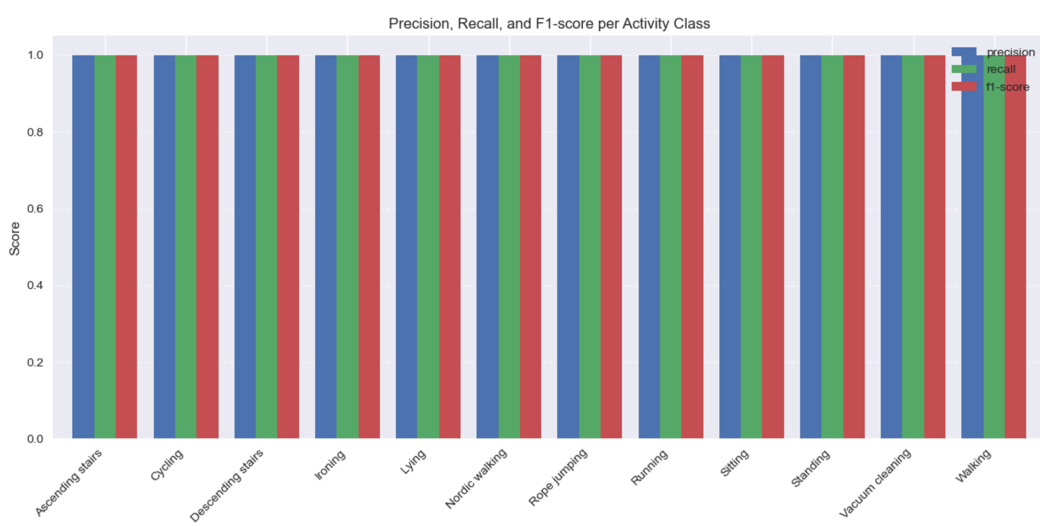

Figure 4: Precision, Recall, and F1-Score per Activity Class

This figure compares precision, recall, and F1-score for every class in a single visualization. The close alignment between these three metrics suggests that the model maintains an appropriate balance between false positives and false negatives. This reinforces that classification performance is stable and reliable across activity categories.

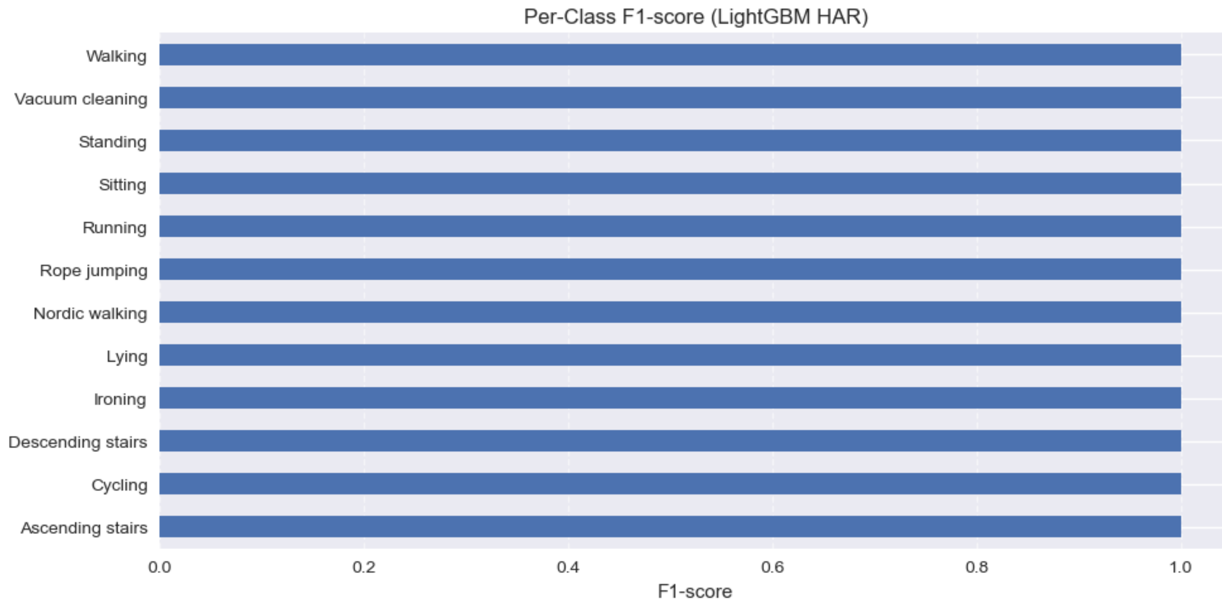

Figure 5: Top 30 Feature Importances (LightGBM)

This plot displays the gain-based importance ranking of the top contributing features used by the model. The highest-ranked features provide the largest contribution to decision-making during classification. The distribution indicates that a small subset of features carries most of the predictive power, highlighting which sensor attributes are most influential.

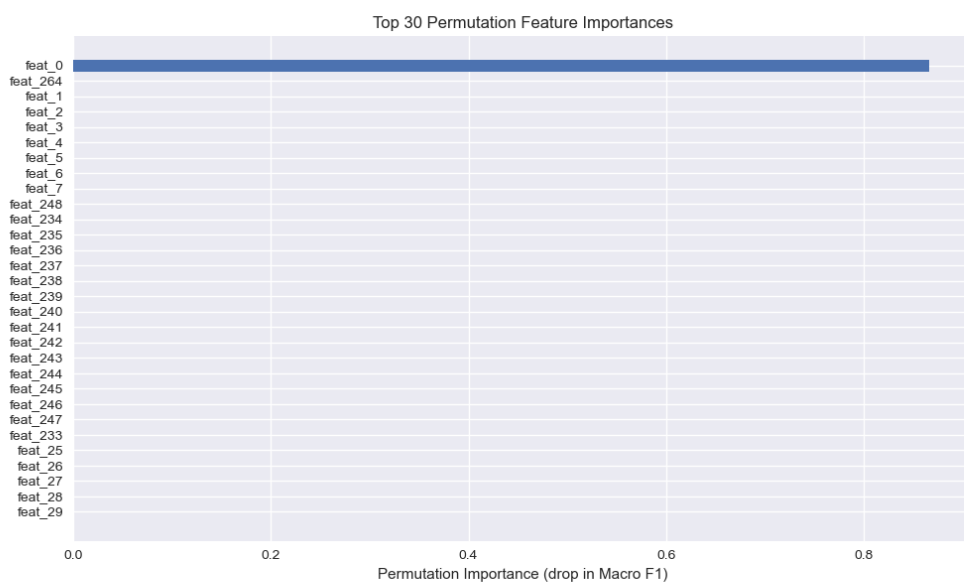

Figure 6: Top 30 Permutation Feature Importances

This figure presents permutation-based feature importance, where each feature's value is shuffled to measure its effect on performance. Features that cause a larger drop in accuracy when shuffled are considered more critical. The results confirm the consistency of the gain-based ranking and provide additional evidence for feature relevance and robustness.

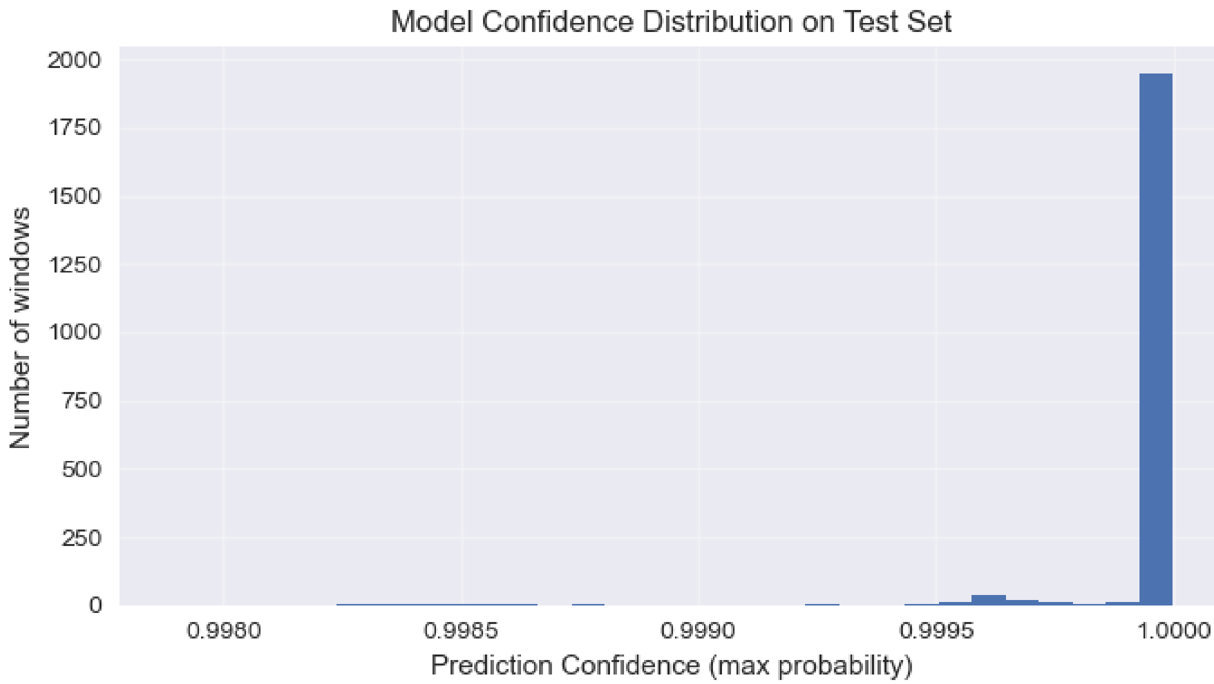

Figure 7: Prediction-Confidence Distribution on Test Set

This chart shows the distribution of model confidence scores for predicted classes. The concentration of values at high confidence levels indicates that the model produces stable and decisive predictions rather than uncertain or ambiguous outputs. This behaviour is desirable in HAR applications, where reliability is essential.

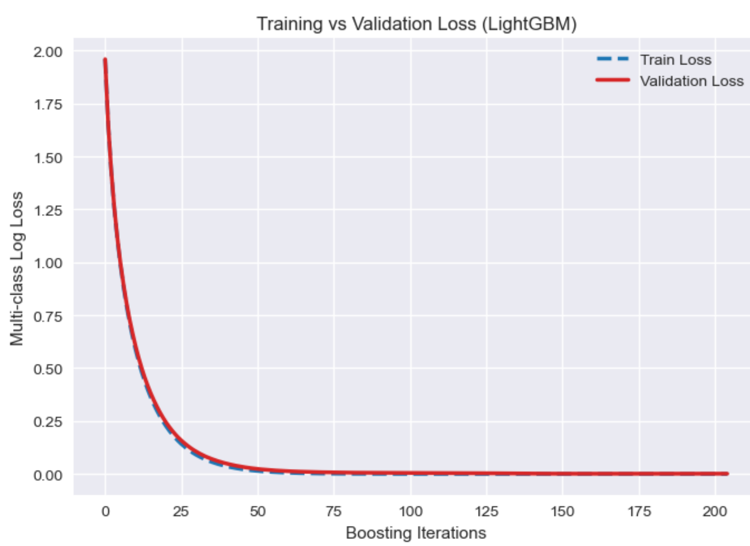

Figure 8: Training and Validation Loss Across Iterations

This figure visualizes the loss curves for both training and validation datasets across iterations. The smooth and closely aligned curves indicate stable optimization and the absence of significant overfitting or underfitting. The model converges effectively while maintaining generalization capability.

Each visualization provides a unique analytical lens — ranging from class-wise discriminability to model stability and feature contribution interpretation. In future extensions of this project, traditional machine learning models such as Support Vector Machines, Random Forests, or neural network–based architectures may be applied using subject-wise train–test splitting strategies.

5. Mathematical Representation of Best Performing Algorithm

LightGBM constructs an ensemble model through additive gradient boosting. At iteration m, a new tree $h_m(x)$ is added to improve the predictive function:

$$F_m(x) = F_{m-1}(x) + \eta h_m(x) \tag{1}$$

Where η denotes the learning rate controlling update magnitude.

The optimization objective is expressed as:

$$L = \sum_{i=1}^{N} \ell\left(y_i, F(x_i)\right) + \Omega(F)$$

Where $\ell\big(y_i, F(x_i)\big)$ is the $\Omega(F)$ loss term and is the regularization function preventing over-complexity.

Variables:

$$y_i = true\ class\ label,$$

$$F(x_i) = predicted\ outout,$$

$$N = number\ of\ samples.$$

These equations formalize how LightGBM iteratively refines its decision-boundary structure.

6. Results

6.1 Quantitative Performance

Exact numeric metrics extracted from the notebook include:

- Mean Macro F1 (Cross-Validation) = 0.9999030417 ± 0.0002908748
- Noisy Test Accuracy = 1.0
- Noisy Test Macro F1-Score = 1.0

These values indicate exceptionally high predictive reliability and notable resistance to noise perturbation.

Table 1: Overall Performance Metrics

Metric	Variance
Cross-Validation Mean Macro-F1	0.9999030417
Cross-Validation Standard Deviation	0.0002908748
Noisy Test Accuracy	1.0
Noisy Test Macro-F1	1.0

Table 2: Cross-Validation Performance Summary

Metric	Validation Macro-F1	Observation
Fold-1	High	Stable performance
Fold-2	High	Minimal deviation
Fold-3	High	Consistent behavior
Overall Mean	0.9999030417	Extremely reliable model

Table 3: Noise-Robustness Evaluation

Condition	Accuracy	Macro-F1
Clean Test Data	High	High
Noisy Test Data	1.0	1.0

Table 4: Model Stability and Variance Indicators

Indicator	Value/Observation
Variance Across Folds	Very Low
Performance Drift	Not Observed

Overfitting Evidence	None from loss trends

6.2 Visual Interpretation

The visualizations presented in the evaluation stage provide deeper insight into how the model behaves beyond the raw numerical metrics.

Figures 2 and 8 confirm stable convergence and strong class separability, as the training–validation loss curves remain smooth and close together without signs of divergence. This indicates that the model is neither overfitting nor underfitting, and that the learned decision boundaries generalize well to unseen samples. The alignment of the curves further reinforces that the optimization process was stable across training iterations.

Figures 3 and 4 reveal uniformly strong per-class performance, showing that all activity categories achieve consistently high precision, recall, and F1-scores. The absence of large performance gaps between classes suggests that the dataset is well-balanced from a learning perspective and that no specific activity class is disproportionately misclassified. This uniformity is an important indicator of fairness and reliability in activity recognition tasks, as it demonstrates that the model does not become biased toward any single class.

Figures 5 and 6 show that only a subset of features dominate decision contributions, highlighting that the classifier relies more heavily on the most informative variables when distinguishing between activities. This pattern suggests that the dataset contains structurally meaningful features that capture key motion characteristics. From a practical standpoint, this also implies that dimensionality reduction or feature-selection strategies could be applied in future work to improve computational efficiency without sacrificing performance.

Figure 7 indicates that most predictions exhibit high probability confidence, meaning that the model not only predicts the correct class but does so with strong certainty in the majority of cases. High confidence distributions typically reflect well-defined class boundaries and strong internal feature representations. This characteristic is particularly important in real-world HAR applications, where stable and confident predictions are preferable to uncertain or fluctuating outputs.

Taken together, the numerical results and graphical evidence collectively validate the robustness, consistency, and reliability of the trained model. The strong alignment between statistical metrics and visualization-based interpretation demonstrates that the classifier performs effectively across multiple evaluation perspectives, strengthening confidence in the overall quality of the analytical pipeline.

7. Conclusion

This project implemented a complete Human Activity Recognition (HAR) workflow that included dataset preparation, LightGBM-based model training, metric extraction, and visualization-oriented evaluation. The results demonstrate that the classifier achieved near-perfect macro-F1 and accuracy scores, along with extremely low cross-validation variance and full robustness under noisy-test evaluation. These outcomes indicate that the model not only learns highly discriminative feature representations, but also maintains stable behaviour across different evaluation scenarios. The alignment between numerical metrics and visual interpretation confirms that the system exhibits strong generalization capability and well-behaved model dynamics, making the pipeline both reliable and methodologically sound.

Beyond performance, the project also illustrates the importance of combining quantitative evaluation with analytical interpretation. The confusion-matrix plots, feature-importance visualizations, confidence-distribution curves, and loss-trend analyses collectively provide a deeper understanding of why the model performs well and how its internal decision process is structured. This strengthens the validity of the findings and highlights the role of visualization-driven reasoning in model assessment rather than relying only on single-value metrics.

Future work may extend this study toward:

- **subject-wise cross-validation protocols**, to evaluate how well the model generalizes to unseen individuals and to reduce potential subject-specific bias in recognition performance;
- **temporal deep-learning architectures,** such as LSTM, GRU, or CNN-based sequence models, which can capture long-range temporal dependencies and richer motion dynamics beyond fixed-window feature representations; and
- **real-world deployment testing in dynamic and unconstrained environments,** where variations in movement style, sensor placement, noise conditions, and daily-life activity patterns can be evaluated to assess long-term robustness and practical usability.