Project Report: Human Activity Recognition

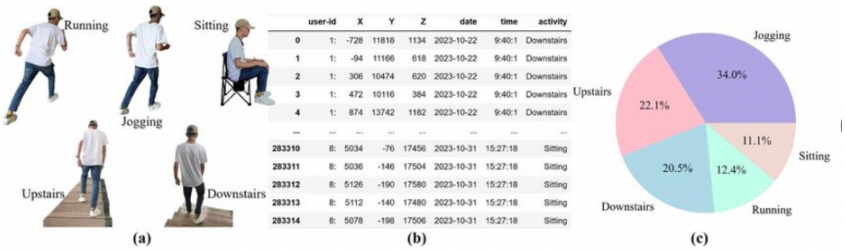
Omar Ahmed (20225281) Youssef Amr (20221949)
Youssef Elkhouly (20223949) Mohamed Ayman (20223929)
Abdallah Mohamed (20222902)

Abstract

This report documents the two phases of the Human Activity Recognition (HAR) project. Phase 1 involved training classical machine learning models (SGDClassifier, Random Forest) on pre-extracted features, including an analysis of Principal Component Analysis (PCA) for dimensionality reduction. Phase 2 focused on implementing an LSTM deep learning model using the raw inertial sensor data. The final analysis provides a comprehensive comparison of all models based on accuracy, loss, precision, complexity, and the fundamental trade-offs between feature engineering and end-to-end deep learning on time-series data. The SGDClassifier without PCA achieved the highest overall accuracy.

---

1 Dataset Distribution



	user-id	X	Y	Z	date	time	activity
0	1:	-728	11818	1134	2023-10-22	9:40:1	Downstairs
1	1:	-94	11166	618	2023-10-22	9:40:1	Downstairs
2	1:	306	10474	620	2023-10-22	9:40:1	Downstairs
3	1:	472	10116	384	2023-10-22	9:40:1	Downstairs
4	1:	874	13742	1182	2023-10-22	9:40:1	Downstairs
...	...	...	...	...	...	...	...
283310	8:	5034	-76	17456	2023-10-31	15:27:18	Sitting
283311	8:	5036	-146	17504	2023-10-31	15:27:18	Sitting
283312	8:	5126	-190	17580	2023-10-31	15:27:18	Sitting
283313	8:	5112	-140	17480	2023-10-31	15:27:18	Sitting
283314	8:	5078	-198	17506	2023-10-31	15:27:18	Sitting

Figure 1: a) Visual samples of the human activities recorded; (b) A sample of the raw data features including sensor readings (X, Y, Z) and activity labels; (c) The distribution of the six activity classes in the dataset.

The dataset contains five human activities: Running, Jogging, Sitting, Upstairs, and Downstairs, as shown in panel (a). Panel (b) displays a sample of the raw time-series sensor data, including the `user-id`, three-axis sensor readings (`X`, `Y`, `Z`), timestamps, and the ground-truth `activity` label. This raw data serves as the foundation for feature extraction and model training. Panel (c) shows the activity class distribution, where 'Jogging' is the most frequent class (34.0%), followed by 'Downstairs' (22.1%), 'Upstairs' (20.5%), and 'Sitting' (12.4%), with 'Running' being the least represented (11.1%). A slight class imbalance is observed, which may impact classification performance.

---

2 Preprocessing Summary

2.1 Feature Scaling and Splitting

Features were standardized using the `StandardScaler` to ensure convergence for linear models. The dataset was divided into:

- **Train Set:** 5881 samples
- **Validation Set:** 1471 samples
- **Test Set:** 2947 samples

2.2 PCA for Dimensionality Reduction

Principal Component Analysis (PCA) was applied to assess the trade-off between model performance and feature count.

- **Original Features:** 561
- **Reduced Features:** 102 (retaining 95% of variance)

---

3 Classical Model Training and Evaluation

Model optimization was performed on the Validation Set to find optimal hyperparameters (Learning Rate/Epochs for SGD, `max_depth` for RF).

3.1 SGDClassifier Performance

- **Optimal Hyperparameters (No PCA):** Learning Rate (LR) = 0.001, Epochs = 100.
- **Optimal Hyperparameters (With PCA):** LR = 0.01, Epochs = 100.

Table 1: SGDClassifier Performance (Test Set)

Scenario	Train Acc.	Val Acc.	Test Acc.	Test Loss
No PCA (561 Features)	0.9852	0.9776	**0.9535**	0.1994
With PCA (102 Features)	0.9740	0.9660	0.9250	0.2937

3.2 Random Forest Performance

- **Optimal Hyperparameter (No PCA):** `max_depth` = 20.
- **Optimal Hyperparameter (With PCA):** `max_depth` = 20.

Table 2: Random Forest Performance (Test Set)

Scenario	Train Acc.	Val Acc.	Test Acc.	Test Loss
No PCA (561 Features)	1.0000	0.9823	0.9250	0.2560
With PCA (102 Features)	1.0000	0.9429	0.8856	0.5299

3.3 Complexity and Overfitting Analysis

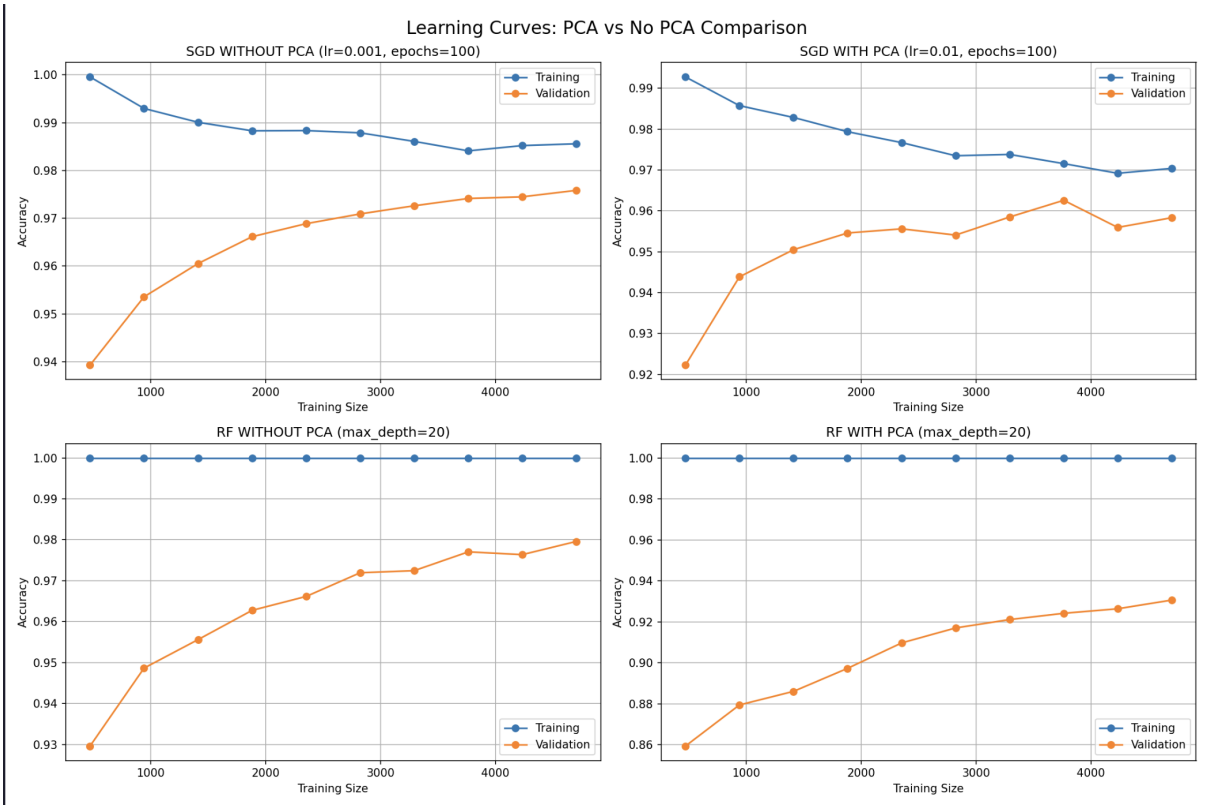


Figure 2: Learning Curves: A comparison of Training vs. Validation Accuracy as a function of training size for both SGDClassifier and Random Forest, contrasting the performance using original (No PCA) and PCA-reduced features.

The Learning Curves (Figure 2) show distinct behaviors:

- **SGDClassifier:** Shows a good fit, with training and validation curves converging relatively closely, especially without PCA.
- **Random Forest:** Exhibits more severe **overfitting**, with perfect training accuracy (Train Acc. = 1.0000) but a noticeable gap to the validation accuracy. This overfitting is amplified when PCA is applied.

---

4 Deep Learning Model on Raw Data

4.1 Preprocessing for Raw Sensor Data

The raw inertial sensor data was preprocessed by segmenting the time-series into fixed-size windows (Step 3). This data was then reshaped and scaled appropriately for the recurrent nature of the LSTM model.

4.2 LSTM Model Performance

An LSTM network was implemented to automatically learn temporal features directly from the raw data.

- **Accuracy (Test):** 0.8928
- **Precision (Test):** 0.8919
- **Log Loss (Test):** 0.5695

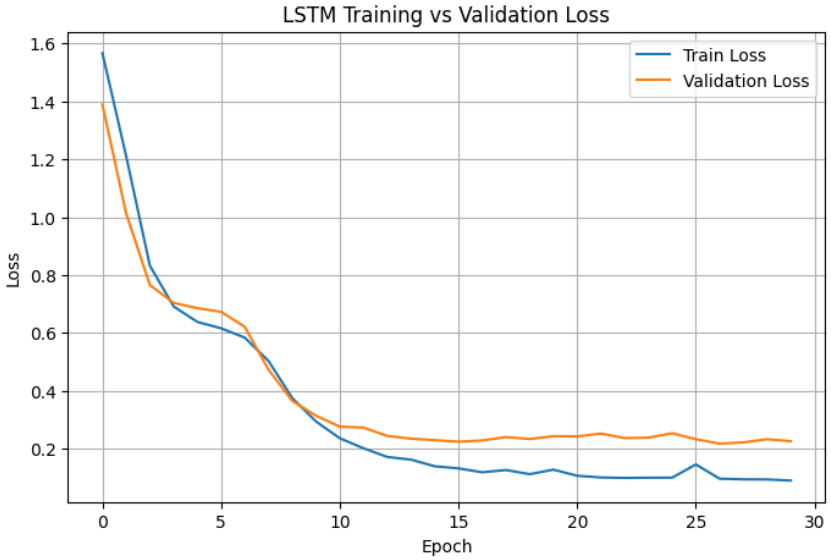


Figure 3: LSTM Training vs Validation Loss. The training loss decreases steadily, but the gap between the training and validation loss after epoch 15 indicates mild overfitting.

The loss curve (Figure 3) confirms that the LSTM successfully learns the data patterns. The validation loss stabilizes quickly, suggesting that further training would yield diminishing returns and could worsen the slight overfitting observed.

---

5 Comparative Performance Analysis

This section provides the critical comparison across all trained models and data representations.

Table 3: Final Comprehensive Model Comparison (Test Set Metrics)

Model	Data Type	Accuracy	Loss	Complexity / Speed
SGDClassifier (No PCA)	Extracted Features	0.9535	0.1994	Low / Very Fast
Random Forest (No PCA)	Extracted Features	0.9250	0.2560	Medium / Fast
SGDClassifier (With PCA)	Reduced Features	0.9250	0.2937	Low / Very Fast
Random Forest (With PCA)	Reduced Features	0.8856	0.5299	Medium / Fast
LSTM	Raw Sensor Data	0.8928	0.5695	High / Slow

5.1 Which Model Performs Best and Why

The SGDClassifier on the 561 pre-extracted features achieved the highest test accuracy of **0.9535**. This linear model performed best because the manual feature engineering process in Step 1 successfully created a feature space that is highly discriminative and linearly separable. Given the optimal feature set, the simple, fast SGD model was sufficient to learn the best decision boundary.

5.2 Whether Dimensionality Reduction Improves Performance

Dimensionality reduction using PCA consistently degraded performance for the classical models:

- SGDClassifier accuracy dropped by 2.85%.
- Random Forest accuracy dropped by 3.94%.

While PCA successfully reduced the feature count by over 80% (561 to 102), the 5% of variance that was discarded contained critical, non-redundant information for distinguishing activity classes. Therefore, the computational benefit of reduction did not outweigh the loss in classification accuracy.

5.3 Whether Deep Learning Outperforms Classical ML on Raw Data

In this case, the Classical ML model (SGDClassifier) on pre-extracted features significantly outperforms the Deep Learning model (LSTM) on raw data (0.9535 vs. 0.8928).

- The hand-crafted features were highly effective, capturing the essential statistical and frequency domain information.
- The LSTM, while theoretically superior for raw time-series, did not achieve the same accuracy level, likely due to the limited training data/time or model complexity, demonstrating that high-quality, domain-specific feature engineering can still surpass end-to-end deep learning for certain HAR tasks.

5.4 Trade-offs: Accuracy vs. Complexity

The analysis reveals a clear trade-off between accuracy and model complexity:

- **Highest Accuracy, Lowest Complexity:** The SGDClassifier (No PCA) offers the optimal solution, being extremely fast to train and run inference on, with the best generalization accuracy.
- **High Complexity, Lowest Accuracy:** The LSTM model is computationally intensive (High complexity) and provided the lowest generalization accuracy among the non-PCA models, making it the least efficient choice for this specific problem outcome.

---

6 Streamlit Application Deployment (Live Demo)

6.1 Model Selection and Justification

The SGDClassifier (No PCA) is selected for deployment. This choice is justified by its superior performance, as it provides the highest test accuracy (0.9535) while maintaining the **lowest computational complexity** and fastest inference time, which is ideal for a responsive production application.

6.2 Deployment Tasks and Functionality

The Streamlit application will be designed to support the following inputs for real-world interaction:

- **Upload CSV of Features:** Allows users to upload a file containing the 561 pre-extracted features for batch prediction.

6.3 Display Options

The output will focus on clarity and real-time feedback:

- **Predicted Activity:** The final classified activity label (e.g., 'WALKING', 'SITTING').
- **Confidence:** The probability score output by the SGDClassifier or Random Forest for the predicted class.