
Assignment 5.2 – Feature Engineering + Classical ML Approach

WESAD Dataset – Stress Detection Project

---

1. Objective

This assignment implements **Approach-1** of the Stress Detection Project using the WESAD dataset. The main objectives were:

- Understand the structure of WESAD dataset
- Perform feature engineering on chest and wrist sensor data
- Extract statistical, temporal, frequency-domain, and physiological features
- Train classical Machine Learning models (Logistic Regression, SVM, Random Forest, XGBoost)
- Evaluate models using Accuracy, Precision, Recall, F1-Score, Confusion Matrix, and ROC–AUC
- Provide analysis and discussion

---

2. Dataset Overview

- **Dataset:** WESAD (Wearable Stress and Affect Detection)
- **Source:** UCI Repository
- **Sensors:**
    - Chest: ACC (X, Y, Z), ECG, EDA, EMG, Respiration, Temperature

    - Wrist: ACC, EDA, Temperature, BVP
- **Problem Type:** Multiclass classification
    - 0 → Baseline
    - 1 → Stress
    - 2 → Amusement
    - 3–7 → Others
- **Data File Used:** `S2.pkl`

---

3. Feature Engineering

Signals were segmented into **10-second windows**, and features were extracted from each window:

- **Statistical Features:** mean, median, std, variance, min, max, 25th & 75th percentile, skewness, kurtosis, RMS
- **Temporal Features:** Zero Crossing Rate (ZCR), Signal Magnitude Area (SMA), energy, peak-to-peak amplitude
- **Frequency-Domain Features:** dominant frequency, spectral energy, spectral centroid, spectral entropy
- **Physiological Features:** EDA peaks, tonic level, ACC SMA

---

4. Modeling

- **Models Used:** Logistic Regression, SVM, Random Forest, XGBoost
- **Train/Test Split:** 80% train, 20% test

- **Scaling:** StandardScaler used for Logistic Regression and SVM
- **Evaluation Metrics:** Accuracy, Precision, Recall, F1-Score, ROC–AUC

---

5. Results

Model	Accuracy	Precision	Recall	F1-Score	ROC–AUC
Logistic Regression	0.746	0.429	0.468	0.444	0.940
SVM	0.770	0.364	0.395	0.376	0.978
Random Forest	0.926	0.949	0.920	0.932	0.984
XGBoost	0.885	0.628	0.623	0.624	0.985

Observations:

- Random Forest achieved the highest overall accuracy (92.6%) and strong F1-score (0.932)
- XGBoost performed slightly lower in accuracy but had competitive ROC–AUC (0.985)
- Logistic Regression and SVM achieved moderate accuracy with lower precision and recall

---

6. Analysis

- **Best Model:** Random Forest, due to its ability to handle non-linear patterns and robustness to feature variance
- **Important Features:** Spectral entropy, RMS, SMA, peak-to-peak amplitude contributed most to model performance
- **Challenges:**
    - Baseline vs Amusement classes showed some confusion

    - Physiological noise affected feature stability
    - Window size influenced model results

---

7. Limitations

- Only one subject’s data used
- HRV and advanced physiological features not included
- Deep learning models not applied yet

---

8. Future Work

- Extend analysis to multiple subjects
- Include deep learning approaches (CNN, LSTM, Transformer)
- Multimodal fusion of chest and wrist signals
- Advanced physiological features such as HRV and respiration patterns

---