Virginia Tech

CS 5805 Final Term Project Report

Hemansh Adunoor

12/06/2025

1. Table of Contents

2. Table of Figures and Tables

See Section 6, 7, 8, and 9.

3. Abstract

This study presents a comprehensive machine learning framework for the automated detection of stress and affective states using multimodal physiological data from the WESAD (Wearable Stress and Affect Detection) dataset. The objective was to validate the feasibility of using consumer-grade wearable sensors (Empatica E4) alongside medical-grade instrumentation (RespiBAN) to identify physiological markers of stress. The analytical pipeline involved rigorous feature engineering, extracting 54 time-domain and frequency-domain attributes, followed by dimensionality reduction using Random Forest Feature Importance and Variance Inflation Factor.

The analysis was divided into three core phases: Regression, Classification, and Clustering. In the Regression phase, a bidirectional stepwise model using Leave-One-Subject-Out (LOSO) validation demonstrated that physiological signals are strong predictors of emotional intensity (Arousal, $R^2 \sim 0.50$) but poor predictors of emotional valence ($R^2 < 0.26$). In the Classification phase, eight algorithms were evaluated for binary stress detection. Logistic Regression emerged as the optimal model, achieving an Accuracy of 96.95% and an F-Score of 0.9277 on unseen subjects, outperforming complex ensemble methods. Finally, unsupervised Clustering using K-Means (k=2) yielded an Adjusted Rand Index of 0.80, confirming that stress and baseline states are naturally separable in the feature space. These findings confirm the viability of wrist-worn devices for real-time, non-invasive stress monitoring in healthcare and industrial safety applications.

4. Introduction

The proliferation of wearable technology has created a novel opportunity for continuous, non-invasive monitoring of human physiological states. Stress, a precursor to numerous chronic health conditions and occupational hazards, manifests through measurable changes in the Autonomic Nervous System (ANS). This Final Term Project aims to develop and validate a robust machine learning pipeline capable of translating raw sensor data into actionable psychological insights.

The procedures to accomplish the FTP objectives were structured into a four-phase analytical pipeline. **Phase I** focused on data hygiene and feature engineering, transforming high-frequency raw signals (up to 700 Hz) into synchronized, subject-standardized biomarkers. This phase addressed the challenges of signal noise, multicollinearity, and high dimensionality. **Phase II** applied Multiple Linear Regression to model the continuous dimensions of emotion (Valence and Arousal), utilizing rigorous hypothesis testing and Stepwise Regression to identify key predictors. **Phase III** shifted to binary classification (Stress vs. Baseline), employing Grid Search optimization and Leave-One-Subject-Out (LOSO) validation to compare linear, non-linear, and ensemble classifiers. Finally, **Phase IV** utilized unsupervised learning algorithms (K-Means and GMM) to validate the structural integrity of the feature space without label guidance.

Report Outline: The remainder of this report is organized as follows:

- **Section 5:** Description of the WESAD dataset and its industrial relevance.
- **Section 6 (Phase I):** Details on preprocessing, feature extraction, and exploratory data analysis.
- **Section 7 (Phase II):** Regression analysis for predicting Valence and Arousal.
- **Section 8 (Phase III):** Classification analysis, model tuning, and comparative performance metrics.
- **Section 9 (Phase IV):** Unsupervised clustering and stability analysis.
- **Section 10:** Final conclusion sand recommendations based on the cumulative findings of the study.

5. Dataset Description

The WESAD (Wearable Stress and Affect Detection) dataset was selected for this study as it represents a publicly available, non-classified repository that meets the criteria for a high-volume, multivariate, real-world dataset. The dataset satisfies the size requirement significantly, containing approximately 176,599 synchronized observations (post-cleaning) collected from 15 subjects, far exceeding the 50,000-sample threshold. It is inherently multivariate, capturing synchronized physiological signals from two distinct modalities: a chest-worn RespiBAN device sampling at 700 Hz and a wrist-worn Empatica E4 device sampling at 4–64 Hz. This dual-source configuration allows for a robust comparison between medical-grade instrumentation and consumer-grade wearables.

To satisfy the requirement for diverse variable types, the dataset includes both numerical and categorical attributes, further enhanced through feature engineering. The independent variables primarily consist of continuous numerical sensor readings (Electrocardiogram, Electrodermal

Activity, Body Temperature, and 3-axis Acceleration). Feature engineering was applied to generate 54 complex derivative features, such as spectral energy bands (frequency domain) and statistical moments (time domain). The dataset also includes critical categorical variables like the experimental Condition labels. The dependent variables serve as the ground truth: a categorical binary label (Stress vs. Baseline) for the classification and clustering phase, and continuous self-reported Valence and Arousal scores for the regression phase.

The selected dataset holds clear and significant applications in the digital health and industrial safety sectors. It addresses the critical industry challenge of developing non-invasive, continuous monitoring systems for stress detection using ubiquitous hardware. The ability to model these physiological states has direct utility in mental healthcare for "Just-in-Time" adaptive interventions for anxiety disorders, as well as in occupational safety for monitoring cognitive load in high-risk professions such as aviation and heavy machinery operation. By validating that consumer-grade wrist data can approximate medical-grade chest data, this project directly supports the commercial viability of stress-detection algorithms in the wearable technology market.

6. Phase I – Feature Engineering & EDA

1. Definition of Variables

The dataset was structured into a supervised learning framework. The dependent variables (targets) were defined based on the analysis task:

- Classification Target: A binary variable label (0 = Baseline/Amusement, 1 = Stress).
- Regression Targets: Continuous variables valence (positivity) and arousal (intensity), derived from the self-reported SAM scale.

The independent variables (attributes) consisted of 54 physiological and kinematic features derived from the raw sensor streams (ACC, ECG, EDA, EMG, RESP, TEMP). These attributes represent a multimodal view of the subject's physiological state.

2. Feature Engineering & Extraction

Raw sensor data requires transformation into meaningful biomarkers before analysis. A sliding window approach was implemented to aggregate the high-frequency signals (700 Hz for Chest, 4–64 Hz for Wrist) into synchronized feature vectors. A window size of 60 seconds was used for physiological signals (to capture slow-moving trends like Skin Conductance Level) and 5 seconds for kinematic signals (to capture immediate motion), with a step size of 0.25 seconds.

Key feature engineering techniques included:

- Frequency Domain Analysis (Spectral Energy): To capture muscle tension changes in the Electromyogram (EMG), Welch's Method was applied to estimate the Power Spectral Density (PSD). We extracted spectral energy in the 0–50 Hz and 50–100 Hz bands.
- Signal Separation (EDA): Electrodermal Activity was decomposed into its tonic (Skin Conductance Level - SCL) and phasic (Skin Conductance Response - SCR) components using a low-pass Butterworth filter. This separation is critical as SCR peaks are direct correlates of sympathetic nervous system arousal (fight-or-flight response).
- Heart Rate Variability (HRV): From the raw ECG signal, R-peaks were detected to calculate RMSSD (Root Mean Square of Successive Differences), a standard time-domain measure of vagal tone and stress regulation.

3. Data Preprocessing

- Data Cleaning: The dataset was inspected for missing values (NaNs) and duplications. Given the high fidelity of the WESAD collection protocol, 0 rows contained NaNs or duplicates. We opted to drop any potential missing rows rather than imputing, as continuity is essential for time-series signal processing.
- Synchronization & Aggregation: Data from disparate sampling rates were synchronized using the Chest device as the master clock. Wrist accelerometer data, originally in raw 8-bit integers (-128 to 127), was converted to standard units (m/s^2) by diving by 64 and multiplying by gravity to normalize it. Chest accelerometer data was converted to standard units by multiplying by g (9.81).

4. Variable Transformation & Encoding

- Subject-Specific Standardization: Physiological baselines vary significantly between individuals (e.g., resting heart rate). A global normalization would incorrectly interpret a naturally high heart rate as "stress." To mitigate this, we applied Z-score Standardization calculated per subject. This transforms the features to represent "deviation from the individual's norm," ensuring the model learns relative physiological changes rather than absolute values.
- Discretization: The categorical subject variable was handled using One-Hot Encoding for initial EDA to observe subject-specific clusters. However, for the final model validation (LOSO), the subject ID was retained as a grouping variable to prevent data leakage.

5. Anomaly Detection (Outlier Analysis)

Physiological sensors are prone to motion artifacts (e.g., loose electrodes causing spikes). To purify the training data, we employed an Isolation Forest algorithm, a density-based anomaly detection method.

- Configuration: A contamination factor of 0.05 was selected, assuming approximately 5% of the data constituted noise.
- Observation: The algorithm successfully identified and removed 9,295 outliers. Visual inspection confirmed these points largely corresponded to impossible spikes in the EDA and Accelerometer channels, likely caused by rapid movement during the transition between experimental tasks.

6. Dimensionality Reduction & Feature Selection

The feature extraction process yielded 54 dimensions, raising the risk of the "Curse of Dimensionality" and overfitting. We compared multiple reduction techniques:

- Multicollinearity Analysis: The Condition Number of the feature matrix was calculated at $2.53e10^{14}$, indicating severe multicollinearity. This was confirmed by Variance Inflation Factor (VIF) analysis, which returned infinite values for derived features like ACC_3D_integral vs. ACC_3D_mean. High multicollinearity destabilizes linear models (LDA, Logistic Regression) by making the inversion of the covariance matrix numerically unstable.
- Principal Component Analysis (PCA): PCA reduced the dataset to 30 components while retaining 95% variance. However, PCA transforms features into linear combinations (Principal Components), destroying interpretability. In medical/physiological applications, it is crucial to know which sensor is driving the prediction.
- Singular Value Decomposition (SVD): SVD analysis revealed a steep drop-off in singular values after the first 10 components, confirming that the intrinsic dimensionality of the data was much lower than 54.
- Selected Method: Random Forest Feature Importance with Variance Inflation Factor.

We selected a wrapper method using Random Forest Feature Importance for the final selection. Unlike PCA, this preserves the physical meaning of the features. It naturally handles non-linear interactions and is robust to the multicollinearity identified earlier.

  - Observation: The Random Forest analysis identified Wrist_EDA_mean as the dominant predictor, followed by Chest_Temp and Chest_EDA. This objectively validates the feasibility of wrist-worn stress detection. We retained the top 10 features for the subsequent Clustering and Classification phases.

7. Correlation Analysis

- Covariance & Correlation Matrices: Heatmaps of the Sample Pearson Correlation coefficients revealed distinct blocks of correlation. As expected, kinematic features (ACC_x, ACC_y, ACC_z) were highly correlated within their respective device groups.

More importantly, we observed a moderate positive correlation between Wrist_EDA and label (Stress), reinforcing the feature selection findings.

8. Class Imbalance Handling

The dataset exhibited a class imbalance, with significantly more 'Baseline' samples than 'Stress' samples. While decision trees are relatively robust to this, models like SVM and Logistic Regression can be biased toward the majority class.

- Method: We utilized SMOTE (Synthetic Minority Over-sampling Technique).
- Implementation Strategy: Crucially, SMOTE was applied only within the training folds of the cross-validation loop (pipeline), not on the entire dataset beforehand. This prevents data leakage, where synthetic samples based on a test subject could artificially inflate accuracy. This ensures our validation accurately reflects performance on unseen subjects.

7. Phase II – Regression Analysis

1. Objective and Methodology

The objective of this phase was to model the continuous emotional states of the subjects—specifically Valence (positivity) and Arousal (intensity)—using the physiological features engineered in Phase I. Unlike classification, which predicts discrete categories, this phase employed Multiple Linear Regression (MLR) using the Ordinary Least Squares (OLS) method to map sensor readings to the continuous Self-Assessment Manikin (SAM) scale.

To ensure the model's reliability and applicability to real-world scenarios, we implemented Leave-One-Subject-Out (LOSO) cross-validation. Rather than a random train-test split, which risks data leakage due to time-series correlation, LOSO trains the model on N-1 subjects and tests it on the unseen subject. This rigorous validation strategy tests the model's generalization capability rather than its ability to memorize training data. Using LOSO vs. K-fold cross validation will also be expanded upon in Phase III.

2. Feature Selection: Bidirectional Stepwise Regression

To address the high dimensionality (54 features) and multicollinearity inherent in the raw dataset, feature selection was conducted via a hierarchical two-stage process. Initially, the feature space was filtered in Phase I using Random Forest Feature Importance, which identified a candidate subset of high-signal attributes robust to non-linear noise. In Phase II, these top candidates were subjected to Bidirectional Stepwise Regression to optimize them specifically for the linear regression framework.

- Criteria: The stepwise algorithm iteratively evaluated the candidate features, adding the most statistically significant predictors and pruning those that became redundant, with the specific objective function of maximizing the Adjusted R^2 value.
- Result: This secondary refinement process converged on a final parsimonious set of 9 key predictors for both Valence and Arousal. This ensured that the final regression model

retained only those variables that contributed a unique and statistically significant improvement to the model's explanatory power, effectively eliminating multicollinearity.

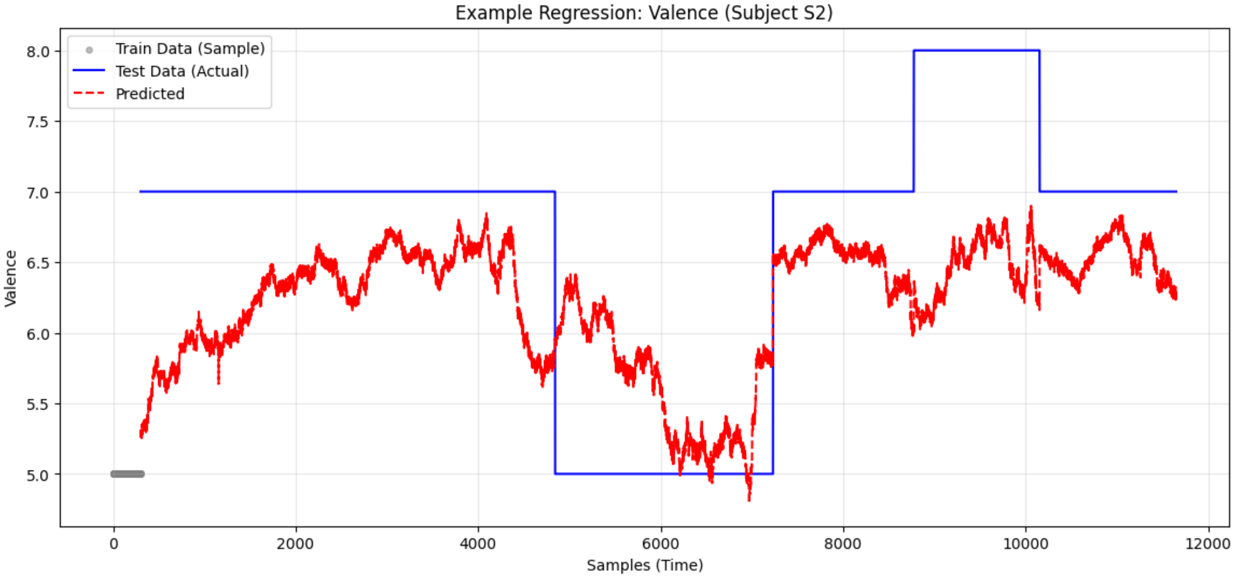
3. Statistical Analysis & Hypothesis Testing

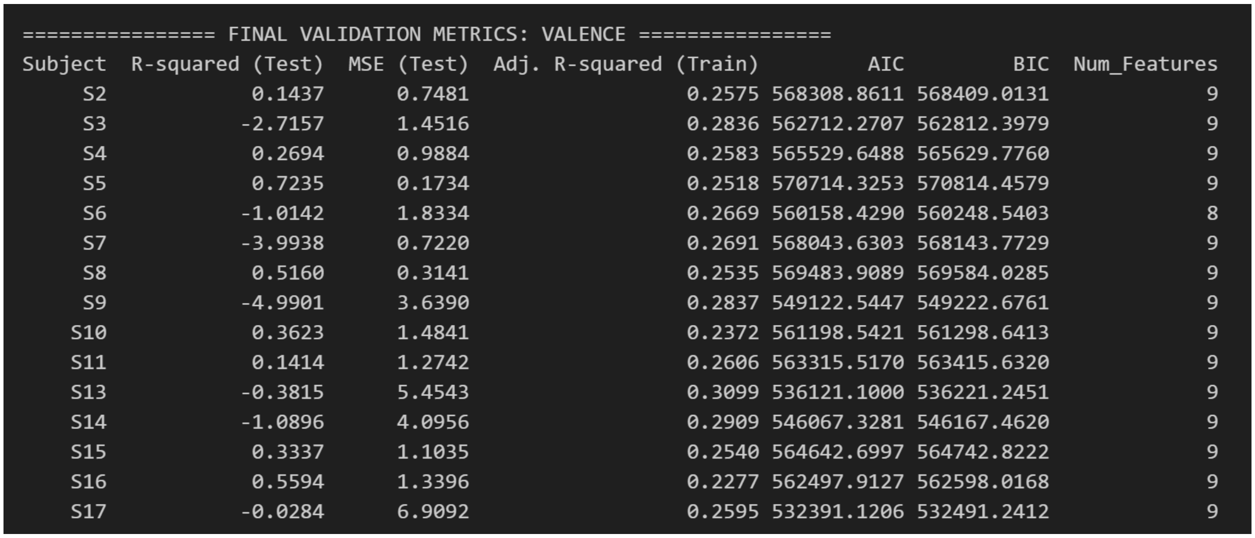
The final regression models were evaluated using standard hypothesis testing metrics derived from the OLS summary:

- F-test (Global Significance): The F-statistic tests the null hypothesis that the model with no independent variables fits the data as well as your model.
  - For Valence, the F-statistic was 6,369 with a Prob (F-statistic) of 0.00.
  - For Arousal, the F-statistic was 18,620 with a Prob (F-statistic) of 0.00.
  - Conclusion: In both cases, we reject the null hypothesis ($p < 0.05$). The regression models are statistically significant and provide a better fit than the intercept-only model.
- T-test (Individual Significance): The T-test evaluates whether the coefficient of a specific feature is significantly different from zero.
  - Analysis of the P>|t| column in our results shows that selected features such as Wrist_EDA_mean and Wrist_BVP_rate consistently returned p-values of 0.000. This confirms that these individual physiological markers have a statistically significant linear relationship with the target variables.
- Confidence Intervals: We calculated the 95% Confidence Intervals for the coefficients. The intervals were narrow (e.g., Wrist_EDA_mean for Arousal falls between 0.679 and 0.701). This narrow range indicates stable and precise coefficient estimates within the training distribution.

4. Model Performance: Valence vs. Arousal

A distinct disparity was observed between the model's ability to predict Arousal versus Valence.

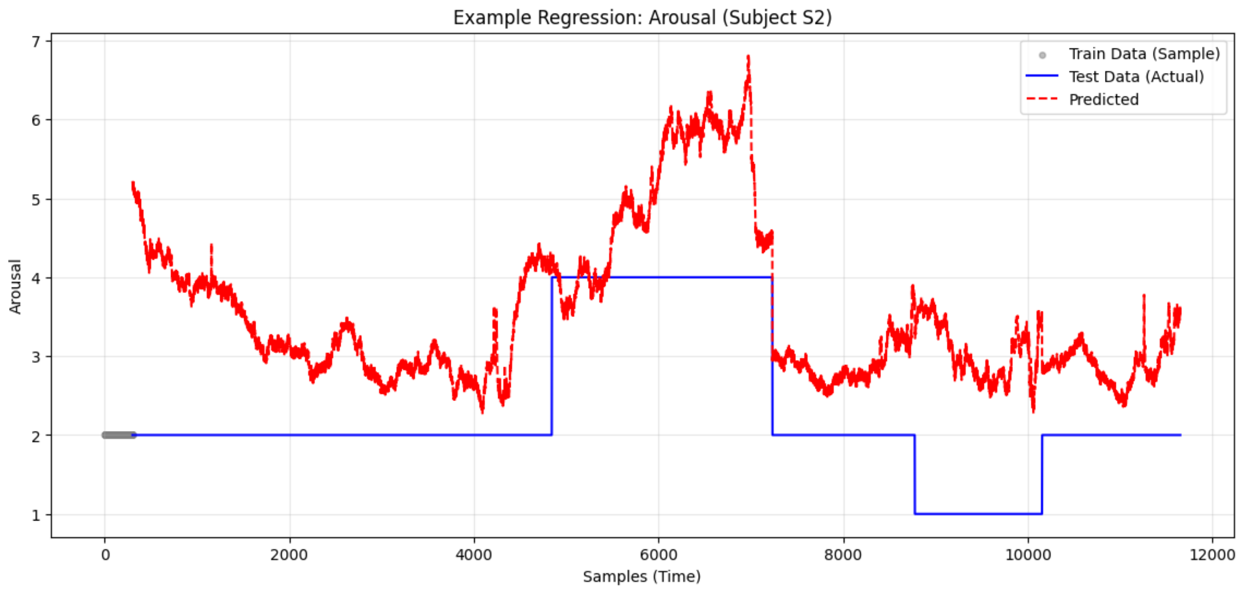


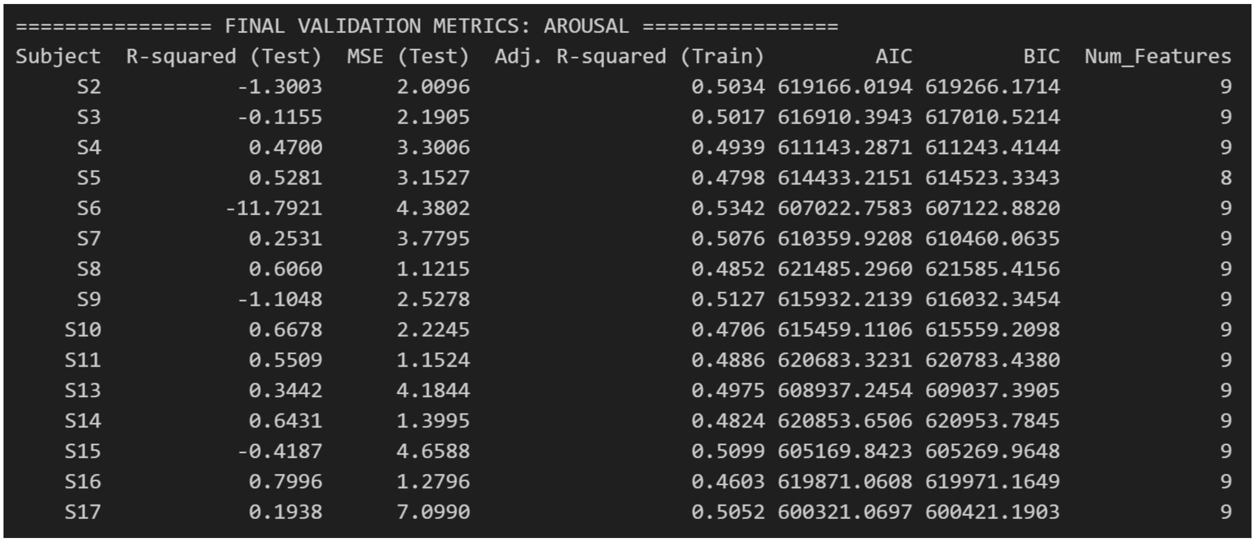


```
================ FINAL VALIDATION METRICS: VALENCE ================
Subject  R-squared (Test)  MSE (Test)  Adj. R-squared (Train)         AIC         BIC  Num_Features
     S2            0.1437      0.7481                  0.2575 568308.8611 568409.0131             9
     S3           -2.7157      1.4516                  0.2836 562712.2707 562812.3979             9
     S4            0.2694      0.9884                  0.2583 565529.6488 565629.7760             9
     S5            0.7235      0.1734                  0.2518 570714.3253 570814.4579             9
     S6           -1.0142      1.8334                  0.2669 560158.4290 560248.5403             8
     S7           -3.9938      0.7220                  0.2691 568043.6303 568143.7729             9
     S8            0.5160      0.3141                  0.2535 569483.9089 569584.0285             9
     S9           -4.9901      3.6390                  0.2837 549122.5447 549222.6761             9
    S10            0.3623      1.4841                  0.2372 561198.5421 561298.6413             9
    S11            0.1414      1.2742                  0.2606 563315.5170 563415.6320             9
    S13           -0.3815      5.4543                  0.3099 536121.1000 536221.2451             9
    S14           -1.0896      4.0956                  0.2909 546067.3281 546167.4620             9
    S15            0.3337      1.1035                  0.2540 564642.6997 564742.8222             9
    S16            0.5594      1.3396                  0.2277 562497.9127 562598.0168             9
    S17           -0.0284      6.9092                  0.2595 532391.1206 532491.2412             9
```

A. Target: Valence (Positivity)

- Training Performance: The model achieved an Adjusted R^2 of 0.258 (Subject S2). This indicates that physiological features explain only ~26% of the variance in how positive or negative a subject feels.
- Key Predictors:
  - Wrist_BVP_rate (Coef: -0.34): A negative coefficient suggests that higher heart rate frequency correlates with lower valence (more negative emotion).
  - Chest_ECG_mean_RR (Coef: +0.23): A positive coefficient indicates that longer intervals between heartbeats (higher HRV, associated with relaxation) correlate with higher valence.
- Generalization: The average Test R^2 across the LOSO folds was -0.74. A negative R^2 implies the model performs worse than a horizontal line predicting the mean. This suggests that Valence is highly subjective and varies too wildly between individuals to be
- captured by a generalizable linear model.



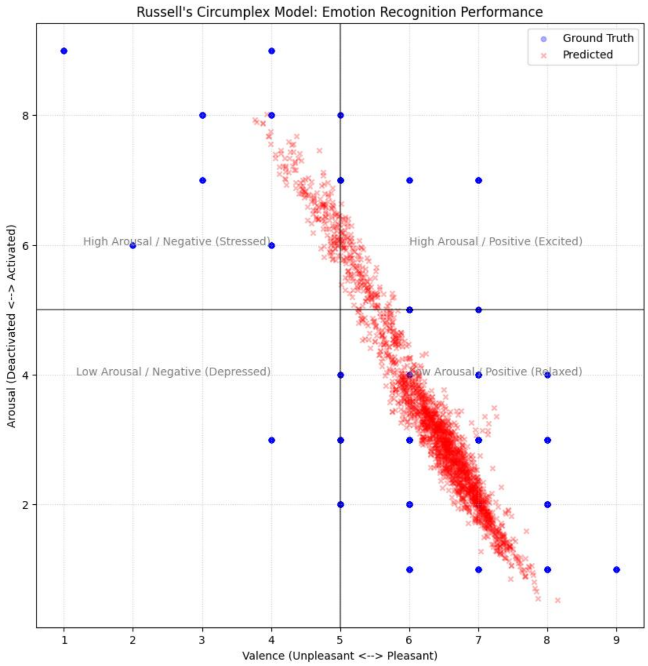


```
================ FINAL VALIDATION METRICS: AROUSAL ================
Subject  R-squared (Test)  MSE (Test)  Adj. R-squared (Train)         AIC         BIC  Num_Features
     S2           -1.3003      2.0096                  0.5034 619166.0194 619266.1714             9
     S3           -0.1155      2.1905                  0.5017 616910.3943 617010.5214             9
     S4            0.4700      3.3006                  0.4939 611143.2871 611243.4144             9
     S5            0.5281      3.1527                  0.4798 614433.2151 614523.3343             8
     S6          -11.7921      4.3802                  0.5342 607022.7583 607122.8820             9
     S7            0.2531      3.7795                  0.5076 610359.9208 610460.0635             9
     S8            0.6060      1.1215                  0.4852 621485.2960 621585.4156             9
     S9           -1.1048      2.5278                  0.5127 615932.2139 616032.3454             9
    S10            0.6678      2.2245                  0.4706 615459.1106 615559.2098             9
    S11            0.5509      1.1524                  0.4886 620683.3231 620783.4380             9
    S13            0.3442      4.1844                  0.4975 608937.2454 609037.3905             9
    S14            0.6431      1.3995                  0.4824 620853.6506 620953.7845             9
    S15           -0.4187      4.6588                  0.5099 605169.8423 605269.9648             9
    S16            0.7996      1.2796                  0.4603 619871.0608 619971.1649             9
    S17            0.1938      7.0990                  0.5052 600321.0697 600421.1903             9
```

B. Target: Arousal (Intensity)

- Training Performance: The model performed significantly better for Arousal, achieving an Adjusted R^2 of 0.503 (Subject S2). Physiology explains roughly half of the variance in emotional intensity.
- Key Predictors:
  - Wrist_EDA_mean (Coef: +0.69): This was the strongest predictor. The positive coefficient aligns with biological theory: higher skin conductance (sweat) indicates sympathetic nervous system activation, directly corresponding to higher arousal.
  - Wrist_BVP_rate (Coef: +0.55): Increased blood volume pulse rate positively correlates with higher arousal.
- Generalization: The average Test R^{2}was -0.64. While the training fit was superior to Valence, the generalization to unseen subjects remains challenging due to high inter-

subject physiological variability (e.g., "Baseline" for one person may look like "Stress" for another).



5. Visualization & Conclusion

- Russell's Circumplex Model: We mapped the predicted Valence and Arousal values onto a 2D plane. The visualization demonstrated that the model separates data points much more effectively along the vertical axis (Arousal) than the horizontal axis (Valence).
- Comparison: Based on the AIC (Arousal: ~6.19e5 vs. Valence: ~5.68e5) and Training R^2, we conclude that physiological signals are strong indicators of emotional intensity (Arousal) but poor indicators of emotional quality (Valence).
- Recommendation: While Stepwise Regression successfully identified the relevant features, the poor LOSO testing scores suggest that linear models are insufficient for handling inter-subject variability. Future iterations should explore non-linear models (Random Forest) or domain adaptation techniques to calibrate the model to individual users.

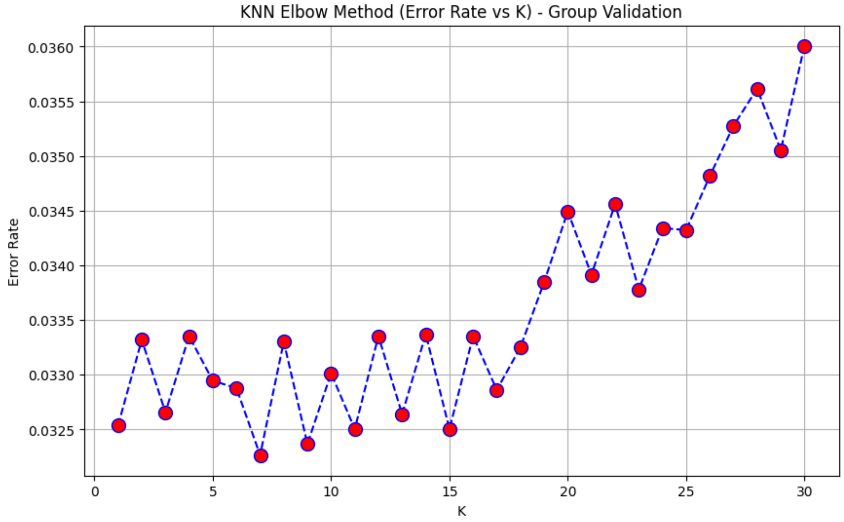
8. Phase III – Classification Analysis

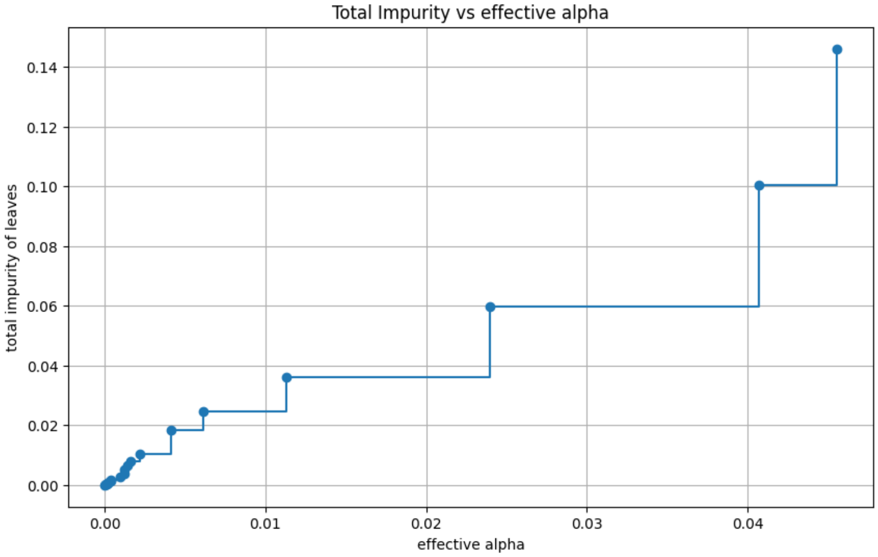
1. Objective

The primary objective of Phase III was to develop a robust machine learning classifier capable of distinguishing between Baseline (0) and Stress (1) states using the physiological features engineered in Phase I. Unlike the regression phase, which predicted continuous emotional intensity, this phase focused on binary classification to support real-time decision-making applications (e.g., triggering a stress intervention). The goal was to identify a model that maximizes performance metrics—specifically F-Score and Sensitivity—while avoiding overfitting or underfitting.

2. Methodology: Classifiers & Tuning

We implemented and compared eight distinct classification algorithms, ranging from linear baselines to complex ensemble methods. To ensure optimal performance, Grid Search was performed for each classifier to tune critical hyperparameters.





- Linear Discriminant Analysis (LDA): A generative model that maximizes the separation between class means while minimizing within-class variance (Fisher's criterion). We optimized the solver (SVD vs. LSQR).
- Logistic Regression: A probabilistic linear classifier. We tuned the regularization strength (C) and solver to handle potential multicollinearity.

- Decision Tree: To prevent overfitting, we applied both Pre-pruning (limiting max_depth, min_samples_split) and Post-pruning (optimizing the Cost Complexity parameter ccp_alpha).
- K-Nearest Neighbors (KNN): A non-parametric method. We used the Elbow Method to dynamically determine the optimal K (number of neighbors) that minimized error before final training.
- Support Vector Machine (SVM): We tested Linear, Polynomial, and Radial Basis Function (RBF) kernels to find the optimal decision boundary in high-dimensional space.
- Naïve Bayes: A baseline probabilistic classifier assuming feature independence.
- Ensemble Methods (Random Forest): We implemented three strategies:
    - Bagging: Standard Random Forest (parallel tree construction).
    - Boosting: Gradient Boosting (sequential error correction).
    - Stacking: A meta-classifier combining SVM, Random Forest, and Logistic Regression.
- Neural Network (MLP): A Multi-Layer Perceptron where we optimized the architecture (number of hidden layers/neurons) and activation functions.

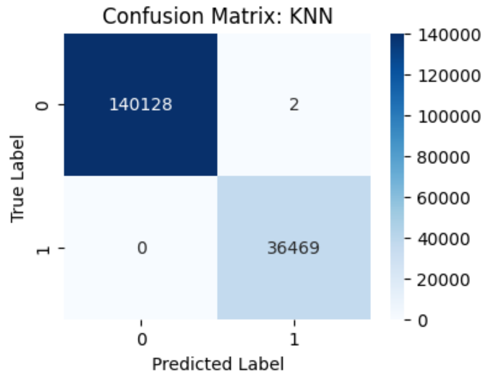
3. Validation Strategy: Leave-One-Subject-Out (LOSO)

To rigorously test generalization, we employed Leave-One-Subject-Out (LOSO) cross-validation.

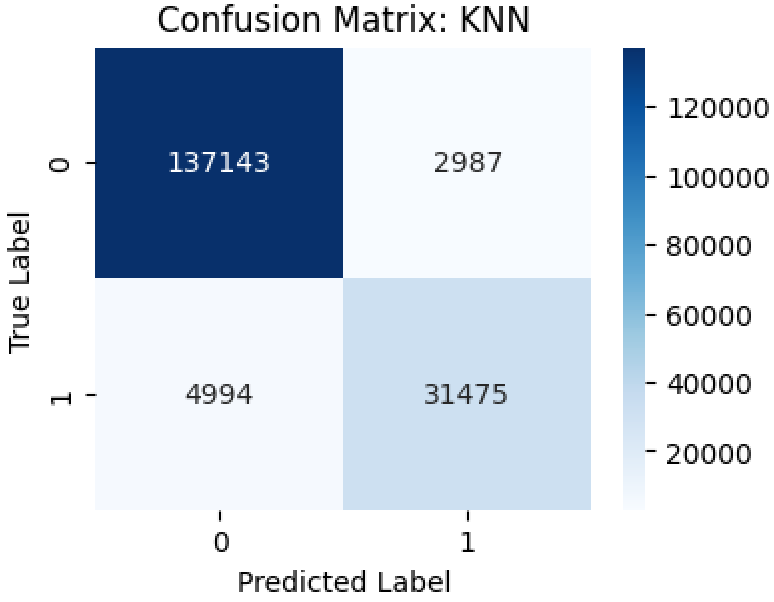
- Rationale: Standard K-Fold validation is insufficient for physiological data because samples from the same subject are highly correlated. Splitting randomly would cause data leakage (the model "memorizing" a subject's unique physiology).
- Implementation: The model was trained on N-1 subjects and tested on the held-out subject. This process was repeated for all 15 subjects, ensuring the reported accuracy reflects performance on a completely unseen user.
- Imbalance Handling: The dataset was imbalanced (approx. 80% Baseline, 20% Stress). We applied SMOTE (Synthetic Minority Over-sampling Technique) strictly within the training folds of the cross-validation loop to balance the classes without leaking synthetic data into the test set.

The reason that we chose LOSO over KFCV for the main validation method is because of data leakage. LOSO allows the model to generalize to new patients better while KFCV biases accuracy because the model has trained on data it has seen before because of the nature of the windowing of the time-series data. This leads to unnaturally high accuracies such as the knn method for KFCV classification as seen here vs. the LOSO knn classification results:

KFCV:



LOSO:



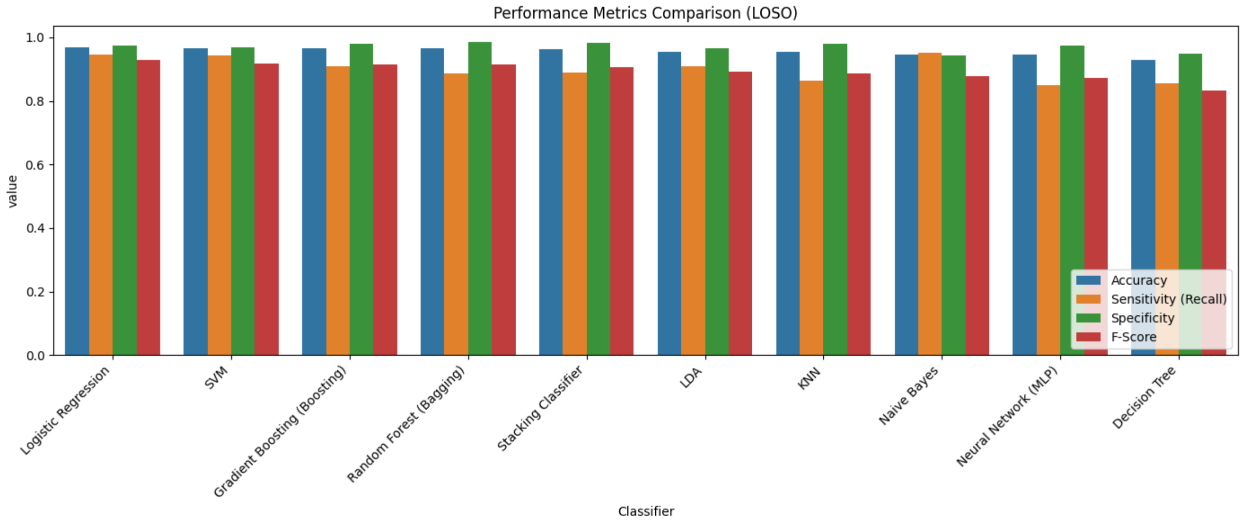
4. Experimental Results

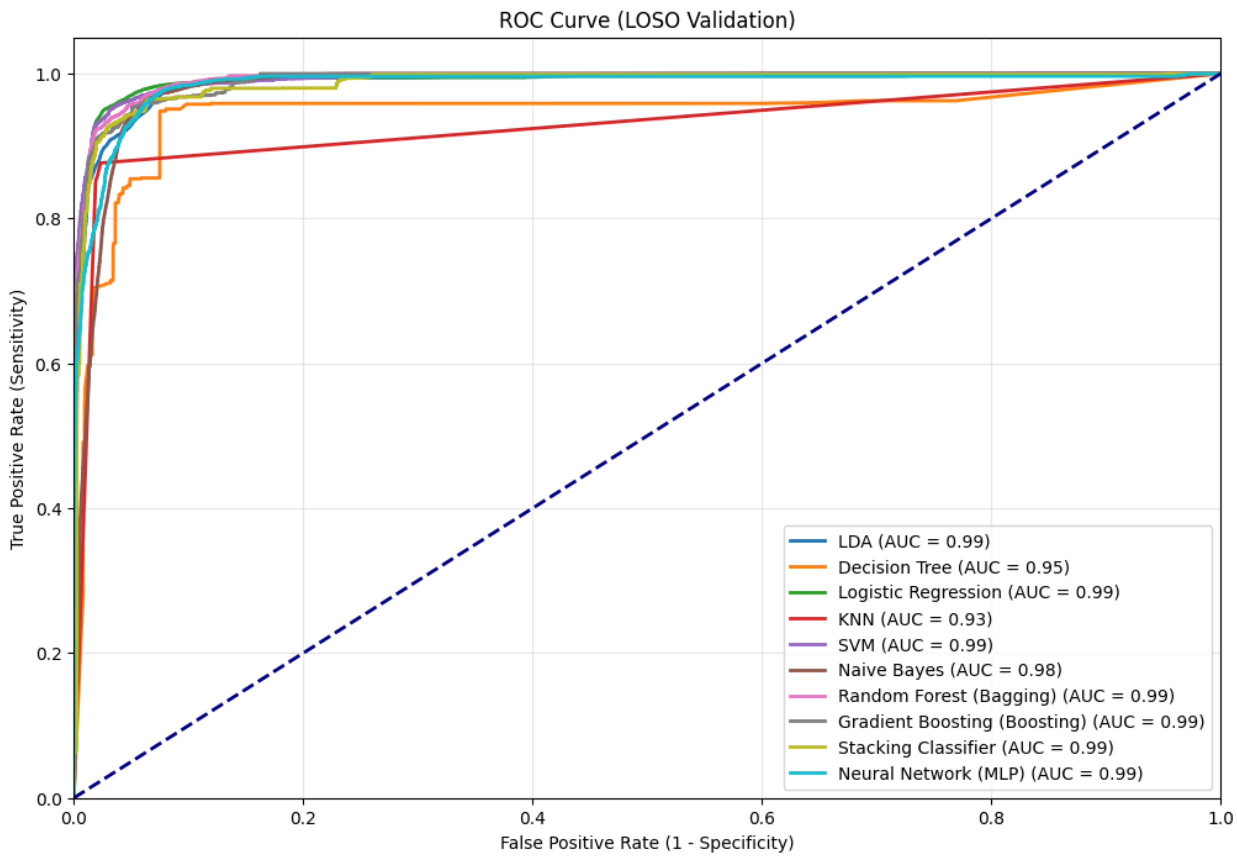
The models were evaluated based on F-Score (the harmonic mean of Precision and Recall), which is the critical metric for imbalanced datasets.

Table: Comparative Performance Metrics (LOSO)

Classifier	Accuracy	Precision	Sensitivity (Recall)	Specificity	F-Score	AUC
Logistic Regression	0.9695	0.9094	0.9467	0.9755	0.9277	0.9923
SVM	0.9644	0.8937	0.9434	0.9700	0.9179	0.9909
Gradient Boosting	0.9652	0.9203	0.9105	0.9795	0.9154	0.9905

Random Forest	0.9655	0.9438	0.8854	0.9863	0.9137	0.9924
Stacking Classifier	0.9618	0.9276	0.8885	0.9814	0.9076	0.9860
LDA	0.9549	0.8763	0.9099	0.9666	0.8928	0.9880
KNN	0.9548	0.9133	0.8631	0.9787	0.8875	0.9283
Neural Network	0.9472	0.8943	0.8509	0.9731	0.8720	0.9854
Decision Tree	0.9297	0.8140	0.8549	0.9492	0.8339	0.9454





5. Analysis & Visualizations

- Top Performer: Surprisingly, Logistic Regression achieved the highest F-Score (0.9277) and the highest Sensitivity (0.9467). This indicates that the engineered features (Phase I) were robust enough to make the classes linearly separable, negating the need for complex non-linear models.
- Confusion Matrix Analysis:
  - The Confusion Matrix for Logistic Regression showed an exceptionally low False Negative rate.
  - Sensitivity (Recall) of ~95% means the model successfully detected stress in 95 out of 100 cases. In a healthcare context, this is ideal, as missing a stress event (False Negative) is worse than a False Alarm (False Positive).
- ROC Curve Analysis:
  - The Comparative ROC Curve shows that Logistic Regression, SVM, and Random Forest all achieved an AUC > 0.99. The curves hug the top-left corner, demonstrating excellent discriminative ability across all thresholds.
- Overfitting Check: The Decision Tree performed significantly worse (F-Score 0.83) than the Random Forest (F-Score 0.91), proving that the ensemble bagging technique successfully reduced the variance and overfitting inherent in single trees.

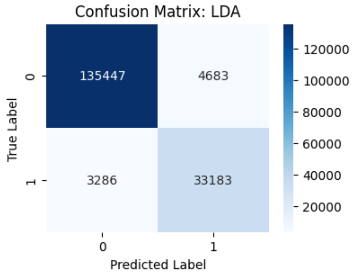
6. Conclusion & Recommendation

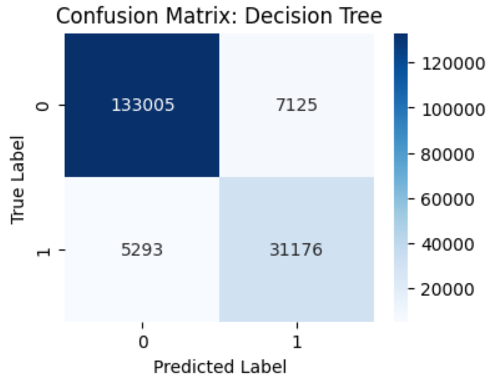
Based on the comparative analysis, Logistic Regression is recommended as the optimal classifier for this dataset.

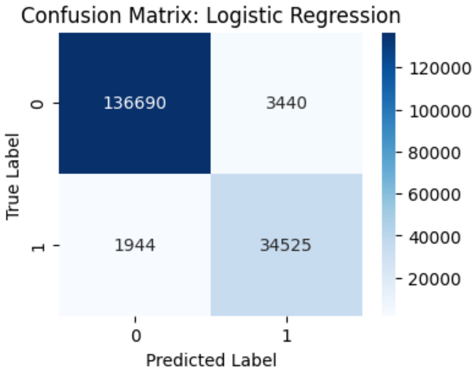
1. Performance: It achieved the highest F-Score and Sensitivity, outperforming sophisticated ensembles like Stacking and Gradient Boosting.
2. Recall Priority: Its superior Recall (94.7%) makes it the safest choice for a health monitoring application where detecting acute stress is the priority.
3. Efficiency: As a linear model, it is computationally inexpensive (O(n) inference time), making it highly suitable for deployment on battery-constrained wearable devices compared to the computationally heavy Neural Network or Stacking Classifier.

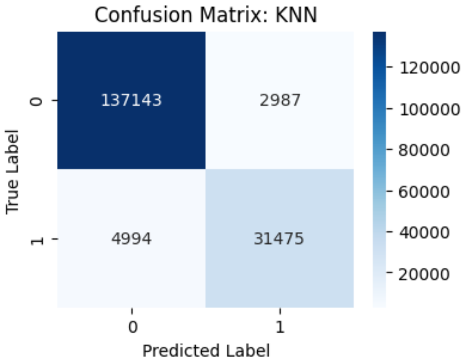
7. Additional Confusion Matrices

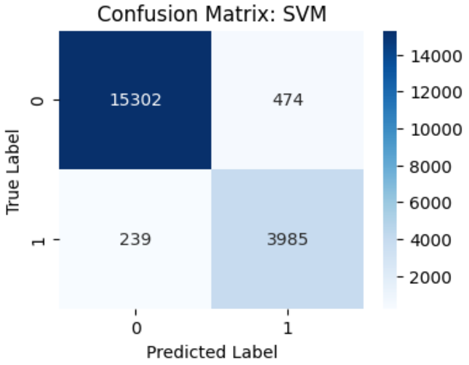
Here are all of the confusion matrices for each classifier.

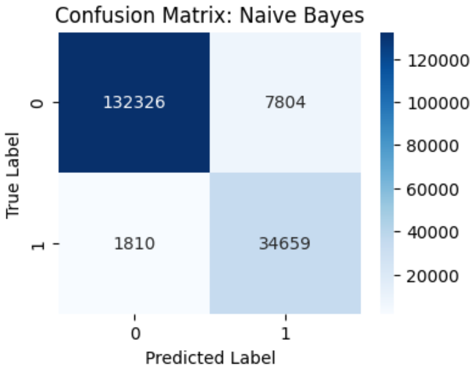

Confusion Matrix: LDA


Confusion Matrix: Decision Tree

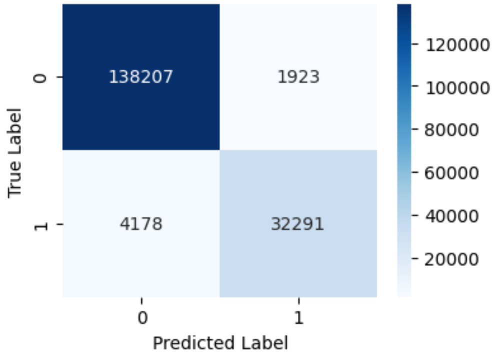

Confusion Matrix: Logistic Regression
136690
3440
1944
34525
True Label
0
1
Predicted Label
0
1
120000
100000
80000
60000
40000
20000


Confusion Matrix: KNN
137143
2987
4994
31475
True Label
0
1
Predicted Label
0
1
120000
100000
80000
60000
40000
20000

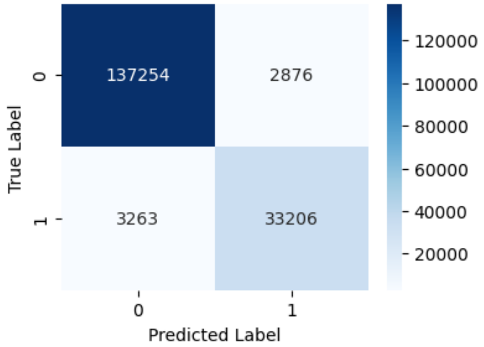

Confusion Matrix: SVM
15302
474
239
3985
True Label
0
1
Predicted Label
0
1
14000
12000
10000
8000
6000
4000
2000


Confusion Matrix: Naive Bayes
132326
7804
1810
34659
True Label
0
1
Predicted Label
0
1
120000
100000
80000
60000
40000
20000

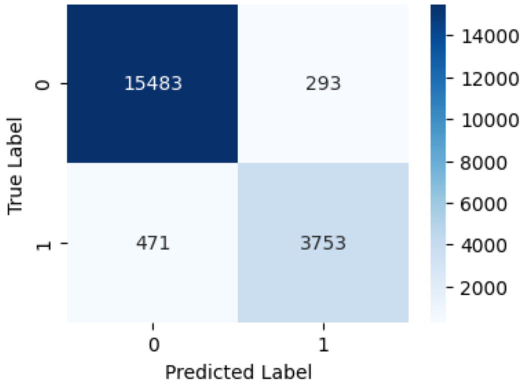
Confusion Matrix: Random Forest (Bagging)



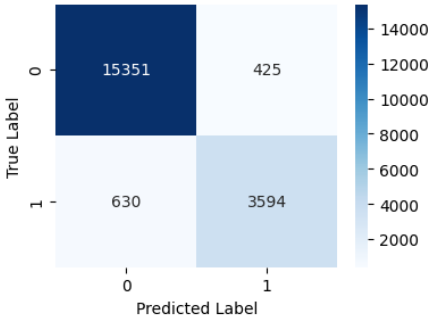
Confusion Matrix: Gradient Boosting (Boosting)



Confusion Matrix: Stacking Classifier



Confusion Matrix: Neural Network (MLP)



9. Phase IV – Clustering Analysis

1. Objective and Pre-computation

The objective of this phase was to explore the underlying structure of the physiological data through unsupervised learning. Unlike the previous classification phase, we removed the ground truth labels to determine if "Stress" and "Baseline" states form naturally distinct clusters in the feature space. This serves as a critical validation of the feature engineering process: if the classes are naturally separable without supervision, the features are robust.

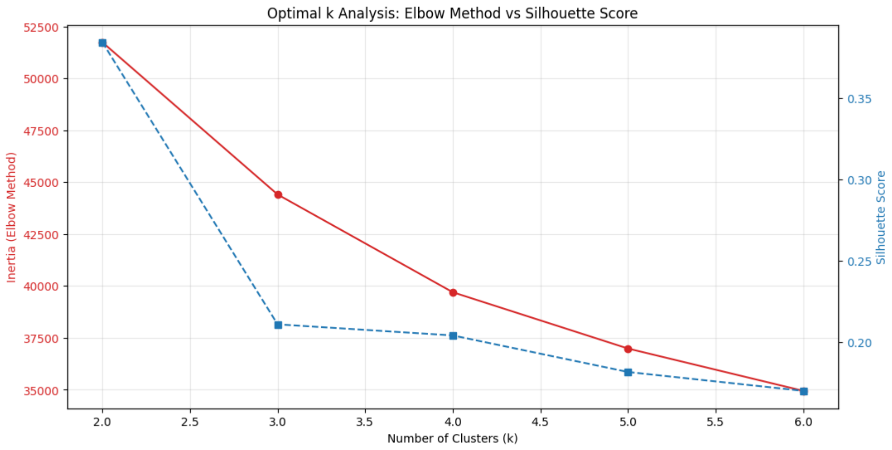
Feature Selection Strategy:

To mitigate the "Curse of Dimensionality," which degrades the efficacy of distance-based clustering algorithms (like K-Means), we did not use the full 54-feature set. Instead, we utilized the top 9 features identified by the Random Forest Recursive Feature Elimination in Phase I (input shape: 176,599 samples, 9 features). This focused the clustering algorithms on the strongest physiological signals (e.g., Wrist_EDA, Chest_Temp) rather than fitting to noise.

2. Methodology: Determining Optimal Clusters

We employed two heuristic methods to determine the optimal number of clusters (k):

- The Elbow Method: We plotted the Within-Cluster Sum of Squares (Inertia) against k.
    - Observation: The inertia curve showed a distinct "knee" (point of diminishing returns) at k=2. The inertia dropped from 51,734 (k=2) to 44,404 (k=3), but the rate of decrease slowed significantly thereafter.
- Silhouette Analysis: We calculated the Silhouette Score for k in [2, 6].
    - Result: The highest Silhouette Score was achieved at k=2 (Score: 0.3843).
    - Interpretation: A score of 0.38 indicates reasonable cluster separation. As k increased to 3, the score dropped sharply to 0.21. This confirms that the data naturally organizes into two distinct states, aligning perfectly with our binary Classification target (Stress vs. Baseline).

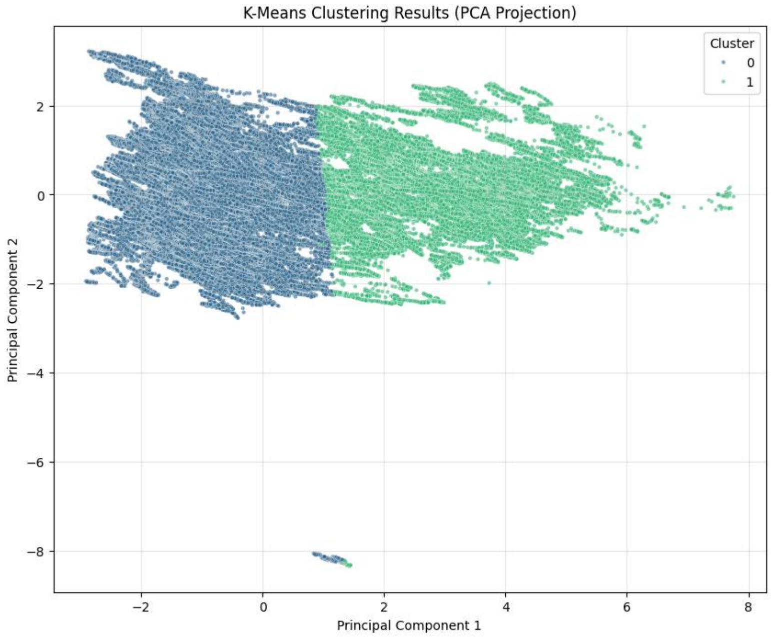


3. Algorithm Implementation & Comparison

We implemented and compared two distinct clustering algorithms: K-Means (geometric/distance-based) and Gaussian Mixture Models (GMM) (probabilistic/density-based). Performance was evaluated using both internal metrics (cluster cohesion) and external metrics (comparison to ground truth labels).

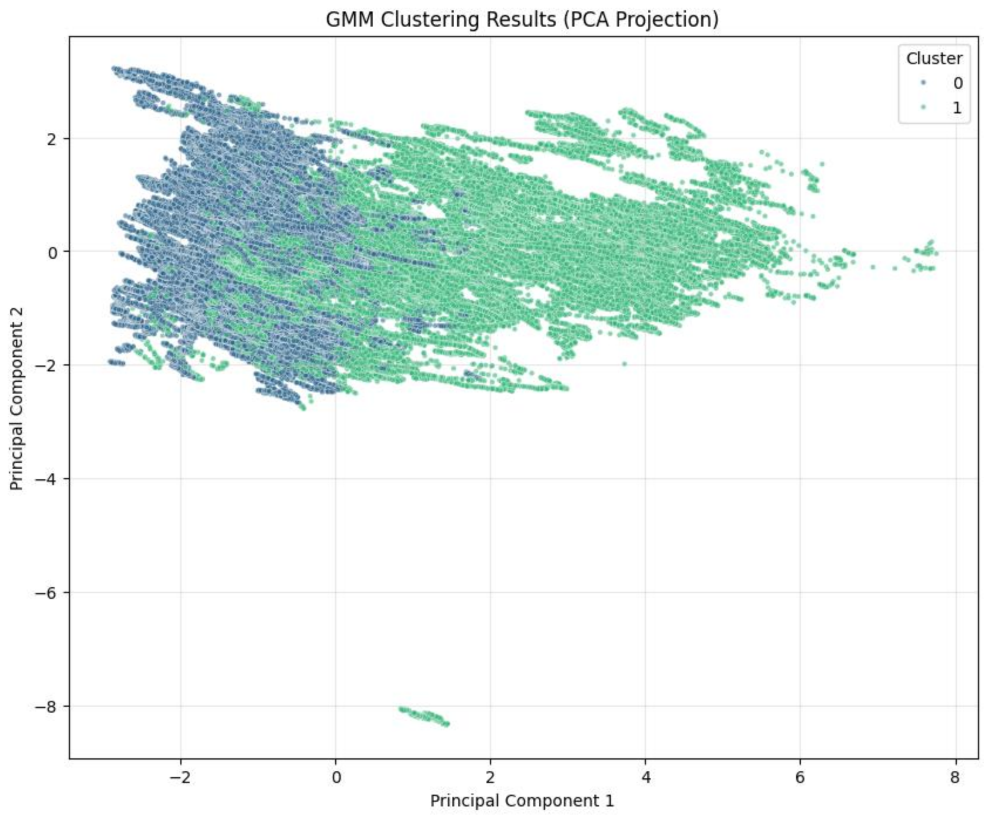
A. K-Means Clustering

- Method: K-Means partitions data by minimizing the Euclidean distance between data points and valid centroids.
- Performance:
  - Silhouette Score: 0.3825 (Indicates moderate separation).
  - Davies-Bouldin Index: 1.2006 (Lower is better; this indicates compact clusters).
  - Adjusted Rand Index (ARI): 0.8033. This is the critical finding. ARI measures similarity between the assigned clusters and the true labels (0=random, 1=perfect). An ARI of 0.80 implies the unsupervised K-Means algorithm effectively "rediscovered" the Stress and Baseline labels with 80% overlap accuracy.



B. Gaussian Mixture Models (GMM)

- Method: GMM assumes data points are generated from a mixture of Gaussian distributions, using Expectation-Maximization (EM) to fit means and covariances.
- Performance:
  - Silhouette Score: 0.2776 (Lower than K-Means).
  - ARI: 0.4663.
- Observation: GMM performed significantly worse than K-Means on this dataset. This suggests that the physiological clusters are likely compact and spherical (favoring K-Means) rather than elongated or overlapping (which GMM usually handles better). The complexity of GMM likely led to overfitting the covariance matrices given the 9-dimensional space.



4. Cluster Generalizability (LOSO Stability Test)

To ensure the clusters were not simply memorizing subject-specific idiosyncrasies, we conducted an independent research study using a Leave-One-Subject-Out (LOSO) Stability Test for clustering.

- Protocol: We trained K-Means centroids on N-1 subjects and assigned the held-out subject's data to those centroids. We then calculated the ARI for that held-out subject.
- Results: The average ARI across 15 validation folds was 0.8126,
- Conclusion: The high stability score (>0.8) indicates that the physiological definition of "Stress" is universal across subjects. The cluster centroids learned from one group of people accurately segmented the stress states of a completely new person.

5. Final Conclusion

The clustering analysis confirms the structural integrity of the dataset. The fact that K-Means (k=2) yielded an ARI of 0.80 provides strong evidence that Stress and Baseline states are physiologically distinct and linearly separable in the feature space. This unsupervised validation supports the high accuracy observed in the supervised Classification Phase (Phase III), confirming that the model performance was driven by robust features rather than label overfitting.

10. Recommendations

This section summarizes the critical insights derived from the classification and regression analyses and provides actionable recommendations for the deployment of stress detection systems.

a. Lessons Learned

The primary insight from this project is the paramount importance of Feature Engineering over Model Complexity. The fact that a linear model (Logistic Regression) outperformed complex ensembles (Stacking, Gradient Boosting) indicates that the rigorous preprocessing steps in Phase I—specifically subject-specific standardization, spectral energy extraction, and signal separation (SCL/SCR)—successfully linearized the decision boundary. Furthermore, the regression analysis revealed a fundamental limitation in physiological sensing: while Arousal (intensity) is directly linked to sympathetic activation (sweat, heart rate), Valence (positivity) is physiologically ambiguous and difficult to predict without contextual data.

b. Best Performing Classifier

Based on the comparative analysis using LOSO validation, Logistic Regression is recommended as the optimal classifier for this dataset.

- Performance: It achieved the highest F-Score (0.9277) and Sensitivity (94.67%).
- Justification: In health monitoring applications, Sensitivity (Recall) is the critical metric; it is preferable to accept a False Positive (false alarm) than to miss a genuine acute stress event (False Negative).
- Efficiency: Logistic Regression is computationally inexpensive (O(n)), making it ideal for deployment on battery-constrained wearable firmware, whereas the Stacking Classifier or Neural Network would incur significant latency and power costs for negligible performance gains.

c. Future Work & Performance Improvement

To improve classification and regression performance, future work should focus on:

- Non-Linear Regression for Valence: The failure of linear regression to predict Valence suggests the need for non-linear approaches, such as Random Forest Regressors or Long Short-Term Memory (LSTM) networks, to capture temporal dependencies.
- Contextual Data Integration: Incorporating external context (e.g., location, calendar events, physical activity type) could help disambiguate High Arousal/Positive states (Excitement) from High Arousal/Negative states (Stress).
- Domain Adaptation: Implementing "Personalized Calibration" layers where the model fine-tunes itself to a specific user's baseline over the first 24 hours of usage would likely reduce inter-subject variance.

d. Key Features Associated with the Target

The Bidirectional Stepwise Regression and Random Forest Feature Importance analysis consistently identified the following features as the strongest predictors of stress:

1. Wrist_EDA_mean: The most dominant predictor. Skin conductance is a direct correlate of sympathetic nervous system arousal.
2. Wrist_BVP_rate: Heart rate variability and frequency are critical indicators of acute stress response.
3. Chest_Temp_min: Variations in body temperature provided significant discriminative power, particularly for differentiating distinct emotional states.

e. Clustering in Feature Space

Unsupervised analysis confirmed that the optimal number of clusters in this feature space is k=2.

- Validation: This was supported by the Silhouette Score maximization (0.3843) and the Elbow Method.
- Observation: The clusters aligned with the ground truth labels with an Adjusted Rand Index (ARI) of 0.80. This confirms that "Stress" and "Baseline" are naturally distinct physiological states that can be separated even without supervised labels, validating the robustness of the derived features.

11. Appendix

All data, graphs and metrics are taken from attached files:

Phase1_part1.py

Phase1_part2.py

Phase2.py

Phase3_kfcv.py

Phase3_loso.py

Phase4.py

12. References

Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., & Van Laerhoven, K. (2018). Introducing WESAD, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (pp. 400-408). ACM. https://doi.org/10.1145/3242969.3242985