Stress Level Prediction

Data Mining Project Report

Nadine Safwat, Hussein Heggi, Mohamed Tarek

Github Link

Introduction

Stress is a major contributor to physical and mental health problems, including cardiovascular disease, anxiety, and weakened immune function. However, stress often goes undetected until it causes serious health issues. Traditional stress assessment relies on self-reported questionnaires or clinical evaluations, which are infrequent, subjective, and fail to capture stress as it occurs in real-time throughout daily life.
Without continuous, objective monitoring, individuals remain unaware of their stress patterns and triggers, missing opportunities for timely intervention through behavioral changes, relaxation techniques, or medical support.

Modern wearable devices offer a solution by continuously monitoring physiological signals that reflect the body's stress response. When stressed, the autonomic nervous system produces measurable changes in:

- Electrodermal Activity (EDA): Increased sweat gland activity
- Heart Rate: Changes in cardiac rhythm
- Blood Volume Pulse: Changes in circulation of blood in the extremities of the body
- Skin Temperature: Changes in skin temperature in the extremities of the body
- Movement Patterns: Activity changes captured by accelerometers

By applying data mining techniques to these multimodal sensor streams, we can develop automated systems that detect stress episodes in real-time

Project Objective

This project aims to build predictive models for stress detection using the Wearable Device Dataset from Induced Stress and Structured Exercise Sessions (v1.0.1). This dataset provides controlled, labeled physiological data from subjects undergoing standardized stress protocols, offering a foundation for training robust stress detection algorithms. Successful models would enable real-time stress alerts, personalized interventions, long-term pattern analysis, and early warning systems for stress-related health risks bridging wearable technology with actionable health insights.

Dataset Overview

The dataset contains physiological recordings from 31 subjects (18 male, 13 female) who participated in three experimental protocols: stress induction, aerobic exercise, and anaerobic exercise. Data was collected using the Empatica E4 wristband, which captures seven sensor channels: accelerometry (ACC), blood volume pulse (BVP), electrodermal activity (EDA), heart rate (HR), inter-beat interval (IBI), skin temperature (TEMP), and event tags. The sensors operate at different sampling rates; ACC at 32Hz, BVP at 64Hz, and EDA/TEMP at 4Hz. The data is segmented into 60-second windows with 30-second overlap, yielding approximately 6,500 labeled windows across all subjects. Stress labels are derived from self-reported stress scores on a 0-10 scale, which are then mapped to either four classes (no stress, low, moderate, high) or binary categories (stress vs. no stress).

Despite its utility for stress detection research, the dataset presents several significant challenges:

- Class Imbalance and Distribution Issues: 82% of samples represent no-stress conditions
- Data Quality and Coverage Problems: Only 27 out of 31 subjects have usable data
- Generalization Limitations: Primarily young adults aged 19-31 years, with 89% physically active participants.
- Label Reliability Concerns: Subjective nature of self-reports introduces inconsistency in ground truth labels

These limitations necessitate careful preprocessing, robust model evaluation strategies, and cautious interpretation of results, particularly regarding the model's ability to generalize to diverse populations and real-world stress scenarios.

Methodology

Model Architecture: Phase-Aware Hybrid LSTM-ResNet

This project employs a deep learning architecture that combines convolutional feature extraction, temporal modeling, and statistical feature fusion to predict stress levels from multimodal physiological signals. Four variants of this core architecture were developed (Models 5-8), differing primarily in their preprocessing strategies and loss functions while maintaining the same fundamental structure.

The architecture processes raw time-series signals through six interconnected components. First, a multi-scale sequence encoder captures features at different temporal scales using three parallel convolutional pathways. The short-term pathway (kernel size 3, dilation 1) captures immediate physiological responses, the medium-term pathway (kernel size 7, dilation 2) captures evolving stress patterns, and the long-term pathway (kernel size 15, dilation 4) captures sustained physiological changes. Each pathway consists of a 1D convolutional layer followed by batch normalization, ReLU activation, and two ResNet blocks with skip connections to prevent gradient degradation. The three scales are concatenated and merged through a 1×1 convolution to produce 128-channel representations.

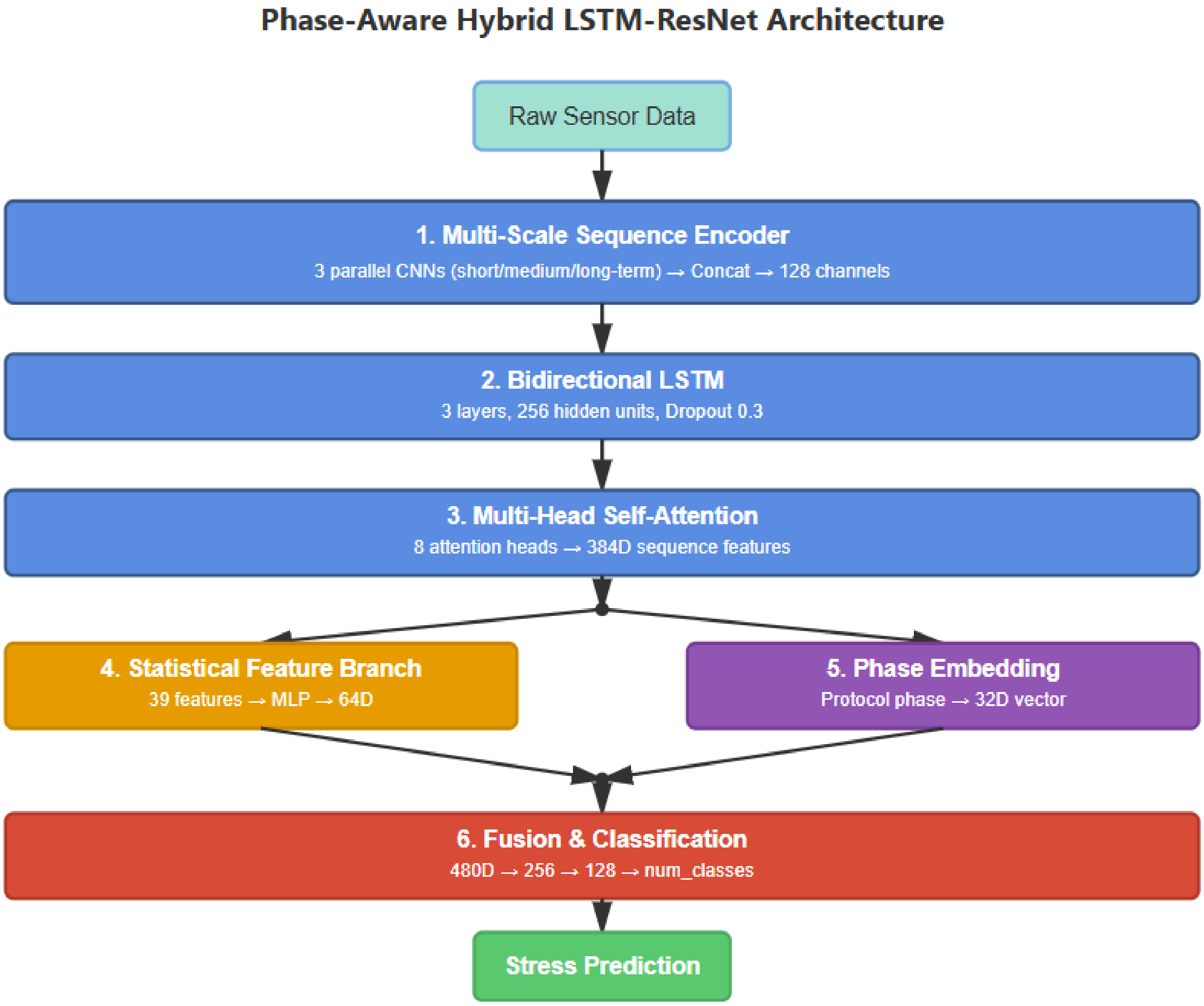
Second, the merged multi-scale features are processed by a 3-layer bidirectional LSTM with 256 hidden units (128 forward and 128 backward). This bidirectional structure captures both past-to-future and future-to-past temporal dependencies, learning sequential stress patterns from the physiological signals. Dropout of 0.3 is applied between layers for regularization.

Third, an 8-head self-attention mechanism enables the model to focus on different aspects of the time series simultaneously, learning which temporal segments are most important for stress detection while reducing noise from irrelevant physiological fluctuations.

Fourth, a parallel statistical feature branch processes 39 engineered features extracted from the raw signals, including mean, standard deviation, minimum, maximum, percentiles, and cross-signal correlations. This branch consists of a 2-layer MLP (128 → 64 units) with batch normalization and dropout (0.3, 0.2), providing interpretable physiological metrics alongside the learned representations from the sequence encoder.

Fifth, protocol phase information (such as Stroop test, TMCT, rest periods, aerobic exercise) is embedded into 32-dimensional vectors. This phase embedding allows the model to learn phase-specific stress signatures and account for contextual differences, such as distinguishing stress during cognitive tasks from stress during physical exercise.

Finally, the fusion and classification component concatenates the sequence features (384 dimensions), statistical features (64 dimensions), and phase embedding (32 dimensions) into a 480-dimensional fusion vector. This vector is processed through a 3-layer classifier (480 → 256 → 128 → num_classes) with dropout regularization (0.3, 0.2) before outputting class logits for stress level prediction.


Phase-Aware Hybrid LSTM-ResNet Architecture
Raw Sensor Data
1. Multi-Scale Sequence Encoder
3 parallel CNNs (short/medium/long-term) → Concat → 128 channels
2. Bidirectional LSTM
3 layers, 256 hidden units, Dropout 0.3
3. Multi-Head Self-Attention
8 attention heads → 384D sequence features
4. Statistical Feature Branch
39 features → MLP → 64D
5. Phase Embedding
Protocol phase → 32D vector
6. Fusion & Classification
480D → 256 → 128 → num_classes
Stress Prediction

Data Preprocessing

All models share a common foundation of enhanced signal processing and normalization before diverging in their augmentation and loss function strategies.

Signal Enhancement and Channel Engineering:

Raw sensor signals were first resampled to a unified 4 Hz sampling rate to standardize temporal resolution across different sensors. The preprocessing pipeline then extracts 30 enhanced channels from the original sensor data to capture multiple physiological perspectives:

- Raw signals: EDA, Temperature, ACC (3-axis), BVP

- First derivatives: $\partial EDA/\partial t$, $\partial Temp/\partial t$, $\partial ACC/\partial t$, $\partial BVP/\partial t$ to capture rate of change
- EDA decomposition: Tonic component (10-second moving average) and phasic component (EDA - Tonic) to separate slow baseline shifts from rapid responses
- EDA acceleration: $\partial^2 EDA/\partial t^2$ to capture sudden arousal changes
- EDA moving averages: 5-second and 15-second windows for multi-scale smoothing
- BVP envelope: Hilbert transform to extract amplitude modulation
- Cross-signal products: EDA × BVP to capture arousal-cardiovascular interactions
- Smoothed ACC: 3-second moving average to reduce motion artifacts
- Respiratory signal: Extracted from BVP fluctuations
- Sample entropy signal: Local complexity measure indicating physiological irregularity
- Wavelet channels: Additional time-frequency domain features

Subject-Specific Baseline Normalization:

Individual physiological baselines vary significantly across subjects, with some individuals showing naturally higher EDA or heart rate than others. To address this, subject-specific normalization was applied. For each subject, mean and standard deviation were computed from their "no_stress" windows, establishing their personal baseline. All windows for that subject were then normalized as (x - baseline_mean) / baseline_std. This normalization handles individual physiological differences and ensures the model learns relative stress responses rather than absolute signal magnitudes.

Train-Test Splitting:

StratifiedGroupKFold with 5 folds was used for splitting, with fold 0 selected for evaluation. This approach ensures that each subject appears exclusively in either the training set or test set, preventing data leakage and providing a realistic assessment of generalization to unseen individuals. The stratification maintains class distribution across folds despite severe imbalance.

Model Variants and Training Strategies

Four model variants were developed to explore different approaches to handling the severe class imbalance and to compare 4-class versus binary classification tasks.

Model 5: 4-Class Classification with Conservative Temporal Augmentation

Model 5 addresses the 4-class stress prediction task (no stress, low, moderate, high) using conservative temporal augmentation focused on the most underrepresented class. Low-stress samples, comprising only 1.3% of the dataset, receive 2× augmentation, while moderate-stress and high-stress samples receive no augmentation. Each augmentation applies time warping (0.95-1.05× speed variation), Gaussian noise (3% of signal standard deviation), and temporal shifts (±10% of window length). The model is trained for 50 epochs with batch size 32 using the AdamW optimizer (learning rate 5e-4, weight decay 1e-4) and cosine annealing with warm restarts. Focal loss ($\gamma=2.0$) with class weights derived from inverse frequency down-weights easy no-stress examples while up-weighting hard minority class examples. Label smoothing of 0.05 and exponential moving average (decay=0.995) are applied for stable predictions.

Model 6: 4-Class Classification with Asymmetric Focal Loss

Model 6 tackles the same 4-class task but completely disables temporal augmentation, instead relying on a sophisticated loss function to handle imbalance. Asymmetric Focal Loss assigns per-class gamma values that prioritize learning from the most challenging classes: low-stress ($\gamma=3.0$), high-stress ($\gamma=2.5$), moderate-stress (γ=2.0), and no-stress (γ=1.0). Higher gamma values increase focus on misclassified examples within that class. To compensate for the absence of augmentation, Model 6 trains for 60 epochs, 10 more than Model 5, with otherwise identical hyperparameters. This design trades data augmentation for loss-driven imbalance handling.

Model 7: Binary Classification with Conservative Temporal Augmentation

Model 7 simplifies the task to binary classification by mapping low, moderate, and high stress to a single "stress" class (class 1) and retaining "no_stress" as class 0. This creates an approximately 82% no-stress versus 18% stress distribution. The stress class receives 2× conservative temporal augmentation using the same techniques as Model 5, while no-stress samples receive no augmentation. Training configuration mirrors Model 5: 50 epochs with Focal loss (γ=2.0) and binary class weights. The binary simplification is expected to yield higher overall performance than 4-class models, though it sacrifices granular stress level information.

Model 8: Binary Classification with Asymmetric Focal Loss

Model 8 combines binary classification with asymmetric loss, paralleling the Model 6 philosophy. Augmentation is disabled, and Asymmetric Focal Loss assigns γ=1.0 to the easy majority class (no-stress) and γ=2.5 to the harder minority class (stress). The model trains for 60 epochs with binary class weights from inverse frequency. This variant is expected to achieve the highest recall for stress detection by combining binary task simplification with loss-driven learning focused on the minority class.

Preprocessing Comparison:

Technique	Model 5	Model 6	Model 7	Model 8
Task	4-class	4-class	Binary	Binary
Temporal Augmentation	2× low-stress	Disabled	2× stress	Disabled
Loss Function	Focal (γ=2.0)	Asymmetric Focal	Focal (γ=2.0)	Asymmetric Focal
Class Gammas	Uniform 2.0	[2.5, 3.0, 2.0, 1.0]	Uniform 2.0	[1.0, 2.5]
Training Epochs	50	60	50	60
Philosophy	Feature-rich + Augmentation	Loss-driven imbalance handling	Binary + Augmentation	Binary + Loss-driven

The rationale behind this design is to compare two distinct philosophies for handling severe class imbalance. Models 5 and 7 trust rich feature engineering and conservative augmentation to maintain signal quality while providing more training examples for minority classes. Models 6 and 8 minimize data manipulation, letting asymmetric loss functions guide the learning process toward minority classes through increased penalties for misclassifications. Additionally, Models 7 and 8 test whether binary simplification can trade granularity for improved overall performance in stress detection.

Results and Performance Analysis

Overall Performance Summary

The experiments across four model variants demonstrate both the challenges and successes of automated stress detection in the presence of severe class imbalance. The best performing model achieved 66.4% macro F1-score (Model 8: Binary Classification with Asymmetric Focal Loss). This performance was achieved despite significant dataset limitations including a small subject pool, severe class imbalance (82% no-stress samples), and data quality issues.

Model	Task	Accuracy	Macro F1
Model 5	4-class	85.9%	54.3%
Model 6	4-class	80.3%	41.8%
Model 7	Binary	82.8%	59.0%
Model 8	Binary	80.7%	**66.4%**

Model 5 achieved the highest accuracy at 85.9%, but this metric can be misleading in imbalanced classification tasks as it is often inflated by correctly predicting the majority class. Macro F1-score, which equally weighs performance across all classes regardless of their frequency, provides a more meaningful measure of balanced performance. By this metric, Model 8 clearly outperforms all other variants.

Key Findings

- Binary Models Outperform 4-Class Models

Binary classification models (Models 7 and 8) consistently achieved higher macro F1-scores than their 4-class counterparts (Models 5 and 6). Model 8 showed a 24.6% improvement over Model 6 (66.4% vs. 41.8% macro F1), while Model 7 improved 4.7% over Model 5 (59.0% vs. 54.3%).

This performance advantage stems from collapsing low, moderate, and high stress into a single "stress" class, which provides more training examples per class and reduces data scarcity. The binary task establishes a clearer decision boundary between stress and no-stress states compared to distinguishing between multiple stress intensity levels. Additionally, binary classification eliminates label ambiguity between adjacent stress levels, where the distinction between low and moderate stress may be subjective.

The trade-off for this improved performance is a loss of granularity. Binary models cannot distinguish between mild stress that might be managed through self-regulation techniques and severe stress that could warrant clinical intervention. For applications requiring stress intensity assessment, such as clinical monitoring or adaptive intervention systems, this limitation is significant.

- Asymmetric Focal Loss Shows Task-Dependent Effectiveness

The asymmetric focal loss strategy produced contrasting outcomes across the two tasks. In the 4-class setting (Model 6), it underperformed the augmentation-based Model 5 by 12.5% macro F1. However, in the binary setting (Model 8), it outperformed the augmentation-based Model 7 by 7.4% macro F1.

This divergence suggests that asymmetric focal loss is more effective for simpler classification tasks with clearer decision boundaries. In the 4-class scenario with extreme imbalance, particularly the low-stress class comprising only 1.3% of samples. The per-class gamma weighting may have over-emphasized rare classes, leading to overfitting on the few available minority class examples or creating unstable training dynamics. In contrast, the binary task benefits from asymmetric loss because the "stress" class, while still a minority at 18%, provides enough examples for the heightened focus (γ=2.5) to guide effective learning without overfitting.

- Model 8 Achieves Best Overall Performance

Model 8's 66.4% macro F1-score establishes it as the top-performing model for balanced stress detection. Its combination of binary task simplification and asymmetric focal loss creates an effective balance between stress and no-stress class performance. The asymmetric focal loss with γ=2.5 for the stress class effectively handles the binary imbalance, while the 80.7% accuracy demonstrates reasonable overall performance. This makes Model 8 particularly suitable for practical applications including real-time wearable stress monitoring, binary alert systems that notify users when stress is detected, and applications prioritizing balanced class performance over raw accuracy.

However, Model 8 has notable limitations. It cannot distinguish between different stress intensity levels, which may be important for tailoring intervention strategies. The 80.7% accuracy, while respectable given dataset constraints, indicates room for improvement before deployment in critical applications. Additionally, the model may confuse physical activity with stress, as both produce elevated electrodermal activity and heart rate, a challenge inherent to physiological stress detection.

- Dataset Limitations Constrain Performance

All models were fundamentally constrained by dataset characteristics. The small subject pool of only 28 training subjects and 8 test subjects restricts the model's ability to learn generalizable patterns across diverse individuals. Severe class imbalance with 82% no-stress samples makes minority class learning difficult despite specialized techniques. Individual variability in physiological stress responses means some subjects show minimal physiological changes despite self-reported stress, creating inconsistencies in the training signal. Data quality issues including sensor disconnections, missing data, and incomplete protocols reduce the amount of clean training data available. These limitations explain why the models achieved 54-66% macro F1 rather than the higher performance levels (often 75-90% F1) reported in larger, more balanced stress detection datasets with controlled laboratory conditions.

- Common Classification Challenges

Analysis of classification errors reveals several primary sources of mistakes that affect all models. Physical activity confusion represents a major challenge, as exercise-induced elevation in EDA and heart rate closely mimics physiological stress responses, causing false positive stress predictions during physical activity periods. Transition windows that span stress onset or offset periods contain ambiguous labels because the 60-second windows with 30-second overlap may capture both stressed and non-stressed states within a single labeled segment. Individual variability means some subjects exhibit minimal physiological responses despite reporting stress, creating training examples where the model observes "no-stress" physiological patterns labeled as "stress." Sensor disconnections and missing data due to Bluetooth issues or improper device placement reduce signal quality and create gaps in coverage. Finally, the small sample size, particularly for minority classes, restricts the model's ability to learn robust representations of stress patterns, as it has fewer examples from which to extract generalizable features.

Conclusion

This project demonstrates the feasibility of automated stress detection from wearable sensor data using deep learning techniques, while highlighting the significant challenges posed by severe class imbalance and limited training data. The Phase-Aware Hybrid LSTM-ResNet architecture, combining multi-scale convolutional feature extraction, bidirectional temporal modeling, and statistical feature fusion, achieved a best performance of 66.4% macro F1-score on binary stress classification.

The comparison of four model variants reveals several important insights for practical stress detection systems. Binary classification substantially outperforms 4-class classification, with Model 8 (binary with asymmetric focal loss) achieving the strongest balanced performance across stress and no-stress classes. This suggests that for real-world applications, trading granular stress intensity levels for improved detection reliability may be worthwhile. The effectiveness of asymmetric focal loss proved task-dependent, working well for binary classification but struggling with extreme multi-class imbalance, indicating that loss function design must be carefully matched to problem complexity.

Despite these achievements, the project reveals fundamental limitations that constrain performance. The small subject pool limits generalization to diverse populations. Severe class imbalance (82% no-stress samples) makes minority class learning difficult even with specialized techniques. Data quality issues including sensor disconnections and incomplete protocols reduce the amount of reliable training data. Individual variability in physiological stress responses creates inconsistencies between self-reported stress and measurable physiological changes. These constraints explain why performance reached 54-66% macro F1 rather than the 75-90% reported in larger, more controlled datasets.

This work demonstrates that automated stress detection from wearable sensors is achievable but requires careful attention to class imbalance, individual variability, and the inherent ambiguity in physiological stress signals. While current performance levels are promising for consumer wellness applications, further research and larger datasets are needed before deployment in clinical or safety-critical contexts.