# Stress Prediction Using Multimodal Data and Machine Learning

**Claire Najjuuko**
c.najjuuko@wustl.edu

**Praneel Panchigar**
p.a.panchigar@wustl.com

## Abstract

Stress prediction and monitoring is critical due to the increasing prevalence of stress-related health issues. This study explores the use of wearable sensor data and survey responses to develop machine learning models for predicting stress levels. By integrating multimodal datasets, including mobile sensor features, and weekly EMA surveys, we classify stress into "Low," "Medium," and "High". Singular Spectrum Analysis was employed to process time-series data, while models such as Random Forest, and Logistic Regression were trained to identify stress patterns. AUROC and AUPRC were used to assess performance. Results indicate that mobile sensor data only marginally improves model's predictive accuracy of stress. This work contributes to the growing body of research on stress prediction, paving the way for real-time, personalized stress monitoring systems. Future research should invest in developing larger datasets that can leverage deep learning methods for time series mobile sensor data.

## 1 Introduction

Mobile sensing technologies have demonstrated the ability to monitor and predict many aspects of physical health, mental health, and even work performance [1]. Stress is a physiological response to various stimuli, making it suitable for monitoring using such mobile sensing technologies. Stress is a significant contributor to physical and mental health issues, including chronic illnesses such as cardiovascular diseases [2] and depression [3]. With the increasing prevalence of stress-related health problems, early detection of high stress would allow more effective management, enhancing overall well-being [4]. Despite advancements in mobile sensing technologies, stress prediction using passive data remains under explored [5]. This research addresses this gap by leveraging the Generalization of LOngitudinal BEhavior Modeling (GLOBEM) dataset—a longitudinal collection of short-survey, wearable and mobile sensor data from 497 participants. This dataset offers a unique opportunity to model and predict stress patterns effectively [6].

This study employs machine learning techniques to integrate multi-modal data, including mobile sensor features, and user-reported stress levels collected through ecological momentary assessment (EMA), to predict future stress levels. Given the limited size of our dataset, we utilize simple machine learning models such as linear regression (LR) and support vector machine (SVM) to predict students' raw stress scores and further classify students into three levels: low, medium, and high stress at the end of the study. To extract meaningful features from the time series mobile sensor data, we utilize an end-to-end feature engineering pipeline proposed in a previous study [7]. Results demonstrate the efficacy of combining diverse data modalities and machine learning techniques for accurate stress prediction. These findings contribute to the development of passive, scalable stress monitoring tools, paving the way for real-time, personalized interventions that promote overall well-being.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

## 2 Related Work

The integration of machine learning techniques with sensor data has significantly advanced detection methodologies for various mental health conditions such as depression, anxiety, and stress. In particular, wearable sensors, capturing physiological signals such as heart rate variability (HRV), electrodermal activity (EDA), and electrocardiogram (ECG) readings, provide rich datasets for analysis and detection of stress. Machine learning algorithms, including decision trees, random forests (RF), and XGBoost (XGB), have been effectively employed to analyze this data, enabling personalized feedback and interventions for users [8]. Deep neural networks have also been widely adopted for stress detection, leveraging physiological signals to predict stress levels. These models analyze data collected from sensors attached to the human body, offering rapid and accurate stress detection [9].

Singular Spectrum Analysis (SSA) has emerged as a powerful tool for time-series analysis, capable of decomposing signals into components such as trends, oscillations, and noise. This decomposition facilitates effective feature extraction, enhancing the performance of machine learning models in applications such as hyper spectral imaging and stress prediction [10]. The application of SSA in time-series analysis has been extensively discussed, emphasizing its versatility in tasks such as smoothing, filtration, noise reduction, and extraction of trends and periodicities. This versatility makes SSA a valuable tool in preprocessing data for stress prediction models [11], [7].

## 3 Methods

### 3.1 The Dataset

In this study, we utilized the GLOBEM dataset which was collected over four-years (2018-2021), from 497 college students. The study collected data for approximately 10 weeks annually. Each student installed a mobile app on their phones and wore a fitness tracker, which together passively captured multiple sensor streams 24/7. These streams included location, phone usage, calls, bluetooth connections, physical activity, and sleep behavior. Additionally, participants completed weekly short surveys and biannual comprehensive surveys (pre- and post-study) that assessed various dimensions of health behaviors, emotional well-being, and mental health. In this study, we aimed to predict the end of semester stress, and used post surveys as ground truths. Additionally, we further aimed to assess the generalizability of our model on making predictions on a novel dataset. We excluded the 2018 dataset because it was missing most of the features considered for model development. We used the 2019 and 2020 dataset for model training and internal validation, and the 2021 dataset for external validation.

### 3.2 Exploratory data analysis (EDA) and handling missing data

We performed data cleaning, preprossessing, and feature engineering to prepare the dataset. We performed shape analysis, missing data checks, and obtained summary statistics to assess data quality. We filtered each student's EMA and sensor data to include only 10 weeks of data which improved data quality. Additionally, we dropped all calls features because they have more than 50% missingness.

### 3.3 EMA feature extraction

The EMA surveys included data on anxiety; depression; negative and positive affect. We excluded students who had more than 10% missing data. EMA data per student was aggregated to obtain 10 summary statistics including mean, median, maximum, minimum, standard deviation, skewness, kurtosis, and interquartile range. Additionally, autocorrelation and root mean squared difference (RMSD) were computed for to capture temporal patterns.

### 3.4 Mobile sensors feature extraction

The GLOBEM dataset includes over 60 sensor features with multiple time segments [6]. For our analysis, we only considered the allday (daily) features. To simplify the analysis, we further selected only 22 semantic time series features. These are grouped as screen usage (total duration of unlock episodes, total number of unlock episodes, minutes until first unlock episode in the morning);

bluetooth (total number of scans, total number of unique scanned devices, total number of scans for most scanned unique device); sleep (duration asleep during main sleep (minutes), duration awake during main sleep (minutes), count of asleep episodes in main sleep, count of awake episodes in main sleep, average sleep efficiency, duration user stayed in bed (minutes)); location (time spent at home, radius of gyration (area covered - meters), number of significant visited locations, circadian routine (measure of routine), fraction of day spent in a pause); and steps (total steps in a day, count of sedentary bouts, count of active bouts, duration of sedentary bouts (minutes), duration of active bouts (minutes)).

We used SSA to extract features from the time series sensor data. We excluded students who had more than 20% missing data and used backward fill to impute missing values for the rest of the data prior to applying SSA. SSA is a useful tool to extract information from time series data via spectral decomposition [7]. SSA decomposes time series data into a sum of interpretable components, such as trend, oscillation, and noise. It consists of three distinct stages: embedding, decomposition, and reconstruction. The embedding stage, transforms the original time series into a series of lagged vectors using a specific window size (7 in our case), creating a trajectory matrix. Next, the decomposition stage applies singular value decomposition to the trajectory matrix, which separates the matrix into a set of rank-one matrices corresponding to different signal components such as trends or cycles. Finally, the reconstruction stage aggregates these rank-one matrices into reconstructed components. We used only the first component for reconstruction which simplifies the data to its most dominant trend, effectively filtering out lesser oscillations and noise. We then fitted spline and linear models on the extracted trend, and extracted 10 high-level features describing the original 71-day long time series per participant. The 10 extracted features were spline features (slope mean, slope variance, curvature mean, curvature max, number of inflection points, total variation, and absolute area under the curve) and linear features (mean, slope, and variance). Together, these derived features represented the temporal dynamics of each sensor modality.

The end-to-end feature engineering pipeline is presented in Figure 1 which includes daily feature extraction, aggregation, handling of missing data, and feature selection.

### 3.5 Deep Learning Models

As part of our exploration, we implemented well known deep learning models to process and extract meaningful information from the time series data. A convolutional neural network (CNN) with three convolutional layers, pooling layers, and dropout regularization to prevent overfitting, was implemented. Inputs were structured as [batch size, time steps, features. A long short term model (LSTM) with a single hidden layer (hidden size = 64) was also implemented to capture temporal dependencies in the sensor data. Inputs were structured as [batch size, features, time steps]. Adam optimizer was used for both models, and evaluation done using adjusted R-Squared (R2). Despite its ability to model sequences, the LSTM exhibited poor generalization, often defaulting to zero predictions and achieving R2 of 0.000. Similarly, the CNN model exhibited poor predictive ability, consistently predicting a narrow range of scores (15–25) and achieving R2 of 0.005. Due to these limitations, we discontinued the use of these models for further predictive modeling.

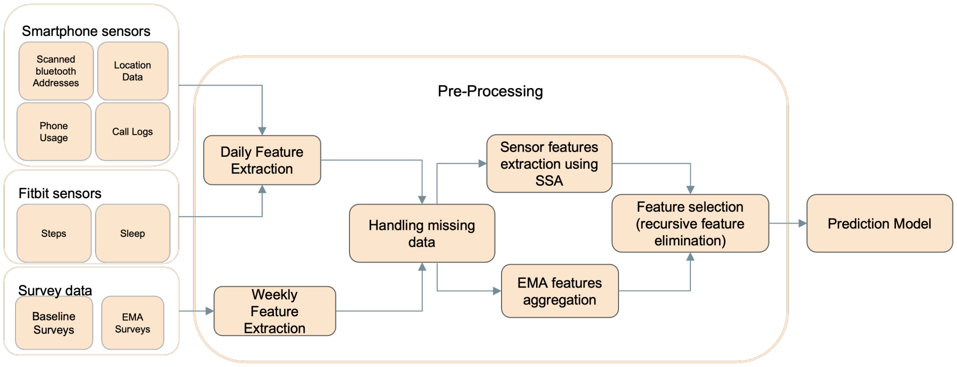

Figure 1: End-to-end feature engineering pipeline.

### 3.6 Hierarchical Feature Selection

A hierarchical feature selection pipeline was implemented to select the most relevant features for stress prediction from each data source. This pipeline utilized recursive feature elimination (RFE) to iteratively select features across multiple stages. Initially, sensor features consisting of 220 spline and linear attributes were reduced to 121 and 17 key attributes for the regression and classification settings respectively, by applying RFE. Similarly, weekly EMA features, originally comprising 52 statistical variables, were subjected to RFE, resulting in selection of 42 and 34 key attributes in the regression and classification settings respectively. For the pre-survey data, 15 and 16 features were selected in the regression and classification settings respectively through a separate RFE process. In the final stage of the pipeline, the selected features from the three sources: pre-survey, EMA, mobile sensors, were combined, and RFE applied. This resulted in a final feature set of 66 and 27 predictive attributes in the regression and classification settings respectively that were used for the final predictive model.

### 3.7 Regression and Classification Modeling

Regression modeling was initially pursued to predict continuous stress scores using various approaches. LR was implemented as a baseline model to evaluate linear trends in the data. To address multicollinearity and mitigate overfitting, Lasso and Ridge regression were employed as regularized techniques. Ensemble-based methods such as RF Regressor were utilized to capture non-linear relationships within the dataset, while XGB Regressor offered a high-performance gradient-boosted tree approach tailored for complex data. SVM Regressor (SVR), a kernel-based method, was also explored to model non-linear patterns. However, despite rigorous optimization efforts, regression modeling was hampered by the clustered nature of stress scores, which were predominantly concentrated in the medium range (14–26). This limited variability resulted in predictions that failed to generalize effectively.

Classification modeling was adopted to better align the stress prediction task with clinical and practical applications. Stress scores were discretized into three categories: low (0–13), medium (14–26), and high (27–40). These groupings reflect meaningful distinctions that facilitate clinical decision-making and intervention strategies. Four classification models were implemented and compared. Logistic Regression (LoR) served as a simple baseline model, while RF Classifier, an ensemble-based method, was employed to capture complex feature interactions. XGB Classifier, a gradient-boosted tree model, was selected for its robustness in handling imbalanced datasets. SVM Classifier (SVC) was also utilized, leveraging its kernel-based approach to separate non-linear classes effectively.

## 4 Experiments and Results

### 4.1 Model training and internal validation

Using the 2019 and 2020 dataset, we employed repeated 10-fold cross validation (with 10 repeats) to train and evaluate each of the models. Hyperparameter tuning was performed using grid search to optimize parameters such as tree depth, learning rate, and regularization strength. For the regression models, assessment of model performance was done using the root mean squared error (RMSE) and the adjusted R-Squared (R2). For the classification models, assessment of model performance was done using the AUROC and the AUPRC.

#### 4.1.1 Evaluation of different data modalities

For each setting; regression and classification, we trained and evaluated machine learning models using individual data sources, in order to ascertain the predictive ability of individual data modalities. The predictive performance for each modality is shown in Table 1 and Table 2 for regression and classification models respectively. In both cases, we observe that the survey features (pre and EMA), have a higher predictive ability compared to the mobile sensor features when used in isolation, as indicated by the R2 and AUPRC for regression and classification tasks respectively.

We further assess the model performance using combined modalities, that is pre and EMA; EMA and mobile sensor; and pre, EMA, and mobile sensor data. See Table 3 and Table 4 for regression and classification models results respectively. For the regression task, we observe that best model performance is achieved by Lasso model using the pre and EMA data modalities. For the classification models, we observe an incremental improvement when we combine all data sources as indicated by the higher AUPRC and AUROC achieved by the RF model in that setting. We observe that ultimately combining all three modalities, we obtain the best predictive performance for the classification task.

Table 1: Regression model performance for individual data sources

| Models | PRE | | EMA | | SENSOR | |
|---|---|---|---|---|---|---|
| | **RMSE** | **R2** | **RMSE** | **R2** | **RMSE** | **R2** |
| LR | 5.033 (0.042) | 0.441 (0.009) | 5.233 (0.067) | 0.337 (0.017) | inf | inf |
| Lasso | 5.057 (0.050) | 0.436 (0.011) | 4.871 (0.024) | 0.426 (0.006) | **6.426 (0.024)** | **0.037 (0.007)** |
| Ridge | 5.041 (0.032) | 0.440 (0.007) | 4.917 (0.032) | 0.415 (0.007) | inf | inf |
| RF | 5.071 (0.029) | 0.433 (0.006) | 4.837 (0.029) | 0.434 (0.007) | 6.472 (0.047) | 0.023 (0.014) |
| XGB | 5.411 (0.094) | 0.354 (0.023) | 5.258 (0.089) | 0.331 (0.023) | 6.980 (0.129) | -0.137 (0.042) |
| SVM | **5.030 (0.033)** | **0.442 (0.007)** | **4.832 (0.029)** | **0.435 (0.007)** | 6.545 (0.040) | 0.001 (0.012) |

Table 2: Classification model performance for individual data sources

| Models | PRE | | EMA | | SENSOR | |
|---|---|---|---|---|---|---|
| | **AUROC** | **AUPRC** | **AUROC** | **AUPRC** | **AUROC** | **AUPRC** |
| LR | 0.808 (0.004) | 0.647 (0.007) | 0.755 (0.008) | 0.550 (0.012) | 0.627 (0.009) | 0.442 (0.011) |
| RF | **0.817 (0.004)** | **0.664 (0.008)** | **0.810 (0.004)** | **0.613 (0.011)** | 0.623 (0.012) | 0.439 (0.011) |
| XGB | 0.804 (0.005) | 0.656 (0.008) | 0.756 (0.015) | 0.558 (0.021) | **0.641 (0.015)** | **0.464 (0.011)** |
| SVC | 0.795 (0.005) | 0.637 (0.012) | 0.741 (0.010) | 0.525 (0.019) | 0.592 (0.016) | 0.419 (0.014) |

Table 3: Regression models performance for combined data sources

| Models | PRE AND EMA | | EMA AND SENSOR | | PRE, EMA, AND SENSOR | |
|---|---|---|---|---|---|---|
| | **RMSE** | **R2** | **RMSE** | **R2** | **RMSE** | **R2** |
| LR | 5.002 (0.084) | 0.397 (0.020) | inf | inf | inf | inf |
| Lasso | **4.639 (0.046)** | **0.482 (0.010)** | **4.764 (0.038)** | **0.427 (0.009)** | 4.617 (0.035) | 0.463 (0.008) |
| Ridge | 4.663 (0.059) | 0.476 (0.013) | inf | inf | inf | inf |
| RF | 4.698 (0.025) | 0.469 (0.006) | 4.857 (0.040) | 0.405 (0.009) | **4.602 (0.061)** | **0.467 (0.014)** |
| XGB | 4.998 (0.081) | 0.398 (0.019) | 5.110 (0.067) | 0.341 (0.017) | 4.977 (0.073) | 0.376 (0.018) |
| SVM | 4.667 (0.056) | 0.475 (0.013) | 6.644 (2.047) | -0.220 (0.741) | 8.883 (2.390) | -1.131 (0.977) |

### 4.2 Model external validation

The 2021 dataset was used to perform external validation of the best classifier – RF Classifier. For this step, only the 27 combined features selected in the previous stage of model training and evaluation, were used. The RF model was trained on the entire 2019 and 2020 dataset (N = 205), and tested on the entire 2021 dataset (N = 155). The model achieved an **AUROC of 0.810** and **AUPRC of 0.690**.

### 4.3 Feature Importance Analysis

To enhance model interpretability, feature importance was assessed using SHapley Additive exPlanations (SHAP) values [12]. SHAP summary plots were generated for the best model highlighting the relative contributions of individual features. Figure 2 shows the 15 most important features for predicting all classes. We notice that the 15 most important features are from the baseline surveys and the weekly EMA surveys only. The most important baseline and EMA features are those describing the mental well being of the student throughout the study. We further interrogate the model's predictions of the "High Stress" class. The 15 most important features include features from baseline surveys, weekly EMA surveys, and the daily mobile sensors. However, low importance is attributed to the sensor features. The key sensor features revealed are sleep duration, physical activity. Particularly, shorter durations of main sleep and higher variation of time spent in bed generally increase the likelihood of high stress, and lower counts of sedentary bouts are associated with an increased likelihood of high stress. By contextualizing these predictors, SHAP analysis provided actionable insights into behavioral patterns associated with stress.

Table 4: Classification models performance for combined data sources

| Models | PRE AND EMA | | EMA AND SENSOR | | PRE, EMA, AND SENSOR | |
|---|---|---|---|---|---|---|
| | AUROC | AUPRC | AUROC | AUPRC | AUROC | AUPRC |
| LR | **0.805 (0.006)** | **0.620 (0.011)** | 0.773 (0.011) | 0.573 (0.017) | 0.840 (0.007) | 0.660 (0.011) |
| RF | 0.823 (0.007) | 0.627 (0.011) | **0.805 (0.011)** | **0.625 (0.018)** | **0.847 (0.009)** | **0.673 (0.018)** |
| XGB | 0.792 (0.011) | 0.612 (0.019) | 0.772 (0.009) | 0.580 (0.018) | 0.829 (0.010) | 0.647 (0.020) |
| SVC | 0.794 (0.012) | 0.600 (0.020) | 0.758 (0.016) | 0.553 (0.022) | 0.801 (0.013) | 0.614 (0.022) |

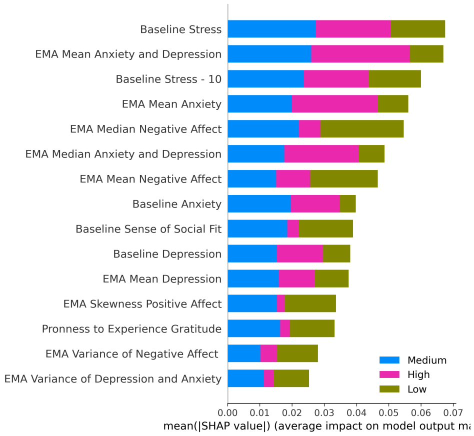

Figure 2: Shap summary plot for all classes

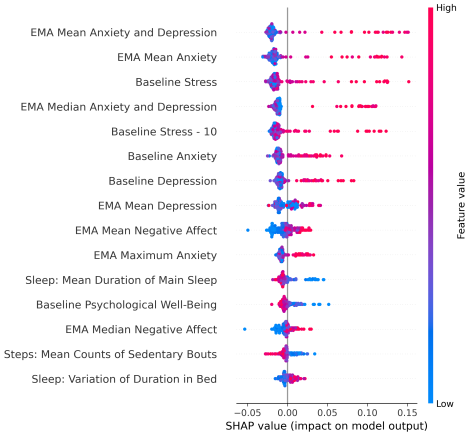

Figure 3: Shap summary plot for "High Stress" class

## 5 Discussion

The present study demonstrates the potential of multi-modal machine learning approaches in predicting student stress levels using a comprehensive longitudinal dataset. Our findings highlight the critical role of survey-based features, particularly those related to mental well-being, in stress prediction. While mobile sensor data contributed to the model, baseline and EMA surveys emerged as the most predictive modalities. The Random Forest classifier showed promising results, achieving an AUROC of 0.810 and AUPRC of 0.690 during external validation, underscoring the model's generalizability across different academic years.

## 6 Future Work

Future research should explore several promising avenues to advance stress prediction methodologies. First, expanding the dataset to include more diverse participant demographics and longer longitudinal tracking could enhance model robustness and generalizability. Additionally, investigating more advanced feature engineering techniques and exploring deep learning architectures that can better capture temporal dependencies could potentially improve predictive performance. Integrating physiological markers such as heart rate variability, sleep quality metrics, and more granular mobile sensor data could provide deeper insights into stress mechanisms. Moreover, translating these predictive models into actionable interventions—such as personalized stress management strategies or early warning systems—represents a critical next step in applying machine learning to support student mental health.

## References

[1] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of*

*the 2014 ACM international joint conference on pervasive and ubiquitous computing*, pages 3–14, 2014.

[2] Ahmed Tawakol, Amorina Ishai, Richard AP Takx, Amparo L Figueroa, Abdelrahman Ali, Yannick Kaiser, Quynh A Truong, Chloe JE Solomon, Claudia Calcagno, Venkatesh Mani, et al. Relation between resting amygdalar activity and cardiovascular events: a longitudinal and cohort study. *The Lancet*, 389(10071):834–845, 2017.

[3] Clementine Maddock and Carmine M Pariante. How does stress affect you? an overview of stress, immunity, depression and disease. *Epidemiology and psychiatric sciences*, 10(3):153–162, 2001.

[4] Sheldon Cohen, Denise Janicki-Deverts, and Gregory E Miller. Psychological stress and disease. *Jama*, 298(14):1685–1687, 2007.

[5] Sohrab Saeb, Mi Zhang, Christopher J Karr, Stephen M Schueller, Marya E Corden, Konrad P Kording, David C Mohr, et al. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *Journal of medical Internet research*, 17(7):e4273, 2015.

[6] Xuhai Xu, Han Zhang, Yasaman Sefidgar, Yiyi Ren, Xin Liu, Woosuk Seo, Jennifer Brown, Kevin Kuehn, Mike Merrill, Paula Nurius, et al. Globem dataset: multi-year datasets for longitudinal human behavior modeling generalization. *Advances in Neural Information Processing Systems*, 35:24655–24692, 2022.

[7] Jingwen Zhang, Dingwen Li, Ruixuan Dai, Heidy Cos, Gregory A Williams, Lacey Raper, Chet W Hammill, and Chenyang Lu. Predicting post-operative complications with wearables: a case study with patients undergoing pancreatic surgery. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(2):1–27, 2022.

[8] Shadab Askar and Shadab Askar. Stress monitoring using machine learning, iot and wearable sensors. *Sensors*, 23(21):8875, 2023.

[9] Russell Li and Zhandong Liu. Stress detection using deep neural networks. *BMC Medical Informatics and Decision Making*, 20(1):285, 2020.

[10] S Sriramprakash, S Vinoth, and S Karthik. Automatic stress detection using wearable sensors and machine learning: A review. In *2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*, pages 1–7. IEEE, 2017.

[11] Wei Wang, Wei Zhang, Li Li, and Liu Liu. Spline regression based feature extraction for semiconductor process fault detection using support vector machine. *Expert Systems with Applications*, 38(5):4820–4826, 2011.

[12] Scott Lundberg. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.