ISSN : 2350-0743

International Journal of Recent Advances in Multidisciplinary Research

Vol. 12, Issue 09, pp.11681-11684, September, 2025

https://doi.org/10.64485/ijramr.6134.09.2025

RESEARCH ARTICLE **CROSSREF** **OPEN ACCESS**

FROM STRESS TO SCORES: A MACHINE LEARNING ANALYSIS OF WEARABLE SENSOR DATA DURING EXAMINATIONS FOR PREDICTING ACADEMIC PERFORMANCE FROM PHYSIOLOGICAL STRESS SIGNALS

***Ria Sharma**

United States

ARTICLE INFO

Article History

Received 24th June, 2025
Received in revised form
28th July, 2025
Accepted 24th August, 2025
Published online 30th September, 2025

Keywords:

Pulse Amplitude,
BVP Readings,
Heart Rate (HR), Blood Volume Pulse (BVP), Skin Surface Temperature (ST), Inter-Beat Interval (IBI).

Corresponding author: **Ria Sharma**

ABSTRACT

Many students suffer from heavy stress when it comes to their exams, often disabling them from performing their best due to a fear of failure. This is significant because it affects one's health, grades, and future endeavors. But how does exam stress affect students' exam performance? Our objective is to decipher which physiological indicators affect the exam score the most. This study investigates whether physiological signals captured by wearables—electrodermal activity (EDA), heart rate (HR), blood volume pulse (BVP), skin surface temperature (ST), inter-beat interval (IBI), and accelerometer (ACC) data—can predict exam scores for three testing events (Midterm 1, Midterm 2, Final). To assess this, we used signal processing and feature importance in measuring and predicting the exam performance. I found that the feature that affected exam scores the most was the BVP readings, also known as the Blood volume pulse amplitude, had an importance of 0.068433.

Citation: Ria Sharma. 2025. "From Stress to Scores: A Machine Learning Analysis of Wearable Sensor Data During Examinations for Predicting Academic Performance from Physiological Stress Signals", International Journal of Recent Advances in Multidisciplinary Research, 12, (09), 11681-11684.

INTRODUCTION

High levels of exam stress negatively impact both student health and academic performance. Stress-related symptoms such as increased heart rate, sweating, and irregular breathing can disrupt focus, ultimately leading to lower exam scores. The central research question guiding this study is: How does exam stress, measured through physiological indicators, influence exam performance, and which features are most predictive of exam scores?. This project approaches the problem as a supervised learning task using regression models, where the input data consists of numerical physiological readings and the output is the exam score. Physiological features were captured through wearable devices, providing continuous measures of stress-related signals during exam sessions. By analyzing these data, the study aims to identify the most influential physiological markers and develop models capable of predicting exam outcomes. The final output is a numerical prediction of exam scores, which can contribute to understanding and potentially mitigating the effects of stress on academic achievement.

Background: The issue of exam stress on academic performance has been widely recognized, with studies examining how physiological states change during real-world exams. One study assessed the effectiveness of using exam periods as a natural stressor, finding that such periods significantly increased both cortisol levels and self-reported stress [5]. Ahmed et al. (2023) investigated the effects of exam-induced stress on memory and blood pressure, showing that heightened stress impaired recall ability and increased physiological strain [1]. While these studies link stress with academic outcomes, they do not incorporate predictive modeling approaches. Psychology literature highlights test anxiety as a prevalent issue, often associated with poor study habits and prior negative experiences, which can amplify stress responses during exams [3,4]. These findings underscore the relevance of examining physiological features, though many rely on self-reported measures rather than objective data. Recent work has applied machine learning to physiological signals to predict academic performance. Xie and Vellido (2023) demonstrated that models such as k-nearest neighbors could achieve high predictive accuracy (ROC-AUC = 0.81), illustrating the potential of physiological markers to serve as

reliable predictors of exam outcomes [6]. Similarly, Yadav and Sano (2025) analyzed a wearable exam stress dataset using hypothesis testing, bootstrapping, and regression tree modeling, uncovering important temporal variations in stress across different exam stages, even when overall stress differences were not significant [7]. This study builds on prior research by utilizing objective physiological data from the Wearable Exam Stress Dataset [2,8] and applying machine learning techniques to predict actual exam scores, bridging the gap between stress biomarkers and academic performance outcomes.

Dataset: The dataset used for this research is the Wearable Exam Stress Dataset for Predicting Cognitive Performance in Real-World Settings, published on PhysioNet and collected by researchers led by Md. Rafiul Amin, Dilranjan S. Wickramasuriya, and Rose T. Faghih. This dataset contains the numerical physiological recordings of 10 different college students(8 male and 2 female) across 3 different exams they took. The data was taken by wearables that all students wore during each exam monitored by time intervals. It monitors factors such as: heart rate, temperature, blood volume pulse, electrodermal activity, and inter-beat interval. The data was split into 80% for training and 20% for testing. The rest were trained on midterm 1/midterm 2 and tested on the final exam.

To interpret the output, I calculated the mean, standard deviation, and extrema of the exam scores. Using that, I created 4 different graphs: Performance Trends Plot, Box Plot Distributions, Correlation Heatmap, Change Pattern Scatter. From this I was able to better visualize the data of the 10 students. I found the graph of the pattern scatter to be the most interesting as it included how students' scores changed between exams. I was able to see that most students performed poorly in the second midterm based mostly on the line graph and that the change in score was the greatest at midterm 2.

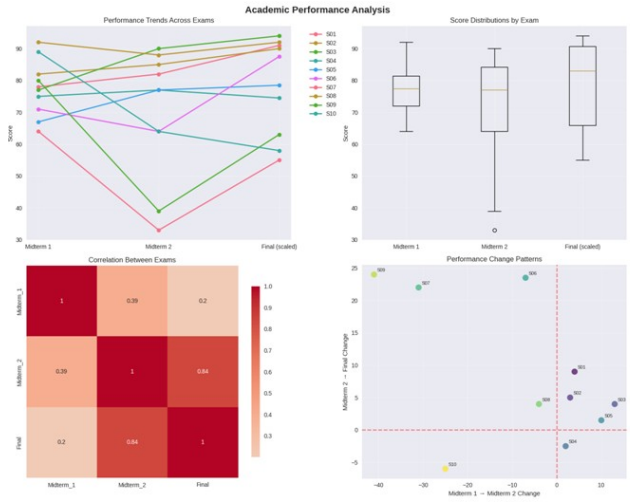

Figure 1.

The next step was to upload the physiological data from the students' wearables. I set up a system to handle the 180 physiological data files (10 students x 3 exams x 6 sensors). A data availability check was run and it was successful in loading all the files. Visualizations such as heatmaps and bar charts in Figure 1 illustrated data availability by student, exam, and sensor. Specifically, the correlation heatmap shown in Figure 1 shows the relationship between each exam. The highest correlations occurred during midterm 2.

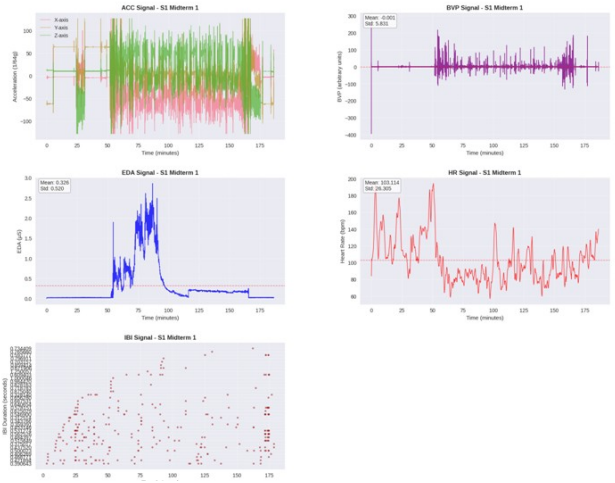

Figure 2

The next section explored the characteristics of the 6 physiological sensors recorded during the exams. ACC, BVP, EDA, HR, IBI, and TEMP all varied in terms of data size and sample rate. EDA had a sample rate of 4 Hz and is a strong indicator of stress. HR data showed average heart rates around 103 bpm, higher than resting levels, possibly indicating exam stress. TEMP readings hovered around 26.4°C, with lower values potentially linked to stress-induced vasoconstriction. ACC data revealed minimal movement, as expected during an exam, while BVP and IBI provided insights into cardiovascular function and heart rate variability. Each samples' visualizations showed their respective trend-with consistent peaks across all sensors at fiscal times of its recording.

METHODOLOGY

To evaluate the predictive power of physiological features, several machine learning models were tested:

To address the challenge of predicting exam scores from physiological stress signals, I tested a variety of machine learning models that ranged from simple baselines to more complex ensemble methods. This approach ensured that I could evaluate whether linear relationships, nonlinear dynamics, or aggregated model predictions best explained the connection between stress indicators and exam performance. Linear Regression was the simplest model applied, serving as a baseline for comparison. This model assumes that exam scores can be predicted as a weighted sum of features such as heart rate, blood volume pulse (BVP), electrodermal activity, and skin temperature. While it does not capture nonlinear relationships, it was useful as a first step because it provided interpretable coefficients that highlighted the direct effect of each physiological feature. By establishing a baseline, I could determine whether more advanced models were truly necessary or whether a simple approach could capture the key patterns. Ridge Regression extended this baseline by adding a regularization term to reduce the impact of highly correlated features. Physiological signals often overlap in meaning (for instance, heart rate and inter-beat interval are closely related), and Ridge Regression helps prevent instability in coefficient estimates when such correlations exist. This model was included to improve the reliability of predictions on a

relatively small dataset, while still retaining the interpretability of a linear framework.

Lasso Regression was also tested, as it not only regularizes coefficients like Ridge but can also shrink some of them to zero entirely. This makes Lasso particularly useful for feature selection, as it highlights which physiological signals contribute most to exam score prediction. In this study, Lasso helped verify that BVP and skin temperature were consistently the most predictive features, effectively filtering out less informative signals. Its ability to automatically focus on the strongest predictors made it a valuable tool in understanding which aspects of stress physiology matter most. Support Vector Regression (SVR) was introduced to explore potential nonlinear relationships between stress features and exam outcomes. Unlike linear models, SVR can use kernel functions to model complex boundaries, which is helpful when the relationship between stress responses and performance is not straightforward. While it did not outperform ensemble methods, SVR provided an important contrast by testing whether nonlinear modeling significantly improved accuracy, and it revealed that the dataset did not strongly benefit from heavy nonlinear transformations.

Random Forest was chosen as a tree-based ensemble method capable of capturing feature interactions and nonlinear effects. By averaging the predictions of multiple decision trees, Random Forest reduces overfitting and offers feature importance scores that help interpret which signals drive predictions. In this project, Random Forest confirmed the high importance of BVP but did not achieve the best predictive accuracy. Its inclusion was still valuable, as it demonstrated how tree-based models compared with both linear and boosting approaches. XGBoost represented a more advanced tree-based algorithm using gradient boosting, where trees are built sequentially to correct the errors of previous ones. Known for its high performance in structured datasets, XGBoost allowed me to test whether boosting could outperform bagging-based methods like Random Forest. Although it did not surpass the simpler regression models or ensembles in this case, XGBoost still contributed by refining the importance of specific features and showing that boosting was not necessarily optimal for this relatively small dataset.

```
for name, (model, param_grid) in models_to_tune.items():
    print(f"Tuning {name}...")

    grid_search = GridSearchCV(
        estimator=model,
        param_grid=param_grid,
        cv=5,  # 5-fold CV instead of LOO for speed
        scoring='neg_mean_absolute_error',
        n_jobs=-1
    )

    grid_search.fit(X_with_prev, y_score)

    best_model = grid_search.best_estimator_
    best_params = grid_search.best_params_
    best_score = -grid_search.best_score_

    print(f"  Best MAE: {best_score:.2f}")
    print(f"  Best params: {best_params}")

    tuned_models[name] = best_model
```

Finally, using hyperparameter turning as shown in the code above, I was able to find the best configuration. The Voting Ensemble combined multiple models to leverage their complementary strengths. In this approach, predictions from models such as Ridge, Lasso, and Linear Regression were aggregated to produce a final score. The ensemble approach reduced individual model weaknesses and provided more stable predictions overall. This strategy proved most successful, achieving the lowest Mean Absolute Error (12.47) and confirming that aggregating multiple methods outperformed any single algorithm. The Voting Ensemble's success reflects the value of combining models that balance interpretability, regularization, and generalization.

RESULTS AND DISCUSSION

The performance of the tested machine learning models varied, but overall results demonstrated that physiological stress signals can serve as predictors of exam scores.

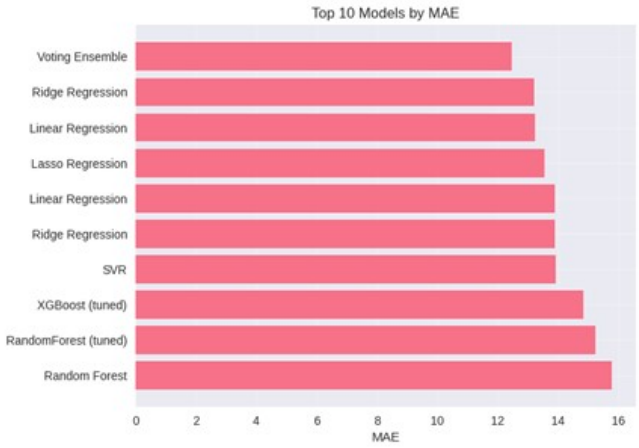

Figure 3

Figure 3 shows the ranking of the top-performing models in terms of Mean Absolute Error (MAE). The Voting Ensemble model achieved the lowest MAE (~12.47), outperforming Ridge, Linear, and Lasso Regression. More complex models such as Random Forest and XGBoost performed worse, suggesting that simpler regression-based models combined in an ensemble better captured the data's structure.

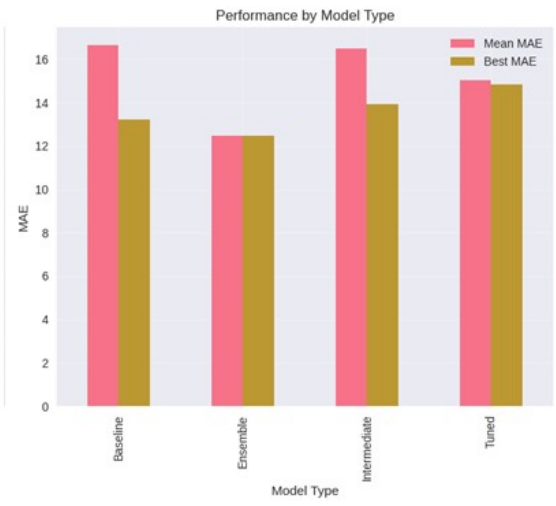

Figure 4

Figure 4 compares the performance of model families (Baseline, Ensemble, Intermediate, and Tuned). The Ensemble approach outperformed all other categories, achieving both the lowest average MAE and the lowest best-case MAE. This confirms that aggregating predictions across models reduces variability and improves robustness in exam score prediction.

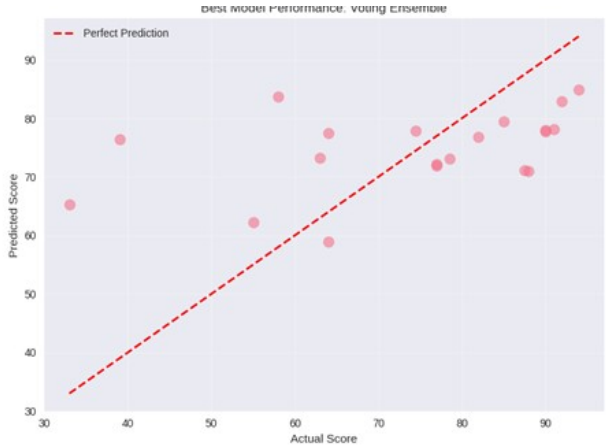

Figure 5

Figure 5 plots actual exam scores against predicted scores for the best-performing model, the Voting Ensemble. The red dashed line represents perfect prediction, where predicted scores exactly equal actual scores. The scatter points, while somewhat dispersed, show that predictions generally follow the trend of actual scores, confirming that the Voting Ensemble captures useful information from the physiological data. However, there is noticeable variance, with some students' scores being overestimated or underestimated by up to 15 points. This aligns with the moderate R^2 value (0.197), which indicates limited but non-trivial predictive power. Across all models, Blood Volume Pulse (BVP) had the highest feature importance (0.068433), followed by skin temperature. These findings support the hypothesis that cardiovascular responses and thermal regulation are key indicators of exam-related stress.

Although the models achieved reasonable predictive accuracy, the results also highlight important limitations. Predictions were weaker for extreme stress cases, suggesting that the small dataset (30 exam sessions from 10 students) limited the models' ability to generalize. Additionally, academic outcomes are influenced by non-physiological factors such as preparation, sleep, and prior knowledge, which were not captured in this study.

The results of this study suggest that wearable devices can be used as practical tools for monitoring stress and predicting exam outcomes, offering potential early-warning systems to identify students at risk of underperforming due to high stress. By incorporating physiological data such as BVP and skin temperature into personalized learning systems, schools could adapt exam conditions, study strategies, or support resources to better address students' needs and reduce performance gaps. Beyond education, these findings highlight broader applications of non-invasive stress monitoring in workplaces, athletics, and other high-pressure environments where performance is impacted by stress. While the predictive power of the models remains modest, combining physiological signals with psychological and behavioral data in future research could lead to more accurate and holistic stress management solutions.

CONCLUSION

This study demonstrates the potential of using wearable physiological data to predict exam performance under stress. Among various machine learning models tested, the Voting Ensemble model achieved the best accuracy, with BVP and temperature emerging as the most important predictors. Although predictive performance was modest, the findings highlight a promising direction for integrating physiological monitoring into educational settings. Future work could involve expanding the dataset, incorporating psychological self-reports, and testing deep learning models for improved prediction. Ultimately, such research could help design personalized interventions to manage stress and enhance student performance.

ACKNOWLEDGMENTS

I'd like to thank my instructor Abhinav Agarwal and the team at Inspirit AI for guiding me in this research endeavor.

REFERENCES

Ahmed, F., Dubey, D. K., Garg, R., and Srivastava, R., 2023, "Effects of Examination-Induced Stress on Memory and Blood Pressure," Journal of Family Medicine and Primary Care, 12(11), pp. 2757–2762. doi:10.4103/jfmpc.jfmpc_925_23.

PhysioNet, 2022, Wearable Exam Stress Dataset for Predicting Cognitive Performance in Real-World Settings, https://physionet.org/content/wearable-exam-stress/1.0.0/#files-panel

Putwain, D. W., 2007, "Test Anxiety in UK Schoolchildren: Prevalence and Demographic Patterns," British Journal of Educational Psychology, 77(3), pp. 579–593.

Verywell Mind, 2024, "Why Are You So Anxious During Test Taking?," https://www.verywellmind.com/what-is-test-anxiety-2795368

Weekes, N., Lewis, R., Patel, F., Garrison-Jakel, J., Berger, D. E., and Lupien, S. J., 2006, "Examination Stress as an Ecological Inducer of Cortisol and Psychological Responses to Stress in Undergraduate Students," Stress, 9(4), pp. 199–206. doi:10.1080/10253890601029751.

Xie, H., and Vellido, A., 2023, "Predicting Students' Exam Scores Using Physiological Signals," arXiv Preprint, https://arxiv.org/abs/2301.12051

Yadav, S., and Sano, A., 2025, "Exploring the Correlation of Physiological Stress Signals: A Wearable Exam Dataset Study," Applied Psychophysiology and Biofeedback, https://link.springer.com/article/10.1007/s10484-025-09685-2

NYU Scholars, 2022, A Wearable Exam Stress Dataset for Predicting Grades Using Physiological Signals, https://nyuscholars.nyu.edu/en/publications/a-wearable-exam-stress-dataset-for-predicting-grades-using-physio

.

****