Nurses Stress Prediction Wearable Sensors

Satwik Garg Saurav Haldar Namit Jain Saarthak Saxena
Indraprastha Institute of Information Technology, Delhi
satwik22461@iiitd.ac.in, saurav22464@iiitd.ac.in
namit22315@iiitd.ac.in, saarthak22421@iiitd.ac.in

Abstract

*This project focuses on predicting stress levels among healthcare professionals, particularly nurses, using machine learning techniques. The SWELL dataset, containing physiological and task-related data from a controlled workplace environment, was utilized. Various models, including Logistic Regression, Decision Trees, Random Forest, Naive Bayes, Perceptron, KNN and GMM were developed to classify stress levels. Data preprocessing involved handling missing values, noise, and outliers, followed by dimensionality reduction using PCA and feature selection through Random Forest. The KNN model demonstrated the highest accuracy of **92.73%**, outperforming other models. These findings highlight the potential of machine learning for stress prediction and its implications for improving healthcare management and intervention strategies. The source code and implementation details are available at our GitHub repository [1].*

1. Introduction

Stress is a significant factor affecting the well-being and performance of nurses, who work in high-pressure environments characterized by long hours, emotional strain, and the need for constant alertness. Prolonged exposure to such stress can lead to both physical and psychological health issues, including burnout, depression, and anxiety, all of which can adversely impact the quality of care provided to patients. Therefore, the ability to monitor and predict stress levels in nurses is essential for maintaining their health and ensuring optimal patient care.

Traditionally, stress levels have been assessed using subjective measures such as self-reported questionnaires, which can be influenced by personal biases and may not always accurately reflect an individual's physiological stress state. Recent advances in machine learning (ML) and wearable technologies, however, provide an opportunity to assess stress more objectively using physiological data. The autonomic nervous system (ANS), which regulates the body's stress response, can be monitored through various signals, such as heart rate variability (HRV), skin conductance, and electrocardiogram (ECG) readings, all of which provide insights into a person's stress levels.

In this project, we aim to develop different ML methods like Logistic Regression, Decision Trees, Random Forest, Naive Bayes and Perceptron to predict stress in nurses based on the SWELL dataset, which includes physiological and task-related data.

2. Literature Survey

- **Stress Detection in Working People by Sriramprakash.S, Prasanna Vadana. D, O. V. Ramana Murthy [2]** The research paper explores machine learning methods, including Logistic Regression, Decision Trees, and Random Forests, for classifying data after thorough pre-processing and feature selection. The methodology involves cleaning and normalizing the data, followed by training models using cross-validation to ensure reliability. Feature selection through Random Forest Importance reduces data complexity, focusing on the most relevant variables. Among the models tested, Random Forest achieved the highest accuracy at **91.34%**, showcasing its strength in handling complex datasets by aggregating multiple decision tree outputs. The study emphasizes Random Forest's superior performance compared to other models in terms of classification accuracy.
- **Multi-CIass Stress Detection Through Heart Rate Variability by Jon Andreas Mortensen, Martin Efremov Mollov, Ayan Chatterjee, Debasish Ghose (Senior Member, IEEE), and Frank Y. Li [3]** In the "Classification Results and Discussions" section of the paper, a 1D Convolutional Neural Network (CNN) model is developed for multi-class stress classification using Heart Rate Variability (HRV) data. Feature selection through ANOVA reduced the features from 34 to 15, achieving **96.5%** accuracy.

3. Dataset

For this project, we utilized the SWELL dataset, which provides preprocessed data tailored for workplace stress analysis. The Swell-KW dataset consists of data from 25 participants who were knowledge workers, collected to study the effects of stress and workload in a controlled work environment. Participants were exposed to three different conditions: low stress, high stress, and interruption. Physiological measurements, including heart rate variability (e.g., RMSSD, SDSD), respiration, and other metrics like KURT, SKEW, and their relative variations (REL_RR), were recorded during various work tasks. We have explained some features of dataset present in the dataset in the Table 1.

Table 1. Dataset features

Feature	Meaning
MEAN_PR	Mean of RR intervals
MEDIAN_PR	Median of RR intervals
SDRR	SD of RR intervals
RMSSD	Root mean square of successive RR interval differences
SDSD	SD of successive RR interval differences
SDRR_RMSSD	Ratio of SDRR over RMSSD
HR	Heart Rate
pNN25	Percentage of successive RR intervals that differ more than 25ms
pNN50	Percentage of successive RR intervals that differ more than 50ms
SD1	Measure short-term HRV in ms and correlates with baroreflex sensitivity(BRS)

Also we plotted the pie chart 1 to show the distribution of different stress classes in dataset.

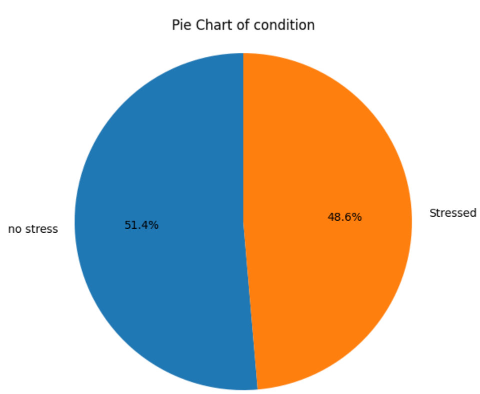

Figure 1. Pie Chart

4. Data Preprocessing

- **Initial Analysis:** Plotted histograms for all features to visualize distributions and detect irregularities.
- **Outlier Removal:** Identified and removed significant outliers to improve data reliability for stress prediction.
- **Handling Missing Values:** Addressed missing data to ensure complete feature representation.
- **Correlation Heatmap:** Generated a heatmap to detect multicollinearity and highlight important features for stress prediction. Figure 2 present the correaltion between 34 features.
- **Conduct t-test:** Determines the critical t-value for a two-tailed test at a 0.05 significance level and conducts t-tests.
- **Feature Combination and PCA:** Reduce 34 feature to 24 features.

These steps enhanced data quality, making it suitable for feature selection and model training.

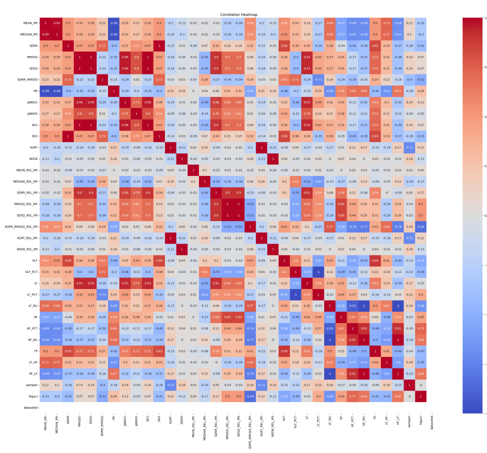

Figure 2. Correlation Heatmap

5. Methodology

The following steps outline the detailed methodology for our machine learning project, covering the feature engineering, and model optimization refer to 3.

5.1. Feature Correlation Analysis

During Exploratory Data Analysis (EDA), we noticed several features were highly correlated, which could negatively impact the models due to multicollinearity. To address this, we performed a t-test for correlation with a significance level of $\alpha = 0.05$. This allowed us to merge or remove highly correlated features, simplifying the feature set.

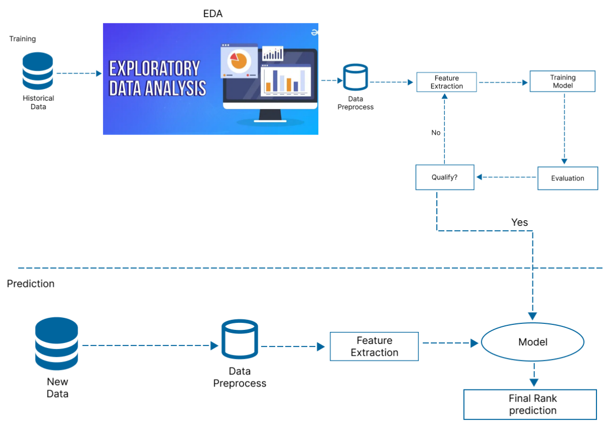

Figure 3. Methodology

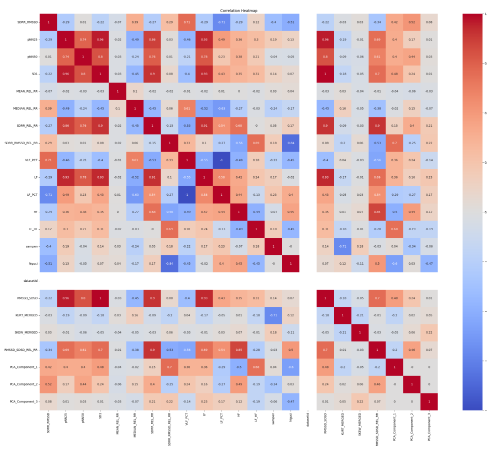

Figure 4. Final correlation heatmap

5.2. Feature Combination

The combined features integrate complementary HRV metrics for improved analysis. RMSSD_SDSD averages RMSSD and SDSD, capturing short-term variability. KURT_MERGED and SKEW_MERGED combine kurtosis and skewness with their relative RR counterparts, enhancing sensitivity to distribution shape. RMSSD_SDSD_REL_RR normalizes RMSSD and SDSD relative to RR intervals, providing a balanced view of HRV dynamics. These combinations offer a more comprehensive interpretation of heart rate variability.

5.3. Dimensionality Reduction with PCA

Principal Component Analysis (PCA) was applied to the 11 original HRV features—MEAN_RR, MEDIAN_RR, HR, SDRR, SD2, VLF, TP, LF_NU, HF_PCT, HF_NU, and HF_LF—to reduce dimensionality while preserving key variance in the data. The result is 3 new composite features, each representing a principal component. These PCA features condense the original 11 variables into 3 orthogonal components, capturing the most significant variations across the dataset. Figure 4 represents the correlation between the 22 features.

5.4. Building Models

Following the feature engineering steps, we retrained all the baseline models (Logistic Regression, Decision Tree, Random Forest, Naive Bayes, and Perceptron) using the reduced feature set. The accuracy and Macro F1 Score for each model after feature reduction are shown in Table 2.

Table 2. Final Accuracy and Macro F1 Scores after Feature Reduction

Model	Accuracy	Macro F1 Score
Logistic Regression	0.6509	0.6482
Decision Tree	0.8650	0.8649
Random Forest	0.9221	**0.9219**
Naive Bayes	0.5708	0.5486
Perceptron	0.5469	0.5463

5.5. Feature Selection with Random Forest

We used the Random Forest model to identify the top 20 most important features. These top 20 features were then used to retrain all the baseline models (Logistic Regression, Decision Tree, Random Forest, Naive Bayes, and Perceptron) to compare their performance on a reduced feature set.

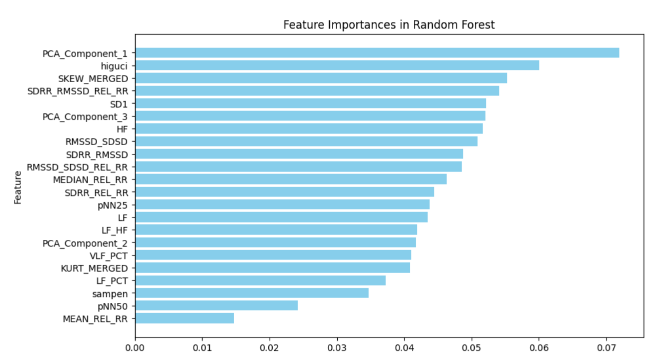

Figure 5. top features from random forest

5.6. Model Training with Varying Number of Top Features

After selecting the top 20 features, we further tested the old models and two new models KNN and GMM by training them with subsets of 20, 15, 10, 7, and 5 top features to determine the optimal number of features that balance performance and simplicity. Table 3 shows the results for each model across different numbers of features.

Table 3. Accuracy of Models with Top Features

Model	Number of Features				
	20	15	10	7	5
Logistic Regression	0.6489	0.6119	0.5939	0.5771	0.5788
Decision Tree	0.8662	0.8647	0.8652	0.8663	0.8647
Random Forest	0.9214	0.9219	0.9215	0.9234	0.9226
Naive Bayes	0.5689	0.5772	0.5813	0.5885	0.5829
Perceptron	0.5180	0.4804	0.5597	0.5058	0.4434
KNN	0.9269	0.9265	**0.9273**	0.9269	0.9267
GMM	0.4729	0.4609	0.4926	0.5408	0.4527

6. Results and Analysis

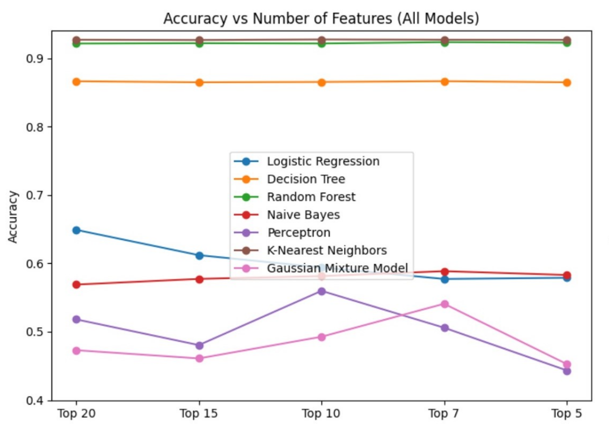

Figure 6. Accuracy vs no of features for all models

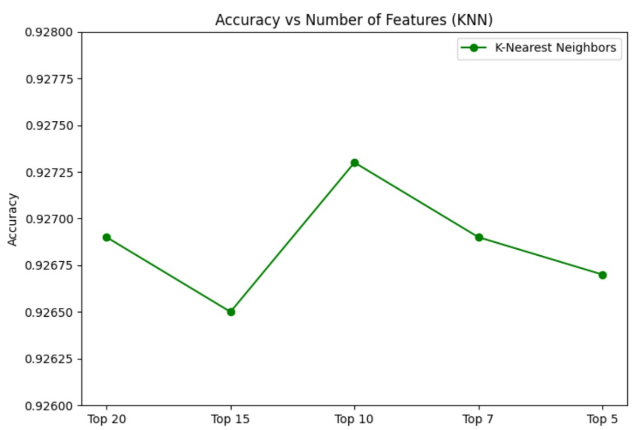

Figure 7. Accuracy vs no of features for KNN

As shown in Table 3, The K-Nearest Neighbors (KNN) model outperformed all other models, achieving the highest accuracy of **0.9273** using a subset of 10 features. The model demonstrated robust performance across various feature subsets, consistently maintaining high accuracy even as the number of features was reduced. n terms of computational efficiency, Naive Bayes was the fastest model, with training times below 0.15 seconds across all feature subsets. However, KNN required significantly more time for larger feature sets, with a peak training time of 20.34 seconds on the Top 20 features subset.

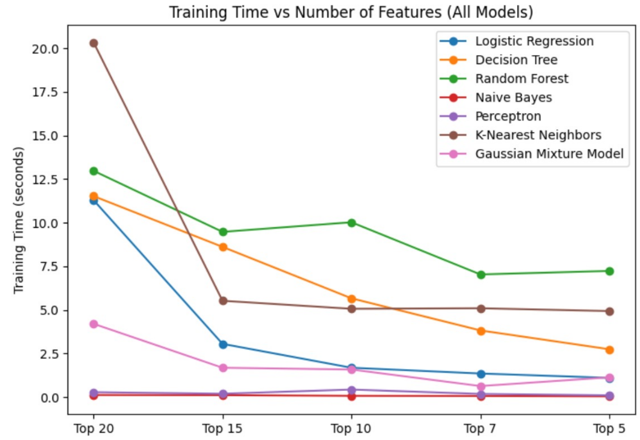

Figure 8. Training time vs no of features for all models

Overall, KNN emerged as the most effective model, offering the highest predictive accuracy. Random Forest followed closely, with an accuracy of 0.9234 on the Top 7 features, while maintaining consistent training times. These findings emphasize the importance of feature selection in improving model performance and efficiency, particularly for computationally intensive models like KNN.

Finally, the K-Fold validation results for the K-Nearest Neighbors (KNN) model demonstrated excellent performance, achieving a Mean Accuracy of 0.9276 ± 0.0011 and a Mean F1-Score of 0.9275 ± 0.0011.

7. Conclusion

The study demonstrates that machine learning models, particularly K-Nearest Neighbors (KNN), can effectively predict stress levels in healthcare workers using physiological data. KNN consistently outperformed other models, achieving the highest accuracy of 0.9273 on the Top 10 features subset. Despite its computational cost for larger feature sets, KNN exhibited robust predictive capabilities across various subsets. Random Forest, while slightly less accurate, balanced computational efficiency and accuracy, making it a strong alternative for real-time applications. Feature selection and dimensionality reduction played a critical role in enhancing model performance, highlighting the need to identify the optimal feature subsets for achieving both accuracy and efficiency. These insights provide a foundation for further research in stress detection, with future efforts focused on optimizing KNN for reduced computational cost and exploring ensemble methods to combine the strengths of multiple models. Developing scalable, real-time stress monitoring systems for healthcare environments remains a promising direction for future work.

8. Contribution

- **Namit Jain (2022315):-** PPT creation, Programming
- **Saurav Haldar (2022464):-** Report Creation, Programming
- **Saarthak Saxena (2022421):-** Report Creation, Programming
- **Satwik Garg (2022461):-** PPT creation, Programming.

References

[1] N. J. S. S. Satwik Garg, Saurav Haldar, "Github repo," github repository for the project. [Online]. Available: https://github.com/satwikgarg2022461/ML_Project_Stress_prediction/tree/main 1

[2] O. V. R. M. Sriramprakash.S *, Prasanna Vadana. D, "Stress detection in working people," *Science Direct*, vol. 115, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S187705091731904X 1

[3] A. C. D. G. S. M. I. Jon Andreas Mortensen, Martin Efremov Mollov and F. Y. Li, "Multi-ciass stress detection through heart rate variability," *IEEE*, vol. 11, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10121432 1

[4] SWELL, "Swell dataset," dataset used for training models. [Online]. Available: https://drive.google.com/file/d/1oNLmZnaretcCT38weDMi9BGfqeTY2Xvi/view 5