Stress Level Prediction from Bio-Sensor Data using Extra Trees

Shibu P
Vellore Institute of Technology, Vellore
Email: shibu.p2025@vitstudent.ac.in*

Abstract

This work presents a machine learning system for evaluating human stress levels using multiple physiological signals collected from wearable devices. The dataset includes electrodermal activity (EDA), skin temperature (TEMP), heart rate (HR), and accelerometer readings (X, Y, Z), which are continuously monitored over time. A full-fledged preprocessing pipeline was developed that includes timestamp conversion, minute-level resampling, skewness correction, feature scaling, and class balancing. These steps ensure a uniform, clean, and interpretable representation of the raw sensor streams.

Compared to the existing method which achieves an accuracy of 93% accuracy the proposed system uses the extra trees classifier achieved a training accuracy of 100%, with test and validation accuracy of 99.30% and 98.74%, respectively, indicating strong generalisation to unseen data. This work demonstrates that tree-based models can effectively interpret low-dimensional biosensor features and are suitable for real-time continuous stress monitoring. The proposed system can be deployed in workplace well-being platforms, digital health applications, and personalised stress-management tools for efficient early detection of psychological stress.

Keywords: Biosensor data, Machine learning, Extra Trees, Stress prediction, Physiological signals

1 Introduction

Advancements in science and technology have enabled more effective solutions in human–machine interaction and biological signal processing. The integration of machine learning with health informatics has significantly enhanced the ability to automatically identify an individual's psychophysiological states [2]. Among these, stress detection using physiological signals has emerged as an important multidisciplinary research field [3].

Stress is a natural response of the human body to demanding or challenging situations [4]. While short-term stress can contribute to improved attention and performance, chronic stress is associated with long-term negative consequences such as cardiovascular diseases, depression, anxiety, and reduced decision-making ability. These risks are particularly significant in high-pressure domains such as healthcare, where the well-being of workers directly affects patient safety [5]. Therefore, continuous and objective stress monitoring has become essential.

Traditional stress assessment relies on self-reported questionnaires, which are often subjective and impractical for real-time monitoring. With the advent of wearable sensor technology, physiological signals such as electrodermal activity (EDA), heart rate (HR), skin temperature (TEMP), and accelerometer data

*This work was supported by the Korean Institute of Electrical Engineers (KIEE).

(X, Y, Z) can be recorded continuously and unobtrusively. These signals reflect autonomic nervous system activity, making them reliable indicators of stress.

The dataset used in this study was collected using the Empatica E4 wearable device from 15 female nurses across two phases: Phase I (15 April–6 August 2020) and Phase II (8 October–11 December 2020). It comprises more than 11.5 million time-series samples containing EDA, HR, TEMP, and triaxial accelerometer readings. Each record is labeled with three categories: low (0), moderate (1), and high stress (2). A significant class imbalance was observed, with the high-stress class contributing approximately 74% of all records. To address this imbalance, the random sampling was applied.

The primary objective of this work is to classify stress levels using physiological biosensor data and evaluate the performance of several machine learning models,including Logistic Regression, SVM, AdaBoost, and extra Trees. Temporal trends were also analysed to investigate stress variations over time and assess the feasibility of deploying real-time stress monitoring systems in healthcare environments.

2 Literature Review

This study presents a machine learning framework to classify nurse stress levels (low,medium,high) using physiological time-series data from wearable sensors (EDA, HR, TEMP, ACCELEROMETER).The researchers resampled the data into one-minute intervals and balanced it using SMOTE.A comparison of Random Forest,XGBoost,K-NN, and LightGBM showed that the ensemble methods,Random Forest and XGBoost , had better predictive performance [1].

This paper addresses the limited generalizability of stress models trained on small, single-study datasets. The authors combined four public datasets (SWELL, NEURO, WESAD, UBFC-Phys) to create "StressData," which includes 99 subjects. They also developed a larger dataset called "SynthesizedStressData." An ensemble model consisting of Gradient Boosting and ANN trained on this synthesized data reached 85% accuracy on new data [2].

This paper presents a hybrid deep learning model that combines CNN and LSTM with an attention mechanism for recognizing three types of stress: stress, neutral, and amusement. Using multimodal physiological data from the WESAD dataset, the model achieved 92.70% accuracy, outperforming standalone CNN, CNN-LSTM, and decision-fusion methods [3].

This research presents a Deep Neural Network (DNN) that automatically classifies employees as "stressed" or "satisfied." Using the Employee Satisfaction Index (ESI) dataset, the 4-layer DNN achieved 88.40% accuracy, outperforming SVM, Decision Tree, and results from earlier studies [4].

This study introduces a new deep neural network called "Shuffled ECA-Net" for feature-level sensor fusion of multimodal bio-signals such as ECG, RESP, and EGG. Using data from 26 subjects with salivary cortisol as a stress reference, the model achieved an accuracy of 91.6%, demonstrating the feasibility of multimodal fusion for stress recognition [5].

This paper presents a simple yet accurate stress detection model using only wrist-worn PPG (BVP) signals from the WESAD dataset. By applying a three-step denoising method, 360-second Hanning window segmentation, and seven HRV features, the SVM classifier reached 95.55% accuracy [6].

This research creates a 1D-CNN model aimed at attaining exceptionally high accuracy for three categories of stress detection (no stress, interruption, time pressure) utilizing solely HRV features from the SWELL-KW dataset. The model reached an accuracy of 99.9% with all 34 HRV features and achieved 96.5% accuracy using a reduced selection of the top 15 features.[7].

This research compares chest-worn and wrist-worn sensors from the WESAD dataset for four-class emotional state prediction. The Random Forest classifier showed the best results, achieving 97.15% accuracy with chest sensors and 95.54% with wrist sensors. ECG and GSR played key roles in differentiating stress levels [8].

This work investigates stress classification in an academic environment using the MIST protocol and evaluates the effect of meditation audio on stress reduction. Using HRV, BVP, and EDA signals from 30 students, the Gradient Boosting classifier achieved 98.28% accuracy after GA and MI-based feature selection and Bayesian optimization [9].

This framework predicts real-time worker stress in high-pressure assembly-line environments using physiological signals collected from an Empatica E4 watch. Among six tested machine learning models, XGBoost performed best, achieving 99.70% accuracy [10].

This study predicts stress levels during sleep using physiological data from the SayoPillow multimodal dataset. After balancing the dataset using ADASYN, both SVM and Gaussian Naive Bayes achieved perfect performance, with 100% accuracy and F1-scores of 1.0 [11].

This research focuses on stress detection using wrist-based EDA signals from four public datasets—CLAS, UTD, VerBIO, and WESAD. The SVM classifier achieved the best accuracy at 92.9%, and EDA signals outperformed multimodal combinations such as ECG + PPG. Additionally, stress classification was found to be more accurate for female subjects [12].

This work introduces "TinyStressNet," a lightweight neural architecture created via Neural Architecture Search (NAS) for on-device stress detection using EDA data. The model achieved 85.98% accuracy across four public datasets while using just 49.40 kB of storage, making it practical for low-resource devices [13].

This paper presents "SELF-CARE," a robust stress detection model that uses context-aware sensor fusion to handle noisy wearable data. By recognizing noise contexts such as motion and muscle artifacts, the model selectively combines sensor inputs. It achieved 94.12% accuracy on the WESAD dataset for wrist-worn sensors in a two-class setting [14].

This study evaluates supervised and unsupervised models for stress detection using heart rate signals from the WESAD and SWELL-KW datasets. While the supervised MLP model achieved 99.03% accuracy on its own dataset, it struggled with cross-dataset transfer. The unsupervised LOF model generalized better, indicating that anomaly detection may be more suitable for real-world stress monitoring [15].

This research focuses on stress detection in free-living environments using the large-scale SWEET dataset, which includes 240 subjects and signals such as ECG, SC, and ST. Among five ML models tested, Random Forest achieved the highest accuracy: 98.29% for binary classification and 97.98% for three-class classification, without requiring data balancing techniques [16].

3 Proposed System

The proposed system presents an end-to-end machine learning pipeline for stress prediction using multi-sensor physiological data collected from wearable devices. The model is designed to transform raw EDA, HR, TEMP, and accelerometer readings into accurate stress-level predictions through structured data preprocessing, feature extraction, and supervised learning.

3.1 Data Source

The dataset used in this study was collected using the Empatica E4 wearable device, a medically certified physiological monitoring sensor. It records various biosignals that are directly associated with autonomic nervous system activity. The dataset includes:

- Electrodermal Activity (EDA)
- Heart Rate (HR)
- Skin Temperature (TEMP)
- Acceleration Signals (X, Y, Z)

Data was gathered from 15 female nurses across two continuous monitoring phases, resulting in nearly 11.5 million time-series samples. Each sample was labelled into three stress levels: low (0), moderate (1), and high (2). Due to real-world working conditions, the dataset exhibits irregular sampling frequencies, noise, and significant class imbalance, which required extensive preprocessing.

3.2 Preprocessing Pipeline

The preprocessing stage is essential for transforming raw biosensor streams into consistent and machine-readable inputs. The following steps were applied:

- **Timestamp Conversion:** Most wearable sensors record timestamps in different formats such as Unix epoch time, device-specific local time strings, or irregularly spaced logging intervals. To ensure consistency across all physiological streams, the raw timestamps were first converted into a standardized `datetime` format (UTC). This process included parsing the original time strings, correcting time-zone offsets, removing duplicated or corrupted timestamps, and sorting the signals in strict chronological order. Standardizing the timestamps is essential for aligning multimodal sensor data and enables accurate time-series resampling and feature synchronization across EDA, HR,TEMP,X,Y,Z signals.
- **Minute-Level Resampling:** Wearable sensors often record data at irregular intervals due to variable sampling rates, signal loss, or device-specific logging. This irregularity can lead to misaligned features and inconsistent input for machine learning models. To address this, the raw time-series signals were resampled into fixed 1-minute intervals. Within each 1-minute window, the measurements for each physiological feature (EDA, HR, TEMP, ACC-X/Y/Z) were aggregated using the mean value. Missing measurements in a given interval were handled using interpolation or forward-filling methods to maintain continuity. This procedure ensures uniform temporal resolution, aligns all sensor modalities, and provides a consistent structure for downstream preprocessing and modeling.
- **Yeo–Johnson Power Transformation(Skewness Correction:**

  The Yeo–Johnson transformation is applied to reduce skewness and make the feature distribution closer to normal.In the dataset the feautre EDA and X has high skewness of 3.03 As shown in fig 1 and 0.97 As shown in fig 2 to avoid this Yeo–Johnson transofrmation method is used . For an input value $y \in R$ and parameter λ, the transformation is defined as:

$$T_\lambda(y) = \begin{cases} \dfrac{(y+1)^\lambda - 1}{\lambda}, & y \geq 0,\ \lambda \neq 0, \\ \log(y+1), & y \geq 0,\ \lambda = 0, \\ -\dfrac{(-y+1)^{(2-\lambda)} - 1}{2-\lambda}, & y < 0,\ \lambda \neq 2, \\ -\log(-y+1), & y < 0,\ \lambda = 2. \end{cases}$$

  The derivative (Jacobian) used in the log-likelihood estimation is:

$$\frac{dT_\lambda(y)}{dy} = \begin{cases} (y+1)^{\lambda-1}, & y \geq 0, \\ (-y+1)^{1-\lambda}, & y < 0. \end{cases}$$

The parameters λ are estimated using maximum likelihood estimation (MLE). Assuming the transformed data $z_i = T_\lambda(y_i)$ are normally distributed, the profile log-likelihood to be maximized is:

$$\ell_{\text{prof}}(\lambda) = -\frac{n}{2} \log \hat{\sigma}^2(\lambda) + \sum_{i=1}^{n} \log \left| \frac{dT_\lambda(y_i)}{dy_i} \right|,$$

where

$$\hat{\mu}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} z_i, \qquad \hat{\sigma}^2(\lambda) = \frac{1}{n} \sum_{i=1}^{n} (z_i - \hat{\mu})^2.$$

This ensures that the selected λ makes the transformed feature as close as possible to a Gaussian distribution.

As shown in Figs. 1 and 2, the features before skewness correction are highly skewed. After applying the correction (Figs. 3 and 4), the distributions become more symmetric.

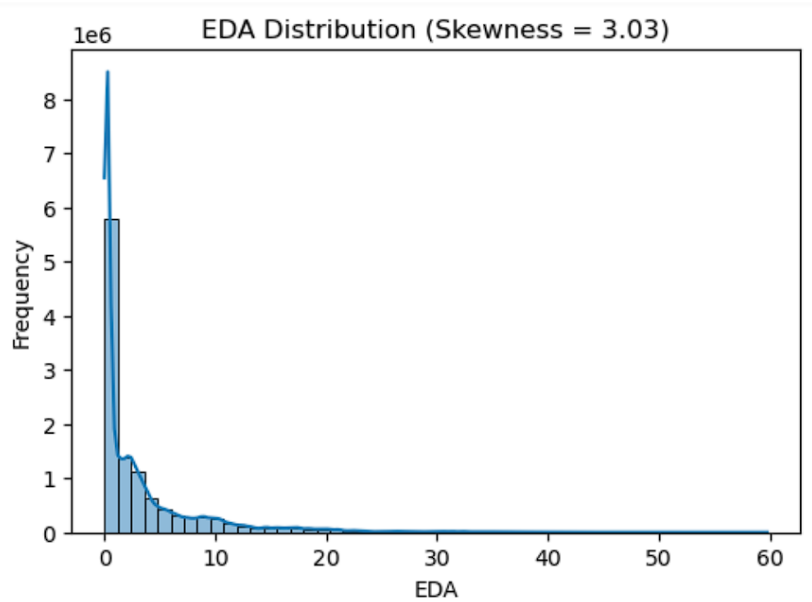

Fig. 1: Before Skewness Correction of EDA Feature

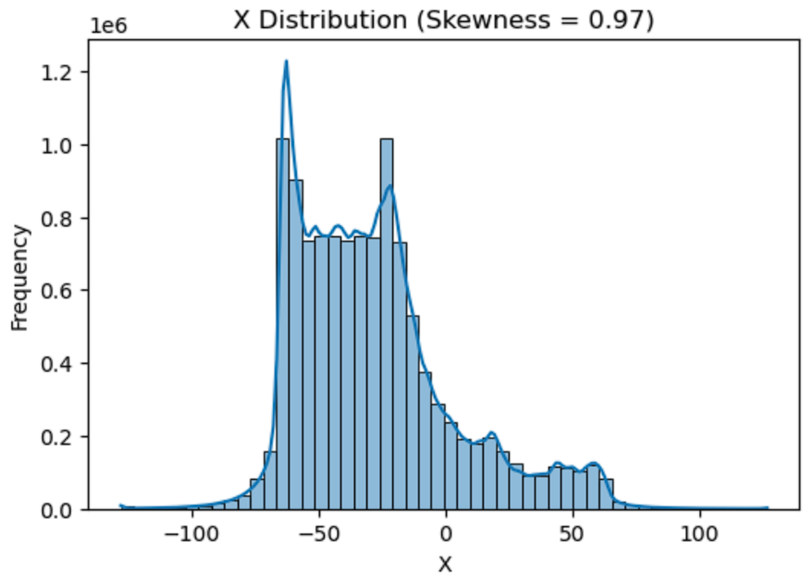

Fig. 2: Before Skewness Correction of X Feature

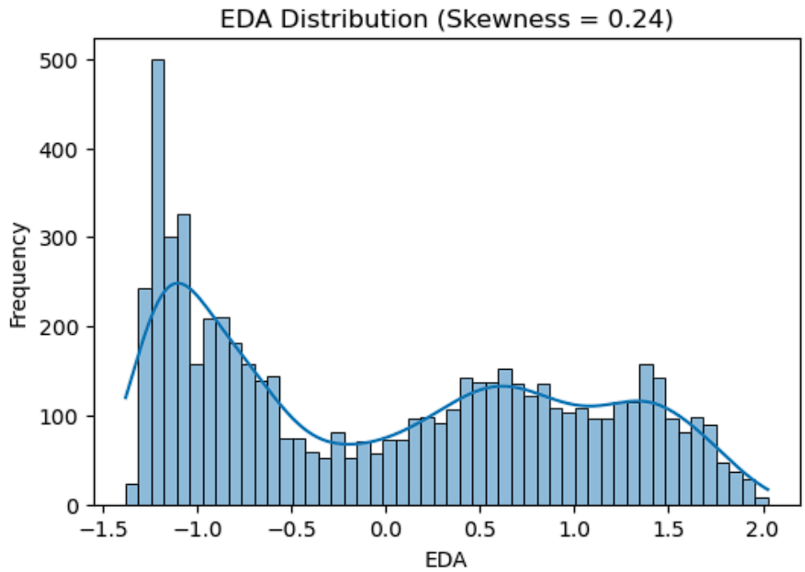

Fig. 3: After Skewness Correction of EDA Feature

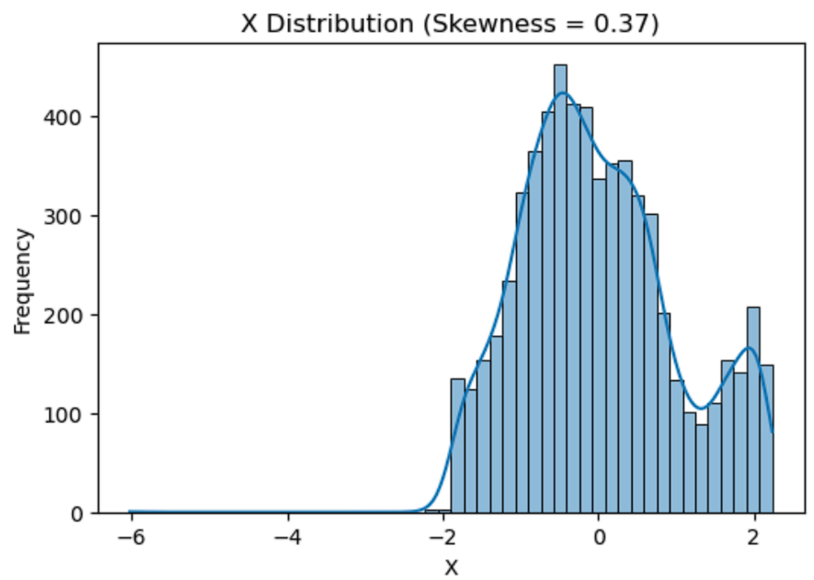

Fig. 4: After Skewness Correction of X Feature

- **Class Balancing:** The dataset exhibited significant class imbalance, with the high-stress class accounting for approximately 74% of the samples. Training machine learning models on imbalanced data can lead to biased predictions toward the majority class and poor performance on minority classes. To mitigate this issue, **RandomOverSampler** was applied.

  RandomOverSampler works by randomly duplicating samples from the minority classes (low and moderate stress) until all classes are roughly equally represented. This ensures that the classifier receives sufficient examples from all stress levels, improving its ability to learn patterns from minority classes and enhancing metrics such as recall and F1-score.

3.3 Extra Tree classifier

The Extra Trees (Extremely Randomized Trees) classifier is an ensemble learning method that constructs a large number of uncorrelated decision trees and aggregates their predictions. Unlike Random Forests, which select the best split among a random subset of features, Extra Trees introduce stronger randomness by selecting split thresholds completely at random. This reduces variance and improves generalization, especially for high-dimensional physiological data.

Mathematical Formulation

Let the training dataset be:

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$$

where $x_i \in R^d$ is the feature vector and $y_i \in \{0, 1, 2\}$ is the stress class label.

1. Random Feature Selection For each tree, a random subset of features $S \subseteq \{1, 2, \ldots, d\}$ is selected:

$$S = \text{RandomSubset}(d, k)$$

where k is the number of randomly chosen features.

2. Random Split Threshold For each selected feature $f \in S$, a random split threshold τ is drawn from the uniform distribution over the feature range:

$$\tau \sim U(\min(x_f), \max(x_f))$$

3. Split Criterion Even though thresholds are random, the best feature-threshold pair is chosen based on impurity reduction. For Gini impurity:

$$G(t) = 1 - \sum_{c=1}^{C} p_c^2$$

where p_c is the proportion of samples of class c in node t.

The optimal split is:

$$(f^*, \tau^*) = \arg\max_{(f,\tau)} \Delta G(f, \tau)$$

where:

$$\Delta G = G(t) - \left(\frac{N_L}{N} G(t_L) + \frac{N_R}{N} G(t_R) \right)$$

4. Ensemble Prediction Given T trees, each tree produces a class prediction:

$$h_t(x), \quad t = 1, \ldots, T$$

The final prediction is obtained by majority voting:

$$\hat{y} = \arg\max_c \sum_{t=1}^{T} I(h_t(x) = c)$$

where $I(\cdot)$ is the indicator function.

Advantages

- Low variance due to high randomization.
- Fast training because no optimal split search is required.
- Handles non-linear relationships in physiological signals.
- Robust to noise and high-dimensional feature spaces.

3.4 Evaluation Metrics

To evaluate the performance of the machine learning models, several standard classification metrics were used. These metrics are computed from the confusion matrix, which consists of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

1. Accuracy

Accuracy measures the proportion of correctly classified samples.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

2. Precision

Precision indicates how many of the predicted positive instances are actually positive.

$$\text{Precision} = \frac{TP}{TP + FP}$$

3. Recall (Sensitivity)

Recall measures the ability of the model to correctly identify positive instances.

$$\text{Recall} = \frac{TP}{TP + FN}$$

4. F1-Score

F1-score is the harmonic mean of precision and recall, providing a balanced metric when classes are imbalanced.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

5. Specificity

Specificity measures the ability of the model to correctly identify negative instances.

$$\text{Specificity} = \frac{TN}{TN + FP}$$

6. Macro Average

Macro averaging computes metrics independently for each class and then takes the average.

$$\text{Macro-Avg} = \frac{1}{K} \sum_{i=1}^{K} M_i$$

where K is the number of classes and M_i is the metric for class i.

7. Weighted Average

Weighted averaging accounts for class imbalance by weighting each class metric by its support.

$$\text{Weighted-Avg} = \frac{\sum_{i=1}^{K} w_i \times M_i}{\sum_{i=1}^{K} w_i}$$

where w_i is the number of true instances of class i.

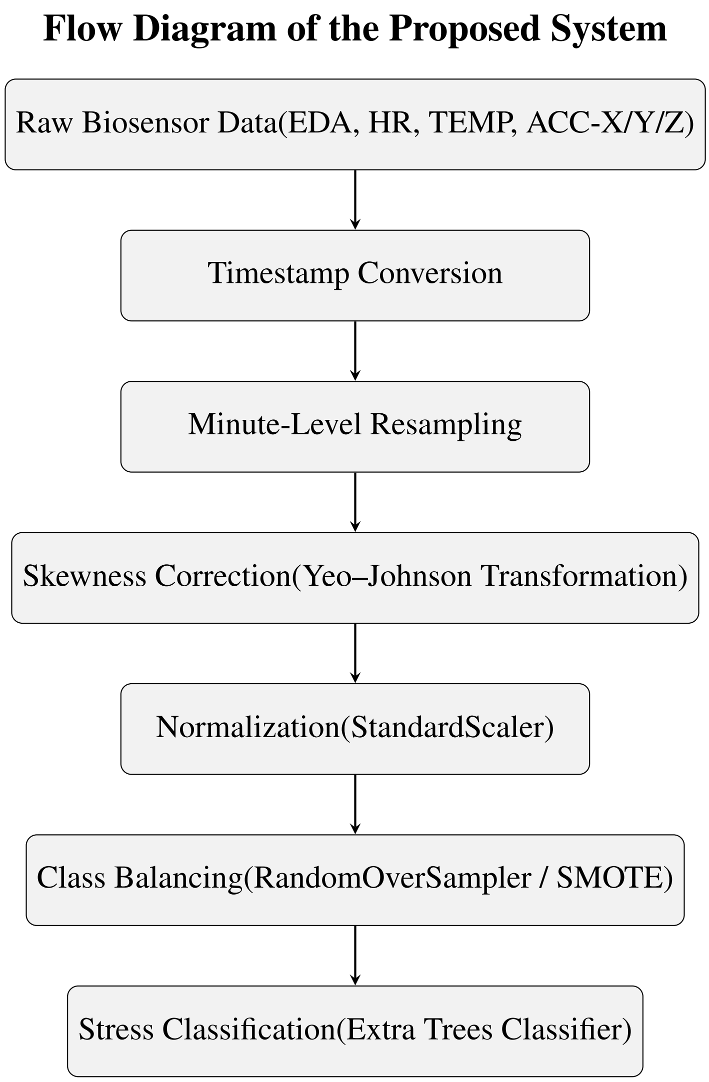

Figure 5: Flow Diagram of the Proposed Stress Prediction System

4 Results

Model	**Train Accuracy**	**Test Accuracy**	**Validation Accuracy (5-Fold CV)**
Logistic Regression	0.3993	0.4035	0.3932
SVC	0.7683	0.7435	0.6874
KNN Classifier	1.0000	0.9106	0.8549
Random Forest	1.0000	0.9858	0.9785
AdaBoost Classifier	0.5906	0.5877	0.6007
Gradient Boosting Classifier	0.9773	0.9453	0.7476
MLP Classifier	0.9984	0.9518	0.7709
Extra Trees	**1.0000**	**0.9931**	**0.9874**

Table 1: Comparison of Train, Test, and Validation Accuracies (5-Fold Cross Validation) for Different Models

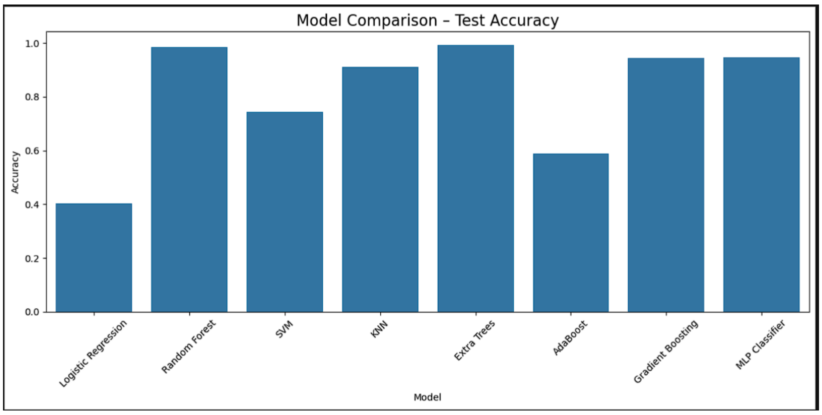

Figure 6: Comparsion of all models

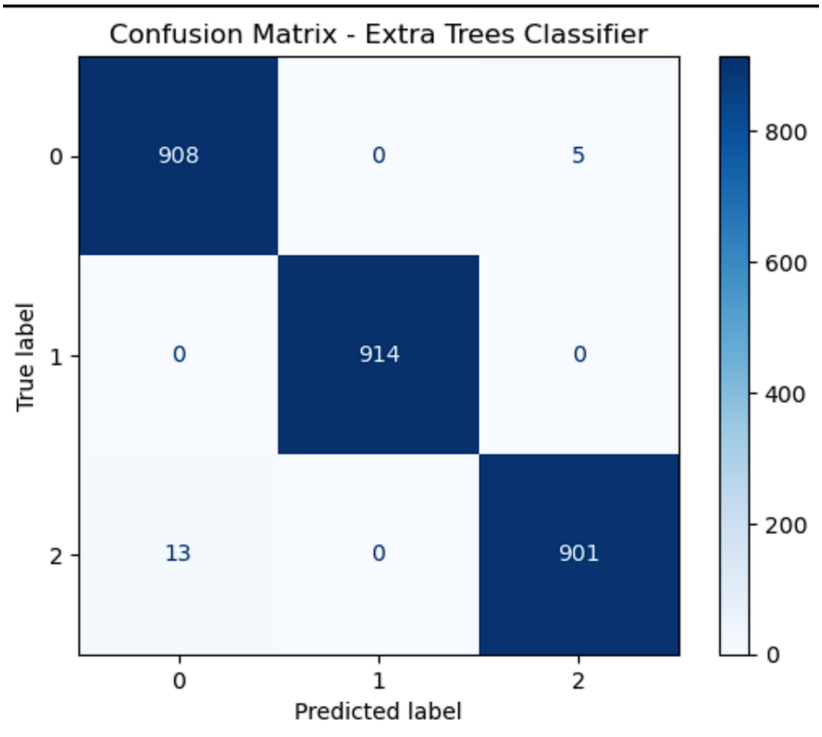

Figure 7: Confusion matrix of Extra Trees Classifier

5 Conclusion

The Proposed Extra Trees Classifier demonstrates significant improvement over the baseline model ,increasing accuracy from 93% [1] to 99.30% using Yeo-Johnson Power Transformation method .This highlight the improvement of the accuracy with reduced data skewness. Implementation of machine learning for stress prediction during sleep has opened a new path for researchers to detect heart attacks and other complications early. We aim to contribute to improving technology in healthcare, machine learning, and IoT. In our study, we presented a method to identify individuals at risk for stress during work. We used a dataset and focused on recognizing the most significant performance improvement approaches. When compared to other works in the same field, our proposed method uses simple and user-friendly models to achieve the best classification performance. With the right configuration, it can even outperform other learning algorithms.

Our goal was not to examine the sensitivity of model hyper-parameters related to these choices. Instead, we aimed to enhance the methodology's predictive power regarding nocturnal stress.

6 References

[1] Korkmaz, Ayşe Çiçek, et al. "Predicting Nurse Stress Levels Using Time-Series Sensor Data and Comparative Evaluation of Classification Algorithms." DOAJ (DOAJ: Directory of Open Access Journals), 22 Aug. 2025, pp. 30–30, https://doi.org/10.3390/engproc2025104030

[2] Vos, Gideon, et al. *Ensemble Machine-Learning Model Trained on a New Synthesized Dataset Generalizes Well for Stress Prediction Using Wearable Devices*. Journal of Biomedical Informatics, vol. 148, 2023. https://doi.org/10.1016/j.jbi.2023.104556

[3] Tanwar, Ritu, et al. *Attention Based Hybrid Deep Learning Model for Wearable Based Stress Recognition*. Engineering Applications of Artificial Intelligence, vol. 127, 2024. https://doi.org/10.1016/j.engappai.2023.107391

[4] Patel, Nikhil, et al. *An Innovative Deep Neural Network for Stress Classification in Workplace*, 2023. https://doi.org/10.1109/icsca57840.2023.10087794

[5] Kim, Namho, et al. *Shuffled ECA-Net for Stress Detection from Multimodal Wearable Sensor Data*. Computers in Biology and Medicine, vol. 183, 2024. https://doi.org/10.1016/j.compbiomed.2024.109217

[6] Jahanjoo, Anice, et al. *High-Accuracy Stress Detection Using Wrist-Worn PPG Sensors*. ISCAS, 2024. https://doi.org/10.1109/iscas58744.2024.10558012

[7] Mortensen, Jon Andreas, et al. *Multi-Class Stress Detection through Heart Rate Variability: A Deep Neural Network Based Study*. IEEE Access, 2023. https://doi.org/10.1109/access.2023.3274478

[8] Gupta, Rohit, et al. *Multimodal Wearable Sensors-Based Stress and Affective States Prediction Model*, 2023. https://doi.org/10.1109/icaccs57279.2023.10112973

[9] Shikha, None, et al. *Optimization of Wearable Biosensor Data for Stress Classification Using Machine Learning and Explainable AI*. IEEE Access, 2024. https://doi.org/10.1109/access.2024.3463742

[10] Hijry, Hassan, et al. *Real Time Worker Stress Prediction in a Smart Factory Assembly Line*. IEEE Access, 2024. https://doi.org/10.1109/access.2024.3446875

[11] Rahman, Md Minhazur, et al. *SleepWell: Stress Level Prediction Through Sleep Data*. AIIoT, 2023. https://doi.org/10.1109/aiiot58121.2023.10174306

[12] Zhu, Lili, et al. *Stress Detection through Wrist-Based Electrodermal Activity Monitoring and Machine Learning*. IEEE JBHI, 2023. https://doi.org/10.1109/JBHI.2023.3239305

[13] Jaiswal, Dibyanshu, et al. *TinyStressNet: On-Device Stress Assessment with Wearable Sensors on Edge Devices*. PerCom Workshops, 2024. https://doi.org/10.1109/percomworkshops59983.2024.10502631

[14] Rashid, Nafiul, et al. *Stress Detection Using Context-Aware Sensor Fusion from Wearable Devices*. IEEE IoT Journal, vol. 10, no. 16, 2023. https://doi.org/10.1109/jiot.2023.3265768

[15] Rashid, Nafiul, et al. *Evaluating Different Configurations of Machine Learning Models and Their Transfer Learning Capabilities for Stress Detection Using Heart Rate*. Journal of Ambient Intelligence and Humanized Computing, vol. 14, 2022. https://doi.org/10.1007/s12652-022-04365-z

[16] *A Machine-Learning Approach for Stress Detection Using Wearable Sensors in Free-Living Environments*. Computers in Biology and Medicine, vol. 179, 2024. https://doi.org/10.1016/j.compbiomed.2024.108918