ANALYZING EEG SIGNALS FOR STRESS DETECTION USING RANDOM FOREST ALGORITHM

Fi Imanur Sifaunnufus Ms[1*], Fitra Abdurrachman Bachtiar[1], Barlian Henryranu Prasetio[1]

[1]Department of Informatics Engineering, Computer Science Faculty, Brawijaya University Malang, East Java, Indonesia

Received: 15th July 2024; Revised: 28th October 2024; Accepted: 29th October 2024

ABSTRACT

Detection of stress using EEG signals has gained much interest because of monitoring and early intervention. As for the contribution of this research, a reliable method for stress identification has been suggested, using a random forest model to categorize stress levels from EEG signals. Data were filtered using a bandpass filter, Independent Component Analysis, and more so using the Z-score to remove outliers and poor signals. Data that has been cleaned from noise and outliers will go through a feature extraction process using Power Spectral Density (PSD). The result of PSD is the power of each frequency of the EEG signal. The number of features used is 20. Random Forest was chosen due to its high accuracy and robustness in handling complex, high-dimensional data, which is common in EEG analysis. Thus, the model obtained an accuracy level of 0.8571, thereby approving the tool's efficiency in distinguishing between different degrees of stress. The computational efficiency of the model, with a classification time of 0.2762 seconds, demonstrates its feasibility for practical applications. Based on these findings, it can be concluded that the Random Forest algorithm can be used to integrate wearable technology and for offering suggestions and timely interventions for better mental health.

Keywords: Stress; EEG; Random Forest; Machine Learning

Introduction

Stress is indeed a problem that presents itself in many facets of one's life, which encompass one's personal life, productivity for occupations, and general health.[1] As evidenced by the Indonesian National Adolescent Mental Health Survey (I-NAMHS), the survey results show that one in three students in Indonesia experiences mental disorders.[2] The identification of stress levels must be precise for the right time for intercessions, which would reduce the impact of stress. EEG has been identified as a potential device for stress measurement without having to penetrate the skin due to its high temporal resolution, which can capture brain electrical activity.

Another research study examined the features of the EEG signal to identify stress by employing various classifiers. Some of the conventional approaches that include Support Vector Machines (SVM), Naive Bayes (NV), Multilayer Perceptron (MLP), and neural networks exist.[3,4] Of these, the most frequent application has been made with SVM since it can produce good results for data operating in high dimensions and is also capable of dealing with non-linearly separable data. Priya[5] examined EEG-based stress classification using SVM polynomials, noting that while SVM polynomials perform well with high-dimensional data, their processing time significantly increases with larger datasets, creating a barrier for real-time analysis. Similarly, research by Li[4] on neural networks has shown that while neural networks can capture complex patterns in EEG data, they often require substantial computational resources and extensive training times, making them less practical for rapid stress detection. Additionally, Arsalan[3] reported that traditional models like Naive Bayes suffer from low accuracy when dealing with more than two stress classes, particularly for

*Corresponding author.
E-Mail: fiimanursifa@student.ub.ac.id

distinguishing subtle differences between moderate and high-stress levels.

In order to overcome these limitations, the current study suggests the employment of a Random Forest model for 3-level stress identification from EEG signals. Random Forest, a technique of ensemble learning, has its own benefits, which include efficiency compared to other models for large datasets, less prone to overfitting, and a lower error rate of function than that of individual decision trees.[6] Through these strengths, we envision increasing the reliability and validity of stress classification, including the discrimination between moderate and high stress.

The research contribution and originality of this paper are provided by the use of the Random Forest model for stress detection regarding EEG signal and comprehensive preprocessing procedures for better data quality and feature selection. Besides, this approach tries to overcome obstacles that arose during prior investigations and prove the possibility of stress monitoring applications.

This research aims to create, test and apply the Random Forest model for stress detection based on EEG signals with regard to accuracy and computational time. It is through this work that we aim to contribute to the improvement of the currently available stress-detection solution that is more valid and feasible for real-world applications.

This paper is organized as follows: The Introduction section provides a foundation by discussing EEG signal processing and the critical role of preprocessing and feature extraction in ensuring high-quality data for accurate analysis. The Methodology section explains each step in the data preparation process, detailing the use of bandpass filtering, Independent Component Analysis (ICA) for artifact removal, and Z-score calculations for outlier elimination, as well as the Power Spectral Density (PSD) feature extraction using Welch's method. In the Results and Discussion section, we present and analyze the outcomes of preprocessing and feature extraction, including visual comparisons of the EEG signal before and after cleaning and a numerical sample of the extracted features. The Conclusion section then summarizes the findings, emphasizing the importance of robust preprocessing in EEG analysis and suggesting directions for future improvements in EEG-based studies.

Methods

The approaches that were used in this study were carefully selected with the aim of improving the reliability of automatic stress detection from the EEG signals. As highlighted in the methodologies section and depicted in Figure 1, the study entailed the collection of EEG data from the participants, and general preprocessing was then applied to improve data quality. Data that is free from noise and outliers will be extracted using Power Spectral Density into several features. The resulting feature is the power value of each frequency band for each electrode. These features will be used by the Random Forest model to classify stress levels. There are three levels of stress used in this research, namely High Stress, Mid Stress, and Not Stress. This decision aligns with previous research, which also categorizes stress into three levels, providing a standardized approach that facilitates comparison and validation across studies. By adopting this widely used classification scheme, the model's outputs are more consistent with established findings in the field, supporting clearer interpretations of stress levels based on EEG power values.

a. EEG Signal

EEG is a beneficial and less invasive technique for measuring the electrical activity of the brain.[7,8] It entails the use of electrodes mounted on the scalp in order to estimate voltages that stem from the ionic currents of neurons. It appears that EEG is used in neuroscience research mainly because it provides data on brain activity and has a high temporal resolution. This research relied on primary data that were obtained from the target population, which in this case was the academic community at Brawijaya University. The sample selected in this study involved a total of 35 students from the Faculty of Computer Science of Brawijaya

University. Descriptions of characteristics of the data sets used in this analysis are shown in Table 1.

Table 1. Detailed Dataset

Dataset	Total of Data
Training Dataset	84
Testing Dataset	21

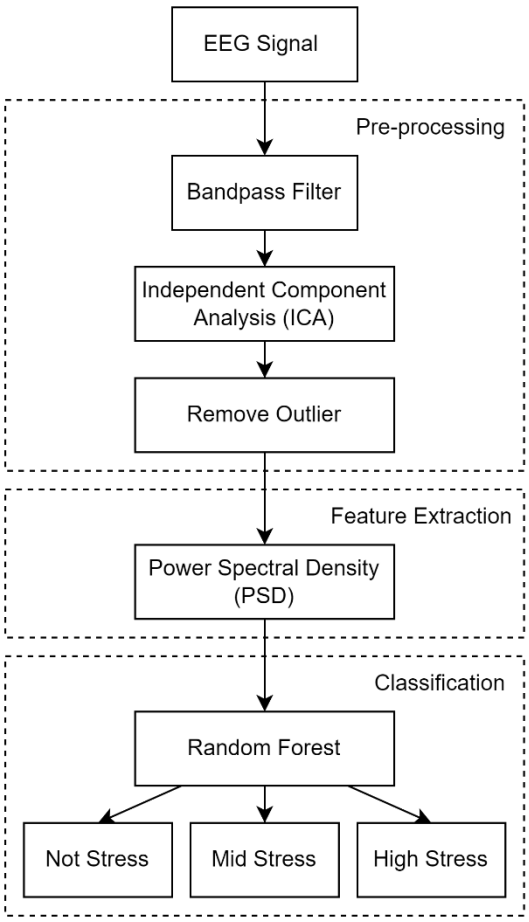

Figure 1. Graphical Abstract

The participants were asked to perform a series of tasks while wearing a Muse 2 device, which records EEG signals from four points: TP9, AF7, AF8 and TP10. The tasks were Pre-Activity, Pauli Test, Translating Test, and Questionnaires. For the pre-activity, the participants had to close their eyes for five minutes and think of anything that made them feel so relaxed. The EEG signals collected during this period were labelled as "Suspect Not Stressed". In the Pauli Test, participants continuously solved simple addition problems for 30 minutes, with the EEG signals collected during this period labelled as "Suspect High Stress. " For the Translating Test, participants spent 8 minutes translating scientific papers from English to Bahasa Indonesia, and the EEG signals collected were also labelled as "Suspect High Stress. " Finally, participants completed the DASS-21, PHQ2, MBI, and SOFI questionnaires within 7 minutes, based on their perceived conditions, with the EEG signals collected during this period labelled as "Suspect Mid Stress. " This study utilized a predefined labelling method to categorize the stress levels based on the tasks performed, resulting in three labels: In this study, you have Suspect Not Stressed, Suspect Mid Stressed and Suspect High Stressed.

b. *Pre*-Processing

Bandpass Filter

A bandpass filter is one of the significant preprocessing techniques commonly used when analyzing the EEG signal and focusing on the desired frequency while eliminating interference. The key application of bandpass filtering is to pass the applied signal only if it is within a certain band of frequencies while rejecting all others outside this range. This step is important for improving the signal-to-noise ratio and ensuring only those signals from the brain activities of interest can be measured.[9]

This study applied the bandpass filter of 1 Hz to 50 Hz to the raw EEG signals. The selection of this bandwidth range in EEG investigations is presented in Table 2.

Independent Component Analysis

Independent Component Analysis (ICA) is a signal processing technique that finds the independent components of a multivariate signal which are added and mixed.[10] This technique is especially useful when dealing with EEG data because sometimes there are disturbances such as eye blink muscle activity, among others, and all these interfere with the data obtained.

In this study, the Fast ICA algorithm was used to decompose the data. Fast ICA is one of the most commonly used methods to implement ICA because it employs a faster approach of fixed point iteration to maximize the independence of the components estimated.[11] Fast ICA calculation is shown in Equation (1) – (4).

$$w^{+} = E\{xg(w^{+}x)\} - E\{g'(w^{T}x)\}w \quad (1)$$

$$w = \frac{w^+}{\|w^+\|} \quad (2)$$

$$w_{p+1} = w_{p+1} - \sum_{j=1}^{p} w_{p+1}^T w_j w_j \quad (3)$$

$$w_{p+1} = \frac{w_{p+1}}{\sqrt{w_{p+1}^T w_{p+1}}} \quad (4)$$

The weight vector (w) calculation is repeated, resulting in an updated weight vector value (w^+). The matrix value w is calculated using an activation function or non-linear function (g) with the expected value or average of the random variable (E). The iteration of the value of w refers to the number of iterations (p) as an update of the vector weight using the transpose function (T). Then, normalizing the vector with the Euclidean norm improves the convergence of the learning algorithm due to the unbalanced scale of features or weights during iteration.

Outlier Removal

The process of outlier identification is an important phase of data preprocessing that excludes certain values that often affect the quality of data and are not considered to be a part of the raw data set.[12] To remove outliers within this study, the steps that were followed included calculating relative values by applying the Z-score. Z score calculation can be seen in Equation (5).

$$Z_{score} = \frac{x - mean}{standart\ deviation} \quad (5)$$

c. Feature Extraction

In order to extract features related to stress detection, the cleaned EEG data was analyzed using Power Spectral Density (PSD). PSD produces features that represent the power of each EEG signal frequency.[13] Welch method was used to calculate the PSD where EEG data is divided into overlapping epochs; each epoch is multiplied with a Hamming window before calculating the periodograms and then averaging the result.[14] The Welch method PSD ($S_x(v)$) the formula is shown in Equation (6)-(9).

$$X_k(v) = \sum_m x[m]w[m]exp\,(-j2\pi vm) \quad (6)$$

$$W = \sum_{m=0}^{M} w^2\,|m| \quad (7)$$

$$P_k(v) = \frac{1}{W}|X_k(v)|^2 \quad (8)$$

$$S_x(v) = \frac{1}{K}\sum_{k=1}^{K} P_k(v) \quad (9)$$

Welch's PSD method divides the data into segments with a certain number of sizes (m) per frequency. Each segment is calculated in its Fourier Transform window ($w[m]$) with a specific frequency (v). Then, the final value of Welch's PSD ($S_x(v)$) method is the average of the modified $P_k(v)$ periodogram values.

d. Classification

A Random Forest classifier was used to predict stress levels from the extracted EEG features. The Random Forest algorithm, consisting of multiple decision trees,[15] was trained on the feature set to classify the EEG data into different stress levels. The Entropy value (E) is a measure of the impurity or disorder in a set of examples. In the context of decision trees, it is used to quantify the homogeneity of a dataset.[16] The Gain value (G_i) is used to decide which feature to split on at each node in the decision tree. It measures the reduction in entropy from a split[16]. Once all the trees are built, they collectively make predictions for new data points. Each tree in the forest casts a vote for the class it predicts. The final classification decision is made based on majority voting, where the class with the most votes from all the trees is selected as the predicted class.[17] The Random Forest formula is shown in Equation (10)-(11).

Table 2. Detailed Dataset

Brain Waves	Frequency (Hz)	Stimulant
Delta (δ)	0.5 – 4	Deep sleep and unconsciousness
Theta (θ)	4 – 8	Fatigue and light sleep
Alpha (α)	8 – 12	Relaxed, relaxed and meditative states
Beta (β)	12 – 30	Focused and centered attention
Gamma (γ)	30-50	Higher-order cognitive processes

$$Entropy(E) = -\sum_{i=0} p_i \times \log_2 p_i \quad (10)$$

$$Gain(G_i) = E - \sum_{i=0} \frac{|S_i|}{|S|} E_i \quad (11)$$

Result and Discussion

a. EEG Signal Pre-processing

In the preprocessing stage, several steps were implemented to enhance the EEG signal quality for further analysis. First, a bandpass filter within the range of 1 Hz to 50 Hz was applied, allowing only the relevant frequency components of neural activity to pass through while removing low- and high-frequency noise. Following this, Independent Component Analysis (ICA) using the Fast ICA algorithm was performed, effectively isolating and removing artifacts caused by eye movements and muscle activity, ensuring that the remaining signals represent true neural activity. Lastly, outliers were removed by calculating Z-scores, identifying and eliminating data points that significantly deviated from typical signal patterns, thus improving the signal's reliability. The results of these preprocessing steps are visually represented in Figure 2. "Before Preprocessing" shows the original EEG signal containing noise and artifacts, and "After Preprocessing" displays the cleaned signal that more accurately reflects neural activity and is ready for further analysis.

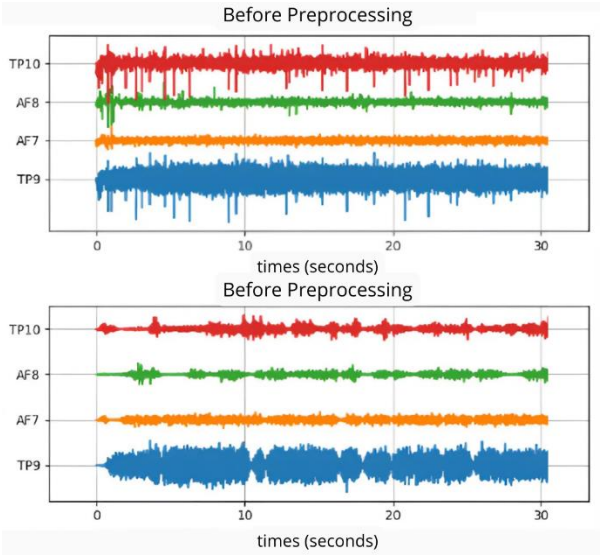

Figure 2. Results of Pre-Processing Steps

b. Feature Extraction

In the feature extraction phase, Power Spectral Density (PSD) analysis was applied using Welch's method. This technique provided a detailed estimation of the power within each frequency band, capturing the essential oscillatory patterns of the brain associated with cognitive states. By focusing on these specific frequency components, PSD analysis yielded numerical features that concisely represent the signal's power distribution, forming the basis for classification or predictive modelling. The sample numerical features extracted using PSD are shown in Figure 3, providing a meaningful summary of the EEG data that enhances further interpretation and analysis.

c. Model Performance

The Random Forest model demonstrated a high degree of accuracy and robustness in predicting stress levels from EEG data. In Figure 4, the model achieved an overall accuracy of 0.8571. The computational efficiency was also notable, with the classification process completed in approximately 0.2762 seconds. Our primary objective was to develop a robust model for stress detection using EEG signals. The high accuracy and rapid computation time achieved by our Random Forest model support this goal, demonstrating its potential for practical applications in wearable devices and mobile applications for continuous stress monitoring.

Delta_Power	Theta_Power	Alpha_Power	...	Label
23.34062	22.54035	22.45844	...	High Stress
23.33720	22.56089	22.48058	...	Mid Stress
23.39151	22.63067	22.55325	...	Not Stress

Figure 3. Sample Features

	precision	recall	f1-score	support
High Stress	1.00	1.00	1.00	8
Mid Stress	0.75	0.60	0.67	5
Not Stress	0.78	0.88	0.82	8
accuracy			0.86	21
macro avg	0.84	0.83	0.83	21
weighted avg	0.86	0.86	0.85	21

Figure 4. Model Performance

The confusion matrix in Figure 5 further illustrates the model's performance, showing high precision and recall for the High-Stress category. This suggests that our Random Forest model is particularly effective in identifying extreme stress levels. Compared to previous studies on stress detection using EEG data, our results indicate a higher precision and recall for the High Stress category. Previous research has reported challenges in accurately classifying high-stress instances due to overlapping features with mid-stress levels. Our study's approach to preprocessing and feature selection appears to mitigate these issues, resulting in improved classification performance.

Our results align with existing literature that emphasizes the effectiveness of EEG-based stress detection. However, the higher performance metrics for High-Stress classification in our study highlight the advantages of our specific preprocessing steps and model tuning. This contrasts with previous studies where overlapping stress levels posed significant classification challenges.

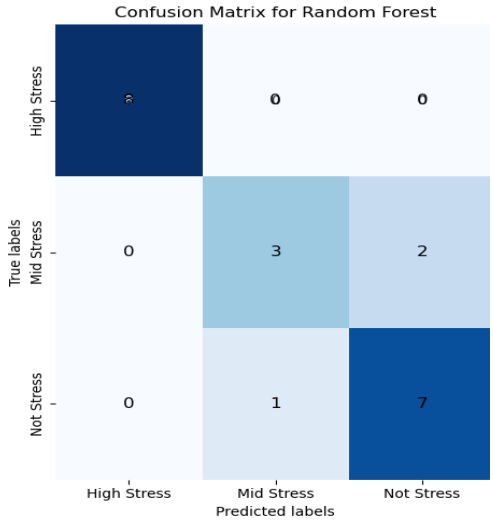

Figure 5. Confusion Matrix

Conclusion

This research was able to achieve the first aim of providing a reliable model for detecting stress from EEG signals by utilizing the Random Forest classifier. The results showed that the model had a general accuracy of 0.8571. The findings obtained in this study effectively illustrate the capability of the model to recognize and distinguish between different degrees of stress, which represents another substantial improvement over the existing model within the area of study.

The conclusion of our study, therefore, contributes to the advancement of existing literature by demonstrating how Random Forest models can be used to accurately and efficiently classify stress levels from EEG data. The model's high computational efficiency, with a classification time of approximately 0.2762 seconds to complete the task, suggests its feasibility for integration into wearable devices and mobile applications, offering stress monitoring capabilities that can aid in timely interventions and potentially improve mental health outcomes.

There is a limitation in using the predefined labels based on the task activities that highlight an area for future research. Further improvements can be made in creating labelling methods that adapt to people's characteristics and variations in real environments to improve the model's precision and usability. In future research, integrating other physiological signals, including the heart rate variability and representing the skin conductance level as GSR, may enhance the stress detection system's reliability.

Future experiments should aim at improving these dynamic labelling methods and test the generalization of the model on much larger and more diverse samples to show its effectiveness across various samples and contexts. Further investigations are completed to consider these directions; it is expected that it will contribute to the increase of the model's stability and applicability in practice in the future. By addressing these areas, we can further improve the

effectiveness and applicability of EEG-based stress detection systems, ultimately contributing to better mental health monitoring and intervention strategies.

Acknowledgement

This journal article is based on research conducted at the Faculty of Computer Science, Brawijaya University. We would like to extend our gratitude to the faculty members and students who participated in the study and provided valuable insights and support. Special thanks to our colleagues for their constructive feedback and to our institution for providing the necessary resources and environment to conduct this research..

References

1. O'Connor DB, Thayer JF, Vedhara K. Stress and Health: A Review of Psychobiological Processes. Annual Review of Psychology. 2021 Sep 4;72(1):663–88.
2. Queensland Centre for Mental Health Research (QCMHR). Indonesia – National Adolescent Mental Health Survey (I-NAMHS) Report (Bahasa Indonesia) [Internet]. qcmhr.org. 2023. Available from: https://qcmhr.org/wp-content/uploads/2023/02/I-NAMHS-Report-Bahasa-Indonesia.pdf
3. Arsalan A, Majid M, Butt AR, Anwar SM. Classification of Perceived Mental Stress Using A Commercially Available EEG Headband. IEEE Journal of Biomedical and Health Informatics. 2019 Nov;23(6):2257–64.
4. Li R, Liu Z. Stress detection using deep neural networks. BMC Medical Informatics and Decision Making. 2020 Dec;20(S11).
5. Priya TH, Mahalakshmi P, Naidu V, Srinivas M. Stress detection from EEG using power ratio. 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE). 2020 Feb;
6. Pandey P, Miyapuram KP. Nonlinear EEG analysis of mindfulness training using interpretable machine learning. 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2021 Dec 9;
7. Malviya L, Mal S, Lalwani P. EEG Data Analysis for Stress Detection. 2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT). 2021 Jun 18;
8. Tasyakuranti AN, Sumarti H, Kusuma HH, Istikomah I, Prastyo IS. ANALYSIS OF THE EFFECT OF ISTIGHFAR DHIKR TO ADOLESCENT ANXIETY AT BETA WAVE ACTIVITY USING ELECTROENCEPHALOGRAM (EEG) EXAMINATION. Jurnal Neutrino:Jurnal Fisika dan Aplikasinya [Internet]. 2022 Sep 26 [cited 2024 Jul 15];15(1):31–7. Available from: https://ejournal.uin-malang.ac.id/index.php/NEUTRINO/article/view/17270/9748
9. Sen D, Bhupati Bhusan Mishra, Prasant Kumar Pattnaik. A Review of the Filtering Techniques used in EEG Signal Processing. 2023 7th International Conference on Trends in Electronics and Informatics (ICOEI). 2023 Apr 11;
10. Tharwat A. Independent component analysis: An introduction. Applied Computing and Informatics. 2018 Aug;17(2).
11. Zhou R, Han J, Li T, Guo Z. Fast Independent Component Analysis Denoising for Magnetotelluric Data Based on a Correlation Coefficient and Fast Iterative Shrinkage Threshold Algorithm. IEEE transactions on geoscience and remote sensing. 2022 Jan 1;60:1–15.
12. Rousseeuw PJ, Hubert M. Robust statistics for outlier detection. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2011 Jan;1(1):73–9.
13. Yi Wen T, Mohd Aris SA. Electroencephalogram (EEG) stress analysis on alpha/beta ratio and theta/beta ratio. Indonesian Journal of Electrical Engineering and Computer Science. 2020 Jan 1;17(1):175.

14. Loza CA, Principe JC. EEG Models and Analysis. Springer eBooks. 2021 Jan 1;1–36.
15. Su C. Heart Rate Variability Feature Selection using Random Forest for Mental Stress Quantification [Internet]. spectrum.library.concordia.ca. 2020. Available from: https://spectrum.library.concordia.ca/id/eprint/987471/
16. Mohammad Zoynul Abedin, M. Kabir Hassan, Hajek P, Mohammed Mohi Uddin. The Essentials of Machine Learning in Finance and Accounting. Routledge; 2021.
17. Tangirala S. Evaluating the Impact of GINI Index and Information Gain on Classification using Decision Tree Classifier Algorithm. International Journal of Advanced Computer Science and Applications. 2020;11(2).
18. Boateng EY, Otoo J, Abaye DA. Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review. Journal of Data Analysis and Information Processing. 2020;08(04):341–57.