Journal of Computer Science

Research Article

XAI_MLPCNN: A Novel Explainable AI-Based Deep Learning Framework for Stress Identification

[1]Fateh Bahadur Kunwar, [1]Rakesh Kumar Yadav, [2]Hitendra Singh and [3]Nitin Tripathi

[1]Department of Computer Science and Engineering, MSOET, Maharishi University of Information Technology, Lucknow, Uttar Pradesh, India
[2]Department of Electronics and Communications Engineering, MSOET, Maharishi University of Information Technology, Lucknow, Uttar Pradesh, India
[3]Department of Computer Science and Engineering, Tata Consultancy Services, India

Article history
Received: 02-11-2024
Revised: 19-02-2025
Accepted: 10-03-2025

Corresponding Author:
Fateh Bahadur Kunwar
Department of Computer Science and Engineering, MSOET, Maharishi University of Information Technology, Lucknow, Uttar Pradesh, India
Email: fateh.kunwar@gmail.com

Abstract: Over the past ten years, there has been a lot of emphasis focused on the development of Artificial Intelligence (AI) and Machine Learning (ML)-based mental health treatments. To increase practitioners' and patients' trust in AI applications, AI systems need to explain their actions. This is called Explainable AI (XAI). While significant progress has been achieved in stress prediction models, XAI has not advanced as much. To overcome this gap, this work presents an explainable AI-based Multi-Layer Pyramid Convolutional Neural Network (XAI_MLPCNN) architecture for stress detection. Multi-channel EEG recordings can be deconstructed into distinct frequency bands and their non-linearity and non-stationarity removed using the Discrete Wavelet Transform (DWT). When processing features, the Power Spectral Density (PSD) is employed. Conversely, the decomposed signals are employed in the automatic feature extraction process through MLPCNN, and the dual BiLSTM with self-attention layer (DBiL_SA) is utilized to predict stress. MLPCNN-DBiL_SA and PSD features are combined to improve prediction. To provide explanations or assess how explainable the predictions are, explainable artificial intelligence techniques like Shapley additive explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) are employed. The Python platform is used to implement the model. Performance is further assessed using a several performance metrics, such as accuracy, recall, precision, and f1-measure. Furthermore, the proposed approach is compared to other methods that are currently in use, like CNN-DWD and PSD, LSTM-DWD and PSD, BI-LSTM-DWD and PSD, RNN-DWD and PSD, and GRU-DWT AND PSD.

Keywords: Explainable AI, Stress Detection, EEG Recordings, Local Interpretable Model-Agnostic Explanations, Shapley Additive Explanations

Introduction

Stress has become a crucial issue of concern in the healthcare sphere since it is a major contributor to a variety of psychological and physical diseases (Amid *et al.*, 2023). To mitigate the negative consequences of stress and improve people's general well-being, early detection and intervention are crucial (Aristizabal *et al.*, 2021). However, the subjectivity and lack of real-time analysis of standard stress assessment techniques (Goumopoulos and Stergiopoulos, 2022), which frequently rely on self-reported measures and clinical evaluations, limit their applicability (Moser *et al.*, 2024). The formation of AI and ML-based methods for stress diagnosis has attracted a lot of attention in response to these limitations (Islam and Washington, 2023).

With models like CNNs and RNNs, AI and ML can provide objective, data-driven insights into stress levels, potentially revolutionizing mental health care (Naegelin *et al.*, 2023). These are technologies that have explainability problems even though they operate promptly and monitor continuously. According to Sasikala and Sachan (2024), in the healthcare sector, XAI is a means of checking AI predictions, establishing confidence among patients (Han *et al.*, 2022), and ensuring equity by removing biases. XAI thus creates opportunities for speedy and non-discriminatory mental health services by making the decisions of the AI understandable and dependable for both patients and medical professionals (Yang *et al.*, 2021). The current paper presents a new framework for the identification of the status of stresses, entrenched in deep-learning-based

explainability. This framework embeds feature attribution, visualization tools, and model-agnostic methods as a compromise between interpretability and accuracy. The subsequent structure for building trust will enhance transparency and interpretation in AI forecasting for users and practitioners. This will finally lead to better clinical results and help integrate AI into mental health treatment.

The research is vital as it addresses the critical gaps in traditional stress assessment methods, which often lack real-time analysis and can lead to misdiagnosis. By leveraging Explainable AI and deep learning, our study not only enhances the precision of stress detection but also fosters transparency and trust in the results. This approach ensures that healthcare professionals and patients alike can understand the reasoning behind stress assessments, ultimately transforming how mental health conditions are diagnosed and managed. As stress continues to pose significant threats to overall well-being, our framework paves the way for timely interventions and personalized care, making it a crucial contribution to the evolving landscape of mental health treatment.

The paper introduces the XAI_MLPCNN framework, which combines a type of neural network called MLPCNN with Explainable AI methods to analyse stress levels from multi-channel EEG recordings. One of the main strengths of this model is that it helps bridge the gap in understanding how AI makes its decisions in mental health applications, providing clear insights for both patients and healthcare professionals. To enhance transparency, the study employs techniques like SHAP and LIME, which make it easier to understand the model's predictions, thus increasing trust in AI-driven mental health treatments. The framework also processes EEG signals by breaking them down into different frequency bands and using Discrete Wavelet Transform (DWT) to handle complexities in the data. By applying Power Spectral Density (PSD) for feature extraction, the model can predict stress levels more accurately through the integration of DBiL_SA and MLPCNN. In light of the innovative approach, the literature review in Part 2 will explore existing research on stress detection methods, highlighting gaps that this new framework aims to address and providing a foundation for understanding its contributions to the field.

Literature Review

Some of the Recent Research Works Related to the Deep Learning Framework for Stress Identification Were Reviewed in this Section.

To classify stress levels, Campanella *et al.* (2023) used machine learning methods, such as Random Forest, Logistic Regression (LR), and Support Vector Machine (SVM), and to analyze data from Empatica E4 bracelet. The chi-square test and Pearson's correlation coefficient were used to choose features. In terms of stress evaluation measures, Random Forest showed the best stability and consistency.

Shahbazi and Byun (2023) investigated the possible long-term health impacts of Early Life Stress (ELS) during pregnancy. The research investigated the relationship between stress and inflammatory imbalance by examining retrospective accounts of childhood or pregnancy challenges in a broad group of women. CNNs are used to identify stress through short-term physiological signals like heart rate and galvanic skin response.

In Kumar *et al.* (2021) concentrated on employing IoT-based wearable sensors for the detection of mental stress. To analyze bio-signals based on the wrist and chest, it presented a multi-level deep neural network with hierarchical convolutional capabilities. The model outperformed previous techniques and advanced early stress detection by combining high-level information to classify stress into three categories.

A unique EEG-based method for stress detection that concentrates on short-duration signals was presented by Sharma *et al.* (2022). Using supervised machine learning methods, entropy-based features from stationary wavelet-transformed EEG data were categorized. SVM performance was improved by using evolutionary-inspired techniques, such as whale optimization, indicating the technique's potential for accurate and timely stress detection.

A mental stress detection system for drivers of automobiles was proposed by Siam *et al.* (2023). It makes use of biosignals such as breathing rate, GSR, ECG, and EMG. The system uses a variety of machine learning models for signal pre-processing, feature extraction, and classification in conjunction with Driver Assistance Systems. When it came to differentiating between levels of stress and relaxation, the Random Forest classifier performed better.

Stress can have serious negative effects on one's physical and mental health, particularly if it is ignored or improperly managed. Studies from Campanella *et al.* (2023); Siam *et al.* (2023) demonstrate the state of current research, which has investigated the use of ML and DL approaches for physiological signal-based stress detection. However, there are still difficulties in obtaining high precision, real-time detection, and application in a variety of settings. The objective of this study is to integrate physiological data from wearable sensors with machine learning models and advanced signal processing techniques to create a robust, dependable, and efficient stress detection system. Improving the model's performance and generalizability will be the main goal, especially in terms of differentiating between different stress and relaxation levels.

Materials

The study proposes an XAI_PCNN framework for stress detection using multi-channel EEG signals.

EEG signals, which are non-linear and non-stationary, are decomposed using Discrete Wavelet Transform (DWT). Feature extraction is performed using Pulse Coupled Neural Network (PCNN) and Power Spectral Density (PSD). These features are fed into a Deep Bidirectional LSTM with Self-Attention (DBiL_SA) model for prediction. The combined PCNN-DBiL_SA-PSD model enhances stress prediction accuracy. To ensure interpretability, SHAP and LIME explainable AI methods are used.

They provide insights into how each feature contributes to the prediction. The implementation is done using the Python programming language. The OpenNeuro dataset, containing emotional EEG signals, is used for model training and testing. All experiments are carried out on a system with Intel Core i7 processor. The system is equipped with 16 GB RAM and an NVIDIA GTX 1660 GPU. This hardware supports efficient training of deep learning models. Fig. 1 illustrates the architecture of the proposed methodology.

Proposed Methodology

In this study, an XAI_PCNN structure for multi-channel EEG recordings-based stress identification is developed. Signals that exhibit non-linearity or non-stationarity are broken down into distinct frequency bands using the DWT. While PCNN is used for automatic feature extraction, PSD is utilized for feature extraction. By using a DBiL_SA, stress prediction is accomplished. Prediction accuracy is improved with the PCNN-DBiL_SA model and integrated PSD. By using SHAP and LIME to make the model's decisions more understandable, explainability is achieved. Fig. 1 depicts the general architecture of the proposed methodology.

Preprocessing

Preprocessing is an essential stage in the study of EEG signals that enhances the dependability and quality of the data prior to conducting any additional analysis.

High-Pass Filter at 1 Hz to Eliminate Low-Frequency Noise

A high-pass filter set at 1 Hz removes the low-frequency noise and baseline points detected for stress. This filter decreases gradual points and artifacts that mask stress-related data by removing frequencies lower than 1 Hz. The refinement sharpens the data toward the pertinent high-frequency components and hence improves the quality of the signal, in which stress may be more accurately identified and analysed.

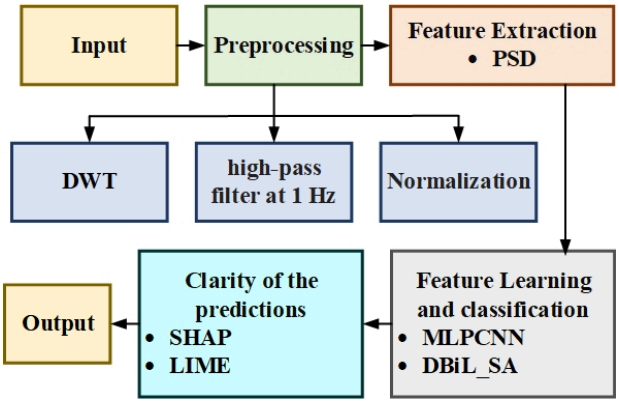

Fig. 1: Overall architecture of the proposed methodology

Normalization

It is a technique of scaling data into a common range or format in order to retain uniformity and comparability. In processing EEG signals, the amplitude in signals is adjusted to lie in a prescribed range, normally between 0 and 1 or -1 and 1. The difference between recordings and sessions can then be reduced by comparing the signals as closely as feasible. Because normalisation standardises the data, it enhances the data's overall interpretability and strengthens any subsequent analytical processes.

Discrete Wavelet Transform

It efficiently manages the non-linearity and non-stationarity characteristics of multi-channel EEG recordings. In most cases, EEG signals show complex temporal fluctuations, which make traditional linear analytical approaches less suitable. Using DWT allows decomposing the EEG data into distinct frequency bands, thus performing a more complete and informative analysis. In a DWT, according to various wavelet functions, the EEG signal is processed to decompose it into a set of coefficients representing the signal at various resolutions or scales. This is produced by passing the signal through two filters, as part of a multilevel decomposition process; the first filter yields the detail coefficients, and the second filter yields the approximation coefficients. The mathematical description of the DWT is expressed in Eq. 1:

$$DWT_{m,n} = \int_{-\infty}^{\infty} x(t)\,\psi_{m,n}(t)\,dt \quad (1)$$

The original continuous-time signal under analysis is denoted by $x(t)$. The wavelet function, which is a translated and scaled version of the mother wavelet, is represented by $\psi_{m,n}(t)\,dt$. The wavelet function's scale and translation are determined by the parameters m and n, respectively. Preprocessing ensures greater quality and more dependable analysis by successfully cleaning EEG data by eliminating errors with DWT, filtering out low-frequency noise, and normalizing signals.

Feature Extraction

Feature extraction is the process of converting unstructured data into a format that highlights significant patterns and characteristics. This would be crucial for effective model training or analysis. Enhancement and convenience of data processing are essential.

Complexity Analysis

3.3.1 Step: High-pass filter at 1 Hz

O(N), where *N* is the number of data points in the input signal. This assumes efficient filtering algorithms like FFT-based filtering.

Step: Normalization

O(N), as it involves computing the mean and variance.

3.3.2 Feature Extraction

Step: PSD (Power Spectral Density)

- Complexity: *O(N log N)*, where N is the number of data points in the input signal.

Step: DWT (Discrete Wavelet Transform)

- Complexity: *O(N),* assuming a single-level DWT. For multi-level decomposition with k levels, complexity becomes *O(N log k)*.

3.3.3 Feature Learning and Classification

MLPCNN (Multilayer Perceptron-CNN)

- Complexity: *O(n_inputs * n_neurons)* per layer.

- CNN: Convolution layers.

- Complexity: *O(N * k^2 * d),* where *N* is the input size, *k* is the kernel size, and *d* is the number of filters.

DBiL-SA (Dual-Bidirectional LSTM with Self-Attention)

- Complexity: *O(T * h^2),* where *T* is the sequence length and *h* is the hidden size.

Computes attention scores.

- Complexity: *O(T^2 * h),* where *T* is the sequence length and h is the hidden size.

3.3.4 Clarity of Predictions

SHAP (SHapley Additive exPlanations):

- Complexity: *O(2^M * N)*, where *M* is the number of features and *N* is the number of samples (approximation methods reduce this).

LIME (Local Interpretable Model-Agnostic Explanations):

- Complexity: *O(K * N),* where *K* is the number of perturbed samples and *N* is the number of features.

Power Spectral Density (PSD)

PSD provides the distribution of signal strength over various frequencies. So, it is a very important feature extraction technique in the case of signal analysis, including EEG analysis. The distribution of power across frequency components helps to identify anomalies and to recognize patterns. First, filtering of the signal is the first preprocessing step, followed by segmenting. Next, the Fourier Transform is applied to transform the signal into the frequency domain, then square the Fourier coefficients so as to obtain the power spectrum. Further, features are extracted based on statistical measures from specific frequency bands on which a feature vector is formed and can further be exploited for more additional analysis. The Expression of the PSD is described in Eq. 2:

$$S_{xx}(f) = \lim_{T\to\infty} E\left(\frac{|X_T(f)|^2}{2T}\right) \quad (2)$$

$S_{xx}(f)$ PSD at frequency f, X_T represents the Fourier transform of the signal over the time interval T, and the expected value is denoted by $E(\cdot)$.

Detailed Feature Learning and Classification

The PCNN architecture, which was created for high-level feature learning, uses PSD properties to increase the efficiency of learning. It makes use of a Dual BiLSTM with a self-attention layer (DBiL_SA), which significantly improves classification performance by enhancing feature representation and model consistency through complex sequence processing.

Fig. 2 shows the MLPCNN architecture used for tasks such as sequence modeling. Its multi-sophistication component architecture leverages the Conv1 to Conv5 layers as feature extractors at progressively higher levels of abstraction, reflecting significant recent improvements in learning long-range dependencies. Each convolutional layer is succeeded by a ReLU activation function, while many of them make use of max pooling for the down-sampling of feature maps. This decreases computational complexity and offers small translations invariance. The network's attention mechanisms are positioned so that the most important portions of an input can be given priority. This ensures important features are given high priority, which significantly increases the model's ability to process tasks with long-range dependencies. This is followed by further reduction in the dimensionality of the feature maps by the Global Average Pool (GAP) layers; these layers get fixed-size vectors that still manage to maintain information critical to the global input. The BiLSTM layers can then model dependencies in both directions and enhance the input with ever-higher context information because they are bidirectional. The skip connections, permit the flow of features from various other layers to combine with the one in regression and, hold the integrity of the pertinent data across the network. These are advantageous because they can learn both high-level and low-level features, learn from data over time, and focus on the most relevant parts. This strategy leads to the final deep learning probability classification; such a fully connected layer computes the SoftMax.

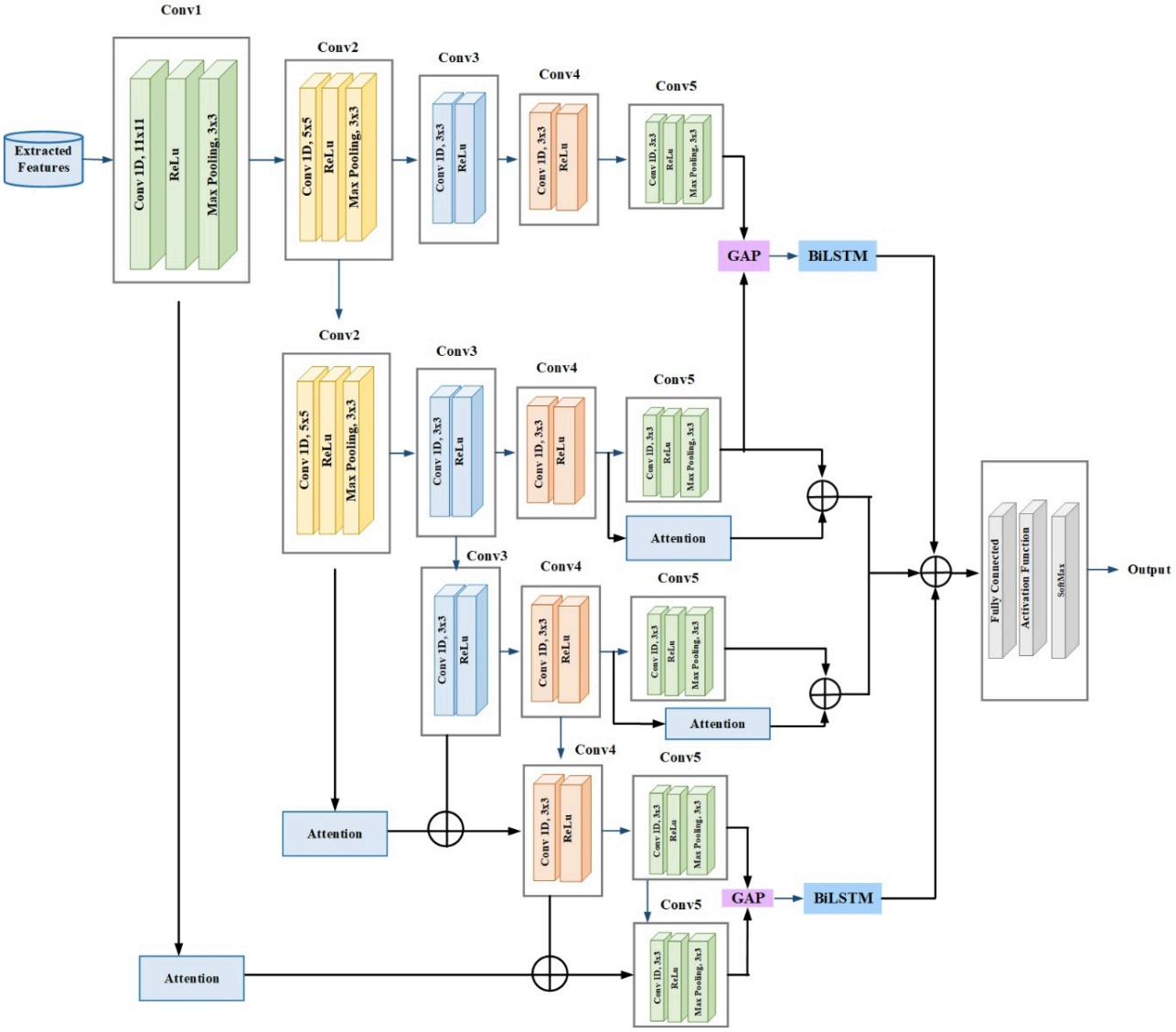

Fig. 2: Architecture of the MLPCNN

Enhancing feature presentation with the addition of a self-attention layer to the Dual BiLSTM architecture adds significantly to classification performance. By capturing intricate patterns and connection features, the context would further strengthen a model by taking advantage of self-attention and bidirectional learning mechanisms.

The Dual BiLSTM architecture is designed to perform effectively with sequential data. The algorithm analyses input sequences using two Long Short-Term Memory (LSTM) layers, represented as $X_{t-1}, X_t, X_{t+1} \ldots \ldots X_T$. The forward LSTM keeps the dependencies from the start to the end of the sequence, and the backward keeps them in reverse order. Merging the outputs of both the Forward and Backward LSTM layers, the Dual BiLSTM layer takes full advantage of the contextual information from forward and backward directions and is therefore able to provide complete comprehension of the sequence. In addition, the input sequence, through the attention layer, assigns the weights to the importance of different elements in the input sequence so that the model could focus on the most relevant parts. After this, BiLSTM is further enhanced using the Normal Distribution function. The general description of the BiLSTM is expressed in Eq. 3:

$$p_t = p_t^f + p_t^b \tag{3}$$

where the network's final probability vector is denoted by p_t. The forward LSTM network's probability vector is denoted by p_t^f and the backward LSTM network's probability vector is represented by p_t^b. An improved weighting mechanism of the normal distribution (W_{N_f}) is integrated into the Bi-LSTM layers and is expressed in Eq. 4. The updated self-attention Dual Bi-LSTM distribution is expressed in Eq. 5 respectively:

$$W_{N_f} = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-\left(N_f-\mu\right)^2}{2\sigma^2}} \quad (4)$$

$$Np_t = W_{N_f} \otimes \left(p_t^f + p_t^b\right) \quad (5)$$

Fig. 3 demonstrates the DBiL_SA process. Prediction accuracy is improved by integrating PSD characteristics with the MLPCNN-DBiL_SA model. Explainable artificial intelligence methods like SHAP and LIME, which provide insights into and assess the reasoning behind the predictions, are used to guarantee visibility and interpretability.

Shapley Additive Explanations (SHAP)

Using Shapley's values from cooperative game theory, SHAP provides a unifying framework for evaluating deep learning models. Assigning an additive linear function to each feature's contribution to a model's output clarifies predictions. SHAP makes sure explanations follow rules like consistency, which states that feature importance does not decrease as a model is more dependent on a feature, and local accuracy, which states that the explanation model matches the original model's prediction. Using linear LIME and Shapley values, kernel SHAP a model-agnostic technique within SHAP provides accurate, efficient local explanations with fewer evaluations.

Fig. 5 shows the stress prediction by SHAP. Input is fed into the model, which then processes and trains on the data. Then, the model generates the output.

Fig. 4 displays a SHAP value plot for model interpretability. It shows how each feature (like AF4, T8, T7) contributes to the model's output for a specific instance. Blue bars indicate a negative impact on the prediction, while red shows a positive impact. The most influential feature here is 'AF4', which lowers the prediction value significantly.

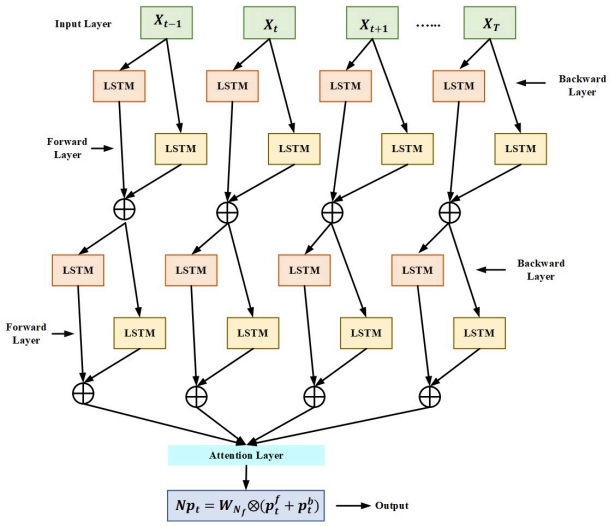

Fig. 3: Dual BiLSTM with Self Attention architecture (DBiL_SA)

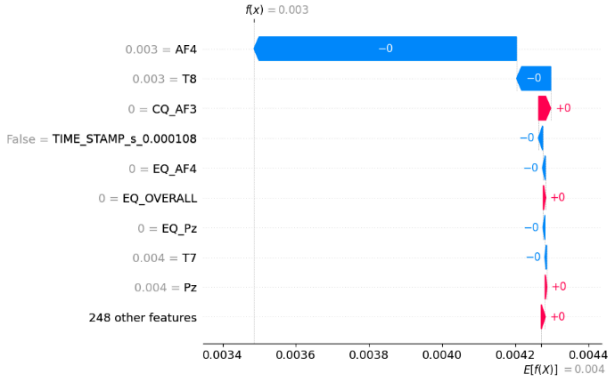

Fig. 4: SHAP Value

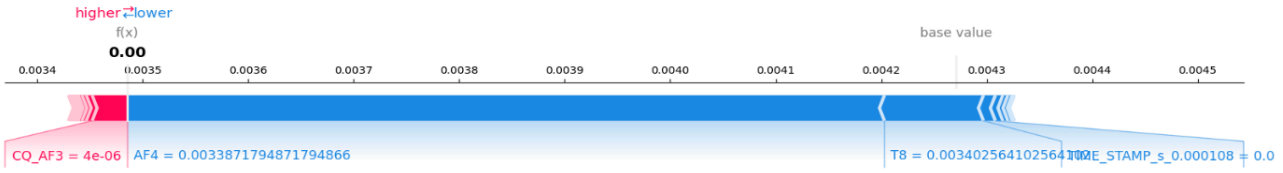

Fig. 5: Stress prediction using SHAP

Local Interpretable Model-Agnostic Explanations (LIME)

Complex models are explained by LIME, which uses simpler, interpretable models to approximate their predictions locally. It creates perturbed samples around a specific case and monitors how modifications impact the predictions of the model. The behavior of the complicated model in the immediate area of the instance is then approximated by LIME using a sparse, interpretable model. The explanation is obtained by minimizing a loss function that achieves a compromise between the interpretable model's complexity and reliability the extent to which the local model accurately represents the complicated model. Fig. 6 demonstrated the stress detection value from https://www.kaggle.com/code/amitvkulkarni/lime-for-explainability-in-python. Prediction accuracy is improved by integrating PSD characteristics with the MLPCNN-DBiL_SA model. By utilizing SHAP and LIME, the results are assured to be interpreted, offering clear insights into the model's decision-making process

and enhancing general comprehension and confidence in the predictions.

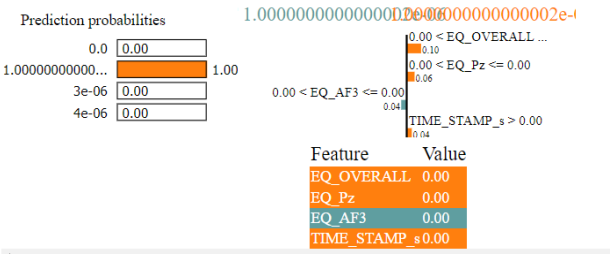

Fig. 6: Stress detection using LIME

Performance Evaluation

Dataset Description

The OpenNeuro dataset focuses on stress research with brain and behavioural data. It involves neuroimaging and other metrics obtained during stress trials in order to assess responses to diverse stressors. This dataset is useful for investigating neural correlates of stress and evaluating models such as XAI-based frameworks for stress detection. The OpenNeuro dataset for stress research has extensive data for exploring stress reactions. It contains neuroimaging data, including fMRI scans, as well as behavioral and physiological variables including heart rate variability, skin conductance, and cortisol levels. These data were acquired during experimental stress trials involving a variety of stressors, such as cognitive demands and social problems, and provide a detailed understanding of stress reactivity. This dataset is an excellent resource for creating and assessing explainable AI (XAI)-based frameworks since it allows for the investigation of brain and behavioral patterns under stress. Researchers can use its multidimensional data to verify models for various stressor kinds and populations, resulting in robust and generalizable stress detection systems.

Results and Discussion

The implementation is done by the python platform. Execution of the recommended method is assessed using a variety of metrics, including Accuracy, Specificity, Sensitivity, F1-Measure, Precision, Matthew's correlation coefficient (MCC), Negative Predicted Value (NPV), False Positive Rate (FPR), and False Negative Rate (FNR). The recently constructed framework is evaluated in terms of how much its performance has increased by comparing it with other models, such as Proposed, SVM, DT, RF, KNN, and LR. Table 1 uses a 70/30 data split to evaluate performance metrics among various. With the best accuracy (0.98542), precision (0.98908), and specificity (0.99025), the proposed model performs better than the others, demonstrating its improved capacity to accurately categorize both positive and negative cases. Additionally, it achieves the greatest F-measure (0.98896), demonstrating a robust equilibrium between sensitivity and precision (0.97248). The robustness of the model's categorization is demonstrated by its 0.98349 MCC. Furthermore, with a 0.98634 NPV which is the highest, it ensures accurate predictions of real negatives. Along with its superior error minimization, the proposed model has the lowest FPR of 0.02518 and FNR (0.01847).

Table 1: Comparative analysis of the performance metrics for data split 70/30

Model	Proposed	SVM	RF	KNN	DT	LR
Accuracy	0.98542	0.93287	0.93861	0.94377	0.95913	0.95729
Precision	0.98908	0.94268	0.93993	0.95583	0.94838	0.96413
Sensitivity	0.97248	0.93907	0.94521	0.95891	0.95384	0.95907
F-measure	0.98896	0.94937	0.95937	0.95683	0.95976	0.96019
Specificity	0.99025	0.94568	0.95082	0.95602	0.96037	0.96114
MCC	0.98349	0.94684	0.95183	0.95891	0.96854	0.96872
NPV	0.98634	0.94166	0.94837	0.94682	0.95543	0.96533
FPR	0.02518	0.06642	0.05896	0.04863	0.04608	0.03961
FNR	0.01847	0.05567	0.04057	0.04186	0.03986	0.03197

Table 2: Comparative analysis of the performance metrics for data split 80/20

Model	Proposed	SVM	RF	KNN	DT	LR
Accuracy	0.99583	0.94496	0.94382	0.95093	0.96831	0.96082
Precision	0.99273	0.95683	0.94682	0.95983	0.95083	0.96869
Sensitivity	0.97986	0.94168	0.95068	0.96195	0.96492	0.96061
F-measure	0.9896	0.95063	0.96381	0.96319	0.96082	0.96952
Specificity	0.99368	0.95193	0.96226	0.95867	0.96659	0.96961
MCC	0.98969	0.95068	0.96682	0.96091	0.97039	0.97093
NPV	0.98986	0.95093	0.95193	0.95096	0.96692	0.96837
FPR	0.01186	0.05293	0.04952	0.03931	0.03894	0.03138
FNR	0.00826	0.04391	0.03951	0.03109	0.02263	0.02093

DOI: 10.3844/jcssp.2025.1156.1167

Fig. 7: Graphical representation of the data splits of 70/30 and 80/20 for the various performance metrics

Using an 80/20 data split, Table 2 offers a comparative examination of several machine learning models based on performance indicators. At 0.99583, the suggested model outperforms SVM (0.94496), RF (0.94382), KNN (0.95093), DT (0.96831), and LR (0.96082) in terms of accuracy. The proposed model's precision, at 0.99273, is also the highest, demonstrating its strong ability to recognize positive situations. While DT (0.96492) and LR (0.96061) also exhibit strong performance in identifying true positives, the proposed model's sensitivity of 0.97986 demonstrates exceptional performance in this area. The proposed approach successfully achieves a balance between sensitivity and precision, maintaining the maximum F-measure (0.9896). The proposed model's specificity, which stands at 0.99368, shows how well it can identify negative cases. The robustness of the model is highlighted by its MCC of 0.98969. To further demonstrate its overall efficiency in this analysis, the proposed model has the lowest FPR (0.01186) and FNR (0.00826), demonstrating its outstanding efficiency in minimizing mistakes.

Fig. 7 demonstrates the comparative analysis of the Graphical Representation of performance matrices of the existing works with the proposed method by data splits 70/30 and 80/20 respectively. Performance metrics like Accuracy, Specificity, Sensitivity, F1-Measure, Precision, MCC, NPV, FPR, and FNR are compared with different methods like Proposed, SVM, DT, RF, KNN, and LR.

With a 70/30 data split, Table 3 shows that the proposed model performs much better than the other models across a range of performance parameters. Its overall excellent performance is demonstrated by the best accuracy (0.98542), precision (0.98908), sensitivity (0.97248), specificity (0.98896), F-measure (0.99025), NPV (0.98349), and MCC (0.98634). Its efficiency in reducing false positives and false negatives is further evidenced by the lowest FPR (0.02518) and FNR (0.01847). Table 4 demonstrates that the Proposed model performs remarkably well across most criteria when the data is split 80/20. It exhibits the lowest false positive rate (0.01186) and false negative rate (0.00826) in addition to achieving high accuracy (0.99583), precision (0.99273), and specificity (0.9896). The F-measure (0.99368) and sensitivity (0.97986) of the proposed model show strong performance, albeit somewhat below the top values of CNN-DWT.

Fig. 8 demonstrates the comparative analysis of the Graphical Representation of performance matrices for the different methods with the proposed method by data splits 70/30 and 80/20 respectively. Performance metrics like Accuracy, Specificity, Sensitivity, F1-Measure, Precision, MCC, NPV, FPR, and FNR are compared with different methods like Proposed, CNN-DWT and PSD, LSTM-DWT and PSD, BI-LSTM-DWT and PSD, RNN-DWT and PSD, and GRU-DWT AND PSD.

Table 3: Comparative analysis of the performance metrics for the different methods by the data split 70/30

Model	CNN-DWT	LSTM-DWT	BI-LSTM-DWT	RNN-DWT	GRU-DWT	Proposed
Accuracy	0.95229	0.95593	0.96921	0.95861	0.96998	0.98542
Precision	0.95972	0.96612	0.95392	0.95591	0.96927	0.98908
Sensitivity	0.95248	0.96954	0.95943	0.95997	0.95591	0.97248
Specificity	0.95153	0.95594	0.95582	0.95947	0.95691	0.98896
F-measure	0.95228	0.94843	0.95922	0.95483	0.96773	0.99025
NPV	0.95896	0.95598	0.96671	0.95165	0.96872	0.98349
MCC	0.95752	0.95834	0.95943	0.95843	0.96742	0.98634
FPR	0.04384	0.05943	0.05412	0.04591	0.04025	0.02518
FNR	0.03796	0.05567	0.04873	0.04889	0.03397	0.01847

Table 4: Comparative analysis of the performance metrics for the different methods by the data split 80/20

Model	CNN-DWT	LSTM-DWT	BI-LSTM-DWT	RNN-DWT	GRU-DWT	Proposed
Accuracy	0.99823	0.96827	0.97083	0.96942	0.97233	0.99583
Precision	0.99023	0.97731	0.96684	0.96542	0.97912	0.99273
Sensitivity	0.99559	0.97086	0.96417	0.96825	0.96628	0.97986
Specificity	0.98469	0.96682	0.96258	0.96256	0.97952	0.9896
F-measure	0.99709	0.95963	0.96089	0.96852	0.97852	0.99368
NPV	0.99082	0.96183	0.97183	0.96815	0.97471	0.98969
MCC	0.99083	0.96084	0.96943	0.96058	0.97972	0.98986
FPR	0.01448	0.04814	0.04912	0.04012	0.03884	0.01186
FNR	0.01189	0.04321	0.04109	0.03951	0.03972	0.00826

DOI: 10.3844/jcssp.2025.1156.1167

Fig. 8: Graphical representation of the data splits of 70/30 and 80/20 for the various performance metrics by the different comparison methods

Conclusion

The XAI_MLPCNN system, which combines advanced deep learning models with explainable AI methodologies, marks a substantial advance in stress prediction. The framework efficiently reduces non-linearity and non-stationarity by preprocessing multi-channel EEG recordings using DWT. This allows for more precise feature extraction using PSD and MLPCNN. To further improve prediction performance, DBiL_SA is used. Most importantly, accessibility is provided by the integration of SHAP and LIME approaches, enabling practitioners and patients to comprehend and have confidence in the AI system's predictions. This approach leads to more dependable and understandable stress detection solutions in clinical settings by increasing prediction accuracy and strengthening trust in AI-driven mental health applications.

A novel explainable AI-based deep learning architecture for stress detection has various drawbacks that must be addressed for effective application. One key problem is the scarcity and quality of labeled data, which frequently lacks diversity and may contain biases, resulting in poor generalizability across groups. Furthermore, the physiological signals and textual inputs employed for stress detection are susceptible to noise, which might impair the model's accuracy. There is also a trade-off between model complexity and interpretability, as simplifying models to increase explainability may result in lower predictive performance. Furthermore, using explainable AI algorithms may raise processing overhead, thereby slowing down real-time stress detection. Finally, models trained in controlled contexts may struggle to generalize effectively to dynamic, real-world scenarios, limiting their practical utility.

The suggested framework, XAI_MLPCNN, has various constraints that should be considered. One major problem is relying on the dataset's quality and variety; insufficient or biased data can result in erroneous stress detection, especially for underrepresented populations. Furthermore, the framework's capacity to generalize across varied groups, cultures, and stressors is limited by differences in stress reactions and measuring methodologies. While deep learning decreases the need for manual feature engineering, finding the right balance between automated learning and domain-specific insights remains challenging. The combination of MLP and CNN designs, while strong, increases computational complexity, which may impede real-time applications in resource-constrained situations, such as wearables. The suggested XAI_MLPCNN framework's ability to generalize successfully across varied populations is an important feature. Stress reactions can vary greatly depending on demographic parameters such as age, gender, ethnicity, and cultural background. These variances are frequently impacted by physiological, psychological, and environmental variables distinct to each group. If the training data lacks representation from varied populations, the model may be biased, resulting in incorrect predictions or explanations for certain groups. To overcome this, it is critical to investigate techniques for improving generalizability, such as developing datasets that include a diverse variety of stressors and demographic characteristics. Furthermore, transfer learning techniques may be used to adapt the model to different populations with little retraining.

Future research should concentrate on using datasets such as Open Neuro to create multimodal explainable AI frameworks that include neuroimaging, physiological, and behavioral data for stress detection. Expanding the dataset to include more diverse populations and stressor types can improve model generalizability. Furthermore, combining real-time data collecting and investigating advanced techniques such as transfer learning and domain adaptation might help models adapt. The emphasis on explainability ensures that the models are interpretable, which fosters trust and facilitates their incorporation into healthcare and workplace stress management systems.

Acknowledgment

Thank you to the publisher for their support in the publication of this research article. We are grateful for the resources and platform provided by the publisher, which have enabled us to share our findings with a wider audience. We appreciate the efforts of the editorial team in reviewing and editing our work, and we are thankful for the opportunity to contribute to the field of research through this publication.

Funding Information

The authors have not received any financial support or funding to report.

Author's Contributions

Fateh Bahadur Kunwar: The research scholar, carried out the core implementation, experimentation, and analysis of the proposed methodology.

Rakesh Kumar Yadav: The supervisor, provided overall guidance, technical direction, and critical revisions throughout the research.

Hitendra Singh: The co-supervisor, contributed to the model design and supported the evaluation process.

Nitin Tripathi: Assisted in conducting the experiments and data preprocessing.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Amid, A., Mahpodz, Z. A., & Abdullah, A. (2023). Stress Monitoring Device and its Future Direction. *IIUM Medical Journal Malaysia*, *22*(2). https://doi.org/10.31436/imjm.v22i2.2154

Aristizabal, S., Byun, K., Wood, N., Mullan, A. F., Porter, P. M., Campanella, C., Jamrozik, A., Nenadic, I. Z., & Bauer, B. A. (2021). The Feasibility of Wearable and Self-Report Stress Detection Measures in a Semi-Controlled Lab Environment. *IEEE Access*, *9*, 102053-102068. https://doi.org/10.1109/access.2021.3097038

Campanella, S., Altaleb, A., Belli, A., Pierleoni, P., & Palma, L. (2023). A Method for Stress Detection Using Empatica E4 Bracelet and Machine-Learning Techniques. *Sensors*, *23*(7), 3565. https://doi.org/10.3390/s23073565

Goumopoulos, C., & Stergiopoulos, N. G. (2022). *Mental stress detection using a wearable device and heart rate variability monitoring*. 261-290. https://doi.org/10.1016/b978-0-323-90585-5.00011-4

Han, X., Wang, L., Seo, S. H., He, J., & Jung, T. (2022). Measuring Perceived Psychological Stress in Urban Built Environments Using Google Street View and Deep Learning. *Frontiers in Public Health*, *10*, 891736. https://doi.org/10.3389/fpubh.2022.891736

Islam, T., & Washington, P. (2023). Individualized Stress Mobile Sensing Using Self-Supervised Pre-Training. *Applied Sciences*, *13*(21), 12035. https://doi.org/10.3390/app132112035

Kumar, A., Sharma, K., & Sharma, A. (2021). Hierarchical Deep Neural Network for Mental Stress State Detection Using IoT Based Biomarkers. *Pattern Recognition Letters*, *145*, 81-87. https://doi.org/10.1016/j.patrec.2021.01.030

Moser, M. K., Ehrhart, M., & Resch, B. (2024). An Explainable Deep Learning Approach for Stress Detection in Wearable Sensor Measurements. *Sensors*, *24*(16), 5085. https://doi.org/10.3390/s24165085

Naegelin, M., Weibel, R. P., Kerr, J. I., Schinazi, V. R., La Marca, R., von Wangenheim, F., Hoelscher, C., & Ferrario, A. (2023). An Interpretable Machine Learning Approach to Multimodal Stress Detection in a Simulated Office Environment. *Journal of Biomedical Informatics*, *139*, 104299. https://doi.org/10.1016/j.jbi.2023.104299

Sasikala, Dr. B., & Sachan, S. (2024). *Decoding Decision-Making: Embracing Explainable AI for Trust and Transparency*. https://doi.org/10.59646/efaimltc3/133

Shahbazi, Z., & Byun, Y.-C. (2023). Early Life Stress Detection Using Physiological Signals and Machine Learning Pipelines. *Biology*, *12*(1), 91. https://doi.org/10.3390/biology12010091

Sharma, L. D., Bohat, V. K., Habib, M., Al-Zoubi, A. M., Faris, H., & Aljarah, I. (2022). Evolutionary Inspired Approach for Mental Stress Detection Using EEG Signal. *Expert Systems with Applications*, *197*, 116634. https://doi.org/10.1016/j.eswa.2022.116634

Siam, A. I., Gamel, S. A., & Talaat, F. M. (2023). Automatic Stress Detection in Car Drivers Based on Non-Invasive Physiological Signals using Machine Learning Techniques. *Neural Computing and Applications*, *35*(17), 12891-12904. https://doi.org/10.1007/s00521-023-08428-w

Yang, K., Wang, C., Gu, Y., Sarsenbayeva, Z., Tag, B., Dingler, T., Wadley, G., & Goncalves, J. (2023). Behavioral and Physiological Signals-Based Deep Multimodal Approach for Mobile Emotion Recognition. *IEEE Transactions on Affective Computing*, *14*(2), 1082-1097. https://doi.org/10.1109/taffc.2021.3100868