INFORMATICS JOURNALS

38

SHRI-LEKH (Multi-disciplinary Research Journal) Vol 2(1), June 2026, p. 38-42

Online ISSN: 3139-5694
DOI:10.18311/SHRI-LEKH/2026/v2i1/53887

Theoretical Framework for Early Mental Disorder Prediction Using Deep Learning

Shruti Joshi[1]*, Shweta M. Barhate[1], and M. P. Dhore[2]

[1]Department of Electronics and Computer Science. RTM Nagpur University, Nagpur – 440001, Maharashtra, India

[2]Department of Electronics and Computer Science. SGBAU, Amravati – 400063, Maharashtra, India

Abstract

Early detection of mental disorders such as Attention Deficit Hyperactivity Disorder (ADHD), Depression, Post Traumatic Stress Disorder (PTSD) and Anxiety in Adolescents is a major public health concern. This paper represents a theoretical deep learning model to identify discriminative Electroencephalography (EEG) features associated with mental disorders. EEG signals contain rich neurophysiological information that can support early diagnosis of mental disorders. This framework addresses the need for clinical interpretable machine learning models by providing accurate predictions and transparent explanations.

Keywords: Adolescents, Deep Learning, EEG, Machine Learning, Mental Disorder, Neuropsychological Biomarkers

INTRODUCTION

Mental disorders (Awan *et.al.*, 2023) among adolescents are increasing globally, as delayed diagnosis is a barrier to early intervention. Adolescents, as defined by the World Health Organisation (WHO) 2023, are individuals aged 10-19. As the WHO gave an insight into statistics that one in seven is suffering from some kind of mental distress. Half of the mental distress conditions start by age 18. 15-19 aged are often involved in panic or excessive worry (Raghu *et.al.*, 2022). Deep Learning (DL), being a subset of Machine Learning (ML), uses multi-layered neural networks to learn patterns from raw data. The DL model (Kheder *et.al.*, 2021; Park *et.al.*, 2024) excels by processing different types of data to manifest in different ways.

EEG is a non-invasive technique that captures electrical brain activity and provides measurable biomarkers linked to cognitive and emotional functioning. While considering EEG (Li *et al.*, 2020) following challenges remain:

- EEG data is high-dimensional.
- Strong noise and subject variability
- Clinical adaptation requires explainability, not just high accuracy, to be considered.

The model is intended for conceptual understanding and which can be later implemented. To address the EEG challenges (Mausavi *et.al.*, 2021) represented above, can be handled through the proposed theoretical model by combining

- Autoencoder - for unsupervised EEG feature compression.
- Latent space classifier - for mental disorder prediction.
- SHAP explainability- to identify the most predictive features.

LITERATURE REVIEW

Zhai *et al.* (2025) proposed a comprehensive multi-centre EEG framework specifically designed for

Email: *shrutig285@gmail.com
shwetab73@yahoo.com
mpdhore@gmail.com

depression grading. The paper focused on signal quality, standardised preprocessing and transparent feature construction instead of relying on opaque deep neural network models. This framework incorporated SHAP-based interpretability, enabling clinicians to understand how specific EEG features contributed to predicting depression severity. Munera *et al.*(2023) showed a systematic mapping study that analyses nearly a decade's worth of research on EEG feature engineering for mental disorders such as ADHD, depression, schizophrenia and anxiety. The paper categorises hundreds of EEG features-frequency-based statistical, connetome-based and non-linear measures and assesses how they are used in classification models. This supports your use of an autoencoder that acts as a learned feature extractor capable of capturing non-linear EEG relationships.

Abbas *et al.*(2022) introduced a Recurrent Autoencoder (RAE) to extract temporal EEG patterns indicative of schizophrenia. It learned a latent space that captures long-term dependencies in EEG signals, which is then fed into a classifier. The study demonstrates that autoencoder-based latent features significantly boost classification performance.

Khan *et al.*(2023) survey paper reviews interpretability and robustness challenges in EEG-based Artificial Intelligence (AI) systems. It categorises interpretability methods, including gradient-based methods, perturbation methods and Shapley values and evaluates their sustainability for clinical settings.

Volker *et al.*(2023) paper introduces SHERPA, an explainable AI method using Shapley values to analyse EEG Event-Related Potentials (ERPs). SHERPA identifies subtle EEG effects that traditional statistical tests fail to capture, demonstrating the ability of SHAP-based frameworks to highlight spatiotemporal patterns.

Awan *et al.*(2023), The authors investigate a range of explainable machine learning techniques for predicting depression severity. They highlighted the trade-off between highly accurate models and those that provide interpretability essential for clinical use.

Sturm *et al.*(2020) applied Shapley value-based explainability to EEG classification tasks. Illustrating how individual EEG features influence model output. The authors show that Shapley-based methods produce both global and local interoperability. Study supports the inclusion of SHAP in your model as a method to interpret latent-feature-based predictions and link them back to original biomarkers.

METHODOLOGY

The proposed framework integrates unsupervised feature learning, then supervised classification and explainable AI (Das *et al.*, 2022) to construct an interpretable prediction model for mental disorders using EEG data. This theoretical model consists of three major components. The architecture of this proposed theoretical model is illustrated in Figure 1, and a discussion of the proposed architecture is provided as follows.

Autoencoder for EEG Feature Compression

This is the first stage of the pipeline that employs an autoencoder to learn a compact representation of high-dimensional EEG feature vectors. Encoder (Saba-Sadiya *et.al.*, 2022; Siddiqui *et.al.*, 2022) will learn an efficient mapping to capture non-linear patterns and correlations across EEG features. The bottleneck layer (latent space) serves as a compressed representation of EEG activity. This latent vector will encapsulate the most critical EEG features. Latent space will suppress irrelevant noise and inter-subject variability. It will potentially align with neurophysiological biomarkers. Then the decoder will expand the latent space back to the original EEG feature dimensions. Decoder attempts to reconstruct the EEG features as accurately as possible. By using an autoencoder(Hammam *et.al.*, 2021; Jain *et.al.*, 2023) along with noise-reduced, information-rich rich and compressed input representation for classification, both interpretability and model stability can be improved.

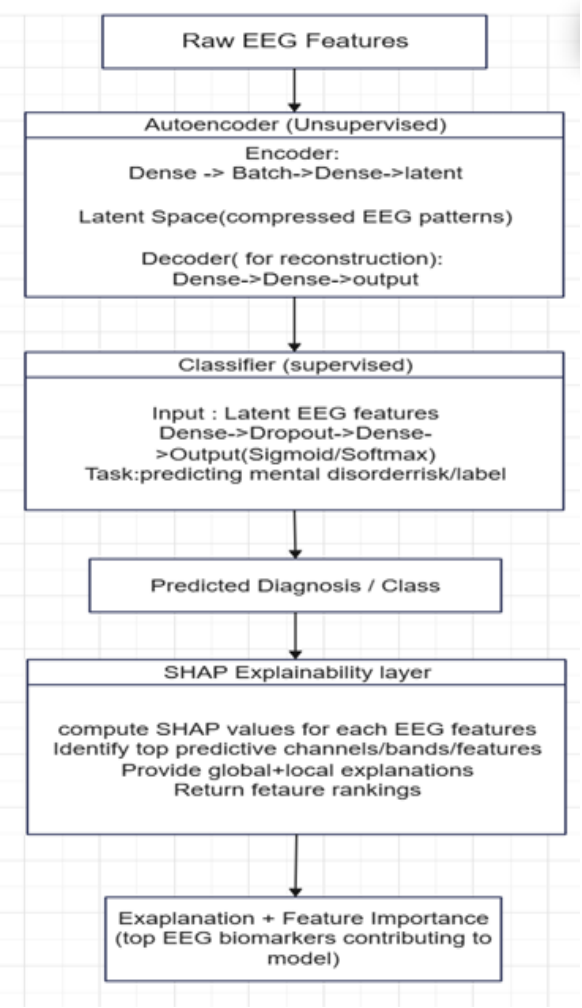


Figure 1. Proposed architecture for the theoretical model using DL.

Classifier on Latent Features

The classifier learns discriminative patterns that separate different mental health categories based on EEG-derived (Zeng *et.al.*, 2021) latent representations. The classifier's purpose is to convert the EEG signal into diagnostic predictions such as ADHD, depression, stress-level categories (Rashid *et.al.*, 2022) or normal/abnormal classification. Compressed latent representations are then passed into a simple feed-forward neural network classifier (Ji *et.al.*, 2024). Input to this layer is a latent vector, but dense layers with dropout prevent overfitting. If the Output needed is multi-class classification, we can use softmax, and if we need binary classification, we can use the sigmoid function.

SHAP explainability layer

SHAP, SHapley Additive exPlanations layers' purpose is to enhance clinical trust and interpretability. SHAP (Fathy *et.al*, 2023; Yin *et.al.*, 2021) is applied to the end-to-end model(EEG --> Encoder --> Classifier). The SHAP algorithm estimates the contribution of each original EEG feature to the model's final prediction. SHAP also enables identification of predictive EEG biomarkers such as abnormal band power distributions or dysfunctional connectivity patterns.

Expected Results

As this is a mere theoretical model, we can only expect outcomes based on evidence in EEG processing and deep learning. It may include

- Low-dimensional latent EEG representations are expected while using autoencoders that will catch neural dynamics, key spectral characteristics and spatial relations between EEG channels. This will reduce redundancy and simplify downstream classification tasks.
- Latent representations are expected to improve prediction accuracy vs raw EEG features, reduce model complexity and enable stable and generalisable predictions.
- The top features list is expected as ranking enhances clinical interpretability and will guide future EEG-based biomarker studies.

Advantages of the Proposed Approach

The proposed framework integrates Autoencoder-based feature learning with SHAP-based explainability to form a clinical interpretable mental disorder prediction model. Using an Autoencoder will reduce EEG noise, and it will capture hidden structure. SHAP will show which EEG features drive predictions. This proposed system is modular in nature.

Limitation

This theoretical model will require a very large EEG dataset, as we are using a deep learning model. Small sample sizes will lead to unstable and latent features that may later affect reliability. Autoencoder may learn dataset-specific artefacts, as its representation learning models that learn complex, non-linear patterns. Using SHAP can be computationally expensive. SHAP complexity grows exponentially with the number of features, even though it is optimised. This theoretical model uses tabular EEG features. However, EEG is fundamentally a time-series signal.

FUTURE SCOPE

This proposed theoretical model establishes a foundation for interpretable EEG-based mental disorder prediction. In future, Exploratory Data Analysis (EDA) and Questionnaire data (Multimodal Fusion) can be integrated. Existing autoencoder learns deterministic latent features. Using a Variational Autoencoder for a more robust latent representation can mitigate overfitting. The current model operates on static EEG feature vectors. Future work could include Long Short Term Memory (LSTM) networks or hybrid approaches, and even transformers can be used. Integrating sequential architectures could significantly improve disorder prediction accuracy. The SHAP explainability is expected to highlight specific EEG features consistently associated with mental disorders. Explainable biomarkers such as frontal theta activity, gamma power changes, and alpha coherence imbalance help bridge the gap between data-driven insights and neurophysiological understanding.

CONCLUSION

This theoretical model presents a hybrid deep learning framework combining autoencoder-based EEG feature compression with SHAP for early detection of mental disorders. The approach emphasises predictive accuracy and interpretability essential for real-world adoption in mental health diagnostics.

REFERENCES

Abbas, A., Khan, Y. D., & Kadry, S. (2022). Schizophrenia detection based on EEG using recurrent auto-encoder framework. arXiv. https://arxiv.org/abs/2207.04262

Amini, L., & Fazel, M. (2022). Interpretable deep learning for Parkinson's disease detection using EEG biomarkers. *IEEE Access, 10*, 34621-34632.

Awan, M. J., Mohammed, M. A., Kadry, S., Abdulkareem, K. H., & Al-Nasif, A. (2023). Explainable artificial intelligence models for predicting depression. *Diagnostics, 15*(11), 1412. https://doi.org/10.3390/diagnostics15111412 PMid:40506984 PMCid:PMC12154222.

Das, S., & Bhowmik, A. (2022). Explainable artificial intelligence for mental state classification using EEG signals. *Expert Systems with Applications, 201*, 117010. https://doi.org/10.1016/j.eswa.2022.117010

Fathy, M. E., & Hossny, M. (2023). SHAP-enhanced deep neural networks for clinical EEG interpretation. *Artificial Intelligence in Medicine, 146*, 102700. https://doi.org/10.1016/j.artmed.2023.102700 PMid:38184363.

Hammam, A., Refaat, M. M., & Sakr, G. E. (2021). EEG-based ADHD detection using sparse autoencoder representations. *Biomedical Engineering Letters, 11*, 147-158.

Jain, R., & Mishra, D. (2023). Autoencoder-based EEG representation learning for ADHD detection in children. *Computers in Biology and Medicine, 161*, 106956. https://doi.org/10.1016/j.compbiomed.2023.106956 PMid:37116241.

Ji, H., Zhang, J., & Hu, B. (2024). Transformer-based neural networks for EEG-driven psychiatric disorder detection. *Neural Networks, 172*, 106178.

Kheder, W., & Gagnon, J. F. (2021). Early detection of PTSD using deep learning models applied to EEG features. *Journal of Affective Disorders, 294*, 487-495. https://doi.org/10.1016/j.jad.2021.06.049 PMid:34252863 PMCid: PMC8410643.

Khan, A., Nisar, H., & Lee, S. (2023). Interpretable and robust AI in EEG systems: A survey. arXiv. https://arxiv.org/abs/2304.10755

Li, X., Song, D., Zhang, P., & Chen, M. (2020). Deep learning approach for depression recognition using EEG features. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 28*(12), 3041-3052. https://doi.org/10.1109/TNSRE.2020.3037321

MousaviU:, Z., & Yoon, E. (2021). EEG abnormality detection using deep autoencoder networks. *Biomedical*

Signal Processing and Control, 68, 102587. https://doi.org/10.1016/j.bspc.2021.102587

Múnera, D. E., Vásquez, G. E., González, C. E., & López, D. M. (2023). Feature engineering of EEG applied to mental disorders: A systematic mapping study. *Applied Intelligence, 53*(9), 10865-10895. https://doi.org/10.1007/s10489-023-04702-5

Park, S., & Lee, J. H. (2024). Interpretable deep learning methods for ADHD diagnosis using EEG biomarkers. *Computers in Biology and Medicine, 155*, 106699. https://doi.org/10.1016/j.compbiomed.2023.106699 PMid:36857941.

Raghu, S., & Sriraam, N. (2022). Multimodal autoencoder fusion for mental health assessment using EEG and physiological signals. *Biomedical Signal Processing and Control, 75*, 103590. https://doi.org/10.1016/j.bspc.2022.103590

Rasheed, A. U., K., Qayyum, A., & Alaa, M. (2022). Stress level classification using EEG deep autoencoder features. *Neural Computing and Applications, 34*, 10523-10536.

Saba-Sadiya, S., AlZoubi, O., & Calvo, R. A. (2022). EEG2Vec: Learning affective EEG representations via variational autoencoders. arXiv.

Siddiqui, M. F., & Ahmad, M. (2022). EEG-based anxiety detection using stacked autoencoders. *Biomedical Signal Processing and Control, 71*, 103198. https://doi.org/10.1016/j.bspc.2021.103198

Sturm, I., Lapuschkin, S., Samek, W., & Müller, K.-R. (2020). Explainable classification of EEG data using Shapley values. *Proceedings of the Cognitive Computational Neuroscience Conference*.

Völker, B., & Hansen, J. C. (2023). Increasing the sensitivity of EEG signals with explainable AI methods. *Behavior Research Methods, 56*, 1231-1248. https://doi.org/10.3758/s13428-023-02335-7 PMid:38453828 PMCid:PMC11335964.

Yin, Z., Zhao, M., & Wang, Y. (2023). Explainable deep learning for EEG-based emotion recognition using SHAP. *IEEE Transactions on Affective Computing*.

Zhai, W., Li, H., Xu, X., Li, J., Xue, J., Liu, Y., & Wang, Z. (2025). A data-centric and interpretable EEG framework for depression severity grading. *Journal of NeuroEngineering and Rehabilitation, 22*(1), Article 1645. https://doi.org/10.1186/s12984-025-01645-5 PMid:40414904 PMCid:PMC12103758.

Zeng, Y., Zhang, J., & Hu, B. (2021). Emotion recognition from EEG using a multi-scale convolutional autoencoder. *Knowledge-Based Systems, 228*, 107235. https://doi.org/10.1016/j.knosys.2021.107235