GenAI-RAG-EEG: A Novel Hybrid Deep Learning Architecture with Retrieval-Augmented Generation for Explainable EEG-Based Stress Classification

Praveen Asthana*§, Rajveer Singh Lalawat†, and Sarita Singh Gond‡ *Independent Researcher, Calgary, Canada †Department of Electronics and Communication Engineering, IIITDM Jabalpur, India ‡Department of Bioscience, Rani Durgavati University, Jabalpur, India §Corresponding Author: Praveenairesearch@gmail.com

***Abstract*—This paper presents GenAI-RAG-EEG, a novel hybrid deep learning architecture integrating Generative AI, Retrieval-Augmented Generation (RAG), and advanced EEG signal processing for explainable stress classification. Our architecture combines a core EEG classifier (1D-CNN, Bi-LSTM, self-attention with 138K parameters) with a RAG-enhanced explanation module. We evaluate on three public datasets: DEAP (32 subjects, arousal-based stress proxy), SAM-40 (40 subjects, cognitive stress), and WESAD (15 subjects, physiological stress). The system achieves 94.7% accuracy on DEAP, 93.2% on SAM-40, and 100% on WESAD. Signal analysis reveals consistent stress biomarkers: alpha suppression (31-33%), decreased theta/beta ratio (8-14%), and frontal alpha asymmetry shift toward right hemisphere dominance. The RAG module provides clinically meaningful explanations with 89.8% expert agreement, though it does not significantly improve classification accuracy ($p = 0.312$). Cross-dataset transfer experiments reveal 14-27% accuracy drops, validating distinct stress paradigms. Statistical validation includes Leave-One-Subject-Out cross-validation, 95% confidence intervals, and multiple comparison corrections. Our results establish a framework for explainable EEG-based stress detection suitable for real-time brain-computer interface applications.**

***Index Terms*—EEG, stress detection, deep learning, RAG, explainable AI, attention mechanism, LSTM, cognitive workload**

I. Introduction

STRESS and cognitive workload significantly impact human health and productivity globally. The World Health Organization reports chronic stress affects over 300 million people, contributing to cardiovascular disease and cognitive impairment [1]. Traditional assessment using self-report questionnaires suffers from recall bias and cannot capture real-time fluctuations.

Electroencephalography (EEG) offers objective, non-invasive stress assessment with millisecond temporal resolution [2]. Stress states manifest as alpha-band (8–13 Hz) suppression, beta-band (13–30 Hz) elevation, and increased frontal theta activity [3]. Deep learning advances enable end-to-end feature learning, achieving improvements over traditional machine learning [4].

Despite impressive classification performance, existing methods lack explainability—a critical barrier to clinical adoption [5]. The emergence of Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) [6] presents opportunities for explainable AI in healthcare.

A. Related Work

Recent EEG-based stress detection methods show promising results but significant limitations. Song et al. [7] achieved 90.4% using DGCNN on SEED/DEAP. Tao et al. [8] reported 88.7% with attention-enhanced CRNN. Chen et al. [9] obtained 89.7% using CNN-LSTM hybrid. Wang et al. [10] explored transformers achieving 91.2%. Li et al. [11] proposed bi-hemisphere networks reaching 92.1%. However, these lack explainability and rigorous statistical validation.

B. Contributions

Our contributions include: (1) A hybrid architecture combining CNN, Bi-LSTM, and self-attention with RAG-enhanced explanations; (2) Systematic evaluation across three datasets with distinct stress paradigms; (3) Comprehensive signal analysis revealing consistent biomarkers; (4) RAG explanation evaluation achieving 89.8% expert agreement; (5) Rigorous statistical validation with LOSO cross-validation and confidence intervals.

II. Methodology

A. Problem Definition

Given EEG segment $\mathbf{X} \in \mathbb{R}^{C \times T}$ where C is channels and T is samples, predict stress label $y \in \{0, 1\}$ and generate explanation E grounded in scientific evidence.

B. Datasets

DEAP [12]: 32 subjects watching music videos, 32 EEG channels at 128 Hz. Arousal ratings binarized as stress proxy (>5 = stress).

SAM-40 [13]: 40 subjects performing cognitive tasks (Stroop, arithmetic, mirror tracing), 32 channels at 256 Hz. Validated with NASA-TLX and skin conductance.

WESAD [14]: 15 subjects undergoing TSST protocol, multimodal signals at 700 Hz. Binary stress/baseline labels from validated protocol.

C. Preprocessing Pipeline

Raw signals undergo: (1) Bandpass filtering (0.5–45 Hz, 4th-order Butterworth); (2) Notch filtering (50 Hz power line removal); (3) Artifact rejection (>100 μV threshold); (4) Segmentation (4-second windows, 50% overlap); (5) Z-score normalization per channel.

D. Model Architecture

The EEG Encoder comprises three convolutional blocks followed by Bi-LSTM and self-attention:

Conv Block 1: Conv1D(32, 32, k=7) → BatchNorm → ReLU → MaxPool(2)

Conv Block 2: Conv1D(32, 64, k=5) → BatchNorm → ReLU → MaxPool(2)

Conv Block 3: Conv1D(64, 64, k=3) → BatchNorm → ReLU → MaxPool(2)

Bi-LSTM: 2 layers, 64 hidden units bidirectional, outputting 128-dimensional sequence.

Self-Attention: Query projection, energy computation via tanh, softmax normalization, weighted context aggregation.

The Text Encoder uses frozen Sentence-BERT (all-MiniLM-L6-v2) [15] with projection layer (384 → 128). Features are concatenated and classified via MLP (256 → 64 → 32 → 2).

Total parameters: 138,081 (EEG) + 49,152 (Text) + 10,402 (Classifier) = 197,635.

E. RAG Explanation Module

The RAG pipeline retrieves relevant scientific literature using FAISS [16] vector search with Sentence-BERT embeddings. Retrieved contexts augment LLM prompts for generating explanations grounded in evidence.

F. Training Configuration

Optimizer: AdamW (β_1=0.9, β_2=0.999); Learning rate: 10^{-4} with ReduceLROnPlateau; Batch size: 64; Epochs: 100 with early stopping (patience=10); Dropout: 0.3; Weight decay: 0.01; Loss: Cross-entropy with class weights.

III. Signal Analysis

A. Band Power Analysis

We computed power spectral density using Welch's method (256-sample windows, 50% overlap) across five frequency bands. Table I shows consistent patterns: delta and theta increase, alpha decreases, beta and gamma increase during stress.

TABLE I: Band Power Effect Sizes (Cohen's d) Across Datasets

Band	**Hz**	**DEAP**	**SAM-40**	**WESAD**	p
Delta	0.5–4	+0.38	+0.42	+0.35	$<.01$
Theta	4–8	+0.62	+0.68	+0.55	$<.001$
Alpha	8–13	−0.82	−0.89	−0.75	$<.001$
Beta	13–30	+0.71	+0.74	+0.58	$<.001$
Gamma	30–45	+0.48	+0.51	+0.41	$<.05$

B. Alpha Suppression

Alpha suppression, a hallmark stress biomarker [3], showed 31.4% (DEAP), 33.3% (SAM-40), and 31.7% (WESAD) reduction during stress (all $p < 0.0001$).

C. Theta/Beta Ratio

The theta/beta ratio (TBR), linked to cognitive load [17], decreased 14.0% (DEAP), 11.2% (SAM-40), and 8.2% (WESAD) during stress, indicating increased cortical arousal.

D. Frontal Alpha Asymmetry

Frontal alpha asymmetry (FAA = ln(Right) − ln(Left)) shifted toward right hemisphere dominance during stress: ΔFAA = −0.26 (DEAP), −0.27 (SAM-40), −0.22 (WESAD), consistent with approach-withdrawal theory [18].

IV. Experimental Results

A. Classification Performance

Table II presents classification metrics using Leave-One-Subject-Out (LOSO) cross-validation.

TABLE II: Classification Performance Across Datasets

Dataset	**Acc**	**F1**	**AUC**	**BA**	κ
DEAP	94.7%	94.3%	96.7%	94.5%	0.894
SAM-40	93.2%	92.8%	95.8%	93.1%	0.864
WESAD	100.0%	100.0%	100.0%	100.0%	1.000

B. Baseline Comparison

Table III compares our method against traditional and deep learning baselines on SAM-40.

TABLE III: Comparison with Baseline Methods (SAM-40)

Method	**Acc**	**F1**	**AUC**
SVM (RBF) [19]	74.8%	87.0%	65.0%
Random Forest [20]	76.2%	86.0%	70.0%
XGBoost [21]	77.5%	86.0%	72.0%
CNN [22]	78.3%	86.0%	74.0%
LSTM [23]	79.1%	87.0%	75.0%
CNN-LSTM [24]	80.2%	87.0%	76.0%
EEGNet [25]	79.8%	87.0%	75.0%
DGCNN [7]	80.6%	87.0%	77.0%
GenAI-RAG-EEG	**93.2%**	**92.8%**	**95.8%**

C. Ablation Study

Table IV shows component contributions to model performance.

TABLE IV: Ablation Study Results

Variant	**Accuracy**	**Δ**
Full Model	93.2%	–
− Text Encoder	91.5%	−1.7%
− Self-Attention	91.1%	−2.1%
− Bi-LSTM	89.6%	−3.6%
− RAG Module	93.0%	−0.2%
CNN Only	89.6%	−3.6%

D. Cross-Dataset Transfer

Table V reveals domain shift between stress paradigms.

TABLE V: Cross-Dataset Transfer Results

Source	Target	Acc	Drop
SAM-40	DEAP	71.4%	−21.8%
DEAP	SAM-40	68.2%	−26.5%
SAM-40	WESAD	78.6%	−14.6%
WESAD	SAM-40	76.8%	−16.4%

E. RAG Explanation Evaluation

The RAG module was evaluated exclusively on SAM-40 with validated stress labels. Expert evaluation (3 domain experts, blinded) showed 89.8% agreement with generated explanations. However, RAG did not significantly improve classification accuracy (Δ = +0.2%, p = 0.312, Wilcoxon signed-rank test).

F. Feature Importance

Gradient-based feature importance revealed top contributors: Frontal Alpha Power (15.6%), Theta/Beta Ratio (14.2%), Frontal Alpha Asymmetry (12.8%), Central Beta Power (11.2%), and Parietal Alpha Power (9.8%).

V. Discussion

A. Performance Analysis

Our model achieves state-of-the-art performance across all datasets. The 12.6% improvement over DGCNN on SAM-40 demonstrates the effectiveness of combining temporal (Bi-LSTM), spatial (CNN), and contextual (attention) processing. Perfect WESAD performance likely reflects the TSST protocol's pronounced physiological stress response [26].

B. Signal Analysis Insights

Consistent biomarker patterns across datasets validate stress detection mechanisms. Alpha suppression aligns with reduced relaxation [3]. Decreased TBR indicates increased cognitive load [17]. FAA shifts toward right dominance support Davidson's approach-withdrawal model [18].

C. Cross-Dataset Generalization

The 14-27% accuracy drops in transfer experiments reveal fundamental differences between arousal-based (DEAP) and cognitive stress (SAM-40) paradigms. This validates treating DEAP arousal as stress proxy rather than equivalent construct [27].

D. Explainability Trade-offs

While RAG provides clinically meaningful explanations, it does not improve predictions. This suggests classification and explanation may benefit from distinct optimization objectives, aligning with recent findings on explanation fidelity [28].

E. Limitations

Key limitations include: (1) Laboratory-controlled data may not generalize to real-world settings; (2) Subject pool demographics may limit applicability; (3) Electrode configurations differ across datasets; (4) RAG explanations require external LLM access.

F. Clinical Implications

The system's high accuracy and explainability support potential clinical applications in stress monitoring, occupational health assessment, and mental health screening. The attention visualization enables clinicians to verify model reasoning.

VI. Conclusion

We presented GenAI-RAG-EEG, a hybrid architecture achieving 93-100% accuracy across three stress datasets with explainable predictions. Signal analysis revealed consistent biomarkers: 31-33% alpha suppression, decreased theta/beta ratio, and right-shifted frontal asymmetry. The RAG module provides clinically meaningful explanations with 89.8% expert agreement. Future work includes real-time implementation, larger clinical validation, and multimodal integration.

References

[1] World Health Organization, "Mental health in the workplace," WHO Technical Report, 2023.

[2] M. Teplan, "Fundamentals of EEG measurement," *Measurement Science Review*, vol. 2, no. 2, pp. 1–11, 2002.

[3] W. Klimesch, "EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis," *Brain Research Reviews*, vol. 29, no. 2-3, pp. 169–195, 1999.

[4] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: a systematic review," *Journal of Neural Engineering*, vol. 16, no. 5, p. 051001, 2019.

[5] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, "What clinicians want: contextualizing explainable machine learning for clinical end use," in *Machine Learning for Healthcare*, pp. 359–380, 2019.

[6] P. Lewis, E. Perez, A. Piktus, et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[7] T. Song, W. Zheng, P. Song, and Z. Cui, "EEG emotion recognition using dynamical graph convolutional neural networks," *IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 532–541, 2020.

[8] W. Tao, C. Li, R. Song, J. Cheng, Y. Liu, and X. Chen, "EEG-based emotion recognition via channel-wise attention and self attention," *IEEE Transactions on Affective Computing*, 2020.

[9] J. Chen, Z. Chen, Z. Chi, and H. Fu, "Accurate EEG-based emotion recognition on combined features using deep convolutional neural networks," *IEEE Access*, vol. 9, pp. 44317–44328, 2021.

[10] Y. Wang, S. Qiu, J. Li, and H. Ma, "EEG-based emotion recognition with transformers and 2D convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 79, p. 104233, 2022.

[11] J. Li, S. Qiu, C. Du, Y. Wang, and H. He, "Domain adaptation for EEG emotion recognition based on latent representation similarity," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 15, no. 4, pp. 1879–1892, 2023.

[12] S. Koelstra, C. Muhl, M. Soleymani, et al., “DEAP: A database for emotion analysis using physiological signals,” *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[13] R. Gupta, K. Laghari, and T. Falk, “Relevance vector classifier decision fusion and EEG graph-theoretic features for automatic affective state characterization,” *Neurocomputing*, vol. 174, pp. 875–884, 2016.

[14] P. Schmidt, A. Reiss, R. Duerichen, C. Marber, and K. Van Laerhoven, “Introducing WESAD, a multimodal dataset for wearable stress and affect detection,” in *International Conference on Multimodal Interaction*, pp. 400–408, 2018.

[15] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using Siamese BERT-networks,” in *EMNLP-IJCNLP*, pp. 3982–3992, 2019.

[16] J. Johnson, M. Douze, and H. Jégou, “Billion-scale similarity search with GPUs,” *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2019.

[17] P. Putman, J. van Peer, I. Maimari, and S. van der Werff, “EEG theta/beta ratio in relation to fear-modulated response-inhibition, attentional control, and affective traits,” *Biological Psychology*, vol. 83, no. 2, pp. 73–78, 2014.

[18] R. J. Davidson, “Anterior cerebral asymmetry and the nature of emotion,” *Brain and Cognition*, vol. 20, no. 1, pp. 125–151, 1990.

[19] A. Subasi, “EEG signal classification using wavelet feature extraction and a mixture of expert model,” *Expert Systems with Applications*, vol. 32, no. 4, pp. 1084–1093, 2010.

[20] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[21] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *ACM SIGKDD*, pp. 785–794, 2016.

[22] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, et al., “Deep learning with convolutional neural networks for EEG decoding and visualization,” *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[23] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[24] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, “Learning representations from EEG with deep recurrent-convolutional neural networks,” in *ICLR*, 2016.

[25] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, “EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces,” *Journal of Neural Engineering*, vol. 15, no. 5, p. 056013, 2018.

[26] C. Kirschbaum, K.-M. Pirke, and D. H. Hellhammer, “The ‘Trier Social Stress Test’–a tool for investigating psychobiological stress responses in a laboratory setting,” *Neuropsychobiology*, vol. 28, no. 1-2, pp. 76–81, 1993.

[27] L. Shu, J. Xie, M. Yang, et al., “A review of emotion recognition using physiological signals,” *Sensors*, vol. 18, no. 7, p. 2074, 2018.

[28] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[29] A. Vaswani, N. Shazeer, N. Parmar, et al., “Attention is all you need,” in *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

[30] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *ICLR*, 2015.