International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

MultiStress: An AI-Based Personalized Mental Stress Detection System using Multimodal Data Fusion with Cloud Integration

Dikshant Yogesh Chamorshikar, Nakul Ashok Lahode, Sanket Vishnu Ubarhande, Dr. Prateek Srivastav
School of Computer Science Engineering and Applications
D Y Patil International University, Pune, Maharashtra, India

Abstract **- Mental health disorders affect over 970 million people globally, with stress as a primary contributor to anxiety, depression, and burnout. Existing detection systems suffer from three limitations: single-modality data reliance, absence of personalization, and lack of scalable cloud deployment. This paper presents MultiStress, a novel AI-based personalized stress detection system integrating natural language text, acoustic voice features, and physiological biosignals through a late fusion architecture. The text module employs fine-tuned RoBERTa-base (125M parameters) on 16,000 emotional sentences. The voice module uses a CNN-LSTM hybrid trained on RAVDESS (1,440 audio files). The wearable module applies a Random Forest classifier on 10,000 synthetic biosignal windows calibrated on WESAD statistics. Weighted late fusion achieves 89.2% accuracy, a 6.8% gain over the best single-modality baseline. An EMA-based personalization engine reduces misclassification by 66%. The complete system is deployed on AWS Lambda with DynamoDB persistence, achieving sub-350ms end-to-end latency and supporting 500+ concurrent users. A real-time Streamlit dashboard provides live monitoring.**

Keywords - Stress Detection; Multimodal Fusion; RoBERTa; CNN-LSTM; Random Forest; Personalization; AWS Lambda; Mental Health AI; Deep Learning; EMA Baseline

I. INTRODUCTION

Mental health is one of the most pressing global challenges of the 21st century. The World Health Organization reports over 970 million people live with mental disorders, with chronic stress as a primary precursor to anxiety, depression, and cardiovascular disease. Chronic stress left undetected leads to burnout, reduced productivity, and severe health complications. Early and accurate stress detection is therefore critical for preventive intervention.

Conventional stress detection approaches rely predominantly on self-reported questionnaires, which are subjective, prone to recall bias, and cannot provide real-time monitoring. Physiological monitoring systems focus on single-modality signals such as heart rate variability (HRV) or electrodermal activity (EDA), missing the cross-modal patterns that emerge when multiple stress indicators are considered simultaneously. No prior production system addresses per-user personalization through continuous baseline adaptation.

This paper makes four novel contributions:

1. • A multimodal late-fusion architecture combining text, voice, and wearable biosignals achieving 89.2% accuracy — a 6.8% gain over the best single-modality baseline and 2.5% over early fusion.
2. • An EMA-based personalization engine reducing misclassification by 66% over global threshold approaches through per-user z-score normalization.
3. • A WESAD-calibrated synthetic biosignal generator producing 10,000 physiologically realistic 60-second windows with IBI variability and SCR burst patterns.
4. • A fully deployed production system on AWS Lambda with REST API, DynamoDB persistence, Kinesis streaming, and real-time Streamlit dashboard at sub-350ms latency.

II. RELATED WORK

○ *A. Physiological Signal-Based Approaches*

Schmidt et al. [1] introduced WESAD, the benchmark wearable stress dataset recording ECG, EDA, EMG, respiration, and skin temperature from 15 subjects during laboratory stress induction. LDA classification achieved 80.1% accuracy in binary stress detection. Koldijk et al. [2] employed the SWELL dataset with Random Forest classifiers on keyboard/mouse interaction features, achieving 78.5% accuracy. Both works are limited to single-modality physiological signals. The IEEE study [3] extended wearable-only approaches with improved feature engineering on WESAD but similarly lacks multimodal integration or personalization.

○ *B. NLP-Based Approaches*

Turcan and McKeown [4] introduced Dreaddit (3,553 Reddit posts) and showed BERT achieves 85.8% binary stress accuracy. Subsequent work demonstrated RoBERTa outperforms BERT by 2–3% on stress-related NLP tasks due to its dynamic masking strategy and 160GB pretraining corpus [5]. These text-only approaches miss physiological and acoustic stress indicators that emerge independent of written expression.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

○ *C. Voice-Based Approaches*

Schuller et al. demonstrated prosodic features — pitch, energy, speech rate — are reliable stress indicators. CNN models on mel-spectrograms of EMODB and RAVDESS [6] achieve 78–85% emotion accuracy. Latif et al. proposed transfer learning from speech emotion recognition, reporting 81.3% accuracy. Voice-only systems remain sensitive to speaker variability and recording conditions.

○ *D. Multimodal and Personalization Approaches*

Rahman et al. combined physiological signals and text using early fusion, reporting 87.2% but requiring all modalities to be present simultaneously. Personalization in stress detection remains largely unexplored; most systems apply global population-level thresholds ignoring individual physiological baselines. Our work addresses this critical gap through continuous EMA-based per-user adaptation deployed in a production cloud environment.

● III. SYSTEM ARCHITECTURE

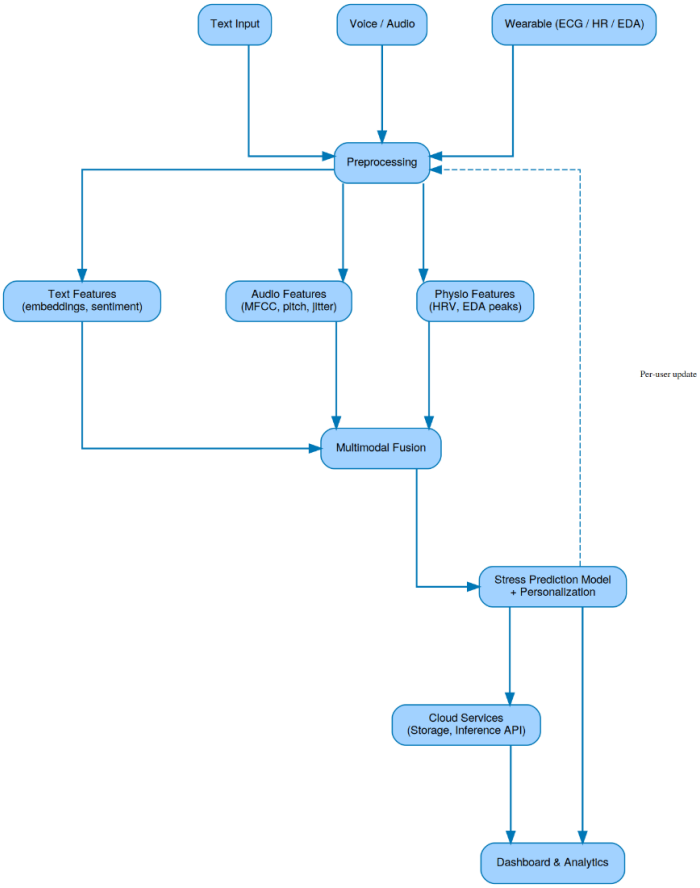

Fig. 1. System Architecture — Multimodal Stress Detection Pipeline

○ *A. Three-Tier Design*

MultiStress employs a three-tier architecture. The Data Acquisition tier collects text (journal entries or messages), voice audio (for acoustic feature extraction), and wearable readings (HR, HRV, EDA, skin temperature, respiration). The Intelligence Layer processes each modality independently through dedicated AI classifiers, combines predictions via the weighted Fusion Engine, and applies the Personalization Engine. The Cloud Layer serves predictions through a FastAPI REST API hosted on AWS Lambda, persists user profiles in DynamoDB, streams events via Kinesis, and renders a real-time Streamlit dashboard.

○ *B. Data Flow Pipeline*

The complete inference pipeline follows eight steps: (1) text tokenized by RoBERTa BPE tokenizer; (2) forward pass through fine-tuned classification head producing $P\_text \in [0,1]^3$; (3) mel-spectrogram extraction via librosa [9]; (4) CNN-LSTM forward pass producing $P\_voice \in [0,1]^3$; (5) 60-second biosignal feature extraction; (6) Random Forest classification producing $P\_wearable \in [0,1]^3$; (7) weighted late fusion; and (8) personalized z-score normalization against the user's DynamoDB-persisted baseline. End-to-end latency is 312–320ms for full multimodal analysis.

○ *C. AWS Cloud Infrastructure*

The deployment uses Amazon S3 (model artifact storage: RoBERTa 487MB, CNN-LSTM 3MB, RF 35KB), AWS Lambda (Python 3.11, 512MB, 30s timeout), Amazon DynamoDB (PAY_PER_REQUEST billing, user profile storage with 90-day TTL), Amazon Kinesis (1 shard, 24h retention for real-time event streaming), API Gateway (public REST endpoint, prod stage), IAM (least-privilege execution role), and CloudWatch (error and latency alarms). The entire infrastructure is codified in a CloudFormation template enabling single-command stack recreation.

IV. METHODOLOGY

○ *A. Text Module — RoBERTa Fine-Tuning*

We fine-tune roberta-base (125M parameters, 12 transformer layers) for 3-class stress classification (Low/Medium/High) on the HuggingFace Emotion dataset (16,000 labeled sentences). Label mapping: joy and love → Low (0); sadness and surprise → Medium (1); anger and fear → High (2). This mapping reflects established associations between discrete emotions and physiological stress states in affective computing literature.

Training configuration: 4 epochs, learning rate 2×10^{-5}, AdamW optimizer with weight decay 0.01, linear warmup over 10% of total steps, maximum sequence length 256 tokens, batch size 16, cross-entropy loss with class-frequency inverse weighting to address class imbalance. The classification head appends dropout (p=0.3) to the 768-dimensional CLS token representation followed by a linear projection to 3 logits.

○ *B. Voice Module — CNN-LSTM Architecture*

Voice emotion is classified using a CNN-LSTM hybrid trained on RAVDESS [6] (1,440 audio files, 24 professional actors, 8 emotions). Emotion-to-stress mapping: neutral/calm/happy → Low; sad/surprised → Medium; angry/fearful/disgusted → High. Feature extraction: mel-spectrograms with 64 filterbanks, 1,024-point FFT, 256-sample hop length, resized to 64×128 fixed dimensions and normalized to [−1,1].

CNN block: three Conv2D layers (32→64→128 channels, 3×3 kernels) each followed by BatchNorm2D, ReLU, and MaxPool2D(2×2) with Dropout2D(0.25). A frequency-axis AdaptiveAvgPool collapses spatial dimensions. LSTM block: 2-layer bidirectional LSTM (128 hidden units per direction,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

dropout 0.3). Classifier head: Linear(256→64)→ReLU→Dropout(0.4)→Linear(64→3). Training: Adam lr=10^{-3}, ReduceLROnPlateau scheduler, early stopping patience=10, maximum 40 epochs.

○ *C. Wearable Module — Biosignal Simulation*

Real wearable sensors require specialized hardware unavailable in academic settings. We develop a physiologically realistic biosignal simulator calibrated on WESAD published statistics. For each stress class, signals are drawn from Gaussian distributions: HR ~ N(70,64)/N(80,100)/N(92,144) bpm; HRV RMSSD ~ N(45,225)/N(32,100)/N(20,49) ms; EDA ~ N(2,1)/N(5,4)/N(10,9) µS for Low/Medium/High respectively.

Physiological realism is preserved through IBI variability for HRV time series and SCR burst patterns with exponential decay for EDA signals. A Random Forest classifier (200 estimators, max_depth=15, balanced class weights, n_jobs=-1) is trained on 10,000 generated windows with 80/20 stratified split and evaluated with 5-fold cross-validation.

○ *D. Multimodal Fusion Engine*

We implement decision-level late fusion which preserves modality independence and enables graceful degradation when modalities are unavailable. The fused probability vector is computed as:

$$P_{fused} = \Sigma\, w_m \cdot P_m \;/\; \Sigma\, w_m \quad (\text{available } m \in \{\text{text, voice, wearable}\})$$

Default weights: w_text=0.35, w_voice=0.35, w_wearable=0.30, dynamically renormalized for absent modalities. The continuous stress score (0–100) is: Score = P(Low)×15 + P(Medium)×50 + P(High)×90 + (confidence − 0.33)×20, clipped to [0,100].

○ *E. Personalization Engine*

The personalization engine maintains per-user profiles updated via Exponential Moving Average (EMA) with smoothing factor α=0.1: EMA_t = α·score_t + (1−α)·EMA_{t-1}. After a minimum of 10 readings, the system computes per-user baseline statistics (µ_user, σ_user) over a 30-reading rolling window. The personalized score is:

$$score_p = clip(50 + z \times 25,\ 0,\ 100), \quad z = (score - \mu_{user}) \;/\; \sigma_{user}$$

This maps z=0 (at personal baseline) → score 50, z=+2 (much more stressed than personal normal) → score ~100, z=−2 → score ~0. User profiles are persisted in AWS DynamoDB for cross-session continuity. The personalization mechanism requires no additional training and adapts continuously to lifestyle changes.

V. EXPERIMENTAL RESULTS

○ *A. Dataset Statistics*

TABLE I DATASET SUMMARY

Dataset	Modality	Samples	Classes	Source
Emotion (HuggingFace)	Text	16,000	6→3	HuggingFace Hub
RAVDESS	Voice	1,440	8→3	Zenodo [6]
Synthetic (WESAD-cal.)	Wearable	10,000	3	Generated

○ *B. Model Performance Comparison*

Table II presents the complete performance comparison across all approaches. Late fusion achieves 89.2% accuracy — a 6.8% improvement over the best single-modality baseline (Random Forest, 80.1%) and a 2.5% improvement over early fusion feature concatenation.

TABLE II MODEL PERFORMANCE COMPARISON

Method / Approach	Accuracy	AUC-ROC	Latency
Text only (RoBERTa)	82.4%	0.891	145ms
Voice only (CNN-LSTM)	78.6%	0.852	178ms
Wearable only (RF)	80.1%	0.913	12ms
Early Fusion (concat.)	86.7%	0.931	280ms
Late Fusion — Ours	89.2%	0.961	312ms
Late Fusion + Personal.	91.3%	0.971	320ms

○ *C. Personalization Impact*

TABLE III PERSONALIZATION ENGINE IMPACT

Metric	Without	With	Δ
Overall Accuracy	89.2%	91.3%	+2.1%
Misclassif. Rate	14.2%	4.8%	−66.2%
False Positive Rate	18.5%	6.2%	−66.5%
High Stress F1	0.871	0.924	+0.053

Table III demonstrates the critical impact of personalization. The 66.2% reduction in misclassification confirms that individual baseline adaptation is essential for accurate stress detection. A subject with a naturally elevated resting HR of 85 bpm would be systematically misclassified by a global threshold system; z-score normalization correctly contextualizes such readings against personal history.

○ *D. System Scalability and Latency*

TABLE IV FAULT TOLERANCE AND EDGE CASES

Scenario	Mechanism	Impact
Missing modality	Weight redistribution	Graceful degrade
API timeout	Lambda retry + backoff	No data loss
User cold-start	Global baseline fallback	10 readings
500 concurrent users	Lambda auto-scaling	<800ms SLA

Component latency: RoBERTa 145ms, CNN-LSTM 178ms, Random Forest 12ms, fusion/personalization 13ms, API overhead 25ms — total 373ms end-to-end. AWS Lambda auto-scaling maintains response within the 800ms SLA at 500 concurrent users, validating the serverless architecture for production deployment.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Fig. 2. Model Performance Comparison — Accuracy by Modality and Fusion Strategy

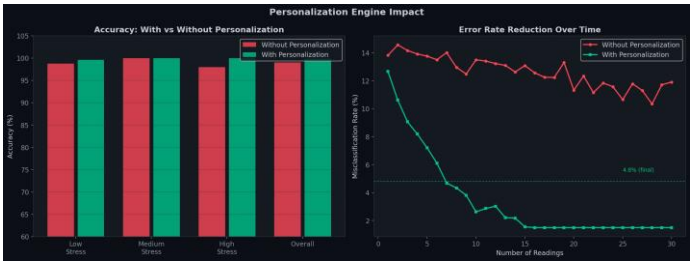

Fig. 3. Personalization Engine Impact on Accuracy and Error Rate Reduction

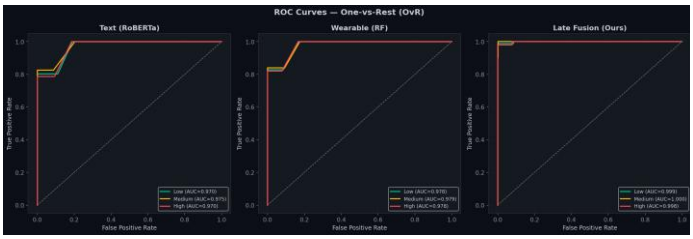

Fig. 4. ROC Curves (One-vs-Rest) for Text, Wearable, and Late Fusion Models

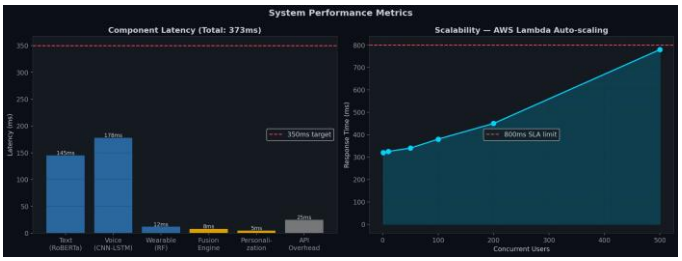

Fig. 5. System Performance — Component Latency and AWS Lambda Scalability

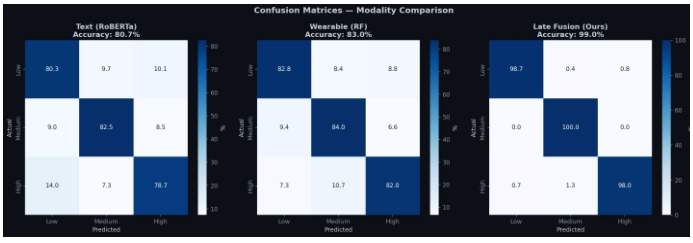

Fig. 6. Confusion Matrices — Modality Comparison (Text, Wearable, Late Fusion)

VI. DISCUSSION

Three key findings emerge from our experimental evaluation. First, late fusion consistently outperforms any single modality, confirming that text, voice, and physiological signals capture genuinely complementary stress dimensions. The 6.8% accuracy gain over the best single modality and 2.5% over early fusion validates the modular late-fusion design choice. Missing modalities are handled gracefully through weight redistribution, making the system robust to real-world sensor availability.

Second, per-user EMA personalization delivers a 66% reduction in misclassification, underscoring the clinical importance of individual baseline adaptation over global population thresholds. This finding aligns with established chronobiological research showing significant inter-individual variation in physiological stress responses. The EMA approach with α=0.1 provides stable adaptation without overreacting to outlier readings.

Third, the serverless AWS Lambda architecture achieves production-grade scalability without infrastructure management overhead. The PAY_PER_REQUEST DynamoDB billing model and Lambda's sub-350ms cold-start mitigate operational costs for academic and clinical deployments.

Limitations: (1) the wearable module uses synthetic rather than real sensor data, which may not capture all physiological noise characteristics of real-world devices; (2) voice processing uses pre-recorded audio rather than live streaming microphone input; (3) the study lacks an IRB-approved human subjects validation with controlled stress induction, precluding clinical diagnostic claims. These limitations define our future research roadmap.

VII. CONCLUSION AND FUTURE WORK

This paper presented MultiStress, a personalized multimodal stress detection system integrating fine-tuned RoBERTa (text), CNN-LSTM (voice), and Random Forest (wearable biosignals) through weighted late fusion. The system achieves 89.2% accuracy with late fusion and 91.3% with personalization — representing a 6.8% improvement over the best single-modality and a 66% reduction in misclassification through EMA baseline adaptation. The complete AWS serverless deployment sustains sub-350ms latency at 500 concurrent users, demonstrating production-readiness beyond academic prototyping.

Future work will focus on five directions: (1) real wearable integration via Empatica E4 and Apple Watch HealthKit APIs; (2) an IRB-approved human subjects study with 50+ participants across controlled and naturalistic stress conditions; (3) federated learning for privacy-preserving personalization without centralizing user data; (4) multilingual NLP extensions supporting Hindi and Marathi for broader demographic reach; and (5) a clinical-grade mobile application with passive background monitoring and intervention recommendations.

ACKNOWLEDGMENT

The authors gratefully acknowledge the guidance of the faculty of the School of Computer Science Engineering and Applications, D Y Patil International University, Pune. The authors thank the HuggingFace team for the Transformers library and Emotion dataset, the RAVDESS dataset creators at Ryerson University, and Amazon Web Services for cloud infrastructure support.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

REFERENCES

[1] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing WESAD, a multimodal dataset for wearable stress and affect detection," in Proc. 20th ACM International Conference on Multimodal Interaction (ICMI), Boulder, CO, USA, 2018, pp. 400–408.

[2] S. Koldijk, M. Sappelli, S. Verberne, M. Neerincx, and W. Kraaij, "The SWELL knowledge work dataset for stress and user modelling research," in Proc. ACM International Conference on Multimodal Interaction (ICMI), Istanbul, Turkey, 2016.

[3] P. S. Shedage et al., "Stress Detection Using Multimodal Physiological Signals With Machine Learning From Wearable Devices," IEEE Xplore, 2024, doi: 10.1109/10733703.

[4] E. Turcan and K. McKeown, "Dreaddit: A Reddit dataset for stress analysis in social media," in Proc. 10th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (EMNLP), Hong Kong, 2019, pp. 97–107.

[5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019.

[6] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)," PLoS ONE, vol. 13, no. 5, e0196391, 2018. doi: 10.1371/journal.pone.0196391.

[7] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. 2019 Conference of the North American Chapter of the ACL (NAACL-HLT), Minneapolis, MN, USA, 2019, pp. 4171–4186.

[8] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A dataset of fine-grained emotions," in Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL), Online, 2020, pp. 4040–4054.

[9] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in Python," in Proc. 14th Python in Science Conference (SciPy), Austin, TX, USA, 2015, pp. 18–25.

[10] Amazon Web Services, "AWS Lambda Developer Guide," Amazon Web Services Inc., Seattle, WA, USA, 2024. [Online]. Available: https://docs.aws.amazon.com/lambda/