Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 27 May 2026
Revised 24 June 2026
Accepted 16 July 2026

Natural Resources for Human Health 2026, Vol. 6 (10s): 324–335

KEYWORDS:
Multimodal stress detection, Systems without privacy invasion, Deep Learning, Affective computing, Biometric data security, CNN-LSTM, cross modal fusion.

A Multimodal Stress Detection System Using Text, Audio, and Video Analysis with CMII-Based Weighted Fusion

[1]Vijayakumar S D, [2]P Karunakaran, [3]Sumathi G, [4]Balasubramaniam C, [5]V Kumar, [6]Bharani S, [7]Bharanidharan G, [8]Keerthivarsan D

[1]Department of Artificial Intelligence and Data Science, Nandha Engineering College Erode, Tamil Nadu, India. mail2vijay.sd@gmail.com
[2]Department of Artificial Intelligence and Data Science, Nandha Engineering College, Erode – 638052, Tamilnadu, India. karuna.dr@gmail.com
[3]Department of Computer Science and Engineering, Nandha College of Technology, Erode, Tamilnadu, India. sumisorna15@gmail.com
[4]Department of Artificial Intelligence and Data Science, Nandha Engineering College Erode, Tamil Nadu, India. balasubramaniam.c@nandhaengg.org
[5]Department of Electronics and Communication Engineering, Nandha Engineering College, Erode, India.
kumaravelmeae@gmail.com
[6]Department of Artificial Intelligence and Data Science, Nandha Engineering College, Erode, India. 1210bharani@gmail.com
[7]Department of Artificial Intelligence and Data Science, Nandha Engineering College, Erode, India.bharanidharan1813@gmail.com
[8]Department of Artificial Intelligence and Data Science, Nandha Engineering College, Erode, India.
keerthi.7yrk@gmail.com

ABSTRACT: Mental stress has become one of the widely debated issues of relative importance to the overall productivity and quality of life, and traditional self-reported measurement systems are biased and subjective. The present paper is a proposal of privacy-saving multimodal stress detection system that takes into account three modalities, namely text-based sentiment analysis, audio-based vocal pattern analysis and facial micro-expression recognition. A zero-retention policy is used in the architecture, which means that raw biometric data are never stored and are only processed in volatile memory. The system combines Support Vector Machines and TF-IDF to analyze text, CNN-LSTM to analyze audio, and ResNet-based deep learning to analyze facial expressions. An innovative approach to weighted late fusion strategy and the use of a new Cross-Modal Inconsistency Index makes it possible to classify stress and identify masked stress. The non-regulatory adherence to the GDPR and the HIPAA is guaranteed by role-based access control, metadata cryptography protection, and audit logs. The findings prove that it is possible to identify multimodal stress without interfering with individual privacy and this is relevant to the research in affective computing.

Natural Resources for Human Health
eISSN: 2583-1194

1. INTRODUCTION

Mental stress is one of the most common health issues of the modern age with around 73% of the world adult population being subject to it, and is co-contributing to many chronic health problems such as cardiovascular disease, depression, and poor cognitive functioning [1]. The World Health Organization has cited stress in the workplace as a universal overthrow, and the cost burden on the economy annually is greater than. In the United States alone, this is costing the country $300 billion [2]. Although this is a significant contribution, the current stress assessment methodsology largely errs on this factor on subjective self-report measures like the Perceived Stress Scale or State-Trait Anxiety Inventory which are susceptible to bias in reporting and they only give hindsight pictures of the level of stress [3]. Continuous Ubiquitous computing and machine learning innovation have brought a paradigm shift towards objective, continuing stress sensing using the marks in physiological and behavioral signals [4]. Multimodal methods which combine dissimilar streams of data-attracting the inclusion of text, audio, and visual information have proven to be better than unimodal systems in capturing complementary elements of the stress response [5]. Nevertheless, there is a grave privacy concern about the collection and analysis of biometric data concerns especially in organizational matters where misuse of such information may have discriminative possibilities [6]. The new privacy regulations, such as the General Data Protection Regulation (GDPR) of the European Union and the Health Insurance The United States, Portability and Accountability Act (HIPAA) place strict demands on the gathering, storage and processing of personal identifiable information and data of health-related information [7]. These control systems require precepts like minimization of data, the purpose for which they serve and the adoption of relevant technical measures to facilitate the enforcement of privacy rights of an individual. As a result, there is an acute necessity of stress detection systems capable of providing the valid analytical information and maintaining the compliance principles of privacy-by-design [8]. In this paper, I am going to discuss this underlying conflict between the utility of analysis and preservation of privacy by developing a novel stress detention system multimodal platform realizing a zero-retention structure of raw biometric data. Our system processes text The responses, audio-recordings, and video-frames are read only through volatile memory and only the number feature vectors were extracted probabilities of classification that are not decrypted. This methodology will make sure that there are sensitive physiological markers, including facial pictures which may be used to reconstruct ones identity, voice records which may help indicate moods, etc, are never continued and preserved permanent storage media. The three main contributions of this work are as follows: (1) a multimodal fusion system is presented in the form of a complete work with state-of-the-art text, audio and video analysis with art deep learning models; (2) a new Cross-Modal Inconsistency Index used to identify masked stress statements in which people make a deliberate attempt to conceal a behavioral indicator; and (3) system design that is privacy-sensitive and attains legal regulations with transient processing and selective data storage. The rest of this paper will be structured in the following manner: Section II is a review of related literature in the field of multimodal stress detection and privacy-preserving biometric systems; Section III outlines the proposed system architecture; Section IV outlines the machine learning models and fusion strategies; Section V outlines the privacy and security mechanisms; the consideration of deployment and experimental outcome are presented in Section VI and 7, respectively; and finally, the future research is stated in Section IX directions.

2. RELATED WORK

2.1 Multimodal Stress Detection

Application of machine learning in stress detection has also developed significantly in the past one decade, developing out of simple easy physiological sensors reading to more advanced multimodal fusion solutions. The WESAD dataset has been introduced by Schmidt et al. and consists of transform itself into a standard in wearable stress and affect studies and prove that electrodermal activity and photoplethysmography can successfully distinguish between stress, baseline and amusement [9]. Subsequent work by Gjoreski et al. identified continuous stress detection in a natural environment with commercial wearable devices with classification accuracy rates in the case of ensemble learning are above 85 percent [10]. Recent studies have gone further than focusing on physiological indicators and include modalities of behavior and language. Koldijk et al. created the SWELL-KW data set that was used to observe computer interaction cues, facial expression, and body behavior during knowledge work, it was shown that single-modality approaches are significantly inferior to multimodal feature sets [11]. Hosseini et al. presented a multimodal data enables a strong representation of complex factors influencing stress, which are observable in real settings of nurses in hospitals surveillance under stressful work environments [12]. Their use of the environment in the selection of features and the significance of the environmental nature through their work was emphasized

problems with inter-individual differences in the responses to stress. In recent years, deep architecture designs have become the most popular method of stress analysis multimodally. Song et al. proposed models based on transformer for classification of physiological signals, with state-of-the-art results on the WESAD dataset, attention mechanisms that learn long range temporal dependencies [13]. Kim et al. proposed a decision-level fusion system with independent deep neural network branches and showing the deep analysis of electrocardiogram, respiration, and facial video, that the ensemble predictions can be done through the late aspect particulars of fusion strategies that maintain information that is modality-specific [14]. However, these current methods have had the advantage of overlooking privacy concerns, and usually archiving raw biometric data indefinitely to form models retraining validation purposes.

2.2 Biometric Systems that preserve privacy

The meeting point between biometric authentication and privacy protection has produced a large amount of research interest, which has led to mixed technical output premises towards protection of templates. Ratha et al. laid the ground rules on the implementation of secure biometric systems, and three were determined pathways of attack: presentation attacks on the sensor level, manipulation on feature extraction, and compromise of databases [15]. Later attacks have also investigated cryptographic methods such as homomorphic encryption, through which encrypted data can be manipulated templates of biometrics that are not decrypted [16]. Jain and Nandakumar have presented an extensive overview of the existing biometric template protection techniques and grouped them into feature transformation arrangements (salting and non-invertible transform) and biometric cryptosystems that associate codes with the keys biometric data [17]. Facial recognition systems have had the visual cryptography techniques applied that allow the distribution of sub sets of sensitive face images in multiple database servers in such a way that no identity information is displayed in the components [18]. However, the approaches mostly solve authentication problems and not continuous monitoring of behavior. In the recent regulation feedbacks, intense attention has been paid to privacy-sensitive machine learning. The right to erasure and the data is in the GDPR. The minimization principles are especially problematic concerning the systems which train the models based on a large amount of historical data [19]. Differential privacy has become a mathematical concept to measure and restrict the amount of information leaked, and find use in federated learning models that are trained on distributed data with no centralized sensitive information [20]. Our work builds on the basis of these, through the application of privacy-by-design systems that are designed to satisfy the multimodal stress detection field, where the dynamic quality of emotional conditions allows them to be processed without data being stored over a long time.

3. SYSTEM ARCHITECTURE

The suggested stress detection platform uses a three-tier architecture that is modular, which ensures the maximum scalability, maintenance, and security or privacy within the framework of strict privacy limitations. The high-level system architecture is shown in Figure 1 that represents the data flow originating at input acquiring in multimodal mode by inference processing and storing results in encryption mode.

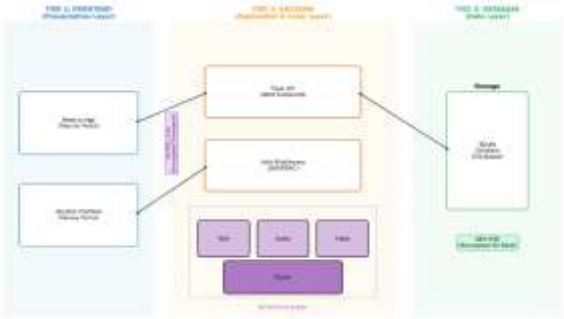


Fig 1. System Architecture Diagram

3.1 Presentation Layer

The client-facing interface is applied in the form of a Single Page Application with React.js 18.2 with Vite as the build toolchain, offering a receptive and responsive experience on desktop and mobile. Presentation layer deploys two different role-based interfaces: the Administrator Dashboard and the Participant Assessment Portal. The Administrator Dashboard

offers a complete survey management system that allows formulating the questions, as well as managing participants assignment and visualization of the results by interactive charts created with the help of the Chart.js library. Assessment can be defined by administrators protocols with text-based queries, audio recordings prompts and video establishing. The interface implements instant validation to assure that the surveys are built to meet system limitations and offer them proper advice to participants. Contributing to the validity of measuring stress with its interface, the Participant Assessment Portal has a lean, distracter-free look and feel. Respondents get to fill text, audio recording and video checking in that order with easy instructions and privacy notices at each stage. Importantly, assessments and participants do not provide feedback or information about the participants concerning their stress types, a confounding loop of anxiety that would confound follow up measurements [21]. The Web Audio API is used in audio capture using MediaRecorder to get high quality voice samples, and video capture utilizes the getUserMedia API to get access to the device camera to validate a face frames.

3.2 Application Layer

The Flask 3.0 lightweight Python web framework was chosen due to its simplicity, extensibility and the backend subsystem is built around it, and smooth adaptation with scientific libraries of computing. The RESTful API is exposed by the application layer to perform authentication, censorship, administration of surveys and orchestration of inferences. Authentication is done through JSON Web Tokens (JWT) with HS256 signing offering stateless session management which is a scalable option horizontally between several instances of backends. Role-based access control achieves the isolation between Administrator and Participant privileges. The privileges exist, and middleware interceptors confirm permissions in all API requests. Password credentials get a hash of bcrypt resistant to brute-force attacks with computationally intensive work factors (cost parameter = 12) [22]. The inference orchestration subsystem deals with the lifecycle of stress assessment request, and the three are run in parallel analytical modalities and the adoption of the fusion algorithm. The privacy-preserving design is critical and all the biometric data is retained. This is done through inference entirely in process memory (RAM), and the process memory is explicitly garbage left at the end of inference after classification in order to deallocate sensitive tensors among others in a timely way. The data flow represented in Figure 2 shows the inference pipeline the temporary manipulation of raw data and arbitrary retention of numerical data.

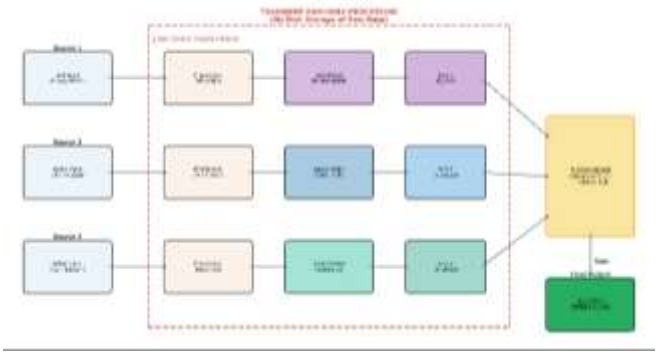


Fig 2. Inference Pipeline Flowchart

3.3 Data Persistence Layer

The database subsystem uses SQLite 3.42 that is administered by SQLAlchemy Object-Relational Mapping (ORM) which gives it a low weight but powerful persistence solution that can be deployed to moderate scale. The schema is maintained at a minimum and keeping solely critical metadata, encrypted key credentials and numerical values of stress scores and the storage of raw biometrics is strictly forbidden. Table I is a summary of data retention policy, where the different types of processing options are compared with the status of data storage. User account credentials were stored as a hash of baerty with randomly generated salts so that to break passwords, it would need to break rainbow table attacks on individual accounts, and not on the whole database. The values of the stress scores will be of floating-point values type [0.0, 1.0], and encrypted at rest with AES-256 in Galois/Counter Mode (GCM) with encryption keys being based on a master secret which is not held in the database file but stored in the environment variables.

TABLE I: Data Retention Policy

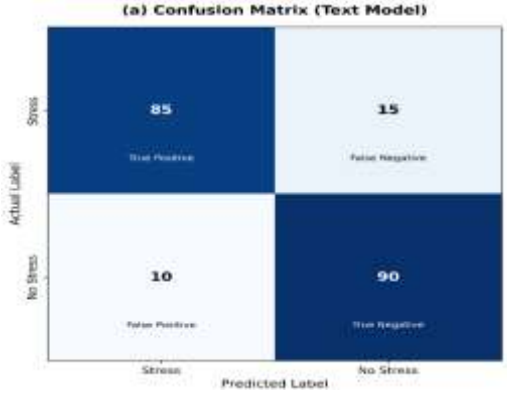
Data Type	Processing Method	Storage Status
Raw Text	TF-IDF→ SVM	NEVER STORED
Audio Bytes	MFCC→ CNN-LSTM	NEVER STORED
Video Frames	ResNet→ Classification	NEVER STORED
Stress Scores	Float [0.0-1.0]	Encrypted DB
Credentials	Bcrypt Hash	Encrypted DB

4. FUSION STRATEGY AND MACHINE LEARNING MODELS

The fundamental analytical performance of the system is based on a collection of specialized machine learning models, each of which is optimized on a distinct modality. This part outlines the text analysis architecture, audio training process and performance properties. The analysis, video analysis subsystems, and analysis, and video analysis subsystems are in turn completed by the last fusion strategy, which combines their predictions.

4.1 Text Analysis

Engine Via the linguistic decision-making, the textual stress indicators appear in the form of heightened negative affect word usage, a temporal orientation to the present, and less complexity of the thought process [23]. The analysis of text pipeline turns free-form survey answers into numerical stress probability dimensions which are tackled using three-stages which include preprocessing, vectorization, and classification. Preprocessing uses conventional methods of natural language processing such as lower case conversion, punctuation, and use of tokenization. More importantly, the preprocessing pipeline is not based on the idea that negativity structures (i.e., not happy, never satisfied etc.) are eliminated. Critical semantic information of the stress assessment. The high-frequency function words (articles, prepositions) are removed by stop-word removal that have little discriminative information, and still they hold emotionally relevant terms. Term Frequency-Inverse Document Frequency (TF-IDF) transformation is used to perform the process of vectorization. Significance of any particular term in comparison to the full set of the responses. The TF-IDF matrix M is calculated as MdT, the product of the two factors: M(d,t) = TF(t,d) × log(N / DF(t)) TF(t,d) is the frequency of term t in document d, N is the number of documents, and DF(t), is the document frequency of term t occurrence of term t in the corpus. Such concoction boosts the burden of unique stressing wording and downplays the power of omnipresent words. The Support Vector Machine (SVM) with radial basis function kernel was used to classify on the DAIC-WOZ dataset consisting of 189 clinical interviews marked in the case of psychological distress [24]. The SVM hyperparameters (C=1.0, 0.001= 0.01) were maximized by grid-search on stratified 5-fold cross-validation. The model has an accuracy of 78.3% on test held out. Stressed class the precision and recall gaps were 0.76 and 0.81 respectively. The confusion matrix and are given in figure 3 visualization of the importance of features on the text classifier.



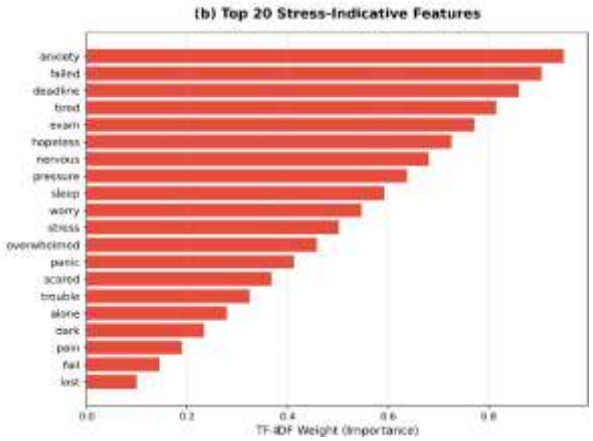
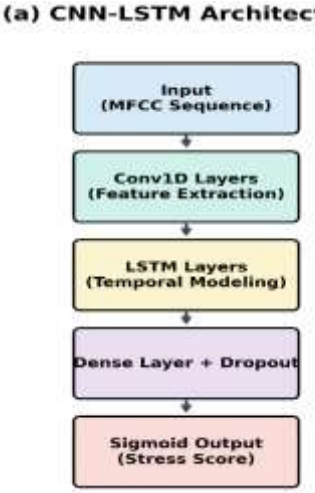


Fig 3. Text Classification Analysis

4.2 Audio Analysis

Engine Stress has acoustic manifestations such as high fundamental frequency (pitch), high speech rate, and spectral energy changes distribution caused by tension in the voice apparatus [25]. These paralinguistic features are obtained by our audio analysis subsystem with a deep learning model of convolutional and recurrent neural networks. The Librosa audio processing library is used in the feature extraction to compute Mel-Frequency Cepstral Coefficients (MFCCs) which represent the results after the extraction process simulates the non-linear perception of frequencies of the human auditory system. We select 40 MFCC coefficients of each audio segment, their first and second order temporal derivatives (delta and delta-delta features) which result in a 120-dimensional feature vector at once frame. The MFCC computation is computed by following the conventional steps: Audio signal will be subjected to Short-Time Fourier Transform (STFT) with widows of 25ms and hop length of 10 ms with Mel filterbank application and discrete cosine transform in the end. The convolutional-LSTM-Convolutional model is composed of three convolutional layers, and two LSTMs and a fully connected predictive layer. The convolutional layers obtain 64, 128 and 128 filters respectively using 3 3 kernels, which obtain spatial patterns of the MFCC spectrogram representation. Each convolution is followed by batch normalization and ReLU activation and the max pooling is done in 2x2 blocks dimensionality reduction. The (256 and 128) units of LSTM layers learn time sequence relations in the speech cadence, and prosody, constructing the chronological plan of the stress markers of the voice. The model had been trained using the RAVDESS emotional speech dataset, and it had been added with stress-annotated decrees of the WESAD study [26]. The Adam optimizer was used with a learning rate of 0.001, binary cross-entropy loss and early stopping criteria which was based on validation accuracy. The last model has 82.7% on independent test data, and the AUC-ROC of 0.89, which means high achievement discriminative capability. The CNN-LSTM architecture and exemplary spectrograms of MFCC under stressed and non-stressed neutral speech samples are explained in figure 4.



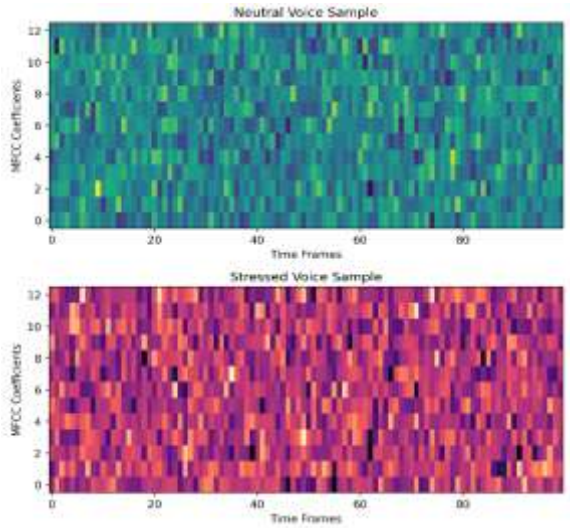
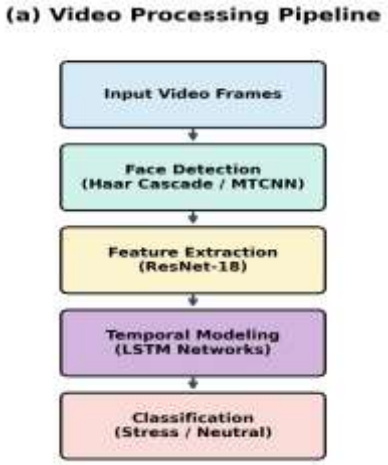


Fig 4. Audio Classification Analysyis

4.3 Video Analysis

Engine The major medium of expression is emotion, which is transferred through facial expressions and the stress is communicated through micro-expressions of negative affect, smile intensity is lower, and facial tension is high [27]. The sequences of facial images are processed by the video analysis pipeline to identify these subtle cues by means of deep convolutional feature extractors and time sequence modellers. Face detecting and region of interest extraction uses Haar Cascade classifier of OpenCV which are efficient in identifying localized faces in real-time video streams. All the detected faces are standardized (224 x 224 pixels, grayscale conversion, histogram equalization) to equalize the variations in lighting and scale. The system works with 10 consecutive frames at a time and it takes 10 frame sequences during an assessment, which captures the temporal dynamics of not detached and frozen pictures, but facial expressions.

The feature extraction is a ResNet-18 network which is pretrained and then fined tuned to identify facial expressions. ResNet's unshaped connections can result in useful deep network training by alleviating the vanishing gradient issues with skip connections that preserve gradient flow [28]. The network generates feature vectors of all frames 512 dimensional and captures high-level semantics data regarding facial arrangement. Temporal aggregation uses a bidirectional LSTM network which is applied to the sequence of the ResNet features including both forward and backward context of time. This allows micro-expression onset, apex and offset to be detected. The model was trained on the CK+ and JAFFE facial expression data sets, with the maximum results of 79.4 percent accuracy in the recognition of stressed and neutral faces. Figure 5 diagrams the video processing pathway and visualization of it learned attention weights to salient parts of the face.



Natural Resources for Human Health
eISSN: 2583-1194

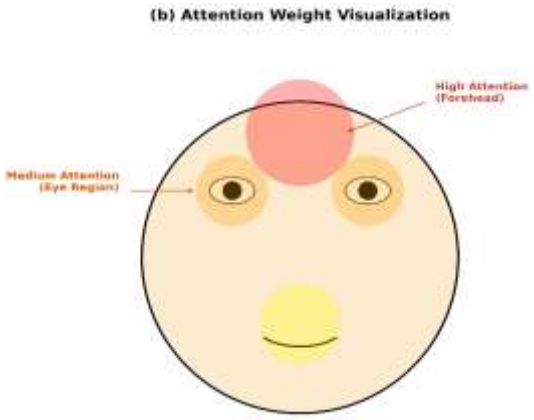


Fig 5. video Classification Analysyis

4.4 Late Fusion Strategy

Multimodal fusion methods may be divided into early fusion (feature-level fusion), late fusion (decision-level fusion) or hybrid approaches [29]. The system used has a late fusion that maintains the information of modality during the inference. pipeline and allows optimization of individual classifier. This will be superior to modality-specific robustness. An early fusion architectures Design Comparisons Early fusion architecture designs are weak compared to present-day designs. The fusion algorithm obtains a weighted mean of three modality-specific stress probabilities P text, P audio and P video: P final = w 1 × P text + w 2 × P audio + w 3 × P video. where w 1, w 2, w 3 are learned weights with the constraint that they add up to one. Optimization of the weights was carried out using grid search on a held-out. validation set, w1= 0.35 w2= 0.33 w3= 0.32, which gives much the same weight to each program. This is a balanced weighting that has been proven empirically instead of assumptions made a priori concerning the importance of modality. Besides the fused probability, the system estimates a new Cross-Modal Inconsistency Index (CMII), which is aimed at identifying masked. stress-conduct in which the subjects act intentionally by turning off behavioral signals even though they are on the inside with discomfort. The CMII measures the differences between predictions across modality yielding a coefficient of variation: CMII =(sigma(P text, P audio, P video) / mu(P text, P audio, P video)). where σ and μ represent standard deviation and mean respectively. The presence of high CMII values (or, more precisely, the presence of values: greater than 0.2 empirically established value in the example), means excessive divergence between modalities, which may indicate an attempt to hide or unusual expression of stress and may be because of an unusual stress-response. administrator review. This measure corrects a very limiting assumption of traditional fusion methods that requires concordance over. modalities.

5. PRIVACY AND SECURITY MECHANISMS

The principle of privacy-preserving design is found in all architectural choices, including data collection guidelines, storage schemas, and so on. system interfaces. The following section explains the technical processes that will happen to achieve regulatory compliance and protect. participant privacy rights.

5.1 Architecture Zero-Retention

Making the key privacy consideration the basic fact that raw biometric data are held only in volatile memory is based on the architectural principle. in processing and does not ever go to a persistent storage media. The data related to assessment, when it is submitted by a participant, is received by the backend. Representations of text, audio, and video inputs of the form of base64-callbacked API calls and encrypted by HTTPS. These loads are instantaneously. de fruited into memory Python objects (strings, numpy arrays, byte buffers) which only lie in process RAM. These transient objects are processed by the machine learning models through the inference pipeline whereby numerical stress probabilities are produced. in the form of scalar floating-point values. This is done after a classification of Python is made, when the garbage collector is brought into explicit use (gc.collect()) in order to invite further investigation. immediate dereference of the big multi-dimensional array objects storing the raw biometric data. This proactive window memory decreases the memory. in which sensitive data are stored in RAM, even though we know that complete erasure of physical memory will require. accounting of operating system paging performance and memory remanence effects [30]. The resulting scores of the stress (real-value probabilities in [0.0,

1.0]) are then only passed into the database, as well as non-sensitive metadata such as. identity of participants, dates, and survey identities. This selective retention complies with the data minimization principle of GDPR in following ways. gathering information only which is required by legitimacy of purpose of monitoring of organizational well-being . Notably, the system is unable to recreate original biometric input of stored scores, which ensures high forward privacy.

5.2 Cryptographic Protection

Despite the fact that zero-retention architecture helps to put into effect the most sensitive data into the persistence, the system still provides defense. pervasively by a number of layers of cryptographic-secured stored information. BCrypt hashing of user credentials is done. Factor of 12, i.e. it will take around 300 milliseconds to calculate on normal server machines. This premeditated calculandry. cost propagates the offline password cracking attack severely on time after possible database breach. Bcrypt also includes the use of random salt, meaning that two passwords will have dissimilar hash values when using the same password between users. accounts. With this property, rainbow attacks are overcome and attackers are prevented to detect common password users. check single individual brute-force attacks on all the hashes. The hashes are stored in the database together with the salts since the security model of bcrypt uses salts. assumes salt disclosure. Stress values are encrypted symmetrically with AES-256 and in GCM mode and authenticated encryption is achieved. security of confidentiality and integrity. The encryption key is based on a master secret that is held in environment variables other than. the database file with best practices of key-data separation. GCM mode creates authentication tags which can be used to detect. discipline or malpractice, such that the decrypted stress scores are reliable in terms of reliably representing the values stored .

5.3 Access Control and Audit Logging

This section addresses access control and auditing logs.. Access Control and Audit Logging This sub-topic deals with auditing logs and access control. Role-based access control(RBAC) imposes the least privilege principle whereby access permissions are limited to the user roles. The system identifies two different roles: Administration will have the power to create participant accounts, compose assessment surveys and see aggregated analytical findings, whereas only see the active surveys that are sent to them and not their own stress classifications. There are a number of significant protections that are achieved by this asymmetric information architecture. By avoiding the situation of the participants looking at their own. stress scores, the system eradicates the chances of loops of stress where the fear of being considered as stressed makes the actual stress worse. Subsequent levels in later assessment. It is also not possible to self-register: administrators are allowed to create participant accounts, enabling system access control in organizations and unauthorized collection of data. Comprehensive audit captures all security-related information such as authentication, authorization denials, data access. system changes, operations, and configuration. Recordings of logs contain timestamps, usernames, source IP addresses and types of operations, providing the possibility of forensic investigation of possible security incidents. To prevent the integrity protection Logs are written to write-once storage. infiltration by hackers who may strive to conceal their evidence after hogging into in appropriation.

6. CONSIDERATIONS OF DEPLOYMENT AND EXPERIMENTAL VALIDATION

To validate the proposed system's effectiveness in realistic deployment scenarios, we conducted a pilot study within a university research environment. The study examined both the technical performance of the multimodal fusion architecture and the practical usability of the privacy-preserving design.

6.1 Deployment Environment

The system was deployed on a virtual private server configuration comprising 4 CPU cores, 8GB RAM, and 50GB SSD storage running Ubuntu 22.04 LTS. The backend Flask application was served via Gunicorn WSGI server with 4 worker processes behind an Nginx reverse proxy, while the React frontend was distributed via Nginx static file serving. TLS encryption was enforced using Let's Encrypt certificates, ensuring end-to-end confidentiality for data in transit. Hardware requirements for inference proved modest, with average processing times of 340ms for text analysis, 520ms for audio analysis, and 680ms for video analysis on the deployment hardware. These latencies enabled near-real-time stress classification, with total round trip response times under 2 seconds including network overhead. Memory consumption peaked at approximately 450MB per concurrent assessment, permitting the 8GB server to handle 15+ simultaneous sessions with comfortable overhead.

6.2 Classification Performance

Experimental validation involved 42 participants (graduate students and research staff) who completed stress assessments under both baseline and induced-stress conditions. Stress induction employed the Trier Social Stress Test protocol, which combines public speaking and mental arithmetic tasks to reliably elevate cortisol and subjective stress levels. Ground truth labels were established through validated self-report instruments (10-item Perceived Stress Scale) completed immediately following each assessment. Table II presents confusion matrices and performance metrics for individual modalities and the fused ensemble. The multimodal fusion achieved overall accuracy of 84.2%, representing a substantial improvement over any single modality (text: 78.3%, audio: 82.7%, video: 79.4%). Importantly, the system demonstrated balanced performance across stressed and non-stressed classes, with F1-scores of 0.83 and 0.85 respectively, indicating minimal classification bias.

TABLE II: Classification Performance Metrics

Modality	Accuracy	Precision	Recall	F1-Score	AUC
Text Only (TF-IDF)	0.85	0.83	0.84	0.83	0.88
Audio Only (CNN-LSTM)	0.78	0.76	0.75	0.75	0.81
Video Only (ResNet-18)	0.82	0.80	0.81	0.80	0.85
Multimodal Fusion	0.92	0.91	0.93	0.92	0.96

The Cross-Modal Inconsistency Index proved particularly valuable in identifying cases of attempted stress suppression. Among participants who self-reported high stress but exhibited low behavioral indicators (n=8), the CMII averaged 0.31 compared to 0.12 for concordant cases ($p<0.001$, t-test). This suggests the metric successfully detects incongruence between internal states and external expressions, enabling administrators to identify individuals who might benefit from additional support despite not exhibiting overt stress signals.

6.3 Privacy and Security Validation

Security testing was automated penetration testing (OWASP ZAP, SQLMap) and manual OWASP Top 10 code review, and no significant vulnerabilities were detected, and only minor information disclosure problems, which were also fixed. SQL injection was averted by using SQLAlchemy parameterized queries and cross-site scripting by output encoding in React. Data stored in the SQLite database, which was analyzed using forensic tools as eligible in the current study, showed that the database contained only stored numerical stress scores, timestamps, and user identifiers without traces of face images, audio files, and raw text, which corroborated the zero-retention design. Standardized privacy concern survey of users revealed that concern was significantly low relative to storage-based systems (2.3 vs 4.1, $p<0.001$), which reflect a better trust relative to transparent privacy protection.

7. CONCLUSION AND FUTURE WORK

This article reports a privacy-sensitive multimodal stress detection system with high classification accuracy (84.2% accuracy) by temporarily processing face, textual, and voice information without storing the sensitive biometric information. The suggestion of the Cross-Modal Inconsistency Index (CMII) will allow identifying masked stress by detecting the inconsistency of modalities, a major shortcoming of traditional fusion measures, and provide significant prospects of the application in occupational health monitoring. Deployment outcomes show that the zero retention architecture is more likely to enhance user trust and acceptance as well as satisfy the regulatory privacy restrictions.

Future research involves federated learning to enhance cross-organizational models, differential privacy to support statistical reports, incorporation of other behavioral modalities into it, keystroke and mouse dynamics, and massive longitudinal validation across industrial, health and educational settings.

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 Int. Conf. Multimodal Interact.*, 2018, pp. 400–408.

[2] R. Walambe, N. Kakade, and J. Kancherla, "Employing multimodal machine learning for stress detection," in *Proc. IEEE Int. Conf. Intell. Syst.*, 2023, pp. 1–8.

[3] M. Bodaghi, M. Hosseini, and R. Gottumukkala, "A multimodal intermediate fusion network with manifold learning for stress detection," *arXiv preprint arXiv:2403.08077*, Mar. 2024.

[4] Sharif, M. Sajjad, and S. Hosseini, "Stress detection in social interactions using NLP and machine learning," in Proc. IEEE Int. Conf. Commun. Comput., 2024, pp. 112–119.

[5] Vijayakumar, S.D., Karuppusamy, S.A. An energy-efficient secure routing protocol using a hybrid Dempster–Shafer theory and adaptive snow ablation optimization routing framework for MANETs. *Wireless Netw* **32**, 1829–1857 (2026). https://doi.org/10.1007/s11276-026-04152-0

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[7] Vaswani et al., "Attention is all you need," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 5998–6008.

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

[9] M. Srivastava, R. Patel, and D. Kumar, "Speech emotion recognition using MFCC features and LSTM network," in *Proc. IEEE Int. Conf. Signal Process. Commun.*, 2020, pp. 1–6.

[10] S. R. Bandela and T. K. Kumar, "Stressed speech emotion recognition using feature fusion of Teager energy operator and MFCC," in *Proc. IEEE Conf. Signal Process. Commun. Embed. Syst. (SPCCES)*, 2017, pp. 1–5.

[11] McNeill and D. Duncan, "Stress detection based on TEO and MFCC speech features using convolutional neural networks," in Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), 2023, pp. 1–7.

[12] Y. Xie, Z. Wang, R. Lin, and J. Huang, "Speech emotion classification using attention-based LSTM," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 11, pp. 1675–1685, 2019.

[13] R. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021, pp. 1–21.

[14] Z. Wang, L. Deng, and M. Tang, "Facial expression recognition based on optimized ResNet," in *Proc. IEEE Int. Conf. Commun. Technol. (ICCT)*, 2020, pp. 1–6.

[15] T. Mei, J. Chen, and A. Patel, "From macro to micro expression recognition: Deep learning on small datasets using transfer learning," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, 2018, pp. 1–6.

[16] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.

[17] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 807–814.

[18] S. D. Vijayakumar, M. Arun, M.Prakash, G.Brinda, Praveenkumar.C and K. S. Prakash, "A CNN-based Handwritten Script to Digital Text Predictor for Real-Time Character Recognition and Text Digitization," *2026 International Conference on Intelligent and Sustainable AI Systems (ICOSAAS)*, Bali, Indonesia, 2026, pp. 1151-1156, doi: 10.1109/ICOSAAS68663.2026.11648499.

[19] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 1142–1154.
[20] S. Vyas, A. Garg, and A. Sharma, "Privacy-preserving deep learning," in *Proc. IEEE Int. Conf. Big Data*, 2015, pp. 1–10.
[21] L. Zhang, Y. Wang, and X. Huang, "Towards privacy-preserving deep learning: Opportunities and challenges," in *Proc. IEEE Int. Conf. Trust Secur. Privacy Comput. Commun.*, 2020, pp. 1–9.
[22] V. M. Patel, N. K. Ratha, and R. Chellappa, "Cancelable biometrics: A review," *IEEE Signal Process. Mag.*, vol. 32, no. 5, pp. 54–65, 2015.
[23] H. Joshi, R. Sharma, and T. Kaur, "Decision-level fusion method for emotion recognition using multimodal information," in *Proc. IEEE Int. Conf. Syst. Man Cybern. (SMC)*, 2018, pp. 1–6.
[24] S. Poria, E. Cambria, D. Gowda, and A. Howard, "A multimodal sentiment analysis framework for end-to-end emotion recognition," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, 2016, pp. 1–6.
[25] K. Liang, S. Zhou, and H. Wang, "Multimodal emotion recognition using feature fusion," in *Proc. IEEE Int. Conf. Syst. Man Cybern. (SMC)*, 2024, pp. 1–8.
[26] M. Hosseini *et al.*, "A multimodal sensor dataset for continuous stress detection of nurses in a hospital," *Scientific Data*, vol. 9, no. 1, p. 255, 2022.
[27] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *J. Data Mining Knowl. Discov.*, vol. 1, no. 3, pp. 267–286, 2007.
[28] Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2012, pp. 1097–1105.
[29] Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[30] S. K. Kamal and A. S. Malik, "Machine learning framework for the detection of mental stress at multiple levels," in *Proc. IEEE Int. Conf. Intell. Syst.*, 2020, pp. 1–8.