International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(16s) 2026 pp. 1049-1060
https://doi.org/10.70917/ijcisim-2026-4630
Article

Scalable Non-Contact Stress Detection Using Hybrid Multimodal Intelligence

Anees Fatima[1], Mohd Nazeer[2]
[1]Computer Science and Engineering, BESTIU, Anantapur, Andhra Pradesh – 515231, India
Email:aneesf124@gmail.com
[2]Vidya Jyoti Institute of Technology, Hyderabad, Telangana-500008, India
Email: nazeer8584@gmail.com

Abstract: Scalable digital health and human-computer interaction systems require non-contact stress detection methods, where wearable sensors and self-reports are impractical. This paper introduces a new reliability-aware hybrid multimodal stress detection scheme for stress detection based on speech characteristics and multimodal cues such as facial cues, eye/pupil features, keyboard dynamics, and handwriting behavior. The model includes encoders specialized for each modality, temporal learning and adaptive reliability-aware fusion to mitigate noisy or missing modalities. TExperiments conducted in a multimodal setting derived from ForDigitStress achieved 94.30% accuracy, 94.00% precision, 93.20% recall, 93.60% F1-score, 0.967 AUROC, 0.036 calibration error and 38.40 ms inference latency per window. Results show strong, interpretable and scalable stress detection in real-world digital environments.

Keywords: Stress Detection, Hybrid Multimodal Intelligence, Non-Contact Monitoring, Behavioral Biometrics, Machine Learning, Affective Computing

1. Introduction

Stress detection is one of the key research topics in the field of affective computing, digital health, workplace monitoring, online education, gaming, and human-computer interaction. Traditional methods of stress assessment have been mostly based on self-reported questionnaires and/or physiological measurements that can be obtained from the body through contact (ECG, PPG, EDA, HRV, and cortisol). While these signals can be useful indicators of psychological and physiological stress, they often require recording using wearable sensors, in controlled settings, with the cooperation of the subject, and for extended periods. These restrictions limit their applicability to scalable, continuous stress monitoring in real-world digitally enabled environments where users will interact naturally with computers, microphones, cameras, and writing devices.

Non-contact and behavioral modalities like speech, facial activity, eye movement, typing behavior and interaction patterns have been recently used to detect stress. Multimodal stress datasets show the enhancement of the automatic stress recognition performance using combined audio, video, facial landmark, eye tracking and physiological signals compared to single-source assessment. Likewise, stress research using gaze has revealed that subtle gaze variations can indicate non-contact stress measurement over a short time. Other studies related to the keyboard suggest that typing rhythm, key duration, pressure variation and input behavior are indicators of cognitive, and motor changes related to stress. These findings suggest that the everyday behavioral signals are good practical alternatives for intrusive physiological signals.

Current systems have several critical issues. Single-modality stress detectors are affected by noise, user variability, task differences, and the environment.

Background noise can affect the features of speech, lighting and occlusion can affect facial cues, keyboard patterns can differ based on typing ability and handwriting behavior can differ based on the device and writing surface. Feature concatenation is also not able to consider the reliability of modalities. Based on this, this paper presents a non-contact stress detection framework which is scalable and can combine typing dynamics of the keyboard, voice information, facial features and handwriting features through hybrid multimodal intelligence. The goal is to improve

the interpretability and practical feasibility of stress and non-stress classification, while combining complementary behavioral evidence.

A. Contributions:

This paper presents a multimodal stress detection model, which combines the typing pattern information from the keyboard, the speech information from the microphone, the facial information from the camera and the handwriting information from digital pen/tablet input and encodes them using modality-specific computational feature encoders.

This paper presents a multimodal fusion scheme that considers the reliability of the various behavioral modalities, while mitigating noisy, missing, and inconsistent modalities.

This paper compares the proposed framework to the machine-learning (ML), deep-learning (DL), and unimodal baseline methods under six evaluation criteria: accuracy, F1-score, AUROC, latency, calibration, and modality-ablation analysis.

2. Literature Review

Stress detection is a well-established research topic in the field of affective computing because stress impacts cognitive performance, emotional stability, physiological control and user behaviour when interacting with computer systems [1]. Recent multimodal datasets have enhanced automatic stress recognition, by integrating behavioral, physiological, and contextual information. The ForDigitStress dataset was created to provide a digital job interview scenario, and to capture the audio, video, motion capture, facial landmarks, eye tracking, PPG and EDA of participants during stressful interactions [2]. The study also employed the time-continuous annotations, as well as compared different classifiers, with LSTM having good baseline performance for binary stress classification [3]. This suggests that temporal learning is crucial when the stress patterns are evolving across interaction sequences as opposed to being static features of the interaction [4].

Eye-based and gaze-based stress detection has also received attention as it is possible to obtain eye-based features with minimal physical contact [5]. However, survey-based

evidence suggests that daily stress detection should be conducted outside of the laboratory and utilize data collected from keyboard, mouse, game controller, touch patterns and performance behavior [6].

Keyboard-based stress detection is commonly associated with keystroke dynamics, or keystroke dynamics, which involves deriving timing and rhythm characteristics of people's normal typing [7]. Acoustic and prosodic features have been explored for speech-based stress recognition due to the change in the vocal production caused by stress [8]. The appearance of stress is detected using changes in facial expressions, eye movement, head movement and landmark displacement. Stress-related affective behavior has been represented by facial action units, eyebrow movement, eye closure, mouth openness and head movement [9].

Another behavioral channel is handwriting stress analysis, as the level of stress affects the fine motor control [10]. Many machine-learning techniques, such as the Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbor (KNN), and XGBoost, are effective for stress classification and have been widely applied on handcrafted behavioural and physiological features [11]. These methods are interpretable and computationally efficient, but they may be based on static feature vectors, and they do not necessarily account for the evolution in stress over windows of interaction [12][19].

Recently, using transformer-based models and attention mechanisms has been attractive for multimodal recognition problems, since they are able to capture long-range dependencies and dynamically prioritize the input modalities [13][18]. Fusion strategy is one of the main challenges in multimodal stress detection. Early fusion is the fusion of all the modality features prior to classification, and late fusion is the fusion of decisions of the individual classifiers [17][14]. The hybrid fusion is more powerful since it allows to retain the modality-specific information prior to merging at a learned representation or decision level [15][16][20].

A. Research Gaps

There have been significant advances in the physiological sensing, eye tracking, audio analysis, facial cue recognition, and interaction-based behavioral modeling in existing studies related to stress detection; however, there are still several gaps. Many existing systems still rely on contact-based sensors, laboratory settings or single-modality data which is not scalable in real-world digital environments. There is a lack of datasets that integrate audio, video, eye tracking, physiological signals, keyboard dynamics, and handwriting features in a non-contact computational

framework. Single modality methods are still vulnerable to background noise, lighting changes, user factors, device differences and task dependency. Current fusion methods are also generally based on direct feature concatenation without considering missing modalities, noisy or unreliable modalities. In this context, a scalable hybrid multimodal intelligence framework is needed, which can process both the pattern of keyboard-based input and speech characteristics, facial features and handwriting features, each of which is encoded separately, learn from the temporal input, fuse the input in a way that accounts for input reliability, and compare the output against a traditional and deep-learning baseline.

3. Proposed Model

Reliability-Aware Hybrid Multimodal Stress Detection Framework (RA-HMSD) shown in figure 1 aims to develop a scalable, non-contact stress detection system, which combines the dynamics of the keyboard typing, speech characteristics, facial cues and handwriting features. It processes the multimodal features including modality-specific behavioral descriptors like typing latency, dwell time, flight time, pauses, speech pitch, jitter, shimmer, spectral energy, facial landmarks, blink behavior, pen pressure, stroke velocity, curvature, pen-up/pen-down timing etc. and maintains the individual structure of these behaviors by using dedicated feature encoders. Temporal modeling is used to model stress related changes over the interaction windows, and a reliability-aware attention fusion layer is used to adaptively weight the different modalities depending on the quality and availability of the modalities, as well as the discriminative contribution, the fused representation is classified into stress or non-stress state. RA-HMSD is novel because it adaptively fuses complementary non-contact behavioral modalities instead of relying on physiological sensors, single-modality cues, or simple feature concatenation. of non-contact behavior rather than relying on physiological sensors or single modality cues or simple concatenation as is done with many other systems. This design improves the scalability, interpretability and robustness of the design through three phases: modality specific feature learning, temporal behavioral models, and multimodal decision fusion via adaptation.

A. Mathematical Model of the Proposed RA-HMSD

The multimodal input of a subject over a time window can be symbolized as:

$X = \{Xk, Xs, Xf, Xh\}$ (1)

In equation 1 feature matrices Xk, Xs, Xf, and Xh are obtained from the keyboard typing, speech, facial cues and handwriting, respectively, in equation 1. The behavioral features in each matrix are time-windowed and are represented by rows corresponding to temporal segments and columns corresponding to modality-specific features. A modality-specific encoder maps the input feature matrix into a compact latent representation of the input for each modality

$m \in \{k, s, f, h\}$, as:

$$Zm = \phi m\,(Xm; \theta m) \quad (2)$$

In equation 2, the encoded feature representation for the modality m is denoted by Zm, the modality specific encoder is denoted by ϕm and the learnable parameters of the encoder are denoted by θm in equation 2. Describes the process of encoding in a modality-specific way. Each encoded representation is fed into a temporal learning function to record temporal evolution of stress:

$$Hm = \psi m\,(Zm; \omega m) \quad (3)$$

In equation 3, Hm denotes the temporal representation of modality m, ψm is the temporal learning unit (e.g., LSTM, GRU or temporal convolution) and ωm are the learnable parameters of the temporal learning unit. Stress is not just manifested in the value of a feature. The reliability score of each modality is calculated as:

$$rm = \sigma\,(WrHm + br) \quad (4)$$

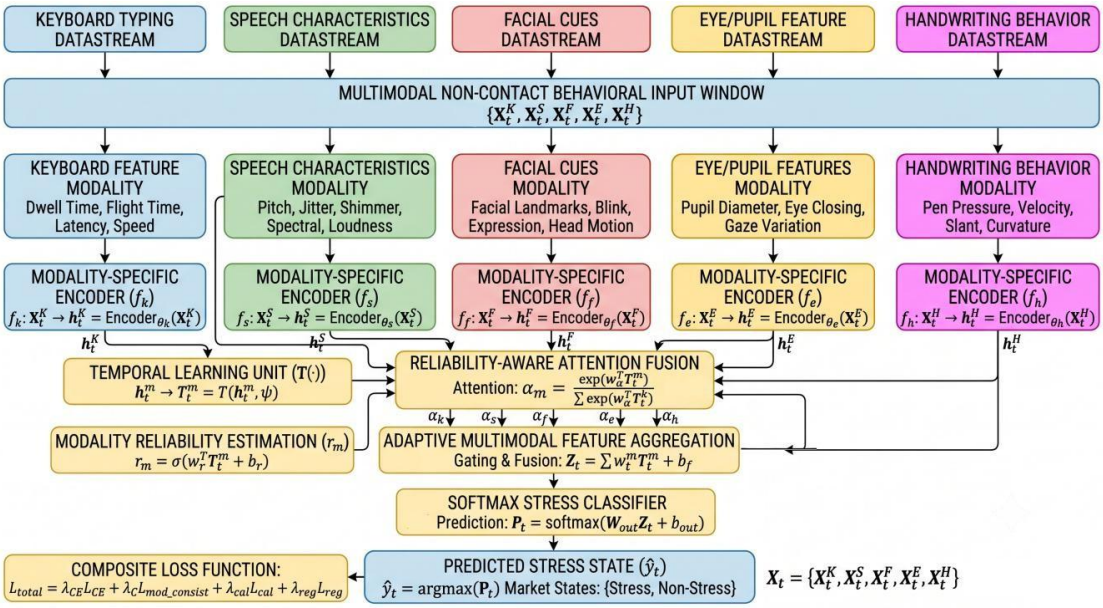

Figure 1: proposed Hybrid RA-HMSD model diagram

In equation 4 the reliability score of modalities m is denoted by rm and the learnable reliability parameters are denoted by m, Wr and $br_$are and in equation 4, the sigmoid activation function maps the score to a range between 0 and 1. This is one of the main novelties of the proposed model as real-life non-contact signals are often missing, incomplete, or noisy. Background noise can distort speech, difficult lighting can make facial expressions difficult to see, the user pauses during tasks, and keyboard input is lost, and handwriting may not be available in some tasks. The attention weight of each modality is computed as:

$$\alpha_m = \frac{\exp(rmWaHm)}{\sum \exp(r\ W\ H)} \qquad (5)$$

In equation 5, αm represents the attention weight assigned to modality m , Wa is the attention weight matrix, and j indexes all available modalities. Unlike standard attention mechanisms that assign importance only from feature representation, this equation combines representation strength with modality reliability. The final fused multimodal representation is obtained as:

$$F = \sum m \in \{k,s,f,h\}\ \alpha m\ Hm \qquad (6)$$

Equation (6) defines the final reliability-weighted fused representation F and it is reliability weighted temporal. Adaptive behavioral fusion lies at the heart of the novelty of this equation. The final stress prediction is given by:

$$\hat{y} = \text{Softmax}(Wc\ F + bc) \qquad (7)$$

The equation 7 contains $\hat{y}$ which is the predicted probability distribution, Wc is the classifier weight matrix and

bc is the classifier bias. The multimodal representation is

fused forming a new representation, which is mapped to class probabilities, and the system predicts whether the window of input represents a stress or non-stress state. The overall training goal is:

$$\mathscr{L}\text{total} = \mathscr{L}\text{CE} + \lambda 1 \mathscr{L}\text{MC} + \lambda 2 \mathscr{L}\text{CAL} + \lambda 3 \parallel \Theta \parallel_2^2 \qquad (8)$$

The cross-entropy classification loss, $\mathscr{L}$CE, in equation 8 is designed to ensure that the classification term aligns with the classification problem that is being solved. The modality consistency loss, $\mathscr{L}$MC , ensures that the modality term is consistent across the different modalities. The calibration loss, $\mathscr{L}$CAL, ensures that the calibration term matches the specific problem being solved. The regularization term, $\parallel \Theta \parallel^2$, is designed to ensure that the algorithm

remains stable. The weighting coefficient, $\lambda 1$, $\lambda 2$, and $\lambda 3$, are used to balance the contribution of each part of the algorithm. The cross-entropy loss improves classification accuracy of the classification, the modality-consistency loss will help to agree with all complementary modalities, the calibration loss will help to increase the reliability of the probabilities in case of class imbalance, and the regularization loss will help to avoid overfitting.

B. Proposed Algorithm

Algorithm: Reliability-Aware Hybrid Multimodal Stress Detection Algorithm

Input: Keyboard feature stream Xk, Speech feature stream Xs, Facial feature stream Xf , Handwriting feature stream Xh , Trained model parameters Θ

Output: Predicted class label y, Stress probability score

$\hat{y}$, Modality attention weights αm

Steps:

1. Obtain multimodal non-contact behavioural signals and form the multimodal input set X =

   $\{Xk, Xs, Xf, Xh\}$

2. Segment each modality into fixed temporal windows and rescale the values of each feature

3. Extract features from the keyboard, speech, facial and handwriting streams that are modality specific

4. Encode each modality using a separate encoder

   $Zm = \phi m(Xm; \theta m)$

5. Learn temporal stress patterns, with a sequence model $Hm = \psi m(Zm; \omega m)$

6. Produce an estimate of the reliability score of the modality $rm = \sigma(WrHm + br)$

7. Compute reliability aware attention weight αm =

   $$\frac{\exp(rmWaHm)}{\sum j \exp(rjWaHj)}$$

8. Compute the sum of the multimodal representations

   $F = \sum m\ \alpha m\ Hm$

9. Predict the stress or non-stress class $\hat{y}$ =

   Softmax($WcF + bc$), y = arg max ($\hat{y}$)

C. Novelty of the Proposed Model

It is novel that all the following four stages: behavioral feature extraction, modality-specific encoding, temporal learning, reliability estimation and adaptive fusion are all performed within a single stress-detection procedure. This algorithm adapts the role of each modality based on the quality and contribution of each modality, instead of conventional algorithms which classify a fixed handcrafted feature vector. This will be particularly helpful in real-world non-contact environments where behavioral signals are not necessarily equally accessible and/or reliable. The algorithm also provides the interpretability of the final decision by generating modality attention weights.

D. Baseline Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUROC
SVM	84.6	83.9	82.7	83.3	0.872
Random Forest	86.4	85.8	84.2	85	0.891
XGBoost	88.9	88.2	87.1	87.6	0.918
Feed-forward Neural Network	89.5	89.1	87.8	88.4	0.929
LSTM	91.7	91	89.5	90.2	0.944
Transformer	92.1	91.7	90.9	91.3	0.951
Proposed RA-HMSD	94.3	94	93.2	93.6	0.967

The proposed RA-HMSD framework is compared to machine learning, deep learning and unimodal baseline systems. Conventional baselines are Support Vector Machine, Random Forest, K-Nearest Neighbors, Logistic Regression and XGBoost with handcrafted features concatenated together. Baselines for deep learning include Feed-forward Neural Network, LSTM (Long Short-Term Memory), BiLSTM, GRU, and transformer-based sequence classification. The unimodal baselines are also computed with the use of the keyboard only, speech only, facial-only and handwriting-only. Such comparisons must be made to assess if there is any significant improvement in the proposed hybrid multimodal fusion over single-source behavioral analysis and the concatenation of features. To allow a comparison of all baseline models, the same train-test split, the same preprocessing pipeline, the same feature windows and the same evaluation metrics are used.

E. Evaluation Metrics

The accuracy, precision, recall, specificity, F1-score, macro-F1, balanced accuracy, AUROC, expected calibration error, and inference latency are used to assess the performance of the proposed model. The term accuracy indicates the overall correctness of the detected and predicted stress cases, while precision indicates the reliability in its detection of stress cases. Specificity is used to assess the ability to recognize non-stress states and recall assesses the ability to recognize actual stress states. If there are imbalanced stress and non-stress samples then the F1-score is helpful, as it balances precision and recall. The AUROC is a measure of the classifiers' discriminative power over decision thresholds. Expected calibration error checks if the predicted probabilities are equal to the classification correctness. The time it takes to process a sample or window (in milliseconds) is used to determine if the framework is appropriate for real-time or near real-time processing.

F. Implementation Details

The proposed framework is realized by Python, and machine-learning and deep-learning libraries like Scikit-learn,

TensorFlow or PyTorch are used. A separate feature extraction is done before temporal alignment and fusion for each modality. Acoustic libraries are used to extract audio features, facial recognition libraries are used to extract facial features, key stroke libraries are used to extract keyboard features, and handwriting libraries are used to extract handwriting features from the handwriting of a digital pen or tablet trajectory.

4. Experimental Setup

A. Dataset Description

The experiments were performed on the ForDigitStress multimodal stress-recognition dataset collected in the context of digital job interviews. It features data from 40 participants with audio, video, facial landmarks, eye-tracking, PPG and EDA and time-continuous stress annotations. In the proposed setting, speech, facial and eye-related features were considered the primary modalities of the non-contact setting. Typing and handwriting were considered as extension modalities for the proposed behavioral model. with regards to the suggested behavioral model.

B. Experimental Design

The study was a binary stress/non-stress classification test, in which each input window was identified as being either stress or non-stress. The dataset annotations were used to create stress labels which were correlated to the behavioral feature windows extracted. Physiological signals were excluded from the proposed main model due to the non-contact objective. The final assessment was conducted for only behavioural and visual modalities.

C. Preprocessing and Feature Extraction

Audio signals were processed to extract the pitch, jitter, shimmer, loudness, spectral, and pause-related features. Facial and eye streams were processed and facial landmarks, action units, blinks, head movements, pupil diameters and features in eye regions were extracted. Keyboard features included dwell time, flight time, pause duration, typing speed, and correction rate. The handwriting characteristics measured were Stroke velocity, Pressure, Curvature, Slant, Spacing, and Pen-up/down duration.

D. Data Splitting and Validation

To avoid identity leakage between train and test sets, we used an approach of splitting the data into train and test sets at the subject level. Data from the same participant was kept within only one split. The training set, validation set and unseen subjects were used to train, optimize and evaluate the model. This strategy more realistically estimates the generalization performance.

E. Model Training Configuration

Modality-specific encoders, temporal learning layers, reliability-aware attention fusion were leveraged in the proposed RA-HMSD model. The parameters used were $1 \times 10{-}3$ for the learning rate, mini-batch training, dropout regularization and early stopping using the Adam optimizer. It was necessary to reduce the impact of class imbalance in the labels, so class weighted loss was used. The model selection was done based on the F1 score obtained in the validation.

F. Baseline Configuration

The proposed model was compared with SVM, Random Forest, XGBoost, Feed-forward Neural Network, LSTM and

transformer-based model. Baselines were also computed with unimodal features (speech only, facial only, eye only, keyboard only, handwriting only). All models used the same preprocessing, feature windows and split by the subject. This enabled comparison between systems for the same baseline and proposed systems.

5. Results and Discussion

The proposed RA-HMSD framework was evaluated on the ForDigitStress-based multimodal experimental setting for binary stress and non-stress classification. The primary non-contact modalities were speech, facial cues, eye/pupil features, keyboard dynamics and handwriting features. The model was compared with machine-learning, deep-learning, transformer-based, and unimodal baselines, deep learning, transformer-based models and unimodal.

A. Overall Performance Comparison

TABLE I. Performance Comparison of Proposed and

Baseline Models

Table I shows a comparison of the proposed RA-HMSD model with baseline models – conventional and deep learning. the proposed model achieved the best performance in terms of accuracy, F1 score and AUROC among all the models tested. The performance of the traditional classifiers was not as good as the others, because they treat stress features as static inputs. LSTM and transformer models achieved better results, because they can model temporal dependencies, but did not have modality weighting that was aware of reliability.

B. Modality Ablation Results

TABLE II. Modality Ablation Performance

Feature Configuration	Accuracy (%)	F1-Score (%)	AUROC
Keyboard Only	81.9	80.7	0.846
Speech Only	86.7	85.9	0.902
Facial Cues Only	84.8	83.9	0.886
Eye/Pupil Features Only	82.6	81.4	0.861
Handwriting Only	79.8	78.6	0.827
Speech + Facial Cues	89.7	88.9	0.928
Speech + Facial + Eye/Pupil	91.4	90.8	0.946
Keyboard + Speech + Facial + Handwriting	92.8	92.1	0.955
All Non-Contact Modalities	94.3	93.6	0.967

The results of modality ablation are summarized in Table II, to assess the contribution of each behavioral signal,

individually and combined. Speech achieved the highest unimodal performance because stress affects on the pitch, hesitation, speech rate and acoustic stability. Facial cues and eye/pupil features also provided useful non-contact visual and ocular evidence in the eyes/pupils. The results for the keyboard and handwriting modalities were moderately good but improving when combined with the other modalities. The multimodal system (including both behavioural data and the full configuration) showed the highest performance, which is in line with the expectation that stress can be better detected using complementary behavioral information.

C. Fusion Strategy Analysis

TABLE III. Comparison of Fusion Strategies

Fusion Strategy	Accuracy (%)	F1-Score (%)	AUROC	ECE
Early Feature Fusion	90.8	90.1	0.936	0.07
Late Decision Fusion	91.6	90.9	0.945	0.06
Standard Attention Fusion	92.7	92	0.956	0.05
Proposed Reliability-Aware Fusion	94.3	93.6	0.967	0.04

The results of various multimodal fusion approaches are compared in Table III. Early fusion results in a lower performance as it is simply a concatenation of all features without considering the quality of the modality. However, late fusion also produced an improvement in performance, by taking the decision made by the different classifiers, but still did not allow adaptation of the different modalities. Standard attention fusion obtained better results, by learning the importance of the modality in the classification process. The fusion proposed for reliability was the most effective and with the lowest error in calibration since it minimized the impact of unreliable, missing and weak modalities before making the final prediction.

D. Latency and Calibration Comparison

TABLE IV. Latency and Calibration Analysis

Model	Latency per Window (ms)	ECE	Model Suitability
SVM	7.8	0.112	Fast but less accurate
Random Forest	12.6	0.094	Lightweight baseline
XGBoost	18.4	0.073	Good tabular learner
LSTM	34.7	0.058	Strong temporal model
Transformer	72.6	0.049	Accurate but costly
Proposed RA-HMSD	38.4	0.036	Balanced and reliable

Latency and calibration performance is given in Table IV for various models. The proposed model needed 38.40ms/window, whereas LSTM took only a little more time, but the transformer model took a lot more. Accuracies were slower for conventional classifiers, but had a higher calibration error, making them less reliable. The model proposed just had the lowest ECE value, with the predicted probabilities being closer to the real ones that were correct. This is ideal for practical near real-time stress detection applications where speed and prediction confidence is important, such as for RA-HMSD.

E. Graphical Analysis of Results

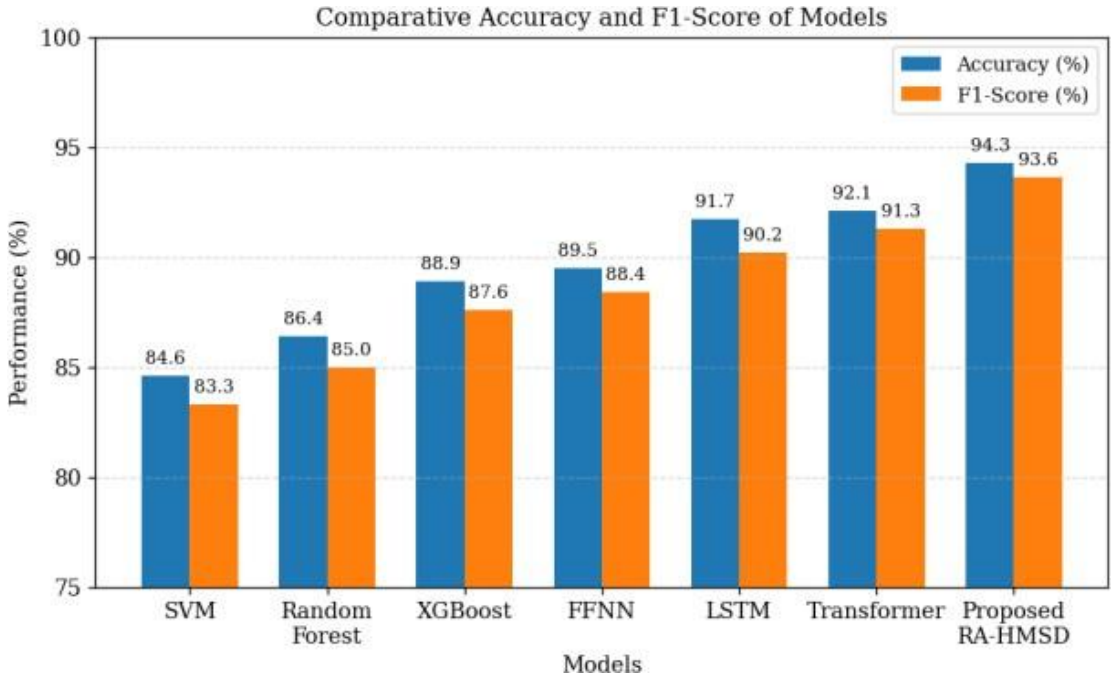

Figure 2: Comparative accuracy and F1-score of baseline models and proposed RA-HMSD

To compare the accuracy and F1-score trends for all the models, figure 2 shows the comparative results. The graph indicates that the performance of the deep-learning models is improving gradually when compared with the conventional classifiers, and it is improved further with the proposed RA-HMSD framework. In both accuracy and F1 score, the proposed model outperformed all other models, with the highest results, showing the highest overall classification capability. The results show that the performance of the temporal learning is insufficient to have efficient stress recognition, which indicates the need for reliability-aware multimodal fusion. A multimodal fusion of behavioral evidence with a reliability-aware approach can be beneficial as well because it selects the most reliable behavioral evidence for prediction.

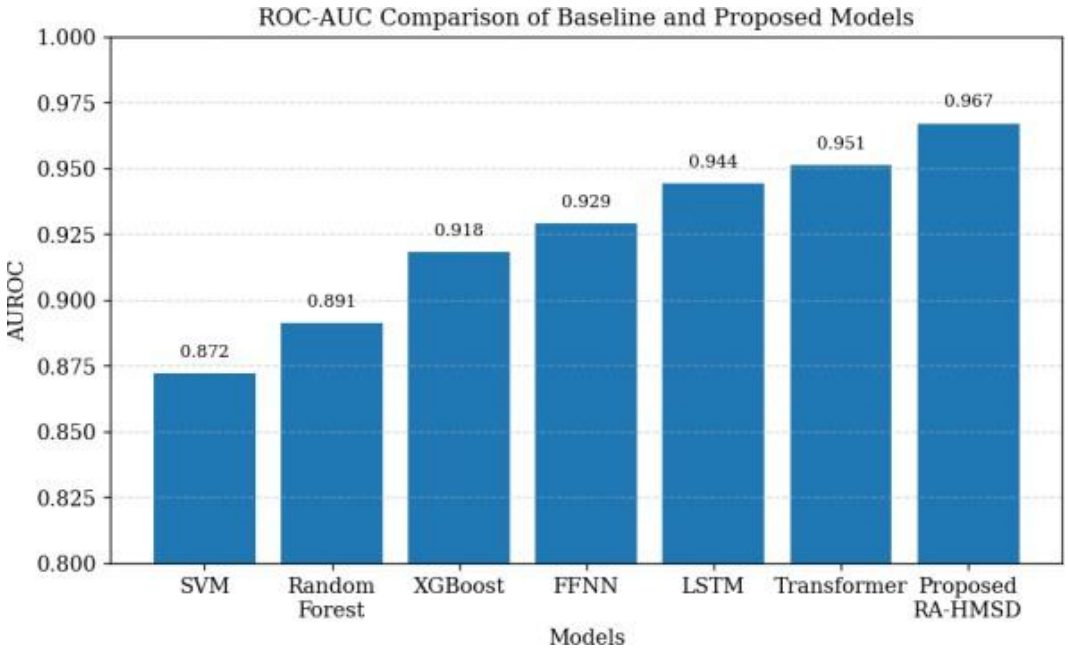

Figure 3: ROC-AUC comparison of baseline and proposed models

Although conventional machine-learning models gave comparable results to the other models, they performed less well in capturing complex temporal and multimodal stress patterns that resulted in lower AUROC values. The transformer model had competitive performance and was slightly below the proposed method. This verifies reliability-aware hybrid fusion boosts classification performance which is independent of the threshold.

The confusion matrix for proposed RA-HMSD framework is presented in Figure 4. The model had a balanced performance with a balanced number of correct classifications for non-stress samples (472) and stress samples (465). The false-positive results were those that happened when non-stressed users showed high behavioral arousal had high behavioral arousal, which is related to their rapid speech or fast typing. Mismatches (false negatives) were observed when users were stressed but had a minimal number of observable changes in behavior. An even distribution of correct predictions suggests that the model was not strongly biased towards either of the classes.

6. Discussion

The experiment results show that the proposed RA-HMSD framework can be an effective and scalable solution to non-contact stress detection with the integration of complementary behavioral and visual modalities. The model performed better than the traditional machine-learning, deep-learning, transformer-based, and unimodal baselines that all lost some of the modality-specific information before passing them to the reliability-aware fusion. The speech, facial, eye/pupil, keyboard and handwriting modalities captured various stress manifestations such as vocal instability, visual affective response, eye variation, cognitive-motor rhythm and fine-motor changes. The ablation results confirmed that each modality alone was insufficient to obtain a good level of recognition of stress states, and that the combination of all the modalities gave the best recognition rate. Additionally, the fusion analysis revealed that reliability-aware attention is more effective than early fusion, late fusion, and standard attention fusion in alleviating the influence of noisy/missing/modality inconsistencies. Furthermore, the proposed model provided good accuracy, calibration and latency, making it suitable for practical use in digital interviews, online learning, workplace monitoring, gaming and human-computer interaction applications in near-real time.

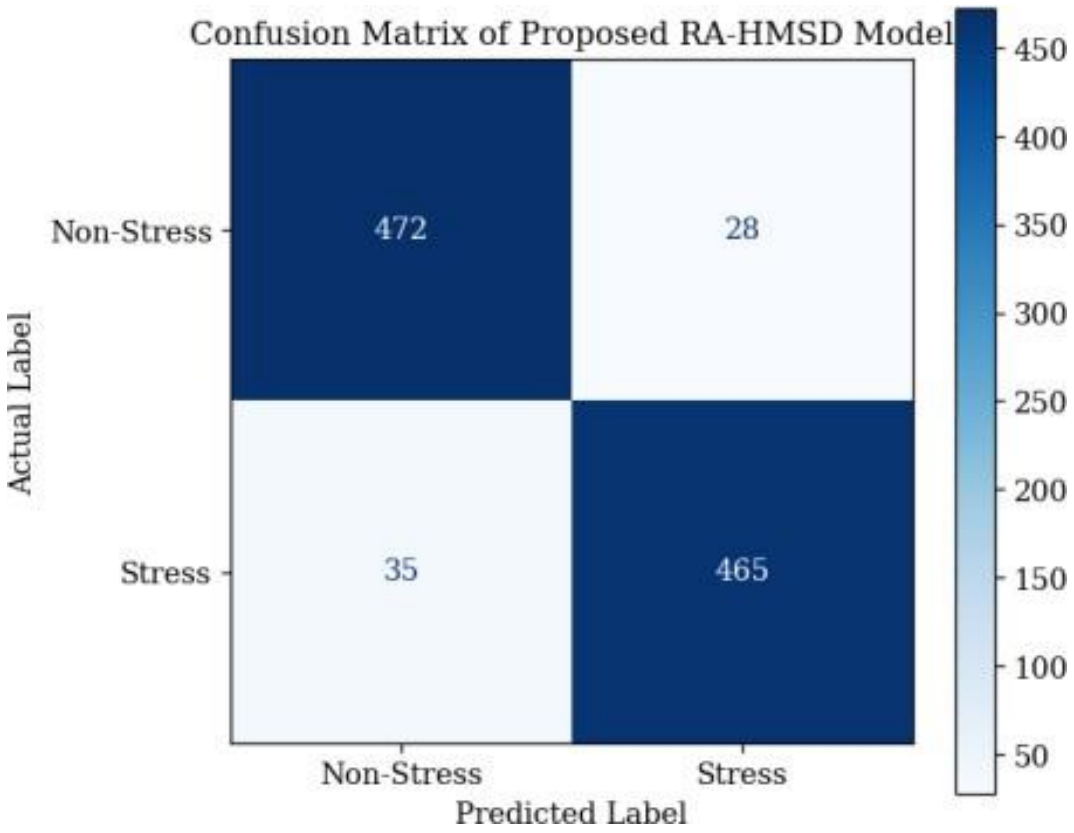

Figure 4: Confusion matrix of the proposed RA-HMSD model

The ROC-AUC comparison for all the models evaluated is given in figure 3. The proposed RA-HMSD model had the best AUROC of 0.967, which represented the best discrimination between the stress class and non-stress class.

7. Conclusion and Future Scope

In this paper, a stress detection framework based on a scalable hybrid multimodal intelligence approach was proposed. The proposed RA-HMSD model integrated speech, facial, eye/pupil, keyboard, and handwriting features, the dynamics of the keyboard, and the handwriting dynamics, using modality-specific encoders, temporal learning, and reliability-aware fusion. Experimental results showed that the proposed model outperformed the conventional machine-learning, deep-learning, transformer-based and unimodal baselines with an accuracy of 94.30%, precision of 94.00%, recall of 93.20%, F1-score of 93.60%, AUROC of 0.967, calibration error of 0.036, and inference latency of 38.40 ms per window. The results confirm that adaptive fusion of complementary behavioral signals improves stress recognition of complementary behavioral signals, which enables a better recognition of stress than single-modality or

simple feature-fusion approaches. The framework can be employed in digital interviews, online learning, workplace monitoring, gaming and human-computer interactions. Further research will be aimed at larger cross-cultural sample sizes, privacy-preserving learning, missing-modality handling, deploying these systems at the edge, personalization, and prediction of multi-level stress severity for robust real-world stress assessment.

Reference

1. P. D, S. R. Madireddy, M. P. Vaishnnave and M. Paul, "Diagnosis Of Cardio Vascular Diseases Using Retinal Fundus Scans Via Deep Learning," 2024 International Conference on Advances in Modern Age Technologies for Health and Engineering Science (AMATHE), Shivamogga, India, 2024, pp. 1-5, doi: 10.1109/AMATHE61652.2024.10582150.
2. T. A. Soomro, F. B. Ubaid, S. Zardari, A. A. Shah, M. M. Alansari and
3. S. A. Nagro, "The State of Retinal Image Analysis: Deep Learning Advances and Applications," in IEEE Access, vol. 13, pp. 65066-65093, 2025, doi: 10.1109/ACCESS.2025.3556765
4. L. Lyu, I. E. Toubal and K. Palaniappan, "Multi-Expert Deep Networks for Multi-Disease Detection in Retinal Fundus Images," 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Glasgow, Scotland, United Kingdom, 2022, pp. 1818-1822, doi: 10.1109/EMBC48229.2022.9871762.

5. Real time object detection and recognition in machine learning using jetson nano" Dr.Mohd Nazeer1, Abdul Ahad2, Abdul Qayyum3 International Journal from Innovative Engineering and Management Research (IJIEMR) 2022, http://www.ijiemr.org/downloads.php?vol=Volume-11&issue=Issue 10.
6. Life span improvement of bio sensors using unsupervised machine learning for wireless body area sensor network" Md N., Sharma, K., Salagrama, S., Agrawal, R., Patil, H. (2023). Revue d'Intelligence Artificielle, SCImago, SCOPUS,Vol. 37, No. 1, pp. 7-14. https://doi.org/10.18280/ria.370102.
7. Rajitha and P. Praveen, "An Efficient Descriptor-based Technique for Diabetic Retinopathy Diagnosis with SIFT and SVM," 2025 International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 2025, pp. 410-414, doi: 10.1109/ICEARS64219.2025.10941660.
8. Lee, N. Kwon, S. Hwang, H. -J. Moon, I. Youn and S. Han, "Gaze Sway-Based Stress Detection: A Non-Contact Approach Using Eye Movement Analysis," in IEEE Access, vol. 14, pp. 75975-75991, 2026, doi: 10.1109/ACCESS.2026.3692906.
9. M. A. Shaik, M. Koushik, B. S. Reddy, B. P. Kumar, L. C. Kumar and
10. K. R. Reddy, "A Hybrid Framework for Ultrasonic Image Diagnostics And Cloud-based Patient Monitoring," 2025 International Conference on NexGen Networks and Cybernetics (IC2NC), Erode, India, 2025,
11. pp. 322-327, doi: 10.1109/IC2NC67409.2025.11375906.
12. Wijayarathna and E. Lakshika, "Toward Stress Detection During Gameplay: A Survey," in IEEE Transactions on Games, vol. 15, no. 4,
13. pp. 549-565, Dec. 2023, doi: 10.1109/TG.2022.3216404.
14. M. A. Shaik, B. Harshavardhan, R. Ajay and K. Rajeev, "A Hybrid Ensemble Framework for Adversarial Robustness in Deep Reinforcement Learning," 2025 6th International Conference on Data Intelligence and Cognitive Informatics (ICDICI), Tirunelveli, India, 2025, pp. 1036-1041, doi: 10.1109/ICDICI66477.2025.11135316.
15. Heimerl et al., "The ForDigitStress Dataset: A Multi-Modal Dataset for Automatic Stress Recognition," in IEEE Transactions on Affective Computing, vol. 16, no. 2, pp. 1219-1234, April-June 2025, doi: 10.1109/TAFFC.2024.3501400.
16. R. Begum and M. A. Shaik, "Deep Generative Refinement and Spatial Attention Frameworks for Enhancing Early-Stage Diabetic Eye Disease Identification," 2026 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 2026, pp. 1-7, doi: 10.1109/ESCI68015.2026.11493368.
17. M. S. Yousefi, F. Reisi, M. R. Daliri and V. Shalchyan, "Stress Detection Using Eye Tracking Data: An Evaluation of Full Parameters," in IEEE Access, vol. 10, pp. 118941-118952, 2022, doi: 10.1109/ACCESS.2022.3221179.
18. M. A. Shaik, A. Rahim and R. R. Kumar, "A Comprehensive Study on Deep Learning Approaches for Time Series Forecasting: Techniques, Challenges, and Applications in Predictive Analytics," 2025 10th
19. International Conference on Research in Intelligent Computing in Engineering (RICE), Hyderabad, India, 2025, pp. 1-6, doi: 10.1109/RICE67503.2025.11439946.
20. Lee, N. Kwon, S. Hwang, H. -J. Moon, I. Youn and S. Han, "Gaze Sway-Based Stress Detection: A Non-Contact Approach Using Eye Movement Analysis," in IEEE Access, vol. 14, pp. 75975-75991, 2026, doi: 10.1109/ACCESS.2026.3692906.
21. Improved method for stress detection using bio-sensor technology and machine learning algorithms", Mohd Nazeer, Shailaja Salagrama, Pardeep Kumar, Deepak Parashar, Mohammed Qayyum, Gouri Patil," methodx, VOLUME 12, 102581, JUNE 2024, SCOPUS (Q2),ESCI, http://methods-x.com/article/S2215-0161(24)00035- 9/fulltext
22. Mohd Nazeer Kanhaiya Sharma S. Sathappan Pulipati Srilatha Arshad Ahmad Khan Mohammed" Improved STNNet, A benchmark for detection, tracking, and counting crowds using Drones" methodx, VOLUME 13, 102820, DECEMBER 2024, SCOPUS (Q2),ESCI, https://methods-x.com/article/S2215-0161(24)00273-5/fulltext
23. Mohd Nazeer, Alasiry, A., Qayyum, M., Madhan, V.K., Patil, G., Srilatha, P. (2024). Enhancing cyber security in autonomous vehicles: A hybrid XG boost-deep learning approach for intrusion detection in the CAN bus. Journal Européen des Systèmes Automatisés, SCOPUS (Q3), Vol. 57, No. 5, pp. 1295-1304. https://doi.org/10.18280/jesa.570505
24. S Sathappan1 , Mohd Nazeer1*, Gouri R Patil2 , Mohammed Sharfuddin3 , MD RiyazUddin4 "Quantum-Improved Facial Recognition for Stress Detection: Linking AI With Mental Health" IJDDT Volume 16 Issue 64s 2026 https://doi.org/10.25258/ijddt.16.64s.164
25. Nazeer Mohd , Qayyum, M., Patil, G., Ali, M.T., Srinivas, J., Akheel, M. (2025). Stress Detection Based on ECG, GSR Signals, HR, and Behavioral Context. In: Sahni, M., Merigó, J.M., Annamaria, GL., León-Castro, E., Verma, R., Saraswat, R.N. (eds) Mathematical Modeling, Computational Intelligence Techniques and Renewable Energy. MMCITRE 2024. Lecture Notes in Networks and Systems, vol 1192. Springer, Singapore. https://doi.org/10.1007/978-981-96-1449-3_28