IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, May 2026
Impact Factor: 8.2

A Multimodal Distress Inference Framework Using Contextual Late Fusion and Suppression-Aware Temporal Supervision

Param Roighare[1], Harshul Yadav[2], Dakshesh Kale[3], Gaurav Gaikwad[4], Prof. Renuka Arbat[5]
[1,2,3,4,5] B.Tech CSE Final Year Student, [5] Mentor and Faculty Guide
MIT ADT University, Pune, India

Abstract: ***Background:*** *Existing speech emotion recognition and wearable distress-monitoring systems predominantly rely on unimodal acoustic inference, making them susceptible to emotionally induced false alarms and to suppression-related acoustic failure. This paper presents a proof-of-concept multimodal distress inference framework that prioritises contextual reliability over raw classification accuracy. The architecture unites three components: (1) a speaker-relative acoustic deviation model that estimates distress confidence from personalised vocal z-scores, (2) a calibrated inertial branch fused at the decision level to suppress physically implausible distress activations such as sports excitement and domestic arguments, and (3) a suppression-aware temporal supervisory layer that monitors multimodal inconsistencies across rolling windows to recover scenarios where acoustic evidence collapses. Experiments on CREMA-D-derived acoustic features and controlled synthetic inertial simulations show an acoustic AUC of 0.808, a reduction in false-positive rate from 43.1 % → 9.4 % after multimodal fusion, with false-negative rates remaining stable (26.0 % → 25.7 %), and successful recovery of 9 out of 10 adversarial suppression scenarios. The results support the proposition that lightweight contextual fusion and temporal supervisory logic can substantially improve the reliability of safety-oriented distress inference beyond conventional acoustic-only approaches.*

Keywords: distress inference; speech emotion recognition; decision-level late fusion; wearable sensing; multimodal safety; suppression-aware inference; speaker-relative normalization

I. INTRODUCTION

Automatic paralinguistic distress inference has attracted sustained interest across affective computing, emergency response, and wearable health technology. Conventional speech emotion recognition (SER) systems extract handcrafted acoustic descriptors—such as Mel-Frequency Cepstral Coefficients (MFCCs), pitch, energy envelope, and spectral features—and classify them using machine learning pipelines ranging from Support Vector Machines to gradient-boosted ensembles and deep neural architectures.[1,4,5] While such systems achieve competitive accuracy on controlled emotional speech benchmarks, their deployment in safety-critical scenarios exposes a critical structural weakness: vocal intensity alone is insufficient to distinguish genuine distress from emotionally expressive but non-dangerous behaviour such as arguments, celebrations, and sports excitement.

A second, less studied failure mode arises when an adversary actively suppresses the user's ability to verbalise distress. Scenarios such as physical restraint, mouth covering, device seizure, or coercive whispering remove the acoustic evidence that unimodal systems depend on entirely. Neither failure mode is adequately addressed in existing literature.[6,10]

This work proposes a three-stage multimodal distress inference framework that addresses both limitations. The acoustic branch employs speaker-relative z-score deviation modeling—normalising each utterance against the speaker's own neutral baseline—to reduce inter-speaker variability.[3] A calibrated inertial branch is fused at the decision level to

DOI: 10.48175/IJARSCT-34939

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 3, May 2026 Impact Factor: 8.2

down-weight false activations in physically implausible scenarios.[6,7] Finally, a temporal supervisory layer monitors rolling windows of voice activity, acoustic confidence, and motion magnitude to detect suppression signatures and recover cases where the acoustic branch fails entirely. The primary contributions are:

1. A speaker-relative acoustic distress pipeline using personalised z-score deviation vectors and calibrated XGBoost classification.
2. A decision-level late fusion framework integrating acoustic and inertial branches to reduce emotionally induced false positives.
3. A suppression-aware temporal supervisory mechanism that recovers adversarial scenarios including abrupt silence, physical restraint, and device seizure.
4. A controlled multimodal evaluation protocol using CREMA-D acoustic data and controlled synthetic inertial simulations designed to quantify contextual reliability under safety-oriented conditions.

II. LITERATURE REVIEW

A. Acoustic Emotion Recognition

SER systems have been studied extensively over the past two decades. Traditional pipelines combine handcrafted descriptors—MFCCs, pitch statistics, zero-crossing rate, energy features, and spectral centroid—with classifiers such as SVMs, Random Forests, and gradient-boosted methods.[4,5] Zhao et al. demonstrated that combining MFCC, ZCR, spectral roll-off, jitter, and RMS energy with a 1-D convolutional architecture achieves strong intra-corpus accuracy on EMODB and RAVDESS.[5] Xiong et al. proposed an ensemble combining XGBoost, KNN, SVM, and LightGBM with automatic weight learning, reporting competitive accuracy across multiple benchmark corpora.[4] Despite these advances, acoustic-only systems exhibit elevated false-positive rates in high-arousal non-distress scenarios because vocal intensity features are not uniquely indicative of danger.

B. Speaker Normalisation in SER

Inter-speaker variability is a well-recognised challenge in SER. Gat et al. proposed a gradient-based adversary learning framework that normalises speaker identity from the self-supervised feature representation, achieving state-of-the-art results on IEMOCAP.[3] An earlier body of work addressed speaker normalisation through feature warping at the acoustic level. The present framework adopts a complementary approach: rather than normalising speaker identity from a learned representation, it computes personalised z-score deviation vectors directly from each speaker's neutral baseline, enabling lightweight inference without pretraining requirements. This design choice is motivated by the deployment target of an on-device wearable system with limited compute resources.

C. Multimodal Fusion and Wearable Safety

Decision-level late fusion has been shown to improve robustness against modality-specific failure modes across a variety of sensor combinations. Rehouma and Boukadoum proposed a bimodal late-fusion fall-detection framework combining IMU abnormality detection with vision-based pose estimation, reducing false-positive rate from 11.3 % (IMU-only) to 3.6 % after fusion—a result structurally analogous to the false-positive suppression demonstrated in this work.[7] Spanoudakis et al. collected a publicly available multimodal dataset integrating ECG, respiration, and IMU signals with speech, demonstrating the feasibility of joint acoustic–physiological fusion for stress detection.[6] However, neither work specifically addresses suppression-induced acoustic failure or adversarial safety scenarios involving active concealment of distress signals.

In wearable safety systems, inertial sensing is a natural complement to acoustic inference because IMU signals remain observable even when acoustic distress is partially suppressed. Ghosh et al. proposed a voice-activated SOS wearable that transmits location upon detecting a distress call; that system does not address the case where the voice call itself is suppressed.[10] US Patent 9,922,537 covers a multi-sensory wearable safety device combining voice, motion, emotion, and impact sensing through threshold-based rules rather than a learned fusion model.[11] The present framework extends beyond threshold logic by using calibrated probabilistic outputs and a temporal supervisory layer, and explicitly targets the adversarial suppression failure mode unaddressed in prior art.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, May 2026
Impact Factor: 8.2

D. Limitations in Existing Research

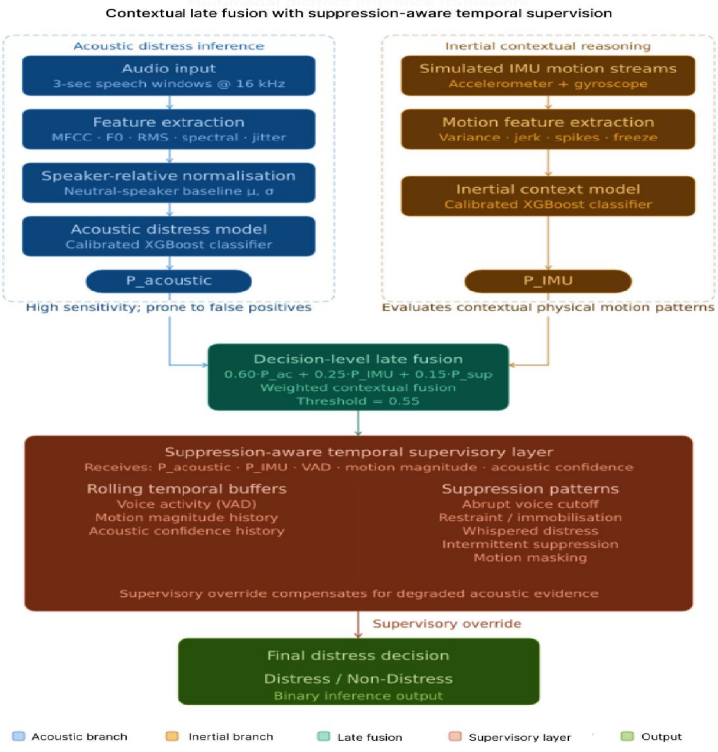
Three gaps motivate this work. First, acoustic SER systems prioritise average classification accuracy rather than contextual reliability, resulting in high false-positive rates in safe high-arousal situations. Second, existing multimodal systems rarely address suppression-related acoustic failure caused by restraint, device seizure, or deliberate silencing. Third, publicly available datasets containing synchronised distress speech and wearable inertial signals do not exist, which forces controlled simulation as a substitute for evaluation—a limitation shared across the field [8,9] and explicitly acknowledged in this paper.

III. RESEARCH GAP AND PROBLEM STATEMENT

Synthesising the literature above, the central problem addressed in this work is: "How can a multimodal wearable distress inference system reduce emotionally induced false alarms while remaining robust to suppression-related acoustic failure under temporally asynchronous multimodal conditions?" Three specific gaps remain unaddressed: (i) false-positive suppression in safe high-arousal scenarios, (ii) recovery when acoustic evidence disappears due to adversarial suppression, and (iii) evaluation methodology that does not require synchronised paired datasets. This work proposes a lightweight framework targeting all three gaps without requiring synchronised paired training data or end-to-end deep learning pipelines.

IV. PROPOSED FRAMEWORK

The system is a hierarchical three-stage pipeline: (1) acoustic distress inference, (2) contextual inertial late fusion, and (3) suppression-aware temporal supervision. Stages 1 and 2 generate an instantaneous multimodal distress estimate; Stage 3 monitors short temporal windows for suppression signatures.



IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, May 2026
Impact Factor: 8.2

Fig. 1. Conceptual overview of the proposed multimodal distress inference framework. The system combines speaker-relative acoustic inference, contextual inertial late fusion, and suppression-aware temporal supervision for binary distress estimation under emotionally ambiguous and acoustically degraded conditions.

A. Acoustic Distress Inference Branch

Audio samples from CREMA-D[1] are resampled to 16 kHz and standardised to three seconds. Emotion labels are mapped binarily: angry (ANG), fearful (FEA), and disgust (DIS) to distress; neutral (NEU), happy (HAP), and sad (SAD) to calm. A 36-dimensional handcrafted feature vector is extracted per sample: 13-coefficient MFCC mean and standard deviation vectors (26 dimensions), pitch mean/standard deviation/maximum (3), voiced ratio (1), RMS energy mean/std/max (3), zero-crossing rate mean (1), spectral centroid mean (1), and jitter (1).

For each speaker, neutral utterances are used to construct a personalised baseline distribution. All feature vectors from that speaker are then transformed via z-score normalisation relative to that baseline, producing deviation vectors that capture how far each utterance departs from that individual's calm vocal norm rather than a population mean.[3] Robust scaling is applied before classification to suppress outlier influence.

An XGBoost classifier[2] is trained using GroupShuffleSplit to enforce speaker-independent evaluation. Isotonic probability calibration is applied via CalibratedClassifierCV to produce reliable distress probabilities for downstream fusion

B. Inertial Context Branch

No publicly available dataset provides synchronised distress speech and wearable IMU recordings. Consistent with methodological precedent in the wearable fall-detection literature, [8,9] controlled synthetic IMU sequences are generated for contextual plausibility evaluation. Eight motion patterns are simulated: calm sitting, calm walking, sports excitement (all calm class), and assault struggle, restrained immobilisation, panic running, phone grabbing, and collapse (all distress class). Statistical descriptors are extracted per window: acceleration magnitude statistics, gyroscope magnitude statistics, jerk metrics, interquartile range, impact indicator (peak > 4 g), stillness indicator (mean $\approx$ 1 g), high-acceleration ratio, and inter-axis correlation (14 features total).

A separate XGBoost classifier with isotonic calibration is trained on these synthetic sequences. The purpose of this branch is not generalised activity recognition but contextual plausibility: does the observed body motion support or contradict an acoustic distress prediction?

C. Decision-Level Late Fusion

Each branch produces an independent calibrated probability. These are combined through weighted late fusion:

$$P\_fusion = 0.60\, P\_acoustic + 0.25\, P\_IMU + 0.15\, P\_suppress$$

The acoustic branch receives the highest weight as the more directly validated modality. Strict frame-synchronous alignment is not required; multimodal evidence is interpreted over short contextual windows to accommodate the temporal asynchrony between vocal and physical distress responses in real emergency scenarios.

D. Suppression-Aware Temporal Supervisory Layer

A lightweight rule-based supervisory layer operates above the fusion stage. It maintains rolling buffers of voice activity detection (VAD) state, acoustic distress probability, and acceleration magnitude. The supervisory layer evaluates six primary suppression-related behavioral patterns alongside baseline distress and false-positive suppression scenarios. These include abrupt voice cutoff during motion spikes, immobilization following speech activity, sudden acoustic confidence loss, whispered distress, intermittent voice disappearance, and chaotic motion with weak acoustic evidence. Additional supervisory evaluation scenarios include sports excitement, emotionally expressive calm-body arguments, and safe high-arousal vocal behavior in order to validate that the supervisory override does not incorrectly escalate ordinary emotional activity.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, May 2026
Impact Factor: 8.2

When a pattern is detected above a confidence threshold, the system activates a supervisory override that elevates final distress confidence using inertial and suppression evidence, compensating for degraded acoustic evidence.

V. EXPERIMENTAL METHODOLOGY AND IMPLEMENTATION

A. Acoustic Dataset

CREMA-D[1] is an audio-visual emotional speech corpus comprising 7,442 clips from 91 demographically diverse actors across six emotion categories. It provides controlled emotional vocal variation suitable for studying distress-like acoustic behaviour under speaker-independent conditions. The acted nature of the corpus—and its consequent departure from naturally occurring emergency speech—is acknowledged as a key limitation (see Section VII).

B. Classifier Training and Calibration

XGBoost[2] was selected over deep neural architectures because the handcrafted feature space is already semantically interpretable, the dataset size is moderate, and the deployment target requires lightweight inference. Gradient-boosted trees offer strong performance on structured tabular data with fewer regularisation requirements than deep networks at this scale.[4]

For the acoustic branch, GroupShuffleSplit enforces speaker-independent splitting (test speakers are fully held out from training). Both acoustic and inertial models undergo isotonic probability calibration, which is necessary because raw XGBoost output scores are systematically overconfident and numerically incompatible across independently trained branches during weighted fusion. The operating threshold of 0.55 was selected empirically by maximising the safety-prioritised score function (1 − FN rate) × 3 − FP rate across threshold values in [0.1, 0.9].

C. Evaluation Protocol

Three sequential experiments were conducted. Experiment 1 evaluated the acoustic branch independently to establish baseline distress inference performance and characterise emotionally induced false-positive behaviour. Experiment 2 introduced inertial late fusion to evaluate whether contextual motion reasoning could reduce false alarms without significantly degrading distress sensitivity. Experiment 3 evaluated the suppression-aware supervisory framework across adversarial suppression scenarios, baseline distress cases, and emotionally expressive false-positive conditions.

Experiment 1 performance was reported using accuracy, AUC, F1-score, false-positive rate (FPR), and false-negative rate (FNR). Experiment 2 primarily evaluated changes in FPR and FNR under controlled multimodal contextual scenarios, while Experiment 3 evaluated supervisory correctness across representative suppression and false-positive conditions.

VI. RESULTS AND DISCUSSION

A. Experiment 1 — Acoustic Distress Inference

The calibrated XGBoost acoustic model achieved AUC = 0.8083, FNR = 7.4 %, and F1 = 0.7487 on the speaker-independent test split. Distress recall was high (93 %), confirming that speaker-relative deviation features effectively capture emotional intensity. However, FPR reached 57.5 %, demonstrating that acoustic inference alone systematically over-predicts distress in high-arousal non-emergency scenarios—consistent with findings in the SER literature on emotionally ambiguous speech.

TABLE I: Acoustic-Only Distress Inference Performance

Metric	Value
Accuracy	68.17%
AUC	0.8083
F1-Score (Distress)	0.7487

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, May 2026
Impact Factor: 8.2

Precision (Distress)	0.63
Recall / Sensitivity (Distress)	0.93
False Positive Rate	57.46%
False Negative Rate	7.42%

The elevated false-positive rate reflects the inherent ambiguity of emotionally intense speech under unimodal acoustic inference, where vocal intensity alone is insufficient to reliably distinguish genuine distress from safe high-arousal behavior. Experiment 1 therefore establishes the baseline limitation that motivates the introduction of contextual multimodal fusion.

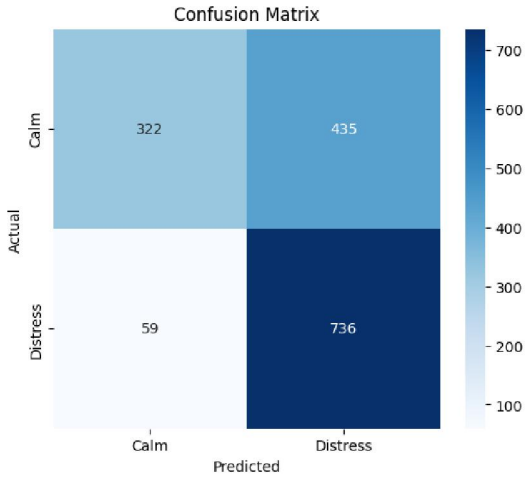


Fig. 2. Acoustic Branch Confusion Matrix

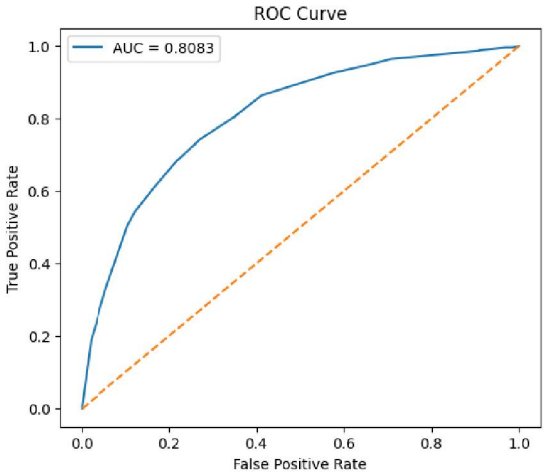


Fig. 3. ROC Curve for Acoustic Branch

B. Experiment 2 — Contextual Late Fusion

Introducing calibrated inertial context through decision-level late fusion reduced FPR from 43.1 % to 9.4 % while FNR remained stable (26.0 % → 25.7 %), representing an approximate 78 % relative reduction in false alarms with negligible sensitivity loss. The result is consistent with the bimodal late-fusion findings of Rehouma and Boukadoum in

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, May 2026
Impact Factor: 8.2

fall detection,[7] and with the general late-fusion literature showing that modality-specific failure modes can be suppressed at the decision layer without requiring feature-level alignment.

The Experiment 2 false-positive and false-negative rates were computed on the controlled scenario-based multimodal evaluation set rather than the full speaker-independent acoustic benchmark used in Experiment 1.

TABLE II: Acoustic vs Late-Fusion Performance

Metric	Acoustic Only	Late Fusion
False Positive Rate	43.1%	9.4%
False Negative Rate	26.0%	25.7%

TABLE III: Scenario-Based Fusion Outcomes

Scenario	Acoustic Score	IMU Score	Fused Score	Final Decision
True Distress	0.689	0.981	0.659	Distress
FP Trap	0.703	0.021	0.427	Calm
Safe Excitement	0.407	0.018	0.249	Calm
True Negative	0.390	0.021	0.239	Calm

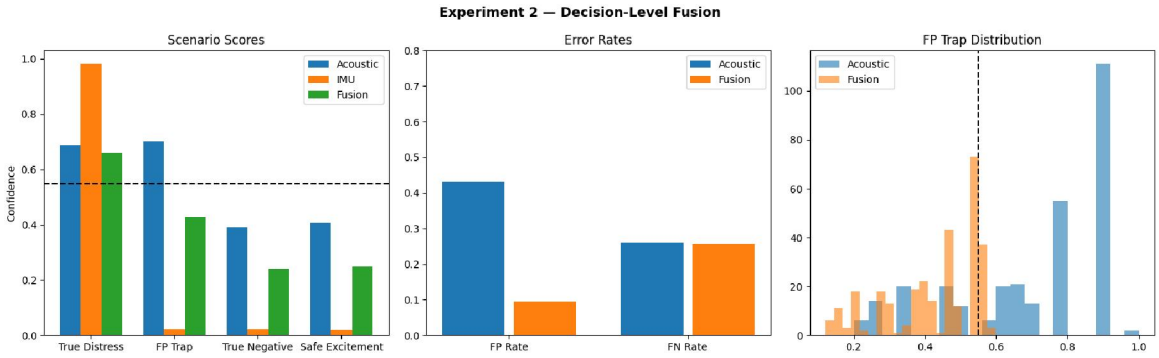


Fig. 4. Decision-level late fusion analysis comparing acoustic-only inference and contextual multimodal fusion. The left panel shows modality-wise confidence scores across representative scenarios. The middle panel compares false-positive and false-negative rates before and after fusion. The right panel visualizes the reduction in false-positive activation after contextual fusion.

The largest improvement occurred in safe high-arousal scenarios such as sports excitement and emotionally expressive calm-body speech. In these situations, the inertial branch produced near-zero distress probability—reflecting the absence of assault-like motion—pulling the fused score below the alert threshold despite elevated acoustic confidence. The late-fusion stage was intentionally designed to tolerate asynchronous modal evidence, reflecting realistic emergency conditions in which panic vocalisations and abnormal body motion may not coincide precisely.

C. Experiment 3 — Suppression-Aware Temporal Recovery

The temporal supervisory layer successfully recovered 9 of 10 adversarial scenarios. All four assault-suppression patterns (mouth covered, restrained, phone seized, intermittent suppression) and two false-positive scenarios (sports excitement, calm-body argument) were handled correctly. Pattern F (chaotic motion masking low voice) was also correctly identified as distress via the suppression override.

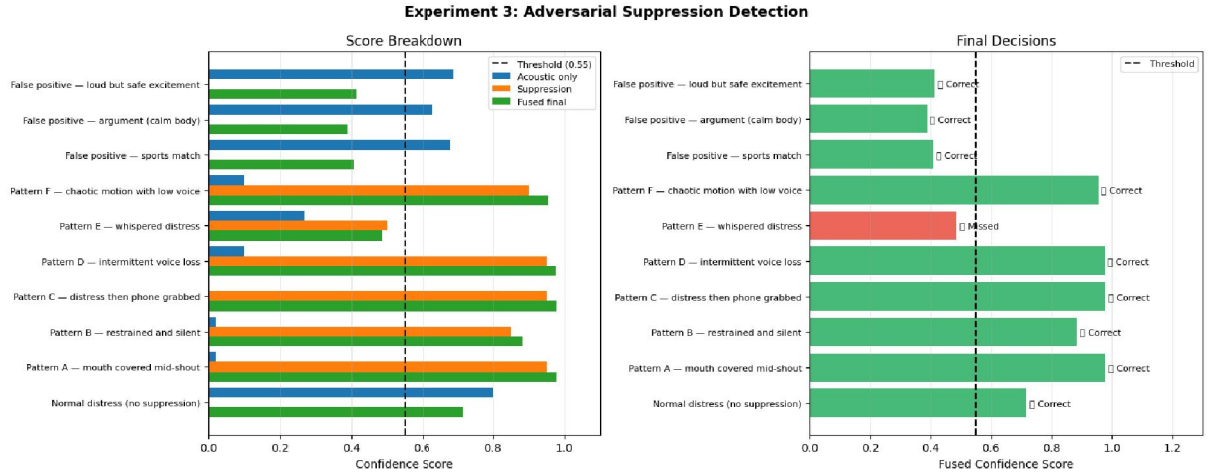
IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, May 2026
Impact Factor: 8.2

TABLE IV: Supervisory Layer Evaluation Across Distress and Suppression Scenarios

The supervisory framework correctly handled 9 out of 10 evaluated scenarios, including suppression-related adversarial conditions, baseline distress inference, and emotionally induced false-positive cases.

Scenario	Acoustic Score	IMU Score	Suppression Score	Final Fused Score	Final Decision	Correct
Normal distress (no suppression)	0.800	0.952	0.000	0.718	Distress	Yes
Pattern A — Mouth covered mid-shout	0.020	1.000	0.950	0.978	Distress	Yes
Pattern B — Restrained and silent	0.020	0.911	0.850	0.884	Distress	Yes
Pattern C — Distress then phone grabbed	0.000	1.000	0.950	0.978	Distress	Yes
Pattern D — Intermittent voice loss	0.100	0.999	0.950	0.977	Distress	Yes
Pattern E — Whispered distress	0.270	1.000	0.500	0.487	Calm	No
Pattern F — Chaotic motion with low voice	0.100	1.000	0.900	0.955	Distress	Yes
False positive — Sports match	0.680	0.000	0.000	0.408	Calm	Yes
False positive — Argument (calm body)	0.630	0.041	0.000	0.388	Calm	Yes
False positive — Loud but safe excitement	0.690	0.000	0.000	0.414	Calm	Yes

The single failure case involved whispered distress (Pattern E). This scenario remained challenging because both the acoustic evidence and the associated inertial abnormality were weak and ambiguous, resulting in a low-observability condition for the multimodal framework. Unlike the other suppression scenarios, the available evidence was insufficient to confidently activate the supervisory override. This result highlights an important limitation of the current framework and suggests that additional physiological sensing modalities may be necessary for reliable detection under extremely low-signal distress conditions.



IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, May 2026
Impact Factor: 8.2

Fig. 5. Suppression-aware temporal supervisory analysis across adversarial and safe scenarios. Left panel shows acoustic, suppression, and final fused confidence scores across representative conditions. Right panel visualizes the final supervisory decisions relative to the activation threshold.

D. Overall Discussion

Collectively, the three experiments demonstrate that safety-oriented distress inference benefits from contextual multimodal reasoning beyond unimodal acoustic classification. The acoustic branch provides emotional evidence; the inertial branch provides contextual plausibility; the temporal supervisory layer provides adversarial recovery capability. Together they implement a hierarchical reliability architecture that reduces FPR by 78 % relative while maintaining distress sensitivity and recovering 90 % of simulated adversarial scenarios. The framework achieves this without requiring synchronised paired datasets or computationally intensive deep learning, making it suitable for edge deployment on resource-constrained wearable hardware. [6,7]

VII. LIMITATIONS

Several limitations affect the generalisability of the current findings. First and most critically, the inertial branch was validated using synthetic motion simulations rather than real synchronised wearable distress recordings. Although the generated patterns were designed to be physically interpretable, they inevitably simplify the biomechanical variability of genuine emergency behaviour. Second, CREMA-D consists of acted emotional speech, which may differ from spontaneous distress in prosodic spontaneity, linguistic variability, and vocal noise characteristics.

Third, the suppression-aware supervisory layer relies on lightweight rule-based heuristics rather than learned temporal sequence models, which may limit its capacity to capture complex long-range dependencies. Fourth, the framework has not been deployed on physical wearable hardware, and real-world challenges including sensor drift, microphone occlusion, motion artefacts, and battery constraints remain unevaluated. Finally, short rolling window operation limits long-term behavioural reasoning; the framework does not model user state across sessions.

VIII. FUTURE WORK

The most important near-term extension is evaluation on real synchronised multimodal datasets combining wearable IMU recordings, physiological signals, and naturally occurring distress audio. Once such data is available, the synthetic inertial branch can be retrained and the calibration adjusted accordingly.

Additional physiological modalities—heart rate variability, electrodermal activity, and skin temperature—could improve contextual robustness, particularly in the whispered-distress regime where both acoustic and inertial evidence remain weak. The supervisory layer could be extended using transformer-based temporal sequence models or recurrent architectures to capture longer-range suppression patterns. On the deployment side, continuous on-device inference, adaptive power management, and ARM Cortex-M NPU deployment via TensorFlow Lite represent concrete engineering targets. Federated personalisation strategies could allow the baseline model to update over time without sharing raw audio.

IX. CONCLUSION

This paper presented a proof-of-concept multimodal distress inference framework combining speaker-relative acoustic deviation modelling, calibrated decision-level late fusion with inertial context, and suppression-aware temporal supervision. Experiments demonstrated a 78 % relative reduction in false-positive activation through contextual fusion (43.1% to 9.4%) with negligible false-negative increase, and successful recovery of 9 of 10 adversarial suppression scenarios. The results support lightweight contextual late fusion and temporal supervisory reasoning as effective tools for improving the reliability of safety-oriented distress inference beyond unimodal acoustic approaches, while remaining tractable for wearable deployment.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, May 2026
Impact Factor: 8.2

ACKNOWLEDGMENT

The authors thank Prof. Renuka Arbat for guidance and the School of Computing at MIT-ADT University, Pune, for supporting this research. The CREMA-D dataset is used under its open database licence.

REFERENCES

[1] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset," IEEE Trans. Affect. Comput., vol. 5, no. 4, pp. 377–390, Oct.–Dec. 2014. doi: 10.1109/TAFFC.2014.2336244.

[2] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining (KDD'16), San Francisco, CA, 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[3] I. Gat, H. Aronowitz, W. Zhu, E. Morais, and R. Hoory, "Speaker normalization for self-supervised speech emotion recognition," in Proc. IEEE Int. Conf. Acoustics, Speech Signal Process. (ICASSP), 2022, pp. 7342–7346. doi: 10.1109/ICASSP43922.2022.9747460.

[4] X. Xiong, H. Li, J. Wang, and Y. Zhang, "A novel speech emotion recognition method based on feature construction and ensemble learning," PLoS ONE, vol. 17, no. 8, p. e0267132, Aug. 2022. doi: 10.1371/journal.pone.0267132.

[5] Y. Zhao, Z. Wang, and Q. Zhao, "Speech emotion recognition based on multiple acoustic features and deep convolutional neural network," Electronics, vol. 12, no. 4, p. 839, Feb. 2023. doi: 10.3390/electronics12040839.

[6] K. Spanoudakis, A. Kalogeras, and M. Giannakopoulou, "A multimodal late fusion framework for physiological sensor and audio-signal-based stress detection: an experimental study and public dataset," Electronics, vol. 12, no. 23, p. 4871, Dec. 2023. doi: 10.3390/electronics12234871.

[7] H. Rehouma and M. Boukadoum, "Fall detection by deep learning-based bimodal movement and pose sensing with late fusion," Sensors, vol. 25, no. 19, p. 6035, Oct. 2025. doi: 10.3390/s25196035.

[8] A. Sucerquia, J. D. López, and J. F. Vargas-Bonilla, "SisFall: A fall and movement dataset," Sensors, vol. 17, no. 1, p. 198, Jan. 2017. doi: 10.3390/s17010198.

[9] E. Casilari, J. A. Santoyo-Ramón, and J. M. Cano-García, "Analysis of public datasets for wearable fall detection systems," Sensors, vol. 17, no. 7, p. 1513, 2017. doi: 10.3390/s17071513.

[10] M. Ghosh et al., "Voice-activated SOS: An AI-enabled wearable device," in Proc. 11th Int. Conf. Emerging Trends Eng. Technol. Signal Inf. Process. (ICETET-SIP), IEEE, 2023, pp. 1–6.

[11] US Patent 9,922,537, "Wearable multi-sensory personal safety and tracking device," filed and granted. [Online]. Available: https://image-ppubs.uspto.gov/dirsearch-public/print/downloadPdf/9922537.