International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

RACF: Reliability-Conditioned Calibrated Fusion for Stress-Related Behaviour Recognition under Missing and Corrupted Modalities

U. Birundhadevi
Department of Artificial Intelligence and Data Science
Jai Shriram Engineering College
Tamil Nadu, India

Abstract **- In a controlled simulation, a reliability-conditioned fusion model (RACF) kept probability calibration stable as input noise grew, but did not beat a strong augmented early-fusion baseline in accuracy. Stress-related recognition from face, voice and typing signals must work when a sensor is missing or corrupted, and its confidence must stay trustworthy when that happens. We propose RACF, which (i) turns a cheap per-window quality indicator (within-window frame jitter) into a learned, quality-conditioned recalibration of each modality's logit, and (ii) fuses the recalibrated logits with a mask-aware linear combiner. We evaluate it against early and late fusion, with and without noise and dropout augmentation, on a synchronized tri-modal simulator with participant-disjoint splits (30 independent repetitions; 26 training, 12 calibration and 10 test participants each). Across a grid of seven availability patterns and four noise levels, RACF holds expected calibration error at 0.04 to 0.05 from clean input to noise far outside its training range, while the augmented early-fusion baseline rises from 0.045 to 0.153 and the clean-trained baseline reaches 0.215. When exactly one modality is corrupted, RACF has the best macro-F1 in five of saix conditions (for example 0.717 against 0.690 with severe face noise). It does not improve macro-F1 over augmented early fusion in a mixed missing-and-noisy scenario (0.688 against 0.687, p = 0.44), and it is slightly worse when every modality is severely corrupted. We also test whether split-conformal prediction sets conditioned on modality availability and estimated reliability improve group-conditional coverage; with reliability-calibrated probabilities they did not, and they enlarged the sets. All results are synthetic and show a mechanism, not performance on human participants. The framework is for non-clinical research and is not a diagnostic tool.**

Keywords **- multimodal learning, missing modalities, calibration, conformal prediction, affective computing, stress-related behaviour, keystroke dynamics.**

I. INTRODUCTION

A stress-related recognition system that is accurate on clean, complete input is not yet usable: cameras are occluded, microphones pick up noise, and typing data may be absent. Three questions follow. Does the model still discriminate? Does its stated confidence still mean what it says? And can a user be told when the output should not be trusted?

Most work on missing modalities reports accuracy only [3]–[6]. Dynamic-fusion work weights modalities by estimated quality, mainly to improve accuracy or stability [7]–[9]. Calibration and set-valued uncertainty under partial input are studied much less, and recent conformal methods for missing modalities target regression, vision-language benchmarks or sentiment, not behavioural stress data [13]–[15].

This paper studies those three questions for a face, voice and typing setting. It uses a controlled simulation because, to our knowledge, no public dataset provides synchronized face, voice and keystroke recordings with stress-related labels from the same participants. Public sources cover subsets: StressID has video, audio and physiological signals [2], WESAD has wearable physiology [1], and keystroke studies use separate cohorts [16]–[18]. Merging unrelated datasets by label would not create person-level multimodal observations, so we do not do it.

The contributions of this paper are as follows.

- *RACF*, a reliability-conditioned calibrated fusion model. A per-window jitter indicator conditions a learned recalibration of each modality's logit; a mask-aware linear combiner fuses the results.
- *A protocol* with participant-disjoint train, calibration and test splits, seven availability patterns, four noise levels (one outside the training range), a one-corrupted-modality test, and a mixed missing-and-noisy scenario. Calibration is reported next to accuracy.
- *A negative result on conformal conditioning.* Availability-conditional and reliability-conditional split-conformal sets did not improve group-conditional coverage over a marginal calibration once the underlying probabilities were reliability-calibrated.
- *Open code.* One script reproduces every number in this paper.

We do not claim that RACF is more accurate than augmented early fusion in the mixed scenario, that availability-conditional conformal prediction is new [13]–[15], or that anything here is evidence about real participants.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

II. RELATED WORK

A. Missing Modalities in Stress and Affect Recognition

ReLearn combines feature selection, outlier detection, imputation and classification for physiological stress detection and reports 86.8% cross-validation accuracy with more than half of the feature values missing [3]. A modality fusion network with personalized attention handles complete and incomplete wearable data (GSR and ECG, 41 participants) for self-reported stress [4]. ADAPT aligns modalities to an anchor and uses a masked multimodal transformer, evaluated on stress induced by triggers and on pilot loss of consciousness [5]. Related work on time-continuous emotion recognition trains with hidden modalities and lets a transformer decoder weigh modalities [22], and Ma et al. study how multimodal transformers degrade when a modality is absent [6]. These papers report accuracy-type metrics; none studies the calibration of the fused probability across noise levels.

B. Quality-Aware Dynamic Fusion

Quality-aware multimodal fusion estimates per-sample modality reliability and reweights evidence [7]. Predictive Dynamic Fusion adds a relative-calibration step for reliability estimates [8]. Unbiased Dynamic Multimodal Learning estimates noise with a controlled-noise estimator and measures modality reliance through modality dropout [9]. RACF is closer in spirit to a light, linear stacking version of these ideas; its contribution is the evaluation of calibration and set size, not a new deep architecture. Calibration is scored with expected calibration error (ECE) [11], and Monte-Carlo dropout is a common source of per-sample uncertainty in dynamic fusion [10]; RACF replaces it with a deterministic quality indicator.

C. Conformal Prediction with Missing Inputs

Split-conformal prediction gives finite-sample marginal coverage under exchangeability [12]. Zaffran et al. obtain coverage conditional on the missingness pattern in regression [13]. Moayedikia proposes Modality-Conditioned Conformal Fusion, which gives group-conditional coverage for every modality-presence subset and reports it on a synthetic task, AVMNIST, CMU-MOSEI and UR-FUNNY [14]. Azizi calibrates multimodal regression intervals using disagreement between per-modality predictors and mask-matched recalibration [15]. Our conformal experiment therefore applies a known idea in a new setting and adds a reliability-conditioned grouping; we report it as an empirical check, including where it did not help.

D. Behavioural Stress Data

Keystroke timing relates to affect: in an online study with 400 participants, dwell time correlated negatively with valence and flight time with arousal [16]. A study relating bigram timing to heart-rate variability concluded that richer, user-specific timing representations are needed [17]. A programmer study found that some timing features changed under time pressure [18]. Public multimodal stress datasets exist for video, audio and physiology (StressID, 65 participants, 11 tasks [2]) and wearable physiology (WESAD [1]); we found none that adds keystrokes for the same participants. A tri-modal system with typing as one modality [19] needs a synchronized collection protocol before real-data claims can be made. Two further multimodal stress datasets combine video, audio and physiology [20] and facial expressions with physiological signals [21].

III. METHOD

RACF has three stages, all linear, so each is cheap to train and inspect. Window i has modalities $m \in$ {face, voice, typing}, an availability flag $a_{im} \in \{0, 1\}$ (at least one is 1), a feature vector x_{im} and a quality indicator q_{im}.

A. Quality Indicator

Each window is a short run of T frame-level feature vectors. The indicator is the log of the mean per-dimension standard deviation across frames, relative to the clean level:

$$q_{im} = \log((1/d_m) \Sigma_k \, \mathrm{std}_t(x_{im,t,k})) - \log \sigma_f \qquad (1)$$

It is near 0 for clean input and grows with noise. In a real system its analogue is face-tracking confidence, audio signal-to-noise ratio or keystroke-timestamp jitter; the paper does not claim that any of these behaves like this simulated indicator.

B. Stage 1: Clean Base Models

One logistic regression per modality is trained on clean windows only, giving a logit z_{im}.

C. Stage 2: Quality-Conditioned Recalibration

The base logit is recalibrated with a second logistic regression whose inputs are the logit, the quality indicator and their product:

$$l_{im} = b_m + w_{m1} z_{im} + w_{m2} z_{im} q_{im} + w_{m3} q_{im} \qquad (2)$$

The product term lets the model shrink a modality's logit as its quality drops. It is trained on out-of-fold logits: the base models are fitted on three quarters of the training participants and applied to degraded copies of the held-out quarter (participant-grouped 4-fold). Degraded copies use random noise levels in {0, 1, 2, 3, 4} and random availability.

D. Stage 3: Mask-Aware Fusion

A logistic regression combines the recalibrated logits of the available modalities:

$$\mathrm{logit}\, p_i = \beta_0 + \Sigma_m a_{im} (\gamma_m l_{im} + \delta_m) \qquad (3)$$

Unavailable modalities contribute nothing, so no imputation is needed.

E. Ablations

Stack-noQ drops q from Stage 2. *Stack-avg* replaces Stage 3 with the mean of the recalibrated logits.

F. Conformal Sets

On calibration participants we compute the score $1 - p$(true class) and the finite-sample quantile at $\alpha = 0.1$. A class is in the set when its score is at most that quantile. We compare four groupings of the calibration scores: marginal; by availability pattern (the Mondrian approach of [13]–[15]); by tertile of the reliability index R_i, the mean q_{im} over available modalities; and the pair (pattern, tertile). A group with fewer than 20 calibration windows falls back to the marginal quantile.

IV. EXPERIMENTAL PROTOCOL

All data are synthetic. The simulator is a stand-in for a synchronized tri-modal recording and is built to have the

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

properties the question depends on: a shared latent stress state, participant-specific offsets, and noise that can be added per modality and per window.

A. Simulator

The simulator settings are listed in Table I. Training uses noise SD in {0, 1, 2, 3, 4}; SD 6 is outside that range. A fresh world (class directions), fresh participants and fresh splits are drawn in each of 30 repetitions. Splits are participant-disjoint, so no participant contributes to more than one of training, calibration and test.

TABLE I. SIMULATOR SETTINGS

Component	Setting
Participants per repetition	48 (26 train, 12 calibration, 10 test), 100 windows each
Labels	binary, 40% positive ("stress-related" proxy)
Window	8 frame vectors per modality; frame jitter SD 0.7
Feature dimension	face 16, voice 16, typing 8
Class separation (per modality)	face 1.6, voice 1.3, typing 0.8
Participant offset	isotropic Gaussian, SD 0.8, constant within a participant
Shared arousal	per-window gain 1 + 0.4 z on the class signal, common to all modalities
Degradation	Gaussian noise added to every frame, SD in {0, 2, 4, 6}; availability by random mask

B. Methods Compared

Table II lists the methods. All models are L2-regularised logistic regressions (C = 1). EF-aug is a deliberately strong baseline: it sees the same noise and dropout as RACF.

TABLE II. METHODS COMPARED

Method	Fusion	Training data
EF-clean	concatenate, zero-fill missing	clean windows
EF-aug	concatenate, zero-fill missing	noise- and dropout-augmented (3 copies)
LF-clean	mean of unimodal logits over available modalities	clean
LF-aug	same	augmented
Stack-noQ, Stack-avg	ablations of RACF (Section III)	augmented, out-of-fold
RACF	quality-conditioned recalibration + mask-aware fusion	augmented, out-of-fold

C. Test Scenarios

- *Grid:* each of seven non-empty availability patterns at noise SD 0, 2, 4 and 6 applied to all available modalities.
- *One corrupted modality:* all three available, one at SD 4 or 6, the others clean.
- *Mixed:* each modality independently available with probability 0.7 (at least one kept), noise SD drawn from {0, 2, 4, 6} with probabilities {0.4, 0.3, 0.2, 0.1}, independently per modality and window. The same generator produces the calibration windows for the conformal experiment.

D. Metrics and Statistics

We report macro-F1 and accuracy at threshold 0.5; ECE (10 equal-width confidence bins); and negative log-likelihood (NLL). For conformal sets we report marginal coverage, mean set size, singleton rate, accuracy of the top class on singleton sets, the largest absolute coverage deviation from 0.9 across availability patterns and across worst-modality noise levels (groups with at least 15 test windows), and the per-participant coverage spread. Intervals in figures are 95% CIs over the 30 repetitions. Paired differences use a two-sided Wilcoxon signed-rank test across repetitions; p-values are not corrected for multiple comparisons.

V. RESULTS

All values are means over 30 repetitions; ± is the SD across repetitions. Every number below is produced by the released script.

A. Clean Input: Modality Contribution and Fusion Gain

Face is the strongest modality, then voice, then typing, as set by the simulator's class separations; this ordering is a design choice and says nothing about real typing signals. Fusing all three raises macro-F1 from about 0.71 (best single modality) to 0.77 for early fusion and RACF. Averaging clean-trained unimodal logits (LF-clean) reaches only 0.736. When only the weakest modality is present, zero-filled early fusion falls to 0.516 (EF-aug) against 0.570 for RACF (Table III).

TABLE III. MACRO-F1 ON CLEAN TEST INPUT BY AVAILABILITY PATTERN

Method	F	V	T	F+V	F+T	V+T	F+V+T
EF-clean	0.702	0.649	0.544	0.754	0.724	0.683	0.770
EF-aug	0.700	0.638	0.516	0.755	0.723	0.678	0.773
LF-clean	0.711	0.666	0.589	0.743	0.709	0.670	0.736
LF-aug	0.706	0.652	0.557	0.716	0.677	0.625	0.684
RACF	0.706	0.657	0.570	0.753	0.725	0.692	0.770

F = face, V = voice, T = typing. Means over 30 repetitions; SD across repetitions ranges from 0.020 to 0.061 and is omitted for space.

B. Uniform Noise on All Modalities

Fig. 1 shows all three modalities available as noise grows. Macro-F1 falls for every method, from about 0.77 to about 0.65 at SD 6. EF-aug is nominally highest at each level (0.773, 0.748, 0.699, 0.659 against RACF 0.770, 0.741, 0.691, 0.649), with differences of at most 0.010. Calibration separates the methods. RACF stays at ECE 0.044, 0.042, 0.041 and 0.050 from SD 0 to SD 6. EF-aug is 0.045, 0.037, 0.086 and 0.153; EF-clean reaches 0.215. LF-clean has a high ECE on clean input (0.102) that falls as noise grows; we did not examine the direction of its miscalibration.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

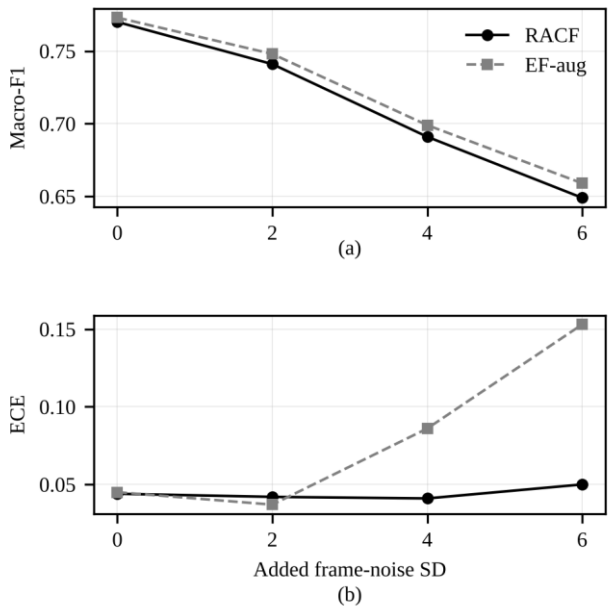


Fig. 1. Macro-F1 (a) and ECE (b) with all three modalities available, against added frame-noise SD (means over 30 repetitions). SD 6 is outside the training range.

The picture is less favourable when averaged over all seven availability patterns at high noise: at SD 6 macro-F1 is 0.595 for RACF against 0.615 for EF-aug, and at SD 4 it is 0.626 against 0.640, while RACF is better at SD 0 (0.696 against 0.683). We did not isolate the cause. Candidates are extrapolation of the recalibration's product term beyond the training noise range and over-shrinking of the logits when all inputs are poor; both are untested here.

C. *One Corrupted Modality*

Fig. 2 corrupts one modality and leaves the other two clean. RACF has the highest macro-F1 in five of six conditions: 0.737 against 0.728 (face, SD 4), 0.717 against 0.690 (face, SD 6), 0.747 against 0.738 (voice, SD 4), 0.742 against 0.711 (voice, SD 6) and 0.754 against 0.742 (typing, SD 6) for EF-aug. With typing at SD 4, EF-aug is level (0.761 against 0.760). RACF's ECE is 0.041 to 0.046 in all six conditions; EF-aug's ranges from 0.037 to 0.101.

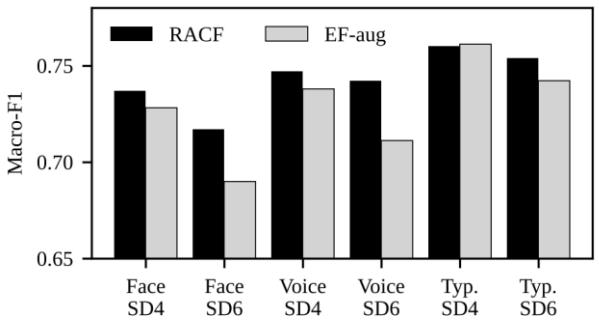


Fig. 2. Macro-F1 with one modality corrupted (SD 4 or 6) and the other two clean; RACF against EF-aug, mean over 30 repetitions.

D. *Mixed Missing and Noisy Input*

TABLE IV. MIXED SCENARIO (RANDOM AVAILABILITY AND NOISE)

Method	Acc.	Macro-F1	ECE	NLL
EF-clean	0.708	0.686	0.102	0.628
EF-aug	0.713	0.687	0.046	0.562
LF-clean	0.710	0.678	0.043	0.568
LF-aug	0.703	0.653	0.059	0.575
Stack-noQ	0.711	0.681	0.040	0.558
Stack-avg	0.702	0.644	0.068	0.576
RACF	**0.717**	**0.688**	**0.037**	**0.544**

Means over 30 repetitions; SD across repetitions ranges from 0.007 to 0.041.

RACF has the best value on all four metrics, but the accuracy-type differences from EF-aug are not distinguishable: macro-F1 +0.001 ($p = 0.44$, RACF higher in 16 of 30 repetitions). The calibration gains are: ECE −0.009 ($p = 0.004$) and NLL −0.017 ($p < 0.001$; lower in 29 of 30) against EF-aug, and NLL −0.084 against EF-clean. Against LF-clean, macro-F1 is +0.010 ($p < 0.001$). Against the ablation without the quality term, macro-F1 is +0.007 ($p < 0.001$) and NLL −0.014 ($p < 0.001$), while the ECE difference is −0.003 ($p = 0.07$). The quality term therefore matters mainly for likelihood and for corrupted-modality cases, and the learned fusion weights matter for accuracy (Stack-avg is 0.044 lower in macro-F1).

E. *Conformal Sets (Target Coverage 0.9)*

TABLE V. CONFORMAL PREDICTION SETS IN THE MIXED SCENARIO

Base	Grouping	Cov.	Size	Single.	Gap (m)	Gap (n)
RACF	marginal	0.899	1.436	0.564	0.068	0.035
RACF	availability	0.900	1.446	0.554	0.072	0.034
RACF	reliability	0.899	1.438	0.562	0.068	0.037
RACF	avail. + rel.	0.907	1.471	0.529	0.069	0.036
EF-aug	marginal	0.898	1.452	0.548	0.068	0.061
EF-aug	availability	0.902	1.466	0.534	0.075	0.056
EF-aug	reliability	0.898	1.447	0.553	0.066	0.044
EF-aug	avail. + rel.	0.906	1.487	0.513	0.066	0.040

Cov. = marginal coverage; Size = mean set size; Single. = singleton rate; Gap (m) and Gap (n) = largest absolute deviation of group coverage from 0.9 over availability patterns (mask) and worst-modality noise levels (noise). Reliability = reliability tertile.

Marginal coverage is at the target for every grouping. No set was empty. With RACF, conditioning on availability or reliability does not change the gaps (all paired $p > 0.3$) and availability plus reliability enlarges the sets by 0.035 ($p < 0.001$). The gaps themselves are about what sampling noise alone produces: a Monte-Carlo check with exact 0.9 coverage

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

and the same group sizes gives a mean gap of 0.051 across availability patterns (5th to 95th percentile 0.026 to 0.086) and 0.029 across noise levels (0.012 to 0.051). With EF-aug, whose probabilities are not reliability-calibrated, reliability conditioning reduces the noise-level gap from 0.061 to 0.044 ($p < 0.001$), and availability plus reliability to 0.040 ($p < 0.001$). In other words, reliability-aware conditioning of the conformal step substitutes for reliability-aware probabilities; it is not additional to them. Per-participant coverage varies: the SD across the ten test participants is 0.04, and about 4% of participants (RACF, marginal) fall below 0.8, which window-level conformal coverage does not control.

VI. DISCUSSION

What the evidence supports. In this simulator, a cheap quality indicator is enough to keep a fused probability calibrated across a wide noise range, including noise outside the training range, and to help most when one modality fails while the others stay clean. The mechanism is plain: the recalibration shrinks a modality's logit when its jitter rises, so a corrupted stream stops dominating the sum. Noise augmentation alone (EF-aug) recovers much of the accuracy but not the calibration at high noise.

What it does not support. RACF is not a more accurate classifier than augmented early fusion in the mixed scenario, and it is somewhat worse when all modalities are severely corrupted. A reader who needs the best accuracy under uniform severe noise should prefer EF-aug in this setting. The accuracy advantage of RACF over LF-clean and over the ablations is real but small (0.007 to 0.010 macro-F1).

Conformal conditioning. Conditioning conformal calibration on availability is established [13]–[15]. Our added grouping by estimated reliability helps only when the probabilities are not already reliability-calibrated, as with EF-aug, and is redundant with RACF. This suggests a practical order of work: fix calibration at the probability level first, then check group coverage; conformal grouping is a patch for miscalibration across groups, not a substitute for it.

A caveat on the quality indicator. Frame jitter detects the added noise by construction. Real corruptions differ: occlusion changes the signal without raising frame-to-frame variance, and a noisy room may not change typing at all. The result shows that a well-behaved indicator is useful; it does not show that one exists for each real modality. Testing that is the main next step.

Participant-level coverage. Window-level conformal coverage averages over participants. With ten test participants the per-participant coverage spread is 0.04 and a few participants fall well below target. Any deployment claim should report this spread.

VII. LIMITATIONS AND ETHICS

- *Synthetic data only.* No human participants were involved, and the results are not evidence about real faces, voices or typing. The simulator's class separations, correlations and noise model are assumptions; different choices could change the ranking of methods.
- *Label proxy.* In real data, emotion or task labels are proxies for stress, not validated stress measurements. The framework is for non-clinical research. It must not be used to diagnose conditions, screen employees, or make high-stakes decisions.
- *Linear models only.* Deep encoders for images and audio were not used; the claim concerns the fusion and calibration layer.
- *Exchangeability.* Conformal guarantees assume calibration and test windows are exchangeable at the group level. Calibration here uses separate participants, but windows within a participant are correlated, and the guarantee is marginal over participants, not per participant.
- *Statistics.* Thirty repetitions with uncorrected paired tests; p-values describe this simulator, not a population of studies.
- *Ethics.* Real-data work needs ethics approval where human participants are recorded: informed consent, secure storage, a withdrawal procedure, and timing-only keystroke capture without typed content.

VIII. CONCLUSION AND FUTURE WORK

A quality-conditioned recalibration step kept fused probabilities calibrated across noise and missing-input conditions in a controlled tri-modal simulation, with the clearest accuracy gain when a single modality is corrupted, and with no accuracy gain over augmented early fusion in a mixed scenario. Conformal grouping by availability and reliability added little once the probabilities were reliability-calibrated.

The next steps, in order, are: (1) replace the simulator with real features from a dataset that has face, voice and physiological streams for the same people, for example StressID [2], with physiology standing in for typing; (2) define a quality indicator for each real modality and test whether it predicts error; (3) collect a small, ethically approved, synchronized face, voice and keystroke-timing set to test the typing claim directly; (4) add participant-level adaptive calibration, since per-participant coverage varied; (5) diagnose the loss at severe uniform noise.

Reproducibility. racf_sim.py generates the data, trains all methods and writes every metric; analyze.py prints the tables and paired tests; plot.py draws the figures. Thirty repetitions run in under a minute on two CPU cores with NumPy and scikit-learn.

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. ACM ICMI*, 2018.

[2] H. Chaptoukaev *et al.*, "StressID: a multimodal dataset for stress identification," *NeurIPS Datasets and Benchmarks Track*, 2023.

[3] A. Iranfar, A. Arza and D. Atienza, "ReLearn: a robust machine learning framework in presence of missing data for multimodal stress detection from physiological signals," arXiv:2104.14278, 2021.

[4] H. Yu, T. Vaessen, I. Myin-Germeys and A. Sano, "Modality fusion network and personalized attention in momentary stress detection in the wild," in *Proc. ACII*, 2021, doi:10.1109/ACII52823.2021.9597459.

[5] "ADAPT: multimodal learning for detecting physiological changes under missing modalities," arXiv:2407.03836, 2024.

[6] M. Ma *et al.*, "Are multimodal transformers robust to missing modality?," in *Proc. CVPR*, 2022.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

[7] Q. Zhang *et al.*, "Provable dynamic fusion for low-quality multimodal data," in *Proc. ICML*, 2023.

[8] "Predictive dynamic fusion," arXiv:2406.04802, 2024.

[9] "Unbiased dynamic multimodal learning (UDML)," arXiv:2603.19681, 2026.

[10] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: representing model uncertainty in deep learning," in *Proc. ICML*, 2016, arXiv:1506.02142.

[11] C. Guo, G. Pleiss, Y. Sun and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. ICML*, 2017, arXiv:1706.04599.

[12] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," arXiv:2107.07511.

[13] M. Zaffran *et al.*, "Conformal prediction with missing values," in *Proc. ICML*, PMLR vol. 202, 2023.

[14] A. Moayedikia, "Conformal fusion under missing modalities," arXiv:2608.07183, 2026.

[15] I. Azizi, "Conformal calibration for multi-modal regression with missing modalities," PMLR vol. 329, 2026, arXiv:2608.07795.

[16] "How you type is what you type: keystroke dynamics correlate with affective content," in *Proc. ACII*, 2019, doi:10.1109/ACII.2019.8925460.

[17] S. Unni, S. Gowda and A. F. Smeaton, "An investigation into keystroke dynamics and heart rate variability as indicators of stress," Dublin City University, 2022.

[18] "Towards detecting programmers' stress on the basis of keystroke dynamics," 2016, doi:10.15439/2016F263.

[19] M. M. Khubrani, "Multimodal Stress-Aware Ensemble Learning (MSAEL): tri-modal behavioral-physiological fusion for real-time proxy-based stress recognition integrating typing, facial, and voice dynamics," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 38, art. 405, 2026, doi:10.1007/s44443-026-00765-9.

[20] A. Heimerl *et al.*, "The ForDigitStress dataset: a multi-modal dataset for automatic stress recognition," *IEEE Trans. Affective Computing*, 2024, doi:10.1109/TAFFC.2024.3501400.

[21] M. Hosseini *et al.*, "A multimodal stress detection dataset with facial expressions and physiological signals," *Sci. Data*, vol. 12, art. 1844, 2025, doi:10.1038/s41597-025-05812-0.

[22] "Accommodating missing modalities in time-continuous multimodal emotion recognition," arXiv:2311.10119, 2023.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)