



iJRASET

International Journal For Research in
Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IX **Month of publication:** September 2026

DOI: <https://doi.org/10.22214/ijraset.2026.84857>

www.ijraset.com

Call: ☎ 08813907089

E-mail ID: ijraset@gmail.com

Multi-Method Explainable Deep Learning for Acute Stress Analysis using Physiological Signals

Nanda Yasaswini¹, Setti Sarika²

¹M.Tech, Department of Information Technology and Computer Applications, Andhra University College of Engineering, Visakhapatnam, India

²Research Scholar, Department of Computer Science and Systems Engineering, Andhra University College of Engineering, Visakhapatnam, India

Abstract: Wearable devices measure body signals such as skin conductance, pulse, skin temperature and movement, and machine-learning models can use these signals to detect acute stress. Recent work has added explainability so that a user can see which signal influenced a decision. However, most studies use only a single explanation method, judge that explanation visually, and rarely report how confident the model is. This raises a practical question: can these explanations be trusted? This paper studies that question on the WESAD dataset under strict subject-independent (Leave-One-Subject-Out) evaluation. Five deep models (LSTM, BiLSTM, GRU, CNN-LSTM, Transformer) and five classical models are trained on identical folds. Explanations are generated with three methods (Integrated Gradients, SHAP, LIME) and their agreement is measured across six held-out participants; explanations are further examined with faithfulness tests, cross-architecture comparison, counterfactual and temporal analysis; and predictive uncertainty is estimated with deep ensembles and assessed for calibration. Three findings are reported. First, the five deep architectures perform almost identically (0.846–0.863 accuracy) and are statistically indistinguishable (all Holm-corrected $p = 1.00$, small effect sizes), while a classical Random Forest is more accurate (0.911); after Holm–Bonferroni correction no individual classical-vs-deep pair is significant at 15 subjects, but the direction is consistent across every comparison, with medium-to-large effect sizes for nine of the ten comparisons (rank-biserial 0.50–0.68) and a smaller effect for the tenth (Logistic Regression vs Transformer, +0.25), so at this data scale increasing architectural complexity is not the productive direction. Second, across six subjects the three explanation methods agree only partially, heart rate is the single most important channel in two to four of six subjects depending on the method, and pairwise method agreement is low-to-moderate and highly variable (Spearman ρ from 0.10 ± 0.45 to 0.70 ± 0.20), which shows that a single explanation method is insufficient on its own and that reliability must be measured rather than assumed; faithfulness is in the expected direction (mean insertion AUC 0.794 > deletion AUC 0.756) but modest and subject-dependent. Third, the ensemble is well calibrated (Expected Calibration Error 0.065) and, under a coverage–risk (deferral) analysis, answering only the most-confident 50% of cases raises selective accuracy from 0.884 to 0.960. The study concludes that trustworthiness, measured explanation reliability together with calibration, rather than accuracy alone, is the appropriate objective for wearable stress detection.

Keywords: Explainable Artificial Intelligence, Acute Stress Detection, Wearable Sensors, WESAD, Deep Learning, Integrated Gradients, SHAP, LIME, Model Calibration, Coverage–Risk Analysis, Leave-One-Subject-Out Cross-Validation.

I. INTRODUCTION

Stress is the body's response to demanding situations. Short-term (acute) stress activates the sympathetic nervous system and produces measurable physical changes: the heart beats faster, the sweat glands become more active, and skin temperature at the fingers falls slightly. A small amount of stress can improve alertness, but repeated or excessive stress harms concentration, decision making and long-term health. Detecting stress early is therefore useful in healthcare, workplace safety and personal wellbeing.

Traditionally stress was measured with questionnaires, which depend on self-judgement, are collected after the event, and cannot be used continuously. Wearable devices remove these limitations because they record physiological signals continuously and comfortably, and a large number of studies now apply machine and deep learning to classify stress automatically.

Most of these studies report high accuracy, yet two practical requirements are usually unmet. The first is explanation: a deep model has many parameters, and when it outputs 'stress' there is no direct way to see why; if it responds to movement rather than physiological arousal, the output is not medically meaningful. The second is confidence: a model that is confidently wrong is dangerous in a health setting, and neural networks are known to be overconfident.

Recent work has begun to address the first requirement by adding explainability; in particular, the base study for this work applied Integrated Gradients to an LSTM model and observed that the highlighted features matched physiological expectations. This is an important step, but it has a limitation: when only one explanation method is used on a single participant, there is no way to check whether the explanation itself is reliable, because a second method, or a different participant, could produce a different answer.

This paper therefore asks a different question: not whether stress can be detected, but whether stress-detection models and their explanations can be *trusted*. The contributions are:

- 1) Five deep architectures and five classical models are compared on identical subject-independent folds, with multiple-comparison correction and effect sizes, showing that the choice of architecture has almost no effect and that a classical model is at least as strong at this data size.
- 2) Three explanation methods are applied to the same trained model and their agreement is measured quantitatively across six held-out participants, demonstrating that a single-method explanation is insufficient on its own and that explanation reliability must be measured, not assumed.
- 3) Explanations are verified rather than assumed, using deletion/insertion faithfulness tests, cross-architecture consistency, counterfactual analysis and temporal occlusion.
- 4) Predictive uncertainty is estimated and calibrated, and a coverage-risk (deferral) analysis shows how selective prediction raises accuracy by abstaining on low-confidence cases.

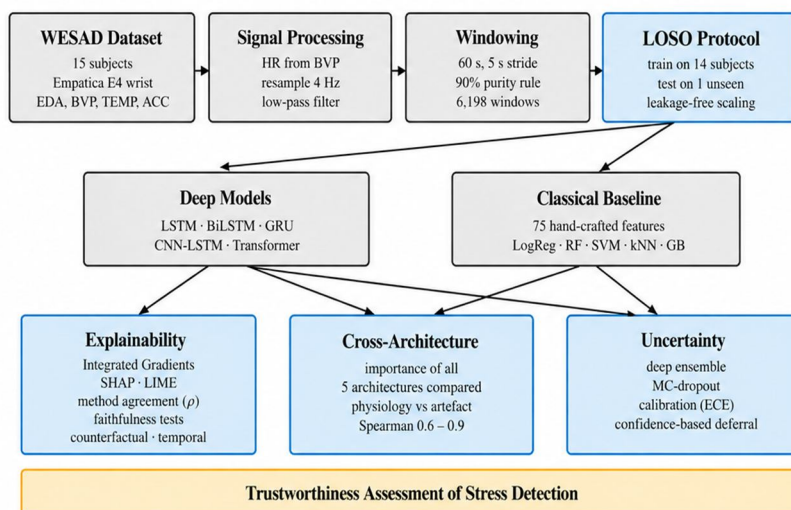


Fig. 1. Overall framework of the proposed trustworthiness assessment.

II. RELATED WORK

A. Machine-learning and deep-learning approaches

Early stress-detection systems used classical machine learning with hand-designed features, SVMs, Random Forests, decision trees, k-NN and boosting, reaching accuracies above 95% in some reports. These methods need domain knowledge for feature design but are stable on small datasets and relatively interpretable. Deep learning removes manual feature design: CNNs capture local signal shapes and recurrent/LSTM networks model temporal change, with reported accuracies frequently above 90%. However, results vary greatly with the validation protocol, and cross-study comparison is difficult because protocols differ.

B. Explainability in stress detection

Explainable-AI methods such as Integrated Gradients, SHAP and LIME estimate how much each input contributes to a prediction. The base study applied Integrated Gradients to an LSTM ensemble and showed that important features agreed with physiological knowledge. Surveys identify interpretability and cross-subject validation as the main open problems. Table 1 positions the present work: the distinguishing feature is not the detection model but the verification of the explanation across methods and subjects, and the reporting of calibrated uncertainty.

Table 1. Positioning of this work relative to representative studies.

Study	Models	Validation	Explanation methods	Explanation verified	Uncertainty
Schmidt et al. (2018)	Classical ML	LOSO	None	No	No
Bobade & Vani (2020)	ML and DL	Mixed-subject	None	No	No
Moser et al. (2024)	LSTM ensemble	Window-level	Integrated Gradients	Visual only	No
This work	5 deep + 5 classical	Strict LOSO	IG, SHAP, LIME	Faithfulness + cross-arch. + multi-subject	ECE + coverage-risk

C. Identified gap

Three gaps follow. First, explanations are usually produced by one method and validated visually on one subject, so their reliability is unknown. Second, uncertainty and calibration are rarely reported, although they determine safe deployment. Third, many studies mix data from the same subject across training and testing, inflating accuracy. This work addresses all three.

III. PROPOSED METHODOLOGY

A. Dataset

The WESAD (Wearable Stress and Affect Detection) dataset is used. It contains recordings from 15 participants (S2–S17, excluding S12) wearing an Empatica E4 wrist device across four controlled conditions: baseline, stress, amusement and meditation. Four wrist signals are used directly, EDA, BVP, skin temperature (TEMP) and three-axis acceleration (ACC), and a fifth channel, heart rate (HR), is derived from BVP by band-pass filtering (0.5–8 Hz), peak detection and inter-beat-interval (IBI) conversion, as in (1). Acceleration is retained as a control: if the model relied mainly on movement, this would indicate activity rather than stress.

Table 2. Summary of the processed dataset.

Property	Value
Participants	15 (S2–S17, excluding S12)
Device	Empatica E4 wrist-worn sensor
Channels used	EDA, BVP, TEMP, ACC, HR (derived)
Common sampling rate	4 Hz
Window length / stride	60 s / 5 s
Samples per window	240 time steps $\times$ 5 channels
Label purity threshold	90%
Total labelled windows	6,198
Class distribution	$\approx$ 30% stress, $\approx$ 70% non-stress

B. Preprocessing and the independence of samples

All channels are resampled to 4 Hz; EDA is low-pass filtered at 1 Hz and temperature at 0.1 Hz. The recording is divided into 60-second windows with a 5-second stride, and a window is kept only if at least 90% of its samples share one condition. Each channel is standardised as in (2), where μ and σ are computed from the current fold's training participants only.

Because the stride (5 s) is much smaller than the window (60 s), consecutive windows overlap heavily and are therefore not statistically independent. Overlap increases the number of training examples but does not increase the number of independent units of observation, which remains the number of participants (15).

For this reason all statistical tests in this paper are computed on the 15 per-subject scores (one value per held-out participant), not on pooled windows. This is the correct unit of analysis and, as discussed in Section 5, it also means the study has inherently limited statistical power.

C. Evaluation protocol

Leave-One-Subject-Out (LOSO) cross-validation is used throughout: for each participant the model is trained on the other 14 and tested on the held-out participant, and results are averaged over the 15 folds. LOSO is essential because a random split would let the model recognise the individual (through overlapping, near-identical windows) rather than the stress state. The full procedure is given in Algorithm 1.

```
Algorithm 1. Trustworthiness assessment protocol
Input : per-subject windows {X_s, y_s} for s = 1..15
for each subject s in 1..15 do
    train_set <- all subjects except s
    mu, sigma <- mean and std of train_set only
    standardise train_set and X_s using mu, sigma
    hold out 2 whole subjects as validation
    train with class-weighted loss + early stopping
    evaluate on X_s -> accuracy, precision, recall, F1
end for; report mean and std over the 15 folds
for six held-out subjects used for explanation do
    A_IG, A_SHAP, A_LIME <- channel attributions
    rho <- Spearman correlation between each pair
    run deletion / insertion faithfulness tests
    compute counterfactual and temporal occlusion
end for; aggregate agreement + faithfulness over subjects
train K = 5 seeds -> deep ensemble
compute predictive entropy, ECE, coverage-risk
```

D. Models

Five deep architectures are implemented, LSTM, BiLSTM, GRU, CNN-LSTM (a 1-D convolution before the recurrent layer) and a self-attention Transformer. Five classical models are trained on the same folds using 15 hand-designed features per channel (75 per window), covering time-domain statistics (mean, std, min, max, range, median, IQR, skewness, kurtosis, slope, Hjorth mobility, RMS) and frequency-domain features (dominant frequency, spectral energy, low-to-high power ratio): Logistic Regression, Random Forest, SVM (RBF), k-NN and Histogram Gradient Boosting.

E. Explainability framework

Four distinct properties are evaluated, and it is important to keep them separate because they answer different questions:

- Explanation agreement: do different attribution methods rank the input channels similarly? (Spearman correlation between method rankings).
- Faithfulness: does the channel an explanation nominates actually drive the prediction? (deletion/insertion curves, comprehensiveness, sufficiency).
- Calibration: do the model's confidence values match its empirical accuracy? (Expected Calibration Error).
- Trustworthiness: the overall property, supported only when agreement, faithfulness and calibration are jointly examined and reinforced by cross-architecture and physiological consistency.

Three attribution methods are applied to the same trained network. Integrated Gradients accumulates the model-output gradient along a straight path from a baseline x' to the input x , as in (3). We use the **standardised zero vector** as the baseline: because each channel is z-scored per fold, zero corresponds to the training-set mean of that channel, a physiologically neutral 'average' input, the natural reference for attributing the deviations that indicate stress. Baseline choice can influence attributions, noted as a limitation; however IG and LIME agree perfectly on the illustrative subject ($\rho = 1.00$) and IG agrees strongly with SHAP on average ($\rho = 0.70$), indicating the rankings are not a baseline artefact. Method rankings are compared with the Spearman rank correlation, (4).

Faithfulness is measured by comprehensiveness (5) and sufficiency (6); counterfactuals search for the smallest additive change δ that flips a stress prediction to non-stress, (7).

Configurations (for reproducibility). Integrated Gradients: Captum, 50 integration steps, zero baseline. SHAP: GradientExplainer with a 100-window background sampled from the fold's training set, stress-class attributions, averaged over up to 64 stress windows. LIME: LimeTabularExplainer in classification mode, continuous features without discretisation, 500 perturbation samples per instance, explaining the flattened 240×5 input and aggregating to per-channel importance over 20 instances.

F. Uncertainty estimation

A deep ensemble trains the same architecture $K = 5$ times from different initialisations and averages the probabilities, (8); Monte-Carlo dropout keeps dropout active at test time. Predictive entropy (9) and the Expected Calibration Error (10) are computed from the averaged probabilities, where B_m are confidence bins.

IV. EXPERIMENTAL SETUP

Experiments used PyTorch (deep models), scikit-learn (classical models) and Captum, SHAP and LIME. A class-weighted cross-entropy loss was used because the stress class is $\approx 30\%$ of windows; two whole participants were held out from each training pool for validation, with early stopping on validation macro-F1. Experiments ran on an NVIDIA T4 GPU. Performance uses accuracy (11), precision (12), recall (13) and F1 (14).

Table 3. Training configuration.

Parameter	Value	Parameter	Value
Hidden units	64	Optimiser	AdamW
Recurrent layers	2	Learning rate	0.001
Dropout rate	0.3	Weight decay	0.0001
Batch size	64	Maximum epochs	40
Early-stopping patience	8	Ensemble size (K)	5
Validation subjects	2	Random seed	42

V. RESULTS AND DISCUSSION

A. Comparison of deep architectures

Table 4 reports the five deep architectures over the 15 LOSO folds (Fig. 2). They lie within 1.7 accuracy points of one another, and the per-participant standard deviation (≈ 0.10) is far larger than the between-model differences. A Wilcoxon signed-rank test on the paired per-subject accuracies of every architecture pair, with Holm-Bonferroni correction across the ten comparisons, finds no significant difference (all corrected $p = 1.00$) and only small effect sizes ($|\text{rank-biserial}| \leq 0.30$). The architectures are therefore statistically indistinguishable on this dataset; notably the plain LSTM used in the base study obtains the lowest F1.

Table 4. Deep-architecture comparison (LOSO, mean $\pm$ standard deviation).

Model	Accuracy	Precision	Recall	F1 score
LSTM	0.848 ± 0.109	0.758 ± 0.263	0.776 ± 0.362	0.702 ± 0.281
BiLSTM	0.863 ± 0.096	0.761 ± 0.266	0.813 ± 0.309	0.745 ± 0.246
GRU	0.846 ± 0.103	0.779 ± 0.172	0.821 ± 0.293	0.736 ± 0.215
CNN-LSTM	0.855 ± 0.100	0.822 ± 0.195	0.817 ± 0.264	0.762 ± 0.165
Transformer	0.856 ± 0.137	0.793 ± 0.246	0.805 ± 0.259	0.772 ± 0.209

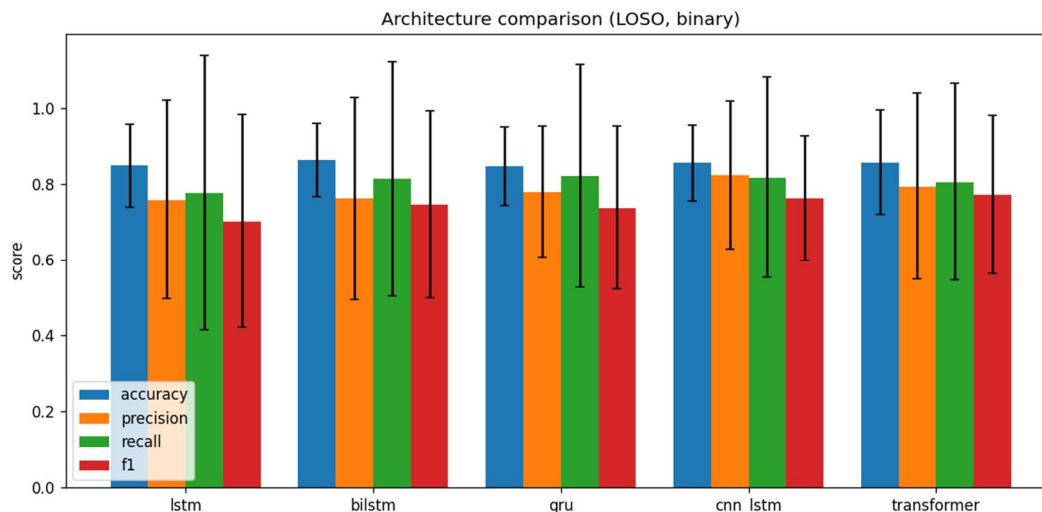


Fig. 2. Performance of the five deep architectures under LOSO cross-validation.

B. Statistical comparison with the classical baseline

Table 5 compares the strongest classical models against each deep architecture on the same 15 paired folds, reporting the raw Wilcoxon p-value, the Holm-corrected p-value, and the matched-pairs rank-biserial effect size. Two points follow. (i) Random Forest is more accurate than every deep model and Logistic Regression obtains a higher F1 than every deep model, with medium-to-large effect sizes for nine of the ten comparisons (rank-biserial 0.50–0.68) and a smaller effect (+0.25) for Logistic Regression vs Transformer. (ii) However, once the multiple comparisons are corrected, no individual pair remains significant at $\alpha = 0.05$ (Holm $p = 0.181$ – 0.421), as expected with only 15 independent units. We therefore do not claim per-pair significance; the consistent direction across every comparison together with the medium-to-large effect sizes supports the conclusion that, on this dataset, a well-regularised classical model is at least as strong as the deep models (Fig. 3).

Table 5. Classical-vs-deep comparison (Wilcoxon signed-rank, Holm–Bonferroni corrected, with rank-biserial effect sizes; 15 paired folds).

Comparison	Metric	Mean diff.	Raw p	Holm p	Effect (rbc)	Sig.
RF vs LSTM	Acc	+0.063	0.022	0.194	+0.67	No
RF vs BiLSTM	Acc	+0.049	0.048	0.248	+0.58	No
RF vs GRU	Acc	+0.065	0.026	0.205	+0.65	No
RF vs CNN-LSTM	Acc	+0.056	0.041	0.248	+0.62	No
RF vs Transformer	Acc	+0.055	0.096	0.289	+0.50	No
LogReg vs LSTM	F1	+0.152	0.030	0.211	+0.63	No
LogReg vs BiLSTM	F1	+0.108	0.018	0.181	+0.68	No
LogReg vs GRU	F1	+0.117	0.041	0.248	+0.60	No
LogReg vs CNN-LSTM	F1	+0.092	0.096	0.289	+0.50	No
LogReg vs Transformer	F1	+0.081	0.421	0.421	+0.25	No

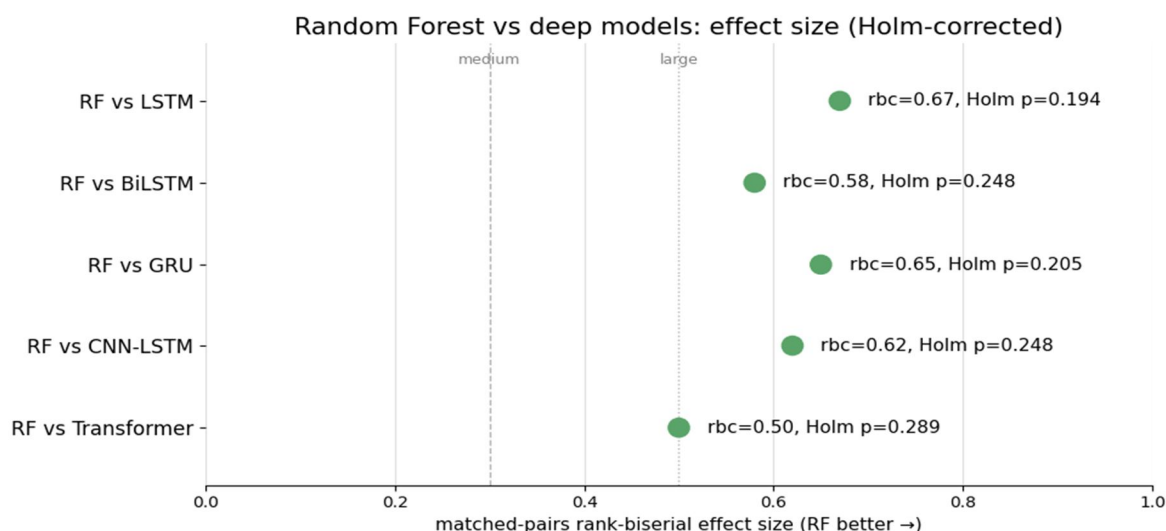


Fig. 3. Random Forest vs deep models: matched-pairs rank-biserial effect size (all Holm-corrected $p > 0.05$; direction consistently favours Random Forest).

C. Classical machine-learning baseline

Table 6. Classical machine-learning baseline (LOSO, mean $\pm$ standard deviation).

Model	Accuracy	Precision	Recall	F1 score
Random Forest	0.911 $\pm$ 0.088	0.858 $\pm$ 0.259	0.797 $\pm$ 0.272	0.811 $\pm$ 0.245
Logistic Regression	0.895 $\pm$ 0.117	0.842 $\pm$ 0.191	0.911 $\pm$ 0.132	0.853 $\pm$ 0.138
SVM (RBF)	0.895 $\pm$ 0.083	0.848 $\pm$ 0.158	0.809 $\pm$ 0.216	0.811 $\pm$ 0.171
k-Nearest Neighbours	0.870 $\pm$ 0.081	0.810 $\pm$ 0.164	0.809 $\pm$ 0.170	0.787 $\pm$ 0.132
Gradient Boosting	0.870 $\pm$ 0.105	0.785 $\pm$ 0.271	0.794 $\pm$ 0.269	0.766 $\pm$ 0.237

Random Forest reaches 0.911 accuracy and Logistic Regression an F1 of 0.853. A plausible explanation is dataset size: deep networks need many subjects, whereas WESAD has only 15. On this 15-subject dataset and in this binary-classification setting, carefully designed statistical features with a well-regularised classical model generalise at least as well to an unseen person. We report this as an observation at this data scale rather than a general property of deep models: with more subjects the picture could change, so the finding is explicitly restricted to the evaluated dataset and configuration and is presented as a reason to direct effort towards reliability and interpretability rather than architectural novelty.

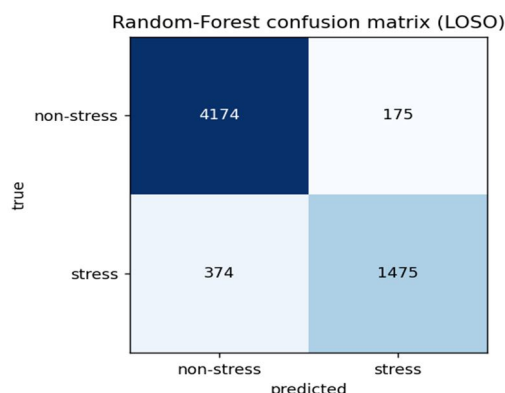


Fig. 4. Confusion matrix of the Random Forest classifier (pooled over all folds).

D. Agreement between explanation methods (six subjects)

Explanations were generated for the CNN-LSTM model on six held-out participants (S2, S5, S6, S9, S13, S16), spanning easy-to-hard folds. Table 7 reports the illustrative per-channel importance on subject S2, and Table 8 reports agreement and faithfulness aggregated across the six subjects.

Table 7. Illustrative normalised channel importance by method (CNN-LSTM, subject S2).

Method	EDA	BVP	TEMP	ACC	HR	Highest
Integrated Gradients	0.19	0.12	0.19	0.12	0.38	HR
SHAP	0.11	0.14	0.18	0.12	0.45	HR
LIME	0.21	0.18	0.19	0.19	0.23	HR

Table 8. Explanation agreement and faithfulness aggregated over six held-out participants (CNN-LSTM).

Quantity	Mean $\pm$ SD across 6 subjects
Agreement: IG vs SHAP (Spearman ρ)	0.70 ± 0.20
Agreement: IG vs LIME (Spearman ρ)	0.10 ± 0.45
Agreement: SHAP vs LIME (Spearman ρ)	0.45 ± 0.42
Comprehensiveness (drop when top channel removed)	0.028 ± 0.040
Sufficiency	-0.144 ± 0.267
Deletion AUC (lower = more faithful)	0.756 ± 0.216
Insertion AUC (higher = more faithful)	0.794 ± 0.211
HR is the single most important channel	IG 3/6 · SHAP 4/6 · LIME 2/6 · faithfulness 4/6

The multi-subject view is more nuanced than any single participant suggests. Heart rate is the *most frequently* dominant channel, the single most important signal in two to four of the six subjects depending on the method, but not universal. Method agreement is low-to-moderate and highly variable: IG and SHAP agree reasonably (0.70 ± 0.20), whereas IG and LIME barely agree on average (0.10 ± 0.45) despite agreeing perfectly on S2, and SHAP–LIME agreement is moderate but unstable (0.45 ± 0.42). This variability is itself the key result: it demonstrates directly, across subjects, that a single explanation method on a single participant cannot be treated as reliable, and that explanation agreement must be measured.

E. Faithfulness of the explanations

Across the six subjects the mean insertion AUC (0.794) exceeds the mean deletion AUC (0.756), the expected direction for a faithful explanation, but the margin is small and comprehensiveness is modest on average (0.028 ± 0.040) with substantial subject-to-subject variation. Subject S2 (Fig. 5, Table 9) is a favourable case in which the nominated channel carries a fifth of the decision (comprehensiveness 0.201); it is retained as an illustrative example, but Table 8 shows such strong faithfulness is not typical of every participant. The honest reading is that the explanations point in the right direction on average, yet their faithfulness is uneven, again supporting measuring reliability rather than assuming it.

Table 9. Faithfulness on the illustrative subject S2 (CNN-LSTM).

Measure	Value	Interpretation
Deletion AUC	0.634	Lower indicates a more faithful explanation
Insertion AUC	0.693	Higher than deletion, i.e. the expected direction
Comprehensiveness	0.201	Removing HR reduces stress probability by 0.201
Sufficiency	-0.464	HR alone yields a higher probability than all channels

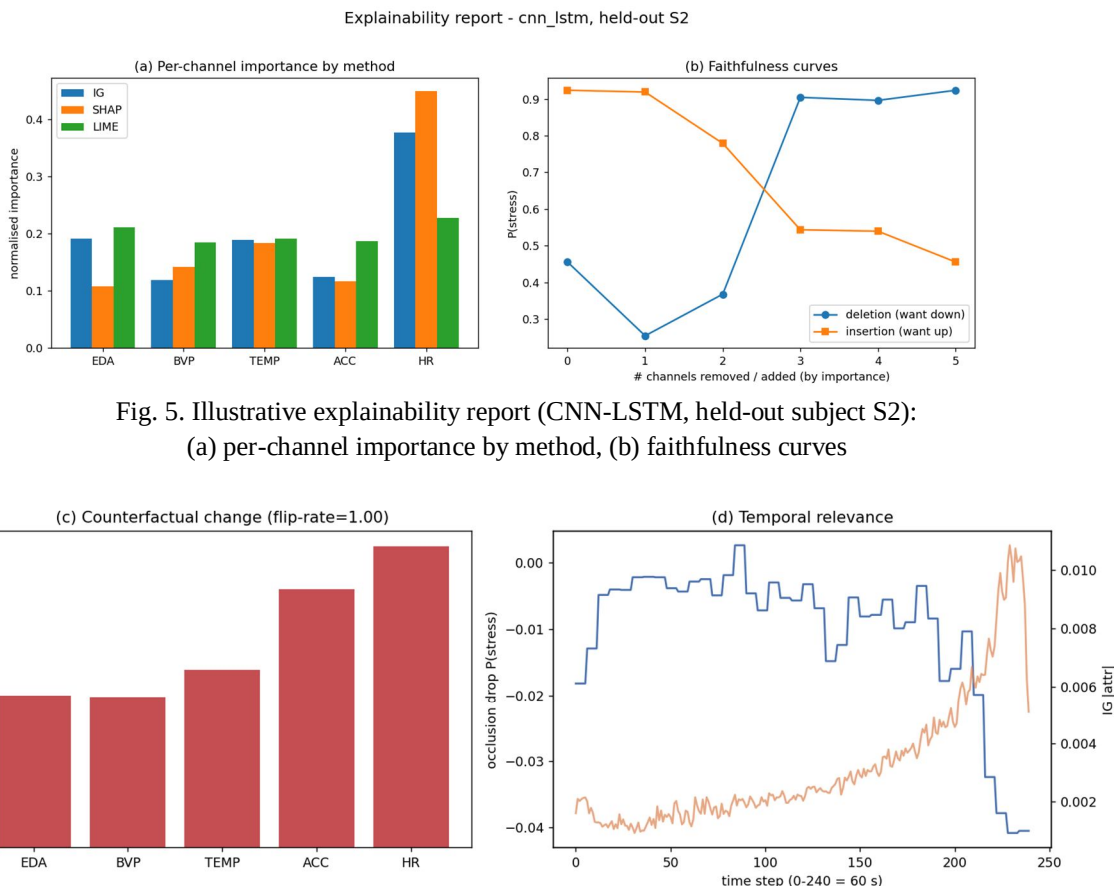


Fig. 5. Illustrative explainability report (CNN-LSTM, held-out subject S2):
(a) per-channel importance by method, (b) faithfulness curves

Fig. 5. Illustrative explainability report (CNN-LSTM, held-out subject S2):
(c) counterfactual channel change, (d) temporal relevance.

F. Consistency across architectures

Table 10 reports the Integrated Gradients channel importance for all five architectures on the same held-out participant. Skin temperature is among the two most important channels for every architecture, and heart rate joins it for four of the five (the exception is CNN-LSTM, where electrodermal activity ranks second); pairwise Spearman correlations range from 0.60 to 0.90 (mostly ≥ 0.80). Because five structurally different models converge on the same physiological channels, the explanation can be attributed to physiology rather than to a particular network. This is reinforced independently by the Random Forest (Fig. 7): HR 0.46, EDA 0.29, TEMP 0.16, ACC 0.06, BVP 0.03; the low weight on acceleration confirms the models are not simply detecting movement.

Table 10. Channel importance across architectures (Integrated Gradients, subject S2).

Model	EDA	BVP	TEMP	ACC	HR
LSTM	0.15	0.06	0.21	0.08	0.51
BiLSTM	0.08	0.04	0.38	0.03	0.47
GRU	0.09	0.08	0.43	0.04	0.36
CNN-LSTM	0.25	0.15	0.28	0.11	0.21
Transformer	0.18	0.04	0.34	0.15	0.29

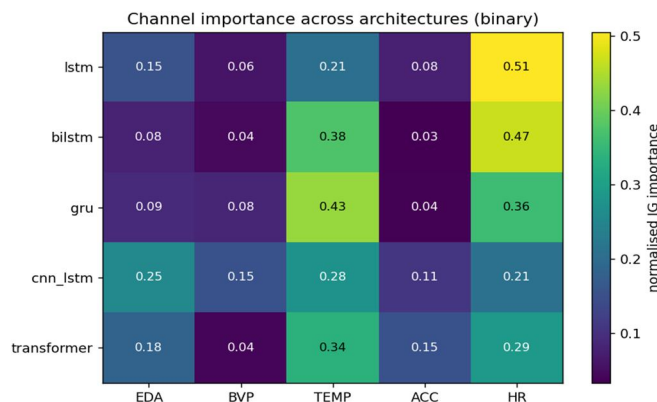


Fig. 6. Channel importance compared across the five deep architectures.

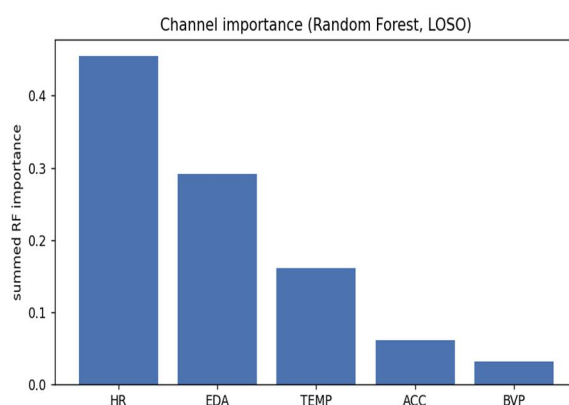


Fig. 7. Channel importance obtained independently from the Random Forest classifier.

G. Counterfactual and temporal analysis

The counterfactual search succeeded for every stress window examined (flip rate 1.00). The channel requiring the largest change to convert a stress prediction into non-stress was heart rate (0.29), then acceleration (0.25), temperature (0.17), EDA (0.15) and BVP (0.14), giving statements of the form 'the window would have been non-stress if the heart rate had been lower.' The temporal-occlusion profile (Fig. 5d) indicates which portions of the 60-second window contribute most.

H. Uncertainty, calibration and coverage-risk analysis

Table 11 reports the deep ensemble across all 15 participants: 0.884 accuracy with Expected Calibration Error 0.065, indicating confidence values close to the accuracy actually achieved. To interpret confidence-based deferral correctly, Table 12 and Fig. 9 give a coverage-risk analysis, where coverage is the fraction of cases answered and risk = 1 – selective accuracy.

Table 11. Uncertainty of the LSTM deep ensemble.

Quantity	Value
Accuracy (all predictions, coverage 100%)	0.884
F1 score	0.810
Expected Calibration Error	0.065
Mean predictive entropy	0.432

Table 12. Coverage–risk analysis of the calibrated ensemble (confidence-based deferral).

Coverage (cases answered)	Selective accuracy	Risk (1 – sel. acc.)
100%	0.884	0.116
90%	0.918	0.082
75%	0.937	0.063
50%	0.960	0.040
25%	0.957	0.043

The 0.960 figure is therefore selective accuracy at 50% coverage, the accuracy over only the most-confident half of predictions when the other half is deferred to a human, and should not be read as overall accuracy, which is 0.884. Risk falls monotonically from 0.116 at full coverage to 0.040 at 50% coverage and does not improve at 25%, so answering roughly the most-confident half is a sensible operating point and a realistic deployment strategy: the model answers when confident and abstains otherwise.

Reliability - lstm (ensemble, binary)

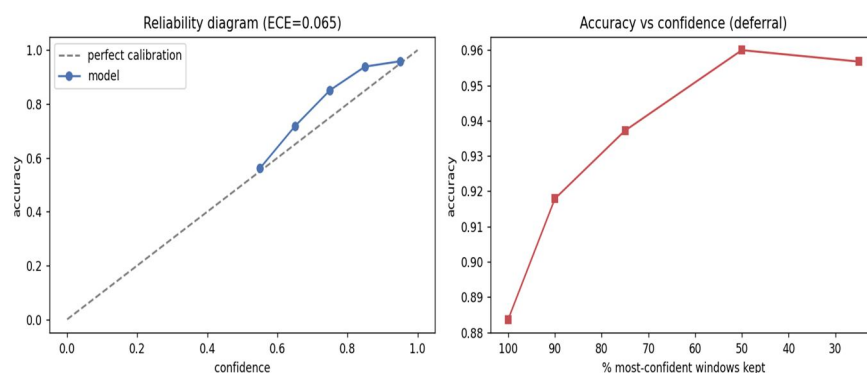


Fig. 8. Reliability diagram (left) and accuracy versus retained confidence (right) for the deep ensemble.

Coverage-risk analysis of the calibrated ensemble (deferral)

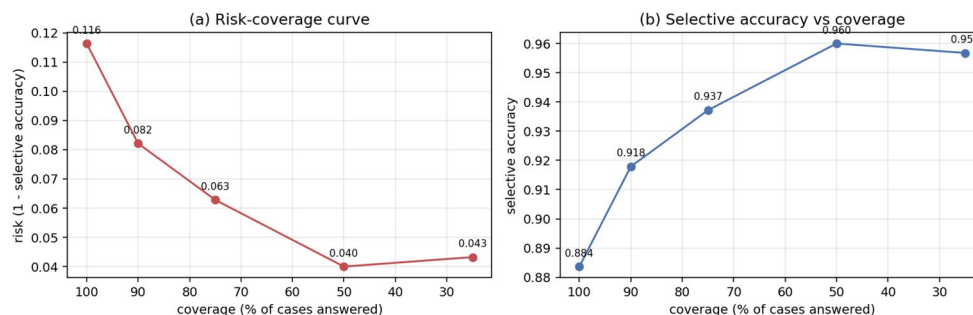


Fig. 9. Coverage–risk analysis: risk–coverage curve (left) and selective accuracy versus coverage (right).

I. Discussion

Three conclusions follow. **First**, on this 15-subject dataset the choice of deep architecture is not an effective lever and a well-regularised classical model is at least as strong; after correction this rests on consistent direction and effect sizes rather than per-pair significance, and effort is better directed towards reliability than architectural novelty. **Second**, explanations must be validated rather than displayed: across six subjects the three methods produced only partially overlapping rankings with high variance, so a single-method explanation cannot be considered dependable; faithfulness and cross-architecture comparison nonetheless establish heart rate and skin temperature as the dominant channels overall, consistent with sympathetic-activation physiology. **Third**, calibrated uncertainty with a coverage–risk deferral strategy converts a moderately accurate system into a usable one by allowing it to abstain when unsure.

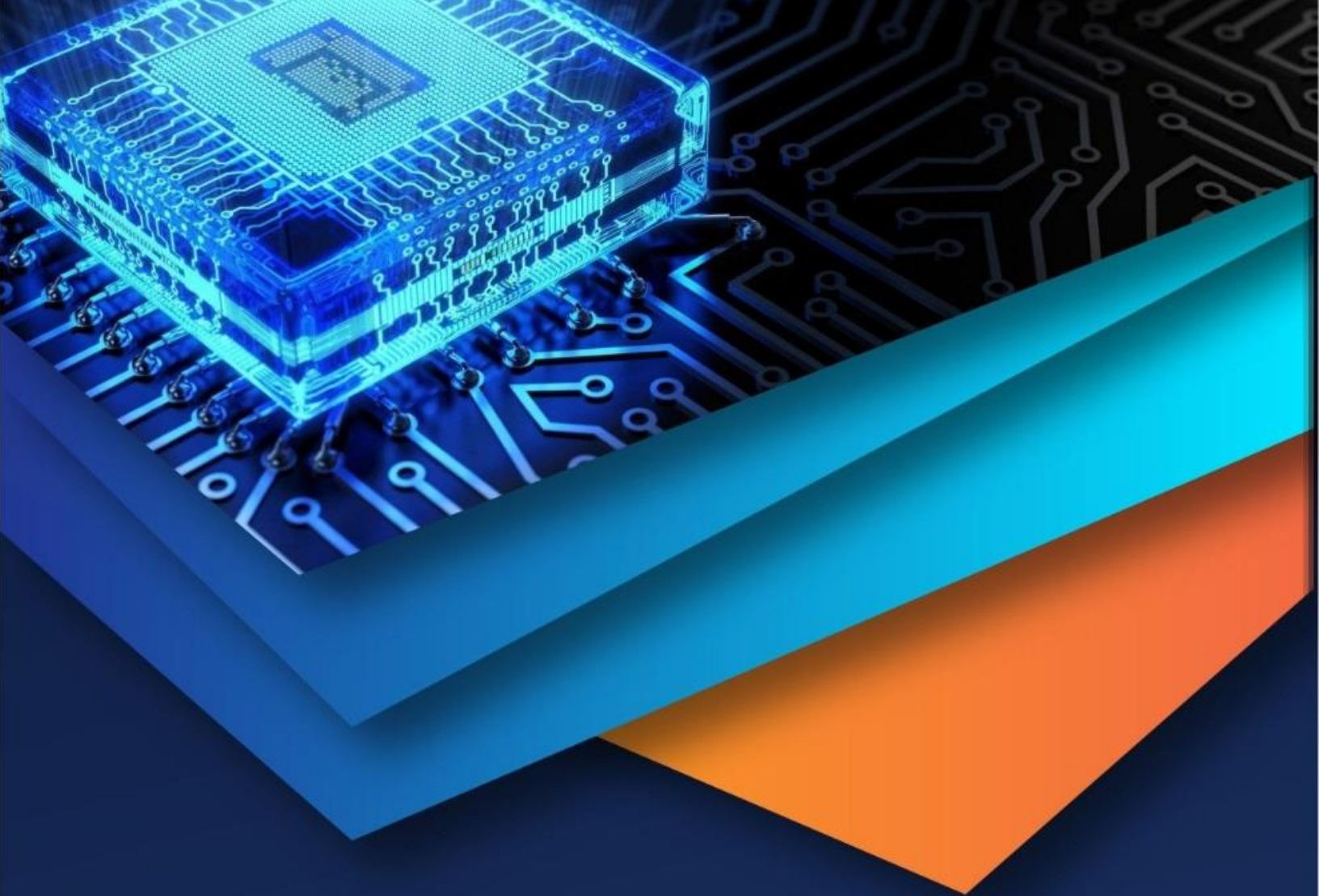
VI. CONCLUSION AND FUTURE WORK

This paper examined whether explanations produced by wearable stress-detection models can be trusted. Using WESAD under strict LOSO, five deep and five classical models were compared with multiple-comparison correction and effect sizes; three explanation methods were applied to the same model and compared across six held-out participants; explanations were verified through faithfulness testing and cross-architecture analysis; and predictive uncertainty was estimated, calibrated and analysed through a coverage–risk curve. The study finds that the deep architectures are statistically indistinguishable, that a classical model is at least as strong at this data scale, that explanation methods agree only partially and with high variance across subjects, so explanation reliability must be measured, not assumed, and that a calibrated ensemble with deferral raises selective accuracy from 0.884 to 0.960 at 50% coverage. The central conclusion is that trustworthiness, measured explanation reliability together with calibration, rather than accuracy alone, is the appropriate objective for this field.

Limitations and future work. The dataset has only 15 participants and was recorded in a controlled laboratory; the task was binary; the multi-subject explainability analysis used six participants and the illustrative faithfulness figures come from one; and, because windows overlap, the effective number of independent observations is the number of subjects, which limits statistical power. Attribution results can also depend on the Integrated Gradients baseline. Future work will extend the evaluation to additional datasets and to all subjects, examine multi-level stress classification, incorporate heart-rate-variability features, investigate per-subject personalisation and calibration, and assess on-device feasibility.

REFERENCES

- [1] M. K. Moser, M. Ehrhart, and B. Resch, “An Explainable Deep Learning Approach for Stress Detection in Wearable Sensor Measurements,” *Sensors*, vol. 24, no. 16, 5085, 2024.
- [2] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, “Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection,” *Proc. 20th ACM ICMI*, 2018, pp. 400–408.
- [3] M. Kyrou, I. Kompatsiaris, and P. C. Petrantonakis, “Deep Learning Approaches for Stress Detection: A Survey,” *IEEE Trans. Affective Computing*, vol. 16, no. 2, pp. 499–517, 2025.
- [4] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic Attribution for Deep Networks,” *Proc. 34th ICML*, 2017, pp. 3319–3328.
- [5] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” *NeurIPS 30*, 2017, pp. 4765–4774.
- [6] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why Should I Trust You? Explaining the Predictions of Any Classifier,” *Proc. 22nd ACM SIGKDD*, 2016, pp. 1135–1144.
- [7] B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles,” *NeurIPS 30*, 2017, pp. 6402–6413.
- [8] Y. Gal and Z. Ghahramani, “Dropout as a Bayesian Approximation,” *Proc. 33rd ICML*, 2016, pp. 1050–1059.
- [9] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On Calibration of Modern Neural Networks,” *Proc. 34th ICML*, 2017, pp. 1321–1330.
- [10] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [11] A. Vaswani et al., “Attention Is All You Need,” *NeurIPS 30*, 2017, pp. 5998–6008.
- [12] S. Gedam and S. Paul, “A Review on Mental Stress Detection Using Wearable Sensors and Machine Learning Techniques,” *IEEE Access*, vol. 9, pp. 84045–84066, 2021.
- [13] G. Giannakakis et al., “Review on Psychological Stress Detection Using Biosignals,” *IEEE Trans. Affective Computing*, vol. 13, no. 1, pp. 440–460, 2022.
- [14] J. DeYoung et al., “ERASER: A Benchmark to Evaluate Rationalized NLP Models,” *Proc. 58th ACL*, 2020, pp. 4443–4458.
- [15] P. Bobade and M. Vani, “Stress Detection with Machine Learning and Deep Learning Using Multimodal Physiological Data,” *Proc. 2nd ICIRCA*, 2020, pp. 51–57.
- [16] Y. S. Can, B. Arnrich, and C. Ersoy, “Stress Detection in Daily Life Scenarios Using Smartphones and Wearable Sensors: A Survey,” *J. Biomedical Informatics*, vol. 92, 103139, 2019.



10.22214/IJRASET



45.98



IMPACT FACTOR:
7.129



IMPACT FACTOR:
7.429



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24*7 Support on Whatsapp)