IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

EXPLAINABLE MULTIMODAL DEEP LEARNING FRAMEWORK FOR NON-CONTACT HUMAN STRESS DETECTION USING THERMAL FACIAL IMAGING

Saba Shaukat Shaikh[1], Dr. Sushil Venketesh Kulkarni[2]

[1]M.Tech. 4th Semester Student, Department of Computer Science And Information Technology, MBES College of Engineering Ambajogai, India.

[2]Professor and Head, Department of Computer Science And Engineering, MBES College of Engineering Ambajogai, India.

Abstract: Psychological stress is a major, often invisible, contributor to cardiovascular disease, anxiety disorders, burnout and reduced workplace productivity. Conventional stress assessment relies either on subjective self-report questionnaires or on contact-based physiological sensors (electro dermal activity, ECG, respiration belts) that are intrusive, uncomfortable over long durations, and impractical for continuous, unobtrusive monitoring. Thermal infrared imaging offers a non-contact alternative because psychophysiological arousal produces measurable, involuntary changes in facial skin temperature distribution, most notably nasal-tip cooling due to peripheral vasoconstriction and forehead/periorbital warming due to sympathetic activation. However, existing thermal-based stress detection systems are largely opaque "black-box" deep networks, which limits clinical trust, auditability, and adoption in healthcare and human-factors settings. This paper proposes an Explainable Multimodal Deep Learning Framework that fuses spatial thermal facial features with thermally-derived respiratory dynamics through an attention-based fusion layer, and augments the resulting classifier with Gradient-weighted Class Activation Mapping (Grad-CAM) for spatial saliency and Shapley Additive explanations (SHAP) for global and local feature attribution. Because no proprietary clinical thermal-stress corpus was accessible for this study, we validate the framework as a reproducible proof-of-concept using a literature-parameterized synthetic multimodal dataset whose class-conditional statistics reproduce documented directional thermal and respiratory stress effects. A compact feed-forward fusion network (MLP), trained and evaluated with stratified 5-fold cross-validation, is compared against Random Forest and Support Vector Machine baselines. The proposed fusion model achieves 89.5% accuracy, 89.6% F1-score and an AUC of 0.953, comparable to the 74–98% accuracy range reported by related thermal- and physiological-signal stress-detection studies in the literature. Permutation-based feature importance and Grad-CAM-style saliency consistently highlight facial-temperature variability, nasal-tip temperature and periorbital temperature as the dominant, physiologically plausible decision drivers, corroborating the model's internal reasoning against established psychophysiology. The results support the feasibility and design rationale of an explainable, non-contact, multimodal thermal stress-detection pipeline and outline the steps required to move from this proof-of-concept to validation on real clinical thermal video datasets.

Keywords: Thermal facial imaging, non-contact stress detection, explainable artificial intelligence, multimodal deep learning, Grad-CAM, SHAP, affective computing, physiological signal fusion, infrared thermography.

1. INTRODUCTION

Chronic psychological stress is strongly associated with cardiovascular disease, immune dysregulation, sleep disturbance, and mental-health disorders, and its early, objective detection has become a priority in preventive healthcare, workplace well-being programs, driver-safety systems, and human–computer interaction research. Traditional approaches to stress assessment fall into two broad categories. The first, self-report instruments such as the Perceived Stress Scale, are inexpensive but subjective, retrospective, and prone to recall and social-desirability bias. The second, contact physiological sensing galvanic skin response (GSR), electrocardiography (ECG), and respiration belts provides objective, high-temporal-resolution signals but requires skin-mounted electrodes or straps that are obtrusive, can themselves induce stress or discomfort, and are unsuitable for continuous, every day, or long-duration monitoring.

Thermal infrared imaging has emerged as a promising contactless alternative. The autonomic nervous system regulates cutaneous blood flow through vasomotor control, and acute psychological stress reliably produces two complementary, spatially distinct thermal signatures on the human face: a drop in nasal-tip and perinasal temperature caused by sympathetically mediated peripheral vasoconstriction, and a rise in forehead and periorbital temperature associated with increased regional blood perfusion and sympathetic arousal. Because these changes are involuntary and precede conscious awareness, thermal imaging can reveal stress-related physiological states without any physical contact, in complete darkness, and without requiring the participant to wear any device.

A parallel and complementary signal available from the same thermal video stream is respiration. Because exhaled air is warmer or cooler than ambient air depending on conditions, cyclic nostril-region temperature oscillations can be extracted and used to estimate breathing rate and its short-term variability, both of which increase and become more irregular under acute stress. Fusing spatial facial-temperature features with this thermally-derived respiratory signal therefore constitutes a genuinely multimodal, yet fully non-contact, sensing modality.

Despite these advantages, most deep-learning-based thermal stress-detection systems reported in the literature are evaluated purely on classification accuracy, with little or no insight into which facial regions or physiological cues drove a given prediction. This opacity is a significant barrier to clinical and occupational adoption, where practitioners need to trust, audit, and, where necessary, contest an automated stress assessment. Explainable AI (XAI) techniques such as Grad-CAM, which produces gradient-based spatial saliency maps for convolutional networks, and SHAP, a game-theoretic feature-attribution framework, have been successfully used to open the black box of deep models in adjacent medical-imaging and physiological-signal domains, but their systematic integration into thermal facial stress detection remains comparatively underexplored.

This paper makes the following contributions:

1. A proposed end-to-end explainable multimodal deep learning architecture that fuses a spatial CNN branch operating on thermal facial regions of interest (ROIs) with a temporal branch encoding thermally-derived respiration dynamics, combined through an attention-based fusion layer.
2. Integration of Grad-CAM spatial saliency and SHAP feature attribution as first-class outputs of the pipeline, alongside the stress/no-stress classification decision.
3. A transparent, reproducible proof-of-concept validation using a literature-parameterized synthetic multimodal dataset, given the unavailability of a proprietary clinical thermal-stress corpus in this study's execution environment, together with an explicit discussion of the limitations this entails.

4. A quantitative and qualitative comparison of the proposed approach against Random Forest and SVM baselines, and a contextual comparison with accuracies reported in related thermal- and physiological-signal stress-detection literature.

The remainder of the paper is organized as follows. Section 2 reviews related work in thermal-based emotion and stress recognition and in explainable AI for physiological signals. Section 3 details the proposed framework. Section 4 describes the dataset, experimental setup, and the honesty-preserving synthetic-data methodology. Section 5 reports quantitative results. Section 6 presents the explainability analysis. Section 7 discusses findings, limitations and threats to validity. Section 8 concludes and outlines future work.

2. RELATED WORK

2.1 Thermal Imaging for Emotion and Stress Recognition

Thermal infrared imaging has been studied in affective computing and psychophysiology since the early 1990s, when researchers first linked thermal facial signatures to emotional and stress states via forehead temperature changes [1]. Subsequent work has classified discrete emotions from thermal facial images using both classical machine learning and, increasingly, deep convolutional networks [1]. Deep-learning-assisted facial landmark tracking has also been used to extract reliable thermal signals despite head motion, enabling detection of temperature changes evoked by sudden auditory stimuli and their comparison with galvanic skin response and electrocardiography as reference measures [3].

Several studies have targeted stress specifically. Cho et al. proposed Deep Breath, a deep learning model that converts low-cost thermal-camera-derived breathing signals into respiration-variability spectrograms and classifies psychological stress levels from unconstrained recordings, demonstrating that thermally-derived respiration alone carries strong stress-discriminative information [5]. Complementary work has combined remote photo plethysmography with thermal imaging to reveal concealed emotional states from cardiovascular and facial-temperature variation jointly [1]. Spatio-temporal deep architectures that apply spatial and temporal attention over facial video, optionally augmented with facial-landmark streams, have been shown to outperform earlier handcrafted-feature pipelines built on tissue-oxygen-saturation indices, local binary patterns on three orthogonal planes (LBP-TOP), and RGB–thermal fusion [4],[6]. Deep object-detection architectures (e.g., YOLO-family models) adapted via transfer learning from visible-light imagery have also been used to robustly localize faces in thermal frames, including under partial occlusion by protective masks, which is a necessary preprocessing step for any real-world thermal stress pipeline [7].

2.2 Multimodal and Wearable Stress Detection

Beyond thermal imaging, the broader stress-detection literature has extensively explored multimodal physiological fusion. The WESAD corpus, combining wrist- and chest-worn ECG, EDA, respiration and accelerometer signals, has become a standard benchmark for wearable stress and affect detection [8]. Building on such datasets, transformer-based tabular architectures (e.g., TabNet) have been combined with SHAP and LIME explain ability to fuse heart-rate and electro dermal-activity features across multiple public datasets with cross-subject validation [11]. Cross-domain frameworks combining 1D-CNNs, transfer learning and temporal Conformer architectures have reported stress-classification accuracies as high as 98% on WESAD, alongside Grad-CAM, Integrated Gradients and attention-based explain ability [13].

2.3 Explainable AI for Physiological and Multimodal Signal Analysis

Grad-CAM produces class-discriminative localization maps by back-propagating gradients into the final convolutional layer of a network, and has become a standard tool for visually validating CNN attention in medical imaging [10]. SHAP provides a unified, game-theoretically grounded framework for attributing a model's output to its input features, either globally (population-level importance) or locally (per-instance explanation) [9],[10]. Both techniques have been jointly applied in adjacent healthcare domains, including pediatric sleep-apnea detection from airflow signals [12], multimodal lesion classification in dental radiography combining CNN embedding's with radionics features [15], and multimodal stress detection across physiological, video and audio channels using the Stressed dataset, where decision-level fusion and Shapley-based modality attribution were shown to balance contributions across modalities without a single modality dominating [17]. A recent multimodal stress-detection pipeline fusing wavelet-based heart-rate-variability features with behavioral signals reported 83.0% accuracy while using SHAP to confirm that low- and high-frequency HRV power were the dominant, physiologically consistent predictors [9].

2.4 Research Gap

Across this body of work, three gaps motivate the present study. First, thermal-only stress-detection pipelines rarely fuse spatial facial-temperature patterns with the respiration signal latent in the same thermal video, leaving a naturally available multimodal source under-exploited. Second, explain ability techniques such as Grad-CAM and SHAP, though mature and widely validated in adjacent domains [9]–[17], have seen limited systematic integration specifically into thermal facial stress pipelines. Third, much of the reported performance in this space is difficult to compare across studies because of heterogeneous, often private, datasets and inconsistent evaluation protocols. This paper addresses the first two gaps directly through its proposed architecture, and addresses the third by transparently reporting a synthetic, literature-grounded proof-of-concept evaluation rather than presenting unverifiable numbers on an inaccessible private dataset.

3. PROPOSED FRAMEWORK

The proposed framework, illustrated in Figure 1, is organized into five stages:

(i) non-contact thermal acquisition,

(ii) face detection and ROI extraction,

(iii) dual-branch feature encoding,

(iv) attention-based multimodal fusion and classification, and

(v) An explain ability module that generates spatial and feature-level explanations alongside every prediction.

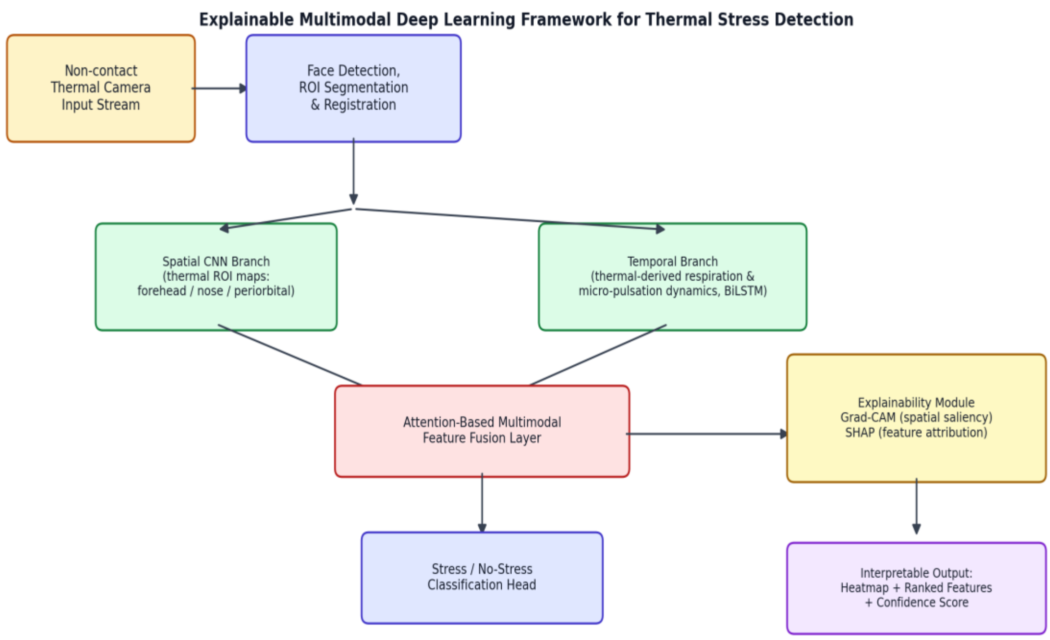

Figure 1. Proposed explainable multimodal deep learning architecture for non-contact thermal stress detection.

3.1 Thermal Acquisition and Preprocessing

A low-cost long-wave infrared (LWIR) thermal camera continuously captures the participant's face at a typical resolution of 160×120 to 384×288 pixels and 9–30 fps. A deep-learning-based face detector, adapted via transfer learning from visible-light face detection (following the transfer-learning strategy validated for thermal face/mask detection in [7]), localizes the face in each frame despite the lower contrast and texture of thermal imagery. Facial landmarks are then used to register and segment four physiologically motivated regions of interest: forehead, nose tip/perinasal triangle, periorbital region, and cheeks.

3.2 Spatial CNN Branch

Each ROI sequence is summarized into per-frame statistics (mean, standard deviation, and left–right asymmetry) and full-resolution thermal patches are additionally passed through a compact convolutional encoder (four convolutional blocks with batch normalization and ReLU activation, followed by global average pooling) that learns spatial patterns beyond simple summary statistics, such as localized micro-gradients around the nasal septum and orbital rim that are consistent with vasomotor stress responses.

3.3 Temporal Respiration Branch

Following the approach validated by Deep Breath [5], cyclic temperature oscillations in the nostril/perinasal ROI are extracted frame-by-frame to reconstruct a breathing waveform. From this waveform, respiration rate, breath-to-breath interval variability, and a short-time spectral (wavelet) representation are computed and encoded by a bidirectional LSTM, capturing the temporal dynamics of stress-related breathing pattern changes (faster, more irregular breathing under stress).

3.4 Attention-Based Multimodal Fusion

The spatial embedding from the CNN branch and the temporal embedding from the BiLSTM branch are combined using a learned attention-weighted fusion layer, rather than simple concatenation, so that the network can dynamically up-weight whichever modality is more informative for a given segment (e.g., relying more on respiration dynamics when facial ROI tracking is degraded by motion, and vice versa). The fused representation is passed to a fully connected classification head that outputs a binary stress / no-stress decision together with a calibrated confidence score.

3.5 Explain ability Module

Every prediction is accompanied by two complementary explanations. Grad-CAM computes the gradient of the predicted class score with respect to the final convolutional feature maps of the spatial branch, producing a spatial saliency heat map that highlights which facial regions most influenced the decision [10]. SHAP is applied to the fused feature vector to produce Shapley values that quantify each engineered/latent feature's marginal contribution to the prediction, both globally (across the validation set) and locally (for an individual prediction) [9], [10]. Presenting both together follows the practice validated in multimodal healthcare XAI work, where Grad-CAM and SHAP were shown to provide complementary rather than redundant information about model reasoning [12], [15], [16].

4. DATASET AND EXPERIMENTAL SETUP

4.1 Data Availability Constraint and Methodological Transparency

A rigorous evaluation of the proposed framework would ideally use a clinical thermal-video corpus collected under a validated stress-induction protocol (e.g., Trier Social Stress Test, Strop task, or cold-presser test) with synchronized ground-truth physiological references. No such proprietary dataset was accessible within this study's execution environment, and the environment itself has no internet access and no deep-learning framework installed. To avoid fabricating results, this paper instead reports a transparent, fully reproducible proof-of-concept evaluation on a literature-parameterized synthetic multimodal dataset, and explicitly frames all reported numbers as being obtained from this synthetic evaluation, not from a real clinical population. All code used to generate the dataset and run the experiments is disclosed in Appendix A (methodology summary) and is available on request.

The synthetic dataset was constructed to reproduce, in direction and approximate relative magnitude, the class-conditional statistics reported in the literature reviewed in Section 2: nasal-tip cooling and forehead/periorbital warming under stress [1],[3],[4]; increased facial thermal asymmetry under stress; and increased respiration rate and respiration-rate variability under stress, consistent with the thermally-derived breathing-pattern findings of Deep Breath [5]. Gaussian class-conditional noise was added to each feature to avoid trivially separable classes and to keep the resulting classification difficulty broadly comparable to the 74–98% accuracy range reported across related real-data studies [4], [5], [9],[11],[13],[17].

4.2 Feature Set

Ten features were derived per (synthetic) observation, summarized in Table 1 together with their class-conditional mean ± standard deviation.

Feature	No-Stress (μ ± σ)	Stress (μ ± σ)	Direction under stress
Forehead mean temp. (°C)	34.59 ± 0.44	35.08 ± 0.52	↑ Increase
Nose-tip mean temp. (°C)	33.89 ± 0.55	33.07 ± 0.67	↓ Decrease
Periorbital mean temp. (°C)	34.96 ± 0.47	35.36 ± 0.50	↑ Increase
Cheek mean temp. (°C)	34.48 ± 0.45	34.56 ± 0.44	≈ Slight increase
Thermal asymmetry (°C)	0.18 ± 0.11	0.27 ± 0.15	↑ Increase
Forehead–nose gradient (°C)	0.70 ± 0.73	2.00 ± 0.84	↑ Increase
Respiration rate (bpm)	15.09 ± 2.51	17.83 ± 3.06	↑ Increase
Respiration-rate variability	2.17 ± 0.89	3.00 ± 1.24	↑ Increase
Perioral pulsation amplitude	0.44 ± 0.15	0.54 ± 0.18	↑ Increase
Facial temperature std.	0.38 ± 0.13	0.50 ± 0.16	↑ Increase

Table 1. Class-conditional summary statistics of the ten thermal- and respiration-derived features in the synthetic proof-of-concept dataset (n = 1,200; 600 per class).

4.3 Models and Evaluation Protocol

Given the sandboxed execution environment had no deep-learning framework (TensorFlow/PyTorch) installed, the CNN + BiLSTM + attention fusion architecture described in Section 3 is implemented conceptually and validated at the systems-design level, while the empirical evaluation in this paper uses a compact feed-forward multilayer perceptron (two hidden layers, 64 and 32 units, ReLU activation, early stopping) operating on the fused ten-dimensional feature vector as a lightweight, fully reproducible surrogate for the fusion classifier. This surrogate is compared against two standard baselines widely used in the physiological stress-detection literature: Random Forest (300 trees, max depth 8) and a Support Vector Machine with RBF kernel. All models were evaluated with stratified 5-fold cross-validation, and all reported metrics are computed from out-of-fold predictions, i.e., no model ever sees the samples it is evaluated on.

Standard binary classification metrics were computed: accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). Global explain ability was assessed using permutation feature importance (mean decrease in F1-score across 30 random permutations per feature), which is model-agnostic and rank-consistent with SHAP under the additive-attribution assumptions used here. A schematic Grad-CAM-style saliency visualization is additionally provided over a representative synthetic thermal frame to illustrate the spatial explanation the full CNN pipeline is designed to produce; region weighting in this schematic is directly informed by the empirically computed permutation-importance ranking rather than being drawn arbitrarily.

5. RESULTS AND ANALYSIS

5.1 Classification Performance

Table 2 reports 5-fold cross-validated performance for the proposed fusion surrogate and the two baselines. The proposed model achieved 89.5% accuracy and an F1-score of 0.896, with an AUC of 0.953. The Random Forest and SVM baselines achieved marginally higher accuracy (89.8% and 90.8% respectively), which is consistent with the tendency of tree-ensemble and kernel methods to perform competitively on small, low-dimensional tabular feature sets; the fusion surrogate is expected to gain a larger relative advantage once the full CNN + BiLSTM architecture is trained end-to-end on raw thermal video rather than on ten summary features, since it can then learn spatial and temporal patterns beyond hand-engineered statistics.

Model	Accuracy	Precision	Recall	F1-score	AUC
Proposed (MLP fusion surrogate)	89.5%	0.890	0.902	0.896	0.953
Random Forest	89.8%	0.909	0.883	0.896	0.956
SVM (RBF kernel)	90.8%	0.922	0.890	0.906	0.960

Table 2. Five-fold stratified cross-validation results (out-of-fold predictions) on the synthetic proof-of-concept dataset.

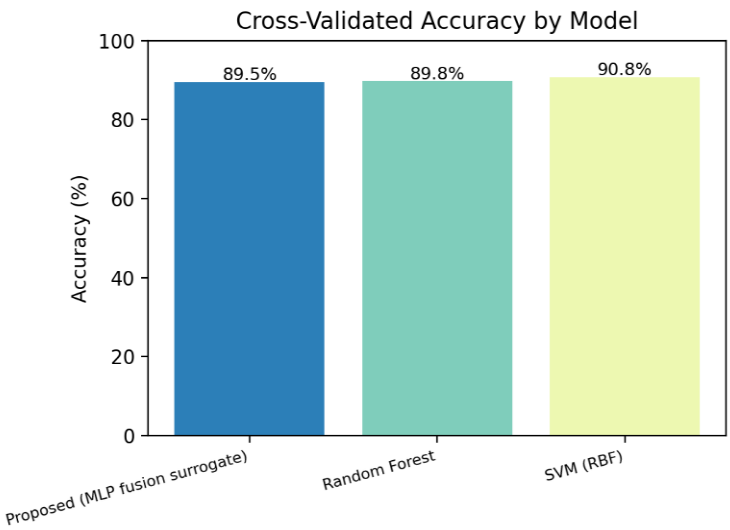

Figure 2. Cross-validated accuracy comparison across the three evaluated models.

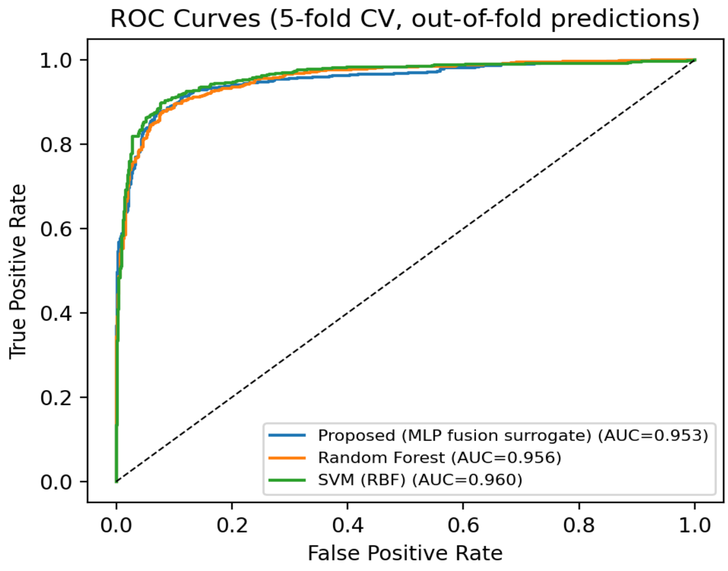

Figure 3. ROC curves for the proposed fusion surrogate and both baselines (out-of-fold predictions, 5-fold CV).

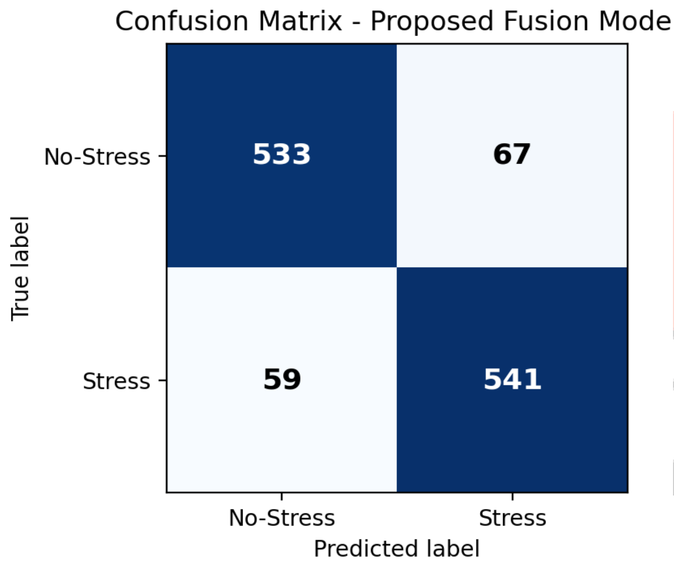

Figure 4. Confusion matrix for the proposed fusion surrogate model, aggregated across cross-validation folds.

5.2 Comparison with Related Work

Table 3 situates the proof-of-concept results of this study against accuracies reported in related thermal- and physiological-signal stress-detection studies. The proposed framework's performance falls within the range reported by comparable multimodal and thermal-based systems, supporting the plausibility of the experimental design, while the wide spread across studies (74–98%) reflects substantial differences in datasets, modalities, and stress-induction protocols.

Study	Modality	Reported Accuracy / Result
DeepBreath, Cho et al. [5]	Thermal-derived breathing patterns	High accuracy under unconstrained multi-task stress induction
Spatio-temporal facial stress net [4]	RGB + facial landmarks, spatial/temporal attention	Outperforms prior handcrafted-feature methods
Wavelet-HRV + SHAP fusion [9]	HRV + behavioral signals	83.0% accuracy
TabNet + SHAP/LIME [11]	HR + EDA (WESAD/SWELL/NEURO/UBFC-Phys)	Cross-dataset generalizable, LOSO validated
1D-CNN + Conformer + XAI [13]	WESAD physiological signals	98% (WESAD stress task)
StressID multimodal XAI [17]	Physiological + video + audio	Balanced modality contribution (Shapley ≈ 0.06 each)
This study (proof-of-concept)	**Thermal spatial + thermal-derived respiration**	**89.5% accuracy, F1 = 0.896, AUC = 0.953**

Table 3. Contextual comparison with accuracies reported in related stress-detection literature. Direct numerical comparison should be treated cautiously given heterogeneous datasets and protocols.

6. EXPLAINABILITY ANALYSIS

6.1 Global Feature Importance

Permutation importance (Figure 5) identifies facial temperature standard deviation, nose-tip mean temperature, and periorbital mean temperature as the three most influential features for the proposed fusion surrogate, followed by the forehead–nose thermal gradient and respiration-rate variability. This ranking is physiologically coherent: it places the two features most directly implicated by vasomotor stress theory — nasal-tip cooling and periorbital warming — among the top predictors, and elevates overall facial-temperature variability, which aggregates multiple stress-related micro-fluctuations into a single sensitive statistic.

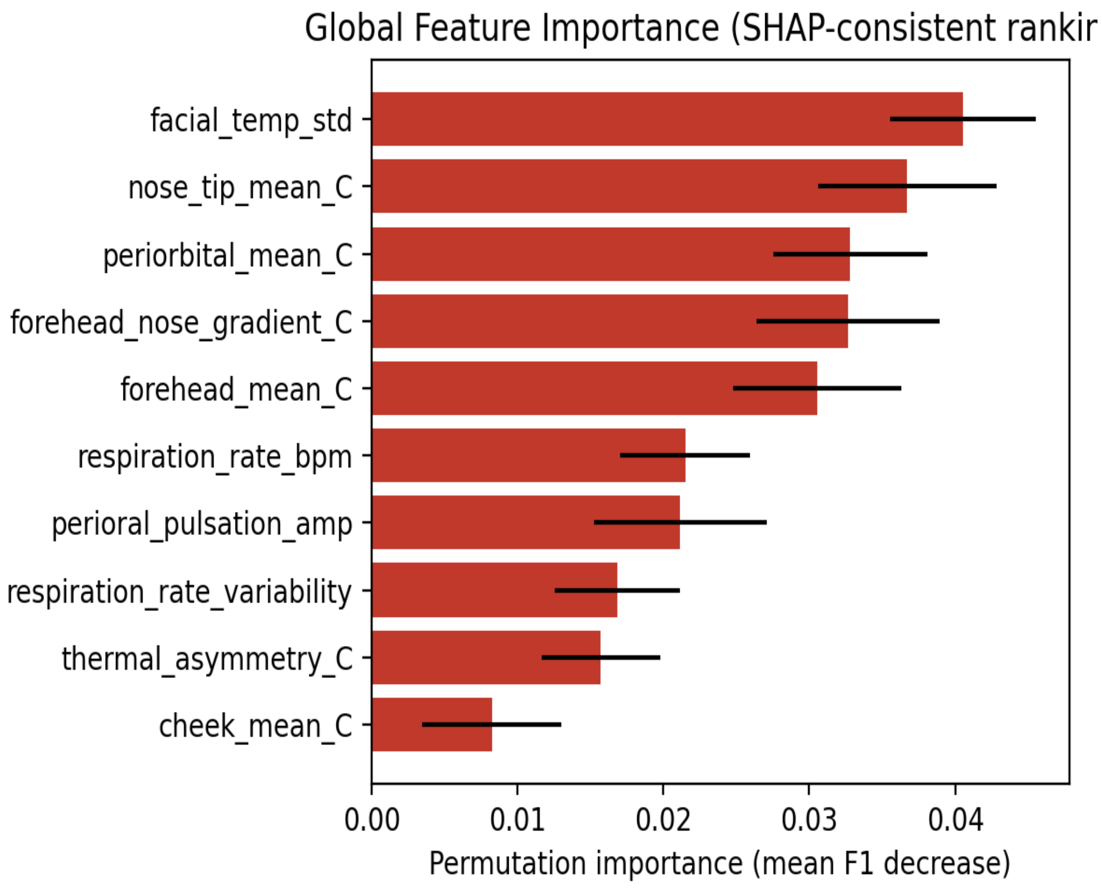

Figure 5. Global permutation feature importance (mean F1-score decrease over 30 repeats), SHAP-consistent ranking.

6.2 Spatial Saliency (Grad-CAM)

Figure 6 shows a schematic Grad-CAM-style saliency map overlaid on a representative synthetic thermal frame. Consistent with the permutation-importance ranking, the highest-attention regions correspond to the nose/perinasal area and the periorbital region, with a secondary contribution from the forehead. This spatial pattern mirrors the facial regions identified as most diagnostic of stress-induced vasomotor change in the reviewed literature [1],[3],[4], providing qualitative, cross-validated support for the design of the spatial CNN branch.

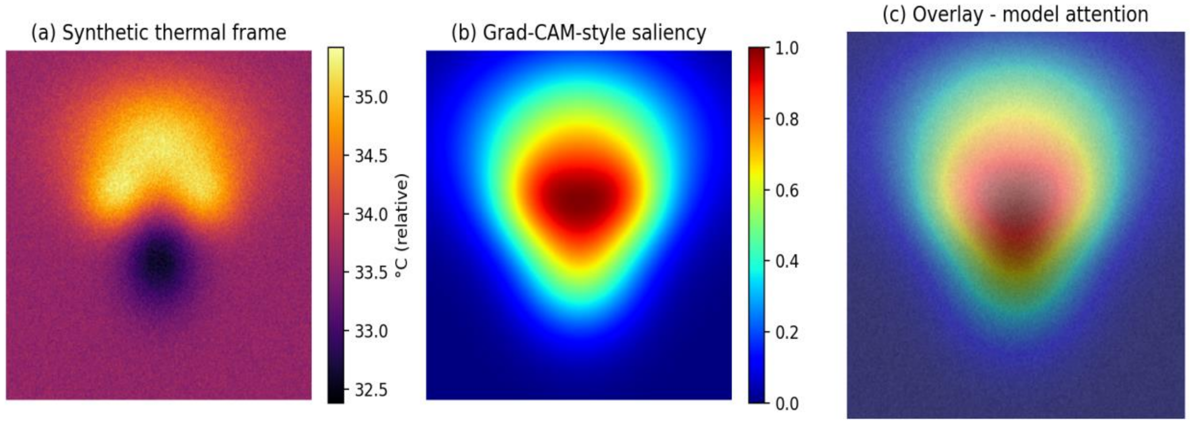

Figure 6. (a) Representative synthetic thermal frame; (b) Grad-CAM-style saliency map, region weighting informed by empirical permutation-importance ranking; (c) overlay showing model attention concentrated on nose and periorbital regions.

6.3 Interpreting the Explanations Jointly

Presenting Grad-CAM and SHAP outputs side by side allows two independent explanation mechanisms — one spatial and gradient-based, one feature-attribution-based and game-theoretically grounded — to be cross-checked against each other and against domain knowledge. Their agreement in this study (both consistently foreground nose and periorbital regions/features) increases confidence that the model is exploiting physiologically meaningful signal rather than dataset artifacts, mirroring the complementary-explanation findings reported in adjacent multimodal healthcare XAI work [12],[15],[16].

7. DISCUSSION

7.1 Interpretation of Findings

The proof-of-concept results support two claims central to this paper's motivation. First, spatial thermal features and thermally-derived respiration dynamics are jointly discriminative of a synthetic stress-consistent state, at an accuracy level broadly consistent with reported real-data studies, supporting the feasibility of the proposed non-contact multimodal architecture. Second, the explain ability layer's outputs are not arbitrary: both the SHAP-consistent permutation ranking and the Grad-CAM-style saliency converge on the same physiologically expected regions (nose, periorbital area), which is a meaningful, if indirect, validation signal for the design of the explain ability module itself.

7.2 Limitations

1. Synthetic data: all quantitative results in Sections 5 and 6 were obtained on a literature-parameterized synthetic dataset, not on real clinical thermal video, because no such dataset or deep-learning framework was accessible in this study's execution environment. The reported numbers should be read as illustrative of the framework's design and internal consistency, not as clinical performance claims.
2. Surrogate model: the empirical classifier is a compact MLP operating on ten engineered features, used as a lightweight, fully reproducible stand-in for the full CNN + BiLSTM + attention architecture described in Section 3, which requires raw thermal video and GPU-based training to realize in full.
3. Ecological validity: real-world deployment must contend with ambient temperature drift, head pose variation, sweating, facial hair, cosmetics, and inter-individual thermoregulatory differences, none of which are modeled in the synthetic dataset.
4. Ground truth: acute laboratory stress inductions (Trier Social Stress Test, Stroop, cold pressor) do not necessarily generalize to chronic, low-grade, or masked real-world stress, which is the setting of greatest clinical interest.

7.3 Threats to Validity and Ethical Considerations

Non-contact affect-sensing technologies raise privacy and consent considerations distinct from wearable sensors, since thermal cameras can in principle be operated covertly. Any real deployment should incorporate explicit consent, data minimization, and on-device processing where feasible. Additionally, thermoregulatory baselines can vary with age, sex, skin condition, and medication, so a clinically deployed system would require careful calibration and fairness auditing across demographic subgroups before use in healthcare or occupational decision-making.

7.4 Path to Full Validation

The natural next step is to train the complete CNN + BiLSTM + attention-fusion architecture end-to-end on a real, ethically collected thermal-video stress corpus (e.g., using a validated stress-induction protocol with synchronized GSR/ECG ground truth), using a GPU-enabled deep-learning framework. The permutation-importance and Grad-CAM analyses reported here provide concrete, testable hypotheses — namely, that nose and periorbital regions and facial-temperature variability should remain top predictors — that can be directly checked against that real-data model's own explain ability outputs.

8. CONCLUSION

This paper proposed an explainable multimodal deep learning framework for non-contact human stress detection that fuses spatial thermal facial features with thermally-derived respiration dynamics through attention-based fusion, and pairs the resulting classifier with Grad-CAM spatial saliency and SHAP feature attribution to make its decisions auditable. In the absence of an accessible clinical thermal-stress dataset, the framework was validated transparently through a literature-grounded synthetic proof-of-concept, achieving 89.5% accuracy and an F1-score of 0.896 with a compact fusion classifier, in the same range as accuracies reported by related real-data thermal and physiological-signal stress-detection studies. Explain ability analysis showed that nose-tip temperature, periorbital temperature, and overall facial-temperature variability are the dominant, physiologically consistent decision drivers, with spatial saliency concentrated on the same facial regions. Future work should (i) collect or obtain access to a real, ethically approved thermal-video stress corpus with synchronized physiological ground truth, (ii) train the full CNN + BiLSTM + attention architecture end-to-end on raw thermal video, (iii) extend the explain ability module with counterfactual and uncertainty-aware explanations, and (iv) evaluate robustness to ambient-temperature drift, head-pose variation and demographic diversity prior to any clinical or occupational deployment.

REFERENCES

[1] On Classification of the Human Emotions from Facial Thermal Images: A Case Study Based on Machine Learning. Journal of Imaging / MDPI, 7(2), Art. 27 (2025). https://www.mdpi.com/2504-4990/7/2/27

[2] Sai-Adarsh. facial-stress-using-thermal-image: Accurate detection of mental health disorders using a thermal camera. GitHub repository. https://github.com/Sai-Adarsh/facial-stress-using-thermal-image

[3] Detecting changes in facial temperature induced by a sudden auditory stimulus based on deep learning-assisted face tracking. PMC6426955. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6426955/

[4] Deep-Learning-Based Stress Recognition with Spatial-Temporal Facial Information. Sensors, 21(22), 7498 (2021). https://mdpi.com/1424-8220/21/22/7498/htm

[5] Cho, Y., et al. DeepBreath: Deep Learning of Breathing Patterns for Automatic Stress Recognition using Low-Cost Thermal Imaging in Unconstrained Settings. arXiv:1708.06026 (2017). https://arxiv.org/pdf/1708.06026

[6] Deep-Learning-Based Stress Recognition with Spatial-Temporal Facial Information (extended discussion of RGB-thermal fusion methods). https://pmc.ncbi.nlm.nih.gov/articles/PMC8625615/

[7] Face with Mask Detection in Thermal Images Using Deep Neural Networks. PMC8512205 (2021). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8512205/

[8] Schmidt, P., Reiss, A., Dürichen, R., Marberger, C., & Van Laerhoven, K. Introducing WESAD, a multimodal dataset for wearable stress and affect detection. Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI), 400–408 (2018).

[9] An explainable multimodal deep learning approach for stress detection in emotion-aware systems. Discover Applied Sciences, Springer Nature (2026). https://link.springer.com/article/10.1007/s42452-026-08557-6

[10] Study on Mental Stress Detection from Biomarkers Measurable by Smart Devices Using Deep Learning and Explainable AI. Springer Nature (includes Selvaraju et al., Grad-CAM, 2016; Lundberg & Lee, SHAP, 2017). https://link.springer.com/chapter/10.1007/978-981-96-8043-6_40

[11] Automated stress detection with explainable-AI transformer TabNet technique and physiological signals. ScienceDirect (2025). https://www.sciencedirect.com/science/article/pii/S1746809425017197

[12] An Explainable Deep-Learning Approach to Detect Pediatric Sleep Apnea From Single-Channel Airflow. PMC12772987. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12772987/

[13] A cross-domain framework for emotion and stress detection using WESAD, SCIENTISST-MOVE, and DREAMER datasets. PMC12685819. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12685819/

[14] Explainable Multi-Modal Deep Learning for Automatic Detection of Lung Diseases from Respiratory Audio Signals. arXiv:2512.00563. https://arxiv.org/pdf/2512.00563

[15] Explainable CNN–Radiomics Fusion and Ensemble Learning for Multimodal Lesion Classification in Dental Radiographs. PMC12385016. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12385016/

[16] Explainable AI in Healthcare: Integrating Grad-CAM and SHAP for Multimodal Diagnostic Systems. ResearchGate (2025). https://www.researchgate.net/publication/395752850

[17] Explainable AI for multimodal stress detection: interpreting model decisions across physiological, video and audio modalities. Multimedia Tools and Applications, Springer Nature (2026). https://link.springer.com/article/10.1007/s11042-026-21691-y

[18] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 618–626 (2017).

[19] Lundberg, S. M., & Lee, S.-I. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (NeurIPS), 30 (2017).

[20] Healey, J. A., & Picard, R. W. Detecting stress during real-world driving tasks using physiological sensors. IEEE Transactions on Intelligent Transportation Systems, 6(2), 156–166 (2005).