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: A Comprehensive Survey

Saba Shaukat Shaikh[1], 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 pervasive determinant of health, productivity, and quality of life, and its unobtrusive, continuous assessment remains an open challenge for affective computing and digital health [1]-[3], [7]. Contact-based physiological sensing (electro dermal activity, electrocardiography, and photo plethysmography) offers reliable ground truth but is intrusive, motion-sensitive, and impractical for continuous, real-world deployment [8], [9]. Thermal facial imaging has emerged as a promising non-contact alternative because sympathetic nervous system activation produces measurable, involuntary vascular and thermoregulatory changes in facial regions such as the nose, forehead, and perioral area, which can be captured passively by infrared cameras without requiring skin contact or user cooperation [18], [20], [27]-[29]. In parallel, deep learning has progressively replaced handcrafted thermal feature extraction with end-to-end convolutional, recurrent, and transformer-based architectures capable of learning discriminative spatio-temporal stress signatures directly from thermal video [21], [23], [25]. However, the clinical and practical adoption of such systems is constrained by the black-box nature of deep networks, which undermines trust, auditability, and regulatory acceptance in health-adjacent applications [57]-[59], [63].

Keywords: Non-contact stress detection; thermal facial imaging; infrared thermography; multimodal deep learning; sensor fusion; explainable artificial intelligence (XAI); Grad-CAM; SHAP; LIME; attention visualization; convolutional neural networks; residual networks; Vision Transformer; long short-term memory; hybrid CNN-Transformer; affective computing; psychophysiology; WESAD; privacy-preserving learning.

1. INTRODUCTION

Stress is the body's adaptive response to perceived physical or psychological demands, mediated by the hypothalamic-pituitary-adrenal axis and the autonomic nervous system [1], [3]. While short-term stress can be adaptive, chronic or unmanaged stress is strongly associated with cardiovascular disease, immune dysfunction, anxiety, depression, and reduced workplace productivity [1], [2]. This has motivated three decades of research into automated stress detection, spanning psychology, biomedical engineering, human-computer interaction, and, most recently, computer vision and deep learning [5], [7], [8].

Historically, stress assessment relied on self-report questionnaires and clinical interviews, which are subjective, retrospective, and unsuitable for continuous monitoring. The introduction of wearable physiological sensors - measuring electro dermal activity (EDA), heart-rate variability (HRV), and skin temperature - enabled objective, continuous stress inference, but at the cost of user compliance, sensor attachment, motion artifacts, and limited scalability outside controlled settings [8], [9], [70]. These limitations have driven growing interest in non-contact sensing modalities, of which thermal facial imaging is among the most physiologically direct: it captures the same autonomic vasomotor responses that drive electro dermal and cardiovascular stress markers, but does so remotely, passively, and without requiring the participant to wear anything [18], [27]-[29].

Concurrently, deep learning has transformed how visual and thermal signals are interpreted. Rather than manually engineering features such as nasal or perinasal temperature differentials, modern systems increasingly rely on convolutional and, more recently, transformer-based encoders that learn stress-relevant spatial and temporal patterns directly from raw thermal sequences, often fused with visible-spectrum video, audio, or wearable physiological streams [21], [23], [25], [30]-[33]. This shift in methodology has been accompanied by markedly improved detection accuracy, but it has also introduced a transparency deficit: end-to-end deep networks are difficult to interrogate, and their internal decision logic is opaque to the clinicians, ergonomists, and end users who would ultimately act on their outputs [57]-[59], [68]. Explainable AI (XAI) techniques such as Grad-CAM, SHAP, LIME, and attention visualization have therefore become an essential complement to high-performing stress-detection architectures, not an optional add-on [57]-[63], [66], [67].

This survey is motivated by the observation that, while thermal imaging, multimodal fusion, and explainable AI have each been surveyed individually in adjacent literatures - facial thermography for affect [18], [24], multimodal wearable stress detection [7], [8], [40], and post-hoc explainability in medical imaging [65], [68] - no existing synthesis integrates all three threads specifically around the problem of non-contact, thermally grounded, and explainable stress detection. The remainder of this paper is organized as follows. Section 2 reviews the physiological and psychological basis of stress. Sections 3 and 4 contrast conventional and contact-based techniques with non-contact alternatives. Section 5 examines thermal facial imaging in depth. Section 6 surveys multimodal deep learning fusion strategies, and Section 7 reviews the deep architectures used across this literature (CNN, ResNet, ViT, LSTM, hybrid CNN-Transformer). Section 8 reviews explainability techniques. Sections 9 and 10 summarize public datasets and compare reported performance. Sections 11 and 12 identify open challenges and research gaps. Section 13 proposes a conceptual explainable multimodal framework, Section 14 outlines future research directions, and Section 15 concludes the survey.

2. BACKGROUND ON HUMAN STRESS

2.1 Physiological and Psychological Basis

Stress responses are coordinated through two principal pathways: the fast-acting sympathetic-adrenal-medullary (SAM) axis, which releases catecholamines and drives immediate autonomic changes (increased heart rate, sweating, altered peripheral blood flow), and the slower hypothalamic-pituitary-adrenal (HPA) axis, which releases cortisol over minutes to hours [1], [3]. Chronic activation of these systems has been linked to impaired prefrontal cortex function, cardiovascular strain, and a broad range of stress-related disorders [1]-[3]. Because these responses are largely involuntary, they provide an objective substrate for automated stress inference that does not depend on a person's willingness or ability to self-report accurately [7].

2.2 Standardized Stress Induction Protocols

Laboratory research on stress detection typically relies on validated stress-induction protocols to generate ground-truth labels. The Trier Social Stress Test (TSST), which combines public speaking and mental arithmetic under evaluative pressure, remains the most widely used paradigm and underlies several of the public multimodal datasets discussed in Section 9 [4], [70]. Other protocols include cognitive load tasks (e.g., color-word Stroop tests), cold-pressor tests, and driving simulations [9], [70].

2.3 Behavioral and Physiological Correlates of Stress

Stress manifests across multiple observable channels: autonomic (EDA, heart rate, skin temperature), facial-expressive (action units associated with tension, per the Facial Action Coding System) [6], vocal (pitch, jitter, speech rate), and behavioral (keystroke dynamics, gaze, posture) [18], [30]. Facial thermal patterns are of particular interest because they are directly downstream of autonomic vasomotor control, offering a physiologically grounded, contactless proxy for the same signals traditionally captured with skin-contact sensors [27]-[29].

3. CONVENTIONAL STRESS DETECTION TECHNIQUES

Conventional stress-assessment techniques can be grouped into three broad categories.

3.1 Self-Report Instruments

Standardized questionnaires such as the Perceived Stress Scale and NASA Task Load Index remain the psychometric gold standard for subjective stress appraisal and are frequently used to generate ground-truth labels for machine learning studies [70]. Their limitations - recall bias, low temporal resolution, and susceptibility to social desirability - motivate the search for objective physiological alternatives.

3.2 Biochemical and Endocrine Assays

Salivary or serum cortisol assays provide a direct biochemical readout of HPA-axis activation but require laboratory analysis, introduce latency of tens of minutes, and cannot support continuous or real-time monitoring [1], [3].

3.3 Physiological Sensor-Based Machine Learning

The dominant contemporary approach uses wearable or attached biosensors - EDA electrodes, ECG leads, photo plethysmography (PPG) wristbands, and respiration bands - combined with classical machine learning (SVM, random forests) or deep networks (CNN, LSTM) to classify stress states [9], [12]-[14], [70]. Systematic reviews report that hybrid CNN-LSTM models fused across EDA, PPG, and accelerometer channels achieve 85-96% classification accuracy on benchmark datasets such as WESAD [40], [70]. These systems have matured considerably, but they remain fundamentally contact-based, motivating the non-contact alternatives discussed next.

4. CONTACT-BASED VS NON-CONTACT STRESS DETECTION

4.1 Contact-Based Sensing

Contact-based systems (chest straps, wrist-worn PPG/EDA sensors, ECG leads) provide high signal fidelity and well-validated feature sets, and underpin most published stress-classification benchmarks [8], [9], [70]. However, they require deliberate donning, are prone to motion artifacts, may cause skin irritation over long wear periods, and are unsuitable for populations or settings where attaching a sensor is impractical (e.g., infants, burn patients, security screening, or naturalistic workplace monitoring) [8], [40], [76].

4.2 Non-Contact Sensing

Non-contact modalities include remote photo plethysmography (rPPG) from RGB video, voice-based stress markers, keystroke and gaze dynamics, and thermal facial imaging [18], [24], [30]. Among these, thermal imaging is distinctive in that it measures a physiological quantity - skin surface temperature governed by autonomic vasomotor tone - rather than an indirect behavioral proxy, giving it a closer mechanistic link to the same autonomic pathways probed by EDA and PPG sensors, while requiring no physical attachment [18], [27]-[29].

4.3 Comparative Trade-offs

The qualitative trade-offs reported across the surveyed literature can be summarized as follows. Contact-based sensing offers higher per-channel signal-to-noise ratio and more mature feature engineering pipelines, but poorer ecological validity and scalability. Non-contact thermal sensing offers superior user acceptability, privacy relative to identifiable RGB video, and applicability in surveillance-style or public-health screening contexts, at the cost of sensitivity to ambient temperature, camera-to-subject distance, and head pose, which combined create a harder face- and landmark-detection problem in the thermal domain [18], [20], [22], [26].

5. THERMAL FACIAL IMAGING FOR STRESS DETECTION

5.1 Physiological Rationale

Infrared thermal cameras capture emitted long-wave radiation, which is converted to a temperature map of the skin surface. Sympathetic activation redirects blood flow away from peripheral vasculature (e.g., the nose) and toward core musculature, producing a measurable nasal and perioral temperature drop under acute stress, while the forehead and periorbital regions can show compensatory warming [18], [27]-[29]. Early work established that peri-nasal thermal imaging could track physiological arousal non-invasively and correlate with concurrent skin-conductance measurements [27], [28], laying the empirical foundation for the field.

5.2 Processing Pipeline

A typical thermal stress-recognition pipeline comprises: (i) thermal face detection and landmark localization, historically performed with Viola-Jones-style detectors, HOG, deformable part models (DPM), or active appearance models adapted to the infrared domain, and increasingly with deep-learning-based face trackers [18], [22]; (ii) region-of-interest (ROI) extraction from the nose, forehead, periorbital, and perioral areas; (iii) feature extraction, either handcrafted (temperature statistics, spatial gradients) or learned end-to-end via CNNs; and (iv) classification into stress/no-stress or graded stress levels [18], [25]. A persistent challenge noted across this literature is that facial landmark detectors trained on visible-spectrum (RGB) faces transfer poorly to thermal imagery because thermal images lack the texture and appearance cues on which such detectors rely, necessitating dedicated thermal-domain detectors and cross-domain image-translation approaches [18], [22].

5.3 Deep Learning for Thermal Stress Recognition

Recent work applies CNNs directly to thermal image sequences, bypassing handcrafted ROI features. Deep Breath, for example, converts low-cost thermal video of breathing patterns into two-dimensional respiration-variability signatures and applies deep feature learning to recognize psychological stress under unconstrained, naturalistic conditions such as verbalization and movement [20]. Other work integrates deep-learning-assisted facial landmark tracking to robustly extract thermal signals under head motion, validating extracted signals against galvanic skin response and electrocardiography [23]. Spatio-temporal deep networks that combine facial landmark streams with attention mechanisms - emphasizing frames and regions most associated with stress - have been shown to outperform earlier, single-frame recognition methods on purpose-built thermal-RGB stress databases [25].

5.4 Practical Constraints

Thermal imaging performance is sensitive to ambient temperature, camera calibration drift, subject distance, and skin-tone-independent but emissivity-dependent surface properties; recent evaluations of thermal versus optical imaging protocols for related physiological sensing tasks confirm that thermal channels are comparatively more robust to lighting and skin-tone variation than RGB channels, though not entirely immune to protocol-induced variability [26]. These constraints must be accounted for in any deployable thermal stress-sensing system.

6. MULTIMODAL DEEP LEARNING APPROACHES

A substantial body of recent work argues that no single modality - thermal, visual, physiological, or behavioral - is sufficient on its own to robustly characterize stress across individuals and contexts, motivating multimodal fusion [30]-[33], [40]. Multimodal deep learning approaches for stress detection can be organized into three fusion paradigms.

6.1 Early (Feature-Level) Fusion

Early fusion concatenates raw or lightly processed signals from multiple modalities before a single model learns a joint representation. Examples include geometric LSTM-based frameworks that combine covariance and cross-covariance representations of physiological and behavioral signals, and networks that fuse ECG, respiration, and facial features at the input level, reporting binary stress-detection accuracies around 73% [33]. Early fusion is architecturally simple but can be brittle to missing or asynchronous modalities and does not explicitly model cross-modal interactions [33], [40].

6.2 Late (Decision-Level) Fusion

Late fusion trains separate unimodal classifiers and combines their outputs (e.g., by averaging, voting, or a meta-learner). This approach is more robust to missing modalities and easier to interpret per-channel, but it can fail to capture fine-grained cross-modal correlations that only emerge when signals are processed jointly at an intermediate representational level [33].

6.3 Intermediate Fusion

Intermediate fusion architectures learn modality-specific encoders whose intermediate representations are merged - via concatenation, attention, or manifold-learning-based alignment - before a shared classification head. Reported results indicate that intermediate fusion of ECG and EDA signals through CNN backbones can reach roughly 75.5% subject-independent stress-detection accuracy, and that manifold-learning-based intermediate fusion offers improved robustness to inter-subject variability compared with naive concatenation [33]. More recent work converts multiple 1-D physiological streams (PPG, EDA, accelerometer) into unified 2-D image representations so that a single CNN can jointly learn inter-signal dependencies, reporting substantial gains over independent per-signal processing on the WESAD benchmark [32].

6.4 Fusion Involving Thermal and Visual Facial Cues

Within the specific context of this survey, multimodal fusion of thermal facial imaging with visible-spectrum video, physiological wearables, or audio is comparatively less mature than fusion among purely physiological wearable channels, but is a rapidly growing direction. Spatio-temporal facial networks that jointly weight informative frames (temporal attention) and informative facial regions (spatial attention) illustrate how attention-based intermediate fusion can be extended across thermal and RGB facial streams to improve robustness over single-channel thermal or RGB models alone [25]. Time-domain and frequency-domain feature fusion (e.g., via Fast Fourier Transform-derived spectral features combined with raw time-series statistics) further illustrates that fusing complementary representations of the same modality, not only across modalities, meaningfully improves robustness and generalization in occupational stress-monitoring settings [31].

6.5 Synthesis

Across the multimodal literature, the consistent finding is that fusion - whether across modalities or across complementary representations of a single modality - improves accuracy and robustness relative to unimodal baselines, at the cost of increased architectural complexity, synchronization requirements, and, critically for this survey, reduced interpretability, since fused representations blend information from heterogeneous sources in ways that are harder for a human to trace back to an original signal or region [30], [33], [40]. This tension between fusion-driven accuracy gains and explainability is a central motivation for the framework proposed in Section 13.

7. DEEP LEARNING ARCHITECTURES

The architectures used across the thermal, multimodal, and affective-computing literature reviewed in this survey can be grouped into five families.

7.1 Convolutional Neural Networks (CNN)

CNNs, popularized for large-scale image classification following the success of AlexNet [44] and formalized earlier in foundational work on gradient-based document recognition [45], remain the default backbone for extracting spatial features from individual thermal or RGB facial frames. Their translation-invariant convolutional filters are well suited to localizing ROI-specific temperature patterns (e.g., peri-nasal cooling) without requiring hand-specified coordinates, and multichannel CNN architectures have been applied directly to physiological and thermal signals for state-of-mind classification [14].

7.2 Residual Networks (ResNet)

Very deep CNNs suffer from vanishing gradients and degraded training accuracy as depth increases; residual networks address this via identity shortcut connections that allow gradients to propagate through very deep stacks of layers [47]. ResNet backbones (and related deep architectures such as VGG [46] and Inception [48]) are widely used as pretrained feature extractors for facial and thermal image classification tasks, including pressure-injury and skin-temperature-change detection, where ResNet50 has been shown to achieve over 90% accuracy on thermal input while remaining comparatively robust to imaging-protocol variation [26].

7.3 Vision Transformer (ViT)

Vision Transformers apply the self-attention mechanism, originally developed for sequence modeling [51], directly to sequences of image patches, dispensing with convolutional inductive biases altogether [52]. When pretrained on sufficiently large datasets, ViTs match or exceed CNN performance on image classification while offering a more global receptive field from the first layer, which is potentially advantageous for capturing distributed thermal gradients spanning multiple facial regions rather than purely local textures [52]-[54]. Hierarchical variants such as the Swin Transformer reintroduce local windowed attention to improve efficiency and scalability for dense prediction tasks [53].

7.4 Long Short-Term Memory (LSTM)

Because stress manifests dynamically over time (e.g., gradually developing peri-nasal cooling, breathing-pattern changes), recurrent architectures capable of modeling long-range temporal dependencies are important complements to per-frame spatial encoders. LSTM networks, which use gated memory cells to overcome the vanishing-gradient problem of plain recurrent networks [49], have been used both independently - e.g., an LSTM-based stress classifier followed by a fully connected softmax layer, evaluated with five-fold cross-validation - and combined with convolutional front-ends (CNN-LSTM) for joint spatial-temporal modeling of wearable and facial signals [50]. Systematic reviews of wearable stress detection consistently identify hybrid CNN-LSTM architectures as the best-performing configuration across benchmark datasets [40].

7.5 Hybrid CNN-Transformer Architectures

A growing trend combines convolutional feature extraction with transformer-based temporal or cross-modal attention, aiming to inherit the local inductive bias and computational efficiency of CNNs together with the long-range dependency modeling and cross-modal fusion capacity of transformers [52]-[54]. In the stress- and affect-detection literature, this typically takes the form of a CNN encoder per modality (thermal, visual, physiological) followed by a transformer-based fusion module that learns cross-modal attention weights, offering a natural mechanism for both improved accuracy and, as discussed in Section 8.4, built-in interpretability via attention-map visualization.

8. EXPLAINABLE AI TECHNIQUES

As deep architectures for stress detection have grown more complex and multimodal, explainability has shifted from a desirable feature to a functional requirement for clinical and human-factors adoption, supporting error detection, bias auditing, and user trust calibration [57], [63], [68]. XAI techniques are broadly divided into post-hoc, model-agnostic methods and architecture-intrinsic (ante-hoc) methods.

8.1 Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) uses the gradients of a target class flowing into the final convolutional layer to produce a coarse, class-discriminative localization heatmap over the input image, highlighting the spatial regions most responsible for a CNN's prediction [57]. Extensions such as Grad-CAM++ generalize the gradient weighting scheme to improve localization of multiple instances of a class within a single image [60]. In the thermal-stress context, Grad-CAM-style visualizations are particularly informative because they can confirm, or challenge, the assumption that a model's prediction is driven by physiologically plausible ROIs (nose, forehead, periorbital region) rather than spurious background or camera artifacts [57], [65]-[67].

8.2 SHAP

SHapley Additive exPlanations (SHAP) draws on cooperative game theory to assign each input feature a Shapley value representing its marginal contribution to a model's output, providing theoretically grounded, additive, and model-agnostic feature attributions [58]. SHAP is well suited to tabular or physiological-signal inputs (e.g., EDA, HRV, and derived thermal ROI statistics) where feature-level attributions are more directly interpretable than pixel-level heatmaps, making it a natural complement to Grad-CAM in multimodal thermal-physiological fusion pipelines [58], [64], [66], [67].

8.3 LIME

Local Interpretable Model-agnostic Explanations (LIME) approximates the local decision boundary of a black-box model around a specific prediction using an interpretable surrogate model (typically sparse linear regression) fit to perturbed samples in the neighborhood of the input [59]. Comparative evaluations across imaging domains report that LIME can achieve higher explanation fidelity than SHAP or Grad-CAM under certain perturbation and noise conditions, though at higher computational cost per explanation, since it requires many perturbed forward passes per instance [65], [68].

8.4 Attention Visualization

For transformer and attention-augmented architectures, the self-attention weights computed during the forward pass can themselves be visualized as a natural, architecture-intrinsic explanation of which input tokens (image patches, physiological time steps, or modality channels) most influenced a given prediction [51]-[54]. Spatial and temporal attention modules applied to facial-thermal sequences have been used both to improve recognition accuracy and to reveal which frames and facial regions the network weights most heavily when recognizing stress, offering a built-in explanatory signal without the additional post-hoc computation required by Grad-CAM, SHAP, or LIME [25]. However, whether raw attention weights constitute a faithful explanation of model behavior remains actively debated in the broader XAI literature, and attention-based explanations are best treated as complementary to, rather than a full substitute for, gradient- or perturbation-based methods [54].

8.5 Comparative Considerations

Meta-analytic comparisons across medical-imaging applications report that LIME tends to achieve the highest fidelity but poorer stability under input noise, Grad-CAM offers a favorable balance of computational efficiency and moderate fidelity, and SHAP provides strong theoretical guarantees but can degrade substantially under noise perturbation in some imaging domains [65], [68]. No single XAI method dominates across all criteria (fidelity, stability, computational cost, and human interpretability), which motivates the ensemble-style, multi-method explainability layer proposed for the framework in Section 13 [64], [67].

9. PUBLICLY AVAILABLE DATASETS

Progress in explainable, thermally grounded stress detection is fundamentally constrained by the scarcity of large, public, multimodal thermal datasets, in contrast to the relative abundance of physiological-wearable datasets. The subsections below summarize the most frequently used public resources referenced across the surveyed literature.

9.1 Physiological/Wearable Datasets

1. WESAD (Wearable Stress and Affect Detection): 15 subjects, chest- and wrist-worn ECG, EDA, EMG, respiration, temperature, and accelerometer signals, collected across baseline, TSST-induced stress, and amusement conditions; the most widely used benchmark in the wearable stress-detection literature [70].
2. A Multimodal Sensor Dataset for Continuous Stress Detection of Nurses in a Hospital: naturalistic, occupational, multi-day physiological recordings intended for ecologically valid stress inference [34].
3. TILES-2018: a large-scale multi-sensor dataset covering personality traits, behavioral states, job performance, and wellbeing over time, used for longitudinal affect and stress modeling [62].
4. AffectiveROAD-style driving datasets: physiological recordings from drivers under real-world or simulated driving stress conditions [9].

9.2 Facial and Thermal Datasets

1. Thermal infrared face databases with facial landmarks and emotion/affect labels, developed to support thermal-domain face detection and landmark localization independent of visible-spectrum detectors [22], [63].
2. Purpose-built spatio-temporal facial-thermal stress databases combining synchronized facial video and physiological ground truth, used to train and evaluate attention-based spatio-temporal recognition networks [25].
3. Recently released multimodal datasets combining facial expressions with physiological signals, intended to bridge the gap between purely physiological and purely facial-visual stress-detection research [35].

9.3 Observations on Dataset Availability

Across this landscape, thermal-specific, richly annotated, publicly released stress datasets remain markedly scarcer than their RGB or physiological counterparts, and most published thermal stress-detection studies rely on smaller, self-collected datasets rather than shared community benchmarks [18], [24], [25]. This scarcity, discussed further in Section 11, is one of the principal bottlenecks limiting reproducible comparison and the training of data-hungry deep architectures such as Vision Transformers specifically on thermal facial imagery.

10. PERFORMANCE COMPARISON OF EXISTING METHODS

Direct, apples-to-apples performance comparison across the surveyed literature is complicated by heterogeneous datasets, labeling protocols (binary stress/no-stress versus graded stress levels), subject-dependent versus subject-independent evaluation splits, and modality combinations. Nonetheless, several consistent patterns emerge.

10.1 Unimodal Physiological and Thermal Methods

LSTM-based classifiers applied to physiological signals with five-fold cross-validation report accuracies around 87% for binary stress classification [50]. EDA-only deep learning approaches have, in some occupational stress-monitoring studies, outperformed multimodal EDA-ECG-PPG fusion, illustrating that naive fusion is not always beneficial and that fusion strategy matters as much as modality count [34]. ResNet-based thermal classification of temperature-change-related physiological states has achieved above 90% accuracy while remaining comparatively robust to imaging-protocol variation, outperforming equivalent optical-image pipelines under the same protocols [26].

10.2 Multimodal Fusion Methods

Early-fusion feature-level networks combining ECG, respiration, and facial features report binary stress-detection accuracy around 73.3% [33]. Intermediate-fusion CNN architectures combining ECG and EDA report approximately 75.5% subject-independent accuracy [33], while systematic reviews of hybrid CNN-LSTM multimodal fusion (PPG, EDA, accelerometer) report a substantially higher 85-96% accuracy range on benchmark datasets such as WESAD, reflecting both stronger fusion architectures and, in several cases, more favorable subject-dependent evaluation protocols [40]. Converting multiple 1-D physiological signals into unified 2-D image representations for CNN-based fusion has been shown to yield further accuracy improvements over independent per-signal processing baselines on WESAD [32].

10.3 Explainability-Integrated Methods

Systems that integrate XAI methods directly into training (e.g., using Grad-CAM- or LIME-derived masks to guide a second training stage) have reported accuracy improvements alongside interpretability gains in adjacent medical-imaging applications, suggesting that explainability and performance are not necessarily in tension and can, in some architectures, be mutually reinforcing [67]. However, meta-analyses caution that imposing interpretability constraints (e.g., inherently interpretable models) can incur a measurable accuracy penalty of several percentage points relative to unconstrained black-box models in medical-imaging tasks, underscoring the accuracy-transparency trade-off that motivates careful XAI method selection rather than blanket adoption [68].

10.4 Summary Observation

Considered together, the surveyed literature indicates that (i) fusion of complementary signals generally, though not universally, improves accuracy over unimodal baselines; (ii) hybrid CNN-recurrent or CNN-transformer architectures currently represent the empirical state of the art for wearable and thermal-adjacent stress detection; and (iii) explainability integration is feasible without prohibitive accuracy loss, provided the choice of XAI method is matched to the input modality (Grad-CAM and attention visualization for image/thermal inputs, SHAP for tabular/physiological features, LIME as a complementary, higher-fidelity but higher-cost cross-check) [40], [64]-[68].

11. CHALLENGES AND LIMITATIONS

11.1 Thermal Dataset Scarcity

Compared with visible-spectrum facial-expression corpora and physiological-wearable datasets such as WESAD, large-scale, publicly available, richly annotated thermal facial stress datasets remain rare, limiting both the training of data-hungry architectures (particularly Vision Transformers) and fair cross-study comparison [18], [24], [25], [70].

11.2 Environmental and Hardware Confounds

Thermal signal quality is sensitive to ambient temperature, airflow, camera-to-subject distance, sensor calibration drift, and head pose, introducing confounds that are difficult to fully control outside laboratory conditions and that can degrade real-world generalization even when validation accuracy is high [18], [26].

11.3 Real-Time Deployment Constraints

Many of the highest-performing architectures reviewed in Sections 6 and 7 - particularly hybrid CNN-Transformer and multimodal fusion networks - are computationally demanding, creating tension between accuracy and the latency, memory, and power constraints of edge or embedded deployment required for continuous, real-world stress monitoring [32], [40].

11.4 Subject and Cross-Domain Generalization

Facial expression and thermal recognition systems exhibit systematic performance degradation under domain shift (different populations, cameras, environments) and demographic bias across age, gender, ethnicity, and skin tone, with accuracy not necessarily correlating with fairness across these axes [79]. Subject-independent evaluation protocols consistently report lower accuracy than subject-dependent splits across the reviewed literature, indicating that many reported results may not generalize to unseen individuals [33], [40].

11.5 Privacy and Ethical Concerns

Thermal and facial data are inherently identity- and health-revealing; capturing continuous physiological arousal data raises distinct privacy, consent, and potential emotional-surveillance concerns beyond those associated with generic image data, particularly as regulatory frameworks around emotion-recognition technologies (e.g., provisions in the EU AI Act) tighten [76]-[78]. Differential-privacy and federated-learning approaches have been explored specifically to mitigate re-identification risk in affect-recognition pipelines, but typically at some cost to recognition accuracy [78].

11.6 Explainability Fidelity and Stability

As discussed in Section 8, no single XAI method is uniformly superior; explanation fidelity can degrade substantially under input noise (particularly for SHAP in some imaging domains), and interpretable-by-design models can incur accuracy penalties, creating a persistent tension between transparency, robustness, and performance that current XAI toolkits do not fully resolve [65], [68].

12. RESEARCH GAPS

1. Absence of a large-scale, standardized, publicly available thermal-plus-multimodal stress benchmark analogous to WESAD but centered on facial thermography, hindering reproducible comparison across studies [18], [24], [70].
2. Limited exploration of Vision Transformer and hybrid CNN-Transformer architectures specifically on thermal facial sequences, as most transformer-based stress work to date has focused on physiological time-series or RGB video rather than thermal imagery [52]-[54].
3. Sparse integration of ante-hoc (built-in) explainability into multimodal thermal-physiological fusion architectures; most XAI applications reviewed here are post-hoc and applied after the fact, rather than co-designed with the recognition architecture [63], [69].
4. Insufficient cross-population and cross-demographic validation of thermal stress-detection systems, given known fairness and generalization concerns identified in adjacent facial-expression-recognition research [79].
5. Few studies jointly optimize for real-time, edge-deployable inference and explainability, despite both being prerequisites for practical, continuous, non-contact stress monitoring outside the laboratory [32], [40].

6. Limited standardized evaluation protocols for comparing the fidelity, stability, and human interpretability of Grad-CAM, SHAP, LIME, and attention visualization specifically in the context of thermal and multimodal stress detection, as opposed to the more heavily studied domains of medical imaging and malware classification [64]-[68].
7. Underdeveloped privacy-preserving pipelines (federated learning, differential privacy, on-device inference) tailored specifically to thermal facial data, which is simultaneously highly informative and highly sensitive [76]-[78].

13. PROPOSED EXPLAINABLE MULTIMODAL FRAMEWORK

Building on the gaps identified in Section 12, this survey proposes a conceptual Explainable Multimodal Stress-Sensing Framework (EMSF) intended to guide future system design rather than to report a single implemented and benchmarked system.

13.1 Architectural Overview

EMSF comprises three stages. First, modality-specific encoders process synchronized thermal facial video, visible-spectrum video (optional), and, where available, wearable physiological streams: a lightweight ResNet-style CNN backbone extracts per-frame spatial thermal features from facial ROIs (nose, forehead, periorbital, perioral regions) [22], [47]; an LSTM or temporal-attention module models the evolution of these features over time to capture gradual thermoregulatory stress signatures [25], [49]; and, where a Vision Transformer or hybrid CNN-Transformer variant is computationally feasible, patch-level self-attention supplements the CNN pathway to capture distributed, non-local thermal gradients across the face [52]-[54]. Second, a cross-modal transformer-based fusion module (Section 6.3, 7.5) combines these modality-specific embeddings via cross-attention, allowing the network to learn, for each prediction, which modality and which spatial/temporal region contributed most [30], [33].

13.2 Integrated Explainability Layer

Rather than applying a single post-hoc XAI method, EMSF proposes a layered explainability design matched to input type, following the comparative findings of Section 8.5: Grad-CAM (or Grad-CAM++) applied to the CNN thermal-image pathway to produce ROI-level saliency maps that can be checked against known physiologically plausible stress regions [57], [60]; native cross-attention weights from the fusion transformer, visualized to show relative modality and temporal contribution per prediction [51]-[54]; and SHAP values computed over any tabular or handcrafted physiological/thermal-ROI summary features feeding into the fusion module, to provide feature-level attributions that complement pixel-level saliency [58]. LIME is retained as a periodic, higher-cost fidelity check rather than a per-inference explanation, given its computational overhead [59], [65].

13.3 Consistency and Trust Auditing

EMSF includes a lightweight consistency-auditing module that flags predictions where Grad-CAM saliency falls outside physiologically plausible facial ROIs, or where attention weight and SHAP attribution disagree substantially, surfacing such cases for human review rather than presenting a single unqualified confidence score [65], [68]. This design directly addresses the concern, raised in Section 11.6, that post-hoc explanations can sometimes provide misleading confidence rather than genuine insight into model behavior.

13.4 Privacy-by-Design

Given the sensitivity of thermal and physiological data, EMSF favors on-device or edge inference for the modality-specific encoders, with only aggregated, privacy-reduced embeddings (rather than raw thermal frames) transmitted for any centralized fusion or auditing step, and recommends evaluation of differential-privacy and federated-learning variants of the fusion module consistent with recent affect-recognition privacy research [78].

14. FUTURE RESEARCH DIRECTIONS

1. Construction and public release of large-scale, demographically diverse, synchronized thermal-visual-physiological stress datasets, ideally collected under standardized protocols such as the TSST across multiple sites, to serve as a WESAD-equivalent benchmark for thermal stress research [4], [18], [70].

2. Systematic benchmarking of Vision Transformer and hybrid CNN-Transformer architectures against CNN and CNN-LSTM baselines specifically on thermal facial sequences, including efficient variants suitable for edge deployment [52]-[54].
3. Development of ante-hoc, architecture-intrinsic explainability mechanisms (e.g., prototype-based or concept-bottleneck layers) co-designed with multimodal fusion architectures, reducing reliance on post-hoc approximations alone [63], [69].
4. Standardized, human-centered evaluation protocols for comparing Grad-CAM, SHAP, LIME, and attention visualization on fidelity, stability under noise, computational cost, and clinician/end-user comprehensibility, specifically within thermal and affective-computing contexts [64]-[68].
5. Federated and differentially private training pipelines tailored to thermal and physiological data, enabling multi-institution collaboration without centralizing sensitive raw signals [78].
6. Cross-population and longitudinal validation studies explicitly measuring fairness and generalization across age, skin tone, and cultural background, extending fairness-auditing practices already emerging in RGB facial-expression recognition to the thermal domain [79].
7. Integration of uncertainty quantification alongside explainability, so that both the confidence and the reasoning behind a stress prediction are communicated to end users and clinicians [63], [68].
8. Exploration of lightweight, quantized, or distilled hybrid CNN-Transformer models to reconcile the accuracy benefits of attention-based fusion with the strict latency and power budgets of real-time, wearable-adjacent, or embedded thermal-camera deployments [32], [40].

15. CONCLUSION

Non-contact, thermally grounded stress detection sits at a productive but still-maturing intersection of psychophysiology, computer vision, multimodal deep learning, and explainable AI. This survey has synthesized evidence that thermal facial imaging provides a physiologically direct, contactless proxy for autonomic stress responses [18], [27]-[29]; that multimodal fusion - whether early, late, or intermediate - generally improves robustness over unimodal approaches, provided fusion strategy and modality selection are chosen carefully [30]-[33], [40]; that CNNs, ResNets, Vision Transformers, LSTMs, and hybrid CNN-Transformer architectures each offer complementary strengths for spatial, temporal, and cross-modal modeling of thermal-physiological signals [44]-[54]; and that Grad-CAM, SHAP, LIME, and attention visualization each address different, only partially overlapping, facets of the explainability problem, with no single method dominant across fidelity, stability, and cost [57]-[68]. Persistent challenges - thermal dataset scarcity, environmental confounds, real-time deployment constraints, cross-population generalization, and privacy - define a clear research agenda for the coming years [76]-[79]. The Explainable Multimodal Stress-Sensing Framework proposed in Section 13 offers one concrete direction for integrating layered, modality-matched explainability directly into multimodal thermal stress-detection pipelines, with the aim of producing systems that are not only accurate but also auditable, trustworthy, and appropriately privacy-preserving. Realizing this vision will require sustained, coordinated effort across dataset creation, architecture design, explainability research, and privacy engineering, but the convergence of these four threads makes non-contact, explainable stress sensing an increasingly tractable and consequential research target.

REFERENCES

[1] B. S. McEwen, "Physiology and Neurobiology of Stress and Adaptation: Central Role of the Brain," Physiological Reviews, vol. 87, no. 3, pp. 873–904, Jul. 2007.

[2] J. M. Nash and R. W. Thebarge, "Understanding Psychological Stress, Its Biological Processes, and Impact on Primary Headache," Headache: The Journal of Head and Face Pain, vol. 46, no. 9, pp. 1377–1386, Oct. 2006.

[3] A. F. T. Arnsten, "Stress Signalling Pathways that Impair Prefrontal Cortex Structure and Function," Nature Reviews Neuroscience, vol. 10, no. 6, pp. 410–422, Jun. 2009.

[4] C. Kirschbaum, K.-M. Pirke, and D. H. Hellhammer, "The 'Trier Social Stress Test' – A Tool for Investigating Psychobiological Stress Responses in a Laboratory Setting," Neuropsychobiology, vol. 28, no. 1–2, pp. 76–81, 1993.

[5] R. W. Picard, Affective Computing. Cambridge, MA, USA: MIT Press, 1997.

[6] P. Ekman and W. V. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement. Palo Alto, CA, USA: Consulting Psychologists Press, 1978.

[7] S. S. Panicker and P. Gayathri, "A Survey of Machine Learning Techniques in Physiology Based Mental Stress Detection Systems," Biocybernetics and Biomedical Engineering, vol. 39, no. 2, pp. 444–469, 2019.

[8] Y. S. Can, B. Arnrich, and C. Ersoy, "Stress Detection in Daily Life Scenarios Using Smart Phones and Wearable Sensors: A Survey," Journal of Biomedical Informatics, vol. 92, p. 103139, 2019.

[9] J. A. Healey and R. W. Picard, "Detecting Stress During Real-World Driving Tasks Using Physiological Sensors," IEEE Transactions on Intelligent Transportation Systems, vol. 6, no. 2, pp. 156–166, Jun. 2005.

[10] K. Plarre et al., "Continuous Inference of Psychological Stress from Sensory Measurements Collected in the Natural Environment," in Proc. 10th ACM/IEEE Int. Conf. Information Processing in Sensor Networks (IPSN), 2011, pp. 97–108.

[11] E. Smets et al., "Large-Scale Wearable Data Reveal Digital Phenotypes for Daily-Life Stress Detection," npj Digital Medicine, vol. 1, no. 1, pp. 1–10, 2018.

[12] H. Lu et al., "StressSense: Detecting Stress in Unconstrained Acoustic Environments Using Smartphones," in Proc. ACM Conf. Ubiquitous Computing (UbiComp), 2012, pp. 351–360.

[13] P. Siirtola, "Continuous Stress Detection Using the Sensors of Commercial Smartwatch," in Adjunct Proc. ACM Int. Joint Conf. Pervasive and Ubiquitous Computing and Proc. ACM Int. Symp. Wearable Computers, 2019, pp. 1198–1201.

[14] S. Chakraborty, S. Aich, M. Joo, M. Sain, and H.-C. Kim, "A Multichannel Convolutional Neural Network Architecture for the Detection of the State of Mind Using Physiological Signals from Wearable Devices," Journal of Healthcare Engineering, vol. 2019, 2019.

[15] P. Schmidt, R. Dürichen, A. Reiss, K. Van Laerhoven, and T. Plötz, "Multi-Target Affect Detection in the Wild: An Exploratory Study," in Proc. 23rd Int. Symp. Wearable Computers (ISWC), 2019, pp. 211–219.

[16] M. T. Uddin and S. Canavan, "Synthesizing Physiological and Motion Data for Stress and Meditation Detection," in Proc. 8th Int. Conf. Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), 2019, pp. 244–247.

[17] J. Birjandtalab, D. Cogan, M. B. Pouyan, and M. Nourani, "A Non-EEG Biosignals Dataset for Assessment and Visualization of Neurological Status," in Proc. IEEE Int. Workshop Signal Processing Systems (SiPS), 2016, pp. 110–114.

[18] D. B. L. Arasu, A. S. A. Mohamed, N. I. R. Ruhaiyem, N. Annamalai, S. L. Lutfi, and M. M. Al Qudah, "Human Stress Recognition from Facial Thermal-Based Signature: A Literature Survey," Computer Modeling in Engineering & Sciences, vol. 130, no. 2, pp. 633–652, 2022.

[19] Y. Cho, N. Bianchi-Berthouze, and S. J. Julier, "DeepBreath: Deep Learning of Breathing Patterns for Automatic Stress Recognition Using Low-Cost Thermal Imaging in Unconstrained Settings," in Proc. 7th Int. Conf. Affective Computing and Intelligent Interaction (ACII), 2017. arXiv:1708.06026.

[20] P. Yuen, K. Hong, T. Chen, A. Tsitiridis, F. Kam, J. Jackman, D. James, M. Richardson, L. Williams, and W. Oxford, "Emotional & Physical Stress Detection and Classification Using Thermal Imaging Technique," in Proc. 3rd Int. Conf. Imaging for Crime Detection and Prevention (ICDP), 2009.

[21] M. Kopaczka, R. Kolk, and D. Merhof, "A Fully Automatic Approach to Facial Feature Detection and Tracking on Thermal Infrared Image Sequences," in Proc. IEEE Int. Symp. Biomedical Imaging (ISBI), 2016.

[22] "Detecting Changes in Facial Temperature Induced by a Sudden Auditory Stimulus Based on Deep Learning-Assisted Face Tracking," Scientific Reports, vol. 9, 2019.

[23] "On Classification of the Human Emotions from Facial Thermal Images: A Case Study Based on Machine Learning," Big Data and Cognitive Computing, vol. 7, no. 2, p. 27, Mar. 2025.

[24] "Deep-Learning-Based Stress Recognition with Spatial-Temporal Facial Information," Sensors, vol. 21, no. 22, p. 7498, Nov. 2021.

[25] "Impact of Imaging Protocols on Convolutional Neural Network-Based Pressure Injury Detection," PMC Open Access Article, 2025.

[26] G. J. Tattersall, "Infrared Thermography: A Non-Invasive Window into Thermal Physiology," Comparative Biochemistry and Physiology Part A, vol. 202, pp. 78–98, 2016.

[27] I. Pavlidis, J. Dowdall, N. Sun, C. Puri, J. Fei, and M. Garbey, "Interacting with Human Physiology," Computer Vision and Image Understanding, vol. 108, no. 1–2, pp. 150–170, 2007.

[28] D. Shastri, M. Papadakis, P. Tsiamyrtzis, B. Bass, and I. Pavlidis, "Perinasal Imaging of Physiological Stress and Its Affective Potential," IEEE Transactions on Affective Computing, vol. 3, no. 3, pp. 366–378, 2012.

[29] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion," Information Fusion, vol. 37, pp. 98–125, 2017.

[30] Z. X. Su et al., "A Multi-Modal Deep Learning Approach for Stress Detection Using Physiological Signals: Integrating Time and Frequency Domain Features," Frontiers in Physiology, 2025.

[31] Y. Hasanpoor, "Multimodal Signal Fusion for Stress Detection Using Deep Neural Networks: A Novel Approach for Converting 1D Signals to Unified 2D Images," arXiv:2509.13636, 2025.

[32] "A Multimodal Intermediate Fusion Network with Manifold Learning for Stress Detection," arXiv:2403.08077, 2024.

[33] W. Seo et al., "Deep Learning Feature-Level Fusion of ECG, Respiration, and Facial Features for Stress Detection," as reported in comparative multimodal-fusion literature, 2020.

[34] S. Hosseini et al., "A Multimodal Sensor Dataset for Continuous Stress Detection of Nurses in a Hospital," Scientific Data, vol. 9, p. 255, 2022.

[35] "A Multimodal Stress Detection Dataset with Facial Expressions and Physiological Signals," Scientific Data, 2025.

[36] J. Lin, S. Pan, C. S. Lee, and S. Oviatt, "An Explainable Deep Fusion Network for Affect Recognition Using Physiological Signals," in Proc. 28th ACM Int. Conf. Information and Knowledge Management (CIKM), 2019, pp. 2069–2072.

[37] Z. Ahmad and N. Khan, "Human Action Recognition Using Deep Multi-Level Multimodal (M2) Fusion of Depth and Inertial Sensors," IEEE Sensors Journal, 2019.

[38] "Multi-Level Stress Assessment Using Multi-Domain Fusion of ECG Signal," arXiv:2008.05503, 2020.

[39] "Multimodal Wearable-Based Stress Detection Using Machine Learning: A Systematic Review of Validation Protocols and Generalization Gaps (2021–2025)," preprint, 2026.

[40] "A Deep Learning Approach for Early Stress Detection Using Electrodermal Activity Through Wearable Devices," in Proc. Int. Conf. Pervasive Computing Technologies for Healthcare, Springer, 2025.

[41] A. Ometov et al., "Reviewing Open-Access Multimodal Datasets and AI Models for Mental State Detection," 2025.

[42] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in Advances in Neural Information Processing Systems (NeurIPS), 2012, pp. 1097–1105.

[43] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied to Document Recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[44] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in Proc. Int. Conf. Learning Representations (ICLR), 2015.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[46] C. Szegedy et al., "Going Deeper with Convolutions," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9.

[47] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[48] F. J. Ordóñez and D. Roggen, "Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition," Sensors, vol. 16, no. 1, p. 115, 2016.

[49] A. Vaswani et al., "Attention Is All You Need," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.

[50] A. Dosovitskiy et al., "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale," in Proc. Int. Conf. Learning Representations (ICLR), 2021. arXiv:2010.11929.

[51] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2021, pp. 9992–10002.

[52] K. Han et al., "A Survey on Vision Transformer," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 1, pp. 87–110, 2023.

[53] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for Simplicity: The All Convolutional Net," in Proc. ICLR Workshop, 2015.

[54] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," International Journal of Computer Vision, vol. 128, pp. 336–359, 2020 (originally in Proc. IEEE ICCV, 2017, pp. 618–626).

[55] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 4765–4774.

[56] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), 2016, pp. 1135–1144.

[57] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks," in Proc. IEEE Winter Conf. Applications of Computer Vision (WACV), 2018, pp. 839–847.

[58] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," PLOS ONE, vol. 10, no. 7, e0130140, 2015.

[59] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," in Proc. ICLR Workshop, 2014.

[60] F. Doshi-Velez and B. Kim, "Towards a Rigorous Science of Interpretable Machine Learning," arXiv:1702.08608, 2017.