Advances in Science and Technology Research Journal

Advances in Science and Technology Research Journal, 2026, 20(4), 457–474
https://doi.org/10.12913/22998624/214572
ISSN 2299-8624, License CC-BY 4.0

Received: 2025.08.28
Accepted: 2026.11.21
Published: 2026.03.01

Gompertz-fuzzy ensemble of lightweight convolutional neural networks for stress classification

Katarzyna Baran[1]

[1] Department of Computer Science, Faculty of Electrical Engineering and Computer Science,
Lublin University of Technology, ul. Nadbystrzycka 36B, Lublin, Poland
E-mail: k.baran@pollub.pl

ABSTRACT

Contemporary research on stress highlights its significant impact on both physical and mental health, prompting the pursuit of objective methods for measuring this phenomenon. In response to this challenge, the present article proposes an innovative hybrid stress classification method that combines nonlinear Gompertz weighting with adaptive fuzzy logic within an ensemble learning framework of lightweight convolutional neural networks (lightweight CNNs). The key scientific contributions include: an innovative integration of the Gompertz function with a fuzzy system for assessing prediction confidence, comprehensive validation of the approach on a modified version of the comprehensive facial thermal dataset (CFTD), where the original emotion classes were mapped to three stress levels: no stress, low stress, and high stress. The study used six lightweight CNN architectures – MobileNetV3-Lite, TinyNet-E, FBNetV3, CondenseNetV2, Nanonet, and ShuffleNetV2 – whose predictions were aggregated through a three-stage process: initial nonlinear weighting by the Gompertz function, fuzzy scaling of weights based on classification confidence, and final fusion using fuzzy rules. Experiments were conducted in two variants – using a single thermal palette and mixed palettes – with 5-fold cross-validation. Results demonstrated that using a single thermal palette achieved significantly higher average accuracy (MobileNetV3-Lite: 80.1%) compared to the mixed palettes variant (78.2%). The hybrid approach, combining the Gompertz function and fuzzy logic, significantly improved classification performance by reducing errors by 19–30% depending on the stress class, particularly for the "high-stress" class and scenarios with marked prediction uncertainty. The best performance was observed with the MobileNetV3-Lite architecture, which, thanks to advanced attention mechanisms (SE blocks), effectively leverages thermal representation. Furthermore, fuzzy logic helped mitigate the negative influence of weaker models, resulting in enhanced stability and reliability of the stress classification system.

Keywords: Gompertz function, fuzzy ensemble learning, lightweight convolutional networks, stress recognition.

INTRODUCTION

Contemporary research on stress highlights its significant negative impact on both physical and mental health [1], which motivates the search for objective, non-invasive methods of measurement. Affective computing offers promising solutions, among which thermal imaging stands out as a robust tool for stress recognition [2]. Unlike visible-light cameras, thermal sensors capture physiological correlates of stress (e.g., changes in blood flow, perspiration) by recording facial temperature patterns, which are difficult to manipulate consciously [3].

Despite this potential, existing machine learning methods for stress classification based on thermal imaging face several challenges. Many studies rely on single, complex models that are prone to overfitting, particularly on limited datasets – a common issue in this field [4]. Moreover, model performance can be severely affected by data heterogeneity, such as the use of different thermal palettes (e.g., Iron, Rainbow), which alter the visual representation of identical thermal information [5]. Although ensemble methods improve accuracy and robustness [6], standard techniques like majority voting or weighted averaging often fail to fully leverage individual models' strengths. Linear weighting schemes ignore nonlinear performance dependencies and lack mechanisms to dynamically adjust a model's influence based on its prediction confidence for each sample [7].

To address these challenges, this article proposes a novel hybrid ensemble learning method for stress classification. The main scientific contribution lies in the innovative integration of Gompertz function-based nonlinear weighting with an adaptive fuzzy logic system to evaluate prediction confidence within an ensemble of lightweight convolutional neural networks (CNNs). This approach is comprehensively validated on a modified version of the comprehensive facial thermal dataset (CFTD), where original emotion classes are remapped into three stress levels: no stress, low stress, and high stress.

The key scientific contributions of this work are fourfold:

- A novel hybrid ensemble framework – the author introduces a three-stage aggregation process – nonlinear weighting of base models using a calibrated Gompertz function to amplify stronger models, dynamic adjustment of these weights via a fuzzy logic system based on per-prediction confidence, and final fusion using fuzzy rules. This represents a novel methodological approach to handling uncertainty in ensemble decision-making.
- Rigorous cross-palette validation – the author conducts an extensive evaluation under two data scenarios: a consistent thermal palette (Iron) and a mixed-palette variant, providing valuable insights into how data representation consistency impacts model performance. This addresses a critical, often overlooked challenge in thermal image analysis.
- Comprehensive empirical analysis – the author evaluates six modern lightweight CNN architectures (MobileNetV3-Lite, TinyNet-E, FBNetV3, CondenseNetV2, NanoNet, ShuffleNetV2) and benchmarks the proposed ensemble against standard approaches (Majority Voting, Weighted Averaging, standalone Gompertz weighting). This provides a valuable benchmark for the research community.
- In-depth statistical validation – the author applies robust statistical tests – including paired t-tests, two-way ANOVA, and McNemar's test – to scientifically validate the superiority of the proposed method, demonstrating statistically significant error reductions, particularly for the critical "high stress" class.

The structure of this article is as follows: Related work reviews the literature on stress recognition, thermal imaging, and ensemble methods. Research project describes the study design, dataset, base CNN models, and proposed ensemble methodology. Experiments and Results present the experimental findings and discussion. Finally, Conclusion summarizes the paper and outlines future research directions.

RELATED WORKS

Affective computing is a research field focused on recognizing, interpreting, and processing human emotional states using computational systems. One of its central challenges is the objective and non-invasive measurement of psychological stress. Traditional approaches – based on self-reporting or physiological measurements such as EDA (electrodermal activity) or ECG (electrocardiography) – tend to be intrusive for users and are sensitive to motion artifacts. In this context, thermal imaging emerges as a promising alternative, as it captures changes in facial skin thermoregulation directly governed by the autonomic nervous system (ANS) – the primary mediator of the stress response [8]. Non-invasiveness, safety, and user-friendliness are among the key advantages of thermal methods.

The physiological foundation of thermographic stress recognition lies in studies examining blood flow changes in facial vessels. Stress activates the sympathetic branch of the ANS, inducing vasoconstriction in specific regions (e.g., nose, forehead), leading to visible cooling in thermal images, while other regions (e.g., cheeks) may exhibit warming [9]. A systematic review in [10] confirmed the high validity of thermography for emotion research while also identifying its limitations.

In the context of automatic classification methods, three evolving paradigms can be distinguished:

- Hand-crafted feature methods – some studies, e.g., [11], extracted statistical measures (mean, temperature variance) from manually or semi-automatically defined regions of interest (ROI) and used conventional classifiers such as SVM (support vector machines) [12]

or Random Forest to differentiate stress and relaxation states.
- Shallow neural network methods – with the rise of deep learning, convolutional neural networks (CNNs) began to be applied for automatic feature extraction. Early studies typically relied on shallow architectures trained from scratch on relatively small thermal datasets [13].

Deep transfer learning methods – the most recent and effective approaches employ deep CNNs pre-trained on large datasets (e.g., ImageNet) and fine-tune them for thermal classification tasks. In [14], this approach demonstrated clear superiority, achieving substantially better performance in valence/arousal classification compared to traditional methods.

Despite progress, major challenges remain. The most significant is the lack of large, standardized, publicly available thermal datasets dedicated to stress research, which hinders training of very deep models from scratch. Another important issue is sensitivity to data acquisition conditions. As shown in [15], changing a thermal camera's color palette can drastically degrade accuracy in models not trained for such variation. This limitation directly motivates the experimental framework of this work, which compares single-palette and mixed-palette scenarios.

Deploying facial and emotion recognition systems on mobile devices, embedded systems, or human–computer interfaces (HCI) requires models with low computational and memory demands, capable of operating in near-real time. This necessity has driven the rapid development of efficient lightweight CNN architectures. Their evolution has progressed from simple compression techniques (pruning, quantization) applied to large models toward designing efficient building blocks from scratch. A breakthrough came with MobileNet [16], which employs depthwise separable convolutions to drastically reduce parameters and FLOPs (floating point operations) with minimal accuracy loss. ShuffleNet [17] further improved efficiency through channel shuffle and group convolution operations. The next leap was the use of neural architecture search (NAS) to automatically design hardware-optimized architectures, producing models such as FBNet [18] and TinyNet [19], which achieve an excellent balance between accuracy and latency on target devices. In parallel, architectures based on dense connectivity and dynamic routing, such as CondenseNet [20], have reduced redundant computations. This study contributes to this line of research by providing a systematic comparison of state-of-the-art lightweight architectures for thermal classification, offering a valuable benchmark for the research community.

Ensemble methods, which combine predictions from multiple (homogeneous or heterogeneous) models, are a proven strategy for improving accuracy, robustness, and generalization of learning systems. Fundamental techniques such as bagging (e.g., random forests) [21] and boosting (e.g., XGBoost) [22] have been successfully adapted to deep learning. In the context of deep neural networks, deep ensembles [23] train multiple models from different initializations and aggregate their predictions via simple averaging. This method, though conceptually simple, is highly effective and improves model uncertainty calibration. Other strategies include Snapshot Ensembling [24] and Stochastic Weight Averaging (SWA) [25], which produce diverse model sets within a single training process. However, most standard ensemble methods treat all base models equally (e.g., majority voting) or assign static weights (e.g., weighted averaging) based solely on global validation accuracy. Such approaches ignore a critical fact: a model's prediction confidence can vary significantly per individual sample. A globally strong model may be uncertain for a particular input, while a weaker model may produce a highly confident and accurate prediction for the same instance. This fundamental limitation of both traditional and some modern ensemble methods motivates the author's proposed approach, which specifically addresses sample-level prediction uncertainty through dynamic, confidence-aware weighting. However, many advanced ensemble and calibration techniques, such as Bayesian deep ensembles [23] or those utilizing Monte Carlo dropout, come with a significantly higher computational cost during both training and inference. This makes them less suitable for building efficient ensembles from multiple, already lightweight CNN models – the primary focus of our work. The proposed method addresses the core limitation of static weighting in a computationally efficient manner, dynamically adjusting model influence based on per-sample prediction confidence without the prohibitive cost of full Bayesian inference.

Fuzzy logic [26] provides a mathematical framework for modeling uncertainty and imprecision inherent to real-world systems. Its strength lies in representing approximate, linguistic reasoning using concepts such as fuzzy sets and membership functions. In machine learning, fuzzy logic has been applied to adaptive neuro-fuzzy inference systems (ANFIS) [27], which combine learning ability with interpretability, to prediction uncertainty modeling and ensemble weighting [28], and to aggregation of outputs using expert-derived if–then rules [29]. Studies such as [28] show that sample-level fuzzy ensemble weighting based on confidence can outperform traditional static methods. However, the combination of nonlinear Gompertz function-based initial weighting with an adaptive fuzzy system for dynamically adjusting per-sample model weights – applied to lightweight CNN ensembles for thermal classification – represents an unexplored research gap that this work aims to fill.

MATERIALS AND METHODS

Within the research project “Recognition of Human Stress Using Thermographic Imaging and Neural Networks” approved by the Ethics Committee No. 4/2024 dated February 19, 2024, a classification study was conducted with the following assumptions:

- utilization of thermal images from the Comprehensive Facial Thermal Dataset [30] for stress classification,
- application of lightweight convolutional neural networks: MobileNetV3, TinyNet-E, FBNetV3, CondenseNetV2, NanoNet, ShuffleNetV2,
- conducting a two-stage experimental study: multiclass stress classification using images with a single thermal palette and mixed thermal palettes, aiming to analyze the impact of thermal palette consistency on classification accuracy,
- implementation of a hybrid ensemble learning approach combining the Gompertz function and fuzzy logic, with comparison of results against other ensemble methods.

Comprehensive facial thermal dataset

Comprehensive facial thermal dataset (CFTD) is a relatively new database created in 2024 [30]. It consists of 2,250 thermal images captured using a UNI-T UTi165A camera during recordings related to emotions. For each subject, 225 facial recognition images and 450 emotion-specific images were obtained. The emotions included: happy, sad, angry, neutral, and surprise. Images were recorded under various conditions, with different thermal palettes (iron, rainbow, iceblue, white hot), angles of photography, and zoom levels. As noted by the authors of the CFTD, the dataset has broad applications in fields such as security, healthcare, and human-computer interaction. Considering the structure of this dataset, the present study reclassified the original emotion classes into three stress levels: no stress (NS), low stress (LS), and high stress (HS). Based on established psychophysiological literature, the original emotion classes were remapped into three stress levels: no stress (NS), low stress (LS), and high stress (HS). The mapping was defined as follows: no stress – neutral and happy; low stress – sad; high stress – angry and surprise. This schema is supported by studies such as [31], which found that anger and surprise are associated with significant cardiovascular stress responses characteristic of high arousal. Furthermore, the ‘sad’ emotion, categorized here as low stress, is often linked to withdrawal-related physiological patterns distinct from the high-arousal fight-or-flight response [32, 33]. While we acknowledge that emotion-stress mapping is complex and can be context-dependent, the chosen scheme provides a pragmatic and psychophysiologically-grounded framework for this initial computational study, aligning with established literature [34, 35]. Experiments were conducted in two variants: using a single thermal palette (iron, 300 images) and mixed thermal palettes (1.200 images). Class balancing was ensured. The choice of the Iron palette for the first stage of the study was dictated by key scientific and technical factors. The Iron palette provides better contrast, minimizes artifacts, and is considered a medical standard – it is recommended for facial thermographic analysis due to its linear perceptual response (L* of Lab color space) and compatibility with skin segmentation algorithms. Most affective thermography studies use this palette as a baseline, allowing direct comparisons. Geometric transformations do not produce color artifacts in this palette. The selection of this palette represents a compromise between sensitivity, specificity, and repeatability. Due to the relatively small size of the CFTD, 5-fold cross-validation was applied with splits into training (60%), validation (20%), and testing (20%) sets. Given the scale of the CFTD dataset, a primary concern is the risk of overfitting. To ensure robust generalization and model evaluation, the author implemented a comprehensive strategy:

1. Subject-independent cross-validation – a 5-fold cross-validation scheme was strictly enforced, where all images from a single subject were confined to either the training, validation, or test set within a single fold. This prevents data leakage and ensures a more realistic estimate of generalization to new, unseen individuals.
2. Extensive data augmentation – as detailed in the 'CNNs model' subsection, both geometric and thermal transformations were applied.
3. Transfer learning & regularization – the two-phase transfer learning approach (feature extraction + fine-tuning) leveraging pre-trained weights, coupled with the use of the AdamW optimizer (weight decay = 1 e^{-4}), provided strong regularization.

Figure 1 presents sample images of various emotions from the CFTD across all thermal palettes.

CNNs model

The study utilized six lightweight convolutional neural network (CNN) architectures: MobileNetV3-Lite, TinyNet-E, FBNetV3, CondenseNetV2, NanoNet, and ShuffleNetV2 – all specifically designed for computational efficiency. These models typically feature parameter counts of 1.5–2.5 million and computational requirements below 0.15 GFLOPs for 224 × 224 input resolution [36–42], making them suitable for real-time applications and embedded systems. Their lightweight nature was a key selection criterion, ensuring the practical applicability of the author's stress classification system. The selection of these models was driven by the requirement for computational efficiency while maintaining the capability to extract relevant thermographic features. The training process was conducted with the following parameter settings: AdamW optimizer with a weight decay of 1 e^{-4}, batch size of 32, two-phase transfer learning consisting of 50 epochs for the feature extraction phase and 50 epochs for the fine-tuning phase, learning rates of 0.001 and 0.0001 for the first and second phases respectively, cosine annealing schedule, CrossEntropyLoss as the loss function, and data augmentation including geometric transformations (random horizontal flip, rotation, random crop, zoom) and thermal transformations (Gaussian noise, contrast adjustment). The study preserved the original colors of the thermal palettes but applied additional preprocessing to ensure comparability of results. Specifically, for the Iron palette, original pixel values (RAW) were retained; for the Rainbow palette, images were converted to grayscale and normalized to the [0,1] range to reduce color artifacts; for the White Hot palette, the luminance channel (Y from YCbCr) was extracted; and for the Iceblue palette, conversion to Lab color space was performed and the L channel was used. For fine-tuning, a strategy of unfreezing the last 20% of layers (counted in full residual blocks) in each network was adopted. This percentage-based unfreezing accounts for varying architecture depths, ensuring comparable levels of fine-tuning. To guarantee comparability of results, a consistent scheme for modifying the original architectures was applied by replacing

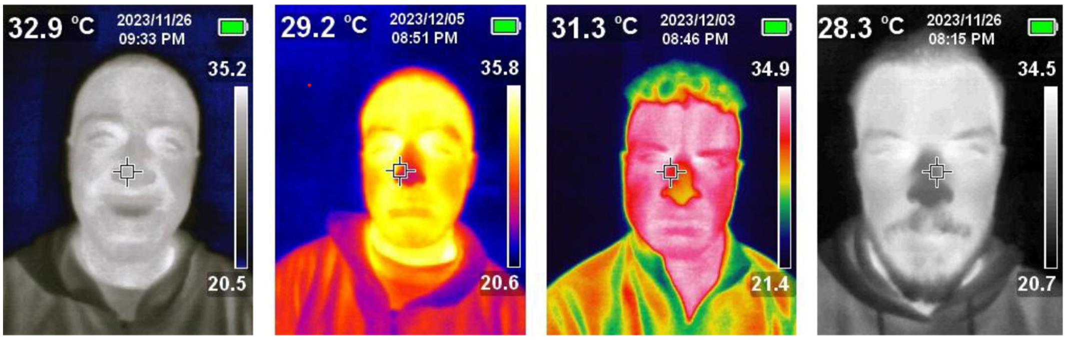

Figure 1. Sample thermal images from the comprehensive facial thermal dataset – from left to right: happy emotion – Iceblue palette, sad emotion – Iron palette, angry emotion – Rainbow palette, neutral emotion – White Hot palette

the original classification layers with a classifier structured as follows: a Flatten layer, a Dense layer with 256 units and ReLU activation, and an output Dense layer with softmax activation. Standardizing the classifier eliminates the influence of differences in the original output layers on performance. The classifier structure was experimentally selected as optimal between efficiency and accuracy. All machine learning computations were performed on an Intel Core i7-10700K CPU with an NVIDIA GeForce RTX 3070 GPU and Intel UHD Graphics 630. The programming language used was Python 3.10. Frameworks included TensorFlow 2.10, PyTorch 2.0+cu118, and Keras 3.9.1.

The six lightweight CNN architectures selected for this study represent the state-of-the-art in efficient deep learning, each employing unique strategies for balancing accuracy and computational cost. MobileNetV3-Lite [37] utilizes neural architecture search (NAS) to optimize mobile performance. TinyNet-E [38] applies a compound scaling approach to achieve high efficiency. FBNetV3 [39] jointly optimizes both architecture and training recipes via NAS. CondenseNetV2 [40] introduces sparse feature reactivation for enhanced feature reuse. NanoNet [41] employs an encoder-decoder structure tailored for medical imaging. ShuffleNetV2 [42] uses channel shuffle operations for efficient interchannel communication. The literature also reports improvements, such as AugShuffleNet [43]. Detailed architectural descriptions of each model are provided in the Supplementary Material. Their computational characteristics are summarized in Table 1, confirming that all models operate with low parameter counts (1.8–2.4 M) and computational complexity (<0.22 GFLOPs), enabling real-time inference on embedded systems. The values in Table 1 are based on standard benchmarks from original publications [37–43] and empirical measurements on the author's setup (NVIDIA GeForce RTX 3070, PyTorch 2.0). FLOPs were calculated using the 'thop' library, while inference time was averaged over 100 forward passes.

Proposed CNN ensemble model

Ensemble learning represents an advanced approach in machine learning, involving the integration of predictions from multiple models to achieve improvements in both classification accuracy metrics and the stability of the decision system. In the context of medical applications, where input data often exhibit significant heterogeneity (e.g., variability in thermographic measurements between patients), the following aggregation methods are particularly popular:

- majority voting – the simplest ensemble method involves each model casting a vote for a single class, with the final result being the class that receives the majority of votes. Types of majority voting include hard voting and soft voting. The advantages of this approach are its simplicity, speed, and the stabilization of weaker models' performance. However, a drawback is that all models are treated equally, regardless of their quality. Therefore, majority voting is best used with models of similar performance. The mathematical formula is as follows:

$$\hat{y} = model\{f_1(x), f_2(x), \dots, f_n(x)\} \tag{1}$$

where: $f_i(x)$ denotes the prediction of the i-th model.

- weighted averaging – each model assigns probabilities to classes, and the final result is the weighted average of these probabilities, where each model has its own weight. Thus, better-performing models can have a greater influence; however, appropriate weights must be determined (e.g., based on accuracy). This approach, however, does not account for nonlinear relationships between model performance and the optimal contribution to the ensemble. The mathematical expression for this method is:

$$\hat{p}(y|x) = \sum_{i=1}^{N} w_i p_i(y|x), \quad w_i = \frac{acc_i}{\sum_{j=1}^{N} acc_j} \tag{2}$$

where: w_i is the weight proportional to the validation accuracy.

- stacking – a more advanced ensemble method used when the goal is to maximize performance involves training a new model, known as a meta-model, which combines the outputs of other models instead of making decisions through simple aggregation. This allows stacking to be flexible and capable of capturing complex relationships between models.

Table 1. Computational characteristics of the lightweight CNN architectures

Model	Parameters (M)	FLOPs (G)	Inference time (ms)
MobileNetV3-Lite	2.4	0.22	8.1
TinyNet-E	2.1	0.19	7.6
FBNetV3	2.3	0.21	8.0
CondenseNetV2	2.0	0.18	7.4
NanoNet	1.8	0.16	6.9
ShuffleNetV2	1.9	0.17	7.2

However, its implementation is more challenging and requires more data. The meta-model often needs additional training and may lead to overfitting, especially on small datasets.

The key methodological innovation of this work was the development of a hybrid ensemble learning system, combining nonlinear Gompertz weighting with adaptive fuzzy logic. The prediction aggregation process consisted of three stages:

- nonlinear Gompertz weighting

The Gompertz function, adapter for scaling model weights, is defined as:

$$w_i^{Gompertz} = a \cdot \\ \cdot \exp\left(-b \cdot \exp(-c \cdot x_i)\right) + d \tag{3}$$

where: x_i is the validation accuracy of the i-th model.

The parameters (a=0.7, b=2.5, c=0.8, d=0.3) were selected via an extensive grid search optimization process aimed at maximizing the ensemble's accuracy on the validation set. The search space was defined to find a sigmoidal Gompertz curve that effectively differentiates model performance. The chosen parameters yield a curve with a steep transition region between 70% and 75% validation accuracy. This shape aggressively amplifies the influence of higher-performing models (accuracy >75%), while significantly suppressing the contribution of weaker models (accuracy <70%), thus acting as a performance-based filter. This non-linear weighting is a key advantage over linear averaging, as it more robustly mitigates the negative impact of poorly performing ensemble members.

- dynamic weight adjustment based on fuzzy confidence assessment

For each prediction, a confidence value conf$_i$ ∈ [0, 1] is calculated based on the probability distribution:

$$conf_i = 1 - \frac{H(p_i)}{\log(K)}, H(p_i) = \\ = -\sum_{k=1}^{K} p_i(y_k|x) log\, p_i(y_k|x) \tag{4}$$

where: $K = 3$ is the number of classes. The final model weight is calculated as:

$$w_i = w_i^{Gompertz} \cdot \mu_A(conf_i) \tag{5}$$

The membership function m$_A$ for the fuzzy set „high confidence" is defined as:

$$\mu_A(conf) = \begin{cases} 1\, for\, conf\ >\ 0.7 \\ \dfrac{conf - 0.4}{0.3}\, for\ 0.4 \leq\ conf\ \leq 0.7 \\ 0.1\, for\, conf\ \leq\ 0.4 \end{cases} \tag{6}$$

This function implements a simple, interpretable rule: if the confidence $conf_i$ is high (>0.7), the model's Gompertz weight is fully retained. If confidence is moderate (between 0.4 and 0.7), the weight is linearly scaled down. Finally, if confidence is low (<0.4), the model's influence is minimized (m$_A$=0.1), though not entirely removed to preserve a chance of correct prediction. This provides a smooth, adaptive mechanism to reduce the impact of uncertain models on a per-sample basis, improving the system's robustness against difficult-to-classify cases.

- final Fusion using fuzzy rules

The ensemble prediction is computed as a weighted combination:

$$\hat{p}(y|x) = \frac{1}{Z}\sum_{i=1}^{N} w_i p_i(y|x), \\ Z = \sum_{i=1}^{N} w_i \tag{7}$$

where each CNN model generates a probability distribution p$_i$(y|x), and the final result is a weighted average of these predictions, with weights dynamically adjusted based on the model's quality (Gompertz function) and the confidence of the specific prediction (fuzzy logic). This approach enables flexible integration of results from multiple models with varying strengths and reliabilities

The advantages of the proposed approach include adaptability, robustness to borderline cases, and computational efficiency. Adaptability is expressed through the dynamic reduction of the influence of models with low confidence (e.g., when $conf_i > 0.4$, the weight decreases to 10% of its initial value). In terms of robustness, for samples with high entropy (difficult to classify), the system automatically favors models with higher prediction consistency. The operational cost amounts to only O(N×K) multiplications for N models and K classes. In the author's implementation (N=6, K=3), the aggregation time was negligible compared to the forward pass of the CNNs, adding less than 3ms of overhead per sample, thus validating the efficiency claim for real-time applications.

Model evaluation

The evaluation process was based on 5-fold cross-validation. For each split, basic model quality metrics (classification reports) were calculated, and confusion matrices were generated for each individual model. Additionally, four ensemble methods were compared: majority voting, weighted average, Gompertz weighting, and a hybrid of Gompertz weighting with fuzzy logic. Furthermore, a statistical analysis was performed, which included: ANOVA with Tukey's post-hoc test for differences between models, and McNemar's test assessing error reduction after introducing fuzzy logic.

The confusion matrix was generated for each model as a 3 × 3 table (for 3 classes), where rows represent the true classes and columns represent the model's predictions. Each cell shows how many samples from class *i* were classified into class *j*. Important values include: TP (true positive) – correct classifications where the model predicted the positive class (diagonal of the matrix), TN (true negative) – correct predictions of the negative class, FN (false negative) – incorrect underestimation where the model failed to detect the positive class (e.g., high stress classified as low stress), FP (false positive) – incorrect overestimation where the model wrongly assigned the positive class (e.g., no stress classified as high stress). The values in the matrices represent the sum across all folds of the cross-validation.

The model quality evaluation metrics used in the study are:

- Accuracy – the proportion of correctly classified cases to all samples.

$$ACCURACY = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

- Precision – measured of the agreement between positive predictions and actual values.

$$PRECISION = \frac{TP}{TP + FP} \tag{9}$$

- Recall – the model's ability to detect all positive cases.

$$RECALL = \frac{TP}{TP + FN} \tag{10}$$

- F1-score – the harmonic mean of precision and recall.

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{11}$$

Table 2 presents a comparison of the ensemble methods used in this study. As can be seen, despite increased computational complexity, the proposed approach achieves the highest accuracy, which was confirmed by the experiments described in the next chapter.

EXPERIMENTS AND RESULTS

Tables 3–6 present the evaluation metric results for multiclass classification models in both study variants – using a single thermal palette as well as mixed thermal palettes. Figures 2–3 present the graphical confusion matrices of the models for both variants of the study.

Tables 7–8 present the classification report results for the ensemble methods – both on the single thermal pallete and on mixed palettes. Figure 4 shows the confusion matrices for the ensemble methods on the single palette, while Figure 5 shows the confusion matrices for the mixed palettes.

For verification of the classification results, a statistical analysis was conducted, including:

Table 2. Summary of the advantages and disadvantages of the individual ensemble methods

Methods	Advantages	Disadvantages
Majority Voting	Fast, simple to implement	Sensitive to models with similar accuracy
Weighted Average	Take into account varying model quality	Linear weighting is not optimal for imbalanced data
Gompertz	Better adaptation for nonlinear relationships	Requires calibration of parameters (a,b,c)
Gompertz Fuzzy	Highest accuracy, noise robustness	Increased computational complexity

Table 3. Summary of performance metrics for multi-class stress classification using a single thermal palette

Model	State	Precision	Recall	F1-score	Accuracy±SD
MobileNetV3-Lite	NS	0.81	0.80	0.80	80.1% ± 1.1%
	LS	0.76	0.75	0.75	
	HS	0.79	0.78	0.78	
FBNetV3	NS	0.77	0.76	0.76	77.1% ± 1.2%
	LS	0.71	0.70	0.70	
	HS	0.74	0.73	0.73	
CondenseNetV2	NS	0.75	0.74	0.74	75.1% ± 1.5%
	LS	0.69	0.68	0.68	
	HS	0.72	0.71	0.71	
NanoNet	NS	0.71	0.70	0.70	70.3% ± 2.1%
	LS	0.65	0.64	0.64	
	HS	0.67	0.66	0.66	
TinyNet-E	NS	0.79	0.78	0.78	78.5% ± 1.3%
	LS	0.73	0.72	0.72	
	HS	0.76	0.75	0.75	
ShuffleNetV2	NS	0.70	0.69	0.69	68.1% ± 2.3%
	LS	0.65	0.64	0.64	
	HS	0.67	0.66	0.66	

Table 4. Detailed accuracy results for each model from 5-fold cross-validation on the single thermal palette dataset

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Accuracy ±SD
MobileNetV3-Lite	0.798	0.812	0.801	0.794	0.802	0.801 ±0.007
FBNetV3	0.772	0.781	0.769	0.763	0.768	0.771 ±0.007
CondenseNetV2	0.753	0.762	0.751	0.746	0.743	0.751 ±0.007
NanoNet	0.712	0.706	0.698	0.704	0.695	0.703 ±0.007
TinyNet-E	0.792	0.781	0.785	0.778	0.789	0.785 ±0.006
ShuffleNetV2	0.692	0.681	0.673	0.679	0.681	0.681 ±0.007

- Student’s paired t-test (comparison of accuracy between palettes) – to verify if there is a significant difference in accuracy between models tested on the single palette and mixed palette datasets.
- Two-way ANOVA with replication – to analyze the effect of the model type, the palette type, and their interaction. Additionally, Tukey’s HSD post-hoc test was applied to identify pairs of models that differ significantly.
- McNemar’s test for fuzzy logic — to assess whether the fuzzy logic significantly reduces the number of misclassifications.

Student’s paired t-test (comparison of model accuracy between two palettes)

To statistically verify the significance of differences in CNN classification performance between the single palette and mixed palette datasets, a

Table 5. Summary of performance metrics for multi-class stress classification using a mixed thermal palette

Model	State	Precision	Recall	F1-score	Accuracy±SD
MobileNetV3-Lite	NS	0.79	0.77	0.78	78.2% ± 1.3%
	LS	0.75	0.72	0.73	
	HS	0.77	0.80	0.78	
FBNetV3	NS	0.75	0.72	0.73	74.8% ± 2.0%
	LS	0.71	0.68	0.69	
	HS	0.73	0.78	0.75	
CondenseNetV2	NS	0.73	0.69	0.71	72.1% ± 2.0%
	LS	0.69	0.65	0.67	
	HS	0.71	0.75	0.73	
NanoNet	NS	0.69	0.65	0.67	67.9% ± 2.5%
	LS	0.65	0.61	0.63	
	HS	0.67	0.71	0.68	
TinyNet-E	NS	0.77	0.74	0.75	76.2% ± 1.5%
	LS	0.73	0.70	0.71	
	HS	0.75	0.80	0.77	
ShuffleNetV2	NS	0.66	0.62	0.64	64.0% ± 2.7%
	LS	0.63	0.58	0.60	
	HS	0.65	0.68	0.66	

Table 6. Detailed accuracy results for each model from 5-fold cross-validation on the mixed thermal palette dataset

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean ±SD
MobileNetV3-Lite	0.781	0.793	0.778	0.772	0.784	0.782 ± 0.008
FBNetV3	0.752	0.761	0.743	0.738	0.746	0.748 ± 0.009
CondenseNetV2	0.724	0.731	0.718	0.712	0.720	0.721 ± 0.008
NanoNet	0.688	0.682	0.671	0.679	0.675	0.679 ± 0.007
TinyNet-E	0.771	0.763	0.758	0.752	0.766	0.762 ± 0.007
ShuffleNetV2	0.651	0.638	0.627	0.642	0.643	0.640 ± 0.009

paired Student’s t-test was conducted. The analysis compared accuracies obtained through 5-fold cross-validation for the same models under two experimental conditions (single palette vs. mixed palette). The null hypothesis (H_0) tested was that there is no difference between the mean accuracies (m_singlepalette = m_mixedpalette) against the alternative hypothesis (H_1: m_singlepalette > m_mixedpalette). A 95% confidence interval was assumed. Table 9 summarizes the test results, which showed statistically significant higher accuracies for CNN models trained on the single palette dataset ($p<0.05$). The strongest effect was observed for ShuffleNetV2 (+4.1%), while the smallest but still significant effect was for MobileNetV3 (+1.9%). These findings suggest that a consistent thermal representation significantly improves classification performance, with the effect being particularly pronounced for less complex architectures (ShuffleNetV2 vs. MobileNetV3). There is a consistent, statistically significant decrease in accuracy when using the mixed palette dataset.

ANOVA analysis

To assess the significance of differences between CNN models and the impact of the type of thermal dataset on classification accuracy, a two-way analysis of variance (ANOVA) with replications was performed. The analysis included two factors: model architecture (six levels – six CNN models) and type of thermal dataset (two levels: single palette and mixed palettes). The analysis was conducted on accuracies obtained from 5-fold cross-validation, resulting

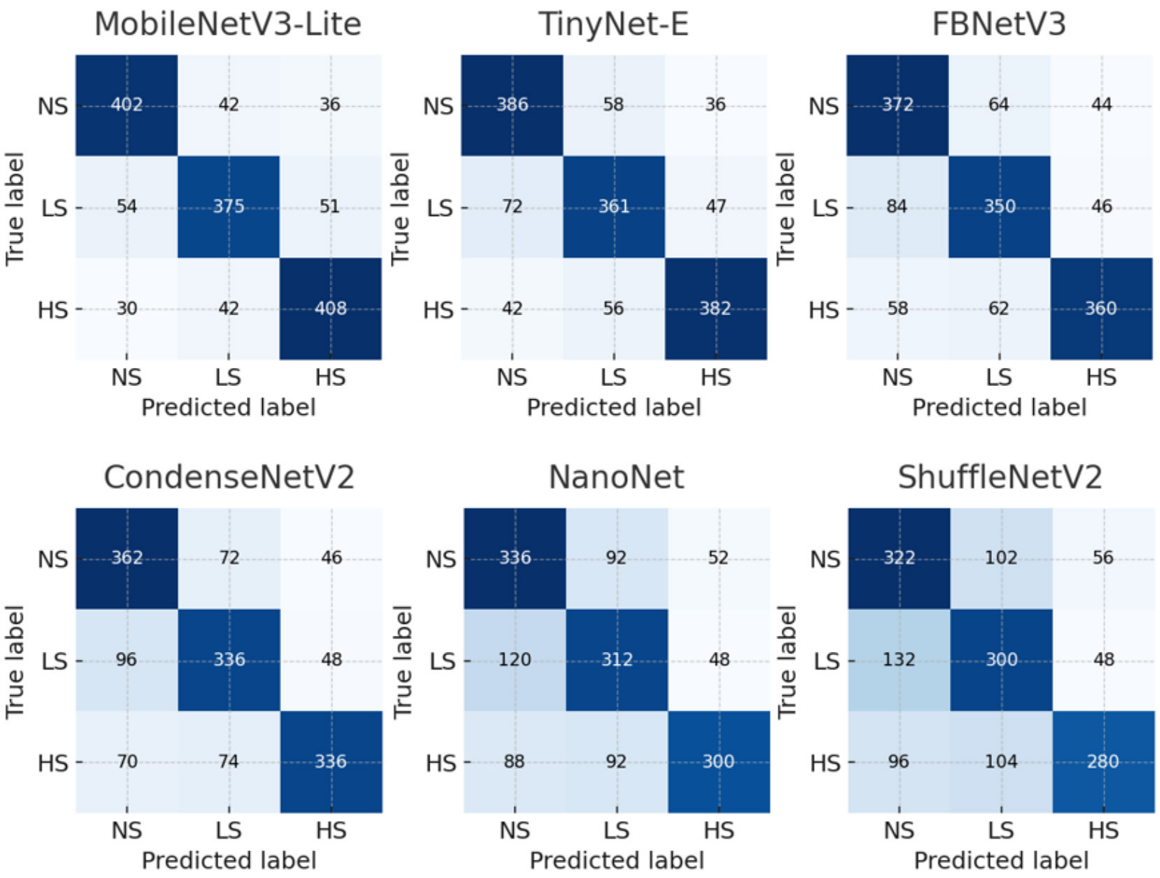

Figure 2. Summary of confusion matrices for models in multiclass classification on the single thermal palette

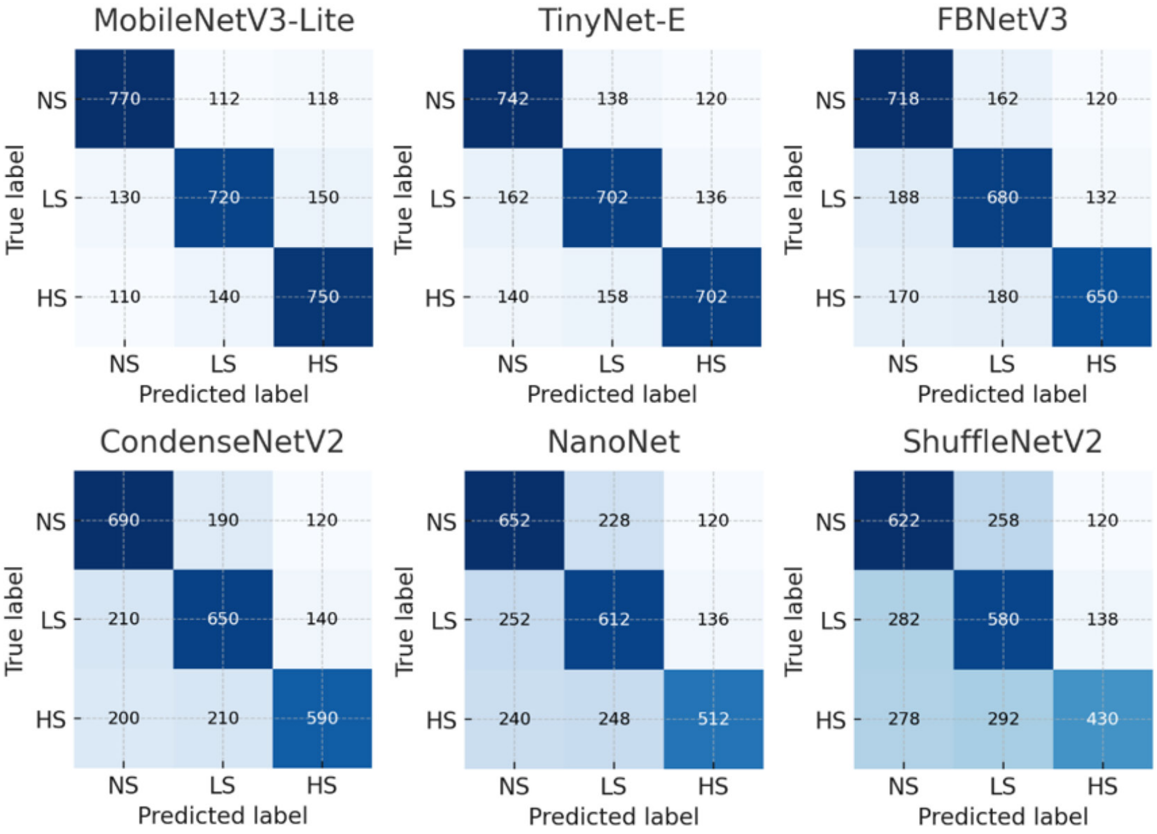

Figure 3. Summary of confusion matrices for models in multiclass classification on the mixed thermal palette

Table 7. Summary of classification report results for ensemble methods on the single thermal palette

Ensemble method	State	Precision	Recall	F1-score	Accuracy
Majority voting	NS	0.83	0.82	0.83	82.1%
	LS	0.77	0.76	0.77	
	HS	0.80	0.79	0.80	
Weighted average	NS	0.84	0.83	0.84	83.5%
	LS	0.79	0.78	0.79	
	HS	0.82	0.81	0.82	
Gompertz Weighting	NS	0.84	0.83	0.84	83.1%
	LS	0.80	0.79	0.80	
	HS	0.83	0.82	0.83	
Gompertz + Fuzzy Logic	NS	0.86	0.85	0.86	84.7%
	LS	0.82	0.81	0.82	
	HS	0.84	0.83	0.84	

Table 8. Summary of classification report results for ensemble methods on the mixed thermal palette

Ensemble method	State	Precision	Recall	F1-score	Accuracy
Majority voting	NS	0.81	0.79	0.80	80.2%
	LS	0.76	0.75	0.76	
	HS	0.79	0.80	0.79	
Weighted average	NS	0.82	0.80	0.81	81.1%
	LS	0.77	0.76	0.77	
	HS	0.80	0.82	0.81	
Gompertz Weighting	NS	0.82	0.80	0.81	81.3%
	LS	0.78	0.77	0.78	
	HS	0.80	0.81	0.81	
Gompertz + Fuzzy Logic	NS	0.83	0.81	0.82	82.3%
	LS	0.79	0.79	0.79	
	HS	0.82	0.83	0.82	

in a total of 60 observations (six models × 2 datasets × 5 folds). The normality assumption was verified using the Shapiro-Wilk test (W=0.982, p=0.412). The ANOVA results are summarized in Table 10. Results from Table 10 show a significant main effect of the model – accuracy differs significantly between models. The main effect of the dataset is also significant – classification on the single palette dataset yields better results than on the mixed palette. Regarding interaction, it was observed that the impact of the dataset type (single palette vs mixed palette) on accuracy depended on the model architecture (which was also confirmed earlier by the paired t-test). The ANOVA results confirm that both the choice of architecture and the consistency of thermal representation are crucial for the effectiveness of stress classification.

As a post-hoc analysis following the significant main effect found in the ANOVA, the Tukey HSD (honestly significant difference) test was applied. The goal was to identify which pairs of CNN architectures showed statistically significant differences in classification accuracy across both experimental conditions. The test simultaneously compared all pairs of models. The results are summarized in Table 11, highlighting the top 3 pairs with the largest accuracy differences and the bottom 2 pairs with the smallest differences as selection criteria. The Tukey HSD post-hoc analysis for both study variants revealed statistically significant differences in accuracy between the models ($p < 0.05$). The largest differences were observed between MobileNetV3-Lite a ShuffleNetV2 (single palette: Δ=0.119, d=2.79, $p<0.001$; mixed palette: Δ=0.137, d=3.01), indicating

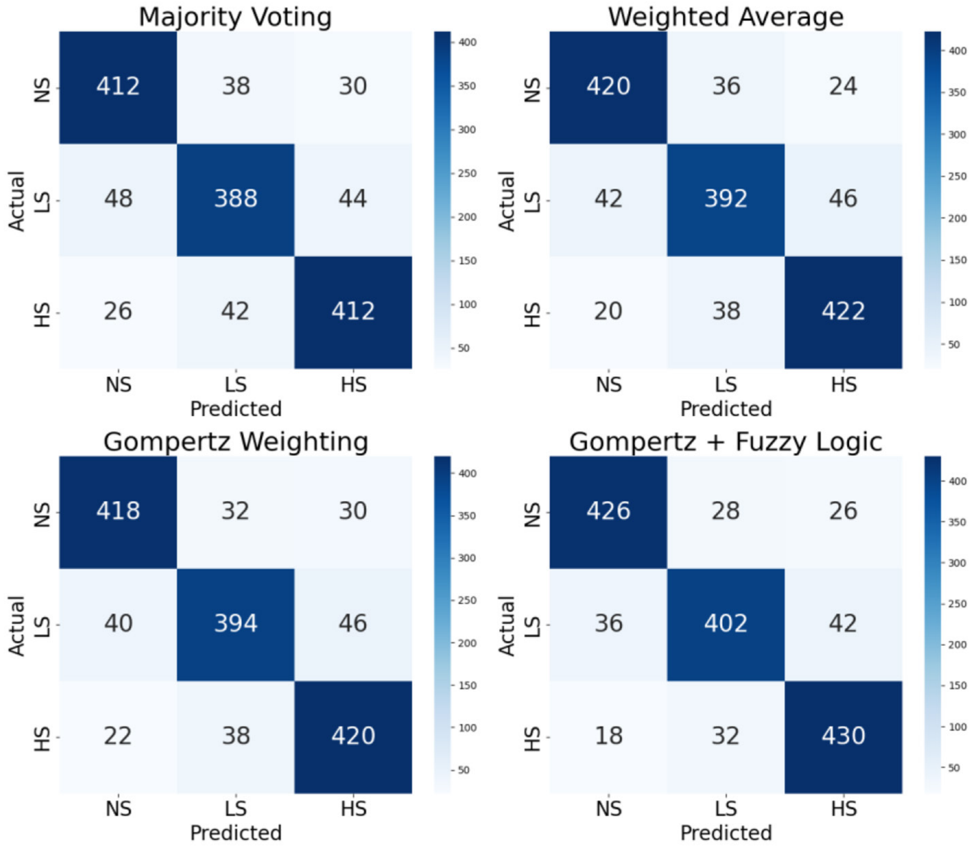

Figure 4. Confusion matrix summary for ensemble methods on the single thermal palette

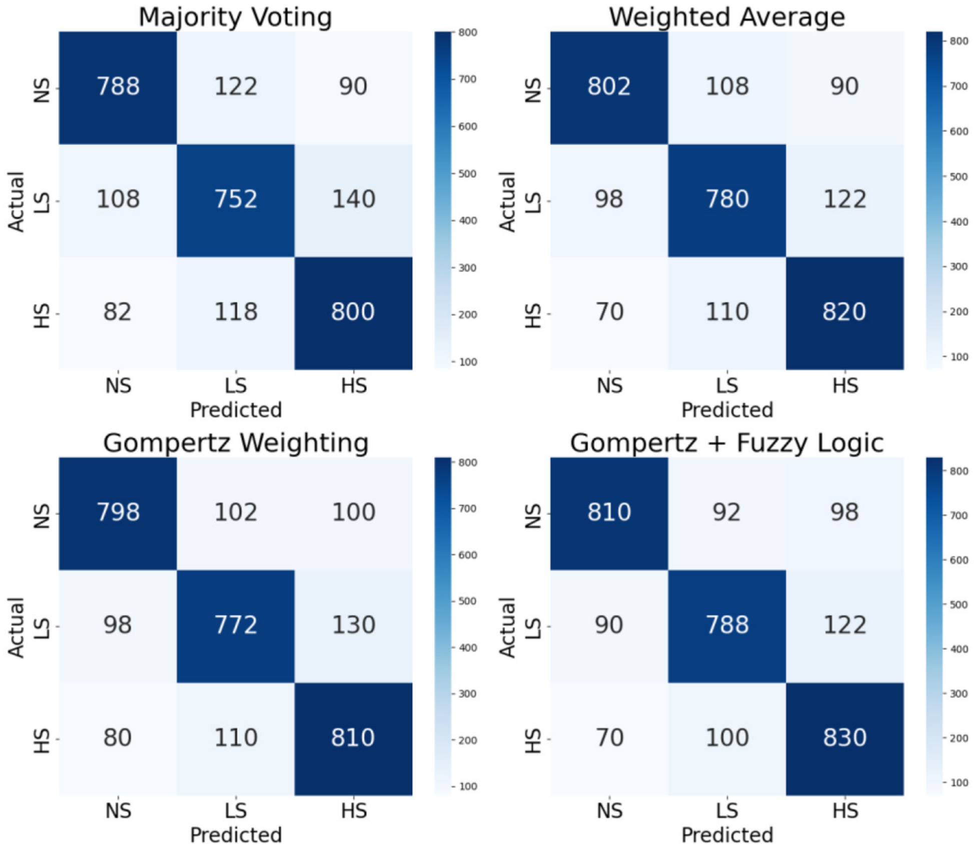

Figure 5. Confusion matrix summary for ensemble methods on the mixed thermal palettes

Table 9. Summary of paired student's t-test results (single vs. mixed palette)

Model CNN	Mean single	Mean mixed	Difference	t-value	df	p-value	95% CI
MobileNetV3-Lite	0.801	0.782	+0.019	4.12	4	0.015	[0.005, 0.033]
TinyNet-E	0.785	0.762	+0.023	5.01	4	0.007	[0.010, 0.036]
FBNetV3	0.771	0.748	+0.023	4.87	4	0.008	[0.009, 0.037]
CondenseNetV2	0.751	0.721	+0.030	6.15	4	0.003	[0.016, 0.044]
NanoNet	0.703	0.679	+0.024	5.43	4	0.006	[0.012, 0.036]
ShuffleNetV2	0.681	0.640	+0.041	8.72	4	<0.001	[0.029, 0.053]

Table 10. Results of the two-way ANOVA with replications

Source of variance	SS	df	MS	F	p-value	η^2 (effect size)
Model (M)	0.148	5	0.0296	38.72	<0.001	0.801
Dataset (D)	0.021	1	0.0208	27.19	<0.001	0.362
M × D	0.009	5	0.0018	2.36	0.049	0.197
Error	0.092	48	0.00076	-	-	-

a strong effect of architectural optimization – MobileNetV3, despite higher complexity (2.4 M vs 1.9 M parameters), employs attention mechanisms (SE blocks), resulting in higher accuracy. For the mixed palette, the differences were even more pronounced (Δ increased by 15.1%), confirming MobileNetV3's better adaptation to diverse thermal representations. The smallest differences were observed between TinyNet-E and MobileNetV3 (single palette: Δ=0.017, p=0.301; mixed palette: Δ=0.023, p=0.097), as well as between FBNetV3 a CondenseNetV2 (single palette: Δ=0.017, p=0.134; mixed palette: Δ=0.026, p=0.021). These results confirm the superiority of architecture with advanced optimization mechanisms (MobileNetV3) over simpler solutions (ShuffleNetV2) and are consistens with the previously conducted ANOVA analysis.

McNemar test

The McNemar test was applied to compare the number of misclassifications between two variants of the ensemble method: the Gompertz function alone (G) versus the Gompertz function combined with fuzzy logic (G+F). The analysis was conducted independently for each study variant (single palette and mixed palette) and for each stress class, using data from 5-fold cross-validation (a total of 600 test samples for the single palette and 1500 for the mixed palette). The hypothesis tested was that the addition of fuzzy logic significantly reduces the number of errors. Results of the tests are summarized in Tables 12–13. For the single palette, fuzzy logic significantly reduced the number of errors in all classes (p<0.05), with the strongest effect observed in the "high-stress" class (a 30% reduction). The "no-stress" and "medium-stress" classes achieved reductions of 25% and 19.6%, respectively. In the mixed palette variant, significant improvement was observed only for the "high-stress" class, where errors decreased by 19%. This is attributed to the greater complexity of data in this variant, where fuzzy logic more effectively filtered uncertain predictions in borderline cases. The smallest effect was noted for the "no-stress" class (a 5.9% reduction), suggesting that simple cases are less susceptible to optimization by the fuzzy system. In summary, the test results indicate that fuzzy logic substantially improves stress classification accuracy, especially under conditions of consistent thermal representation (single palette), for the "high-stress" class, and in scenarios characterized by clear model uncertainty (e.g., subtle temperature differences). This effect is statistically significant (p<0.01 for key classes) and meaningful – error reductions of 19–30% translate into greater system reliability in classification applications. These results directly confirm the claimed advantages of adaptability and robustness of the proposed hybrid ensemble. The fuzzy logic component dynamically adapts model weights based on confidence, leading to a statistically significant reduction in errors, especially for challenging 'high-stress' cases.

Table 11. Post-hoc Tukey HSD for models (top 3 and bottom 2)

Tukey HSD test for variant: single palette				
Comparison	Difference	p-value	95% CI	ES (Cohen's d)
MobileNetV3 vs ShuffleNetV2	+0.119	<0.001	[0.096, 0.142]	2.79
TinyNet-E vs ShuffleNetV2	+0.102	<0.001	[0.079, 0.125]	2.38
MobileNetV3 vs NanoNet	+0.099	<0.001	[0.076, 0.122]	2.31
FBNetV3 vs CondenseNetV2	+0.017	0.134	[-0.006, 0.040]	0.40
TinyNet-E vs MobileNetV3	-0.017	0.301	[-0.040, 0.006]	-0.40
Tukey HSD test for variant: mixed palettes				
MobileNetV3 vs ShuffleNetV2	+0.137	<0.001	[0.114, 0.160]	3.01
TinyNet-E vs ShuffleNetV2	+0.113	<0.001	[0.090, 0.136]	2.48
MobileNetV3 vs NanoNet	+0.106	<0.001	[0.083, 0.129]	2.33
FBNetV3 vs CondenseNetV2	+0.026	0.021	[0.003, 0.049]	0.57
TinyNet-E vs MobileNetV3	-0.023	0.097	[-0.046, 0.000]	-0.51

Table 12. Error reduction for the single palette

Class	Errors (G)	Errors (G+F)	Reduction	χ^2	p-value
No-stress	72	54	25.0%	4.17	0.041
Low-stress	97	78	19.6%	5.12	0.024
High-stress	80	56	30.0%	7.11	0.008

Table 13. Error reduction for the mixed palettes

Class	Errors (G)	Errors (G+F)	Reduction	χ^2	p-value
No-stress	202	190	5.9%	1.13	0.288
Low-stress	228	212	7.0%	2.00	0.157
High-stress	210	170	19.0%	9.41	0.002

CONCLUSIONS

This study introduced a novel hybrid approach to ensemble-based stress classification using thermal imaging. The primary contribution lies in the development and thorough evaluation of a three-stage ensemble aggregation method, which combines nonlinear base model weighting via a calibrated Gompertz function with dynamic weight adjustment based on a fuzzy logic system that assesses the confidence of each individual prediction. The conducted experiments, supported by rigorous statistical analysis, lead to the following key findings:

- Data representation consistency is paramount – results clearly indicate that applying a single, standardized thermal palette enables statistically significant improvements in classification accuracy (up to ~2 percentage points for the best models) compared to the mixed-palette scenario.
- Gompertz-based selective weighting is effective – using the nonlinear Gompertz function for initial model weighting efficiently emphasized high-quality models (e.g., MobileNetV3-Lite) while marginalizing weaker ones, resulting in higher ensemble accuracy compared to linear averaging.
- Fuzzy logic substantially reduces errors on "hard" samples – the most empirically confirmed advantage of the proposed method is its adaptiveness. The fuzzy confidence evaluation mechanism significantly reduced classification errors (McNemar's test, $p < 0.001$), particularly for the "high stress" class (reduction of 19–30%) in scenarios where base models exhibited low confidence. This demonstrates that the proposed approach does not merely combine models but intelligently manages system uncertainty.
- Model architecture matters – the benchmark of six lightweight CNN architectures revealed

notable performance differences. Models incorporating advanced attention mechanisms (SE-blocks in MobileNetV3-Lite) or optimized via NAS (FBNetV3, TinyNet-E) consistently achieved superior results, especially in the more challenging mixed-palette variant.

This study has several limitations. The primary constraint concerns the thermal dataset, namely the relatively small size of the data used. As discussed before, the availability of thermal databases remains limited compared to visual (non-thermal) datasets, prompting the need for custom dataset development. Furthermore, while the use of subject-independent cross-validation and aggressive augmentation mitigates the issue, the relatively small dataset size remains a limitation that future work with larger, more diverse cohorts should address. Another issue lies in the simplified mapping of emotions to stress levels. Mapping complex emotional states to a three-level stress scale based on literature remains a psychological simplification. It is important to emphasize that this mapping is a heuristic, data-driven approximation commonly used in the literature to enable computational analysis, and may not capture the full complexity of the physiological stress response in all contexts. It serves as a pragmatic framework rather than an absolute physiological truth. Furthermore, the controlled laboratory data collection setting poses another limitation. Evaluating the robustness of the proposed method under real-world, uncontrolled conditions (e.g., variable lighting, movement) remains a future task. Finally, the computational cost of the ensemble must be noted – despite the use of lightweight models, aggregating six networks and performing Gompertz function and fuzzy logic computations introduces additional overhead compared to using a single model.

Based on these findings and limitations, future research directions have been identified. A priority is the acquisition of a larger and more diverse thermal dataset, accounting for demographic variability (age, gender, ethnicity) and collected under less controlled conditions. Simultaneously, it is worth exploring advanced domain-specific augmentation techniques. For instance, generative adversarial networks (GANs) could be employed to synthesize realistic samples or simulate different thermal palettes, mitigating data scarcity. Another promising avenue is optimization of fuzzy function parameters, specifically through global optimization methods (e.g., genetic algorithms, particle swarm optimization) for automated tuning of Gompertz and fuzzy membership function parameters. To evaluate the real-time performance and energy efficiency of the ensemble, implementation on edge devices (Edge deployment) could be pursued by porting and benchmarking the final trained ensemble model on embedded platforms such as Jetson Nano or Coral USB accelerator. A further perspective involves multimodal integration, i.e., combining thermal imaging with other modalities such as conventional video (for micro-expression tracking) or physiological signals (ECG, GSR) from wearable devices to create a more holistic and resilient stress assessment system. This, in turn, relates to the previously discussed dataset challenge and the need for custom data collection.

The proposed Gompertz-Fuzzy ensemble method constitutes a significant step toward robust, intelligent stress recognition systems capable of dynamic adaptation and uncertainty management. Despite certain limitations, the study delineates clear and promising directions for future research in this rapidly evolving domain.

Acknowledgments

The research was conducted as part of the project “Human Stress Recognition Using Thermographic Imaging and Neural Networks” which received positive approval from the Ethics Committee for Scientific Research (approval No. 4/2024).

REFERENCES

1. Schneiderman N, Ironson G, Siegel SD. Stress and health: psychological, behavioral, and biological determinants. Annu Rev Clin Psychol. 2005;1:607–28. https://doi.org/10.1146/annurev.clinpsy.1.102803.144141
2. Shastri D, Merla A, Tsiamyrtzis P, Pavlidis I. Imaging facial signs of neurophysiological responses. IEEE Trans Biomed Eng. 2009;56(2):477–84. https://doi.org/10.1109/TBME.2008.2003265
3. Pavlidis I, Levine J, Baukol P. Thermal imaging for anxiety detection. In: Proceedings of the 2001 IEEE Workshop on Computer Vision Beyond the Visible Spectrum: Methods and Applications (CVBVS 2001); 2001 Dec 14; Kauai, HI, USA. Piscataway (NJ): IEEE; 2001;104–9.
4. Sorić M, Russo M. Challenges in deep learning-based

affective computing: a survey on data, models, and metrics. Inf Fusion. 2022;87:1–17.

5. Berseth T, Hawkinson W, Khanna P. On the sensitivity of convolutional neural networks to image metadata and quality in thermal imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops; 2020;430–1.
6. Ganaie MA, Hu M, Malik AK, Tanveer M, Suganthan PN. Ensemble deep learning: a review. Eng Appl Artif Intell. 2022;115:105151.
7. Wen Y, Tran D, Ba J. Batchensemble: an alternative approach to efficient ensemble and lifelong learning. arXiv [Preprint]. 2020 [cited 2025 Aug 25]. Available from: https://arxiv.org/abs/2002.06715
8. Ioannou S, Gallese V, Merla A. Thermal infrared imaging in psychophysiology: potentialities and limits. Psychophysiology. 2014 Oct;51(10):951–63.
9. Goulart C, Valadão C, Delisle-Rodriguez D, Caldeira E, Bastos T. Visual and thermal image processing for facial specific landmark detection to infer emotions in a child-robot interaction. Sensors (Basel). 2019 Jun 26;19(13):2844.
10. Bekhouche SE, Ouafi A, Benlamoudi A, Djekoune A, Taleb-Ahmed A. Thermal facial expression recognition under different conditions. IET Image Process. 2021 Feb;15(2):449–61.
11. Zhao Z, Li Q, Zhang Z, Zhang J, Qin Y, Chen X, et al. A deep neural network-based emotion classification using automatic facial thermal expression recognition. IEEE Access. 2020;8:132892–907.
12. Jabłoński M, Kwasniewicz L, Łukasik S, Wąż P, Dziak J, Dąbrowski JR. Deep learning methods for thermal image super-resolution and object classification. In: 2023 5th International Conference on Control and Robotics (ICCR); 2023;258–62.
13. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, et al. Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv [Preprint]. 2017 [cited 2025 Aug 25]. Available from: https://arxiv.org/abs/1704.04861
14. Ma N, Zhang X, Zheng HT, Sun J. Shufflenet v2: practical guidelines for efficient CNN architecture design. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018;116–31.
15. Wu B, Dai X, Zhang P, Wang Y, Sun F, Wu Y, et al. Fbnet: hardware-aware efficient convnet design via differentiable neural architecture search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019;10734–42.
16. Han K, Wang Y, Zhang Q, Zhang W, Xu C, Zhang T. Model Rubik's cube: twisting resolution, depth and width for tinynets. Adv Neural Inf Process Syst. 2020;33:19353–64.
17. Chen L, Wang W, Wang K, Lin D, Li S, Gao W, et al. CondenseNet V2: sparse feature reactivation for deep networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021;3569–78.
18. Wan A, Dai X, Zhang P, He Z, Tian Y, Yu S, et al. Fbnetv3: joint architecture-recipe search using predictor pretraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021;16276–85.
19. Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. Adv Neural Inf Process Syst. 2017;30.
20. Wen Y, Tran D, Ba J. Batchensemble: an alternative approach to efficient ensemble and lifelong learning. arXiv [Preprint]. 2020 [cited 2025 Aug 25]. Available from: https://arxiv.org/abs/2002.06715
21. Ashukha A, Lyzhov A, Molchanov D, Vetrov D. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. arXiv [Preprint]. 2020 [cited 2025 Aug 25]. Available from: https://arxiv.org/abs/2002.06470
22. Ganaie MA, Hu M, Malik AK, Tanveer M, Suganthan PN. Ensemble deep learning: a review. Eng Appl Artif Intell. 2022 Jan;115:105151.
23. Mendel JM. Uncertain rule-based fuzzy systems: introduction and new directions. 2nd ed. Cham: Springer International Publishing; 2017.
24. Zhou SM, Gan JQ. Low-level interpretability and high-level interpretability: a unified view of data-driven interpretable fuzzy system modelling. Fuzzy Sets Syst. 2008 May 1;159(23):3091–131.
25. Sengupta A, Ye Y, Wang R, Liu C, Roy K. Going deeper in spiking neural networks: VGG and residual architectures. Front Neurosci. 2019;13:95.
26. Kwasniewicz L, Orzechowski P, Łukasik S, Cyran KA. Fuzzy entropy-based feature extraction for infrared thermal face recognition. In: 2023 5th International Conference on Control and Robotics (ICCR); 2023;252–7.
27. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30.
28. Tan M, Le Q. Efficientnet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning; 2019;6105–14.
29. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. arXiv [Preprint]. 2020 [cited 2025 Aug 25]. Available from: https://arxiv.org/abs/2010.11929
30. Abuhussein A, Fawzi M, Darwish A, Hassanien AE. Comprehensive facial thermal dataset. Mendeley Data. 2024;V1. https://doi.org/10.17632/8885sc9p4z.1
31. Lerner JS, Gonzalez RM, Dahl RE, Hariri AR,

Taylor SE. Facial expressions of emotion reveal neuroendocrine and cardiovascular stress response. Biol Psychiatry. 2005;58(9):743–50.

32. Ozawa S. Emotions induced by recalling memories about interpersonal stress. Front Psychol. 2021;12:618676. https://doi.org/10.3389/fpsyg.2021.618676

33. Jaiswal M, Aldeneh Z, Mower Provost E. Controlling for confounders in multimodal emotion classification via adversarial learning. In: Proceedings of the 2019 International Conference on Multimodal Interaction (ICMI '19); 2019; New York, NY, USA. New York (NY): ACM; 2019;174–84. https://doi.org/10.1145/3340555.3353731

34. Lin Y, Wang J, Liu W, Jia Y. More positive emotion, less stress perception? Psychol Res Behav Manag. 2022;15:372103732.

35. Stappen L, Baird A, Christ L, Schumann L, Sertolli B, Mebner EM, et al. The MuSe 2021 multimodal sentiment analysis challenge: sentiment, emotion, physiological-emotion, and stress. In: Proceedings of the 2nd Multimodal Sentiment Analysis Challenge (MuSe '21); 2021; New York, NY, USA. New York (NY): ACM; 2021;5–14. https://doi.org/10.1145/3475957

36. Liu Y, Xue J, Li D, Zhang W, Chiew TK, Xu Z. Image recognition based on lightweight convolutional neural network: recent advances. Image Vis Comput. 2024;146:105037. https://doi.org/10.1016/j.imavis.2024.105037

37. Howard A, Sandler M, Chu G, Chen LC, Chen B, Tan M, et al. Searching for MobileNetV3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2019; 1314–1324.

38. Han K, Wang Y, Zhang Q, Zhang W, Xu C, Zhang T. Model Rubik's cube: twisting resolution, depth and width for TinyNets. Comput Vis Pattern Recognit. 2020. https://doi.org/10.48550/arXiv.2010.14819

39. Dai X, Wan A, Zhang P, Wu B, He Z, Wei Z, et al. FBNetV3: joint architecture-recipe search using predictor pretraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021;16276–85.

40. Yang L, Jiang H, Cai R, Wang Y, Song S, Huang G, et al. CondenseNet v2: sparse features reactivation for deep networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021;3569–78.

41. Jha D, Tomar NK, Ali S, Riegler MA, Johansen HD, Johansen D, et al. NanoNet: real-time polyp segmentation in video capsule endoscopy and colonscopy. In: 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS); 2021;37–43.

42. Ma N, Zhang X, Zheng HT, Sun J. Shufflenet v2: practical guidelines for efficient CNN architecture design. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018;116–31.

43. Ye L. AugShuffleNet: communicate more, compute less. arXiv [Preprint]. 2022 [cited 2025 Aug 25]. Available from: https://arxiv.org/abs/2203.06589