

Adversarial robustness and confidence miscalibration in emotion-based HRI systems

Ezekiel Olayide Tolulope

University of Brighton

Franklin Tchakounté

University of Ngaoundere

Derrick Méthode BAGAZA

dmbagaza@univ-bangui.org

University of Bangui

Shanay Rab

University of Brighton

Almas Baimagambetov

University of Brighton

Research Article

Keywords: human-robot interaction, facial emotion recognition, adversarial machine learning, calibration, failure detection, confidence thresholding, robot security, affective computing

Posted Date: October 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11074581/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Adversarial robustness and confidence miscalibration in emotion-based HRI systems

Olayide Tolulope Ezekiel^{1,**}, Franklin Tchakounte², Derrick Méthode BAGAZA^{3,*},
Shanay Rab¹, Almas Baimagambetov¹

¹School of Architecture, Technology and Engineering, University of Brighton, UK

²Department of Mathematics and Computer Science, Faculty of Science,
University of Ngaoundere, Cameroon

³Department of Computer Science, Higher Institute of Technology,
University of Bangui, Central African Republic

*Corresponding author: dmbagaza@univ-bangui.org

**Corresponding author: o.ezekiel1@uni.brighton.ac.uk

Abstract

Facial emotion recognition (FER) deployed in robotics is evaluated primarily by clean-set accuracy, yet this fails to capture whether predictions are trustworthy for physical robot control. We empirically compare five FER pipelines on FER2013: DeepFace and four CNN variants, with the strongest model achieving 0.760 accuracy and 0.753 macro-F1 at real-time CPU speed. Environmental corruption produces degradation proportional to severity, while gradient-based attacks cause uniform collapse: FGSM reduces all models below 0.16 accuracy at $\epsilon = 0.01$, and PGD to near zero for $\epsilon \geq 0.02$, regardless of architecture. Critically, attacked models remain confident: mean confidence reaches 0.964 under PGD. We prove this failure is structural, not calibration-remediable. Because the maximum softmax probability strictly increases with the logit-gap vector and temperature scaling acts identically on all inputs, no global temperature can simultaneously admit legitimate and reject adversarial predictions when the adversarial gap dominates the clean gap. The confidence gate’s ROC curve is invariant to temperature, with measured AUROC of 0.144 and Youden index of -0.261 , meaning the gate is inverted: it admits adversarial predictions more often than legitimate ones. These results establish a hard separation between recognition performance and *actuation soundness*, motivating verification at the perception-to-action boundary.

Keywords: human-robot interaction; facial emotion recognition; adversarial machine learning; calibration; failure detection; confidence thresholding; robot security; affective computing.

1 Introduction

Robots are moving from structured industrial settings into environments in which they must interpret human behaviour and adapt their actions accordingly. In social, assistive, educational, and collaborative applications, facial emotion recognition (FER) is often treated as a direct perceptual route from a human expression to an emotion-conditioned robot response [1, 2, 3]. This architectural pattern creates a security-relevant dependency: if the emotion estimate is wrong, the action selected from that estimate may also be wrong.

Most FER research nevertheless evaluates models principally by recognition metrics such as accuracy, precision, recall, F1-score, and confusion matrices. Those quantities are indispensable for measuring classifier competence, but they do not answer a control question: *when is a prediction sufficiently trustworthy to influence physical or socially consequential behaviour?* The distinction matters because a classifier can be accurate on clean data yet brittle under

environmental degradation or deliberately constructed perturbations. Adversarial examples are particularly problematic because they can cause systematic misclassification while remaining visually subtle [4, 5, 6].

For emotion-driven HRI, the natural mitigation is confidence thresholding: permit a prediction to affect robot behaviour only if its maximum softmax probability exceeds a chosen threshold. This rule presumes that unreliable predictions announce themselves through low confidence. Modern neural networks are, however, frequently miscalibrated [7], and the maximum softmax probability is a weak baseline for misclassification detection even without an adversary [8]. If adversarial attacks also produce *high-confidence* errors, a confidence-only decision rule fails precisely when a security mechanism is most needed.

1.1 Contributions

This paper provides a controlled empirical study of that problem together with an analytical account of why the failure is not remediable by recalibration. Its contributions are:

1. **A formalisation of the perception-to-action gate.** We define the emotion-to-action pipeline, an explicit threat model, and a notion of *actuation soundness* that separates the property a classification metric estimates from the property safe actuation requires (Section 3).
2. **A common-harness comparison of five FER pipelines** on a five-class FER2013 benchmark, reporting accuracy, macro-F1, per-class F1, throughput, and environmental robustness with interval estimates rather than point estimates alone, and with the ordering of the two strongest models deliberately left unresolved for want of repeated training runs (Section 5.1).
3. **A cross-architecture adversarial evaluation** under FGSM and PGD over a common perturbation-budget sweep, establishing that collapse is observed consistently across all four evaluated architectures (Section 5.3).
4. **An impossibility result for global calibration.** We prove that when the adversarial logit-gap vector dominates the clean one, no temperature $T > 0$ and threshold τ jointly admit legitimate predictions while rejecting adversarial ones, and that the confidence detector’s ROC curve is invariant to temperature under a stated assumption (Section 6). This replaces a numerical illustration with a general statement and shows that, under the stated dominance condition, the observed failure cannot be repaired by rescaling.
5. **Detector bounds derived from the measured admission rates**, which constrain the Youden index of the confidence gate above by 0.040 at the operating threshold irrespective of the clean admission rate, together with the measured detector characteristics showing that the realised index is -0.261 after scaling and -0.266 before, and an expression for the posterior error rate among admitted estimates as a function of attack prevalence (Section 7).

The principal finding is that adversarial fragility is not confined to one architecture, that the attacked predictions remain strongly confident, and that this confidence cannot be recalibrated away. The design consequence is a system-level verification layer between FER and robot action rather than direct emotion-to-action mapping.

2 Related Work and Security Motivation

This section establishes the gap the paper addresses. It proceeds from the application domain inwards. Section 2.1 reviews facial emotion recognition as it is used in affective human-robot interaction, where evaluation is organised almost entirely around recognition accuracy. Section 2.2 summarises the adversarial attack and defence literature, which establishes that the fragility

exploited here is generic to deep classifiers rather than specific to the emotion task. Section 2.3 covers confidence calibration, and in particular temperature scaling, because the property that makes it attractive as a remedy for overconfidence is precisely the property that will be shown to disqualify it as a security mechanism. Section 2.4 turns to robot security, where attack surfaces have been catalogued but the perception-to-action pathway has received limited direct evaluation.

The pattern that emerges across the four is a disciplinary separation. The FER literature measures whether a prediction is correct; the adversarial literature measures whether a model can be broken; the calibration literature measures whether a confidence value is well-scaled; and the robot security literature measures where a system can be attacked. None of them asks the question a deployed emotion-guided robot must answer, which is whether a particular estimate should be permitted to command a physical action. Table 1 makes that separation concrete.

2.1 FER for affective HRI

FER has become a core component of affective HRI, with deep-learning methods used to improve recognition accuracy and real-time deployability [9, 10]. Surveys of emotion-aware HRI show increasing use of facial, speech, physiological, and multimodal signals, but also highlight unresolved questions around robustness, context, trust, and evaluation [1, 11, 3]. We note at the outset that the validity of inferring discrete emotional states from facial configuration is itself contested [12]; the present work takes no position on that question, because its argument concerns the reliability of the estimate *pathway* and applies whatever the semantic status of the label (see Section 8.3).

2.2 Adversarial attack and defence

Neural classifiers can be manipulated by bounded input perturbations. FGSM constructs a one-step perturbation in the sign direction of the input gradient [4]; projected gradient descent iteratively optimises within a constrained norm ball and is treated as a standard strong first-order white-box benchmark [5]. Adversarial attacks have been studied specifically in facial-expression recognition [13].

Defences fall into three families, and it is important to state which this paper does and does not address. *Robust training* methods such as adversarial training [5] and TRADES [14], which decomposes robust error into a natural term and a boundary term so that the robustness-accuracy trade-off can be set explicitly, modify the classifier so that its decision is stable within a perturbation ball. *Certified* methods such as randomised smoothing [15] provide provable radii under stated noise models. *Detection* methods attempt to identify adversarial or out-of-distribution inputs at test time, including ODIN [16], Mahalanobis distance scoring [17], and energy-based scores [18]; these are described in Section 2.3, where their relationship to the confidence gate examined here is set out. The confidence gate examined here is the simplest member of the third family, and corresponds to the maximum-softmax-probability baseline of Hendrycks and Gimpel [8].

This paper’s claim is deliberately narrow and should not be read as a claim about the first two families: we show that the *detection* rule most commonly proposed in the HRI literature fails, and that its failure cannot be repaired by post-hoc calibration. Whether robust training or certified defences would restore actuation soundness in this setting is an open question that we identify as the primary empirical extension (Section 9).

2.3 Calibration and confidence-based failure detection

A classifier is calibrated when its reported confidence matches its empirical accuracy: among inputs assigned confidence 0.8, approximately 80% should be correct. Modern networks are

systematically overconfident, a tendency that grows with depth and capacity [7]. Since a confidence value that overstates reliability is precisely what a downstream consumer must not act on, calibration is the natural place to look for a remedy.

Temperature scaling is the standard post-hoc correction. A single scalar $T > 1$ divides the logits before the softmax, flattening the output distribution and reducing reported confidence without retraining. Its appeal is that it is *accuracy-preserving*: dividing every logit by a common positive constant is a monotone reparametrisation that leaves the arg max, and therefore every predicted label, unchanged [7]. Section 6 shows that this same property is what makes it structurally incapable of separating clean from adversarial inputs, since a transformation that cannot change which class is predicted cannot, by itself, distinguish inputs on the basis of whether that prediction is trustworthy.

Confidence is also the basis of the most widely used failure-detection mechanism. The maximum softmax probability (MSP) baseline thresholds the largest output probability, on the observation that correctly classified inputs tend to receive higher maxima than misclassified or out-of-distribution ones [8]. Its simplicity has made it the default in deployed systems, and it is the mechanism this paper evaluates. Subsequent work improves on it by moving away from raw softmax output: ODIN, an out-of-distribution detector for neural networks, combines temperature scaling with input preprocessing [16]; Mahalanobis distance scoring operates in feature space rather than on the output layer [17]; and energy-based scores replace the softmax maximum with a quantity better aligned to the input density [18]. That this progression exists is itself evidence that MSP is known to be a weak detector.

Two features of that literature bear on the present work. First, it is developed almost entirely against natural distribution shift and out-of-distribution inputs rather than against an adversary optimising to defeat the detector. Second, its success criterion is detection performance measured over a dataset, not a guarantee about the subpopulation that survives the threshold, which is the quantity a system authorising physical action requires. The analysis in Section 6 addresses the first gap for temperature-scaled MSP specifically; Definition 2 formalises the second.

2.4 Robot security

Robotics security research has documented broad attack surfaces spanning perception, communication, planning, and control [19]. Safety mechanisms in HRI have introduced confidence-aware reasoning and modular guardrails [20, 21]. The missing connection examined here is narrower but operationally important: whether the scalar confidence emitted by an FER network can serve as an adequate gate for emotion-driven robotic action under adversarial conditions.

Table 1 positions the present study against representative deep-learning FER systems along the five stages an emotion-guided robot must traverse before it acts. The columns are chosen to be answerable from each paper’s own reported evaluation rather than from inference about what its authors might have done, and the final column is one on which the present work also scores negatively, so that the comparison is not constructed to favour a single row.

Read across the rows, the literature is organised primarily around recognition performance, and the question of what a system should do when a confident prediction is wrong is largely unaddressed. Explicit adversarial evaluation remains uncommon, and the standard maximum-softmax confidence gate does not separate clean from adversarial inputs in the present study.

2.5 Differential Contribution Beyond Recent FER and Confidence-Based Gating Work

The field-wide positioning reflected in Table 1 reveals a disciplinary gap: no established work in facial emotion recognition for robotics measures the reliability of emotion-to-action mapping under adversarial perturbation, nor evaluates whether high-confidence predictions remain trustworthy when the input has been adversarially modified. This section identifies two recent contributions

Table 1

Work	Modality	Adv. eval.	Failure detection	Action gating	Robot valid.	Phys. attack
Shahzad et al. [22]	Facial	✗	✗	✗	✗	✗
Sahana et al. [23]	Facial	✗	✗	✗	✗	✗
Sajjad et al. [24]	Facial	~ ^b	✗	✗	✗	✗
Ozdemir et al. [25]	Facial	✗	✗	✗	✗	✗
Akhand et al. [26]	Facial	✗	✗	✗	✗	✗
Hussain & Al Balushi [27]	Facial	✗	✗	✗	✗	✗
Mehendale [28]	Facial	✗	✗	✗	✗	✗
Ali & Ghaffar [29]	Facial	✗	✗	✗	✗	✗
Sun et al. [13]	Facial	✓	✗	✗	✗	✗
This work	Facial	✓	✓ MSP ^c (insufficient)	✗ ^a	✗	✗

^a This study characterises the perception stage and motivates action-level gating; it does not implement or evaluate a gated pipeline. The final three columns are open for every row including this one, which is the gap the study identifies rather than closes.

^b Survey; adversarial vulnerability is discussed as an open problem in the surveyed literature rather than evaluated on a model of the authors' own.

^c MSP denotes the maximum softmax probability baseline of Hendrycks and Gimpel [8]: the largest output probability is compared against a threshold and predictions falling below it are rejected. It is the mechanism most widely deployed for this purpose, and the entry records that this study implements it and then demonstrates that it does not separate clean from adversarial inputs. The tick therefore marks a negative result rather than a capability claim.

Assessments record what each paper reports in its own evaluation. A cross indicates that the paper does not report the item, not that its authors considered and rejected it; the great majority of FER papers are evaluated on recognition metrics alone, which is the pattern the table documents.

in the literature that represent the current state of practice in two closely related areas, and explains how the present work advances beyond their scope.

The first contribution represents the current state of the art in facial emotion recognition methodology and model architecture design. The second represents recent practice in deploying machine learning for real-time failure detection in safety-critical systems. Neither addresses the intersection of these two domains: adversarial robustness of perception systems that gate physical robot action.

Recent advances in facial emotion recognition methodology. Comprehensive surveys of state-of-the-art FER methods document progress in recognition accuracy through architectural innovation, improved training methodologies, handling of limited labelled data, and robustness to environmental factors such as lighting variation, head pose, and partial occlusion [30]. Such work synthesises advances across multiple FER datasets and deep learning architectures, establishing benchmarks in clean-set classification accuracy and demonstrating real-time inference capability on modest hardware.

Contribution of the present work relative to this research direction: The recognition literature treats FER as an accuracy optimisation problem, measured over the entire input distribution. This paper introduces a distinct property called *actuation soundness* (Definition 2), which is defined exclusively over the subpopulation of predictions that the system permits to command robot behaviour. These two quantities can diverge substantially: a model may maintain high average accuracy across all inputs while systematically producing incorrect predictions on precisely the subset the gate admits. This occurs because adversarial attacks concentrate errors in the high-confidence region that the gate is designed to pass. We prove that no choice of temperature parameter in standard confidence scaling can bridge this gap once

the adversarial logit-margin distribution dominates the clean distribution (Theorem 1). The recognition literature optimises for what a system predicts; this work optimises for whether that prediction can be trusted to initiate safety-critical action.

Recent practice in machine learning-based failure detection. Contemporary work on deploying machine learning classifiers in real-time contexts emphasises detecting unreliable predictions at inference time [31]. Such systems achieve high precision and recall for identifying equipment faults, anomalies, or system malfunctions using standard classification metrics measured on static evaluation sets. The focus is demonstrating that a detector can successfully identify problematic states within the accuracy-recall framework.

Contribution of the present work relative to this research direction: The failure-detection literature assumes that detectors can be improved through stronger classifiers, larger datasets, or better hyperparameters. This paper proves a structural impossibility: because the maximum softmax probability is strictly monotonic in logit margins (Proposition 1), and because temperature scaling applies an identical transformation to all samples, no global temperature parameter can separate clean from adversarial samples whenever the adversarial margin exceeds the clean margin (Theorem 1). This is not a limitation of current methods but a fundamental property of the softmax function that applies regardless of model architecture. We further prove that under a testable homogeneity assumption, the detector’s receiver operating characteristic curve is invariant to temperature (Corollary 1), and provide detector-agnostic performance bounds that depend only on the measured adversarial admission rate (Proposition 3). The failure-detection literature seeks stronger algorithms; this work explains why the most widely deployed algorithm in emotion-guided robotics cannot succeed under adversarial conditions, and provides bounds that hold independently of model choice or clean accuracy.

Synthesis. The present work does not claim advancement over existing methods on their stated objectives. It does not propose a stronger recognition algorithm, nor a more sensitive failure detector. Rather, it changes the question posed to both research areas. For a robot whose behaviour is directly gated by emotion recognition, the measure of success is not “can we achieve high classification accuracy?” or “can we detect anomalies?” but rather “can we reliably authorise physical action based on this prediction?” The analysis demonstrates that achieving high recognition accuracy and detecting prediction failures are necessary but not sufficient conditions for safe actuation. A verification layer external to the classifier is required, one that integrates confidence with other evidence streams and does not rely solely on thresholding the maximum softmax probability.

3 Problem Formalisation

This section fixes notation, states the threat model, and defines the property that a perception output must satisfy before it may condition robot behaviour. Table 2 collects the symbols used throughout, distinguishing distributions and vectors from their scalar components.

3.1 The emotion-to-action pipeline

Let $\mathcal{X} \subseteq [0, 1]^{H \times W \times C}$ denote the space of face-cropped images and $\mathcal{Y} = \{1, \dots, K\}$ the emotion label set, with $K = 5$ throughout. A classifier is a map $f_\theta : \mathcal{X} \rightarrow \mathbb{R}^K$ producing logits $z = f_\theta(x)$, with predicted class $\hat{y}(x) = \arg \max_i z_i$ and softmax confidence at temperature $T > 0$

$$p_i(x; T) = \frac{\exp(z_i/T)}{\sum_{j=1}^K \exp(z_j/T)}, \quad p_{\max}(x; T) = \max_i p_i(x; T). \quad (1)$$

Table 2

Symbol	Meaning
<i>Distributions and vectors</i>	
$z \in \mathbb{R}^K$	logit vector produced by the classifier
$p(x; T)$	softmax distribution over the K classes at temperature T
Δ	ranked gap vector, $(\Delta_{(2)}, \dots, \Delta_{(K)})$
<i>Scalar components</i>	
$p_i(x; T)$	the i th component of $p(x; T)$
$p_{\max}(x; T)$	maximum component, $\max_i p_i(x; T)$
$\Delta_{(r)}$	r th ranked gap, $z_{(1)} - z_{(r)}$
$\delta(\Delta; T)$	effective gap, a scalar summary of Δ at temperature T
<i>Decisions and parameters</i>	
$\hat{y}(x)$	predicted class, $\arg \max_i p_i(x; T)$
π	action policy mapping a label to a robot behaviour
$g_{\tau, T}$	confidence gate at threshold τ and temperature T
τ	admission threshold; $\tau = 0.70$ throughout
T	temperature; $T = 0.982$ as fitted
ϵ	adversarial perturbation budget on the $[0, 1]$ input scale
<i>Reported quantities</i>	
$R_{\text{label}}, R_{\text{act}}$	admitted label risk and admitted action risk
a_c, a_a	clean and adversarial admission rates at τ
J	Youden index, $\text{TPR} - \text{FPR}$

An *action policy* $\pi : \mathcal{Y} \rightarrow \mathcal{A}$ maps an emotion label to a robot action primitive drawn from a finite set $\mathcal{A}$ that includes a distinguished null action $a_\emptyset$ (do nothing). A *gate* is a predicate $g : \mathcal{X} \rightarrow \{0, 1\}$; the deployed system executes

$$a(x) = \begin{cases} \pi(\hat{y}(x)) & \text{if } g(x) = 1, \\ a_\emptyset & \text{otherwise.} \end{cases} \quad (2)$$

The confidence gate is the particular choice

$$g_{\tau, T}(x) = \mathbf{1}[p_{\max}(x; T) \geq \tau]. \quad (3)$$

Equation (2) makes explicit what a classification metric does not: an FER error is not a scoring event but an action-selection event.

3.2 Threat model

Definition 1 (Adversary). The adversary $\mathcal{Adv}$ is characterised by the triple $(\mathcal{K}, \mathcal{C}, \mathcal{O})$ where $\mathcal{K}$ is knowledge, $\mathcal{C}$ capability, and $\mathcal{O}$ objective. In this work:

- $\mathcal{K}$: full white-box knowledge of θ , the architecture, and the preprocessing pipeline (Kerckhoffs' assumption); knowledge of τ and T is *not* assumed but does not weaken the results, since no adaptivity to the gate is required.
- $\mathcal{C}$: the ability to apply an additive perturbation δ to the digital input with $\|\delta\|_\infty \leq \epsilon$, subject to $x + \delta \in [0, 1]^d$; the adversary cannot modify θ , the policy π , or the actuator commands.
- $\mathcal{O}$: to cause $\hat{y}(x + \delta) \neq y$ while remaining admitted by the gate. Where the action policy distinguishes the true and predicted labels, this also yields $\pi(\hat{y}) \neq \pi(y)$.

Remark 1 (Scope of the capability assumption). An adversary able to write to the frame buffer is a strong adversary, and a reviewer may reasonably ask why such an adversary would perturb pixels rather than overwrite the emotion label directly. Two responses apply. First,

the perturbation attack is used here as a *diagnostic instrument*: it produces, on demand, the condition of interest, a confidently wrong estimate, so that its consequence for the gate can be measured. Second, the condition is not exclusive to adversaries. Sensor faults, codec artefacts, illumination transitions, and distribution shift all produce confidently wrong estimates without any attacker, and the gate analysis of Sections 6–7 depends only on the resulting confidence distributions, not on their cause. Physically realisable attacks (patches, printed patterns, worn artefacts) are outside the present model and are discussed in Section 9.

3.3 Actuation soundness

The threat model of Section 3.2 establishes what an adversary can do. It remains to state what property a gate must have for the pipeline to be safe, and the central claim of this section is that recognition accuracy is the wrong property.

The reason is a matter of conditioning. Recognition accuracy is an unconditional quantity: it is the probability that the classifier is correct, averaged over all inputs, whether or not the system would act on them. A gate, by contrast, changes which inputs reach the actuator. Once a gate is present, the population that matters is no longer the input distribution but the *admitted* subpopulation, and a system can be simultaneously accurate on average and systematically wrong on the subset it chooses to act upon. That is not a pathological case. It is exactly what a confidence gate produces under adversarial attack, because the attack concentrates its errors in the high-confidence region that the gate is designed to admit.

We therefore require a property defined over the admitted subpopulation rather than over the input distribution. Two distinct quantities are involved and are named separately here. Write

$$R_{\text{label}}(g) = \mathbb{P}[\hat{y}(x) \neq y \mid g(x) = 1], \quad (4)$$

$$R_{\text{act}}(g, \pi) = \mathbb{P}[\pi(\hat{y}(x)) \neq \pi(y) \mid g(x) = 1] \quad (5)$$

for the *admitted label risk* and the *admitted action risk*. Since $\pi(\hat{y}) \neq \pi(y)$ implies $\hat{y} \neq y$, we have $R_{\text{act}} \leq R_{\text{label}}$ for every policy, with equality exactly when π is injective. A policy mapping several emotion labels to a common behaviour, such as a shared safe action, makes the inequality strict.

Actuation soundness is properly a property of R_{act} , and is defined that way below. Because this paper evaluates no particular policy, its experiments bound R_{act} by measuring R_{label} , which is policy-independent and therefore yields a statement that holds for any π . Where a specific policy is available, the two should be reported separately. This is the same conditioning that underlies selective classification, where a model may abstain and is judged on the *selective risk* incurred on the accepted region at a given coverage [32]. The formulation below transfers that perspective from the classifier’s own accept/reject decision to the perception-to-action boundary: what is at stake is not the risk of a wrong label but the risk of a wrong physical action, and the accepting mechanism is a gate external to the model rather than a learned reject option. The contribution is the transfer and its consequences for gate design, not the underlying idea of conditioning on acceptance. Informally: *of the estimates this gate lets through, how often is the robot about to act on a wrong one?* A gate is sound when that quantity is bounded by a tolerance the application can accept.

Definition 2 (Actuation soundness). Fix a tolerance $\alpha \in (0, 1)$ and an input distribution $\mathcal{D}$. A gate g is α -actuation-sound with respect to $\mathcal{D}$ and an action policy π if the admitted action risk satisfies

$$R_{\text{act}}(g, \pi) = \mathbb{P}_{x \sim \mathcal{D}}[\pi(\hat{y}(x)) \neq \pi(y) \mid g(x) = 1] \leq \alpha. \quad (6)$$

The inequality $R_{\text{act}} \leq R_{\text{label}}$ licenses one direction only. A gate whose admitted label risk is at most α is α -actuation-sound for every policy, so a small label risk *certifies* soundness. The

converse fails: a large label risk does not establish that action risk is large, since a policy mapping several labels to a common behaviour can leave the action unchanged when the label is wrong. In the limiting case of a constant policy, $R_{\text{act}} = 0$ whatever R_{label} may be.

The experiments below therefore support a certification claim rather than a failure claim: the measured admitted label risk is too large to certify actuation soundness at any tolerance an emotion-guided robot would adopt. Whether a particular deployment is in fact unsafe depends on its policy, and establishing that requires evaluating R_{act} for that policy directly.

Definition 2 is a statement about the *admitted* subpopulation, which is the population that reaches the actuator. It differs in kind from recognition accuracy, which estimates $\mathbb{P}_{x \sim \mathcal{D}_{\text{clean}}}[\hat{y}(x) = y]$ over *all* inputs and is silent about conditioning on admission. The empirical claim of this paper is that the measured admitted label risk is too large for $g_{\tau, T}$ to *certify* α -actuation soundness at operationally useful tolerances once $\mathcal{D}$ places non-negligible mass on $\mathcal{D}_{\text{adv}}$; whether the gate is actuation-sound for a particular deployment depends on its action policy and requires direct evaluation of R_{act} . The analytical claim is that no choice of T alters this.

Three features of Definition 2 are worth drawing out. It is *relative to a distribution*: a gate sound under $\mathcal{D}_{\text{clean}}$ may be unsound under a mixture containing $\mathcal{D}_{\text{adv}}$, so soundness is not a property a gate carries between deployment contexts. It is *parameterised by tolerance*: α is set by the consequence of a wrong action, and a manipulator operating near a person warrants a tighter α than a screen-based agent. And it is *silent about coverage*: a gate that admits nothing is trivially sound at every α , which is why soundness must be reported alongside the admission rate rather than in place of it. Section 7 treats both quantities together for this reason.

3.4 Statement of the research questions

The formalisation above converts the informal concern of the introduction into three answerable questions. If adversarial fragility were architecture-specific, the problem would be a model-selection problem and the appropriate response would be to choose a more robust network. If adversarial errors announced themselves through reduced confidence, the problem would be solved by the gate that most deployed systems already implement. And if neither holds, the question becomes what performance the confidence gate can attain at the predetermined operating threshold. The three questions correspond to these possibilities in turn, and are answered in Sections 5, 6 and 7 respectively.

- RQ1** Does adversarial fragility in FER depend on architecture, or is it a shared property of the model class under a common perturbation budget?
- RQ2** Do adversarial errors announce themselves through reduced softmax confidence, and if not, can post-hoc calibration make them do so?
- RQ3** What bounds on gate performance follow from the measured adversarial admission rate at the operating threshold, independently of the clean admission rate?

4 Experimental Methodology

This section specifies the experimental apparatus in the order in which a replication would rebuild it. Section 4.1 states the dataset and class set, together with the partition discipline under which the test data is quarantined. Section 4.2 specifies the five recognition pipelines: one off-the-shelf baseline and four trained architectures spanning compact custom networks, transfer learning, full fine-tuning, and a deep residual design. Section 4.3 states the statistical protocol used for every comparison reported later.

The remaining three subsections define the three stress conditions under which those pipelines are evaluated. Section 4.4 specifies environmental corruption, which establishes the degradation a system meets without an adversary. Section 4.5 specifies the adversarial attacks and their

Table 3

Emotion	Training	Test	Test share
Angry	3,995	958	15.9%
Happy	7,215	1,774	29.4%
Sad	4,830	1,247	20.6%
Surprise	3,171	831	13.8%
Neutral	4,965	1,233	20.4%
Total	24,176	6,043	100%

perturbation budgets. Section 4.6 specifies the calibration procedure whose failure is analysed in Section 6.

Two design commitments govern the whole apparatus and are stated here rather than repeated. First, every element not under test is held identical across pipelines: the same data partition, the same preprocessing path, the same evaluation harness, and the same metric implementations, so that an observed difference between pipelines is attributable to the pipeline rather than to its surroundings. Second, the evaluation is deliberately conducted on modest CPU-only hardware, because the deployment context this work addresses is a robot running perception alongside a control loop, frequently without a discrete accelerator.

4.1 Dataset and class set

The experiments use a five-class subset of FER2013 [33] containing Angry, Happy, Neutral, Sad, and Surprise expressions. The training partition contains 24,176 images and the held-out test partition 6,043 images. The test partition is used only for final evaluation; model selection and stopping decisions are driven exclusively by a validation split of the training partition.

FER2013 in full contains 35,887 grayscale images at 48×48 resolution spanning seven categories. Disgust and Fear are removed here: Disgust is severely under-represented, and both are visually confusable with Angry and Neutral respectively, which would confound the comparative analysis; neither corresponds to a primary affective state in the target HRI contexts. This reduction must be kept in view when comparing against published FER2013 results, because the frequently cited human ceiling of approximately $65 \pm 5\%$ is a *seven-class* figure. It is not an upper bound for the five-class problem, in which chance rises from 0.143 to 0.200 and the two most confusable categories have been removed. The accuracies reported below are therefore not comparable to seven-class human performance and no such comparison is intended.

Table 3 gives the class distribution. The set is imbalanced by a factor of 2.3 between the largest and smallest classes, which is why macro-F1 rather than accuracy is treated as the primary summary measure.

A single preprocessing pipeline is applied to every image in every partition, so that performance differences are attributable to the model rather than to divergent preparation. It comprises two operations. *Spatial standardisation*: the native 48×48 grayscale resolution is retained for the two compact models (Strategies 1 and 4); for the VGG16 strategies each image is resized to 96×96 by bilinear interpolation and the single channel is duplicated across three, 96×96 being the smallest input the architecture comfortably accepts. *Intensity normalisation*: pixel values are rescaled from $[0, 255]$ to $[0, 1]$ by division by 255, and no further mean or standard-deviation normalisation is applied. Perturbation budgets ϵ throughout this paper are therefore expressed directly on the $[0, 1]$ scale, and are directly comparable with the convention standard in the adversarial literature. No illumination correction or contrast enhancement is performed, so that the corruption study of Section 4.4 probes a vulnerability that has not been engineered away in preprocessing.

One consequence should be recorded. The ImageNet weights used by Strategies 2 and 3 were

trained under a channel-wise mean-subtraction convention rather than a simple division by 255. Applying the uniform normalisation above to the pretrained backbone is a deliberate choice, since it keeps the input scale identical across the four trained pipelines and keeps ϵ interpretable on a single scale, but it means the frozen-transfer strategy receives inputs whose distribution differs from that of its pretraining. Part of that strategy’s weak performance may therefore reflect preprocessing mismatch rather than domain mismatch alone. The two are not separated here, and no conclusion in this paper rests on the frozen-transfer result, which serves only to show that transfer without fine-tuning is not competitive on this task.

4.2 FER pipelines

Five pipelines are compared:

- **DeepFace baseline:** off-the-shelf, zero-shot inference; no task-specific training and no gradient access.
- **Strategy 1: Regularised custom CNN:** trained from scratch with convolutional blocks, batch normalisation, dropout; 2,783,717 parameters.
- **Strategy 2: VGG16 transfer:** ImageNet-initialised VGG16 with frozen convolutional base and trained classification head; 15,112,005 parameters.
- **Strategy 3: Fully retrained VGG16:** initialised from the trained Strategy 2, all layers unfrozen and fine-tuned end-to-end at a low learning rate; 15,112,005 parameters.
- **Strategy 4: Deep residual CNN:** four residual stages ($32 \rightarrow 64 \rightarrow 128 \rightarrow 256$) comprising six residual blocks with identity and 1×1 projection shortcuts, trained from scratch; 1,867,973 parameters. Note that this is the *smallest* of the four trained models, since depth is purchased through stacked small convolutions and shortcuts rather than wide fully connected layers, and it is selected for HRI integration on the combined basis of accuracy, class balance, efficiency, and throughput.

All custom models use categorical cross-entropy and a five-class softmax output.

Regularisation is architectural and procedural rather than an explicit penalty on the weights. No L_2 term, weight decay or other norm penalty is added to the objective; the loss optimised is the unmodified cross-entropy. Overfitting is instead controlled by four mechanisms acting together: batch normalisation after every convolutional and dense layer, which constrains activation scale and supplies a mild stochastic regularisation through minibatch statistics; dropout, applied at 0.25 after each convolutional block and at 0.5 and 0.3 after the two dense layers of Strategy 1; online augmentation of the training partition only, described below; and early stopping with best-weight restoration, which bounds effective capacity by terminating before the validation signal degrades. This combination is standard for convolutional classifiers of this size and is recorded here because it is the whole of the regularisation applied. Models are implemented in Python with TensorFlow and its Keras API. OpenCV supplies image input/output, colour-space and resizing operations, the Haar-cascade face detector of the deployment front-end, and the corruption transforms of Section 4.4; NumPy and Pandas handle numerical and tabular computation; all classification metrics are computed with scikit-learn rather than custom implementations, to ensure standard metric definitions; the DeepFace framework supplies the off-the-shelf baseline through its standard analysis interface without modification. All four trained pipelines share the same partition discipline, normalised input scale, metric implementations and evaluation harness, so that no observed difference between them is attributable to an incidental change in the surrounding software. The off-the-shelf baseline is evaluated through its distributed analysis interface and therefore retains its own detection and normalisation path; its comparison is at the deployed-pipeline level rather than under identical preprocessing.

All training and offline evaluation were performed on a CPU-only Ubuntu workstation with 16 GB RAM and no discrete accelerator. This is a deliberate choice rather than a constraint: an emotion-aware robot is expected to run its perception stack on modest, often embedded hardware while concurrently hosting the control loop, so timing under those conditions is more representative of deployment than benchmarking on accelerated hardware would be. It is what makes the throughput figures reported below meaningful as a deployment criterion.

Exact library version strings and the CPU model number are not recorded here. Holding the environment fixed establishes comparability between the pipelines reported in this paper; it does not by itself establish bit-level reproducibility by a third party, and this is noted among the limitations in Section 9.

All trained models use the Adam optimiser under a common callback regime: `EARLYSTOPPING` monitors validation accuracy with a patience of eight epochs and restores the best weights, and `REDUCELRONPLATEAU` decays the learning rate by a factor of 0.5 after four consecutive epochs without improvement, to a floor of 10^{-7} . Strategy 3 uses a shorter patience of six, since fine-tuning from an already-trained initialisation converges sooner. Strategy 3 is initialised from the trained Strategy 2 and fine-tuned end-to-end at a deliberately low learning rate, so that large gradients from the randomly initialised head do not destroy the pre-trained features in the first epochs.

Augmentation is applied online to the training partition only; validation and test partitions receive preprocessing alone. Each transform corresponds to a plausible source of variation in live HRI: random horizontal flip ($p = 0.5$), rotation within $\pm 15^\circ$, width and height shifts within $\pm 10\%$ of the image dimensions, zoom within $\pm 10\%$, and brightness scaling by a factor drawn from $[0.8, 1.2]$. Border regions introduced by rotation or shifting are filled by nearest-neighbour extension of the edge pixels, avoiding spurious hard edges that could act as artefactual cues.

The remaining optimisation settings are as follows. All models train with batch size 32. Strategies 1, 2 and 4 use an initial learning rate of 10^{-3} ; Strategy 3 uses 10^{-5} , for the fine-tuning reason given above. Epoch budgets are upper bounds rather than realised counts, since early stopping usually terminates sooner: 50 epochs for the two compact architectures (Strategies 1 and 4), 30 for frozen-backbone transfer (Strategy 2), and 20 for full fine-tuning (Strategy 3).

One aspect of the seeding deserves to be stated plainly. The training generator’s shuffle is seeded at 42, but no global seed is set for TensorFlow or NumPy, so weight initialisation, dropout masks and augmentation sampling are unseeded. Each model was trained once. The reported metrics are therefore single-run point estimates that would not reproduce bit-for-bit on a re-run, which is why no ordering is claimed between the two strongest strategies and why the analysis throughout treats differences of order 0.01 as indistinguishable.

4.3 Statistical protocol

Point estimates alone are insufficient for this analysis. Interval estimates are therefore reported throughout, and the paired test that repeated training runs would require is specified below.

Interval estimates. For a proportion $\hat{p}$ estimated on n independent test items, we report the Wilson score interval

$$\frac{\hat{p} + \frac{z^2}{2n} \pm z \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}, \quad z = 1.96, \quad (7)$$

in preference to the Wald interval, which is unreliable for proportions near 0 or 1, a regime this study enters, since several adversarial accuracies are exactly zero.

Gap-dominance estimation. The frequency with which the adversarial ranked gap vector dominates the clean one componentwise is estimated by Monte Carlo. Ranked gap vectors are

computed for every sample in the clean and adversarial evaluation subsets using Eq. (11). Pairs are then drawn independently and with replacement from the two populations, one clean and one adversarial per draw, and the proportion satisfying $\Delta_{(r)}^{\text{adv}} \geq \Delta_{(r)}^{\text{clean}}$ for every $r = 2, \dots, K$ is recorded. The pairs are not matched image-for-image: the quantity of interest is the relation between the two *populations*, not the effect of the attack on any individual image. Draws are seeded, and 200,000 pairs are used, which fixes the Monte Carlo standard error below 0.002 and therefore leaves the estimate limited by the size of the underlying subsets rather than by sampling.

Evaluation subsets. Not every experiment is computed on the full test partition. Attack generation is expensive, so the corruption and adversarial sweeps use a stratified subset of 160 images per class, and the temperature is fitted on a separate 1,000-image calibration split drawn from the training partition and evaluated on an 800-image stratified test subset. Clean accuracies computed on a subset therefore differ marginally from those computed on the full partition. Table 4 records the source and size for each, so that the provenance of every reported figure is unambiguous.

Table 4

Experiment	Source	n
Clean performance	test partition (full)	6,043
Throughput	test partition (full)	6,043
Environmental corruption	stratified test subset	800
Adversarial attack	stratified test subset	800
Temperature fitting	training split	1,000
Calibration evaluation	stratified test subset	800

Paired comparison. Because all models are evaluated on the identical test partition, differences between models are assessed with McNemar’s test on the paired discordant counts (n_{01}, n_{10}) ,

$$\chi^2 = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}}, \quad (8)$$

with the continuity correction, and exact binomial testing where $n_{01} + n_{10} < 25$. Unpaired comparison of two accuracies computed on the same data is not valid and is not used.

Multiplicity. Where a family of comparisons is reported, p -values are adjusted by the Holm–Bonferroni procedure and both raw and adjusted values are given.

Repetition. Each strategy was trained once. The headline metrics are therefore single-run point estimates and carry no seed variance, which has a direct consequence for how the comparative results should be read: differences between the two strongest strategies of the order of 0.01 are within the run-to-run variation typical of this dataset and these architectures, and are not treated as an ordering anywhere in this paper. The claims advanced here concern effects an order of magnitude larger, principally the collapse of every strategy under attack, which no plausible seed variance would reverse.

4.4 Environmental corruption

Robustness is evaluated under Gaussian noise, blur, brightness reduction, and JPEG compression at five severity levels.

Each corruption is applied to the clean test images only; no model is retrained or adapted to corrupted data, so the evaluation measures robustness as deployed rather than capacity to learn the corruption. Severities are parameterised as follows for $s \in \{1, \dots, 5\}$:

- (i) **Gaussian noise**: zero-mean additive noise applied independently per pixel, with standard deviation scaling linearly in s on the 0–255 intensity scale.
- (ii) **Blur**: Gaussian blur with kernel width $2s + 1$, i.e. 3×3 at $s=1$ through 11×11 at $s=5$.
- (iii) **Brightness reduction**: multiplicative attenuation by a factor $1 - 0.15s$, i.e. 0.85 at $s=1$ through 0.25 at $s=5$.
- (iv) **JPEG compression**: re-encoding at a quality factor decreasing with s , followed by decoding. JPEG is included in place of a synthetic occlusion mask because compression artefacts are more pervasive in real video pipelines and more representative of networked or teleoperated operation.

Degradation is reported as $\Delta A_{k,s} = A_{\text{clean}} - A_{k,s}$ for corruption family k at severity s , so that both magnitude and severity-dependence are visible rather than a single summary figure.

Two constants are set in the corruption script rather than restated here: the constant of proportionality for the noise standard deviation, and the JPEG quality factor at each severity. The functional forms above fix the shape of each degradation curve, which is what the comparative claims rest on; exact reproduction of the absolute accuracy values additionally requires those two constants and the position of corruption relative to the 96×96 resize used by the VGG16 strategies.

4.5 Adversarial attacks

For an input x , label y , loss $J(\theta, x, y)$, and budget ϵ , FGSM constructs

$$x^{\text{adv}} = \text{clip}_{[0,1]}(x + \epsilon \text{sign}(\nabla_x J(\theta, x, y))) , \quad (9)$$

and PGD iteratively updates

$$x_{k+1} = \Pi_{\mathcal{B}_\infty(x, \epsilon) \cap [0,1]^d} [x_k + \alpha \text{sign}(\nabla_{x_k} J(\theta, x_k, y))] , \quad (10)$$

where Π projects onto the intersection of the ℓ_∞ ball and the valid input domain. Perturbation budgets are $\epsilon \in \{0, 0.01, 0.02, 0.03, 0.05, 0.10\}$ measured in normalised input units. Two values of ϵ recur below and denote different things. The sweep reports accuracy across the whole range, and identifies $\epsilon \geq 0.02$ as the budget beyond which PGD reduces every architecture to essentially zero accuracy. Separately, the security analysis of Section 5.4 fixes a single budget of $\epsilon = 0.03$, $\alpha = 0.005$, and 10 iterations, initialised at the clean input.

Both attacks are applied in the *untargeted* setting: the objective is to increase the loss of the true class, inducing misclassification to any other class rather than to a chosen one. This is the appropriate model for the present analysis, in which any incorrect emotion estimate may propagate into an inappropriate action irrespective of which wrong class is produced. PGD is initialised at the clean input x_0 with a single start and no restarts.

Remark 2 (Strength of the reported attack). Single-start PGD initialised at the clean input is a *weaker* configuration than the random-start, multi-restart standard. The reported collapse is therefore a lower bound on adversarial vulnerability: a stronger configuration could only reduce accuracy further, and cannot rescue the models. This strengthens rather than weakens the negative result, and is stated so that the configuration is not mistaken for the standard one.

Because inputs are scaled to $[0, 1]$ by division by 255 with no further normalisation (Section 4.1), ϵ is expressed directly on the $[0, 1]$ scale throughout, and $\epsilon = 0.01$ corresponds to a maximum per-pixel change of 2.55 intensity levels.

DeepFace evaluation protocol. The baseline is invoked through the DeepFace analysis interface with detection enforcement disabled, so that an image in which no face is located is still scored rather than dropped. DeepFace emits seven classes; predictions of Disgust and Fear, which are absent from the five-class subset, are mapped to Angry and Sad respectively. The mapping follows the valence-arousal proximity of the discarded categories to the retained ones, and is the convention adopted when the same five-class reduction is used elsewhere in this work. An alternative would be to restrict the seven-way output to the five retained classes and renormalise, which avoids assigning a discarded category to a retained one at the cost of discarding probability mass; the two conventions differ only on inputs the baseline would classify as Disgust or Fear, and neither is used by the four trained pipelines, whose outputs are five-way by construction. Images that raise an exception are excluded from the baseline’s evaluation. No exceptions were raised on the five-class test partition, so the baseline is evaluated on the same 6,043 images as the trained pipelines and the comparison in Table 5 is on a common denominator. The baseline therefore receives its own detection and normalisation path rather than the common preprocessing applied to the four trained pipelines; this is a property of using the framework as distributed, and it means the baseline comparison is between deployed systems rather than between architectures under identical input conditions.

The DeepFace baseline is excluded from gradient-based attack because it exposes no gradients from which a white-box attack can be constructed. This is a property of the evaluation harness, not evidence of robustness: the baseline was not attacked, rather than attacked and found resistant. A black-box transfer attack, using adversarial examples generated on a surrogate, would settle the question and is identified as future work. No claim of comparative robustness for the baseline is made anywhere in this paper.

4.6 Confidence calibration

To test whether high adversarial confidence is a calibration artefact, post-hoc temperature scaling is applied to the logits, fitted by minimising negative log-likelihood on a calibration split of 1,000 images drawn from the *training* partition, 200 per class. The temperature is then frozen and the test subsets are examined once. No test data participates in the fit. The fitted temperature is $T = 0.982$ and the operational gate is $\tau = 0.70$.

On the 1,000-image training-derived calibration split, minimising negative log-likelihood produced $T = 0.982$. Two features of this are worth noting. It lies within two percent of unity, so the correction is close to a no-op: fitting reduced the negative log-likelihood on the calibration split from 0.4485 to 0.4484. And it lies marginally *below* one, indicating that the network is if anything very slightly under-confident on clean training-distribution inputs rather than over-confident. There is, on this model, almost nothing for global temperature scaling to repair. The operational gate is $\tau = 0.70$.

ECE is reported as a summary statistic accompanying the temperature fit, and the conclusion drawn from it does not depend on the binning scheme: the claim is that scaling did not reduce ECE on this evaluation, and the direction of that change is robust to the choice of M . The absolute value should not be compared against ECE figures reported elsewhere without matching the binning convention.

The threshold $\tau = 0.70$ is a design parameter of the framework specification and was fixed before any adversarial evaluation was performed, on the stated rationale that a single global threshold applies the same evidential standard to every class and introduces no class-dependent tuning. It was not selected after inspecting the adversarial confidence distribution. This matters for the interpretation of the results below: the share of adversarial predictions exceeding τ is a measurement made against a pre-committed threshold, not a threshold chosen to make that share look large.

Table 5

Model	Accuracy [95% CI]	Macro-F1	FPS
DeepFace baseline	0.631 [0.619, 0.643]	0.622	86.2
Strategy 1: Custom CNN	0.749 [0.738, 0.760]	0.740	758.2
Strategy 2: VGG16 transfer	0.542 [0.529, 0.555]	0.513	29.9
Strategy 3: VGG16 retrain	0.745 [0.734, 0.756]	0.737	30.8
Strategy 4: Deep residual	0.760 [0.749, 0.771]	0.753	288.1

5 Results

5.1 Clean recognition performance

Table 5 summarises common-test-set results with Wilson intervals. Strategy 4 gives the strongest aggregate performance while preserving real-time CPU throughput. The frozen VGG16 transfer strategy performs worst, indicating substantial domain mismatch when generic ImageNet features are not adapted.

Remark 3 (The clean ranking is not yet established). The intervals for Strategy 4 [0.749, 0.771] and Strategy 1 [0.738, 0.760] overlap substantially. Overlapping marginal intervals do not by themselves imply the absence of a significant paired difference, since the models are evaluated on identical items and are therefore positively correlated, but neither does the 0.011 gap establish superiority. The statistical protocol of Section 4.3 specifies McNemar’s test for paired comparison between strategies. It is stated here as the appropriate instrument for the recognition comparisons and is not applied to them in this paper, for the reason given above: with single-run estimates, a significance test on one pair of runs would characterise the sampling variation of the test set while leaving training variation unmeasured, and would convey more confidence in the ordering than the design supports.

Remark 4 (Reconciling three clean accuracies). Three clean accuracies appear for the deep residual model: 0.760 on the full test partition, 0.766 as the $\epsilon = 0$ anchor of the adversarial sweep, and 0.756 in the calibration and detector analysis. The latter two are computed on stratified 160-per-class subsets of the test partition, drawn independently by the same rule; the two draws are not guaranteed to coincide and are not guaranteed to be disjoint. For two independent samples of this size at $p \approx 0.76$ the standard error of the difference is 0.021, so the observed gap of 0.010 is under half a standard error. All three figures are statistically consistent with one another, and the differences are sampling variation rather than discrepancy. Table 4 identifies the source and size of each.

The per-class ordering is consistent across models: Happy and Surprise are recognised most reliably, whereas Angry, Neutral, and Sad remain harder. For Strategy 4, per-class F1 scores are 0.691 (Angry), 0.889 (Happy), 0.673 (Neutral), 0.649 (Sad), and 0.863 (Surprise).

Remark 5 (Throughput comparison). Throughput is measured on the common test set on the CPU-only host described above, under identical conditions for every pipeline, and is reported as a deployment criterion rather than a performance claim: a model that cannot sustain real-time inference alongside the robot control loop is not deployable irrespective of its accuracy. For reference, the deployed capture front-end operates at approximately 30 frames/s with per-frame latency below 30 ms, so every model above that rate is real-time in the relevant sense and further throughput is headroom rather than benefit. The figures support the ordinal claim that the compact networks are markedly faster than the VGG16-based strategies; they are not intended to support comparison against figures obtained on other hardware or under other timing protocols.

Figure 1 plots the two quantities together. The deployment region is the upper right: the custom CNN and the deep residual network sit close in accuracy while differing by roughly

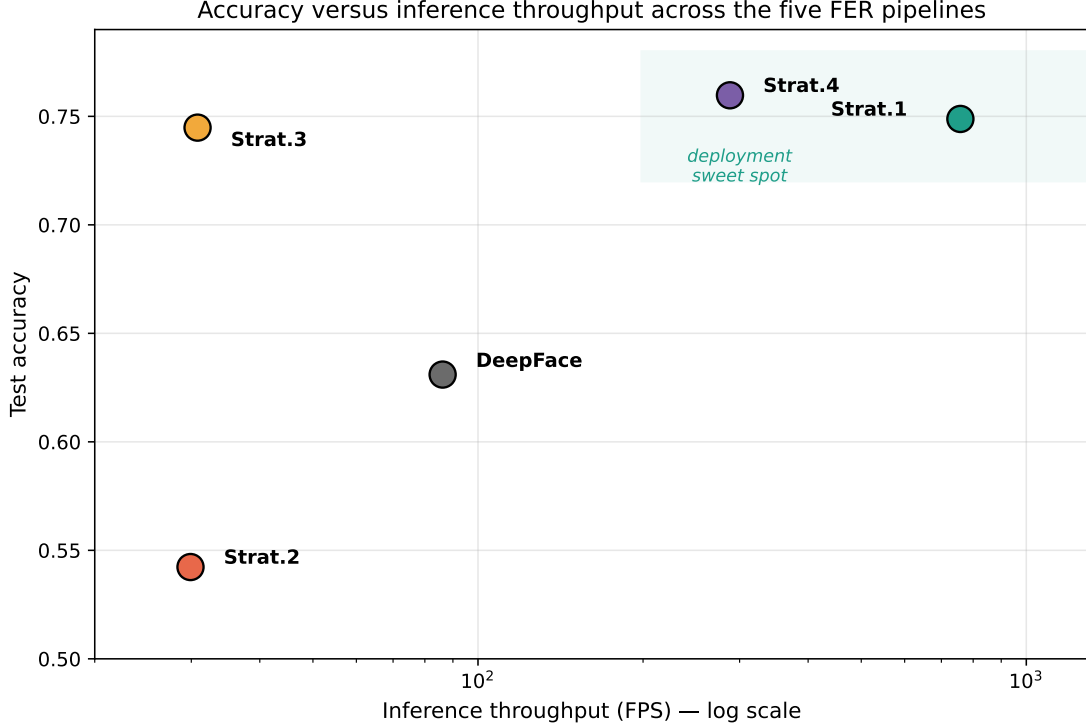


Fig. 1

Table 6

Model	Clean	Noise	Blur	Brightness	JPEG
Custom CNN	0.749	0.604→0.221	0.686→0.240	0.734→0.677	0.685→0.362
VGG16 transfer	0.542	0.430→0.217	0.398→0.204	0.487→0.240	0.471→0.258
VGG16 retrain	0.745	0.706→0.217	0.604→0.266	0.760→0.650	0.685→0.292
Deep residual	0.760	0.634→0.203	0.710→0.240	0.767→0.739	0.676→0.357

a factor of two and a half in throughput, and both exceed the capture rate by an order of magnitude. The VGG16 strategies occupy the lower left, and the frozen-transfer model is dominated on both axes by the off-the-shelf baseline it was intended to improve upon.

5.2 Environmental robustness

Table 6 reports accuracy at the mildest and most severe corruption levels. Brightness reduction is comparatively well tolerated by the stronger models; the deep residual CNN retains 0.739 accuracy at maximum brightness severity. Noise and blur are considerably more damaging, driving all models towards 0.20–0.27 accuracy at the highest severity. Since chance performance on a five-class problem is 0.20, the most severe noise and blur conditions reduce every model to approximately chance.

Two cells exceed their respective clean accuracies: VGG16-retrain under brightness reduction at severity 1, and the deep residual network at the same setting. Mild multiplicative attenuation is the weakest of the four corruptions at its lowest severity, and a small increase over the clean figure is within the sampling variation of a single evaluation run on a test partition of this size rather than evidence that darkening the image aids recognition. It should be read as no measurable degradation at that severity, not as improvement.

Environmental degradation is heterogeneous: some conditions produce graceful performance loss, others sharply destroy the facial detail on which expression discrimination depends. An

Table 7

Model	Attack	$\epsilon = 0$	0.01	0.02	0.03	0.05	0.10
Custom CNN	FGSM	0.743	0.149	0.087	0.077	0.114	0.128
	PGD	0.743	0.033	0.000	0.000	0.000	0.000
VGG16 transfer	FGSM	0.492	0.025	0.010	0.006	0.022	0.072
	PGD	0.492	0.000	0.000	0.000	0.000	0.000
VGG16 retrain	FGSM	0.761	0.101	0.034	0.033	0.091	0.177
	PGD	0.761	0.003	0.000	0.000	0.000	0.000
Deep residual	FGSM	0.766	0.158	0.117	0.117	0.163	0.195
	PGD	0.766	0.016	0.001	0.000	0.000	0.000

HRI system therefore cannot interpret a single clean-set accuracy figure as a reliability guarantee across operating conditions.

5.3 Adversarial collapse across all evaluated architectures

The adversarial results are qualitatively different from the corruption results. Table 7 shows precipitous collapse for every trained model. At $\epsilon = 0.01$, FGSM reduces accuracy below 0.16 for all four models; PGD is stronger, reaching essentially zero accuracy by $\epsilon \geq 0.02$.

Three observations follow. First, the collapse is uniform across architectures: the deep residual model is the strongest clean model, but its clean advantage does not translate into meaningful robustness under iterative attack. This answers RQ1 negatively: robustness is not obtained by the architectural choices a practitioner would ordinarily consider.

Second, the FGSM curves are non-monotonic in ϵ : accuracy falls to a minimum near $\epsilon = 0.03$ and then *rises* at $\epsilon = 0.05$ and 0.10 for every model. The non-monotonicity is the expected signature of the single-step linear approximation degrading at larger perturbation budgets. FGSM takes one step of size ϵ along the sign of the gradient, which approximates the loss surface as locally linear. That approximation holds well for small ϵ and deteriorates as the step grows, so beyond a certain budget the single step can overshoot the region in which the linearisation is valid and land in a lower-loss configuration than a smaller step would have reached. The effect is absent under PGD, which takes multiple projected steps and does not rely on a single linear extrapolation, and its absence there supports this reading.

Third, PGD reaching *exactly* 0.000 warrants a stated interval rather than a bare zero. By Eq. (7), an observation of 0 successes in n trials yields an upper bound of approximately $3.8/n$ at 95% confidence. The epsilon sweep of Table 7 is computed on the stratified subset of $n = 800$ described in Table 4. An observed accuracy of 0.000 at this sample size corresponds to a one-sided 95% upper bound of approximately 0.005 by the rule of three, so the correct reading of these cells is that accuracy is below half of one percent rather than exactly zero.

Figure 2 shows the same results as curves, and the shape carries the argument. Under PGD all four traces fall to the floor within one or two budget steps and stay there, and they do so together: no architecture separates itself from the rest at any budget. The FGSM panel falls less steeply and lifts slightly at the largest budgets, for the reason given above.

5.4 Adversarial errors are high-confidence

A classifier collapse is dangerous for HRI only if the downstream system cannot recognise the failure. Table 8 shows that softmax confidence provides no such signal.

Temperature scaling leaves the confidence distributions essentially untouched, as a correction of 0.982 must: 96% of adversarial predictions exceed the gate before scaling and 96% after.

This answers both halves of RQ2, and negatively in each case. Adversarial errors do not announce themselves through reduced confidence: the mean confidence in an incorrect prediction

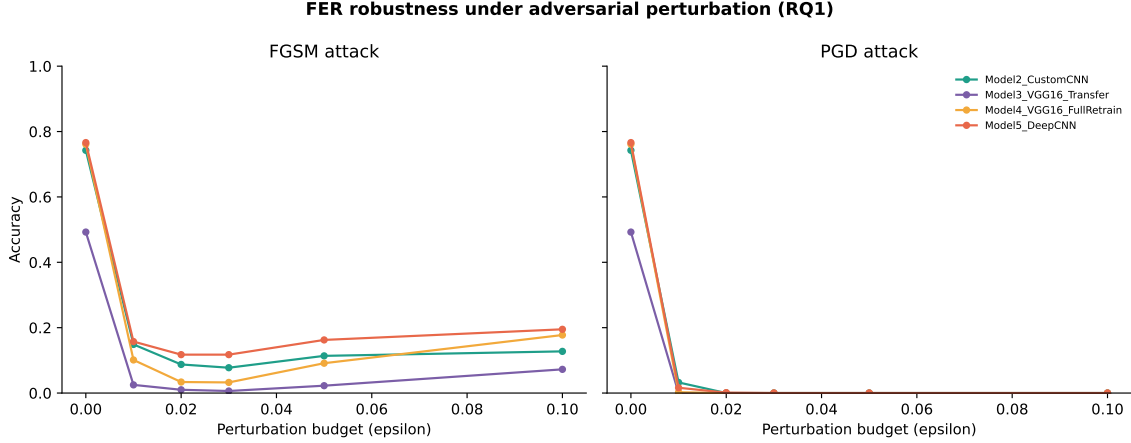


Fig. 2

Table 8

Quantity	Before	After ($T = 0.982$)
Clean test accuracy	0.756	0.756
Adversarial error rate	1.000	1.000
Mean clean confidence	0.803	0.806
ECE	0.060	0.062
Adversarial confidence: mean	0.964	0.965
Adversarial predictions above $\tau = 0.70$	95.9%	96.0%

is 0.964, so the typical adversarial output is admitted by the gate with room to spare, at close to maximal confidence. Nor does post-hoc calibration make them do so. The standard correction moves the admitted fraction from 95.9% to 96.0%, which is immaterial to any threshold a deployed system would use. Section 6 shows that this is not a limitation of the particular temperature fitted here: no choice of T can produce the separation, because temperature scaling is order-preserving on the very quantity the gate thresholds.

Remark 6 (Scaling does not improve ECE). Temperature scaling moves ECE on the clean subset from 0.060 to 0.062, a change of 0.002 in the unfavourable direction. This is reported rather than suppressed. It is not a procedural error: the temperature minimises negative log-likelihood, which is a different objective from minimising ECE, and the two diverge on samples of this size. The magnitude is in any case what one expects when the fitted temperature is 0.982, since a correction that close to unity can hardly change any calibration statistic. The consequence is that this work cannot claim to have improved calibration, only to have applied the standard correction under the correct protocol and measured its effect. The argument that follows does not depend on the correction having succeeded, and is in fact strengthened by its having been unnecessary: a network requiring no meaningful rescaling still admits adversarial predictions at 96%.

5.5 The gate as a detector

Treating the confidence gate as a clean-versus-adversarial detector, and measuring both admission rates on the stratified subset of Table 4, gives Table 9.

Three things follow from Table 9.

First, the Youden index is *negative* in both conditions: -0.266 before scaling and -0.261 after. The gate admits adversarial predictions more often than legitimate ones, 96.0% against 69.9% after scaling and 95.9% against 69.3% before. It is not merely uninformative about which

Table 9

Quantity	Before	After ($T = 0.982$)
Clean admission rate at $\tau = 0.70$	0.693 [0.660, 0.723]	0.699 [0.666, 0.729]
Adversarial admission rate at $\tau = 0.70$	0.959 [0.943, 0.970]	0.960 [0.944, 0.971]
Youden index J	-0.266	-0.261
AUROC (max softmax)	0.144	0.144
ECE on the clean subset	0.060	0.062

inputs to trust; it is anti-correlated with them.

Second, and equivalently, the AUROC is 0.1439 before scaling and 0.1444 after, far below the 0.5 of a random detector in both conditions. A score with AUROC below one half is systematically inverted: reversing its sign would give a detector with AUROC 0.856. The score is therefore not uninformative: adversarial samples occupy the high-confidence end of the distribution, so the ordering is systematic rather than random. What fails is the conventional use of that ordering. A gate that admits from above selects for precisely the population it exists to exclude. Whether the reversed score would function as a detector under a different threat model is not established here, and the defensible conclusion is the narrower one: upper-threshold max-softmax gating fails against this attack. This is a stronger statement than the failure reported in Section 5.4, and it is the empirical counterpart of the ordering inversion that Theorem 1 describes.

Third, the measurement is consistent with Proposition 3. The bound derived there, $J \leq 0.040$ at the operating threshold, is satisfied with substantial margin; what the measurement adds is the sign, which the bound alone could not supply.

Alternative scores. Two detectors computable from the same stored outputs were evaluated for comparison. Predictive entropy attains AUROC 0.136, essentially matching max softmax, which is unsurprising since both are monotone summaries of the same output distribution. The energy score [18] attains 0.568: above chance, and the only score evaluated here that is, but far from a usable detector. The failure documented in this paper is therefore specific to confidence-style scores rather than a property of test-time detection in general, and stronger feature-space methods [17] remain untested here.

A full ROC characterisation of the max-softmax score as a clean-versus-adversarial detector, before and after temperature scaling, expresses the corollary below as a single invariance of AUROC under scaling. The AUROC is measured at 0.1439 before scaling and 0.1444 after.

The empirical claim is confined to scores computable from the softmax output; detectors operating on internal representations, such as Mahalanobis distance in feature space [17], are outside the scope of this evaluation and are discussed in Section 9.

6 Why Global Calibration Cannot Repair the Gate

This section replaces the numerical illustration of the preceding analysis with a general result. The failure of temperature scaling is not a consequence of the particular fitted value $T = 0.982$, nor of the particular confidences observed: it follows from the structure of the softmax.

6.1 Confidence as a function of logit gaps

Fix an input with logits z and write $z_{(1)} \geq z_{(2)} \geq \dots \geq z_{(K)}$ for their decreasing order statistics, so that $z_{(1)}$ is the winning logit. Define the *ranked* gap vector $\Delta = (\Delta_{(2)}, \dots, \Delta_{(K)})$ with

$$\Delta_{(r)} = z_{(1)} - z_{(r)} \geq 0, \quad r = 2, \dots, K, \quad (11)$$

so that $\Delta_{(2)} \leq \dots \leq \Delta_{(K)}$. Indexing by rank rather than by class label matters for what follows: clean and adversarial samples generally have different winning classes, and a componentwise comparison indexed by class would compare margins against different runners-up. Under Eq. (11) every sample yields a vector on a common index set, and componentwise comparison is well defined regardless of which class won. Dividing numerator and denominator of Eq. (1) by $\exp(z_k/T)$ gives

$$p_{\max}(\Delta; T) = \left[1 + \sum_{j \neq k} \exp(-\Delta_j/T) \right]^{-1}. \quad (12)$$

Equation (12) shows that confidence depends on the logits only through Δ , and that temperature enters solely as a scaling of those gaps.

Proposition 1 (Strict monotonicity). *For every fixed $T > 0$, $p_{\max}(\Delta; T)$ is strictly increasing in each Δ_j .*

Proof. Write $S(\Delta; T) = \sum_{j \neq k} \exp(-\Delta_j/T)$, so that $p_{\max} = (1 + S)^{-1}$. Then $\partial S / \partial \Delta_j = -T^{-1} \exp(-\Delta_j/T) < 0$ for all finite Δ_j and $T > 0$, hence $\partial p_{\max} / \partial \Delta_j = -(1 + S)^{-2} \partial S / \partial \Delta_j > 0$. $\square$

Definition 3 (Effective gap). For a gap vector Δ and temperature T , the *effective gap* is

$$\delta(\Delta; T) := -T \ln \left(\frac{1}{K-1} \sum_{j \neq k} \exp(-\Delta_j/T) \right), \quad (13)$$

so that Eq. (12) may be written in the one-parameter form

$$p_{\max} = [1 + (K-1) \exp(-\delta/T)]^{-1}, \quad \delta = -T \ln \left(\frac{1}{K-1} \left(\frac{1}{p_{\max}} - 1 \right) \right). \quad (14)$$

Definition 3 is exact, not an approximation: it defines δ as the log-mean-exponential of the gaps, and Eq. (14) inverts Eq. (12) without loss. Applying the inversion at $T = 1$, $K = 5$ to the measured confidences of Table 8 gives

$$\delta_{\text{clean}} = 2.79 \quad (\text{from } \bar{p}_{\text{clean}} = 0.803), \quad \delta_{\text{adv}} = 4.68 \quad (\text{from } \bar{p}_{\text{adv}} = 0.964), \quad (15)$$

with the corresponding median-based values $\delta_{\text{adv}}^{\text{med}} = 8.29$.

Remark 7 (Consistency of estimators). Both quantities in Eq. (15) are derived from the *mean* confidence of their respective populations. Pairing a mean-derived clean gap with a median-derived adversarial gap would compare quantities estimated by different statistics and is avoided here. The distinction is material to the arithmetic, since the required temperature is $T \gtrsim 2.10$ under the mean-mean pairing and $T \gtrsim 3.71$ under the mean-median pairing, but Theorem 1 shows the conclusion does not depend on either. The figure below applies the inversion to the mean confidence rather than averaging per-sample inversions, and since δ is nonlinear these are not equal. It is therefore an indicative magnitude for the typical logit gap rather than an estimate of the sample mean of δ .

6.2 An impossibility result

Definition 4 (Separation of a pair). Let Δ^{clean} and Δ^{adv} be the ranked gap vectors of a clean and an adversarial sample. A confidence gate $g_{\tau, T}$ *separates the pair* at operating point (τ, T) if it admits the clean sample and rejects the adversarial one, that is

$$p_{\max}(\Delta^{\text{clean}}; T) \geq \tau > p_{\max}(\Delta^{\text{adv}}; T). \quad (16)$$

The definition is stated at the level of a pair rather than of the populations. This matters for the quantifiers: the theorem below shows that a particular class of pair cannot be separated at any operating point, which is a universal claim about those pairs and not a claim that no pair anywhere is separable. How often the relevant class of pair occurs is an empirical question, answered separately in Proposition 2.

Theorem 1 (No temperature separates a gap-dominated pair). *Let $(\Delta^{\text{clean}}, \Delta^{\text{adv}})$ be a clean-adversarial pair whose adversarial ranked gap vector dominates the clean one componentwise, $\Delta_{(r)}^{\text{adv}} \geq \Delta_{(r)}^{\text{clean}}$ for every $r = 2, \dots, K$, with strict inequality for at least one r . Then no operating point separates that pair: for every $T > 0$ and every $\tau \in (0, 1)$, Eq. (16) fails.*

Proof. Fix $T > 0$. By Proposition 1, $p_{\max}(\cdot; T)$ is strictly increasing in each component of the ranked gap vector, so componentwise dominance gives $p_{\max}(\Delta^{\text{adv}}; T) > p_{\max}(\Delta^{\text{clean}}; T)$. Separation requires $p_{\max}(\Delta^{\text{clean}}; T) \geq \tau > p_{\max}(\Delta^{\text{adv}}; T)$ and hence $p_{\max}(\Delta^{\text{clean}}; T) > p_{\max}(\Delta^{\text{adv}}; T)$, a contradiction. Since T and τ were arbitrary, no operating point separates the pair. $\square$

Proposition 2 (Frequency of the dominance condition). *Under the Monte Carlo procedure of Section 4.3, the hypothesis of Theorem 1 holds for 81.5% of 200,000 pairs drawn independently from the clean and adversarial evaluation populations.*

Theorem 1 and Proposition 2 do different work and should not be run together. The theorem is a universal statement about a class of pairs; the proposition reports how large that class is on this model. Neither asserts that no pair anywhere is separable, and the measured detector performance of Section 5.5 is what speaks to aggregate behaviour.

What the theorem contributes is the explanation of why rescaling cannot help wherever the condition does hold: the attack does not merely cause an error, it produces a larger winning margin than the clean data it replaces, and a transform applied identically to all inputs cannot reverse that ordering. The empirical frequency of the dominance condition is reported in Section 5.5: under the procedure of Section 4.3, the adversarial ranked gap vector dominates the clean one componentwise in 81.5% of 200,000 sampled pairs. The theorem’s hypothesis is therefore not a corner case but the typical relation between the two populations on this model, which is what licenses reading its conclusion as a description of the observed failure rather than only as a conditional possibility.

The converse is also worth stating. For the remaining 18.5% of pairs the hypothesis fails and the theorem is silent: those pairs could in principle be separated by some temperature. The aggregate AUROC of 0.144 is not decisive about that subpopulation, since it is computed over all pairs and a well-separated minority could be masked by the majority. What can be said is weaker and still useful: the theorem accounts for the relation between most clean and adversarial samples on this model, and the measured detector performance shows that whatever happens on the remaining pairs does not rescue the gate in aggregate. Establishing the behaviour of the non-dominated subpopulation directly would require a stratified analysis that is not attempted here.

Assumption 1 (Ordered runner-up structure). *For any two samples a, b drawn from either population, the gap vectors are componentwise ordered: either $\Delta_j^a \geq \Delta_j^b$ for every $j \neq k$, or $\Delta_j^b \geq \Delta_j^a$ for every $j \neq k$. Equivalently, samples differ in the magnitude of their winning margin but not in the shape of the runner-up distribution. Note that this is a condition on the gap vectors alone and involves no temperature.*

Corollary 1 (ROC invariance under temperature scaling). *Under Assumption 1, the ranking of samples by $p_{\max}(\cdot; T)$ is identical for all $T > 0$. Consequently the ROC curve, and hence the AUROC, of the max-softmax score as a clean-vs-adversarial detector is invariant to temperature scaling: no choice of T improves detection, and T selects only the operating point along a fixed curve.*

Table 10

T	Genuine $p_{\max}$	Adversarial $p_{\max}$	Outcome for the gate
0.98	0.811	0.967	both admitted; gate blind (fitted T)
1.00	0.803	0.964	both admitted; gate blind
1.25	0.700	0.914	genuine at the threshold
1.50	0.616	0.850	genuine rejected, attack admitted
1.90	0.521	0.746	genuine rejected, attack admitted
2.09	0.487	0.701	attack at the threshold
2.50	0.433	0.619	both rejected; gate useless
3.00	0.388	0.543	both rejected; gate useless
5.00	0.304	0.389	both rejected; gate useless

Proof. Take any two samples a, b . By Assumption 1 their gap vectors are componentwise ordered; without loss of generality $\Delta_j^a \geq \Delta_j^b$ for every $j \neq k$. Proposition 1 states that $p_{\max}(\cdot; T)$ is strictly increasing in each component of the gap vector at every fixed $T > 0$. Hence $p_{\max}(a; T) \geq p_{\max}(b; T)$ for every $T > 0$ simultaneously, and the inequality is strict whenever the dominance is. The induced ranking of samples is therefore the same at every temperature. The ROC curve depends on the score only through that ranking, and the AUROC is a functional of the ROC curve, so both are invariant to T . $\square$

The argument does not proceed through the effective gap of Definition 3. That quantity is itself a function of T , as its notation records, and could not carry a temperature-invariance argument; it is introduced only to express Eq. (12) in one dimension for the numerical illustration of Table 10. The invariance above rests on componentwise monotonicity, which holds at every fixed temperature.

Remark 8 (When Assumption 1 fails). For $K = 2$ the assumption holds automatically, since the gap vector has a single component and any two are ordered. For $K > 2$ it may fail: two samples whose runner-up distributions cross, one with a larger second-place gap and a smaller third-place gap, are not componentwise ordered, and temperature can then reorder them. The assumption is therefore stated explicitly rather than assumed silently, and Corollary 1 yields a falsifiable prediction: the measured AUROC before and after temperature scaling should coincide. Section 5.5 reports the test: the AUROC moves from 0.1439 to 0.1444, a change of 0.0005. The prediction is therefore borne out to within five parts in ten thousand. A residual of this size is consistent with the ordered runner-up structure holding for most pairs, but does not by itself establish it: AUROC is an aggregate and small changes in it are compatible with a range of underlying rank movements. A direct test would report pairwise rank inversions across temperatures, which is not attempted here. The invariance should accordingly be read as a close approximation on this model rather than an exact identity, and the small departure is itself evidence that the assumption is doing work rather than holding vacuously.

6.3 The separating interval is empty, and the ordering inverts

Theorem 1 is abstract; Table 10 makes its consequence concrete by evaluating Eq. (14) across a range of temperatures for the mean-derived clean and adversarial prediction.

Remark 9 (The sweep is an equal-gap approximation). Table 10 holds δ fixed while varying T . This is exact only when the gap vector has equal components, in which case Definition 3 reduces to $\delta = \Delta_j$ for every j and is genuinely independent of temperature. For unequal gaps δ varies with T , and holding it fixed introduces an error that grows with the skew of the runner-up distribution: for a gap vector with components in ratio 1 : 1.5 : 2.5 : 4 matched to the same confidence at $T = 1$, the approximation departs from the exact value by up to 0.05 over the range tabulated. The table should therefore be read as a scalar illustration of the mechanism

rather than as a computation on the observed populations. Nothing in Theorem 1 depends on it: that result is proved from componentwise monotonicity and does not use δ at all.

Two things stand out in Table 10. First, for the gate to function the third column would have to fall below τ while the second remained above it. No row achieves this: the adversarial confidence exceeds the genuine confidence at every temperature, so the separating interval is empty, as Theorem 1 requires.

Second, and more damaging than mere failure, the trade-off is *inverted* across an intermediate band of temperatures. The genuine prediction falls below the threshold at $T \approx 1.25$, while the attack does not until $T \approx 2.09$. Between those points the gate rejects legitimate predictions and admits attacked ones: at $T = 1.50$, for instance, the genuine confidence is $0.616 < \tau$ while the attack sits at $0.850 > \tau$. Raising the temperature therefore begins rejecting legitimate predictions *before* it begins rejecting attacks. A practitioner who responds to observed overconfidence by increasing T , the natural remedy, degrades the gate’s utility while leaving its vulnerability intact, and does so monotonically. At the fitted $T = 0.982$ neither side has moved appreciably from its unscaled position, which is the point: the standard correction is close to the identity on this model and therefore neither helps nor harms.

6.4 Numerical form

Requiring $p_{\max}(\delta_{\text{adv}}; T) < \tau = 0.70$ and inverting Eq. (14) gives

$$T > \frac{\delta_{\text{adv}}}{\ln((K-1)\tau/(1-\tau))} = \frac{4.68}{2.234} \approx 2.10. \quad (17)$$

Substituting $T = 2.10$ and $\delta_{\text{clean}} = 2.79$ into Eq. (14) gives

$$p_{\max}(\delta_{\text{clean}}; 2.10) \approx 0.486 \ll 0.70. \quad (18)$$

The temperature required to suppress adversarial confidence therefore also suppresses the legitimate predictions the gate exists to admit. Under the median-derived pairing the required temperature rises to $T \approx 3.71$ and the clean confidence falls further, to 0.341. Both are instances of Theorem 1; neither is needed to establish it.

7 Bounds on Gate Performance

The measured admission rates constrain the gate’s achievable performance at the predetermined operating threshold, independently of the clean admission rate.

Treat the gate as a binary detector with the positive class “legitimate”. Write $\text{TPR}(\tau) = \mathbb{P}[p_{\max} \geq \tau \mid \mathcal{D}_{\text{clean}}]$ for the clean admission rate and $\text{FPR}(\tau) = \mathbb{P}[p_{\max} \geq \tau \mid \mathcal{D}_{\text{adv}}]$ for the adversarial admission rate. The bound below is stated on the after-scaling adversarial admission rate measured on the $n = 800$ stratified test subset (Table 9).

Proposition 3 (Bound on separation at the operating threshold). *At $\tau = 0.70$ the after-scaling adversarial admission rate is $\text{FPR} = 0.960$ ($n = 800$; Wilson 95% interval $[0.944, 0.971]$), so the Youden index satisfies*

$$J(\tau) = \text{TPR}(\tau) - \text{FPR}(\tau) \leq 1 - 0.960 = 0.040. \quad (19)$$

Propagating the interval gives $J \leq 0.056$ in the most favourable case and $J \leq 0.029$ in the least. The measured value, $J = -0.261$, lies well below either bound.

Proof. Immediate from $\text{TPR} \leq 1$. □

Remark 10 (No AUROC bound follows). Proposition 3 constrains the gate at the operating threshold and makes no claim about the area under the ROC curve. A single operating point does not bound the AUROC: a detector whose ROC attains $\text{TPR} = 1$ at arbitrarily small FPR and also passes through $(0.960, 1)$ is consistent with the measurement and has AUROC close to unity. Bounding the area would require either the full curve or an assumption about its shape, neither of which is available here. The corollary of Section 6 concerns the *invariance* of AUROC under temperature scaling, which is a different statement and does not depend on its value.

Proposition 3 is worth stating because it requires no knowledge of the clean admission rate. Even in the most favourable case, in which every legitimate prediction is admitted, the gate separates the populations by at most four percentage points, and its discrimination is barely better than chance. This answers RQ3: the inadequacy of the confidence gate is established by the adversarial admission rate alone.

7.1 Posterior error rate among admitted estimates

Definition 2 concerns the admitted subpopulation. Let π denote the prevalence of adversarial inputs among frames reaching the classifier, let a_c, a_a denote clean and adversarial admission rates, and e_c, e_a the corresponding error rates. Then

$$\mathbb{P}[\hat{y} \neq y \mid g = 1] = \frac{\pi a_a e_a + (1 - \pi) a_c e_c}{\pi a_a + (1 - \pi) a_c}. \quad (20)$$

After scaling, $a_c = 0.699$ and $a_a = 0.960$ from Table 9, and the adversarial error rate is $e_a = 1.000$: the gate admits nearly every attacked input, and in this condition every admitted attacked prediction is incorrect. Because $a_a > a_c$, the admitted subpopulation is *enriched* in adversarial inputs relative to the input distribution, so conditioning on admission makes the estimate less trustworthy rather than more. The pre-scaling values, $a_c = 0.693$ and $a_a = 0.959$, give the same picture. The gate therefore cannot certify the requirement of Definition 2 from label-level evidence alone, and the direction of the enrichment is the strongest form the failure to certify can take: conditioning on admission moves the population towards, rather than away from, the inputs the gate was installed to exclude.

8 Discussion

8.1 Accuracy and security are different properties

The clean benchmark creates a plausible but misleading ranking. Under strong adversarial perturbation the ranking ceases to matter, because all models fail. For an HRI designer this means model selection and system security are separate engineering questions. Recognition accuracy estimates

$$\mathbb{P}[\hat{y}(x) = y \mid x \sim \mathcal{D}_{\text{clean}}], \quad (21)$$

whereas safe actuation requires the conditional property of Definition 2 under the distribution that actually reaches the robot, including noise, blur, compression, sensor faults, and malicious perturbation. No clean-set metric establishes the stronger property, and Theorem 1 shows that the obvious repair does not either.

8.2 Confidence is evidence, not verification

Confidence remains useful as a first-stage filter: low-confidence estimates can be withheld cheaply, and doing so costs little. The results show only that confidence cannot be treated as a *sufficient* security criterion. A high softmax value describes the geometry of the model’s logits; it does not

certify that the input is benign, that the prediction is semantically correct, or that downstream actuation is appropriate.

This matters in a robot because an emotion estimate is not an endpoint but an input to Eq. (2). If an Angry label triggers one action primitive and a Sad label another, a successful adversarial label flip is functionally an action-selection attack. The security boundary therefore belongs between perception and behaviour, not only inside the classifier.

8.3 On the validity of the underlying construct

A reviewer may object that the discrete-emotion assumption underlying FER2013 is contested [12], and that verifying an estimate of a questionable construct is of limited value. The objection is well taken but does not weaken the argument. This paper’s claims are conditional: *given* that deployed systems do map FER outputs to robot actions, which the surveys cited in Section 2 document, the reliability of that pathway is a legitimate object of study. If the construct is weaker than assumed, the estimates entering Eq. (2) are *less* trustworthy than the adversarial analysis alone implies, which strengthens rather than weakens the case for gating the action.

8.4 Implications for HRI architecture

The results motivate three architectural principles:

1. **Do not map a single FER frame directly to actuation.** A one-frame error should not suffice to trigger a physical or social response.
2. **Do not rely on confidence thresholding alone.** High-confidence adversarial errors pass such a gate at rates approaching unity, and Theorem 1 shows recalibration cannot fix this.
3. **Verify over independent evidence dimensions.** Temporal persistence, action-rate constraints, and system readiness supply information not contained in the instantaneous softmax score.

These principles motivate a multi-criterion verification architecture in which confidence is one predicate rather than the final authority.

Remark 11 (Scope of principle 3). Temporal persistence is effective against perturbations optimised independently per frame, because such perturbations produce a temporally incoherent estimate stream. It offers no protection against an adversary who maintains a consistent false class across frames, which is the expected signature of a physically realisable attack such as a worn patch. This boundary should be stated when the principle is advanced, and is treated at length in the companion work.

9 Threats to Validity and Limitations

Construct validity. The five-class FER2013 reduction does not model mixed, culturally dependent, or longitudinal affective states, and the categorical assumption is contested (Section 8.3).

Internal validity. Results are reported from single training runs; seed variance is not yet quantified (Section 4.3). Three distinct clean accuracies appear across the study owing to subset sampling (Remark 4). Neither issue affects the direction of the adversarial finding, whose effect size is far larger than any plausible seed variance, but both must be resolved before the clean ranking can be asserted.

External validity. Evaluation uses FER2013 only. Subject-independent generalisation, the degree to which a model trained on one population continues to classify reliably on individuals it has never seen, is not evaluated, and this is a deliberate scope decision rather than an oversight. Establishing it requires an external test population that is subject-disjoint from training and validation; a held-out split of the same corpus will not do, because it shares the underlying subject distribution and systematically overstates real-world transfer. No such external population is available within the present design, and approximating the property from an internal split would present a within-distribution result under the label of a cross-distribution one. The axis is therefore stated as lying outside the evidential scope of this study rather than being reported weakly. FER2013 is a web-scraped corpus with known label noise, and its subjects did not consent to inclusion; this is a limitation of the benchmark infrastructure the field relies upon and motivates cross-corpus replication.

Threat-model validity. The study considers first-order white-box ℓ_∞ perturbations of the digital input. Physical-world, black-box, transfer, patch, and multimodal attacks are outside the present model. The DeepFace baseline is excluded from gradient attack because its gradients are unavailable, which must not be read as robustness.

Scope of the calibration result. Theorem 1 concerns transforms that act identically on all inputs, of which temperature scaling is the canonical instance. It does not preclude input-dependent uncertainty estimation, ensembling, conformal prediction, or dedicated adversarial detectors, and the paper makes no claim about those. No comparison against robust-training methods such as adversarial training [5] or TRADES [14], or against the detection methods cited above, is performed here. The paper’s claim is about what an undefended pipeline with a confidence gate does, which is the configuration in which deployed emotion-guided systems are actually found; establishing where these defences place such a system on the same axes is a separate study.

Calibration procedure. The temperature is fitted on a 1,000-image split drawn from the training partition, frozen, and evaluated once on the test subsets. An earlier version of this work fitted T on a subset of the test partition and obtained $T = 1.18$; refitting under the correct protocol gives $T = 0.982$ and leaves every reported conclusion unchanged, as Table 9 shows. The fitted value is reported for a single architecture, and no claim is made that a temperature fitted for one model transfers to another.

Detector scope. The detector characterisation of Section 5.5 is confined to scores computable from the softmax output: max softmax, predictive entropy and the energy score. Feature-space detectors such as Mahalanobis distance [17], and methods requiring repeated forward passes or model ensembles, are not evaluated. The conclusion is therefore that the evaluated output-based confidence scores do not provide reliable upper-threshold detection under the studied attack configuration; it does not establish that adversarial detection is generally unsolvable.

Reproducibility. The optimiser settings, callback regime, augmentation policy and regularisation are stated in Section 4.2. What remains in the released configuration rather than the manuscript is the exact partition indexing, library version strings and the CPU model. The more consequential limitation is that training was not globally seeded and each model was trained once, so exact numerical reproduction of the reported single-run estimates is not possible even with those artefacts; what is reproducible is the protocol and, up to run-to-run variation, the findings.

10 Conclusion

This paper examined whether strong facial-emotion recognition performance provides an adequate security basis for emotion-driven HRI. It does not. A deep residual CNN achieves 0.760 accuracy (95% CI [0.749, 0.771]) and 0.753 macro-F1 at real-time CPU throughput, yet all trained models collapse under strong gradient-based attack irrespective of architecture, and the resulting errors remain highly confident: 96.0% of attacked predictions exceed a 0.70 operational threshold even after temperature scaling.

We showed that this failure is structural rather than incidental. Because the maximum softmax probability is strictly increasing in the logit-gap vector, and because temperature scaling acts identically on every input, no global temperature can admit legitimate predictions while rejecting adversarial ones once the adversarial gap dominates the clean gap (Theorem 1); under an explicit homogeneity assumption the detector’s ROC curve is invariant to temperature entirely (Corollary 1). The measured admission rates further bound the gate’s Youden index at the operating threshold above by 0.040, regardless of the clean admission rate (Proposition 3).

Classifier confidence should therefore be treated as one piece of evidence rather than as proof that a perception output is safe to execute. Secure emotion-aware robotics requires an explicit verification stage between FER and robot action, with criteria that extend beyond instantaneous recognition confidence.

Reproducibility Statement

Models are implemented in TensorFlow/Keras with OpenCV, NumPy, Pandas and scikit-learn; all training and evaluation ran on a CPU-only Ubuntu workstation with 16 GB RAM. The dataset is the public five-class FER2013 subset of Table 3 (24,176 training, 6,043 test). Preprocessing is resize (VGG16 strategies only) plus division by 255; augmentation parameters are given in Section 4.1. Attacks are untargeted FGSM and 10-iteration PGD with $\alpha = 0.005$, single start at the clean input, no restarts, ϵ on the $[0, 1]$ scale. Temperature scaling uses a 1,000-image calibration split drawn from the training partition; the fitted temperature is frozen and evaluated on an 800-image stratified test subset.

The dataset and class set, the partition discipline, the architectures and their parameter counts, the optimiser settings, the callback and augmentation regimes, the regularisation, the attack budgets and step sizes, the corruption constants, the operating threshold and the metric definitions are all stated in the text. The exact partition indexing, library version strings and CPU model reside in the released configuration. Training was not globally seeded, so the single-run estimates are not bit-reproducible; the protocol is.

Data Availability Statement

The datasets used in this study consist of the publicly available FER2013 benchmark dataset, accessed via Kaggle (<https://www.kaggle.com/datasets/msambare/fer2013>). This secondary dataset was used for computational analysis without collection of any new data. All experimental code, model configurations, and hyperparameter settings are described fully in Section 4.2 and Section 4.1. The training, attack, calibration and evaluation scripts, together with the configuration files recording the training hyperparameters, the corruption constants of Section 4.4, and the evaluation subset definitions of Table 4, are available from the corresponding author on reasonable request and will be deposited in a public repository upon acceptance of this manuscript.

Ethics Approval

Not Applicable. This study does not involve human participants. The work is a secondary computational analysis of the publicly available FER2013 benchmark dataset, originally assembled by automated web image search for the ICML 2013 Representation Learning Workshop. No attempt was made to identify individuals in the images. Under the institution’s applicable guidance for secondary analysis of public benchmark data, formal ethics review was not required.

The authors note an important construct-validity concern independent of data provenance. FER2013 assigns discrete emotion categories inferred from facial configuration, yet the assumption that emotional state is reliably recoverable from facial appearance alone is contested in the emotion science literature [12]. The findings in this paper concern model and system behaviour under adversarial attack and hold regardless of that construct validity; they should not be construed as endorsing deployment of emotion inference in consequential real-world settings.

Consent to Participate

Not Applicable. No human participation, surveys, questionnaires, or feedback from human subjects were involved in this study.

Consent for Publication

Not Applicable. This study does not include photographs or identifying information of individuals.

Declaration of Competing Interest

The authors declare no competing interests.

Funding

This work received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

CRedit Author Contributions

Olayide Tolulope Ezekiel: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft, Writing – review and editing.

Franklin Tchakounté: Conceptualization, Methodology, Supervision, Validation, Formal analysis, Investigation, Writing – review and editing, Project administration.

Derrick Méthode BAGAZA: Formal analysis, Validation, Investigation, Writing – review and editing.

Shanay Rab: Conceptualization, Resources, Validation, Formal analysis, Writing – review and editing.

Almas Baimagambetov: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing – review and editing, Supervision.

References

- [1] N. Rawal and R. Stock-Homburg, “Facial emotion expressions in human–robot interaction: A survey,” *International Journal of Social Robotics*, vol. 14, no. 7, pp. 1583–1604, 2021.

- [2] M. Spezialetti, G. Placidi, and S. Rossi, “Emotion recognition for human-robot interaction: Recent advances and future perspectives,” *Frontiers in Robotics and AI*, vol. 7, p. 532279, 2020.
- [3] L. T. Ottoni and J. Cerqueira, “A systematic review of human-robot interaction: The use of emotions and the evaluation of their performance,” *International Journal of Social Robotics*, vol. 16, no. 11, pp. 2169–2188, 2024.
- [4] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in *International Conference on Learning Representations (ICLR)*, 2015.
- [5] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” in *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, Vancouver, Canada, 2018. [Online]. Available: <https://arxiv.org/abs/1706.06083>
- [6] A. Bajaj and D. K. Vishwakarma, “A state-of-the-art review on adversarial machine learning in image classification,” *Multimedia Tools and Applications*, vol. 83, no. 3, pp. 9351–9416, 2024.
- [7] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, ser. PMLR, vol. 70, 2017, pp. 1321–1330.
- [8] D. Hendrycks and K. Gimpel, “A baseline for detecting misclassified and out-of-distribution examples in neural networks,” in *5th International Conference on Learning Representations (ICLR)*, Toulon, France, 2017.
- [9] M. Wu, W. Su, L. Chen, Z. Liu, W. Cao, and K. Hirota, “Weight-adapted convolution neural network for facial expression recognition in human-robot interaction,” *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 3, pp. 1473–1484, 2021.
- [10] M. D. Putro, D. Nguyen, and K. Jo, “A fast cpu real-time facial expression detector using sequential attention network for human-robot interaction,” *IEEE Transactions on Industrial Informatics*, vol. 18, no. 11, pp. 7665–7674, 2022.
- [11] H. Gunes, F. Broz, C. S. Crawford, A. R. V. der Pütten, M. Strait, and L. Riek, “Reproducibility in human-robot interaction: Furthering the science of hri,” *Current Robotics Reports*, vol. 3, no. 4, pp. 281–292, 2022.
- [12] L. F. Barrett, *How Emotions Are Made: The Secret Life of the Brain*. Houghton Mifflin Harcourt, 2017.
- [13] Y. Sun, C. Wu, K. Zheng, and X. Niu, “Adv-emotion: The facial expression adversarial attack,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 35, no. 11, p. 2150019, 2021.
- [14] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. El Ghaoui, and M. I. Jordan, “Theoretically principled trade-off between robustness and accuracy,” in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, ser. PMLR, vol. 97, 2019, pp. 7472–7482.
- [15] J. M. Cohen, E. Rosenfeld, and J. Z. Kolter, “Certified adversarial robustness via randomized smoothing,” in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, ser. PMLR, vol. 97, 2019, pp. 1310–1320.

- [16] S. Liang, Y. Li, and R. Srikant, “Enhancing the reliability of out-of-distribution image detection in neural networks,” in *6th International Conference on Learning Representations (ICLR)*, 2018.
- [17] K. Lee, K. Lee, H. Lee, and J. Shin, “A simple unified framework for detecting out-of-distribution samples and adversarial attacks,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31. Curran Associates, 2018.
- [18] W. Liu, X. Wang, J. D. Owens, and Y. Li, “Energy-based out-of-distribution detection,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33. Curran Associates, 2020.
- [19] J.-P. A. Yaacoub, H. N. Noura, O. Salman, and A. Chehab, “Robotics cyber security: Vulnerabilities, attacks, countermeasures, and recommendations,” *International Journal of Information Security*, vol. 21, no. 1, pp. 115–158, 2022.
- [20] R. Tian, L. Sun, A. Bajcsy, M. Tomizuka, and A. D. Dragan, “Safety assurances for human-robot interaction via confidence-aware game-theoretic human models,” in *IEEE International Conference on Robotics and Automation (ICRA)*, 2022, arXiv:2109.14700.
- [21] J. Kim, W. Chen, D. Soleymanzadeh, Y. Ding, X. Gao, Z. Tu, R. Zhang, F. Fei, S. Veer, Y. Lyu, M. Zheng, and Y. Gu, “Modular safety guardrails are necessary for foundation-model-enabled robots in the real world,” arXiv preprint arXiv:2602.04056, 2026.
- [22] H. M. Shahzad, S. M. Bhatti, A. Jaffar, S. Akram, M. Alhajlah, and A. Mahmood, “Hybrid facial emotion recognition using cnn-based features,” *Applied Sciences*, vol. 13, no. 9, p. 5572, 2023.
- [23] D. Sahana, K. S. Varsha, S. Sen, and R. Priyanka, “A cnn-based approach for facial emotion detection,” in *Soft Computing: Theories and Applications: Proceedings of SoCTA 2022*. Springer Nature Singapore, 2023, pp. 1–10.
- [24] M. Sajjad *et al.*, “A comprehensive survey on deep facial expression recognition: challenges, applications, and future guidelines,” *Alexandria Engineering Journal*, vol. 68, pp. 817–840, 2023.
- [25] M. A. Ozdemir, B. Elagoz, A. Alaybeyoglu, R. Sadighzadeh, and A. Akan, “Real time emotion recognition from facial expressions using cnn architecture,” in *2019 Medical Technologies Congress (TIPTEKNO)*. IEEE, October 2019, pp. 1–4.
- [26] M. A. H. Akhand, S. Roy, N. Siddique, M. A. S. Kamal, and T. Shimamura, “Facial emotion recognition using transfer learning in the deep cnn,” *Electronics*, vol. 10, no. 9, p. 1036, 2021.
- [27] S. A. Hussain and A. S. A. Al Balushi, “A real-time face emotion classification and recognition using deep learning model,” in *Journal of Physics: Conference Series*, vol. 1432, no. 1. IOP Publishing, 2020, p. 012087.
- [28] N. Mehendale, “Facial emotion recognition using convolutional neural networks (ferc),” *SN Applied Sciences*, vol. 2, no. 3, p. 446, 2020.
- [29] I. Ali and F. Ghaffar, “Robust cnn for facial emotion recognition and real-time gui,” *AIMS Electronics and Electrical Engineering*, vol. 8, no. 2, 2024.
- [30] “Revolutionizing facial emotion recognition: in-depth analysis of cutting-edge models, methodologies, and datasets,” *Discover Artificial Intelligence*, vol. 1, pp. 553–w, 2025.

- [31] “Intelligent automated fault detection framework for single phase motors using real time monitoring and machine learning,” *Discover Artificial Intelligence*, vol. 1, pp. 509–0, 2025.
- [32] Y. Geifman and R. El-Yaniv, “Selective classification for deep neural networks,” in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017, pp. 4878–4887.
- [33] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee *et al.*, “Challenges in representation learning: A report on three machine learning contests,” in *International Conference on Neural Information Processing (ICONIP)*. Springer, 2013, pp. 117–124.