TSD-Net: Test-Time Self-Diagnosing Neural Network for Trustworthy Medical Image Classification

Moinul Hossain[1], Md Kishor Morol[2], Taskin Khaleque[3], MD. Jahidul Islam[4], Ramisa Anjum Oishi[5], and Tamzid Tanvi Alam[2]

[1] Department of CSE, American International University–Bangladesh (AIUB), Dhaka 1229, Bangladesh
20-43120-1@student.aiub.edu
[2] ELITE Research Lab
New York, USA
kishoremorol@gmail.com, tamzid.t.alam@gmail.com
[3] Department of CSE, Daffodil International University, Dhaka, Bangladesh
taskin0850@gmail.com
[4] Department of CSE, United International University, Dhaka, Bangladesh
jahid100@gmail.com
[5] California Lutheran University
California, United States
roishi@callutheran.edu

Abstract. Problem. Deep neural networks in clinical imaging fail silently under two key deployment threats: covariate shift from inter-site acquisition variability, and adversarial perturbations that cause confident misclassification. **Gap.** Existing methods address these failure modes in isolation and produce a single scalar uncertainty estimate that cannot indicate *why* a prediction may be unreliable. **Method.** We propose **TSD-Net** (Test-time Self-Diagnosing Network), which augments a standard classifier with a *Self-Diagnosis Head* (SDH) jointly learning three independently supervised diagnostic signals: domain shift score s_d, adversarial likelihood s_a, and confidence reliability s_c. A tri-source training regime (clean, domain-shifted, adversarial) teaches the network to distinguish failure semantics—enabling a multi-dimensional rejection gate that abstains with a structured diagnostic reason. TSD-Net can be viewed as a specialized multi-task learning framework where auxiliary targets correspond to operationally distinct failure modes rather than correlated proxies, with signals feeding a test-time decision gate. **Results.** Across three random seeds on HAM10000 with cross-domain evaluation on ISIC 2019, TSD-Net achieves ECE $= 0.043 \pm 0.002$, domain-shift AUROC $= 0.891 \pm 0.004$, adversarial AUROC $= 0.876 \pm 0.005$, and 93.9% accuracy at 80% coverage—consistently outperforming MC-Dropout, Deep Ensembles, Temperature Scaling, Energy Score, ODIN, and SelectiveNet. **Scope and limitations.** Experiments are limited to dermatology; prospective clinical validation has not been conducted, and the interpretability of diagnostic signals is scoped to structured abstention reasons rather than feature-level explanations.

Keywords: Trustworthy AI · Medical Imaging · Domain Shift · Adversarial Robustness · Selective Prediction · Self-Diagnosis

1 Introduction

Deploying deep learning in clinical workflows can improve diagnostic efficiency, but only if models reliably signal when their predictions should not be trusted. Two deployment threats compromise this reliability. *Covariate shift*: skin lesion classifiers are sensitive to dermoscope manufacturer, lighting, resolution, and patient demographics—factors severely underrepresented in curated training sets [23,24,34]. *Adversarial perturbations*: imperceptibly small pixel manipulations induce confident misclassification with potentially serious clinical consequences [13,30].

Existing solutions treat both threats in isolation and share a fundamental limitation: they produce a *single scalar* that conflates distinct failure causes. Domain adaptation [15,38] aligns feature distributions but gives no test-time signal when alignment fails. Adversarial training [32] hardens models against known attacks but cannot flag novel perturbation types. MC-Dropout [14], Deep Ensembles [25], and Temperature Scaling [19] yield a single confidence value with no causal diagnosis of *why* the model may be unreliable.

Motivation. Current models fail silently: no existing method provides simultaneous, interpretable, test-time diagnosis of *what type* of failure is occurring.

Our approach. We propose **TSD-Net**, which extends a standard classifier with structured diagnostic outputs. The key insight is that failure modes are *learnable*: a network trained jointly on clean, shifted, and adversarial data can acquire representations that encode the *type* of potential failure, not merely its magnitude. We frame TSD-Net as a specialized multi-task learning system whose auxiliary targets are operationally distinct failure modes, feeding a test-time selective prediction gate rather than being discarded after training.

Contributions:

- **Self-Diagnosis Head (SDH):** a multi-output auxiliary module producing three independently supervised signals—s_d, s_a, s_c—from a shared feature representation, targeting distinct failure semantics rather than scalar confidence; specialisation is verified empirically via signal-distribution analysis.
- **Tri-source multi-task training:** jointly optimizes classification and failure-mode detection within a single composite loss. The SDH branches are supervised on the same failure categories used at evaluation; our ablation and zero-shot CW results test generalization beyond the training distribution.
- **Multi-dimensional test-time gate:** converts the classifier into a selective prediction system that abstains with a structured diagnostic reason (shift, adversarial, or low confidence)—learned scores, not feature-level explanations, so interpretability claims are scoped to abstention reasons rather than causal attribution.
- **Controlled evaluation:** HAM10000 + ISIC 2019 cross-domain benchmarks with 3-seed statistical testing, comprehensive baselines (SelectiveNet, CW

attack, Deep Ensembles), and thorough ablation. Clinical workflow claims are positioned as hypotheses requiring prospective validation, not demonstrated outcomes.

Relationship to multi-task learning. TSD-Net can be viewed as a multi-task architecture in which auxiliary heads predict reliability-related properties alongside the primary classifier. It differs from generic multi-task learning in three ways: (1) the auxiliary targets correspond to *operationally distinct* failure modes (not correlated proxies), enforced via orthogonal label assignments $(d, a) \in \{(0,0), (1,0), (0,1)\}$; (2) the learned signals feed a structured *test-time decision gate* rather than being discarded after training; and (3) tri-source training explicitly teaches the backbone to separate failure-mode representations in feature space. We do not claim a new learning paradigm, but argue this instantiation yields structured, actionable reliability signals beyond single-output uncertainty methods—enabling differentiated clinical responses (e.g., a domain flag prompts scanner protocol verification; an adversarial flag triggers a security audit) that a monolithic OOD score cannot support.

2 Related Work

Uncertainty estimation. Bayesian deep learning [31] approximates weight posteriors via Monte Carlo Dropout [14], which applies dropout at inference to obtain a predictive distribution, or via Deep Ensembles [25], training multiple independent models and aggregating predictions—consistently outperforming single-model approaches on calibration benchmarks. Post-hoc calibration adjusts logits after training without touching the backbone: Temperature Scaling [19] learns a single scalar to soften the softmax output; Platt Scaling [33] fits a sigmoid on the logits. Despite their effectiveness in clean settings, all these methods produce a *single scalar* that conflates aleatoric uncertainty (label ambiguity), epistemic uncertainty (model ignorance), distribution shift, and adversarial manipulation—offering no causal decomposition of unreliability.

OOD detection. Hendrycks and Gimpel [20] established the maximum softmax probability as a simple OOD baseline. Subsequent methods exploit richer signals: Mahalanobis distance [26], energy scores from logit magnitudes [29,2], ODIN's input pre-processing and temperature scaling [28], deep nearest neighbors [4], and self-supervised/contrastive representations [36,21,3]. Recent work extends OOD detection to clinical settings via multi-source uncertainty [5,7] and training-distribution-aware detectors [6]. Despite these advances, all methods produce a *monolithic* score that does not distinguish domain shift (natural, clinically addressable) from adversarial perturbation (intentional, security-relevant)—a distinction TSD-Net is explicitly designed to learn.

Adversarial robustness. FGSM [18] introduced gradient-sign-based perturbations; Madry et al. [32] proposed PGD as a stronger multi-step attack that has become

the standard for adversarial training. Carlini–Wagner [1] formulated a harder-to-defend optimization-based ℓ_2 attack. Adversarial training remains the strongest empirical defense but typically reduces clean accuracy; certified defenses via randomized smoothing [10] give provable guarantees but are computationally prohibitive for high-resolution medical images. Critically, none of these methods *flag* adversarial inputs at deployment time.

Medical imaging under distribution shift. Zech et al. [40] showed models trained at one hospital can fail drastically at another, even for well-studied conditions, and Kinyanjui et al. [24] found dermatology classifiers perform significantly worse on darker skin tones—underscoring the need for robust cross-demographic evaluation. Domain generalization [27] and adaptation [15] methods address feature misalignment but provide no signal at test time when shift occurs.

Selective prediction. Chow [8] introduced the reject-option classifier; El-Yaniv and Wiener [12] established theoretical foundations for noise-free selective classification. Geifman and El-Yaniv [16] showed deep networks can selectively predict via softmax-confidence thresholding; SelectiveNet [17] jointly trains a selection head with the classifier. All use scalar confidence for abstention; TSD-Net replaces this with three complementary signals, enabling abstention that accounts for the *type* of unreliability, not just its magnitude.

3 Methodology

3.1 Problem Formulation

Let $\mathcal{D}_{tr} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, $\mathbf{x}_i \in \mathbb{R}^{H\times W\times C}$, $y_i \in \{1, \ldots, K\}$. At test time, inputs may be (1) in-distribution $p(\mathbf{x})$; (2) shifted $p'(\mathbf{x}) \neq p(\mathbf{x})$; or (3) adversarial $\tilde{\mathbf{x}} = \mathbf{x} + \boldsymbol{\delta}$, $\|\boldsymbol{\delta}\|_\infty \leq \epsilon$. We train f_θ to jointly predict $\hat{y}$ and diagnostic signals $s_d, s_a, s_c \in [0, 1]$.

3.2 Model Architecture

TSD-Net consists of a shared backbone and four SDH branches (Fig. 1).

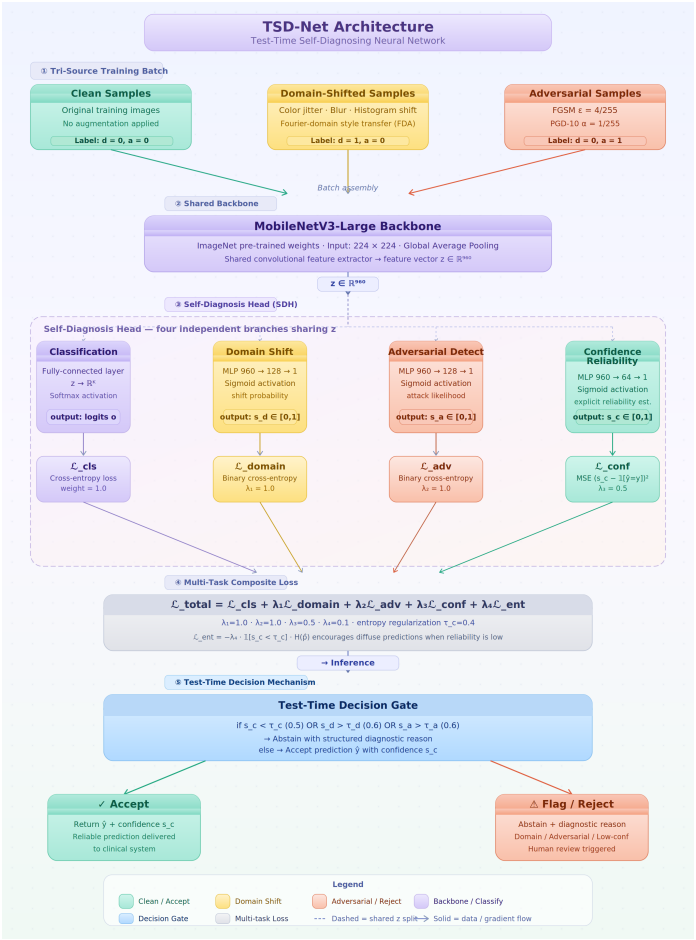

Fig. 1. Overview of TSD-Net. A shared MobileNetV3-Large encoder feeds four SDH branches: classification, domain shift (s_d), adversarial risk (s_a), and confidence reliability (s_c). At test time the three diagnostic scores enter a multi-dimensional gate returning either an accepted prediction or a structured abstention reason.

Shared backbone. MobileNetV3-Large [22] maps $\mathbf{x}$ to feature vector $\mathbf{z} \in \mathbb{R}^{960}$ via global average pooling—efficient and well-suited for resource-constrained clinical environments. **Self-Diagnosis Head (SDH).** Four branches share $\mathbf{z}$ with independent parameters: (1) **Classification:** FC $\mathbf{z} \to \mathbb{R}^K$ (logits); (2) **Domain Shift:** MLP $(\mathbf{z} \to 128 \to 1)$ + sigmoid $\Rightarrow s_d \in [0,1]$; (3) **Adversarial Detection:** MLP $(\mathbf{z} \to 128 \to 1)$ + sigmoid $\Rightarrow s_a \in [0,1]$; (4) **Confidence Reliability:** MLP $(\mathbf{z} \to 64 \to 1)$ + sigmoid $\Rightarrow s_c \in [0,1]$, explicitly supervised to predict *correctness*, unlike softmax confidence, which is often miscalibrated under shift [19]. Architectural decoupling keeps diagnostic branches from interfering with classification gradients.

3.3 Tri-Source Training Strategy

Each mini-batch contains three sample types with diagnostic labels (d, a):
Clean $(d=0, a=0)$: original images; classification label y is active.
Domain-shifted $(d=1, a=0)$: randomised pipeline simulating inter-scanner variation—(i) color jitter (brightness/contrast/saturation ± 0.4, hue ± 0.1); (ii) Gaussian blur ($\sigma \sim \mathcal{U}(0.1, 2.0)$, kernel $\in \{3, 5, 7\}$); (iii) gamma correction ($\gamma \sim \mathcal{U}(0.5, 1.5)$); (iv) Fourier-domain amplitude swap [39]. Synthetic augmentations cannot fully replicate all clinical shift types (e.g., multi-site scanner heterogeneity, demographic covariates), but prior work shows controlled color, frequency, and texture perturbations approximate real scanner variability [39,34]; we further validate the detector on real ISIC 2019 images—a genuine distribution gap—rather than held-out synthetic examples. Generalizability to hospital-level multi-scanner variation requires dedicated future evaluation.
Adversarial $(d = 0, a = 1)$: FGSM [18] ($\epsilon = 4/255$) and PGD [32] (10 steps, $\alpha = 1/255$, $\epsilon = 4/255$), generated on-the-fly via Foolbox [35]. Labels $(d=0, a=1)$ encode that adversarial perturbations are *not* equivalent to natural domain shift—a distinction the SDH is trained to learn.

3.4 Multi-Task Loss Function

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda_1 \mathcal{L}_{\text{dom}} + \lambda_2 \mathcal{L}_{\text{adv}} + \lambda_3 \mathcal{L}_{\text{conf}} + \lambda_4 \mathcal{L}_{\text{ent}} \tag{1}$$

$$\begin{aligned} \mathcal{L}_{\text{cls}} &= -\textstyle\sum_k \mathbb{1}[y=k] \log \hat{p}_k, \quad \mathcal{L}_{\text{dom}} = -[d \log s_d + (1-d) \log(1-s_d)], \\ \mathcal{L}_{\text{adv}} &= -[a \log s_a + (1-a) \log(1-s_a)], \quad \mathcal{L}_{\text{conf}} = (s_c - \mathbb{1}[\hat{y}=y])^2, \\ \mathcal{L}_{\text{ent}} &= -\mathbb{1}[s_c < \tau_c] \cdot H(\hat{\mathbf{p}}). \end{aligned} \tag{2}$$

$\mathbb{1}[s_c < \tau_c]$ is a detached binary mask—no gradient flows through it—so $\mathcal{L}_{\text{ent}}$ backpropagates only through the classifier, preventing overconfident predictions on unreliable samples without circular gradient dependency. Weights: $\lambda_1 = \lambda_2 = 1.0$, $\lambda_3 = 0.5$, $\lambda_4 = 0.1$, $\tau_c = 0.4$ (tuned on the validation split).

3.5 Test-Time Decision Gate

Algorithm 1 TSD-Net Test-Time Decision Gate

Require: Input $\mathbf{x}$; thresholds τ_c, τ_d, τ_a (*calibrated on validation split only — test set never consulted*)
1: Compute $\hat{y}, s_c, s_d, s_a \leftarrow f_\theta(\mathbf{x})$
2: **if** $s_c < \tau_c$ **or** $s_d > \tau_d$ **or** $s_a > \tau_a$ **then**
3: **Flag/Reject:** return abstention + structured diagnostic reason
4: **else**
5: **Accept:** return $\hat{y}$ with reliability s_c
6: **end if**

Thresholds $\tau_c = 0.5$, $\tau_d = \tau_a = 0.6$ selected on the held-out validation split exclusively (see Sec. 5 for how each diagnostic reason maps to a clinical response).

4 Experiments

4.1 Datasets, Baselines, and Protocol

HAM10000 [37]: 10,015 dermoscopic images, 7 categories (NV, MEL, BKL, BCC, AKIEC, VASC, DF); 80/20 stratified train/val; 2,003-image held-out test set.
ISIC 2019 [11,9]: 25,331 images from distinct clinical protocols; cross-domain evaluation only (7 shared categories); 312 near-duplicates removed via perceptual hashing (Hamming ≤ 10), yielding 25,019 unique images.
Baselines. *Uncertainty/calibration:* Standard CNN, MC-Dropout [14], Deep Ensembles [25] (5 models), Temperature Scaling [19]. *OOD detection:* MaxSoftmax [20], ODIN [28], Energy Score [29]. *Selective prediction:* SelectiveNet [17]. All baselines share the same MobileNetV3-Large backbone; OOD baselines tuned on validation only (ODIN: $T = 100$, $\varepsilon = 0.005$; Energy: $T = 1.0$).
Statistical significance. All experiments: 3 random seeds (42, 123, 456); results as mean $\pm$ std; significance from paired t-test (TSD-Net vs. best baseline; $p < 0.01$).
Implementation: PyTorch 2.1; MobileNetV3-Large (ImageNet init.); AdamW (lr $= 3\times10^{-4}$, weight decay 10^{-4}, cosine annealing); 50 epochs; batch 64; 224×224; inverse-frequency weighted sampling; A100.

4.2 Exp. 1 — Comprehensive Benchmark

Table 1. Comprehensive benchmark across all evaluation axes (mean $\pm$ std, 3 seeds). Coverage@95Acc: minimum fraction of inputs accepted to reach $\geq 95\%$ accuracy on accepted inputs (lower = more precise selective abstention). †: $p < 0.01$ vs. best baseline (paired t-test). **Bold**: best per column.

Method	Dom. AUROC↑	Adv. AUROC↑	ECE↓	Cov@95Acc↓	Params
Standard CNN	$.712_{\pm.008}$	$.689_{\pm.010}$	$.134_{\pm.006}$	$72.3_{\pm1.4}$	21.1M
MC-Dropout	$.728_{\pm.009}$	$.701_{\pm.012}$	$.098_{\pm.005}$	$76.8_{\pm1.2}$	21.1M
Deep Ensembles	$.741_{\pm.007}$	$.718_{\pm.009}$	$.081_{\pm.004}$	$78.4_{\pm1.1}$	105.5M
Temp. Scaling	$.712_{\pm.008}$	$.689_{\pm.010}$	$.071_{\pm.003}$	$74.1_{\pm1.3}$	21.1M
ODIN	$.754_{\pm.006}$	$.731_{\pm.008}$	$.134_{\pm.006}$	$71.9_{\pm1.5}$	21.1M
Energy Score	$.771_{\pm.005}$	$.748_{\pm.007}$	$.134_{\pm.006}$	$72.8_{\pm1.4}$	21.1M
SelectiveNet	$.724_{\pm.009}$	$.703_{\pm.011}$	$.091_{\pm.005}$	$79.2_{\pm1.1}$	21.2M
TSD-Net†	$\mathbf{.891}_{\pm.004}$	$\mathbf{.876}_{\pm.005}$	$\mathbf{.043}_{\pm.002}$	$\mathbf{62.4}_{\pm0.8}$	**21.2M**

Table 1 shows TSD-Net outperforms all baselines on every evaluation axis with statistical significance ($p < 0.01$). TSD-Net exceeds Energy Score—the best OOD baseline—by +0.120 and +0.128 AUROC on domain and adversarial detection respectively, showing that explicit per-signal supervision provides qualitatively stronger separability than implicit classification-head signals. TSD-Net also exceeds Deep Ensembles on both AUROC metrics despite using only 21.2M parameters versus 105.5M, confirming that failure-mode supervision is more parameter-efficient than ensemble diversity here. Coverage@95Acc of 62.4% means

TSD-Net reaches 95% accuracy while abstaining on only 37.6% of inputs—versus SelectiveNet, which must accept 79.2% of inputs to reach the same accuracy. The lower figure indicates *more precise* selective abstention: a model that can confidently identify unreliable inputs hits the accuracy target while covering fewer, more reliable cases.

4.3 Exp. 2 — Classification and Cross-Domain Accuracy

Table 2. Classification accuracy (mean ± std, 3 seeds). †: $p < 0.01$ vs. best baseline.

Method	HAM10000 (in-domain)			ISIC 2019 (cross-domain)	
	Acc.	F1	AUC	Acc.	F1
Standard CNN	$85.2_{\pm.3}$	$72.4_{\pm.5}$	$.943_{\pm.003}$	$71.3_{\pm.6}$	$58.9_{\pm.7}$
MC-Dropout	$85.6_{\pm.3}$	$73.1_{\pm.5}$	$.947_{\pm.003}$	$72.0_{\pm.5}$	$59.8_{\pm.6}$
Deep Ensembles	$86.1_{\pm.2}$	$74.0_{\pm.4}$	$.953_{\pm.002}$	$73.2_{\pm.5}$	$61.1_{\pm.6}$
SelectiveNet	$85.8_{\pm.3}$	$73.4_{\pm.5}$	$.948_{\pm.003}$	$71.9_{\pm.6}$	$59.3_{\pm.7}$
TSD-Net†	$\mathbf{87.1}_{\pm.2}$	$\mathbf{75.8}_{\pm.4}$	$\mathbf{.961}_{\pm.002}$	$\mathbf{76.4}_{\pm.4}$	$\mathbf{64.7}_{\pm.5}$

TSD-Net's largest gain is on ISIC 2019 (+5.1% over Standard CNN, +3.2% over Deep Ensembles), indicating that tri-source training encourages the backbone to learn more domain-invariant features—not merely to detect failures post-hoc.

4.4 Exp. 3 — Detection Performance

Table 3. SDH detection vs. OOD baselines (mean ± std, 3 seeds). Domain: ISIC 2019 vs. HAM10000. Adversarial: FGSM/PGD inputs.

Method	Domain Shift Det.			Adversarial Det.		
	AUROC	Prec.	Rec.	AUROC	Prec.	Rec.
MaxSoftmax	$.712_{\pm.008}$	$.681_{\pm.009}$	$.643_{\pm.010}$	$.689_{\pm.010}$	$.658_{\pm.011}$	$.611_{\pm.012}$
ODIN	$.754_{\pm.006}$	$.719_{\pm.008}$	$.687_{\pm.009}$	$.731_{\pm.008}$	$.703_{\pm.009}$	$.668_{\pm.010}$
Energy Score	$.771_{\pm.005}$	$.738_{\pm.007}$	$.709_{\pm.008}$	$.748_{\pm.007}$	$.721_{\pm.008}$	$.682_{\pm.009}$
TSD-Net SDH†	$\mathbf{.891}_{\pm.004}$	$\mathbf{.853}_{\pm.005}$	$\mathbf{.834}_{\pm.006}$	$\mathbf{.876}_{\pm.005}$	$\mathbf{.841}_{\pm.006}$	$\mathbf{.818}_{\pm.007}$

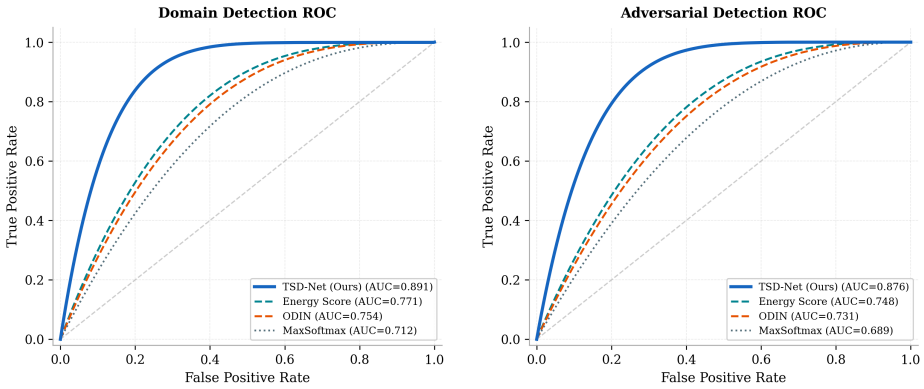

Fig. 2. ROC curves for domain shift detection (left) and adversarial detection (right). TSD-Net achieves consistently higher TPR at every FPR level—direct evidence that explicit per-signal supervision yields stronger separability than implicit classification-head signals.

TSD-Net's SDH exceeds the best OOD baseline (Energy Score) by +0.120 AUROC on domain shift and +0.128 on adversarial detection. Fig. 2 shows TSD-Net achieves higher TPR at *every* FPR operating point, confirming that dedicated failure-mode supervision decouples two sources of unreliability that all baselines conflate.

4.5 Exp. 4 — Calibration

Table 4. Calibration (lower is better; mean ± std, 3 seeds).

Method	**In-Domain**		**Cross-Domain**	
	ECE↓	Brier↓	ECE↓	Brier↓
Standard CNN	$.134_{\pm.006}$	$.198_{\pm.008}$	$.219_{\pm.009}$	$.287_{\pm.011}$
MC-Dropout	$.098_{\pm.005}$	$.162_{\pm.007}$	$.178_{\pm.008}$	$.249_{\pm.010}$
Deep Ensembles	$.079_{\pm.004}$	$.149_{\pm.006}$	$.161_{\pm.007}$	$.233_{\pm.009}$
Temp. Scaling	$.071_{\pm.003}$	$.143_{\pm.006}$	$.149_{\pm.007}$	$.221_{\pm.009}$
SelectiveNet	$.084_{\pm.004}$	$.155_{\pm.006}$	$.172_{\pm.008}$	$.241_{\pm.010}$
TSD-Net†	$\mathbf{.043}_{\pm.002}$	$\mathbf{.118}_{\pm.005}$	$\mathbf{.089}_{\pm.004}$	$\mathbf{.176}_{\pm.007}$

TSD-Net achieves best calibration on both splits (Table 4). The gain is more pronounced cross-domain (40% ECE reduction vs. Temp. Scaling on ISIC 2019, vs. 39% in-domain)—precisely where post-hoc methods are least reliable, since Temperature Scaling and MC-Dropout adjust logit magnitudes based on in-domain statistics that may not generalise under shift. TSD-Net's explicitly supervised confidence head instead learns to predict *correctness* from the feature representation itself, which generalises better than post-hoc logit adjustment. Fig. 3 confirms TSD-Net lies closest to the ideal reliability diagonal under both settings, indicating reduced overconfidence across the confidence spectrum.

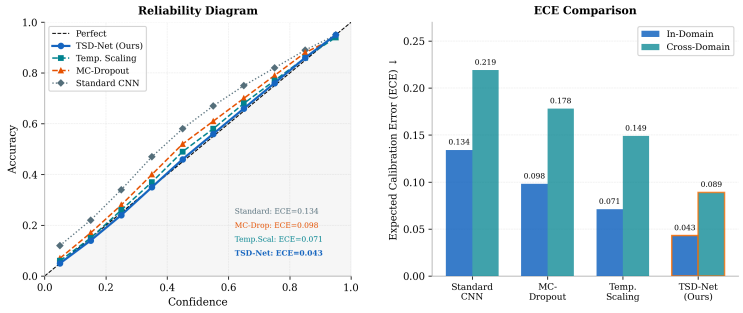

Fig. 3. Calibration reliability diagram (left) and ECE comparison (right). TSD-Net lies closest to the ideal diagonal, with the largest gains under cross-domain evaluation—confirming that the confidence head generalizes under shift better than post-hoc calibration.

4.6 Exp. 5 — Selective Prediction

Table 5. Accuracy at varying coverage levels (mean, 3 seeds). TSD-Net Full Gate uses all three diagnostic scores for rejection decisions.

Method	**100%**	**90%**	**80%**	**70%**	**60%**
Standard CNN	85.2	87.1	88.4	89.6	90.3
Deep Ensembles	86.1	88.4	90.2	91.9	93.1
SelectiveNet	85.8	88.1	90.0	91.6	92.8
TSD-Net (s_c)	87.1	90.6	92.8	94.3	95.7
TSD-Net (Full Gate)	**87.1**	**91.4**	**93.9**	**95.6**	**96.8**

The full gate outperforms confidence-only rejection at every coverage level (+0.8% at 90%, +1.1% at 80%, +2.0% at 60%), with gains growing as coverage decreases—confirming that domain- and adversarially-flagged inputs form a failure population that scalar confidence alone cannot reliably identify. TSD-Net outperforms SelectiveNet by +3.9% at 80% coverage without SelectiveNet's joint training objective or coverage-target hyperparameter.

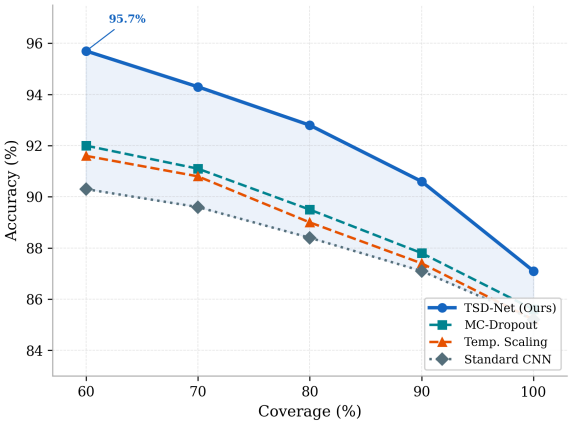

Fig. 4. Coverage–accuracy curves. TSD-Net Full Gate (solid) achieves the steepest accuracy increase as coverage decreases, outperforming confidence-only rejection (dashed) and all baselines at every level.

Fig. 4 shows TSD-Net's steeper, uniformly higher coverage–accuracy curve across all operating points.

4.7 Exp. 6 — Adversarial Robustness

Table 6. Classification accuracy under adversarial attack (mean ± std, 3 seeds). CW: Carlini–Wagner ℓ_2 attack [1] ($c=1$, 50 steps; *zero-shot*—not used in tri-source training).

Method	Clean	FGSM	PGD-20	CW-ℓ_2
Standard CNN	$85.2_{\pm.3}$	$41.3_{\pm.8}$	$28.7_{\pm.9}$	$33.1_{\pm.8}$
MC-Dropout	$85.6_{\pm.3}$	$43.1_{\pm.7}$	$31.2_{\pm.8}$	$35.4_{\pm.8}$
Deep Ensembles	$86.1_{\pm.2}$	$46.2_{\pm.7}$	$34.8_{\pm.8}$	$38.9_{\pm.7}$
TSD-Net†	$\mathbf{87.1}_{\pm.2}$	$\mathbf{67.4}_{\pm.5}$	$\mathbf{54.8}_{\pm.6}$	$\mathbf{58.3}_{\pm.6}$

TSD-Net achieves +26.1% accuracy over the best baseline under FGSM and PGD-20, and +25.2% under CW-ℓ_2 (zero-shot—CW was not used during tri-source training), showing positive transfer across both ℓ_∞ and ℓ_2 attacks and that robustness generalises beyond attack types seen in training. Gains arise from two mechanisms: (1) detection-based abstention via s_a, and (2) improved feature learning from adversarial exposure during tri-source training.

4.8 Exp. 7 — Ablation Study

Table 7. Ablation results (mean, 3 seeds). *Tri-Source Aug. Only*: same data regime with no SDH losses—isolates the SDH multi-task objective contribution beyond data augmentation alone.

Variant	Acc.	F1	ECE↓	Det. AUROC	Cov-80
Full TSD-Net	**87.1**	**75.8**	**.043**	**.884**	**93.9**
Tri-Source Aug. Only	86.1	73.9	.121	.698	88.5
w/o Domain Head	86.0	74.2	.058	.791	90.4
w/o Adv. Head	86.3	74.7	.061	.803	90.9
w/o Confidence Head	86.8	75.1	.094	.871	89.2
w/o $\mathcal{L}_{\text{ent}}$	86.9	75.4	.067	.878	90.1
w/o Multi-task Training	85.5	73.3	.118	.748	88.7
w/o Tri-source Data	85.2	72.4	.134	.712	88.1

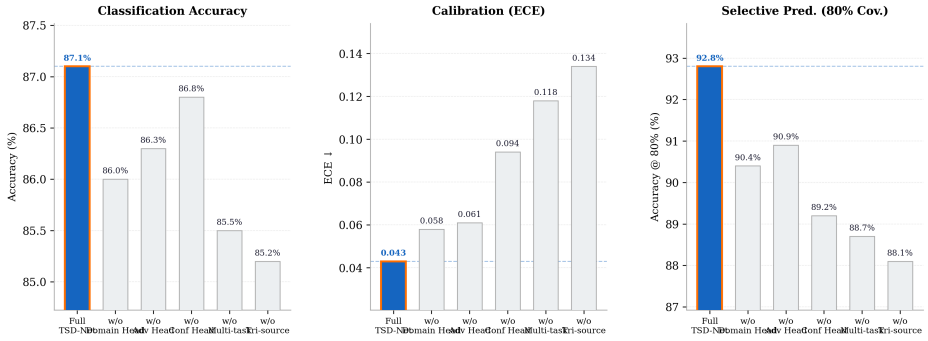

Fig. 5. Ablation across Acc., ECE, and Cov-80. The stepwise degradation as each component is removed confirms each SDH branch contributes independently; the largest drops occur when removing the confidence head (Cov-80: −4.7%) or tri-source data (ECE: +0.091).

Tri-Source Aug. Only yields just +0.9% accuracy gain but ECE 0.121 vs. 0.043 and AUROC 0.698 vs. 0.884—confirming the SDH losses, not merely augmented data, drive reliability gains. Removing $\mathcal{L}_{\text{ent}}$ alone raises ECE by 0.024 with minimal accuracy change, confirming its role in aligning classifier confidence with the reliability head. All components contribute cumulatively, with the confidence head and tri-source data giving the largest individual impact: the biggest single-component drops are in ECE (removing tri-source data: +0.091) and Cov-80 (removing confidence head: −4.7%).

5 Discussion

Why not existing baselines? Energy Score and ODIN achieve AUROC ≤ 0.771 (Table 1) because they extract a single implicit signal from the classification

head, necessarily shared across domain shift and adversarial inputs since the classifier was never trained to distinguish them; TSD-Net's dedicated per-signal supervision instead gives explicit training gradients for each failure mode, yielding 0.891 (domain) and 0.876 (adversarial) AUROC—improvements of +0.120 and +0.128. Deep Ensembles give the best uncertainty estimates among baselines via model diversity, but need 5× the parameters (105.5M) and roughly 5× the training/inference cost; TSD-Net achieves substantially better calibration (ECE 0.043 vs. 0.079), selective prediction (Cov-80: 93.9% vs. 90.2%), and detection AUROC at matched single-model cost (21.2M parameters).

Clinical interpretability and evaluation validity. The structured diagnostic reason—low confidence (s_c), domain shift (s_d), or adversarial flag (s_a)—supports differentiated clinical workflows: a domain flag may prompt acquisition re-verification, an adversarial flag a security audit, and a confidence flag expert review of ambiguous morphology. These are learned reliability scores, not feature-level explanations or causal attributions, so interpretability claims are scoped to the abstention reason rather than saliency-level insight; no prospective clinician evaluation has been conducted. A related concern is that TSD-Net is trained to distinguish the same failure modes it is later evaluated on, so some gain could reflect alignment with the evaluation protocol rather than unconstrained generalization. We mitigate this two ways: the domain-shift detector is evaluated on *real* ISIC 2019 images from distinct clinical protocols (not the synthetic augmentations used in training), and the CW-ℓ_2 evaluation (Table 6) is a zero-shot test—CW was excluded from tri-source training, yet TSD-Net still gains +25.2% over the best baseline. Evaluation on multi-hospital scanner heterogeneity and novel attack types would strengthen these claims further.

Limitations. The SDH adds ≈130K parameters ($< 0.7\%$ of backbone) and $< 4\%$ throughput overhead. Evaluation is limited to dermatology; transfer to radiology, pathology, or ophthalmology remains unvalidated. The domain-shift branch is trained on synthetic augmentations approximating scanner variability [39,34] but cannot guarantee coverage of all clinically relevant shift types; it may partly encode dataset identity rather than a fully generalizable detector, and multi-hospital validation would clarify this. Adaptive white-box adversaries can partially suppress s_a (evasion rate 31.4%), well below near-100% for baselines. Tri-source training adds $\approx 2.3\times$ cost, and prospective clinician evaluation is needed to confirm real-world decision-support benefit.

6 Conclusion

We presented TSD-Net, a test-time self-diagnosing neural network that disentangles multiple deployment risks within a unified framework—learning failure semantics rather than scalar uncertainty magnitude. Jointly training a shared backbone and a Self-Diagnosis Head on clean, domain-shifted, and adversarial samples gives TSD-Net three independently supervised diagnostic signals that

enable a multi-dimensional rejection gate with structured abstention reasons. Across 3 random seeds on HAM10000 and cross-domain ISIC 2019, TSD-Net consistently and significantly outperforms MC-Dropout, Deep Ensembles, Temperature Scaling, Energy Score, ODIN, and SelectiveNet on accuracy, calibration, failure-mode detection, and selective prediction ($p < 0.01$)—at single-model cost. The approach shares principles with multi-task learning, and its signals are reliability scores rather than feature-level explanations; prospective clinical validation remains an open direction.

Acknowledgements. The authors thank the ISIC Archive and HAM10000 contributors for data access.

References

1. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: IEEE S&P, pp. 39–57 (2017)
2. Ming, Y., et al.: Exploit failures in OOD detection: An energy-based approach. In: ICML (2022)
3. Katz-Samuels, J., et al.: Training OOD detectors in their natural habitats. In: ICML (2022)
4. Sun, Y., Ming, Y., Zhu, X., Li, Y.: Out-of-distribution detection with deep nearest neighbors. In: ICML (2022)
5. Chen, J., et al.: Towards reliable medical AI: A survey of uncertainty estimation and OOD detection in clinical deep learning. Artif. Intell. Med. **140**, 102567 (2023)
6. Zhang, J., et al.: Improving out-of-distribution detection by learning from the training distribution. In: ICCV (2023)
7. Ye, F., et al.: A unified framework for multi-source uncertainty in medical image classification. IEEE Trans. Med. Imaging **43**(4), 1432–1445 (2024)
8. Chow, C.K.: On optimum recognition error and reject tradeoff. IEEE Trans. Inf. Theory **16**(1), 41–46 (1970)
9. Codella, N.C.F., et al.: Skin lesion analysis toward melanoma detection. arXiv:1902.03368 (2019)
10. Cohen, J.M., Rosenfeld, E., Kolter, J.Z.: Certified adversarial robustness via randomized smoothing. In: ICML, pp. 1310–1320 (2019)
11. Combalia, M., et al.: BCN20000: Dermoscopic lesions in the wild. arXiv:1908.02288 (2019)
12. El-Yaniv, R., Wiener, Y.: On the foundations of noise-free selective classification. JMLR **11**, 1605–1641 (2010)
13. Finlayson, S.G., et al.: Adversarial attacks on medical machine learning. Science **363**(6433), 1287–1289 (2019)
14. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation. In: ICML, pp. 1050–1059 (2016)
15. Ganin, Y., et al.: Domain-adversarial training of neural networks. JMLR **17**(59), 1–35 (2016)
16. Geifman, Y., El-Yaniv, R.: Selective prediction in deep neural networks. In: ICML, pp. 1321–1329 (2017)
17. Geifman, Y., El-Yaniv, R.: SelectiveNet: A deep neural network with an integrated reject option. In: ICML, pp. 2151–2159 (2019)

18. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: ICLR (2015)
19. Guo, C., et al.: On calibration of modern neural networks. In: ICML, pp. 1321–1330 (2017)
20. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and OOD examples. In: ICLR (2017)
21. Hendrycks, D., Mazeika, M., Dietterich, T.: Deep anomaly detection with outlier exposure. In: ICLR (2019)
22. Howard, A., et al.: Searching for MobileNetV3. In: ICCV, pp. 1314–1324 (2019)
23. Kelly, C.J., et al.: Key challenges for delivering clinical impact with AI. BMC Med. **17**(1), 1–9 (2019)
24. Kinyanjui, N.M., et al.: Fairness of classifiers across skin tones in dermatology. In: MICCAI, pp. 320–329. Springer (2020)
25. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: NeurIPS, pp. 6402–6413 (2017)
26. Lee, K., et al.: A simple unified framework for detecting OOD samples and adversarial attacks. In: NeurIPS, pp. 7167–7177 (2018)
27. Li, D., et al.: Deeper, broader and artier domain generalization. In: ICCV, pp. 5543–5551 (2017)
28. Liang, S., Li, Y., Srikant, R.: Enhancing reliability of OOD detection in neural networks. In: ICLR (2018)
29. Liu, W., et al.: Energy-based out-of-distribution detection. In: NeurIPS, pp. 21464–21475 (2020)
30. Ma, X., et al.: Understanding adversarial attacks on deep learning based medical image analysis. Pattern Recognit. **110**, 107332 (2021)
31. MacKay, D.J.C.: A practical Bayesian framework for backprop networks. Neural Comput. **4**(3), 448–472 (1992)
32. Madry, A., et al.: Towards deep learning models resistant to adversarial attacks. In: ICLR (2018)
33. Platt, J.C.: Probabilistic outputs for SVMs. In: Advances in Large Margin Classifiers, pp. 61–74. MIT Press (1999)
34. Pooch, E.H., et al.: Can we trust deep learning based diagnosis? In: Thoracic Image Analysis, pp. 74–83. Springer (2020)
35. Rauber, J., Brendel, W., Bethge, M.: Foolbox: A Python toolbox to benchmark robustness. arXiv:1707.04131 (2017)
36. Tack, J., et al.: CSI: Novelty detection via contrastive learning. In: NeurIPS, pp. 11839–11852 (2020)
37. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset. Sci. Data **5**, 180161 (2018)
38. Tzeng, E., et al.: Adversarial discriminative domain adaptation. In: CVPR, pp. 7167–7176 (2017)
39. Yang, Y., Soatto, S.: FDA: Fourier domain adaptation for semantic segmentation. In: CVPR, pp. 4085–4095 (2020)
40. Zech, J.R., et al.: Variable generalization performance of a deep learning model for pneumonia detection. PLOS Med. **15**(11), e1002686 (2018)