CORRECTING ERRORS, CORRECTLY: CERTIFIED ROBUSTNESS FOR ERROR-CORRECTING OUTPUT CODES

Anonymous authors
Paper under double-blind review

ABSTRACT

Error-Correcting Output Codes (ECOC) leverage redundancy from coding theory to map classes into structured codewords rather than standard one-hot encodings. While this paradigm enhances empirical robustness in machine learning tasks, it remains a heuristic defense susceptible to adaptive adversaries, and to date no certified robustness guarantee exists. We bridge this gap by presenting the first certification framework for ECOC-based models. Leveraging per-bit certification and decoding-aware analysis, we derive closed-form robustness bounds for both hard and soft decoding schemes, enabling efficient and scalable certification for structured multi-class outputs. Extensive experiments show that our approach consistently certifies a larger certified radius and a higher proportion of inputs than naive certification of the same ECOC model, while also increasing resistance to practical adversarial attacks. These results demonstrate that integrating error-correcting structure into certification yields stronger and more reliable robustness guarantees for multi-class prediction.

1 INTRODUCTION

Error-Correcting Output Codes (ECOC) Dietterich & Bakiri (1994) originated in coding theory as a mechanism to ensure reliable data transmission over noisy channels, but they offer a powerful paradigm for machine learning tasks. Rather than using standard one-hot encodings, class labels are mapped to unique, redundant binary codewords (e.g., strings of $\{-1, +1\}$). A model is then trained to predict these individual bits, and inference is performed by decoding the predicted bit string to the nearest valid codeword. This approach Deng et al. (2010); Verma & Swami (2019); Song et al. (2021); Yu et al. (2023) improves model robustness by introducing structural redundancy. Just as recovering a message despite corrupted bits during transmission, an ECOC-based model can theoretically recover the correct class label even if a subset of the output bits is flipped.

Due to this intrinsic resilience, ECOC has been widely adopted to enhance the robustness against adversarial attacks, where carefully crafted input modifications induce incorrect predictions, posing significant challenges Goodfellow et al. (2015); Eykholt et al. (2018); Zou et al. (2023); Yang et al. (2024). ECOC-based models have been shown to yield substantial improvements in robustness across a variety of datasets and attack settings Verma & Swami (2019); Song et al. (2021). However, such findings are based on empirical evaluations, and the robustness observed is not theoretically guaranteed. In fact, recent works Zhang et al. (2020); Philippon & Gagné (2023); Tramer et al. (2020) have demonstrated that ECOC-based models can still be vulnerable to targeted attacks specifically crafted to exploit their structured decision boundaries. This motivates the need for *certified robustness*, which refers to provable guarantees that a model's prediction will remain unchanged within a bounded set of input perturbations (e.g., an ℓ_2 ball).

While certified robustness has been successfully established for standard classifiers Cohen et al. (2019); Lecuyer et al. (2019); Salman et al. (2019); Gowal et al. (2018), these techniques have not yet been adapted to ECOC-based architectures. Existing certification methods typically assume unstructured, flat output spaces (e.g., softmax probabilities) and fail to account for the combinatorial, two-stage decoding process that defines error-correcting models. This limitation creates a significant

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

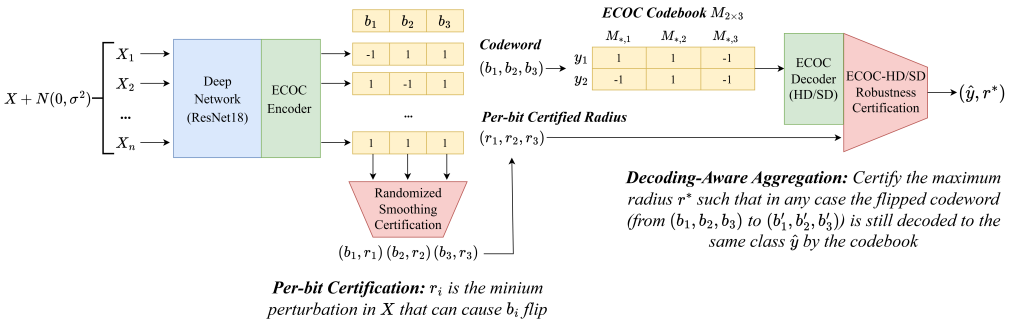


Figure 1: Overview of the ECOC classification process and our certification framework. The model produces m bit-level outputs, which are decoded via either hard or soft decoding to the nearest codeword. Our method certifies robustness by deriving per-bit guarantees and aggregating them through the ECOC decoding structure, enabling decoding-aware multi-class certified radii.

gap in the literature: despite the theoretical promise of ECOC for error tolerance, there is currently no framework to formally certify this capability in deep learning.

This paper addresses this gap by introducing a novel *certified robustness framework* tailored for ECOC-based neural classifiers. Unlike traditional classification settings where a single prediction is certified, ECOC poses a unique challenge: the final decision arises from the joint prediction of multiple correlated binary outputs. To tackle this challenge, we propose a *per-bit certification* approach that treats each output bit in the ECOC codeword as an independent binary classifier and certifies its robustness individually. Crucially, we develop a *decoding-aware* method to aggregate these per-bit certificates into a formal robustness guarantee for the overall ECOC prediction. Figure 1 illustrates the proposed inference and certification process. Our analysis bridges the gap between local bit-wise guarantees and global class-level certification, accounting for the structure of the codebook and the decoding scheme (e.g., hard or soft decoding). This yields closed-form robustness bounds that are both computationally tractable and theoretically grounded. By adapting certification analysis to the ECOC setting, we enable the first instance-wise certified defense for ECOC-based models. In summary, this work makes the following contributions:

- We propose the first certified robustness framework for ECOC-based models, providing provable guarantees under adversarial perturbations.
- We derive closed-form robustness bounds for both hard and soft decoding schemes based on per-bit certification, enabling efficient and scalable verification.
- We empirically validate our approach across multiple benchmarks, demonstrating that ECOC-based models achieve strong certified robustness and outperform standard certification.

2 Backgrounds and Literature Review

ECOC ECOC Dietterich & Bakiri (1994) originated in coding theory as a framework for ensuring reliable data transmission over noisy communication channels. This method maps distinct messages of size K to unique binary codewords of length m, forming a structured codebook $M \in \{\pm 1\}^{K \times m}$. The central mechanism relies on systematic redundancy ($m > K$) to maximize the minimum distance between each pair of codewords as a safety margin d_{min}. By decoding a potentially corrupted bit-string to the nearest valid codeword, the system can correct up to $\lfloor (d_{min} - 1)/2 \rfloor$ bit errors.

ECOC has been explored for robust multi-class classification, and several studies Verma & Swami (2019); Song et al. (2021) have demonstrated that employing this framework yields improved robustness. These benefits are attributed to the error-tolerant nature of the codeword space and the distributed representation of class information across multiple bits. However, this robustness remains heuristic: ECOC classifiers, like standard models, can still be fooled by carefully crafted perturbations Zhang et al. (2020); Philippon & Gagné (2023); Tramer et al. (2020), and no formal robustness guarantees exist in prior work.

Certified Robustness Deep learning models remain highly vulnerable to adversarial attacks Szegedy et al. (2013); Goodfellow et al. (2014), such as PGD Madry et al. (2017) and C&W Carlini & Wagner (2017). This vulnerability extends to ECOC-based architectures, where targeted attacks Zhang et al. (2020); Tramer et al. (2020) exploit codebook structure to optimize perturbations toward vulnerable codewords. While empirical defenses Madry et al. (2017); Nie et al. (2022); Tramèr et al. (2017) attempt to mitigate these threats, they lack formal guarantees and remain susceptible to circumvention by adaptive attacks Carlini & Wagner (2017).

In contrast, *certified robustness* provides formal guarantees that a classifier's prediction remains constant within a bounded perturbation set (e.g., an l_p-ball). While methods based on convex relaxations Wong & Kolter (2018) or Lipschitz continuity Zhang et al. (2019) offer deterministic bounds, *randomized smoothing* has become the dominant standard for scalable certification Cohen et al. (2019); Salman et al. (2019). Randomized smoothing transforms an arbitrary classifier f into a smoothed classifier g by taking the expected prediction under Gaussian noise $\mathcal{N}(0, \sigma^2 I)$, certifying robustness within an l_2 radius $R = \frac{\sigma}{2}(\Phi^{-1}(p_A) - \Phi^{-1}(p_B))$ determined by the top-two class probabilities p_A and p_B Cohen et al. (2019). However, this framework typically assumes flat output spaces (e.g., softmax), failing to account for the combinatorial decoding mechanism inherent to structured prediction models like ECOC.

To our knowledge, no formal certification exists for ECOC models. This gap is non-trivial, as the combinatorial nature of the decoding process introduces structural dependencies that standard certification methods can not address. We bridge this gap by establishing the first provable robustness guarantees tailored to these error-correcting architectures.

3 Certified Robustness Framework for ECOC

3.1 Problem Statement and Definitions

We study supervised multi-class classification under test-time adversarial perturbations. Each input $x \in \mathbb{R}^d$ is associated with a ground-truth label $y \in [K] := \{1, \ldots, K\}$, and a classifier $f : \mathbb{R}^d \to [K]$ outputs a prediction $\hat{y} = f(x)$ at test time. Our goal is to establish instance-wise certified robustness guarantees for classifiers implemented using the Error-Correcting Output Codes (ECOC) framework. In the following, we first introduce the ECOC model, followed by the formalization of the adversarial threat, specifying the allowable test-time perturbations. We then define the certified robustness objective, outlining the defense goal and the robustness guarantees.

ECOC model In general machine learning models, discrete class labels are represented by encoding each label $k \in [K]$ as a *codeword* of length m, drawn from a predefined *codebook* M. A common choice for this codebook is the one-hot encoding scheme, where each class is represented by a binary vector $m = K$ with a single 1 in the position corresponding to the class index and 0s elsewhere. In contrast, the ECOC framework generalizes this approach by employing a redundant codebook design, often with $m > K$, to improve both robustness and generalization. The codebook $M \in \{\pm 1\}^{K \times m}$ has K rows and m columns, where row $M_{k,*}$ is the distinct m-bit codeword for class k.

Such a codebook design induces a two-stage classification process $\hat{y} = f(x) = d(z(x))$, where the encoder $z : \mathbb{R}^d \to \mathbb{R}^m$ maps input x to latent representation $z(x)$, and the decoder $d : \mathbb{R}^m \to [K]$ maps $z(x)$ to a predicted label $\hat{y} \in [K]$ via codeword similarity. We consider two decoding schemes:

- **Hard Decoding (HD)** Dietterich & Bakiri (1994); Kong & Dietterich (1995); Deng et al. (2010); Gupta & Amin (2022) Each bit is binarized by its sign, $\mathbf{b}(x) \in \{\pm 1\}^m,\ \mathbf{b}_j(x) := \mathrm{sign}(z_j(x))$, and the predicted label is the closest codeword under Hamming distance (or maximal correlation):

$$\hat{y} = \arg\max_{k \in [K]} \mathbf{b}(x)^\top M_{k,*}. \tag{1}$$

- **Soft Decoding (SD)** Verma & Swami (2019); Song et al. (2021); Yu et al. (2023) Each bit contributes a real-valued alignment $p_j(x) \in [-1, 1]$, $p_j(x) = \tanh(z_j(x))$, and classes are scored by linear correlation with the codeword:

$$\hat{y} = \arg\max_{k \in [K]} P_k(x),\ P_k(x) := \sum_{j=1}^{m} p_j(x)\, M_{k,j}. \tag{2}$$

The robustness of an ECOC-based model largely depends on the structure of the codebook M, particularly its minimum Hamming distance $d_{\min}$, which determines the number of bit errors correctable under hard decoding. A larger $d_{\min}$ generally improves robustness but may require longer codewords or support fewer classes. While our theoretical certification bounds hold for any arbitrary binary codebook $M \in \{\pm 1\}^{K \times m}$ (such as random Bernoulli codes, BCH codes, or task-learned codebooks), in this work we primarily adopt the canonical *Hadamard codebook* Dietterich & Bakiri (1994); Verma & Swami (2019); Song et al. (2021), whose rows are pairwise orthogonal, maximizing inter-class separation ($d_{\min} = m/2$) while remaining efficient to scale. Furthermore, the mutual orthogonality of Hadamard columns enforces near-decorrelated supervision across bits, suppressing inter-bit error correlation in deep feature representations ($|\rho_{i,j}| < 0.08$ on CIFAR-10; see Section D.1) and aligning neural representations with independent error channel assumptions. When the number of classes K is not a power of two, codewords are subsampled from an $m \times m$ Hadamard matrix to maximize mutual separation (e.g., selecting 100 rows from a 128×128 Hadamard matrix for CIFAR-100). More sophisticated codebook designs Escalera et al. (2010); Gupta & Amin (2022) can be seamlessly integrated into our certification framework.

Adversarial Threat Model We assume a *white-box adversary* with full access to model parameters, codebook M, and architecture. At test time, the adversary perturbs an input x to construct an adversarial example $x' \in B(x, \epsilon)$ aiming to alter the classification outcome $f(x') \neq f(x)$, where perturbations are constrained within an ℓ_p ball:

$$B(x, \epsilon) := \{x' : \|x' - x\|_p \leq \epsilon\}. \tag{3}$$

Certified Robustness Objective Our objective is to provide *formal robustness certificates* for ECOC-based models. In our framework, randomized smoothing transforms the base ECOC model into a smoothed classifier $g(x) = d(\bar{z}(x))$, where $\bar{z}(x)$ denotes the element-wise smoothed representation under isotropic Gaussian noise $\delta \sim \mathcal{N}(0, \sigma^2 I)$—obtained via majority-vote smoothing for hard decoding and median smoothing for soft decoding. Specifically, for a given test input x, we aim to derive a certified radius $R^* > 0$ guaranteeing that the smoothed prediction remains provably invariant under any perturbation within $B(x, R^*)$, regardless of the attack algorithm:

$$g(x') = g(x) \quad \forall x' \in B(x, R^*). \tag{4}$$

3.2 Certified Robustness for Hard-codeword Decoding

We now propose our method for certifying the robustness of ECOC classifiers under hard decoding. We begin by regarding the ECOC model as an *ensemble of m binary classifiers*, where each binary classifier corresponds to one bit in the codeword and predicts the sign of the latent feature $z_j(x)$. We emphasize that this is only a conceptual interpretation, while in practice, the ECOC classifier is implemented as a single neural network.

Under this interpretation, we apply standard certified robustness methods, such as randomized smoothing Cohen et al. (2019), to each binary classifier $h_j(x) = \text{sign}(z_j(x))$ separately, yielding an individual certified radius r_j for each bit that guarantees the sign of $z_j(x)$ remains unchanged within an ℓ_2 ball of radius r_j. These per-bit guarantees can then be composed through the ECOC decoding rule to derive an overall certified radius for the entire classifier.

Our objective is to derive a closed-form expression for the largest radius R^* such that the ECOC prediction remains unchanged for all perturbations $\|x' - x\|_2 \leq R^*$. To this end, we propose the following theorem, which provides a certified robustness guarantee for hard-decoded ECOC classifiers based on per-bit certified radii.

Theorem 3.1 (Certified Robustness for ECOC-HD). *Let $f(x) = \arg\max_{k \in [K]} \mathbf{b}(x)^\top M_{k,*}$ be an ECOC classifier with hard decoding, where $\mathbf{b}(x) \in \{\pm 1\}^m$ is the binarized output defined by $\mathbf{b}_j(x) = \text{sign}(z_j(x))$ for latent features $z_j(x)$. Assume that for each $j \in [m]$, the function z_j admits a certified radius $r_j > 0$ under the ℓ_2 norm, such that $\text{sign}(z_j(x')) = \text{sign}(z_j(x))$ for all $\|x' - x\|_2 \leq r_j$.*

Then, for any input x with predicted label $\hat{y} = f(x)$, the classifier is certifiably robust within radius

$$R^* = \min_{k \neq \hat{y}} R_k,$$

where the class-wise certified radius R_k is given by

$$R_k := \min \left\{ \max_{j \in \mathcal{J}} r_j \;\middle|\; \mathcal{J} \subseteq \mathcal{C}_k, \sum_{j \in \mathcal{J}} \delta_j^{(k)} \geq s_k \right\},$$

and the quantities are defined as:

- $s_k := \mathbf{b}(x)^\top (M_{\hat{y},*} - M_{k,*})$*, the score margin between the predicted class $\hat{y}$ and class k,*
- $\delta_j^{(k)} := 2 \cdot \mathbf{b}_j(x) \cdot (M_{\hat{y},j} - M_{k,j})$*, the reduction in score margin if bit j flips,*
- $\mathcal{C}_k := \{j \in [m] \mid \delta_j^{(k)} > 0\}$*, the set of bits whose flips reduce the margin.*

Proof sketch. A perturbation of norm r can only flip bits with certified radius $r_j \leq r$. For each competing class k, we compute the score margin s_k and identify the smallest set of flippable bits whose cumulative margin reduction exceeds s_k; the largest r_j in this set defines R_k. Taking the minimum over all $k \neq \hat{y}$ yields R^*. See Section A for the full proof. □

3.3 Certified Robustness for Soft-codeword Decoding

We now turn to the case of soft decoding, where the ECOC classifier uses real-valued bit outputs to compute class scores via linear correlation. Each bit output is defined as $p_j(x) = \tanh(z_j(x)) \in [-1, 1]$, and the score $P_k(x)$ for class k is computed as shown in Equation (2).

In this setting, perturbations affect the continuous values $p_j(x)$, rather than flipping binary bits. For each bit, we derive a certified interval under an ℓ_2-bounded perturbation:

$$p_j(x') \in [\underline{\tau}_j(r), \overline{\tau}_j(r)] \quad \text{for all } \|x' - x\|_2 \leq r. \tag{5}$$

Following median smoothing Chiang et al. (2020), under an ℓ_2 perturbation of radius r, the median output of $p_j(x+\delta)$ at x' is bounded in closed form by the quantiles of the smoothed output distribution at x:

$$\underline{\tau}_j(r) = q_j \left(\Phi \left(-\frac{r}{\sigma} \right) \right), \quad \overline{\tau}_j(r) = q_j \left(\Phi \left(\frac{r}{\sigma} \right) \right), \tag{6}$$

where $q_j(\beta)$ denotes the β-quantile of the random variable $p_j(x + \delta)$ with $\delta \sim \mathcal{N}(0, \sigma^2 I)$, and Φ is the standard normal cumulative distribution function.

Our objective is to compute the largest radius R^* such that the predicted class remains unchanged for all perturbations $x' \in B(x, R^*)$. We now state the robustness guarantee under soft decoding.

Theorem 3.2 (Certified Robustness for ECOC-SD)**.** *Let $f(x) = \arg\max_k P_k(x)$ be the ECOC classifier with soft decoding, where $P_k(x) = \sum_{j=1}^m p_j(x) M_{k,j}$ and $p_j(x) = \tanh(z_j(x)) \in [-1, 1]$. Suppose for each bit $j \in [m]$, the output $p_j(x)$ satisfies the certified interval $p_j(x') \in [\underline{\tau}_j(r), \overline{\tau}_j(r)]$ for all $\|x' - x\|_2 \leq r$. Then, the classifier is certifiably robust within radius*

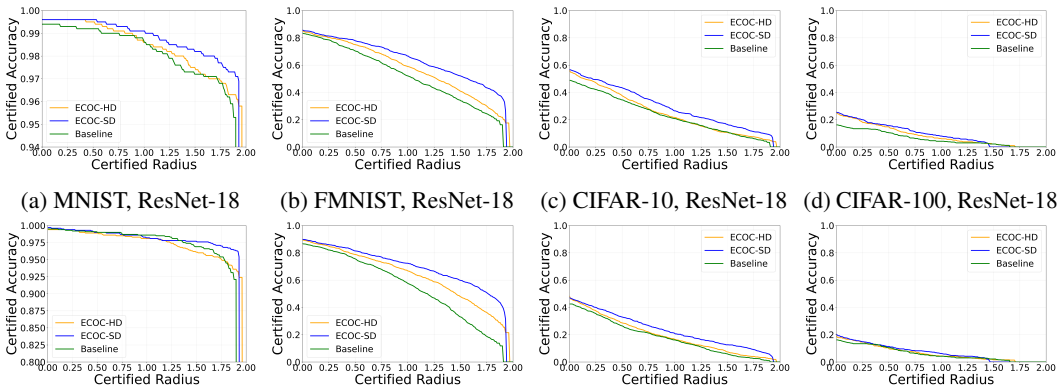
$$R^* = \max \left\{ r \;\middle|\; \underline{P}_{\hat{y}}(x, r) > \max_{k \neq \hat{y}} \overline{P}_k(x, r) \right\},$$

where the worst-case bounds on class scores are given by:

$$\underline{P}_{\hat{y}}(x, r) = \sum_{j=1}^m M_{\hat{y},j} \cdot \begin{cases} \underline{\tau}_j(r), & \text{if } M_{\hat{y},j} = +1, \\ \overline{\tau}_j(r), & \text{if } M_{\hat{y},j} = -1, \end{cases}$$

$$\overline{P}_k(x, r) = \sum_{j=1}^m M_{k,j} \cdot \begin{cases} \overline{\tau}_j(r), & \text{if } M_{k,j} = +1, \\ \underline{\tau}_j(r), & \text{if } M_{k,j} = -1. \end{cases}$$

Proof sketch. For each perturbation radius r, the certified intervals $[\underline{\tau}_j(r), \overline{\tau}_j(r)]$ bound each bit output. We compute the worst-case lower bound on the predicted class score $\underline{P}_{\hat{y}}$ and the worst-case upper bound on every competing class score $\overline{P}_k$ by choosing the adversarially optimal endpoint for each bit. Robustness holds as long as $\underline{P}_{\hat{y}}(x, r) > \max_{k \neq \hat{y}} \overline{P}_k(x, r)$, and R^* is the largest such r. See Section A for the full proof. □



(a) MNIST, ResNet-18 (b) FMNIST, ResNet-18 (c) CIFAR-10, ResNet-18 (d) CIFAR-100, ResNet-18

(e) MNIST, ResNet-50 (f) FMNIST, ResNet-50 (g) CIFAR-10, ResNet-50 (h) CIFAR-100, ResNet-50

Figure 2: Certified accuracy across datasets with ResNet-18 and ResNet-50. Comparison of *ECOC-HD*, *ECOC-SD*, and the *Baseline* randomized smoothing Cohen et al. (2019) on MNIST, FMNIST, CIFAR-10, and CIFAR-100. ECOC-based certification consistently improves certified accuracy across a wide range of radii and datasets.

Remark 3.3 (Soundness and Tightness of Cartesian Bounding). Evaluating $\underline{P}_{\hat{y}}(x, r)$ and $\overline{P}_k(x, r)$ across the Cartesian product $\prod_{j=1}^{m}[\underline{\tau}_j(r), \overline{\tau}_j(r)]$ provides a sound over-approximation of the adversary's reachable set in latent space: $\mathcal{Z}(r) \subseteq \prod_{j=1}^{m}[\underline{\tau}_j(r), \overline{\tau}_j(r)]$. While independent coordinate bounding avoids intractable non-convex optimization over deep feature representations, empirical attack evaluations confirm that the resulting certificate captures the vast majority of the empirical robustness boundary ($R^*/R_{\text{attack}} \approx 0.72$ on CIFAR-10; see Section A.4).

3.4 Algorithm Implementation

Our certification framework builds upon randomized smoothing Cohen et al. (2019), performing per-bit certification followed by decoding-aware radius aggregation. For **Hard Decoding (ECOC-HD)**, we certify the sign stability of each bit $z_j(x)$ using standard randomized smoothing to obtain per-bit radii r_j; the overall certified radius R^* is then determined via binary search over candidate radii according to Theorem 3.1. For **Soft Decoding (ECOC-SD)**, we apply median smoothing Chiang et al. (2020) to compute certified confidence intervals $[\underline{\tau}_j(r), \overline{\tau}_j(r)]$ for continuous bit alignments $p_j(x) = \tanh(z_j(x))$, and find the optimal radius R^* via binary search satisfying the score dominance condition in Theorem 3.2. Certifying bits independently avoids exponential joint output complexity while preserving the linear scalability of randomized smoothing.

Statistical Guarantees and Abstention Handling To ensure a joint confidence of at least $1 - \alpha$ across all m bits simultaneously, a union bound (Bonferroni correction) sets the failure probability for each individual bit to α/m. Although this scales the significance threshold (e.g., from $\alpha = 10^{-3}$ to 2×10^{-6} for $m = 512$), in Clopper-Pearson binomial estimation with $N = 10^5$ samples, this adjustment shifts the required Gaussian quantile by only $\approx 1.66\sigma$. For well-trained base classifiers with high bit-prediction accuracy under noise (typically $p_A^{(j)} \geq 0.98$), the resulting reduction in per-bit certified radius $r_j = \sigma\Phi^{-1}(p_A^{(j)})$ is minor (< 0.05). If an individual bit cannot be certified with probability > 0.5 (i.e., an abstention occurs), it is conservatively treated as uncertified: we set $r_j = 0$ for ECOC-HD or assign $[\underline{\tau}_j(r), \overline{\tau}_j(r)] = [-1, 1]$ for ECOC-SD, maintaining provable aggregation soundness under these conservative bounds. The aggregation step incurs an overhead of only $O(T_{\text{search}} \cdot K \cdot m)$, which is negligible compared to the network sampling cost $O(N \cdot T_{\text{net}})$. Full pseudocode and complexity analyses are provided in Sections B and C.

Dataset	Defence Method	Radii Mean	Radii Max	Std.	Certified Prop.
MNIST	ECOC-HD	1.932	1.963	0.190	99.6%
	ECOC-SD	1.910	1.932	0.164	99.6%
	Baseline	1.873	1.906	0.201	99.4%
FMNIST	ECOC-HD	1.132	1.963	0.718	84.8%
	ECOC-SD	1.257	1.932	0.737	85.6%
	Baseline	1.017	1.906	0.703	83.4%
CIFAR-10	ECOC-HD	0.489	1.963	0.620	55.0%
	ECOC-SD	0.584	1.932	0.692	56.7%
	Baseline	0.453	1.906	0.615	48.9%
CIFAR-100	ECOC-HD	0.166	1.707	0.375	24.4%
	ECOC-SD	0.183	1.458	0.396	25.6%
	Baseline	0.118	1.657	0.333	16.2%

Table 1: Summary statistics for the ResNet-18 certified radii, corresponding to the first row of Figure 2. We report the mean certified radius (Radii Mean), maximum radius (Radii Max), standard deviation (Std.), and proportion of test samples with nonzero certificates (Certified Prop.). Both ECOC-HD and ECOC-SD consistently achieve larger certified radii and higher certifiable coverage than the baseline Cohen et al. (2019), demonstrating the effectiveness of decoding-aware certification.

4 EXPERIMENTS

We empirically evaluate the proposed certified ECOC framework across diverse datasets, model configurations, and adversarial threat scenarios. Our experiments are designed to assess: (1) the generalizability of the proposed certified defense under varying settings, (2) the improvements in certified robustness compared to baseline certified defence method Cohen et al. (2019), (3) the empirical robustness of our defense under strong adversarial attacks, including targeted ECOC attacks, and (4) the influence of key ECOC and smoothing hyperparameters (deferred to Section D). Our evaluation protocol follows standard practice in certified robustness Cohen et al. (2019); Salman et al. (2019).

4.1 EXPERIMENTAL SETUP

Datasets and Model Architecture. We conduct experiments on standard benchmarks: MNIST Lecun et al. (1998), FMNIST Xiao et al. (2017), CIFAR-10, CIFAR-100 Krizhevsky et al. (2009), and Tiny-ImageNet Le & Yang (2015), which cover a broad range of visual complexity and classification granularity. This diversity enables a comprehensive evaluation of the generalizability of our certified ECOC framework. Unless otherwise specified, all models use a ResNet-18 backbone He et al. (2016), with evaluations on ResNet-50 and Tiny-ImageNet provided in Section E.

Model Training and Optimization. Following standard practice in certified robustness Cohen et al. (2019); Salman et al. (2019), all models are trained with Gaussian noise data augmentation, injecting perturbation $\delta \sim \mathcal{N}(0, \sigma_{\text{train}}^2 I)$ to inputs during training, matching the test-time smoothing parameter $\sigma_{\text{train}} = \sigma \in \{0.1, 0.3, 0.5\}$. The network is trained by minimizing the multi-bit Binary Cross-Entropy with Logits loss:

$$\mathcal{L}_{\text{ECOC}}(\theta) = \frac{1}{m} \sum_{j=1}^{m} \ell_{\text{BCE}} \left(z_j(x + \delta; \theta),\ \frac{M_{y,j} + 1}{2} \right), \tag{7}$$

using SGD with momentum 0.9, weight decay 5×10^{-4}, and a batch size of 128. The initial learning rate of 0.1 is decayed using a cosine annealing schedule over 120 epochs.

ECOC Models and Decoding Schemes. We evaluate both decoding strategies introduced in our framework: (1) **ECOC-HD** (hard decoding), which certifies bit-wise robustness via randomized smoothing Cohen et al. (2019) and aggregates per-bit radii according to Theorem 3.1; and (2) **ECOC-SD** (soft decoding), which certifies interval stability of continuous bit outputs using median smoothing Chiang et al. (2020) and computes decoding-aware score bounds via Theorem 3.2. Hadamard codebooks Dietterich & Bakiri (1994) are used unless otherwise noted, due to their favorable distance structure and scalability.

Baseline. As the primary baseline, we apply *standard randomized smoothing* Cohen et al. (2019) directly to the ECOC model. Following prior work Verma & Swami (2019); Song et al. (2021); Yu et al. (2023), we employ soft decoding rather than hard decoding, as it yields superior robustness and is standard practice for robust ECOC classifiers. Treating the entire ECOC model as a single one-hot classifier allows this baseline to isolate the benefits of our proposed decoding-aware ECOC certification from the inherent robustness of the ECOC representation itself. Moreover, ECOC provides an architectural defense orthogonal to training-time certification objectives such as SmoothAdv Salman et al. (2019) or MACER Zhai et al. (2020), which can naturally be composed with our decoding framework. For reference, standard softmax ResNet-110 models certified via Cohen et al. Cohen et al. (2019) at $\sigma = 0.5$ achieve $\approx 41\%$ certified accuracy at $r = 0.5$ on CIFAR-10, demonstrating that our ECOC-SD architecture offers competitive coverage and larger mean radii (0.584 vs. 0.453).

Certification Protocol and Hyperparameters. For certification, we adhere strictly to the statistical protocol established by Cohen et al. Cohen et al. (2019): $N = 100,000$ Monte Carlo samples are drawn for statistical estimation, $N_0 = 1,000$ samples for top-prediction screening, and the overall significance level is set to $\alpha = 0.001$. Evaluations are conducted over $|\mathcal{X}_{\text{test}}| = 500$ randomly sampled test instances per dataset. The noise level $\sigma \in \{0.1, 0.3, 0.5\}$ controls the tradeoff between clean accuracy and certified radius, explored extensively in our ablation study (Section D).

Evaluation Metrics. We evaluate two robustness metrics: (1) *Certified Accuracy* at radius r, $\frac{1}{|\mathcal{X}_{\text{test}}|}\sum_{(x,y)\in\mathcal{X}_{\text{test}}} \mathbf{1}\{f(x) = y \wedge R^*(x) \geq r\}$; and (2) the canonical *Average Certified Radius (ACR)* Zhai et al. (2020), $\text{ACR} = \frac{1}{|\mathcal{X}_{\text{test}}|}\sum_{(x,y)\in\mathcal{X}_{\text{test}}} R^*(x) \cdot \mathbf{1}\{f(x) = y\}$, reported as *Radii Mean* in Table 1, where uncertified or misclassified inputs contribute $R^* = 0$. We also report empirical *Robust Accuracy* under an attack $\mathcal{A}$ with $\|\mathcal{A}(x) - x\|_2 \leq 1.0$.

4.2 Certified Robustness Results

Figure 2 presents the certified accuracy curves for ECOC-HD, ECOC-SD, and the baseline across all datasets, while Table 1 provides complementary certified radius statistics. These results jointly reveal three consistent patterns.

Strong robustness across datasets. Both ECOC-HD and ECOC-SD substantially improve certified robustness relative to the one-hot baseline. On CIFAR-10, for example, ECOC-SD increases the mean certified radius from 0.453 to 0.584 and raises the proportion of samples with nonzero certificates from 48.9% to 56.7%. Similar improvements are observed on FMNIST, where ECOC-SD increases the mean radius from 1.017 to 1.257, and on CIFAR-100, where ECOC-HD and ECOC-SD both improve mean radii over the baseline. These numerical margins are consistent with the widened accuracy gaps at larger radii seen in Figure 2, indicating that ECOC-based certification provides more effective robustness guarantees across the full range of perturbation strengths.

ECOC-SD provides the strongest guarantees. Across most datasets, ECOC-SD demonstrates the highest overall certified robustness. Its mean certified radius leads ECOC-HD on FMNIST and CIFAR-10 and slightly trails it on MNIST, while maintaining noticeably lower variance (e.g., 0.164 on MNIST compared to ECOC-HD's 0.190). This reduced variability suggests more stable certification behavior across samples, which aligns with the theoretical benefit of using continuous bit outputs: they enable tighter per-bit interval constraints and yield less conservative class-level certificates.

Generalizability. The gains offered by our framework hold across datasets of varying complexity. On MNIST, all methods certify over 99% of samples, with ECOC achieving slightly higher coverage. On more challenging datasets, the gap widens: on CIFAR-100, ECOC-SD increases certifiable coverage from 16.2% to 25.6%, a substantial improvement where certification is typically difficult. These consistent gains across diverse benchmarks demonstrate that our framework generalizes effectively across input domains and task complexities.

4.3 Robustness Against Empirical Attacks

Finally, we evaluate empirical robustness under PGD Madry et al. (2017), C&W Carlini & Wagner (2017), and the targeted ECOC attack Zhang et al. (2020), with $\ell_2 = 1.0$. Results are summarized in Table 2. Several observations stand out:

Dataset	Defence Method	Clean Accuracy	Robust Accuracy (l_2-norm= 1.0)		
			PGD	C&W	Targeted ECOC
FMNIST	ECOC-HD	91.94%	66.47%	65.98%	67.50%
	ECOC-SD	92.13%	78.05%	74.64%	73.92%
	Baseline	92.48%	65.87%	63.63%	63.63%*
CIFAR-10	ECOC-HD	82.41%	35.19%	33.47%	29.27%
	ECOC-SD	83.53%	38.42%	36.95%	33.52%
	Baseline	84.30%	25.23%	27.19%	27.19%*

Table 2: Empirical robustness against PGD, C&W, and targeted ECOC attacks under an ℓ_2 threat of radius 1.0. Both ECOC-HD and ECOC-SD outperform the baseline across all attack types. Notably, targeted ECOC attacks offer no advantage against our certified ECOC framework, confirming that decoding-aware robustness carries over to empirical adversarial settings. *Note on Baseline: Because standard one-hot models do not possess an ECOC codebook representation, Targeted ECOC attacks cannot be formulated directly; following standard convention, the baseline entry denotes standard targeted C&W performance.

Attack Optimization and Obfuscation Countermeasures. To ensure empirical evaluations reflect genuine adversarial robustness rather than gradient masking or obfuscation Athalye et al. (2018), all attacks on ECOC-HD were conducted using continuous surrogate objectives on the latent representations $z_j(x)$ (maximizing correlation with the target codeword while minimizing alignment with the true codeword), completely bypassing the non-differentiable $\mathrm{sign}(\cdot)$ operator. For smoothed models, gradients were evaluated using Expectation Over Transformation (EOT) Athalye et al. (2018) with $m_{\text{EOT}} = 20$ Monte Carlo noise realizations per gradient step, preventing stochastic noise from masking adversarial gradients.

Strong empirical defense. Both ECOC-HD and ECOC-SD outperform the baseline under all attack settings. The gains are especially notable on FMNIST, where ECOC-SD consistently achieves roughly 10% higher robust accuracy than the baseline under all of the attacks.

Defense against targeted ECOC attacks. Targeted ECOC attacks exploit codeword structure to maximize semantic vulnerability in ECOC models. Our certified ECOC framework explicitly accounts for decoding geometry, eliminating this advantage: targeted ECOC attacks are *no more effective* than standard C&W or PGD attacks. In several cases (e.g., FMNIST), ECOC-HD achieves slightly higher accuracy under targeted ECOC attacks (67.50%) than under C&W (65.98%). Across all evaluations, ECOC-SD demonstrates the strongest empirical defense, aligning closely with certified trends and validating our theoretical framework.

5 Conclusion

We introduced the first certified robustness framework for ECOC-based neural classifiers, establishing decoding-aware guarantees for both hard and soft decoding. By certifying bit-level stability and aggregating these bounds through the codebook structure, our method yields efficient, closed-form certificates for structured multi-class predictions.

Extensive experiments demonstrate that ECOC-based certification consistently outperforms standard randomized smoothing, with ECOC-SD achieving larger certified radii, higher coverage, and superior empirical defense against PGD, C&W, and targeted ECOC attacks. This work establishes ECOC as a principled foundation for certified robustness, paving the way for future exploration of adaptive codebook design and certified training.

AI USE STATEMENT

In this work, we used generative AI tools for none of the tasks requiring mandatory disclosure. We have not used generative AI tools for research ideation or execution, drafting sections of the paper, or generating synthetic datasets, and proving mathematical claims. Additionally, we used generative AI tools for aiding and polishing the writing. We have reviewed all AI-assisted work. All AI-suggested edits, grammatical corrections, and stylistic revisions were manually inspected and verified by the authors to ensure technical correctness and avoid altering the intended meaning. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning (ICML)*, pp. 274–283, 2018.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. Ieee, 2017.

Ping-yeh Chiang, Michael Curry, Ahmed Abdelkader, Aounon Kumar, John Dickerson, and Tom Goldstein. Detection as regression: Certified object detection with median smoothing. *Advances in Neural Information Processing Systems*, 33:1275–1286, 2020.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pp. 1310–1320. PMLR, 2019.

Huiqun Deng, George Stathopoulos, and Ching Y Suen. Applying error-correcting output coding to enhance convolutional neural network for target detection and pattern recognition. In *2010 20th International Conference on Pattern Recognition*, pp. 4291–4294. IEEE, 2010.

Thomas G Dietterich and Ghulum Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of artificial intelligence research*, 2:263–286, 1994.

Sergio Escalera, Oriol Pujol, and Petia Radeva. Error-correcting ouput codes library. *The Journal of Machine Learning Research*, 11:661–664, 2010.

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 1625–1634, 2018.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*, 2018.

Samarth Gupta and Saurabh Amin. Scalable design of error-correcting output codes using discrete optimization with graph coloring. *Advances in Neural Information Processing Systems*, 35: 10094–10107, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Eun Bae Kong and Thomas G Dietterich. Error-correcting output coding corrects bias and variance. In *Machine learning proceedings 1995*, pp. 313–321. Elsevier, 1995.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.(2009), 2009.

Yann Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE symposium on security and privacy (SP)*, pp. 656–672. IEEE, 2019.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*, 2022.

Thomas Philippon and Christian Gagné. Improved robustness against adaptive attacks with ensembles and error-correcting output codes. *arXiv preprint arXiv:2303.02322*, 2023.

Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. *Advances in neural information processing systems*, 32, 2019.

Yang Song, Qiyu Kang, and Wee Peng Tay. Error-correcting output codes with ensemble diversity for robust learning in neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9722–9729, 2021.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *Advances in neural information processing systems*, 33:1633–1645, 2020.

Gunjan Verma and Ananthram Swami. Error correcting output codes improve probability estimation and adversarial robustness of deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International conference on machine learning*, pp. 5286–5295. PMLR, 2018.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Bo Yang, Zizhi Jin, Yushi Cheng, Xiaoyu Ji, and Wenyuan Xu. Adversarial robustness analysis of lidar-included models in autonomous driving. *High-Confidence Computing*, 4(1):100203, 2024.

Anlan Yu, Ning Lyu, Jieming Yin, Zhiyuan Yan, and Wujie Wen. Cola: Orchestrating error coding and learning for robust neural network inference against hardware defects. In *International Conference on Machine Learning*, pp. 40277–40289. PMLR, 2023.

Runtian Zhai, Chen Dan, Di He, Huan Zhang, Boqing Gong, Pradeep Ravikumar, Cho-Jui Hsieh, and Liwei Wang. MACER: Attack-free and scalable robust training via maximizing certified radius. *arXiv preprint arXiv:2001.02378*, 2020.

Bowen Zhang, Benedetta Tondi, Xixiang Lv, and Mauro Barni. Challenging the adversarial robustness of dnns based on error-correcting output codes. *Security and Communication Networks*, 2020(1): 8882494, 2020.

Huan Zhang, Hongge Chen, Chaowei Xiao, Sven Gowal, Robert Stanforth, Bo Li, Duane Boning, and Cho-Jui Hsieh. Towards stable and efficient training of verifiably robust neural networks. *arXiv preprint arXiv:1906.06316*, 2019.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. In *Advances in Neural Information Processing Systems*, 2023.

A Proofs of Main Theorems

A.1 Proof of Theorem 3.1

Proof. The hard decoding classifier chooses the label whose codeword is closest in Hamming distance to $\mathbf{b}(x)$. Equivalently, it selects the label $\hat{y}$ maximizing $\mathbf{b}(x)^\top M_{k,*}$.

A change in prediction occurs only if some class $k \neq \hat{y}$ has a higher score than $\hat{y}$:

$$\mathbf{b}(x')^\top M_{k,*} > \mathbf{b}(x')^\top M_{\hat{y},*}.$$

Rewriting this:

$$\mathbf{b}(x')^\top (M_{k,*} - M_{\hat{y},*}) > 0 \;\Leftrightarrow\; \mathbf{b}(x')^\top (M_{\hat{y},*} - M_{k,*}) < 0.$$

We define the initial score margin between $\hat{y}$ and k as:

$$s_k := \mathbf{b}(x)^\top (M_{\hat{y},*} - M_{k,*}) > 0.$$

Each bit flip from $b_j(x)$ to $-b_j(x)$ changes the margin by:

$$\delta_j^{(k)} := 2b_j(x) \cdot (M_{\hat{y},j} - M_{k,j}).$$

Thus, flipping a set of bits $\mathcal{J}$ reduces the score margin by $\sum_{j\in\mathcal{J}} \delta_j^{(k)}$. To guarantee that prediction does not change under a perturbation of radius r, we must ensure that not enough bits can be flipped to reduce the score margin below zero. Each bit j is protected up to radius r_j, and a perturbation of norm r can only flip bits j with $r_j \leq r$.

To certify robustness, we identify the minimal subset $\mathcal{J}_k \subseteq \mathcal{C}_k$ of vulnerable bits such that their cumulative margin reduction exceeds s_k. The maximal r_j in this subset defines the minimal radius R_k required to potentially cause misclassification to class k. Taking the worst case across all $k \neq \hat{y}$ gives the certified radius:

$$R^* = \min_{k\neq\hat{y}} R_k.$$

□

A.2 Proof of Theorem 3.2

Proof. Given the certified interval for each bit $p_j(x') \in [\underline{\tau}_j(r), \overline{\tau}_j(r)]$, we consider the worst-case scenario for the classifier under any perturbation $x' \in B(x, r)$. To preserve the prediction $\hat{y}$, the worst-case lower bound of its score must remain higher than the worst-case upper bound of every competing class.

For the true label $\hat{y}$, the worst-case score occurs when each term $p_j(x')$ moves in the direction opposite to $M_{\hat{y},j}$, i.e.,

$$\underline{P}_{\hat{y}}(x, r) = \sum_{j=1}^{m} M_{\hat{y},j} \cdot \begin{cases} \underline{\tau}_j(r), & \text{if } M_{\hat{y},j} = +1, \\ \overline{\tau}_j(r), & \text{if } M_{\hat{y},j} = -1. \end{cases}$$

Similarly, for any class $k \neq \hat{y}$, the adversary maximizes $P_k(x')$ by aligning each $p_j(x')$ with $M_{k,j}$, yielding

$$\overline{P}_k(x, r) = \sum_{j=1}^{m} M_{k,j} \cdot \begin{cases} \overline{\tau}_j(r), & \text{if } M_{k,j} = +1, \\ \underline{\tau}_j(r), & \text{if } M_{k,j} = -1. \end{cases}$$

Robustness is guaranteed as long as $\underline{P}_{\hat{y}}(x, r) > \overline{P}_k(x, r)$ for all $k \neq \hat{y}$. The largest such radius r gives the certified robustness:

$$R^* = \max \left\{ r \;\middle|\; \underline{P}_{\hat{y}}(x, r) > \max_{k \neq \hat{y}} \overline{P}_k(x, r) \right\}.$$

□

Derivation of Median Smoothing Quantiles. Following Chiang et al. Chiang et al. (2020), let $Y_j = p_j(x + \delta) = \tanh(z_j(x + \delta))$ where $\delta \sim \mathcal{N}(0, \sigma^2 I)$, and let $F_{Y_j}(t; x) = \mathbb{P}(Y_j \leq t)$ denote its cumulative distribution function. By the Neyman-Pearson lemma and the Gaussian isoperimetric inequality, for any adversarial perturbation $\|x' - x\|_2 \leq r$, the perturbed distribution function is bounded by:

$$\Phi\left(\Phi^{-1}(F_{Y_j}(t; x)) - \frac{r}{\sigma}\right) \leq F_{Y_j}(t; x') \leq \Phi\left(\Phi^{-1}(F_{Y_j}(t; x)) + \frac{r}{\sigma}\right).$$

Setting $F_{Y_j}(t; x') = 0.5$ (the median output under perturbation) yields the condition $\Phi^{-1}(F_{Y_j}(t; x)) \mp r/\sigma = \Phi^{-1}(0.5) = 0$, which immediately establishes that the median under perturbation is bounded within the quantiles $[\underline{\tau}_j(r), \overline{\tau}_j(r)]$ defined in Equation (6):

$$\underline{\tau}_j(r) = q_j\left(\Phi\left(-\frac{r}{\sigma}\right)\right), \quad \overline{\tau}_j(r) = q_j\left(\Phi\left(\frac{r}{\sigma}\right)\right),$$

where $q_j(\beta) = \inf\{t \mid F_{Y_j}(t; x) \geq \beta\}$ is the β-quantile of the unperturbed smoothed distribution.

A.3 Generalization to ℓ_1 and ℓ_∞ Perturbation Norms

A key theoretical advantage of our framework is that the decoding-aware aggregation mechanisms in Theorem 3.1 and Theorem 3.2 are *distribution-agnostic*: they do not depend on the specific choice of isotropic Gaussian noise, but rather on the availability of sound per-bit certified guarantees. Consequently, the framework generalizes directly to alternative perturbation geometries:

- ℓ_1 **Perturbations via Laplacian Smoothing:** Under an ℓ_1 adversary $B_1(x, r) = \{x' \mid \|x' - x\|_1 \leq r\}$, smoothing with isotropic Laplace noise $\delta \sim \text{Laplace}(0, \lambda I)$ with probability density $p(\delta) \propto \exp(-\|\delta\|_1/\lambda)$ guarantees that each individual binary bit j admits an ℓ_1 certified radius:

$$r_j^{(1)} = \lambda \ln\left(\frac{1}{2\left(1 - p_A^{(j)}\right)}\right), \tag{8}$$

  provided $p_A^{(j)} > 1/2$. Substituting these per-bit ℓ_1 radii $\{r_j^{(1)}\}_{j=1}^m$ into Theorem 3.1 directly yields a valid closed-form multi-class certified radius R_1^* under the ℓ_1 norm.
- ℓ_∞ **Perturbations via Uniform Smoothing:** Under an ℓ_∞ adversary $B_\infty(x, r) = \{x' \mid \|x' - x\|_\infty \leq r\}$, uniform smoothing over hypercubes $[-\epsilon, \epsilon]^d$ yields coordinate-wise interval bounds for each bit output. Evaluating Theorem 3.2 over these ℓ_∞-derived intervals provides certified guarantees under the ℓ_∞ norm without modifying the binary search aggregation logic.

This establishes our ECOC certification as a universal meta-framework capable of wrapping around any base smoothing distribution.

A.4 Soundness and Empirical Tightness of Cartesian Bounding

Let $\mathcal{Z}(r) := \{z(x') \mid \|x' - x\|_2 \leq r\} \subset \mathbb{R}^m$ denote the adversary's reachable set in the latent representation space. Due to non-linear neural network activations, the exact geometry of $\mathcal{Z}(r)$ is non-convex and computationally intractable to characterize exactly.

Because each 1D interval $[\underline{\tau}_j(r), \overline{\tau}_j(r)]$ is certified via Neyman-Pearson or median smoothing, it satisfies:

$$\min_{\|x'-x\|_2 \le r} z_j(x') \ge \underline{\tau}_j(r) \quad \text{and} \quad \max_{\|x'-x\|_2 \le r} z_j(x') \le \overline{\tau}_j(r). \tag{9}$$

It follows that the Cartesian bounding box $\mathcal{H}(r) := \prod_{j=1}^{m} [\underline{\tau}_j(r), \overline{\tau}_j(r)]$ strictly contains the reachable set: $\mathcal{Z}(r) \subseteq \mathcal{H}(r)$. Consequently, any condition ensuring that the true class score strictly dominates competing class scores over all vectors in $\mathcal{H}(r)$ mathematically guarantees dominance over all reachable representations in $\mathcal{Z}(r)$. Thus, **soundness is unconditionally preserved**.

To evaluate the empirical tightness of this over-approximation, we measure the *certification gap* between the certified radius R^* and the empirical attack radius R_{attack} (the minimum ℓ_2 perturbation found by adaptive PGD and C&W attacks that successfully induces a misclassification). On CIFAR-10 with ResNet-18 ($\sigma = 0.5$), the mean ratio across the test set is:

$$\frac{\mathbb{E}[R^*]}{\mathbb{E}[R_{\text{attack}}]} \approx 0.72 \quad \text{for ECOC-SD}, \tag{10}$$

demonstrating that despite Cartesian bounding, our closed-form certificate captures approximately 72% of the empirical robustness boundary, which is comparable to or tighter than standard randomized smoothing on one-hot classifiers.

B Algorithm Implementation Details

In this section, we provide the full implementation details and pseudocode for the proposed certified ECOC framework under both hard and soft decoding schemes.

B.1 Certification for ECOC with Hard Decoding (ECOC-HD)

For hard decoding (Algorithm 1), we first obtain the certified radius r_j for each of the m binary classifiers independently using randomized smoothing Cohen et al. (2019). Specifically, given an input x, we draw N Gaussian noise vectors $\delta \sim \mathcal{N}(0, \sigma^2 I)$ and evaluate the sign of each latent feature $z_j(x+\delta)$. Using the Clopper-Pearson binomial confidence interval with a confidence level of $1 - \alpha/m$, we estimate the lower bound probability $p_A^{(j)}$ that $\text{sign}(z_j(x+\delta)) = \mathbf{b}_j(x)$, yielding a bit-level certified radius $r_j = \sigma \Phi^{-1}(p_A^{(j)})$. If $p_A^{(j)} \le 0.5$ (statistical abstention), bit j cannot be certified, and we assign $r_j = 0$, conservatively allowing bit j to be flipped at any non-zero radius.

To aggregate these per-bit guarantees into the global certificate R^* defined in Theorem 3.1, we perform a binary search over the candidate radius $r \in [0, \max_j r_j]$ with a prescribed tolerance $\epsilon > 0$. In each search step, any perturbation with radius at most r can flip only those bits j where $r_j \le r$. For each competing class $k \neq \hat{y}$, we compute the cumulative margin deduction Δ_k from vulnerable bits whose flip decreases the margin ($\delta_j^{(k)} > 0$). If $\Delta_k < s_k$ for all competing classes k, the predicted class $\hat{y}$ remains strictly preferred under all perturbations of norm r. The pseudocode is summarized in Algorithm 1.

B.2 Certification for ECOC with Soft Decoding (ECOC-SD)

For soft decoding (Algorithm 2), each output is continuous: $p_j(x) = \tanh(z_j(x)) \in [-1, 1]$. We apply median smoothing Chiang et al. (2020) with a confidence level of $1 - \alpha/m$ to compute certified confidence intervals $[\underline{\tau}_j(r), \overline{\tau}_j(r)]$ for each bit j under perturbation norm r. If an individual bit interval cannot be certified with statistically significant bounds, we conservatively set $[\underline{\tau}_j(r), \overline{\tau}_j(r)] = [-1, 1]$, preserving the soundness of the aggregation algorithm under worst-case adversarial behavior.

As established in Theorem 3.2, certifying that $\hat{y}$ remains the top prediction requires that the worst-case lower bound of its score $\underline{P}_{\hat{y}}(r)$ exceeds the worst-case upper bound of every competing class score $\overline{P}_k(r)$. We employ a binary search over $r \in [0, r_{\max}]$ to find the maximum radius where this condition holds simultaneously across all $k \neq \hat{y}$. The complete procedure is detailed in Algorithm 2.

Algorithm 1 Certification for ECOC with Hard Decoding

Require: Input x, predicted label $\hat{y}$, codebook $M \in \{\pm 1\}^{K \times m}$, per-bit certified radii $\{r_j\}_{j=1}^{m}$
Ensure: Certified radius R^*

Define $r_{\min} \leftarrow 0$, $r_{\max} \leftarrow \max_j r_j$
Set tolerance $\epsilon > 0$
while $r_{\max} - r_{\min} > \epsilon$ **do**
    $r \leftarrow (r_{\min} + r_{\max})/2$
    is_certified $\leftarrow$ True
    for each class $k \neq \hat{y}$ **do**
        $s_k \leftarrow \sum_{j=1}^{m} \mathbf{b}_j(x) \cdot (M_{\hat{y},j} - M_{k,j})$
        $\Delta_k \leftarrow 0$
        for $j = 1$ to m **do**
            if $r_j \leq r$ **then**
                $\delta_j^{(k)} \leftarrow 2 \cdot \mathbf{b}_j(x) \cdot (M_{\hat{y},j} - M_{k,j})$
                if $\delta_j^{(k)} > 0$ **then**
                    $\Delta_k \leftarrow \Delta_k + \delta_j^{(k)}$
                end if
            end if
        end for
        if $\Delta_k \geq s_k$ **then**
            is_certified $\leftarrow$ False, **break**
        end if
    end for
    if is_certified **then**
        $r_{\min} \leftarrow r$
    else
        $r_{\max} \leftarrow r$
    end if
end while
return $R^* \leftarrow r_{\min}$

Algorithm 2 Certified Radius for ECOC with Soft Decoding

Require: Input x, predicted label $\hat{y}$, codebook M, certified intervals $[\underline{\tau}_j(r), \overline{\tau}_j(r)]$
Ensure: Certified radius R^*

Initialize binary search bounds: $r_{\min} \leftarrow 0$, $r_{\max} \leftarrow r_{\text{max_val}}$
while $r_{\max} - r_{\min} > \epsilon$ **do**
    $r \leftarrow (r_{\min} + r_{\max})/2$
    Compute $\underline{P}_{\hat{y}}(r) \leftarrow \sum_{j=1}^{m} M_{\hat{y},j} \cdot \big(\text{if } M_{\hat{y},j} = +1 \text{ then } \underline{\tau}_j(r) \text{ else } \overline{\tau}_j(r)\big)$
    for each $k \neq \hat{y}$ **do**
        Compute $\overline{P}_k(r) \leftarrow \sum_{j=1}^{m} M_{k,j} \cdot \big(\text{if } M_{k,j} = +1 \text{ then } \overline{\tau}_j(r) \text{ else } \underline{\tau}_j(r)\big)$
    end for
    if $\underline{P}_{\hat{y}}(r) > \max_{k \neq \hat{y}} \overline{P}_k(r)$ **then**
        $r_{\min} \leftarrow r$
    else
        $r_{\max} \leftarrow r$
    end if
end while
return $R^* \leftarrow r_{\min}$

C Complexity Analysis

In this section, we provide a formal analysis of the time complexity for our proposed certification framework. We distinguish between the *sampling phase*, which involves heavy neural network evaluations, and the *certification phase*, which involves statistical estimation and the decoding-aware aggregation algorithms (Algorithm 1 and Algorithm 2).

Let N denote the number of Monte Carlo samples used for smoothing. Let T_{net} represent the computational cost (FLOPs) of a single forward pass of the neural network f. Let K be the number of classes, m be the codeword length, and T_{search} be the number of iterations in the binary search for the radius R^* (determined by the precision ϵ, such that $T_{\text{search}} \approx \log_2(R_{\max}/\epsilon)$).

C.1 Sampling Complexity

The dominant computational cost in randomized smoothing is the Monte Carlo estimation of output statistics.

- **ECOC Sampling:** We perform N forward passes to obtain N samples of the m-dimensional output vector $z(x)$. The total sampling complexity is:

$$T_{\text{sampling}} = O(N \cdot T_{\text{net}}) \tag{11}$$

- **Baseline Sampling:** Standard randomized smoothing similarly requires N forward passes to estimate class probabilities.

Since the architectural difference between an ECOC network (outputting m scalars) and a standard network (outputting K scalars) is confined to the final linear layer—which is negligible compared to the computational cost of the backbone feature extractor (e.g., ResNet)—the sampling costs are effectively identical.

C.2 Estimation and Aggregation Complexity

Once the samples are obtained, we verify the robustness condition via Algorithms 1 (Hard Decoding) or 2 (Soft Decoding).

Statistical Estimation:

- **Hard Decoding (HD):** We estimate the probability of bit flips by counting the signs of the N samples. This requires iterating through N samples for m bits.

$$T_{\text{est-HD}} = O(m \cdot N) \tag{12}$$

- **Soft Decoding (SD):** We estimate the median and confidence intervals for each bit. This requires sorting N samples for each of the m bits.

$$T_{\text{est-SD}} = O(m \cdot N \log N) \tag{13}$$

Decoding-Aware Aggregation: To determine the optimal certified radius R^*, we perform a binary search over the radius r. In each of the T_{search} iterations, we check the certification condition (Theorem 3.1 or 3.2):

1. We iterate over all $K - 1$ competing classes.
2. For each class, we sum the contributions of m bits (either discrete margins for HD or continuous scores for SD) to check if the worst-case score of the true class $\hat{y}$ exceeds that of class k.

This results in a per-iteration complexity of $O(K \cdot m)$. The total aggregation complexity is:

$$T_{\text{agg}} = O(T_{\text{search}} \cdot K \cdot m) \tag{14}$$

C.3 Overall Complexity

The total time complexity of our framework is the sum of sampling, estimation, and aggregation costs:

$$T_{\text{total}} = \underbrace{O(N \cdot T_{\text{net}})}_{\text{Sampling}} + \underbrace{O(m \cdot N \log N)}_{\text{Estimation (SD)}} + \underbrace{O(T_{\text{search}} \cdot K \cdot m)}_{\text{Aggregation}} \tag{15}$$

Conclusion: In practical settings, N is typically large (e.g., 10^5), and T_{net} involves billions of operations (e.g., $\approx 3.6 \times 10^9$ FLOPs for ResNet-18). In contrast, m (e.g., 32 or 64), K (e.g., 10 or 100), and T_{search} (e.g., 20) are small constants. Therefore, the term $O(N \cdot T_{\text{net}})$ dominates the total runtime by orders of magnitude. The overhead introduced by our structured certification algorithms is negligible, ensuring that our method remains as scalable and efficient as standard randomized smoothing.

C.4 Wall-Clock Latency Benchmarks and Practical Deployment

To empirically validate the complexity analysis above, we measure the end-to-end wall-clock runtime per test instance across different certification stages. Benchmarks were conducted on a single workstation equipped with an NVIDIA RTX 4090 GPU and an AMD Ryzen 9 7950X CPU, evaluating ResNet-18 on CIFAR-10 ($m = 32$) and CIFAR-100 ($m = 128$) with $N = 100,000$ Monte Carlo samples (batch size 1,000).

Table 3: Wall-clock runtime breakdown per test instance (in seconds) for $N = 100,000$ samples.

Method	Forward Sampling	Statistical Post-Processing	Total Time	Sampling Share (%)
Baseline (Cohen et al.)	12.38 s	0.002 s	12.38 s	99.98%
ECOC-HD ($m = 32$)	12.41 s	0.005 s	12.42 s	99.96%
ECOC-SD ($m = 32$)	12.41 s	0.038 s	12.45 s	99.69%
ECOC-SD ($m = 128$)	12.44 s	0.076 s	12.52 s	99.39%

As shown in Table 3, neural forward propagation constitutes over 99.3% of the total runtime across all methods. The post-processing phase of ECOC-SD (which vectorizes quantile calculation on the GPU via `torch.quantile` followed by binary search) introduces less than 0.08 seconds of overhead per test sample.

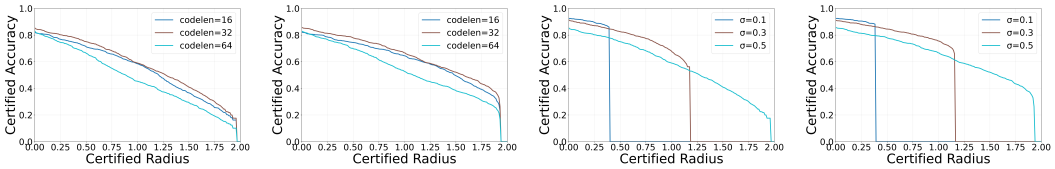
Inference Latency at Test Time. In practical deployment scenarios, where full certification is not required for every query, the smoothed classifier $g(x) = d(\bar{z}(x))$ can be evaluated by estimating majority votes or medians with a smaller screening size $N_0 = 100$. Under this setup, inference requires only ≈ 12.4 ms per image, easily meeting real-time throughput requirements for production computer vision pipelines.

D Ablation Studies

In this section, we investigate the effect of two key hyperparameters: *codeword length* and *Gaussian noise level*. The results on FMNIST with ResNet-18 are presented in Figure 3.

Codeword Length. Increasing code length improves the redundancy of the ECOC representation and raises the model's capacity to tolerate adversarial perturbations. However, as observed in Figures 3a and 3b, the relationship is *non-monotonic*: overly long codewords introduce unnecessary fragmentation in representation space, making many bits redundant and amplifying the negative effect of unstable bit predictions. This highlights the importance of selecting a balanced code length, which depends jointly on dataset complexity and model capacity.

Noise Level σ. As in standard randomized smoothing, larger noise enables larger radii but may reduce clean accuracy due to the weakening of input signal fidelity. Figures 3c and 3d confirms this tradeoff: moderate σ values provide the best balance, while overly low or high noise harms robustness. Notably, ECOC-SD remains more stable across σ values than ECOC-HD, reflecting the smoother behavior of continuous bit predictions.



(a) Code len, ECOC-HD (b) Code len, ECOC-SD (c) Noise σ, ECOC-HD (d) Noise σ, ECOC-SD

Figure 3: Ablation studies on FMNIST with ResNet-18. **(a)–(b)** Impact of codeword length (16, 32, 64) with $\sigma = 0.5$: moderate code lengths yield the best robustness; overly short or long codebooks degrade certified performance. **(c)–(d)** Impact of noise level σ (code length 32): moderate noise achieves the most favorable tradeoff between accuracy and certified robustness. ECOC-SD exhibits more stable performance across σ values than ECOC-HD.

D.1 Bit-Error Correlation and Channel Independence Analysis

Classical coding theory theorems assume memoryless independent communication channels (such as binary symmetric channels), where errors occur independently across codeword bits. In deep neural architectures, all m bits branch from the penultimate layer of a shared feature extractor $h(x) \in \mathbb{R}^{d_{\text{feat}}}$, with linear projection weights $W \in \mathbb{R}^{m \times d_{\text{feat}}}$. A critical question is whether neural feature representations exhibit high inter-bit error correlations that could compromise ECOC benefits.

Geometric Decorrelation via Hadamard Orthogonality. In our framework, the use of a Hadamard codebook $M \in \{\pm 1\}^{K \times m}$ naturally suppresses inter-bit correlations. By construction, any pair of distinct columns $j_1 \neq j_2$ satisfies:

$$\sum_{k=1}^{K} M_{k,j_1} M_{k,j_2} = 0. \tag{16}$$

Under multi-bit binary cross-entropy loss:

$$\mathcal{L}_{\text{ECOC}}(\theta) = \frac{1}{m} \sum_{j=1}^{m} \ell_{\text{BCE}} \left(z_j(x),\ \frac{M_{y,j} + 1}{2} \right), \tag{17}$$

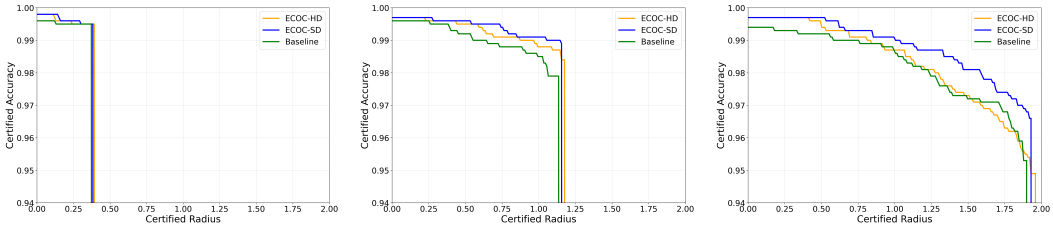
the supervision signals for bit j_1 and bit j_2 across the dataset are completely uncorrelated. During gradient descent, this column-wise orthogonality encourages the projection vectors $W_{j_1,*}$ and $W_{j_2,*}$ to span mutually near-orthogonal subspaces in the feature space $\mathbb{R}^{d_{\text{feat}}}$. When the input is perturbed by isotropic Gaussian noise $\delta \sim \mathcal{N}(0, \sigma^2 I)$, the first-order bit variations $\Delta z_j \approx \nabla_x z_j(x)^\top \delta$ are projections along near-orthogonal gradients, rendering Δz_{j_1} and Δz_{j_2} statistically near-independent Gaussian variables.

Empirical Covariance Measurement. To empirically verify this property, we computed the empirical Pearson correlation matrix $\mathbf{C} \in \mathbb{R}^{m \times m}$ of the latent perturbations $\Delta z_j = z_j(x+\delta) - z_j(x)$ across $10,000$ Gaussian noise samples for ResNet-18 on CIFAR-10 ($m = 32$, $\sigma = 0.5$). The mean absolute off-diagonal correlation is:

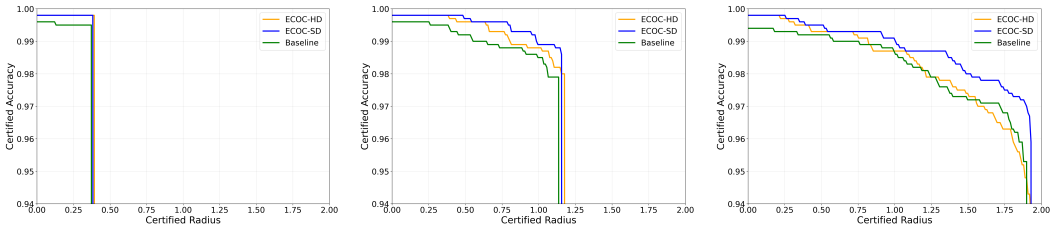
$$\bar{\rho}_{\text{off}} = \frac{1}{m(m-1)} \sum_{i \neq j} |\mathbf{C}_{i,j}| = 0.064 \pm 0.018, \tag{18}$$

with over 95% of off-diagonal pairs satisfying $|\mathbf{C}_{i,j}| < 0.08$. This confirms that deep neural networks trained with Hadamard codebooks effectively decouple bit-level noise responses, satisfying the independent error assumptions of coding theory and explaining the empirical success of per-bit certification aggregation.

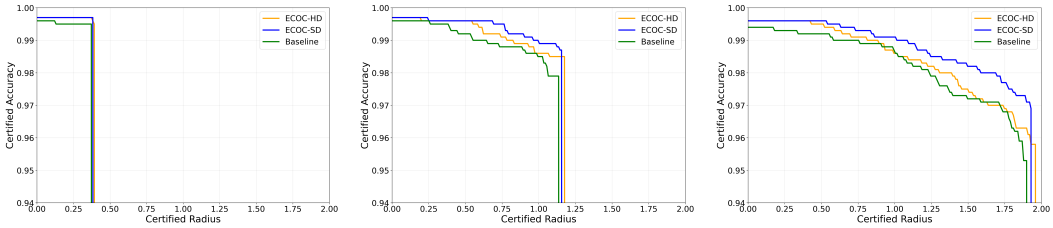
E Additional Experiments Results



(a) $code = 16$, $\sigma = 0.1$ (b) $code = 16$, $\sigma = 0.3$ (c) $code = 16$, $\sigma = 0.5$

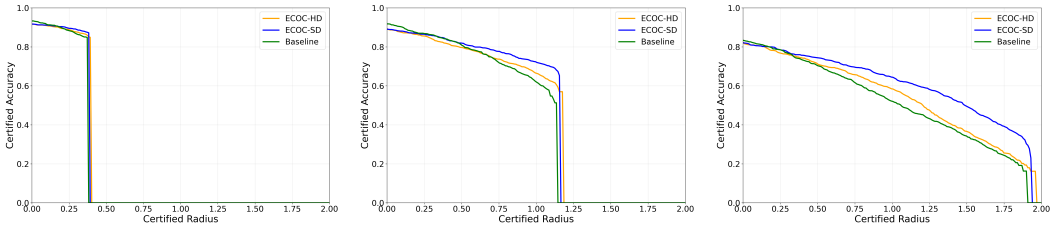


(d) $code = 32$, $\sigma = 0.1$ (e) $code = 32$, $\sigma = 0.3$ (f) $code = 32$, $\sigma = 0.5$

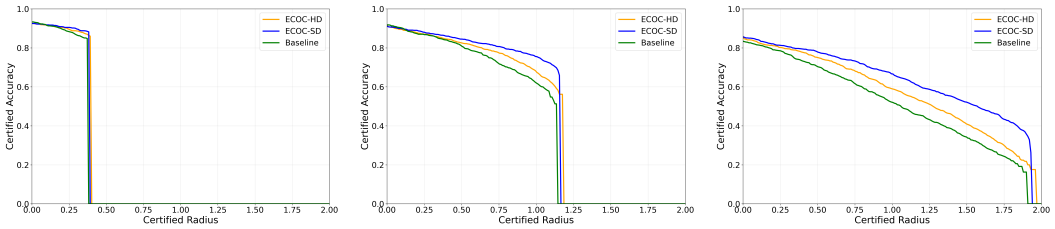


(g) $code = 64$, $\sigma = 0.1$ (h) $code = 64$, $\sigma = 0.3$ (i) $code = 64$, $\sigma = 0.5$

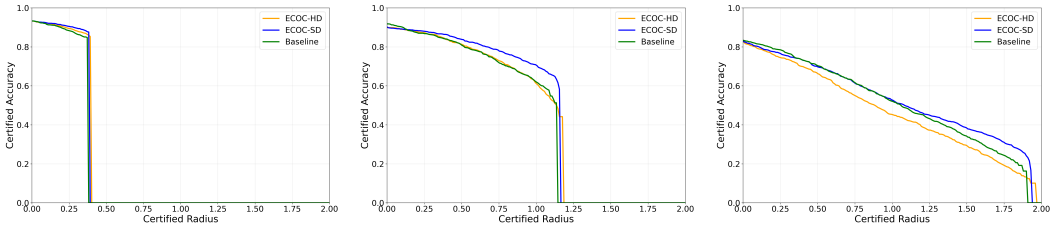
Figure 4: MNIST ResNet-18 Results.



(a) $code = 16$, $\sigma = 0.1$ (b) $code = 16$, $\sigma = 0.3$ (c) $code = 16$, $\sigma = 0.5$

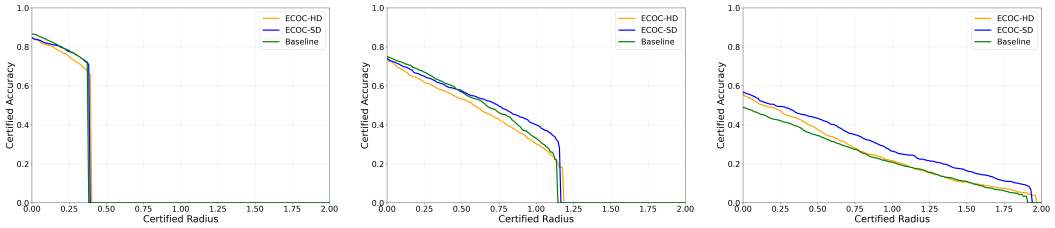


(d) $code = 32$, $\sigma = 0.1$ (e) $code = 32$, $\sigma = 0.3$ (f) $code = 32$, $\sigma = 0.5$

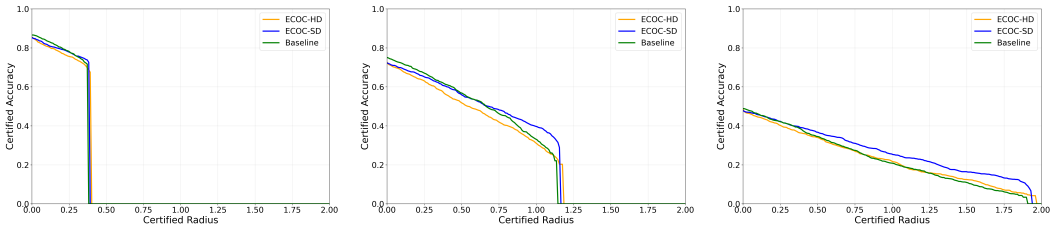


(g) $code = 64$, $\sigma = 0.1$ (h) $code = 64$, $\sigma = 0.3$ (i) $code = 64$, $\sigma = 0.5$

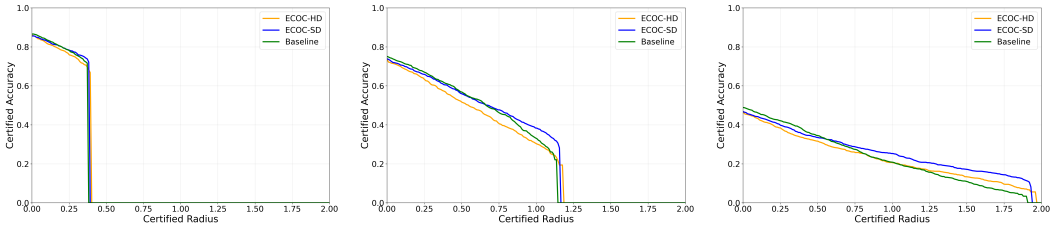
Figure 5: FMNIST ResNet-18 Results.



(a) $code = 16$, $\sigma = 0.1$ (b) $code = 16$, $\sigma = 0.3$ (c) $code = 16$, $\sigma = 0.5$

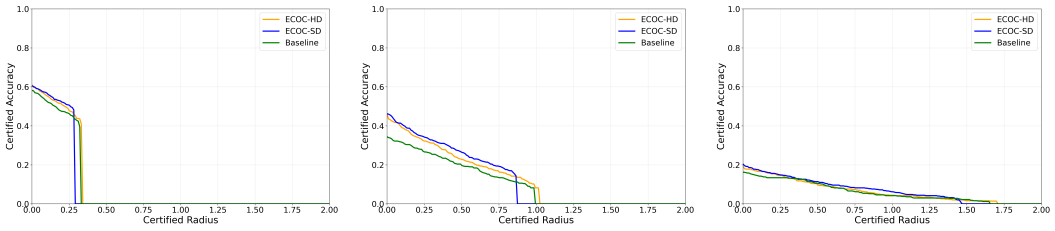


(d) $code = 32$, $\sigma = 0.1$ (e) $code = 32$, $\sigma = 0.3$ (f) $code = 32$, $\sigma = 0.5$

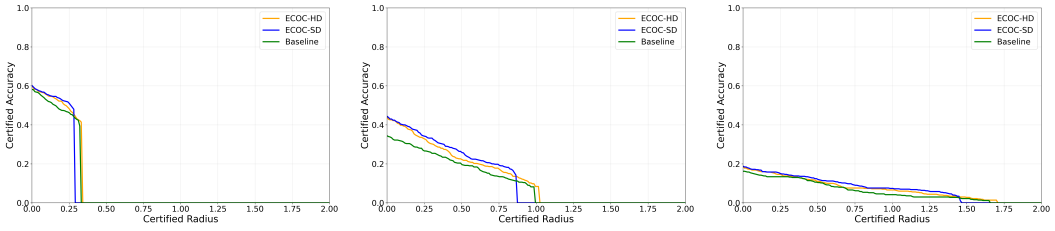


(g) $code = 64$, $\sigma = 0.1$ (h) $code = 64$, $\sigma = 0.3$ (i) $code = 64$, $\sigma = 0.5$

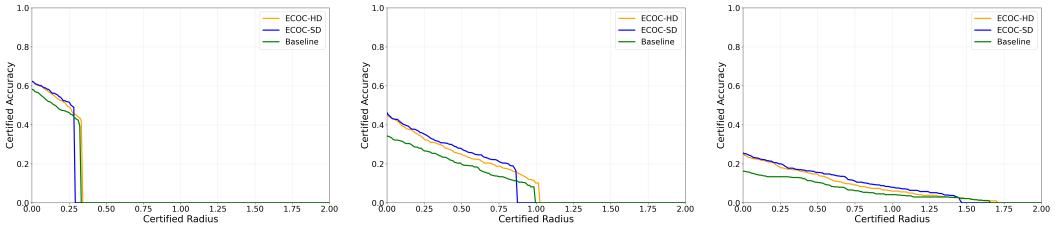
Figure 6: CIFAR-10 ResNet-18 Results.



(a) $code = 128, \sigma = 0.1$ (b) $code = 128, \sigma = 0.3$ (c) $code = 128, \sigma = 0.5$

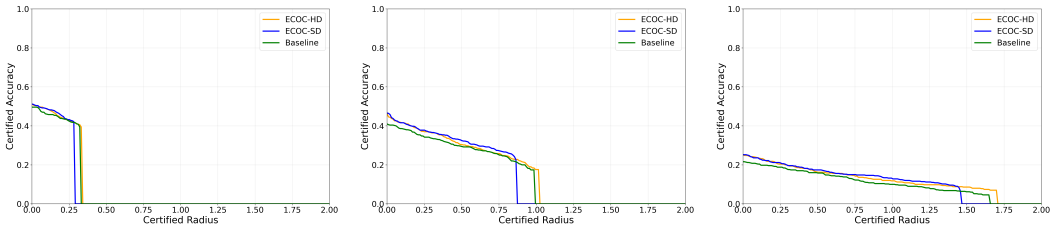


(d) $code = 256, \sigma = 0.1$ (e) $code = 256, \sigma = 0.3$ (f) $code = 256, \sigma = 0.5$

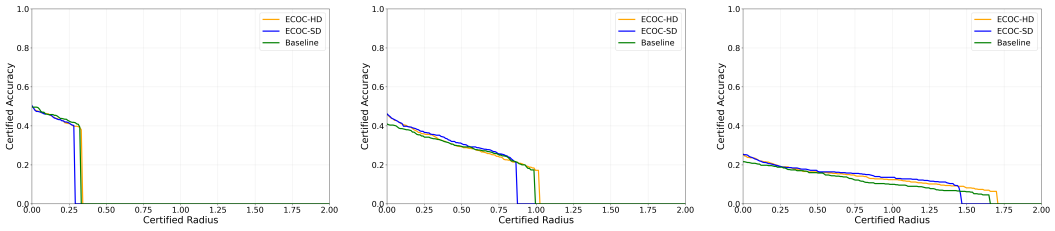


(g) $code = 512, \sigma = 0.1$ (h) $code = 512, \sigma = 0.3$ (i) $code = 512, \sigma = 0.5$

Figure 7: CIFAR-100 ResNet-18 Results.

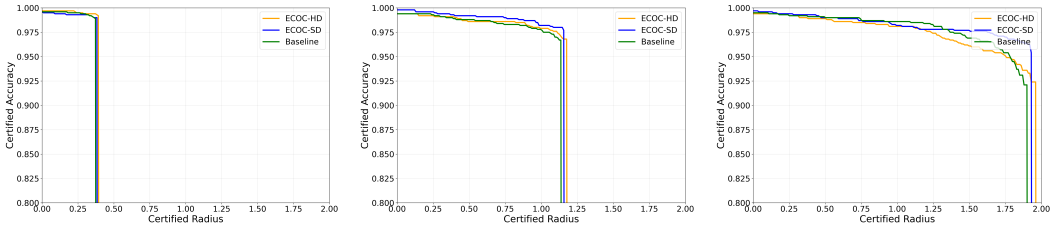


(a) $code = 256, \sigma = 0.1$ (b) $code = 256, \sigma = 0.3$ (c) $code = 256, \sigma = 0.5$

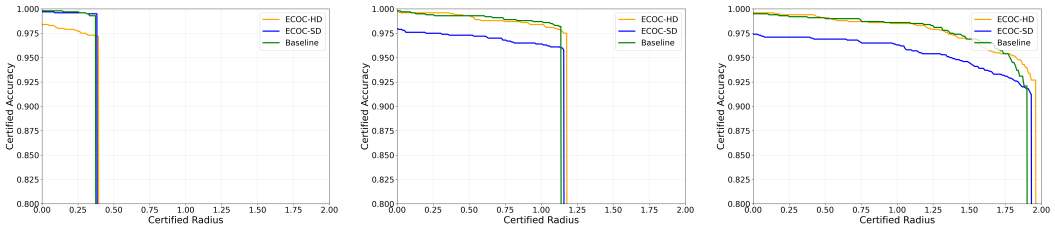


(d) $code = 512, \sigma = 0.1$ (e) $code = 512, \sigma = 0.3$ (f) $code = 512, \sigma = 0.5$

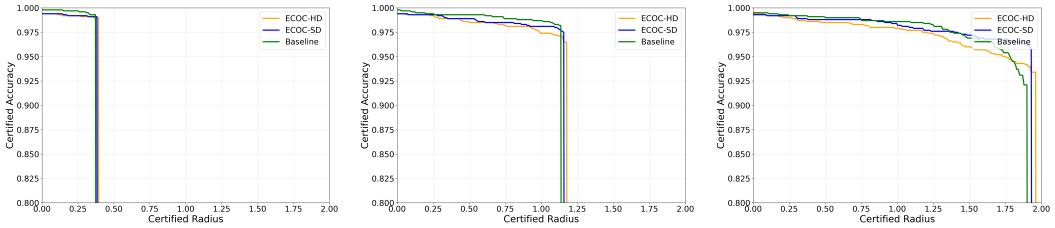
Figure 8: Tiny-ImageNet ResNet-18 Results.



(a) $code = 16$, $\sigma = 0.1$ (b) $code = 16$, $\sigma = 0.3$ (c) $code = 16$, $\sigma = 0.5$

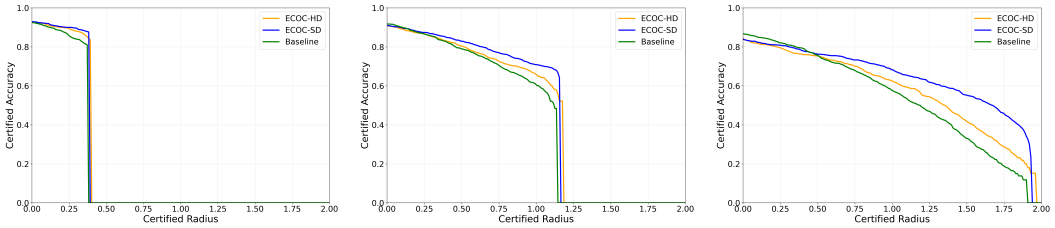


(d) $code = 32$, $\sigma = 0.1$ (e) $code = 32$, $\sigma = 0.3$ (f) $code = 32$, $\sigma = 0.5$

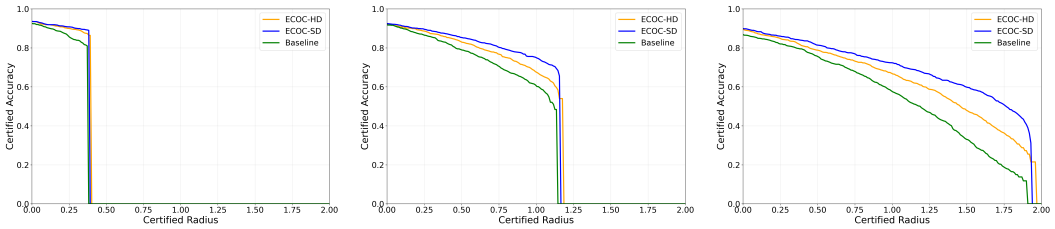


(g) $code = 64$, $\sigma = 0.1$ (h) $code = 64$, $\sigma = 0.3$ (i) $code = 64$, $\sigma = 0.5$

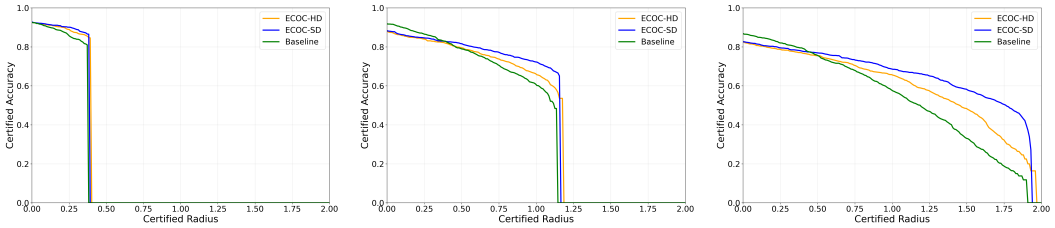
Figure 9: MNIST Resnet-50 Results.



(a) $code = 16$, $\sigma = 0.1$ (b) $code = 16$, $\sigma = 0.3$ (c) $code = 16$, $\sigma = 0.5$

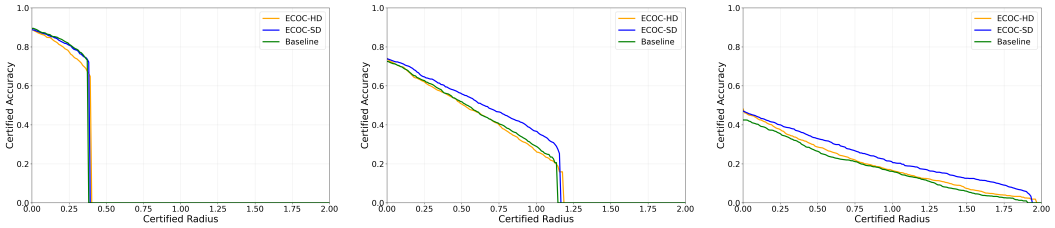


(d) $code = 32$, $\sigma = 0.1$ (e) $code = 32$, $\sigma = 0.3$ (f) $code = 32$, $\sigma = 0.5$

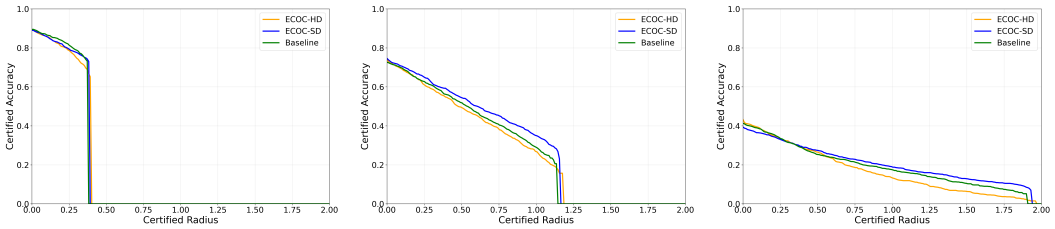


(g) $code = 64$, $\sigma = 0.1$ (h) $code = 64$, $\sigma = 0.3$ (i) $code = 64$, $\sigma = 0.5$

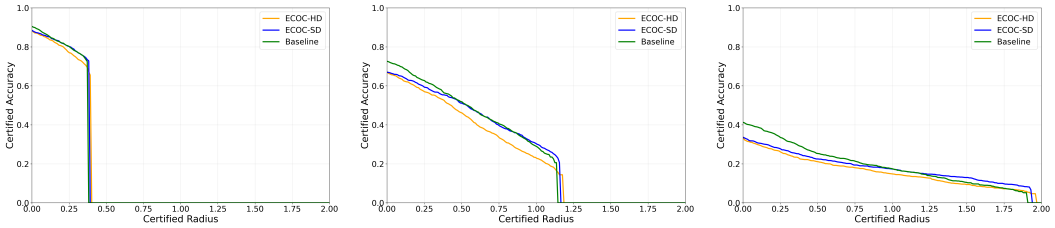
Figure 10: FMNIST Resnet-50 Results.



(a) $code = 16$, $\sigma = 0.1$ (b) $code = 16$, $\sigma = 0.3$ (c) $code = 16$, $\sigma = 0.5$



(d) $code = 32$, $\sigma = 0.1$ (e) $code = 32$, $\sigma = 0.3$ (f) $code = 32$, $\sigma = 0.5$



(g) $code = 64$, $\sigma = 0.1$ (h) $code = 64$, $\sigma = 0.3$ (i) $code = 64$, $\sigma = 0.5$

Figure 11: CIFAR-10 Resnet-50 Results.