Supplement to "Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem"

1. Proofs

Lemma 3.1. *Let $\{Q_i\}_{l=1}^R$ be the set of linear regions associated to the ReLU-classifier $f : \mathbb{R}^d \to \mathbb{R}^K$. For any $x \in \mathbb{R}^d$ there exists $\alpha \in \mathbb{R}$ with $\alpha > 0$ and $t \in \{1, \dots, R\}$ such that $\beta x \in Q_t$ for all $\beta \geq \alpha$.*

Proof. Suppose the statement would be false. Then there exist $\{\beta_i\}_{i=1}^\infty$ with $\beta_i \geq 0$, $\beta_i \geq \beta_j$ if $i \leq j$ and $\beta_i \to \infty$ as $i \to \infty$ such that for $\gamma \in [\beta_i, \beta_{i+1})$ we have $\gamma x \in Q_{r_i}$ with $r_i \in \{1, \dots, R\}$ and $r_{i-1} \neq r_i \neq r_{i+1}$. As there are only finitely many regions there exist $i, j \in \mathbb{N}$ with $i < j$ such that $r_i = r_j$, in particular $\beta_i x \in Q_{r_i}$ and $\beta_j x \in Q_{r_i}$. However, as the linear regions are convex sets also the line segment $[\beta_i x, \beta_j x] \in Q_{r_i}$. However, that implies $\beta_i = \beta_j$ as neighboring segments are in different regions which contradicts the assumption. Thus there can only be finitely many $\{\beta_i\}_{i=1}^M$ and the $\{r_i\}_{i=1}^M$ have to be all different, which finishes the proof. □

Theorem 3.1. *Let $\mathbb{R}^d = \cup_{l=1}^R Q_l$ and $f(x) = V^l x + a^l$ be the piecewise affine representation of the output of a ReLU network on Q_l. Suppose that V^l does not contain identical rows for all $l = 1, \dots, R$, then for almost any $x \in \mathbb{R}^d$ and $\epsilon > 0$ there exists an $\alpha > 0$ and a class $k \in \{1, \dots, K\}$ such that for $z = \alpha x$ it holds*

$$\frac{e^{f_k(z)}}{\sum_{r=1}^K e^{f_r(z)}} \geq 1 - \epsilon.$$

Moreover, $\lim_{\alpha \to \infty} \frac{e^{f_k(\alpha x)}}{\sum_{r=1}^K e^{f_r(\alpha x)}} = 1$.

Proof. By Lemma 3.1 there exists a region Q_t with $t \in \{1, \dots, R\}$ and $\beta > 0$ such that for all $\alpha \geq \beta$ we have $\alpha x \in Q_t$. Let $f(z) = V^t z + a^t$ be the affine form of the ReLU classifier f on Q_t. Let $k^* = \arg\max_k \langle v_k^t, x \rangle$, where v_k^t is the k-th row of V^t. As V^t does not contain identical rows, that is $v_l^t \neq v_m^t$ for $l \neq m$, the maximum is uniquely attained up to a set of measure zero. If the maximum is unique, it holds for sufficiently large $\alpha \geq \beta$

$$\langle v_l^t - v_{k^*}^t, \alpha x \rangle + a_l^t - a_{k^*}^t < 0, \ \forall l \in \{1, \dots, K\} \backslash \{k^*\}. \tag{1}$$

Thus $\alpha x \in Q_t$ is classified as k^*. Moreover,

$$\frac{e^{f_{k^*}(\alpha x)}}{\sum_{l=1}^K e^{f_l(\alpha x)}} = \frac{e^{\langle v_{k^*}^t, \alpha x \rangle + a_k^t}}{\sum_{l=1}^K e^{\langle v_l^t, \alpha x \rangle + a_l^t}} \tag{2}$$

$$= \frac{1}{1 + \sum_{l \neq k^*}^K e^{\langle v_l^t - v_{k^*}^t, \alpha x \rangle + a_l^t - a_k^t}}. \tag{3}$$

By inequality (1) all the terms in the exponential are negative and thus by upscaling α, using $\langle v_{k^*}^t, x \rangle > \langle v_l^t, x \rangle$ for all $l \neq k^*$, we can get the exponential term arbitrarily close to 0. In particular,

$$\lim_{\alpha \to \infty} \frac{1}{1 + \sum_{l \neq k}^K e^{\langle v_l^t - v_{k^*}^t, \alpha x \rangle + a_l^t - a_k^t}} = 1.$$

□

Theorem 3.2. *Let $f_k(x) = \sum_{l=1}^N \alpha_{kl} e^{-\gamma \|x - x_l\|_2^2}$, $k = 1, \dots, K$ be a RBF-network trained with cross-entropy loss on the training data $(x_i, y_i)_{i=1}^N$. We define $r_{\min} = \min_{l=1,\dots,N} \|x - x_l\|_2$ and $\alpha = \max_{r,k} \sum_{l=1}^N |\alpha_{rl} - \alpha_{kl}|$. If $\epsilon > 0$ and*

$$r_{\min}^2 \geq \frac{1}{\gamma} \log\Big(\frac{\alpha}{\log(1 + K\epsilon)}\Big),$$

then for all $k = 1, \dots, K$,

$$\frac{1}{K} - \epsilon \ \leq \ \frac{e^{f_k(x)}}{\sum_{r=1}^K e^{f_r(x)}} \ \leq \ \frac{1}{K} + \epsilon.$$

Proof. It holds $\frac{e^{f_k(x)}}{\sum_{r=1}^K e^{f_r(x)}} = \frac{1}{\sum_{r=1}^K e^{f_r(x)-f_k(x)}}$. With

$$|f_r(x) - f_k(x)| = \Big|\sum_l (\alpha_{rl} - \alpha_{kl}) e^{-\gamma\|x-x_l\|_2^2}\Big| \tag{4}$$

$$\leq e^{-\gamma r_{\min}^2} \sum_l |\alpha_{rl} - \alpha_{kl}| \tag{5}$$

$$\leq e^{-\gamma r_{\min}^2} \alpha \leq \log(1 + K\epsilon), \tag{6}$$

where the last inequality follows by the condition on $r_{\min}$. We get

$$\frac{1}{\sum_{r=1}^K e^{f_r(x)-f_k(x)}} \geq \frac{1}{\sum_{r=1}^K e^{|f_r(x)-f_k(x)|}} \tag{7}$$

$$\geq \frac{1}{K e^{\alpha e^{-\gamma r_{\min}^2}}} \tag{8}$$

$$\geq \frac{1}{K}\frac{1}{1+K\epsilon} \geq \frac{1}{K} - \epsilon, \tag{9}$$

where we have used in the third inequality the condition on $r_{\min}^2$ and in the last step we use $1 \geq (1-K\epsilon)(1+K\epsilon) = 1 - K^2\epsilon^2$. Similarly, we get

$$\frac{1}{\sum_{r=1}^K e^{f_r(x)-f_k(x)}} \leq \frac{1}{\sum_{r=1}^K e^{-|f_r(x)-f_k(x)|}}$$

$$\leq \frac{1}{K e^{-\alpha e^{-\gamma r_{\min}^2}}}$$

$$\leq \frac{1}{K}(1 + K\epsilon) \leq \frac{1}{K} + \epsilon.$$

This finishes the proof. □

2. The effect of Adversarial Confidence Enhanced Training

In this section we compare predictions of the plain model trained on MNIST (Figure 1) and the model trained with ACET (Figure 2). We analyze the images that receive the lowest maximum confidence on the original dataset (MNIST), and the highest maximum confidence on the two datasets that were used for evaluation (EMNIST, grayCIFAR-10).

Evaluated on MNIST: We observe that for both models the lowest maximum confidence corresponds to hard input images that are either discontinous, rotated or simply ambiguous.

Evaluated on EMNIST: Note that some handwritten letters from EMNIST, e.g. 'o' and 'i' may look exactly the same as digits '0' and '1'. Therefore, one should not expect that an ideal model assigns uniform confidences to all EMNIST images. For Figure 1 and Figure 2 we consider predictions on letters that in general do not look exactly like digits ('a', 'b', 'c', 'd'). We observe that the images with the highest maximum confidence correspond to the handwritten letters that *resemble* digits, so the predictions of both models are justified.

Evaluated on Grayscale CIFAR-10: This dataset consists of the images that are clearly distinct from digits. Thus, one can expect uniform confidences on such images, which is achieved by the ACET model (Table 1), but not with the plain model. The mean maximum confidence of the ACET model is close to 10%, with several individual images that are scored with up to 40.41% confidence. Note, that this is much better than for the plain model, which assigns up to 99.60% confidence for the images that have nothing to do with digits. This result is particularly interesting, since the ACET model has not been trained on grayCIFAR-10 examples, and yet it shows much better confidence calibration for out-of-distribution samples.

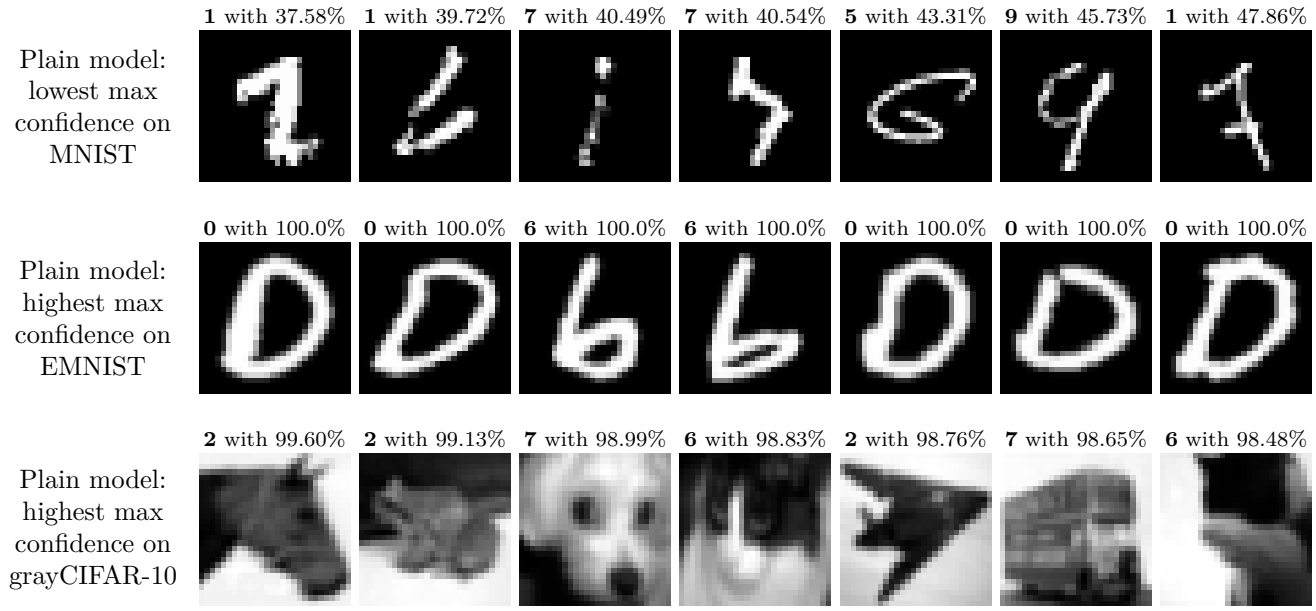

Figure 1: Top Row: predictions of the plain MNIST model with the lowest maximum confidence. Middle Row: predictions of the plain MNIST model on letters 'a', 'b', 'c', 'd' of EMNIST with the highest maximum confidence. Bottom Row: predictions of the plain MNIST model on the grayscale version of CIFAR-10 with the highest maximum confidence. Note that although the predictions on EMNIST are mostly justified, the predictions on CIFAR-10 are overconfident on the images that have no resemblance to digits.

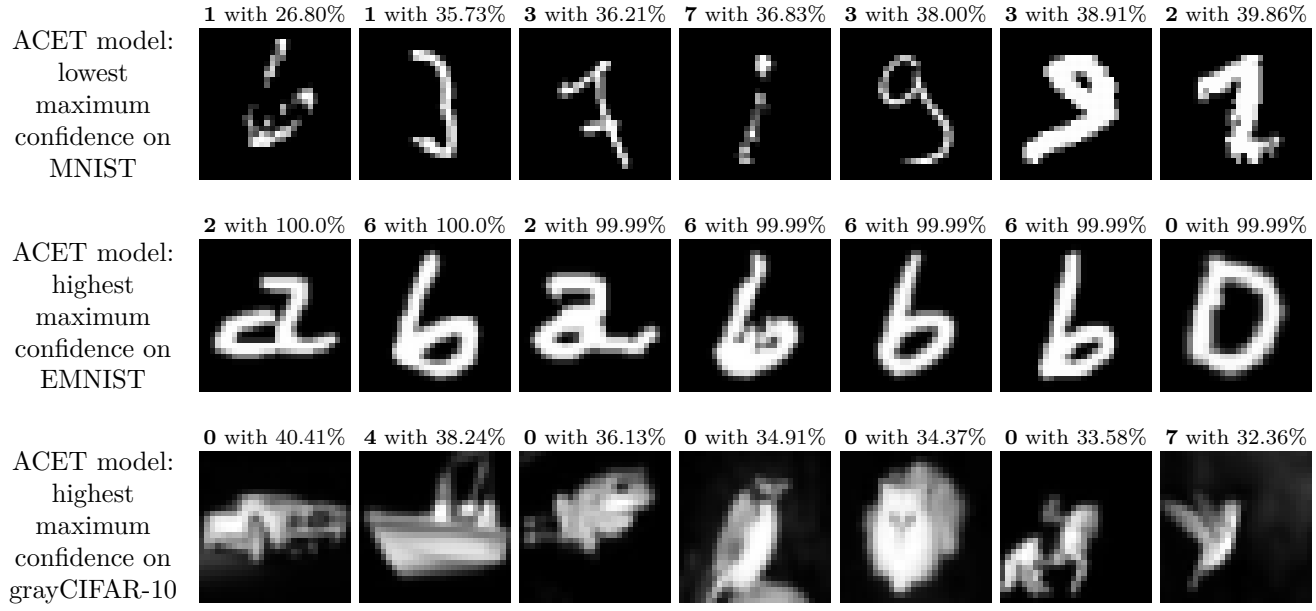

Figure 2: Top Row: predictions of the ACET MNIST model with the lowest maximum confidence. Middle Row: predictions of the ACET MNIST model on letters 'a', 'b', 'c', 'd' of EMNIST with the highest maximum confidence. Bottom Row: predictions of the ACET MNIST model on the grayscale version of CIFAR-10 with the highest maximum confidence. Note that for the ACET model the predictions on both EMNIST and grayCIFAR-10 are now justified.

3. ROC curves

We show the ROC curves for the binary classification task of separating *True* (in-distribution) images from *False* (out-distribution) images. These correspond to the AUROC values (area under the ROC curve) reported in Table 1 in the main paper. As stated in the paper the separation of in-distribution from out-distribution is done by thresholding the maximal confidence value over all classes taken from the original multi-class problem. Note

that the ROC curve shows on the vertical axis the True Positive Rate (TPR), and the horizontal axis is the False Positive Rate (FPR). Thus the FPR@95%TPR value can be directly read off from the ROC curve as the FPR value achieved for 0.95 TPR. Note that a value of 1 of AUROC corresponds to a perfect classifier. A value below 0.5 means that the ordering is reversed: out-distribution images achieve on average higher confidence than the in-distribution images. The worst case is an AUROC of zero, in which case all out-distribution images achieve a higher confidence value than the in-distribution images.

ROC curves for the models trained on MNIST

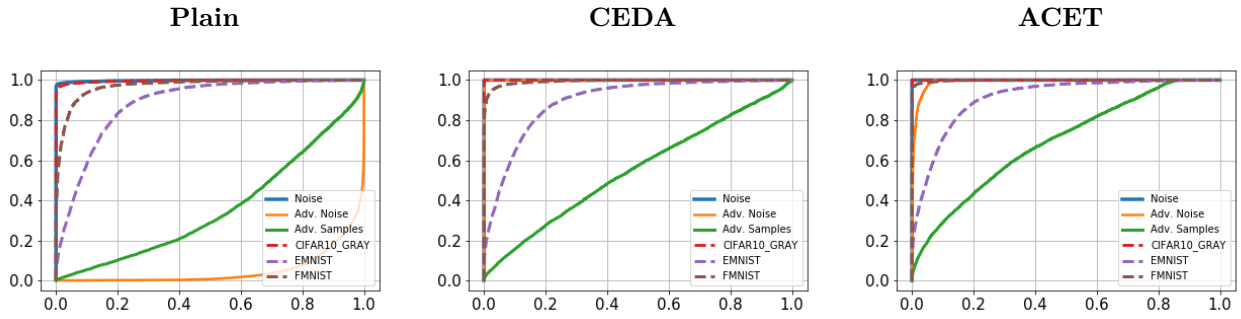

Figure 3: ROC curves of the MNIST models on the evaluation datasets.

In the ROC curves for the plain, CEDA and ACET models for MNIST that are presented in Figure 3, the different grades of improvements for the six evaluation datasets can be observed. For noise, in the plain model the curve is quite far away from the upper left corner, while for the models trained with CEDA and ACET, it reaches that corner, which is the ideal case. For adversarial noise, the plain model is worse than a random classifier, which manifests itself in the fact that the ROC curve runs below the diagonal. While CEDA is better, ACET achieves the ideal result here as well.

ROC curves for the models trained on SVHN

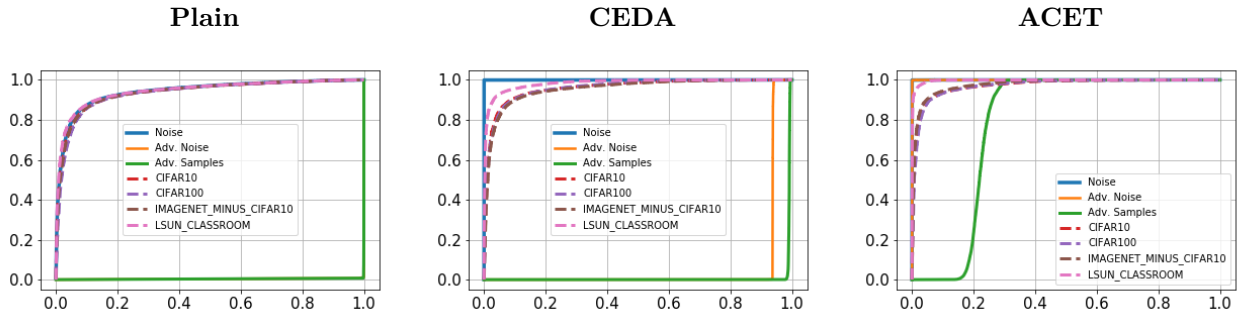

Figure 4: ROC curves of the SVHN models on the evaluation datasets.

CEDA and ACET outperform significantly plain training. While CEDA and ACET perform similar on CIFAR-10, LSUN and noise, ACET outperforms CEDA clearly on adversarial noise and adversarial samples.

ROC curves for the models trained on CIFAR-10

The ROC curves for CIFAR10 show that this dataset is harder than MNIST or SVHN. While CEDA and ACET improve on SVHN, the difference is small. For LSUN even plain training is slightly better (only time for all three datasets). However, on noise and adversarial noise ACET outperforms all other methods.

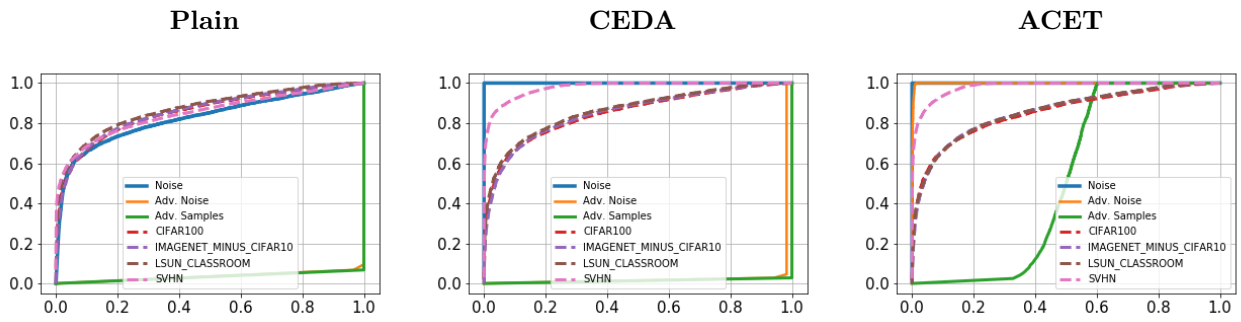

Figure 5: ROC curves of the CIFAR-10 models on the evaluation datasets.

ROC curves for the models trained on CIFAR-100

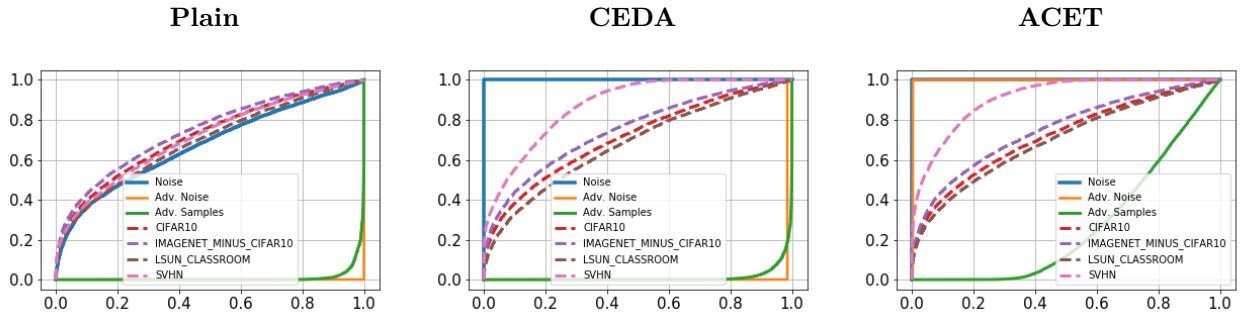

Figure 6: ROC curves of the CIFAR-100 models on the evaluation datasets.

4. Histograms of confidence values

As the AUROC or the FPR@95%TPR just tell us how well the confidence values of in-distribution and out-distribution are ordered, we also report the histograms of achieved confidence values on the original dataset (in-distribution) on which it was trained and the different evaluation datasets. The histograms show how many times the maximum confidence for an image had a certain value between minimal possible 0.1 (0.01 for CIFAR-100) and maximal possible 1.0, for the test set. They give a more detailed picture than the single numbers for mean maximum confidence, area under ROC and FPR@95% TPR. As visible in the top row, the confidence values for clean MNIST test images don't change significantly for CEDA resp. ACET.

4.1. Histograms of confidence values for models trained on MNIST

As visible in the top row of Figure 7, the confidence values for clean MNIST test images don't change significantly for CEDA resp. ACET. For FMNIST, gray CIFAR-10 and Noise inputs, the maximum confidences of CEDA are generally shifted to lower values, and those of ACET even more so. For EMNIST, the same effect is observable, though much weaker due to the similarity of characters and digits. For adversarial noise, CEDA is very successful in lowering the confidences, with most predictions around 10% confidence. As discussed in the main paper, CEDA is not very beneficial for adversarial images, while ACET lowers its confidence to an average value of 85.4% here.

4.2. Histograms of confidence values for models trained on SVHN

Figure 8 shows that both CEDA and ACET assign lower confidences to the out-of-distribution samples from SVHN house numbers and LSUN classroom examples. CEDA and ACET also improve on noise samples. While a large fraction of adversarial samples/noise still achieve high confidence values, our ACET trained model is the only one that lowers the confidences for adversarial noise and adversarial SVHN samples significantly.

4.3. Histograms of confidence values for models trained on CIFAR-10

In Figure 9, CEDA and ACET lower significantly the confidence on noise, and ACET shows improvement for adversarial noise, which fools the plain and CEDA models completely. For CIFAR-10 all models yield very high confidence values on adversarial images. Compared to MNIST ACET leads only to a very small change.

4.4. Histograms of confidence values for models trained on CIFAR-100

In Figure 10, we see similar results to the other datasets. It is noticable in the histograms that for adversarial noise, the deployed attack either achieves 100% confidence or no improvement noise at all. For CEDA, the attack succeeds in most cases, and for ACET only rarely.

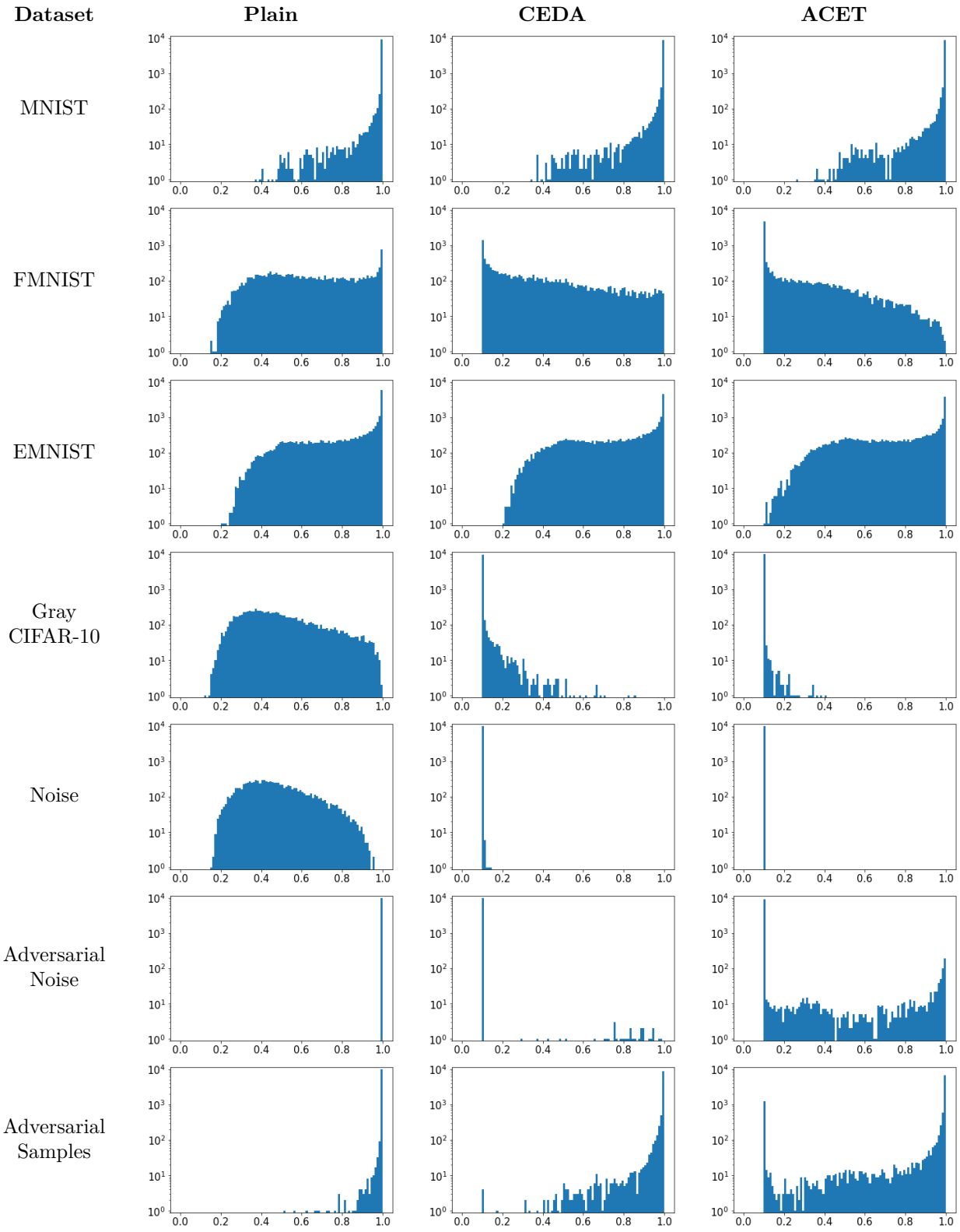

Figure 7: Histograms (logarithmic scale) of maximum confidence values of the three compared models for MNIST on various evaluation datasets.

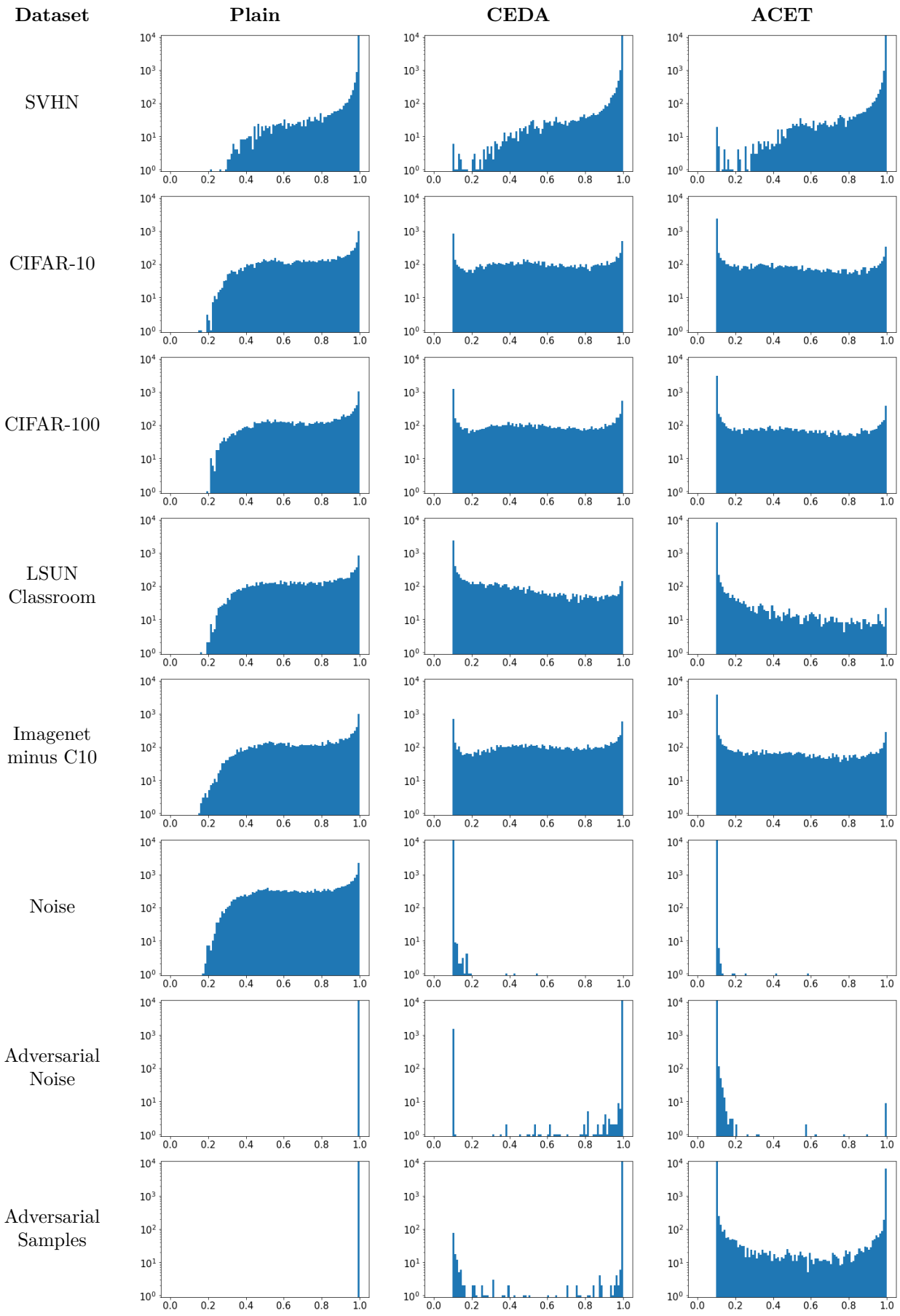

Figure 8: Histograms (logarithmic scale) of maximum confidence values of the three compared models for SVHN on various evaluation datasets.

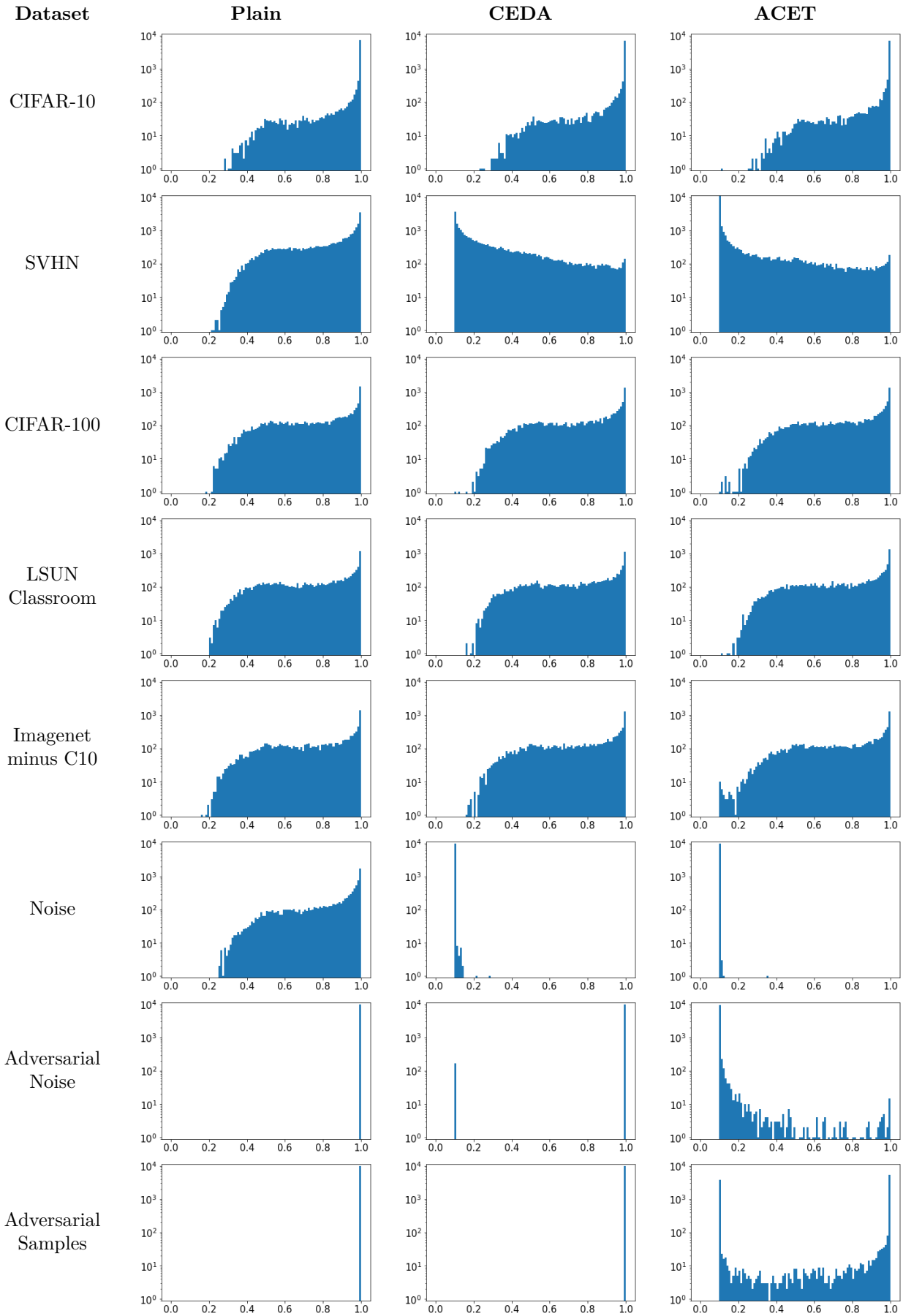

Figure 9: Histograms (logarithmic scale) of maximum confidence values of the three compared models for CIFAR-10 on various evaluation datasets.

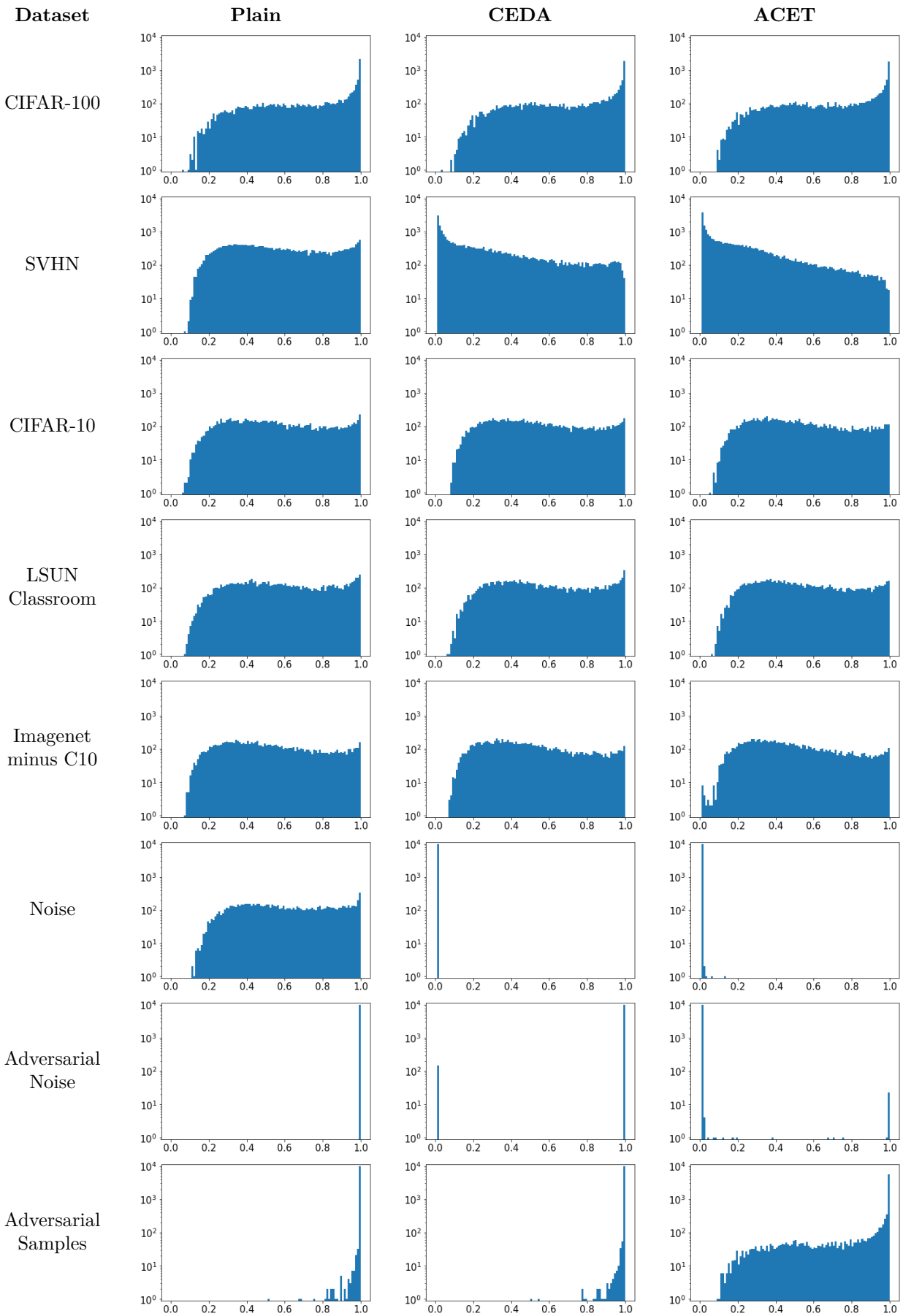

Figure 10: Histograms (logarithmic scale) of maximum confidence values of the three compared models for CIFAR-10 on various evaluation datasets.