Table 1. Comparison of the baseline method (Hendrycks & Gimpel, 2017) and confidence-based thresholding. All results are averaged over 5 runs. All values are shown in percentages. ↓ indicates that lower values are better, while ↑ indicates that higher scores are better.

Model In-distribution Dataset	**Out-of-distribution Dataset**	**Classification Error** ↓	**FPR (95% TPR)** ↓	**Detection Error** ↓	**AUROC** ↑	**AUPR In** ↑	**AUPR Out** ↑
		Baseline (Hendrycks & Gimpel, 2017)/Confidence Thresholding					
DenseNet-BC SVHN	TinyImageNet (resize)	2.89/**2.77**	7.2/**1.5**	5.3/**2.8**	98.4/**99.5**	99.4/**99.8**	95.6/**98.7**
	LSUN (resize)		6.0/**1.0**	4.9/**2.3**	98.6/**99.7**	99.5/**99.9**	96.0/**99.0**
	iSUN		6.0/**0.9**	4.9/**2.3**	98.6/**99.7**	99.5/**99.9**	95.7/**98.8**
	All Images		12.2/**4.2**	7.2/**4.5**	97.3/**98.9**	95.1/**97.4**	98.4/**99.4**
WRN-16-8 SVHN	TinyImageNet (resize)	2.77/**2.66**	10.6/**1.5**	6.1/**2.7**	97.8/**99.6**	99.2/**99.8**	93.6/**99.2**
	LSUN (resize)		9.5/**0.6**	5.8/**1.8**	98.0/**99.8**	99.3/**99.9**	94.0/**99.5**
	iSUN		9.6/**0.8**	5.9/**2.1**	98.0/**99.8**	99.3/**99.9**	93.4/**99.4**
	All Images		15.7/**5.3**	7.9/**5.0**	96.7/**98.7**	94.1/**96.8**	97.9/**99.4**
VGG13 SVHN	TinyImageNet (resize)	3.05/**2.98**	11.4/**1.8**	6.2/**3.1**	97.8/**99.6**	99.2/**99.8**	93.7/**99.1**
	LSUN (resize)		9.4/**0.8**	5.7/**2.0**	98.1/**99.8**	99.3/**99.9**	94.3/**99.6**
	iSUN		10.0/**1.0**	6.0/**2.2**	98.0/**99.8**	99.3/**99.9**	93.7/**99.5**
	All Images		14.2/**4.3**	7.1/**4.6**	97.3/**99.2**	95.9/**98.5**	98.2/**99.6**
DenseNet-BC CIFAR-10	TinyImageNet (resize)	**4.17**/4.39	44.9/**33.8**	12.8/**12.3**	93.2/**94.2**	94.6/**95.0**	91.2/**93.0**
	LSUN (resize)		38.6/**30.7**	10.8/**10.3**	94.6/**95.4**	95.9/**96.4**	92.8/**93.9**
	iSUN		41.4/**31.6**	11.6/**11.0**	94.1/**95.0**	95.8/**96.3**	91.3/**93.0**
	All Images		40.9/**28.9**	11.6/**10.9**	94.1/**95.3**	87.6/**88.1**	98.3/**98.7**
WRN-28-10 CIFAR-10	TinyImageNet (resize)	**3.25**/3.46	41.0/**26.6**	14.3/**11.6**	91.0/**94.5**	88.9/**94.1**	90.5/**94.0**
	LSUN (resize)		34.7/**24.0**	11.7/**9.1**	93.7/**96.0**	93.4/**96.6**	92.7/**94.5**
	iSUN		36.7/**24.9**	12.6/**9.8**	92.8/**95.7**	92.6/**96.5**	91.1/**94.0**
	All Images		36.1/**23.3**	12.4/**9.7**	92.9/**95.7**	73.3/**86.7**	98.1/**98.8**
VGG13 CIFAR-10	TinyImageNet (resize)	**5.28**/5.44	43.8/**18.4**	12.0/**9.4**	93.5/**97.0**	94.6/**97.3**	91.7/**96.9**
	LSUN (resize)		41.9/**16.4**	11.5/**8.3**	94.0/**97.5**	95.1/**97.8**	92.2/**97.2**
	iSUN		41.2/**16.3**	11.4/**8.5**	94.0/**97.5**	95.5/**98.0**	91.5/**96.9**
	All Images		41.6/**19.2**	11.7/**9.1**	93.9/**97.1**	85.5/**92.0**	98.2/**99.3**

we equally weight P_{in} and P_{out} as if they have the same probability of appearing in the test set.

AUROC: Measures the Area Under the Receiver Operating Characteristic curve. The Receiver Operating Characteristic (ROC) curve plots the relationship between TPR and FPR. The area under the ROC curve can be interpreted as the probability that a positive example (in-distribution) will have a higher detection score than a negative example (out-of-distribution).

AUPR: Measures the Area Under the Precision-Recall (PR) curve. The PR curve is made by plotting precision $=$ TP$/$(TP $+$ FP) versus recall $=$ TP$/$(TP $+$ FN). In our tests, AUPR-In indicates that in-distribution examples are used as the positive class, while AUPR-Out indicates that out-of-distribution examples are used as the positive class.

4.2.4. Model Training

We evaluate our confidence estimation technique by applying it to several different neural network architectures: DenseNet (Huang et al., 2017), WideResNet, (Zagoruyko & Komodakis, 2016), and VGGNet (Simonyan & Zisserman, 2015). Following Liang et al. (2018), we train a DenseNet with depth $L = 100$ and growth rate $k = 12$ (hereafter referred to as DenseNet-BC), as well as a WideResNet. We use a depth of 16 and a widening factor of 8 for the SVHN dataset, and a depth of 28 and widening factor of 10 for CIFAR-10 (referred to as WRN-16-8 and WRN-28-10 respectively). We also train a VGG13 model to evaluate performance on network architectures without skip connections.

All models are trained using stochastic gradient descent, with Nesterov momentum of 0.9. We also apply standard data augmentation (cropping and flipping) and Cutout. Following Huang et al. (2017), DenseNet-BC is trained for 300 epochs with batches of 64 images, and a weight decay of 1e-4. The learning rate is initialized to 0.1 at the beginning of training and is reduced by a factor of 10$\times$ after the 150th and 225th epochs. We use the settings of Zagoruyko & Komodakis (2016) for both WideResNet and VGG13, training for 200 epochs with batches of size 128, and a weight decay of 5e-4. The learning rate is initialized to 0.1 at the beginning of training, and reduced by a factor of 5$\times$ after the 60$^{\text{th}}$, 120$^{\text{th}}$, and 160$^{\text{th}}$ epochs.

We train two sets of models, each with 5 runs per model. The first set of models are trained without any confidence branch, and are used to evaluate the baseline method (Hendrycks & Gimpel, 2017) and ODIN (Liang et al., 2018). The second set of models is trained with a confidence estimation branch, and are used to evaluate confidence thresholding with and without input preprocessing. A budget value of $\beta = 0.3$ is