DOCTOR

A Simple Method for Detecting Misclassification Errors

Federica Granese * [2,3], Marco Romanelli * [1], Daniele Gorla [3], Catuscia Palamidessi [2] and Pablo Piantanida [1]

2021 Conference on Neural Information Processing Systems (NeurIPS 2021), December 6th to 14th, 2021

[2]Lix, Inria, Institute Polytechnique de Paris [3]Sapienza University of Rome

[1]L2S, CentraleSupélec, CNRS, Université Paris-Saclay

* = equal contribution

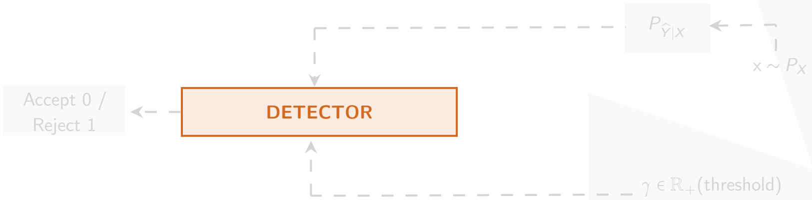
Main Goal of this Work

* Deep Neural Networks (DNNs) exhibit **unwanted behaviors** as they tend to be overconfident even in presence of wrong decisions.

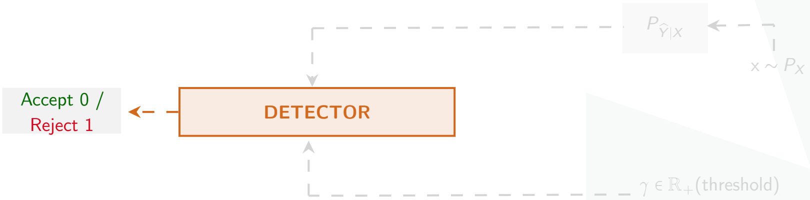
Main Goal of this Work

- **DOCTOR** is a simple method to detect whether a model's prediction is likely to be correct (**accept**), or not (**reject**).

Detection Framework



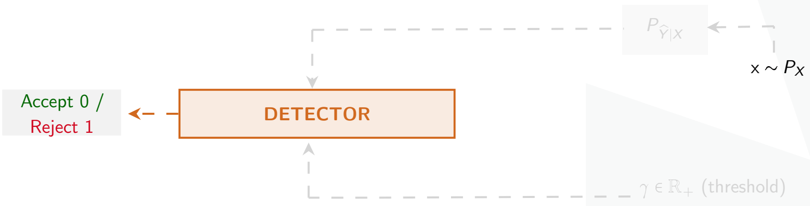
Detection Framework



Detection Framework

Given a pre-trained model and let

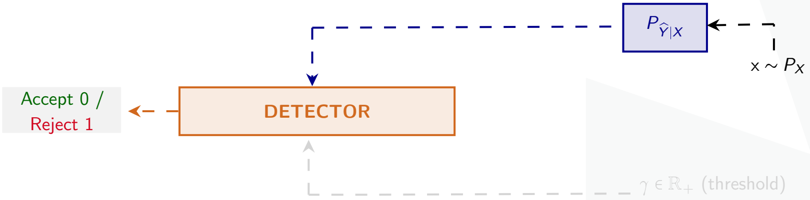
- P_X be the (unknown) probability distribution over $\mathcal{X}$ (feature space);



Detection Framework

Given a pre-trained model and let

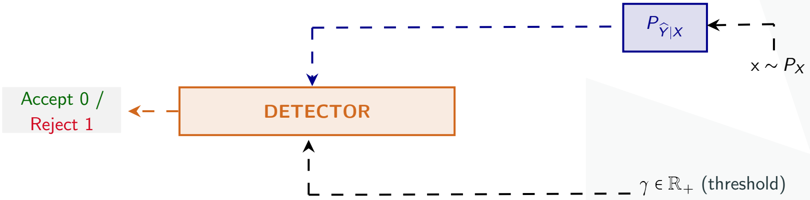
- P_X be the (unknown) probability distribution over $\mathcal{X}$ (feature space);
- $P_{\widehat{Y}|X}$ be the posterior probability given a sample (model distribution).



Detection Framework

Given a pre-trained model and let

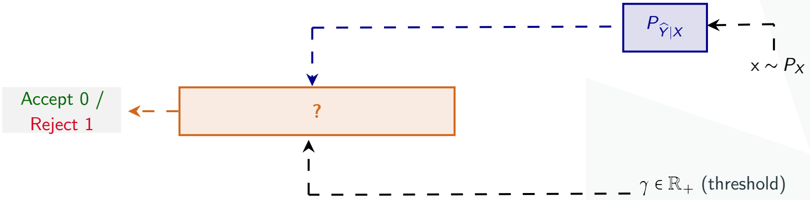
- P_X be the (unknown) probability distribution over $\mathcal{X}$ (feature space);
- $P_{\widehat{Y}|X}$ be the posterior probability given a sample (model distribution).
- γ threshold.



Detection Framework

Given a pre-trained model and let

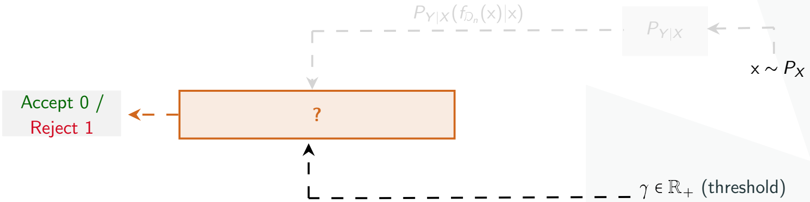
- P_X be the (unknown) probability distribution over $\mathcal{X}$ (feature space);
- $P_{\widehat{Y}|X}$ be the posterior probability given a sample (model distribution).
- γ threshold.



Optimal (Oracle) Detector

Denote

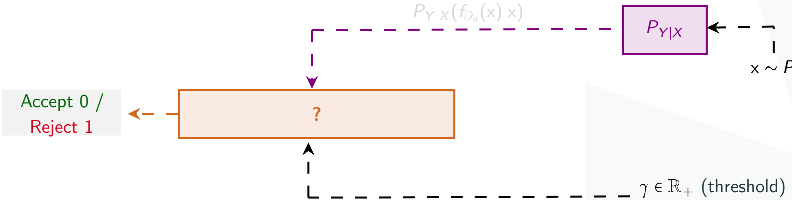
- P_X be the probability distribution over X (feature space)



Optimal (Oracle) Detector

Denote

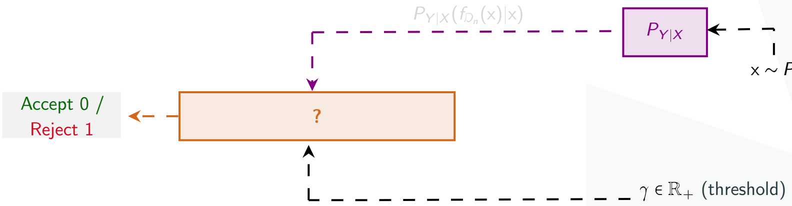
- P_X be the probability distribution over X (feature space)
- $P_{Y|X}$ be the true posterior distribution of the labels given the samples.



Optimal (Oracle) Detector

Denote

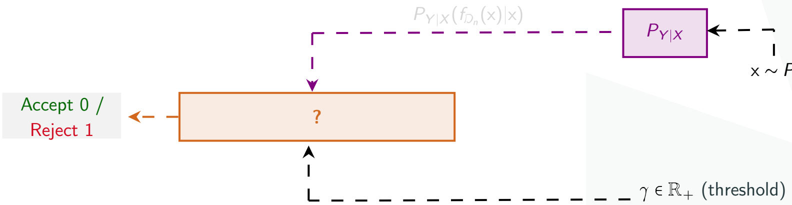
- P_X be the probability distribution over X (feature space)
- $P_{Y|X}$ be the true posterior distribution of the labels given the samples.
- $f_{\mathcal{D}_n} : \mathcal{X} \to \mathcal{Y}$ be the predictor.



Optimal (Oracle) Detector

For a given $x \in \mathcal{X}$,

- $E(x) \triangleq \mathbb{1}[Y \neq f_{\mathcal{D}_n}(x)]$ denotes the **true-error variable**;

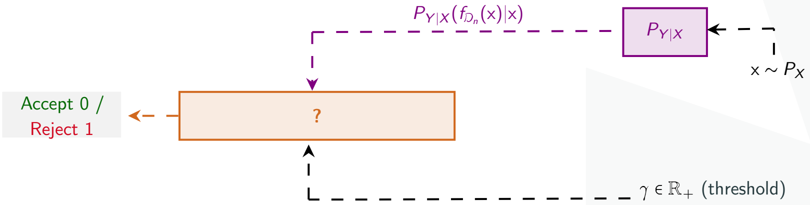


Optimal (Oracle) Detector

For a given $\mathrm{x} \in \mathcal{X}$,

- $E(x) \triangleq \mathbb{1}[Y \neq f_{\mathcal{D}_n}(\mathrm{x})]$ denotes the **true-error variable**;
- The probability of misclassification w.r.t. $P_{Y|X}$ is given by

$$Pe(\mathrm{x}) \triangleq \mathbb{E}[E(\mathrm{x})|\mathrm{x}] = 1 - P_{Y|X}(f_{\mathcal{D}_n}(\mathrm{x})|\mathrm{x}).$$

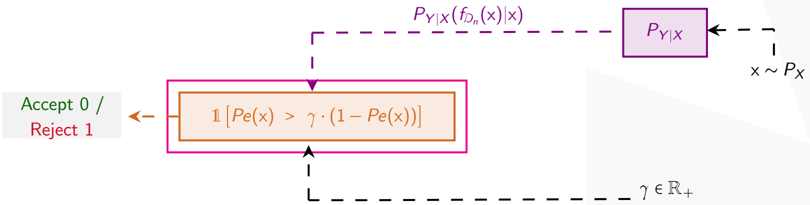


Optimal (Oracle) Detector

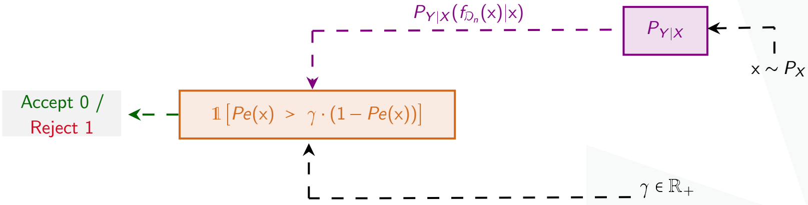
For a given $x \in \mathcal{X}$,

- $E(x) \triangleq \mathbb{1}[Y \neq f_{\mathcal{D}_n}(x)]$ denotes the **true-error variable**;
- The probability of misclassification w.r.t. $P_{Y|X}$ is given by

$$Pe(x) \triangleq \mathbb{E}[E(x)|x] = 1 - P_{Y|X}(f_{\mathcal{D}_n}(x)|x).$$



Optimal (Oracle) Detector



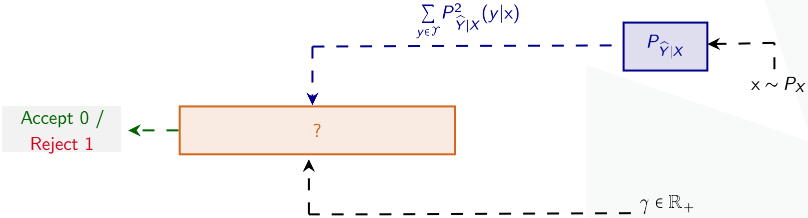
Unfortunately, in practice $P_{Y|X}$ is not available: we need to find a way to approximate $Pe(x)$.

DOCTOR: $D_\alpha(\mathrm{x},\gamma)$

For a given $\mathrm{x} \in \mathcal{X}$,

$$1 - \widehat{g}(\mathrm{x}) \triangleq \sum_{y \in \mathcal{Y}} P_{\widehat{Y}|X}(y|\mathrm{x})\mathbb{P}(\widehat{Y} \neq y|\mathrm{x}) = 1 - \sum_{y \in \mathcal{Y}} P^2_{\widehat{Y}|X}(y|\mathrm{x})$$

is the probability of incorrectly classifying x if it was randomly labeled according to $P_{\widehat{Y}|X}$.

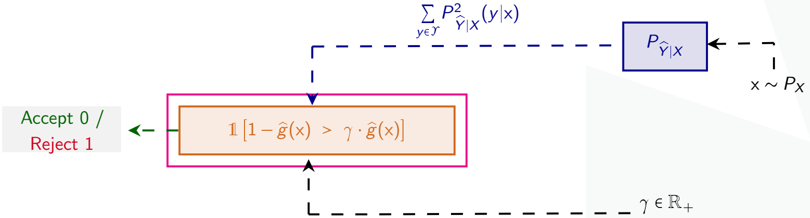


DOCTOR: $D_\alpha(x, \gamma)$

For a given $x \in \mathcal{X}$,

$$1 - \widehat{g}(x) \triangleq \sum_{y \in \mathcal{Y}} P_{\widehat{Y}|X}(y|x)\mathbb{P}(\widehat{Y} \neq y|x) = 1 - \sum_{y \in \mathcal{Y}} P^2_{\widehat{Y}|X}(y|x)$$

is the probability of incorrectly classifying x if it was randomly labeled according to $P_{\widehat{Y}|X}$.

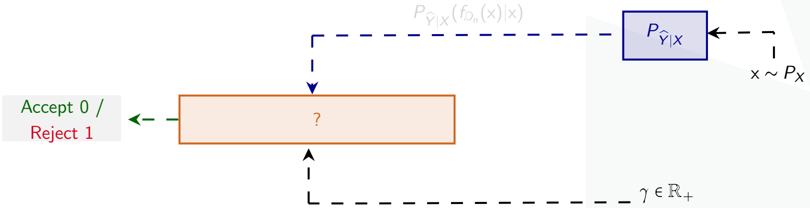


DOCTOR: $D_\beta(x, \gamma)$

For a given $x \in \mathcal{X}$,

- The **self-error variable** is defined as:

$$\widehat{E}(x) \triangleq \mathbb{1}[\widehat{Y} \neq f_{\mathcal{D}_n}(x)];$$



DOCTOR: $D_\beta(x, \gamma)$

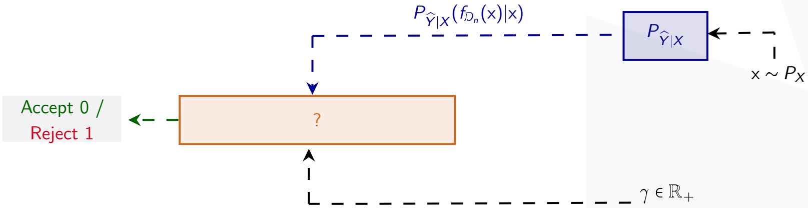
For a given $x \in \mathcal{X}$,

- The **self-error variable** is defined as:

$$\widehat{E}(x) \triangleq \mathbb{1}[\widehat{Y} \neq f_{\mathcal{D}_n}(x)];$$

- The probability of error classification w.r.t. $P_{\widehat{Y}|X}$ is given by

$$\widehat{Pe}(x) \triangleq \mathbb{E}[\widehat{E}(x)|x] = 1 - P_{\widehat{Y}|X}(f_{\mathcal{D}_n}(x)|x).$$



DOCTOR: $D_\beta(x, \gamma)$

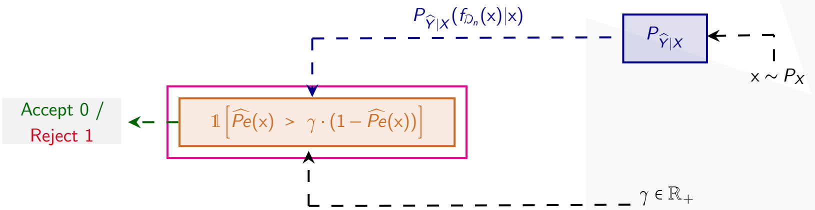
For a given $x \in \mathcal{X}$,

- The **self-error variable** is defined as:

$$\widehat{E}(x) \triangleq \mathbb{1}[\widehat{Y} \neq f_{\mathcal{D}_n}(x)];$$

- The probability of error classification w.r.t. $P_{\widehat{Y}|X}$ is given by

$$\widehat{Pe}(x) \triangleq \mathbb{E}[\widehat{E}(x)|x] = 1 - P_{\widehat{Y}|X}(f_{\mathcal{D}_n}(x)|x).$$



Metrics

FRR versus TRR

The false rejection rate (FRR) represents the probability that a hit (sample correctly classified) is rejected, while the true rejection rate (TRR) is the probability that a miss (sample wrongly classified) is rejected.

Metrics

FRR versus TRR

The false rejection rate (FRR) represents the probability that a hit (sample correctly classified) is rejected, while the true rejection rate (TRR) is the probability that a miss (sample wrongly classified) is rejected.

AUROC

The area under the Receiver Operating Characteristic curve (ROC) depicts the relationship between TRR and FRR. The perfect detector corresponds to a score of 100%.

TRR is at 95%.

Metrics

FRR versus TRR

The false rejection rate (FRR) represents the probability that a hit (sample correctly classified) is rejected, while the true rejection rate (TRR) is the probability that a miss (sample wrongly classified) is rejected.

AUROC

The area under the Receiver Operating Characteristic curve (ROC) depicts the relationship between TRR and FRR. The perfect detector corresponds to a score of 100%.

FRR at 95% TRR

This is the probability that a hit is rejected when the TRR is at 95%.

Totally Black Box & Partially Black Box

Totally Black Box (TBB) Scenario

In TBB only the output of the last layer of the network is available, hence gradient-propagation to perform input pre-processing is not allowed.

Partially Black Box (PBB) Scenario

In PBB we allow method-specific inputs perturbations and the possibility of doing temperature scaling.

Competitors - PBB

1) ODIN [Liang et al., 2018]
It comprises temperature scaling and input pre-processing via perturbation. It compares the maximum softmax probability with a threshold $\delta \in [0, 1]$ to decide whether to accept or to reject the input sample.

2) Mahalanobis distance for OOD (MHLNB) [Lee et al., 2018]
It consists in calculating the distance between the input sample and training distribution. It compares the distance with a threshold $\zeta \in \mathbb{R}$ to decide whether to accept or to reject the input sample.

3) Energy Score (ENERGY) [Liu et al., 2020]
It comprises the the denominator of the softmax activation and it compares it with a threshold $\xi \in \mathbb{R}$.

Competitors - TBB

1) Softmax Response
(SR) [Hendrycks and Gimpel, 2017, Geifman and El-Yaniv, 2017]
ODIN without temperature scaling and input pre-processing.

2) Mahalanobis distance for OOD (MHLNB) [Lee et al., 2018]
Mahalanobis distance without input pre-processing and with the softmax output in place of the logits.

Discrimination Performance - PBB

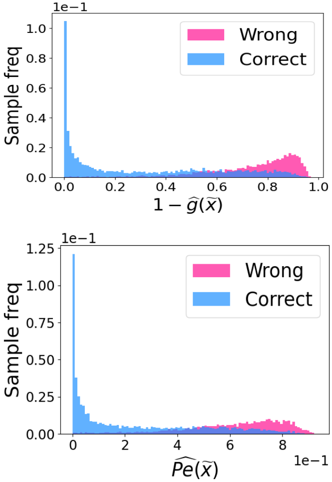


Figure 2. DOCTOR, ODIN and MHLNB to split data samples in TinyImageNet under PBB. Histograms for wrongly classified samples and correctly classified samples.

Discrimination Performance - PBB

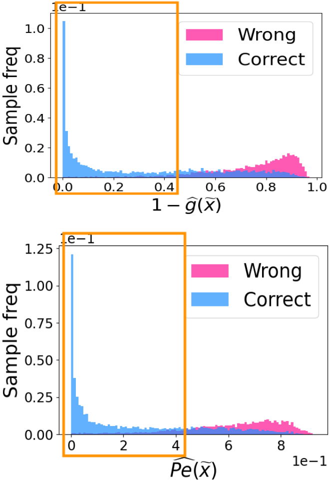


Figure 2. DOCTOR, ODIN and MHLNB to split data samples in TinyImageNet under PBB. Histograms for wrongly classified samples and correctly classified samples.

Discrimination Performance - PBB

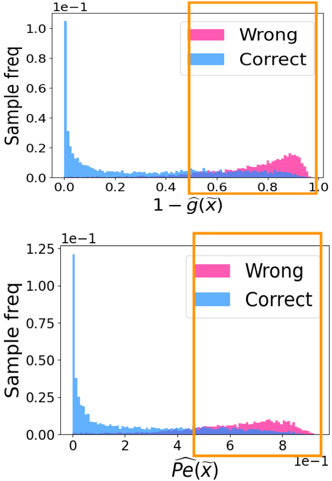


Figure 2. DOCTOR, ODIN and MHLNB to split data samples in TinyImageNet under PBB. Histograms for wrongly classified samples and correctly classified samples.

Discrimination Performance - PBB

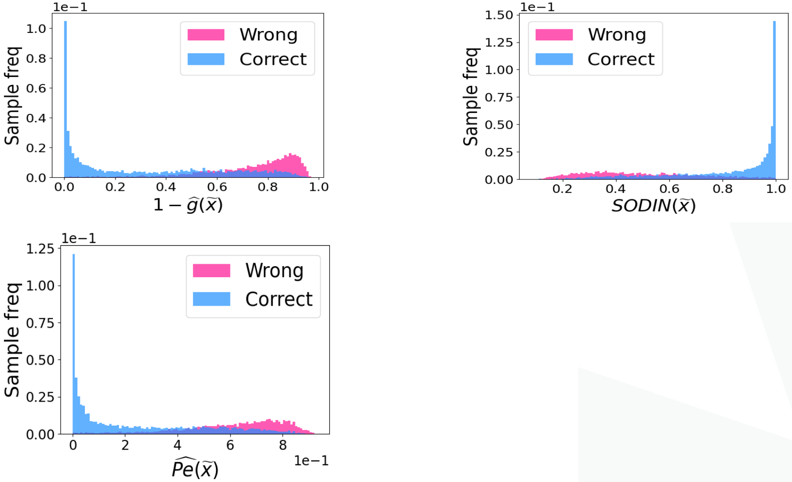


Figure 2. DOCTOR, ODIN and MHLNB to split data samples in TinyImageNet under PBB. Histograms for wrongly classified samples and correctly classified samples.

Discrimination Performance - PBB

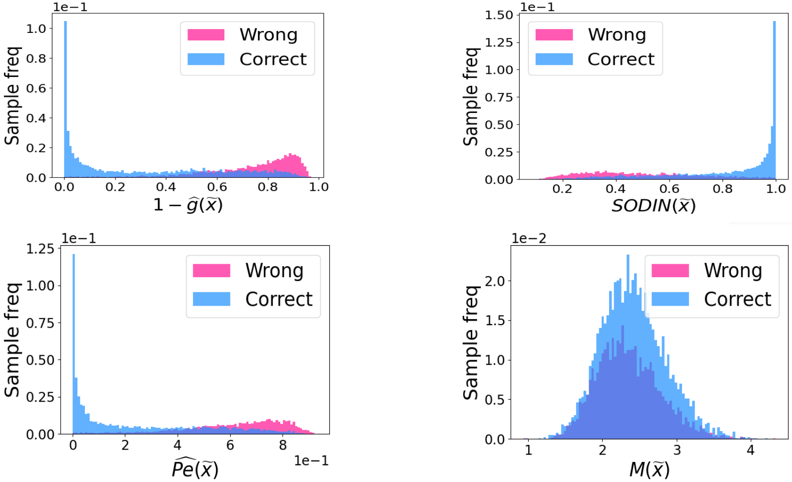


Figure 2. DOCTOR, ODIN and MHLNB to split data samples in TinyImageNet under PBB. Histograms for wrongly classified samples and correctly classified samples.

PBB: ROCs

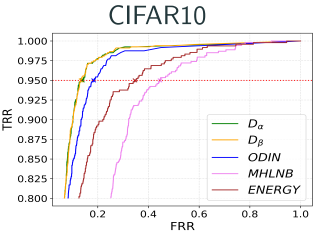
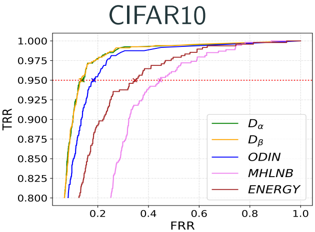


Figure 3. ROC curves. Comparison between DOCTOR, ODIN and MHLNB. The red dashed line marks the 95% threshold of TRR.

PBB: ROCs



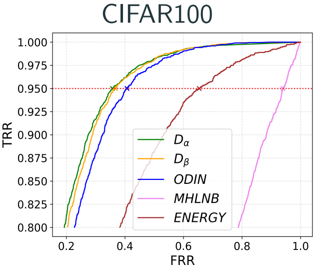


Figure 3. ROC curves. Comparison between DOCTOR, ODIN and MHLNB. The red dashed line marks the 95% threshold of TRR.

PBB: ROCs

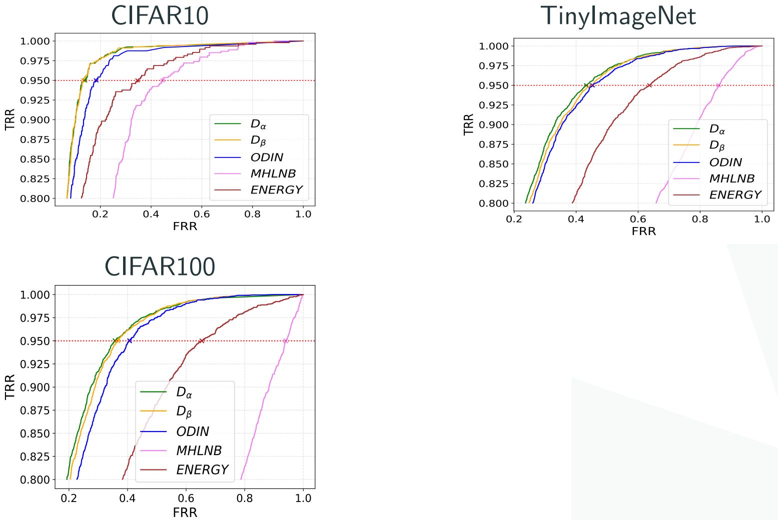


Figure 3. ROC curves. Comparison between DOCTOR, ODIN and MHLNB. The red dashed line marks the 95% threshold of TRR.

PBB: ROCs

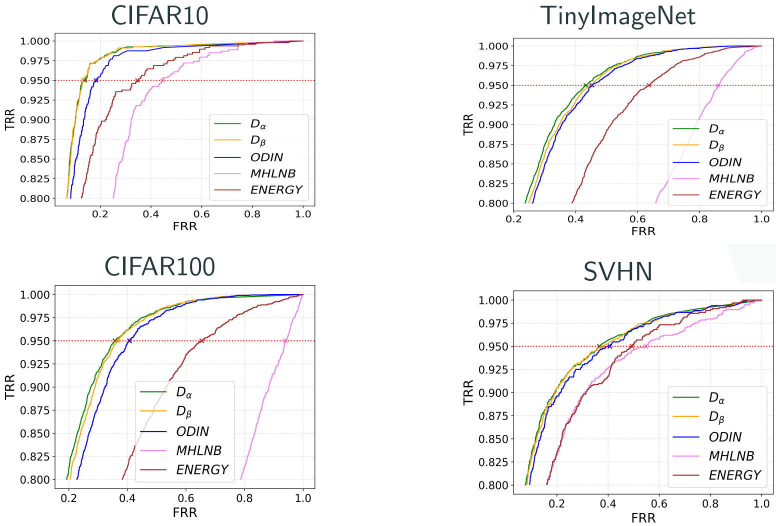


Figure 3. ROC curves. Comparison between DOCTOR, ODIN and MHLNB. The red dashed line marks the 95% threshold of TRR.

Overall Results: TBB & PBB

DATASET	METHOD	AUROC %		FRR % (95 % TRR)	
		TBB	PBB	TBB	PBB
CIFAR10 Acc. 95%	D_α	**94**	**95.2**	**17.9**	13.9
	D_β	68.5	94.8	18.6	**13.4**
	ODIN	93.8	94.2	18.2	18.4
	SR	93.8	-	18.2	-
	MHLNB	92.2	84.4	30.8	44.6
	ENERGY	-	91.1	-	34.7
CIFAR100 Acc. 78%	D_α	**87**	**88.2**	40.6	**35.7**
	D_β	84.2	87.4	40.6	36.7
	ODIN	86.9	87.1	40.5	<u>40.7</u>
	SR	86.9	-	**40.5**	-
	MHLNB	82.6	50	66.7	94
	ENERGY	-	78.7	-	65.4
Tiny ImageNet Acc. 63%	D_α	**84.9**	**86.1**	**45.8**	**43.3**
	D_β	**84.9**	85.3	**45.8**	45.1
	ODIN	**84.9**	84.9	**45.8**	<u>45.3</u>
	SR	**84.9**	-	**45.8**	-
	MHLNB	78.4	59	82.3	86
	ENERGY	-	78.2	-	63.7

DATASET	METHOD	AUROC %		FRR % (95 % TRR)	
		TBB	PBB	TBB	PBB
SVHN Acc. 96%	D_α	**92.3**	**93**	**38.6**	**36.6**
	D_β	92.2	92.8	39.7	38.4
	ODIN	**92.3**	92.3	**38.6**	40.7
	SR	**92.3**	-	**38.6**	-
	MHLNB	87.3	88	85.8	54.7
	ENERGY	-	88.9	-	49.4
Amazon Fashion Acc. 85%	D_α	**89.7**	-	27.1	-
	D_β	**89.7**	-	**26.3**	-
	SR	87.4	-	50.1	
	ENERGY	-	-	-	-
Amazon Software Acc. 73%	D_α	**68.8**	-	**73.2**	-
	D_β	**68.8**	-	**73.2**	-
	SR	67.3	-	86.6	-
	ENERGY	-	-	-	-
IMDb Acc. 90%	D_α	**84.4**	-	**54.2**	-
	D_β	**84.4**	-	54.4	-
	SR	83.7	-	61.7	-
	ENERGY	-	-	-	-

Table 1. Collection of the results in both TBB and PBB.

Overall Results: TBB & PBB

DATASET	METHOD	AUROC %		FRR % (95 % TRR)	
		TBB	PBB	TBB	PBB
CIFAR10 Acc. 95%	D_α	**94**	**95.2**	**17.9**	13.9
	D_β	68.5	94.8	18.6	**13.4**
	ODIN	93.8	94.2	18.2	18.4
	SR	93.8	-	18.2	-
	MHLNB	92.2	84.4	30.8	44.6
	ENERGY	-	91.1	-	34.7
CIFAR100 Acc. 78%	D_α	**87**	**88.2**	40.6	**35.7**
	D_β	84.2	87.4	40.6	36.7
	ODIN	86.9	87.1	40.5	40.7
	SR	86.9	-	**40.5**	-
	MHLNB	82.6	50	66.7	94
	ENERGY	-	78.7	-	65.4
Tiny ImageNet Acc. 63%	D_α	**84.9**	**86.1**	**45.8**	**43.3**
	D_β	**84.9**	85.3	**45.8**	45.1
	ODIN	**84.9**	84.9	**45.8**	45.3
	SR	**84.9**	-	**45.8**	-
	MHLNB	78.4	59	82.3	86
	ENERGY	-	78.2	-	63.7

DATASET	METHOD	AUROC %		FRR % (95 % TRR)	
		TBB	PBB	TBB	PBB
SVHN Acc. 96%	D_α	**92.3**	**93**	**38.6**	**36.6**
	D_β	92.2	92.8	39.7	38.4
	ODIN	**92.3**	92.3	**38.6**	<u>40.7</u>
	SR	**92.3**	-	**38.6**	-
	MHLNB	87.3	88	85.8	54.7
	ENERGY	-	88.9	-	49.4
Amazon Fashion Acc. 85%	D_α	**89.7**	-	27.1	-
	D_β	**89.7**	-	**26.3**	-
	SR	87.4	-	50.1	
	ENERGY	-	-	-	-
Amazon Software Acc. 73%	D_α	**68.8**	-	**73.2**	-
	D_β	**68.8**	-	**73.2**	-
	SR	67.3	-	86.6	-
	ENERGY	-	-	-	-
IMDb Acc. 90%	D_α	**84.4**	-	**54.2**	-
	D_β	**84.4**	-	54.4	-
	SR	83.7	-	61.7	-
	ENERGY	-	-	-	-

Table 1. Collection of the results in both TBB and PBB.

Overall Results: TBB & PBB

DATASET	METHOD	AUROC %		FRR % (95 % TRR)	
		TBB	PBB	TBB	PBB
CIFAR10 Acc. 95%	D_α	**94**	**95.2**	**17.9**	13.9
	D_β	68.5	94.8	18.6	**13.4**
	ODIN	93.8	94.2	18.2	18.4
	SR	93.8	-	18.2	-
	MHLNB	92.2	84.4	30.8	44.6
	ENERGY	-	91.1	-	34.7
CIFAR100 Acc. 78%	D_α	**87**	**88.2**	40.6	**35.7**
	D_β	84.2	87.4	40.6	36.7
	ODIN	86.9	87.1	40.5	40.7
	SR	86.9	-	**40.5**	-
	MHLNB	82.6	50	66.7	94
	ENERGY	-	78.7	-	65.4
Tiny ImageNet Acc. 63%	D_α	**84.9**	**86.1**	**45.8**	**43.3**
	D_β	**84.9**	85.3	**45.8**	45.1
	ODIN	**84.9**	84.9	**45.8**	45.3
	SR	**84.9**	-	**45.8**	-
	MHLNB	78.4	59	82.3	86
	ENERGY	-	78.2	-	63.7
SVHN Acc. 96%	D_α	**92.3**	**93**	**38.6**	**36.6**
	D_β	92.2	92.8	39.7	38.4
	ODIN	**92.3**	92.3	**38.6**	40.7
	SR	**92.3**	-	**38.6**	-
	MHLNB	87.3	88	85.8	54.7
	ENERGY	-	88.9	-	49.4
Amazon Fashion Acc. 85%	D_α	**89.7**	-	27.1	-
	D_β	**89.7**	-	**26.3**	-
	SR	87.4	-	50.1	
	ENERGY	-	-	-	-
Amazon Software Acc. 73%	D_α	**68.8**	-	**73.2**	-
	D_β	**68.8**	-	**73.2**	-
	SR	67.3	-	86.6	-
	ENERGY	-	-	-	-
IMDb Acc. 90%	D_α	**84.4**	-	**54.2**	-
	D_β	**84.4**	-	54.4	-
	SR	83.7	-	61.7	-
	ENERGY	-	-	-	-

Table 1. Collection of the results in both TBB and PBB.

Overall Results: TBB & PBB

DATASET	METHOD	AUROC %		FRR % (95 % TRR)	
		TBB	PBB	TBB	PBB
CIFAR10 Acc. 95%	D_α	**94**	**95.2**	**17.9**	13.9
	D_β	68.5	94.8	18.6	**13.4**
	ODIN	93.8	94.2	18.2	18.4
	SR	93.8	-	18.2	-
	MHLNB	92.2	84.4	30.8	44.6
	ENERGY	-	91.1	-	34.7
CIFAR100 Acc. 78%	D_α	**87**	**88.2**	40.6	**35.7**
	D_β	84.2	87.4	40.6	36.7
	ODIN	86.9	87.1	40.5	40.7
	SR	86.9	-	**40.5**	-
	MHLNB	82.6	50	66.7	94
	ENERGY	-	78.7	-	65.4
Tiny ImageNet Acc. 63%	D_α	**84.9**	**86.1**	**45.8**	**43.3**
	D_β	**84.9**	85.3	**45.8**	45.1
	ODIN	**84.9**	84.9	**45.8**	45.3
	SR	**84.9**	-	**45.8**	-
	MHLNB	78.4	59	82.3	86
	ENERGY	-	78.2	-	63.7

DATASET	METHOD	AUROC %		FRR % (95 % TRR)	
		TBB	PBB	TBB	PBB
SVHN Acc. 96%	D_α	**92.3**	**93**	**38.6**	**36.6**
	D_β	92.2	92.8	39.7	38.4
	ODIN	**92.3**	92.3	**38.6**	<u>40.7</u>
	SR	**92.3**	-	**38.6**	-
	MHLNB	87.3	88	85.8	54.7
	ENERGY	-	88.9	-	49.4
Amazon Fashion Acc. 85%	D_α	**89.7**	-	27.1	-
	D_β	**89.7**	-	**26.3**	-
	SR	87.4	-	50.1	
	ENERGY	-	-	-	-
Amazon Software Acc. 73%	D_α	**68.8**	-	**73.2**	-
	D_β	**68.8**	-	**73.2**	-
	SR	67.3	-	86.6	-
	ENERGY	-	-	-	-
IMDb Acc. 90%	D_α	**84.4**	-	**54.2**	-
	D_β	**84.4**	-	54.4	-
	SR	83.7	-	61.7	-
	ENERGY	-	-	-	-

Table 1. Collection of the results in both TBB and PBB.

Overall Results: TBB & PBB

We observe a reduction of up 4% of FRR in the PBB scenario.

DATASET	METHOD	AUROC %		FRR % (95 % TRR)	
		TBB	PBB	TBB	PBB
CIFAR10 Acc. 95%	D_α	**94**	**95.2**	**17.9**	13.9
	D_β	68.5	94.8	18.6	**13.4**
	ODIN	93.8	94.2	18.2	18.4
	SR	93.8	-	18.2	-
	MHLNB	92.2	84.4	30.8	44.6
	ENERGY	-	91.1	-	34.7
CIFAR100 Acc. 78%	D_α	**87**	**88.2**	40.6	**35.7**
	D_β	84.2	87.4	40.6	36.7
	ODIN	86.9	87.1	40.5	40.7
	SR	86.9	-	**40.5**	-
	MHLNB	82.6	50	66.7	94
	ENERGY	-	78.7	-	65.4
Tiny ImageNet Acc. 63%	D_α	**84.9**	**86.1**	**45.8**	**43.3**
	D_β	**84.9**	85.3	**45.8**	45.1
	ODIN	**84.9**	84.9	**45.8**	45.3
	SR	**84.9**	-	**45.8**	-
	MHLNB	78.4	59	82.3	86
	ENERGY	-	78.2	-	63.7

DATASET	METHOD	AUROC %		FRR % (95 % TRR)	
		TBB	PBB	TBB	PBB
SVHN Acc. 96%	D_α	**92.3**	**93**	**38.6**	**36.6**
	D_β	92.2	92.8	39.7	38.4
	ODIN	**92.3**	92.3	**38.6**	40.7
	SR	**92.3**	-	**38.6**	-
	MHLNB	87.3	88	85.8	54.7
	ENERGY	-	88.9	-	49.4
Amazon Fashion Acc. 85%	D_α	**89.7**	-	27.1	-
	D_β	**89.7**	-	**26.3**	-
	SR	87.4	-	50.1	
	ENERGY	-	-	-	-
Amazon Software Acc. 73%	D_α	**68.8**	-	**73.2**	-
	D_β	**68.8**	-	**73.2**	-
	SR	67.3	-	86.6	-
	ENERGY	-	-	-	-
IMDb Acc. 90%	D_α	**84.4**	-	**54.2**	-
	D_β	**84.4**	-	54.4	-
	SR	83.7	-	61.7	-
	ENERGY	-	-	-	-

Table 1. Collection of the results in both TBB and PBB.

Misclassification Detection in Presence of Out-Of-Distribution (OOD) Samples

- DOCTOR is not tuned for OOD detection (differently from ODIN).
- We test ODIN and DOCTOR when one sample to reject out of five (♣), three (♢), or two (♠) is OOD.

Misclassification Detection in Presence of Out-Of-Distribution (OOD) Samples

- DOCTOR is not tuned for OOD detection (differently from ODIN).
- We test ODIN and DOCTOR when one sample to reject out of five (♣), three (♢), or two (♠) is OOD.

DATASET-In	DATASET-Out	AUROC %				FRR % (95 % TRR)			
		D_α	D_β	ODIN	ENERGY	D_α	D_β	ODIN	ENERGY
CIFAR10 ♣	iSUN	**95.4** / 0.1	95.1 / 0.1	94.6 / 0.1	92.4 / 0	14 / 0.5	**13.5** / 0.4	17.2 / 0.3	32.2 / 0.1
	Tiny (res)	**95.2** / 0.1	94.9 / 0	94.6 / 0.1	92.3 / 0.1	**14** / 0.4	**14** / 0.5	17.8 / 0.4	32.2 / 0.1
CIFAR10 ♢	iSUN	**95.5** / 0.1	95.3 / 0.1	94.9 / 0.1	92.9 / 0	14.4 / 0.6	**13.4** / 0.2	16.8 / 0.5	27 / 1
	Tiny (res)	**95.4** / 0.1	95 / 0.1	94.8 / 0.1	92.8 / 0	15 / 0.1	**14.8** / 0.7	17 / 0.5	28.8 / 1.9
CIFAR10 ♠	iSUN	**95.6** / 0.1	**95.6** / 0	95.4 / 0	93.6 / 0.1	15.1 / 0.1	**13.6** / 0.5	16.1 / 0.2	25.1 / 0.2
	Tiny (res)	**95.5** / 0.1	95.2 / 0.1	95.1 / 0.1	93.5 / 0	**14.7** / 0.3	14.8 / 0.5	17.1 / 0.4	25.6 / 0.3

Table 2. Results in terms of *mean / standard deviation*.

Misclassification Detection in Presence of Out-Of-Distribution (OOD) Samples

- DOCTOR is not tuned for OOD detection (differently from ODIN).
- We test ODIN and DOCTOR when one sample to reject out of five (♣), three (♢), or two (♠) is OOD.

DATASET-In	DATASET-Out	AUROC %				FRR % (95 % TRR)			
		D_α	D_β	ODIN	ENERGY	D_α	D_β	ODIN	ENERGY
CIFAR10 ♣	iSUN	**95.4** / 0.1	95.1 / 0.1	94.6 / 0.1	92.4 / 0	14 / 0.5	**13.5** / 0.4	17.2 / 0.3	32.2 / 0.1
	Tiny (res)	**95.2** / 0.1	94.9 / 0	94.6 / 0.1	92.3 / 0.1	**14** / 0.4	**14** / 0.5	17.8 / 0.4	32.2 / 0.1
CIFAR10 ♢	iSUN	**95.5** / 0.1	95.3 / 0.1	94.9 / 0.1	92.9 / 0	14.4 / 0.6	**13.4** / 0.2	16.8 / 0.5	27 / 1
	Tiny (res)	**95.4** / 0.1	95 / 0.1	94.8 / 0.1	92.8 / 0	15 / 0.1	**14.8** / 0.7	17 / 0.5	28.8 / 1.9
CIFAR10 ♠	iSUN	**95.6** / 0.1	**95.6** / 0	95.4 / 0	93.6 / 0.1	15.1 / 0.1	**13.6** / 0.5	16.1 / 0.2	25.1 / 0.2
	Tiny (res)	**95.5** / 0.1	95.2 / 0.1	95.1 / 0.1	93.5 / 0	**14.7** / 0.3	14.8 / 0.5	17.1 / 0.4	25.6 / 0.3

Table 2. Results in terms of *mean / standard deviation*.

Thanks for your attention.

See you soon at the poster session.

Geifman, Y. and El-Yaniv, R. (2017).
Selective classification for deep neural networks.
In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 4878–4887.

Hendrycks, D. and Gimpel, K. (2017).
A baseline for detecting misclassified and out-of-distribution examples in neural networks.
In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings.

Lee, K., Lee, K., Lee, H., and Shin, J. (2018).
A simple unified framework for detecting out-of-distribution samples and adversarial attacks.
In Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, pages 7167–7177.

Liang, S., Li, Y., and Srikant, R. (2018).
Enhancing the reliability of out-of-distribution image detection in neural networks.
In 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings.

Liu, W., Wang, X., Owens, J. D., and Li, Y. (2020).
Energy-based out-of-distribution detection.
In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.

Supplementary: Optimal (Oracle) Discriminator

- $E \triangleq \mathbb{1}[Y \neq f_{\mathcal{D}_n}(\mathrm{X})]$ denotes the error variable corresponding to $f_{\mathcal{D}_n}$
- $\mathrm{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$ drawn from the unknown distribution p_{XY}
- $p_{XY}(\mathrm{x}, y) \equiv P_E(1) p_{XY|E}(\mathrm{x}, y|1) + P_E(0) p_{XY|E}(\mathrm{x}, y|0)$
- $p_X(\mathrm{x}) \equiv P_E(1) p_{X|E}(\mathrm{x}|1) + P_E(0) p_{X|E}(\mathrm{x}|0)$
- $\mathrm{Pe}(\mathrm{x}) \triangleq \mathbb{E}[E(\mathrm{x})|\mathrm{x}] = 1 - P_{Y|X}(f_{\mathcal{D}_n}(\mathrm{x})|\mathrm{x})$ is the probability of error classification w.r.t. $P_{Y|X}$

$$
\begin{aligned}
D(\mathrm{x}, \gamma) &= \mathbb{1}[p_{X|E}(\mathrm{x}|1) > \gamma \cdot p_{X|E}(\mathrm{x}|0)] \\
&= \mathbb{1}[P_{E|X}(1|\mathrm{x}) P_E(0) > \gamma \cdot (1 - P_{E|X}(1|\mathrm{x})) P_E(1)] \\
&= \mathbb{1}[\mathrm{Pe}(\mathrm{x}) P_E(0) > \gamma \cdot (1 - \mathrm{Pe}(\mathrm{x})) P_E(1)] \\
&= \mathbb{1}[\mathrm{Pe}(\mathrm{x}) > \gamma' \cdot (1 - \mathrm{Pe}(\mathrm{x}))],
\end{aligned}
$$

where $\gamma' = \frac{P_E(1)}{P_E(0)}$.