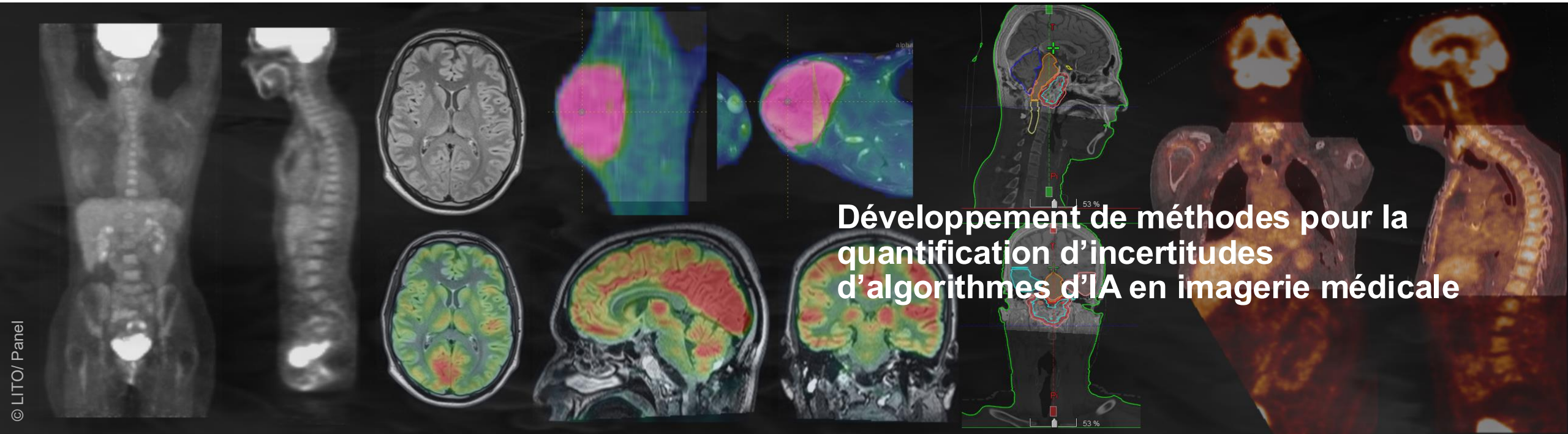
LITO



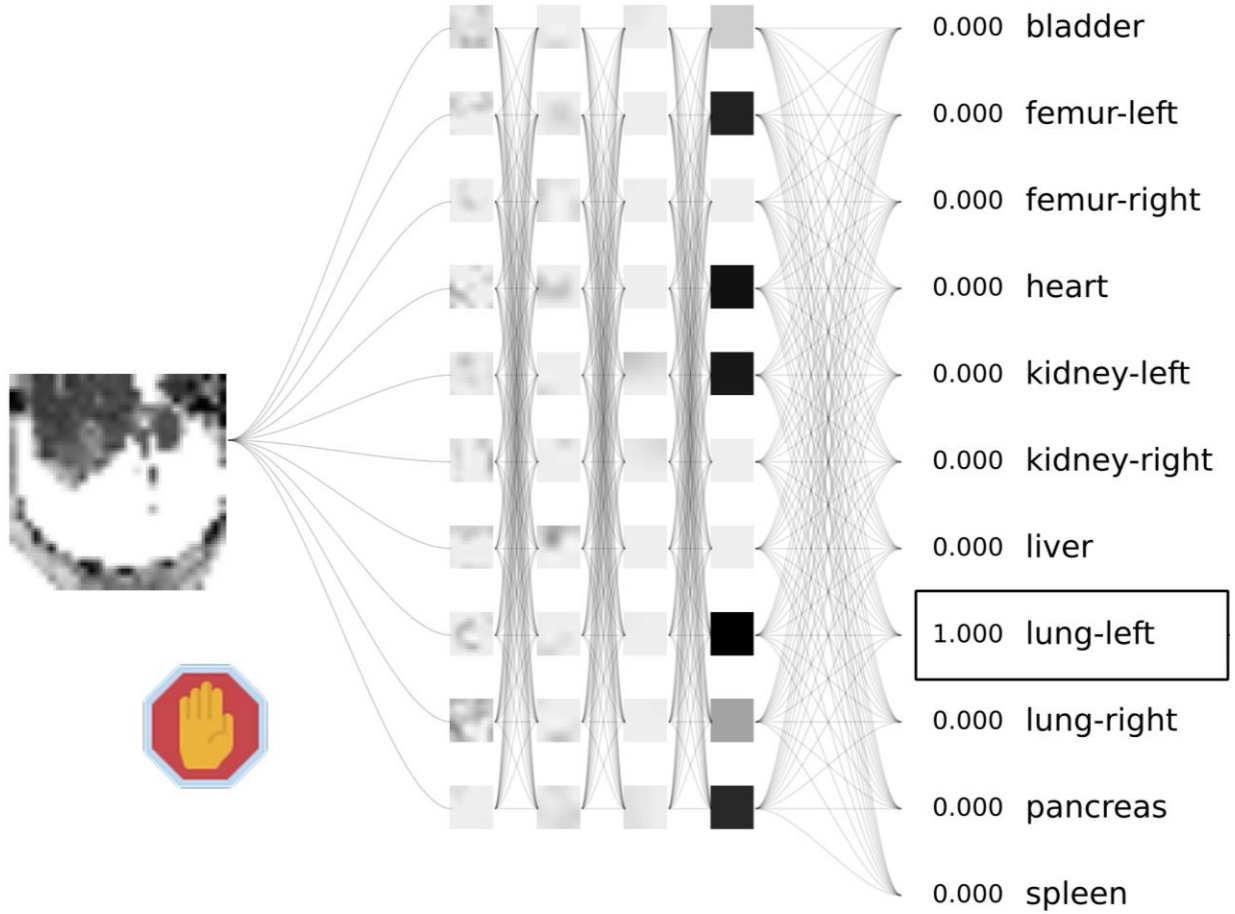
Développement de méthodes pour la quantification d'incertitudes d'algorithmes d'IA en imagerie médicale

© LITO/ Panel

Uncertainty Quantification – Definition

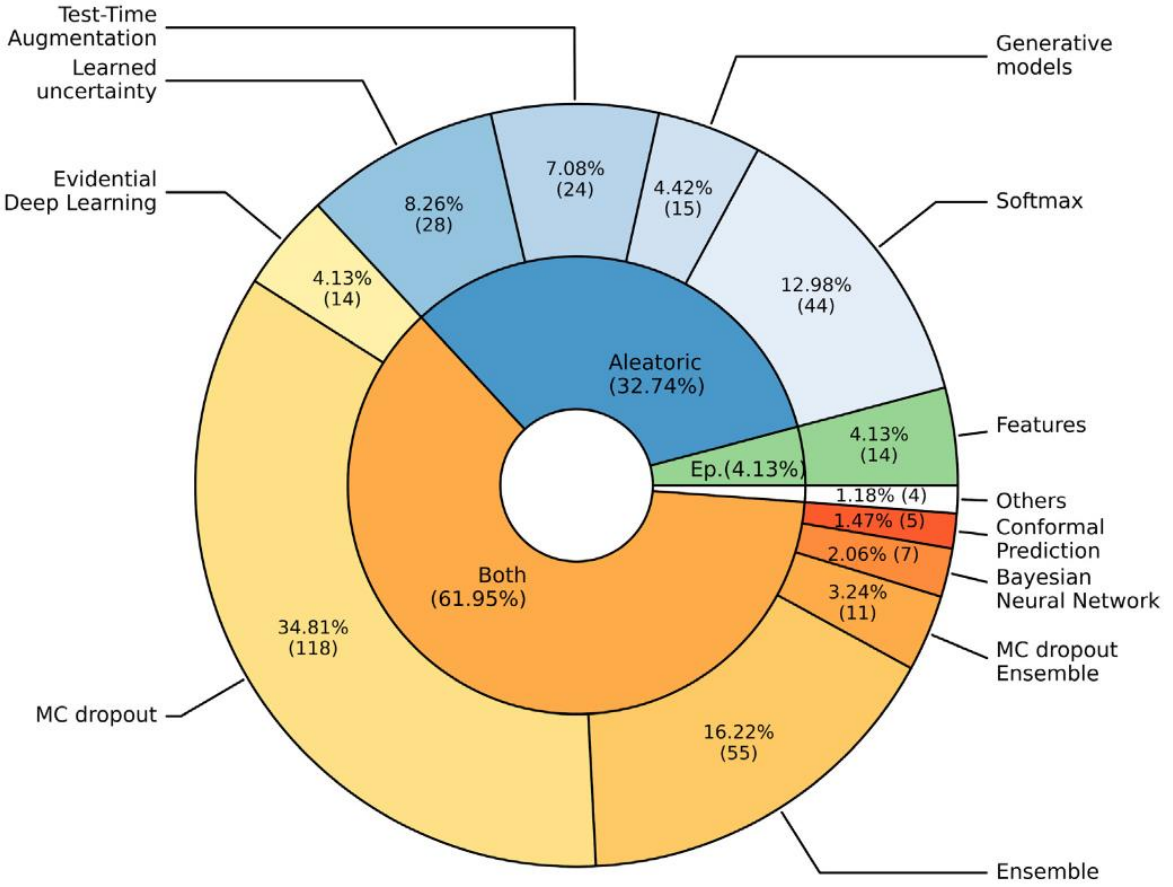
- Increasing use of ML in medical images analysis (classification, segmentation)
- Despite good performances, lack of interpretability and capacity to anticipate failure predictions → poor adoption in clinical practice

- Failures prediction → help increasing confidence in ML algo usage

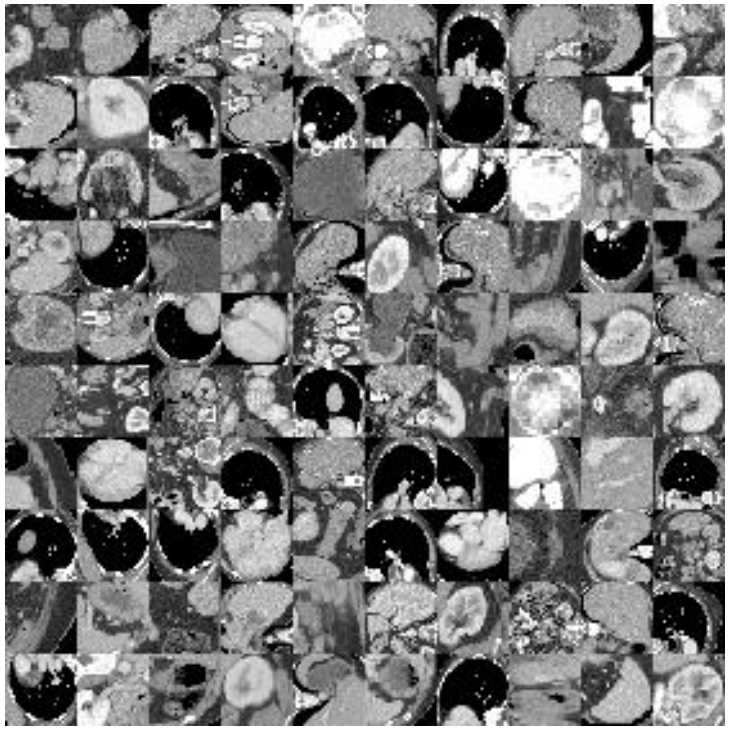


Uncertainty Quantification – Definition

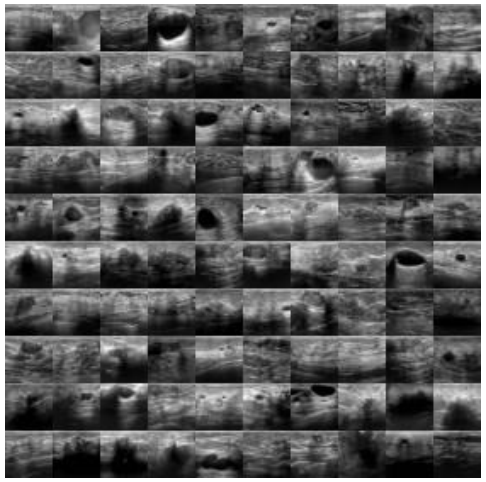
- Already a lot of methods proposed
- Intrinsic vs Post-Hoc
- Posthoc Uncertainty Quantification : Methods to predict algorithm failures with no model alteration

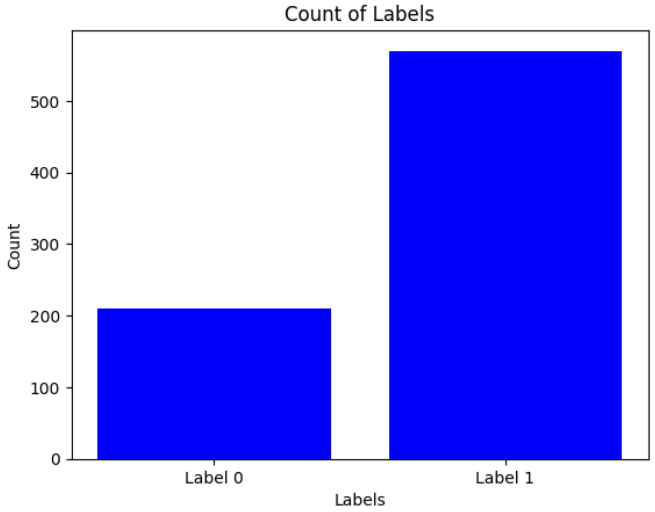


medMNIST – Classification use cases



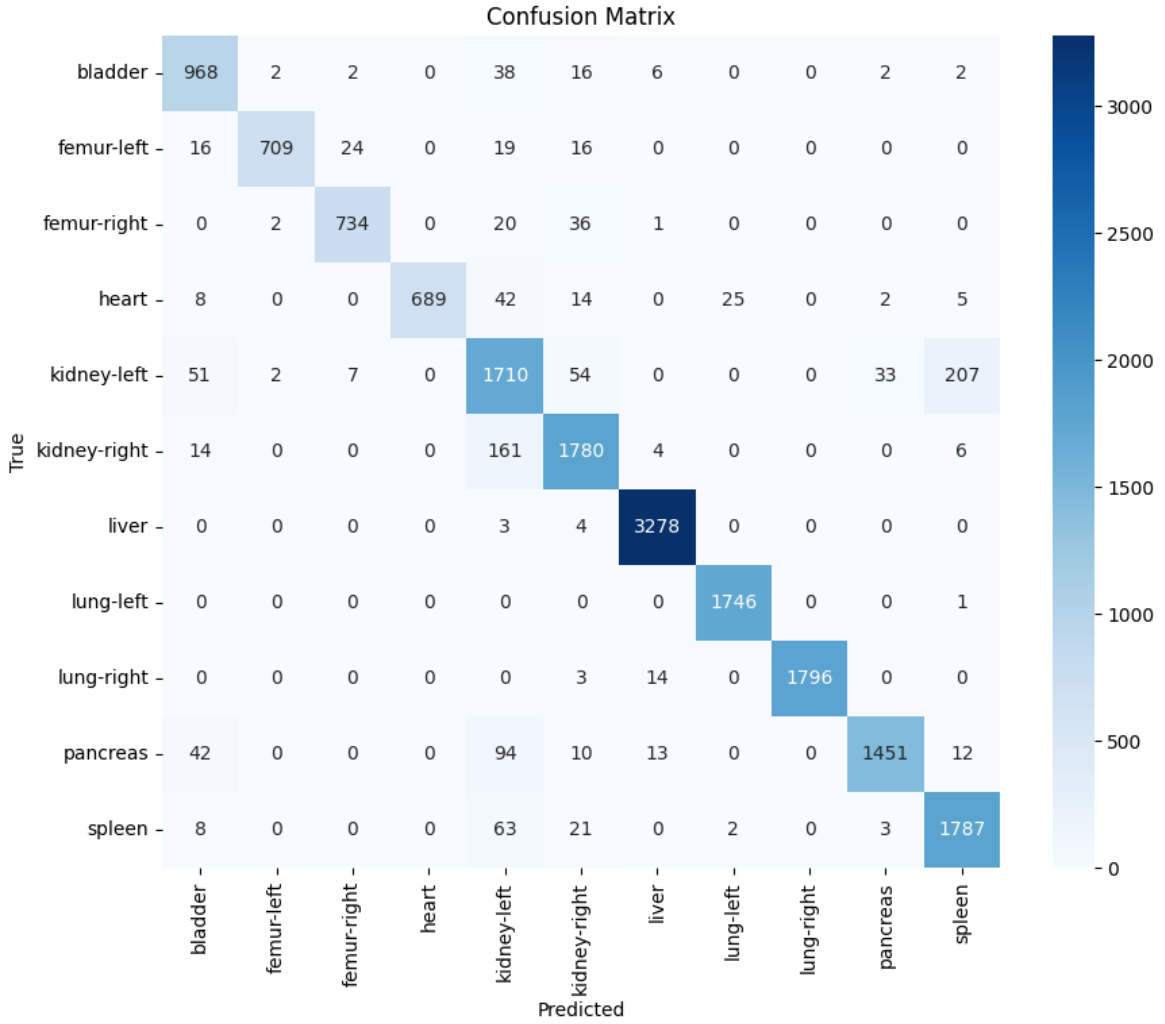
- OrganAMNIST: 11-class classification of 58830 organs CT sections (28x28 pixels)
- Resnet18 training – 5X cross validation (train/val: 55% calibration: 15% test: 30%)



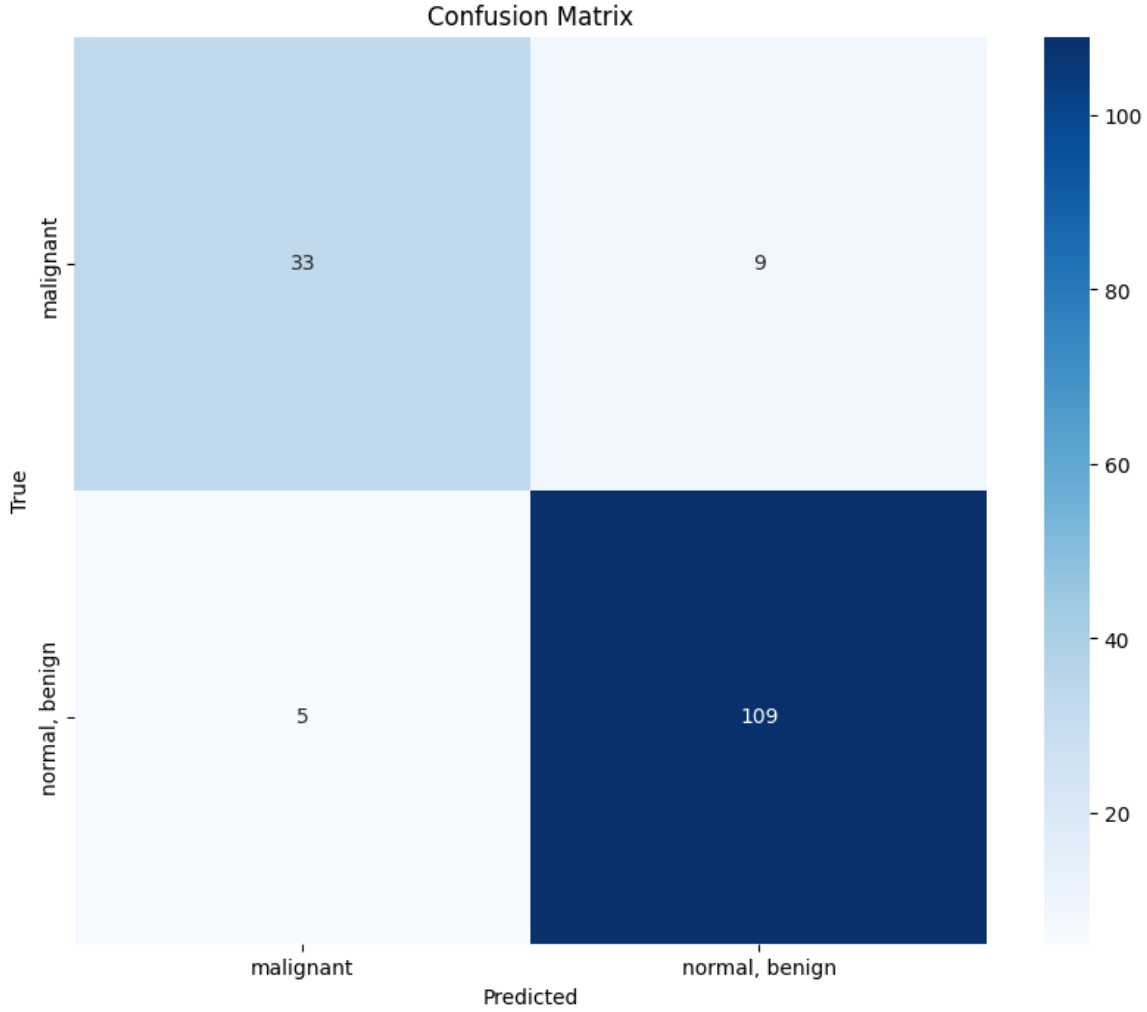


- BreastMNIST: binary classification of 780 US sections of breast tumors (28x28 pixels), malignant vs normal, benign
- Resnet18 training – 5X cross validation (train/val: 55% calibration: 15% test: 30%)

Resnet18 results



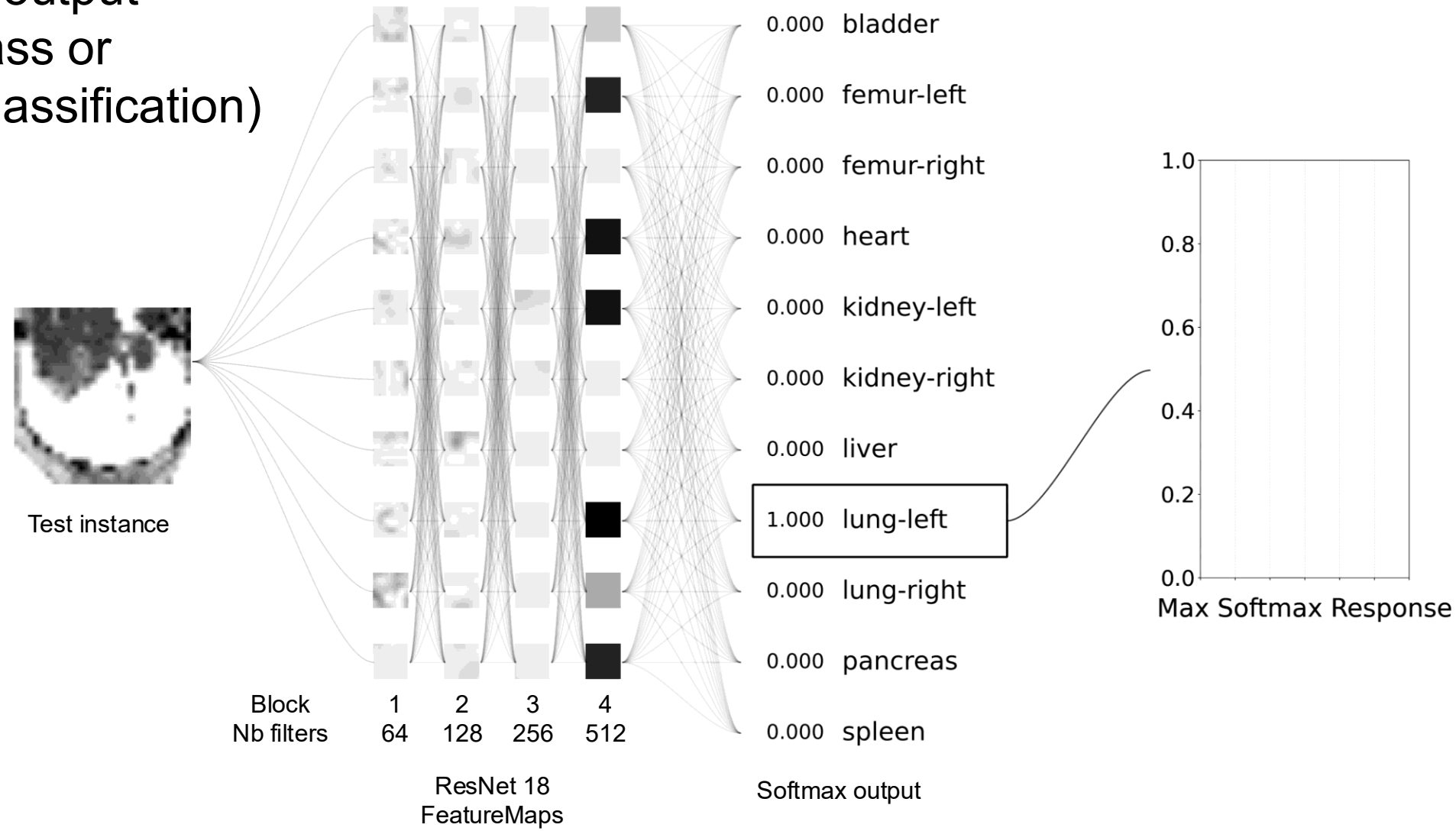
- Sensitivity: 0.93 | Specificity : 0.99 | bACC : 0.93 | AUC : 0.995



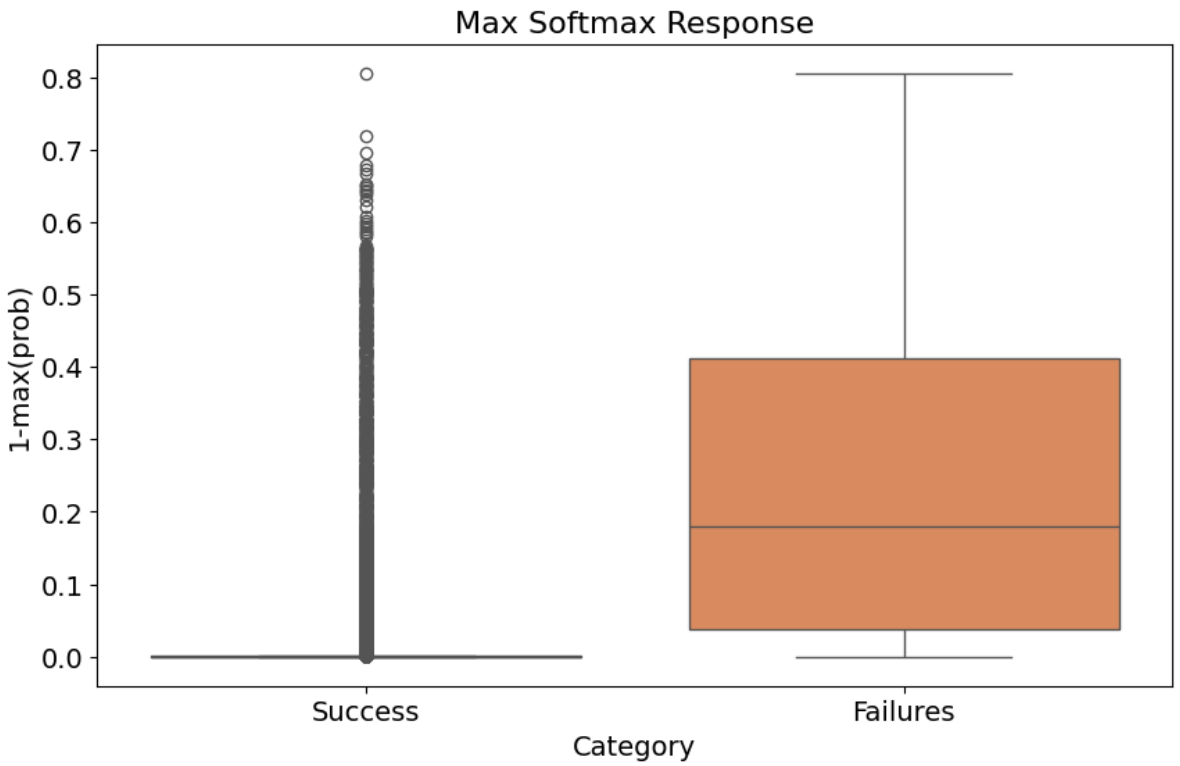
- Sensitivity: 0.96 | Specificity : 0.87 | bACC : 0.87 | AUC : 0.93

Max Softmax Response (MSR) – Distance to hard labels (DHL)

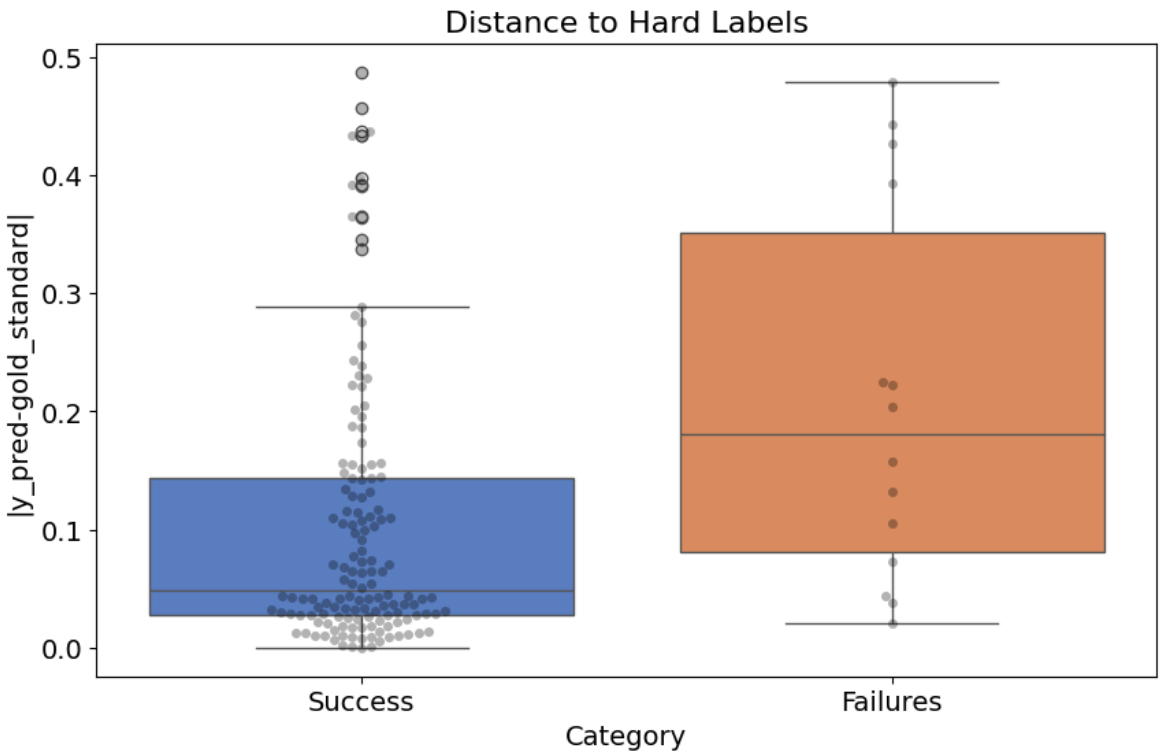
Direct use of model output (softmax for multiclass or sigmoid for binary classification)



Max Softmax Response (MSR) – Distance to Hard Labels (DHL)



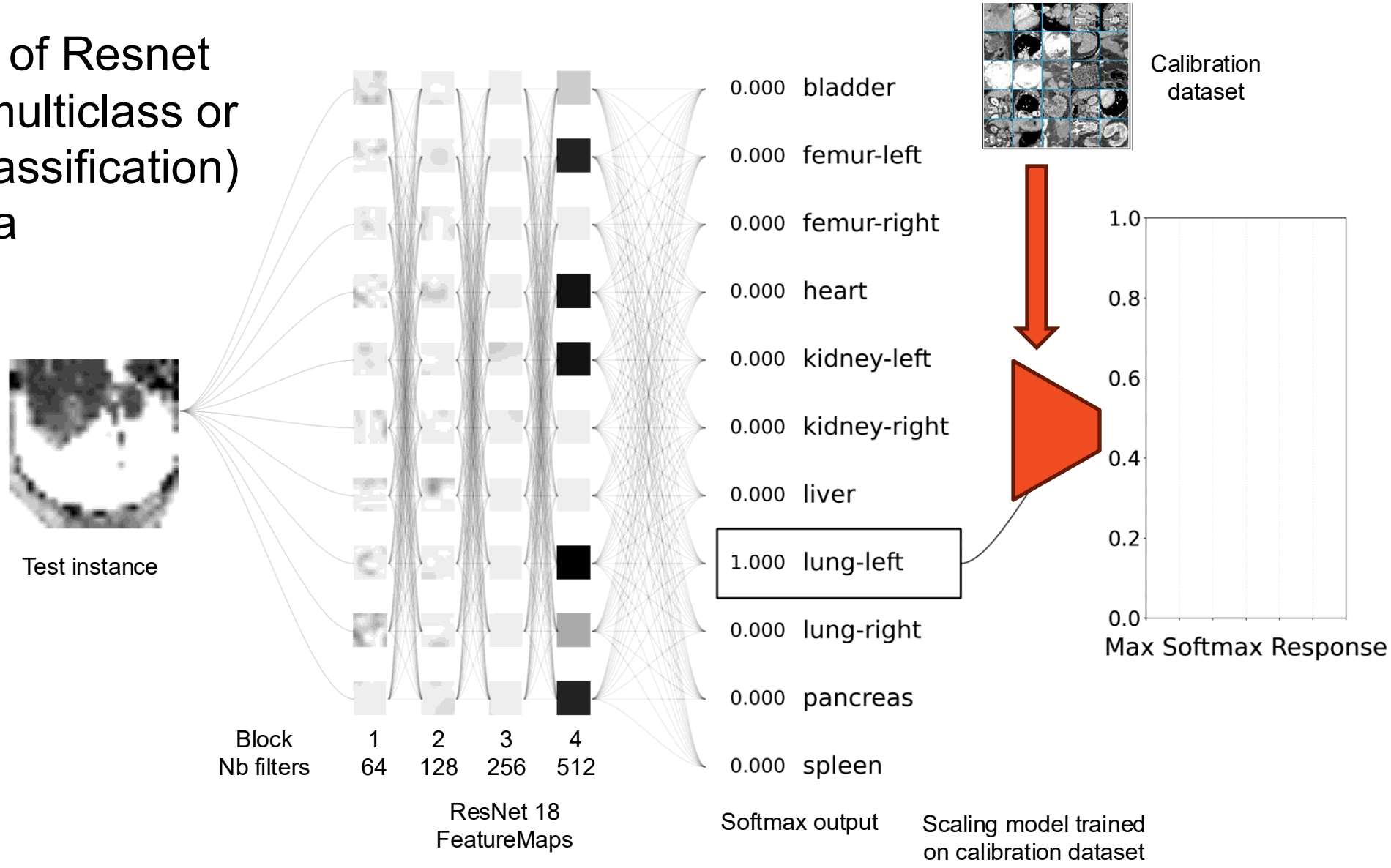
MSR OrganAMNIST
(AUC=0.94)



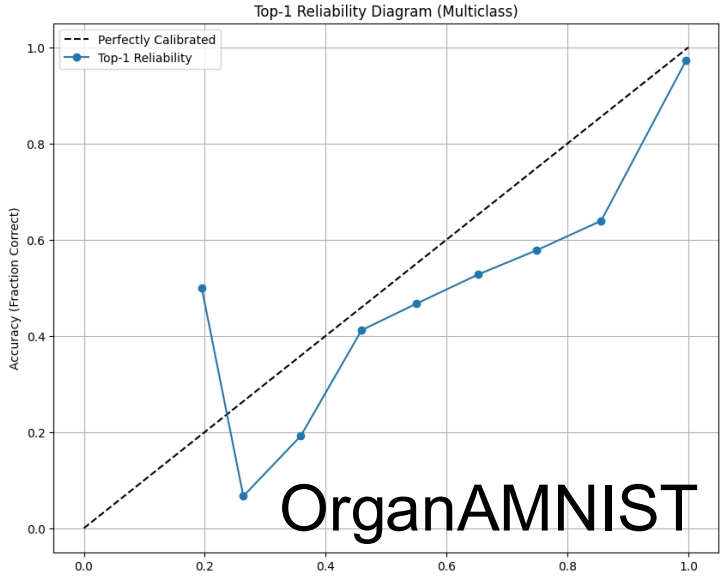
DHL BreastMNIST
(AUC=0.73)

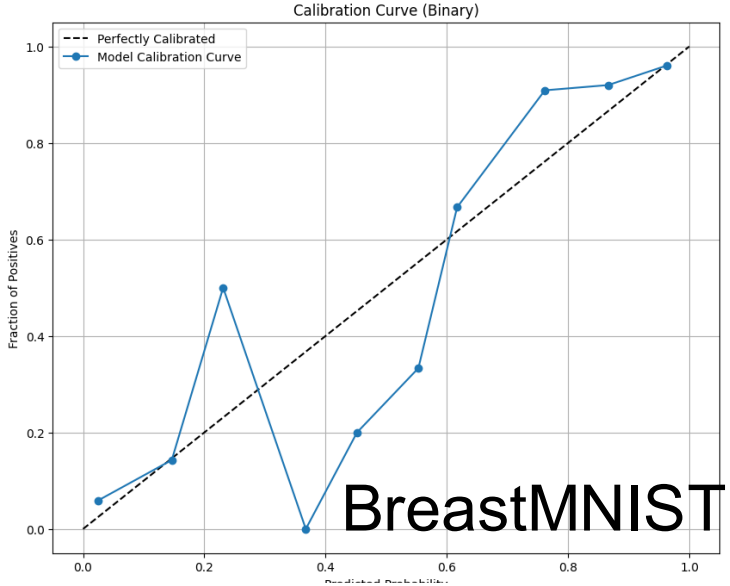
MSR – DHL after Recalibration using calibration set

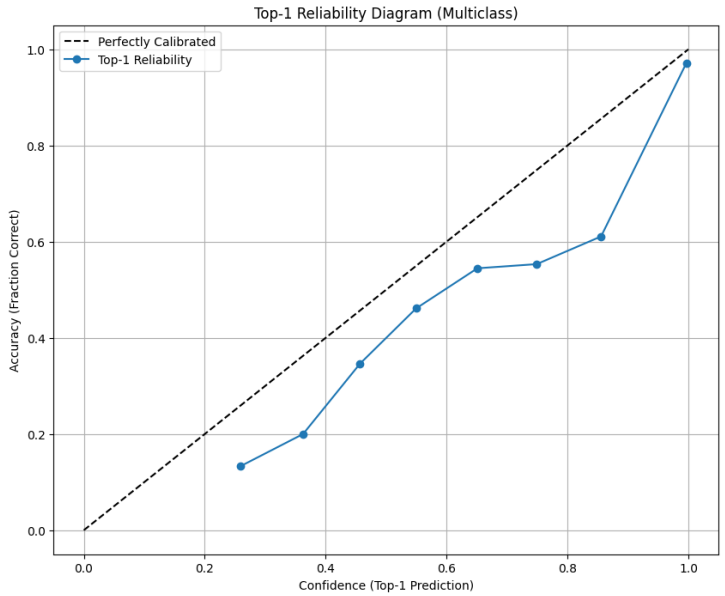
Temperature scaling of Resnet output (softmax for multiclass or sigmoid for binary classification) using calibration data

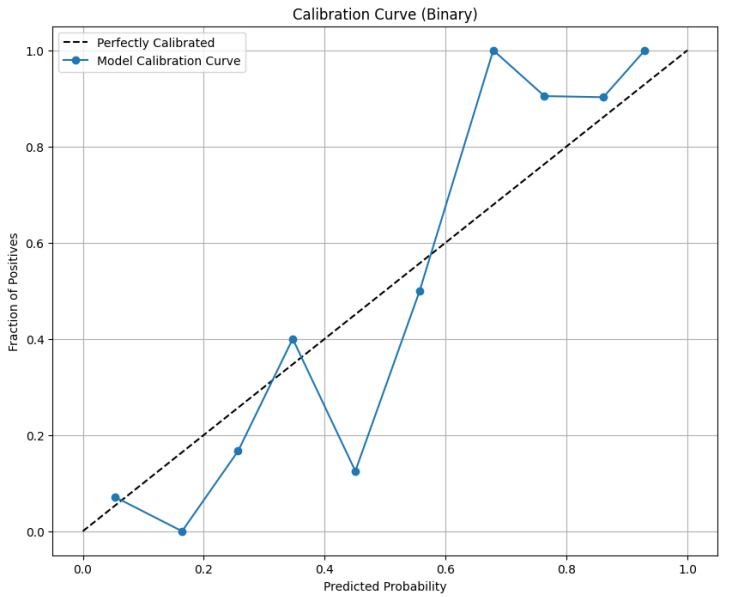


Before vs After model calibration

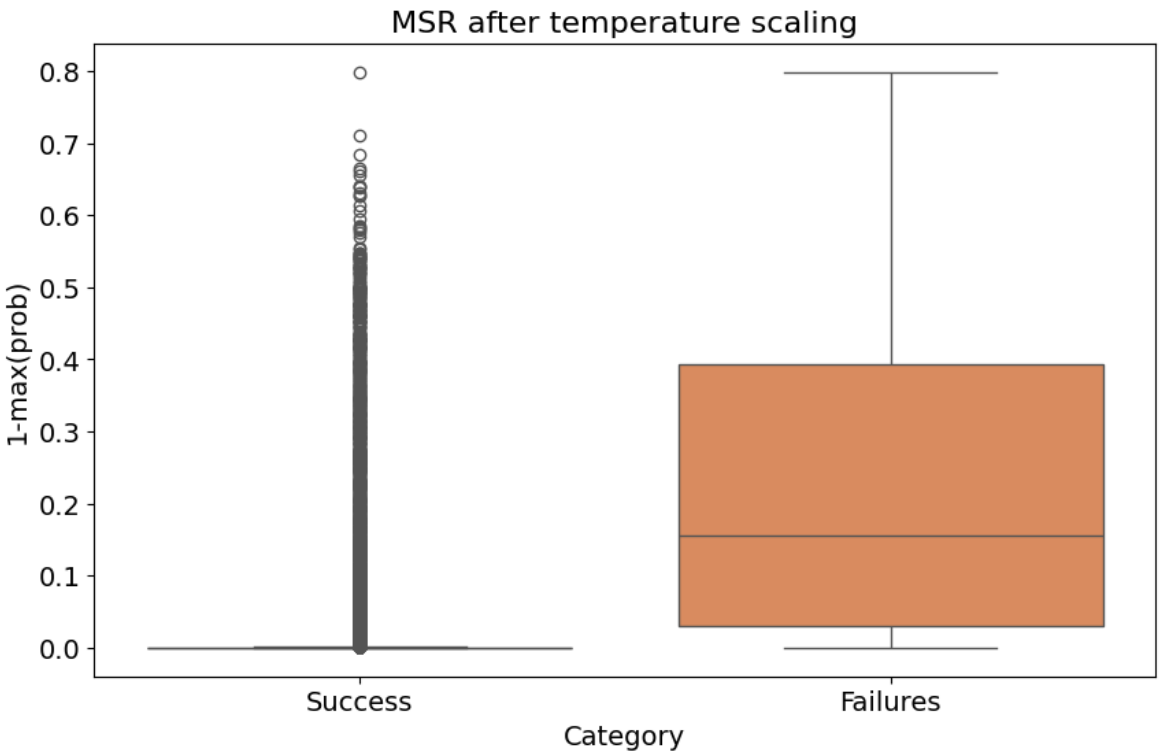


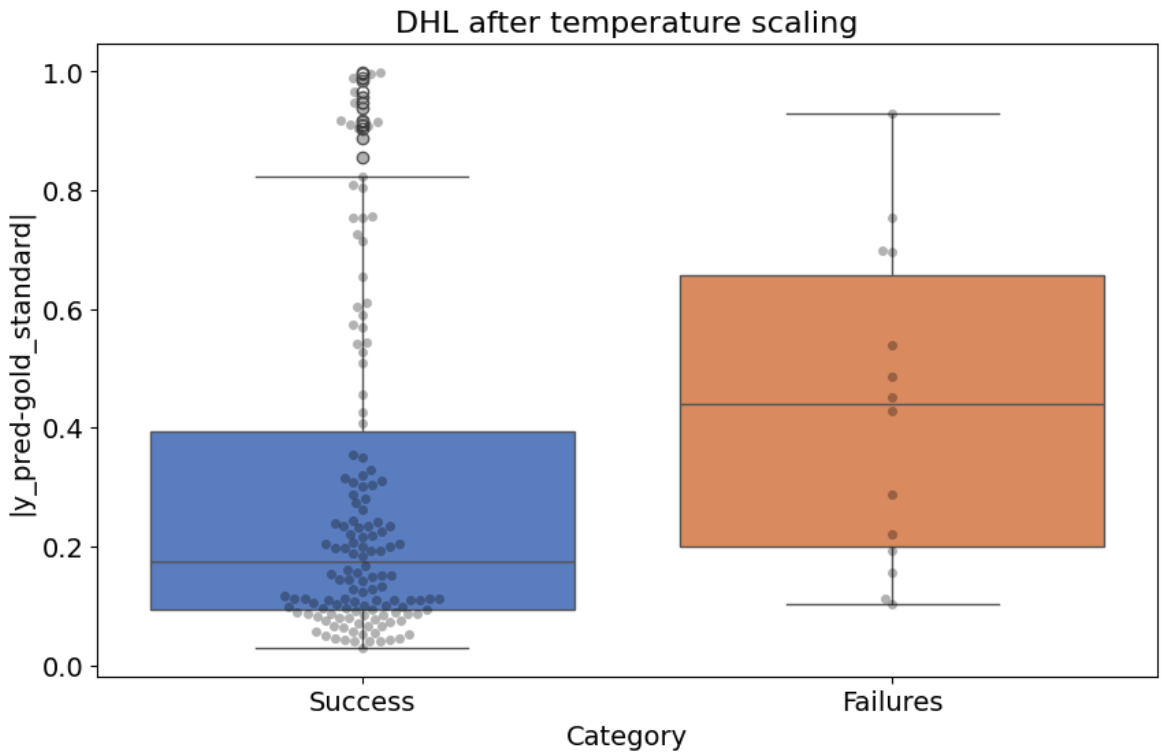






MSR – DHL after temperature scaling



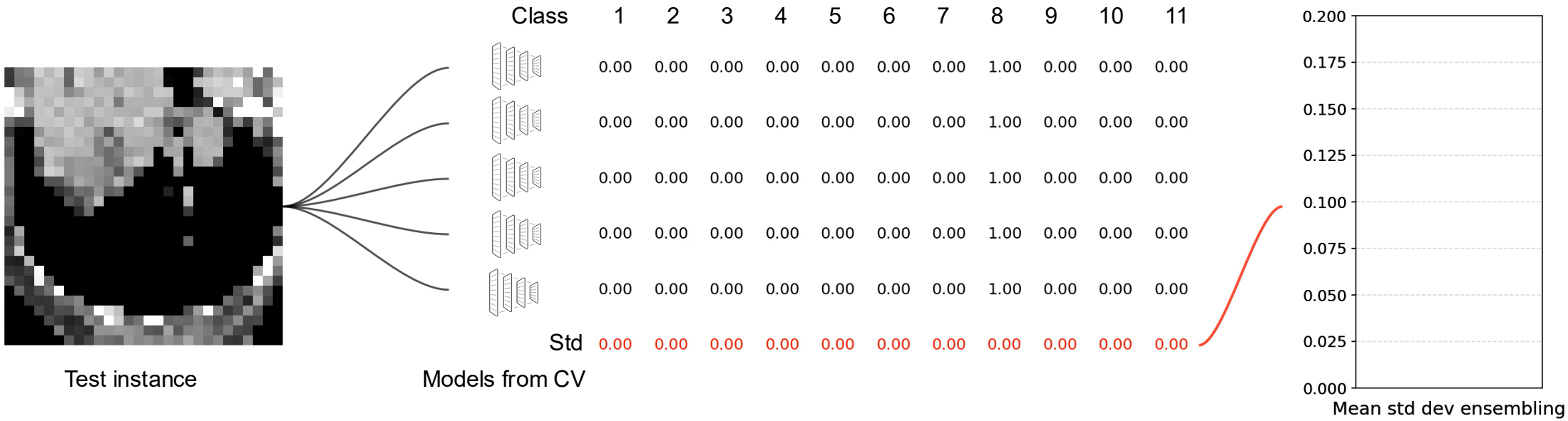


MSR OrganAMNIST
(AUC=0.94)

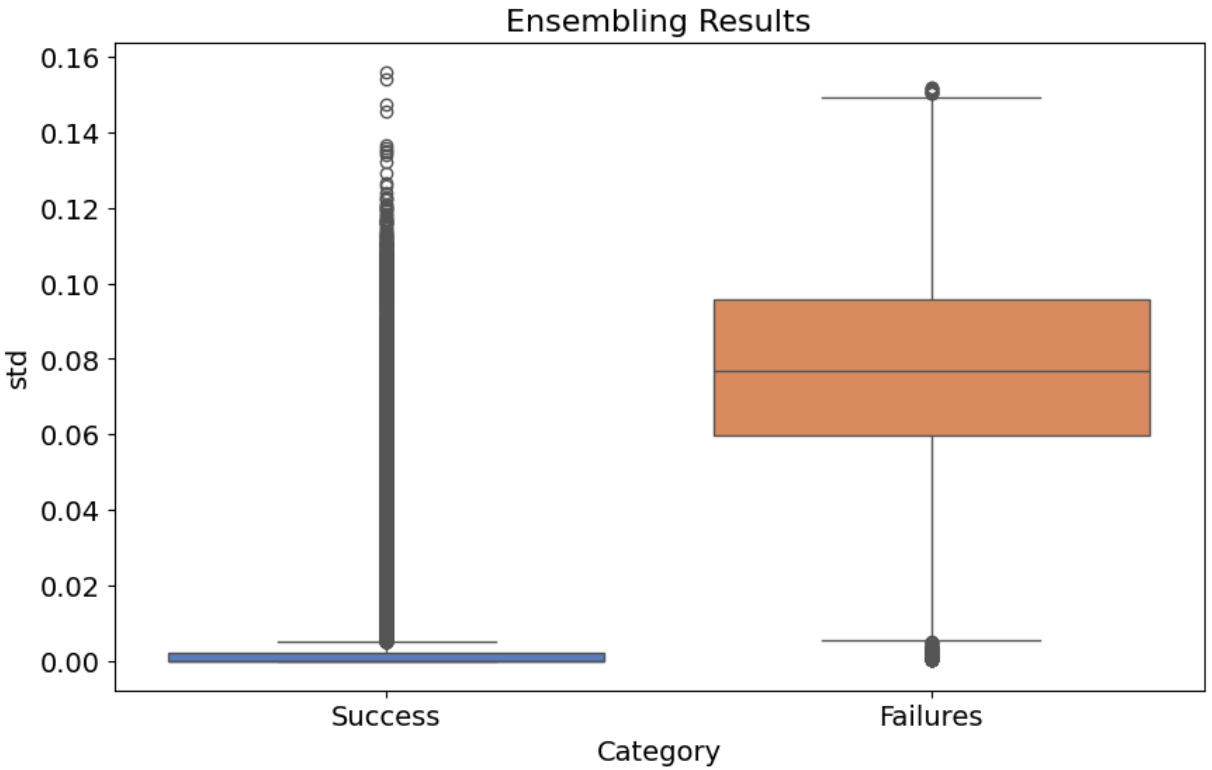
DHL BreastMNIST
(AUC=0.68)

Ensembling (Ens)

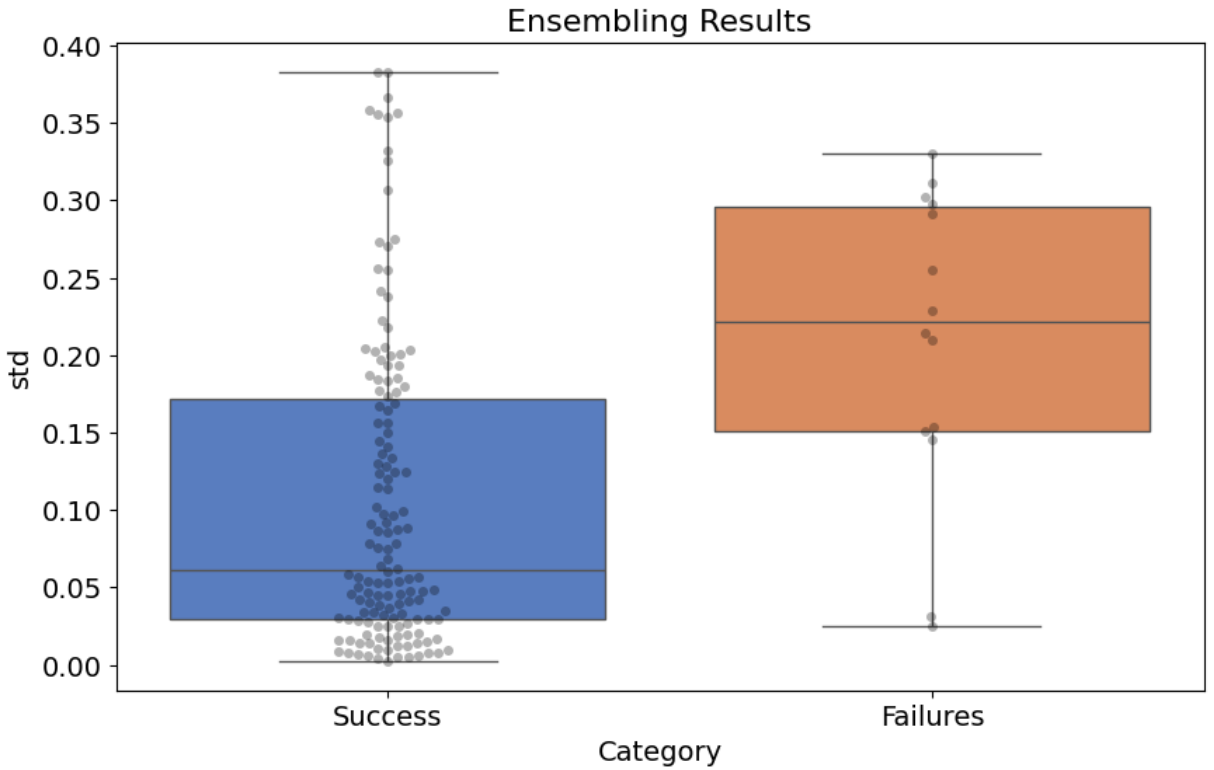
Standard deviation of models output (softmax for multiclass – sigmoid for binary classification)



Ensembling SD



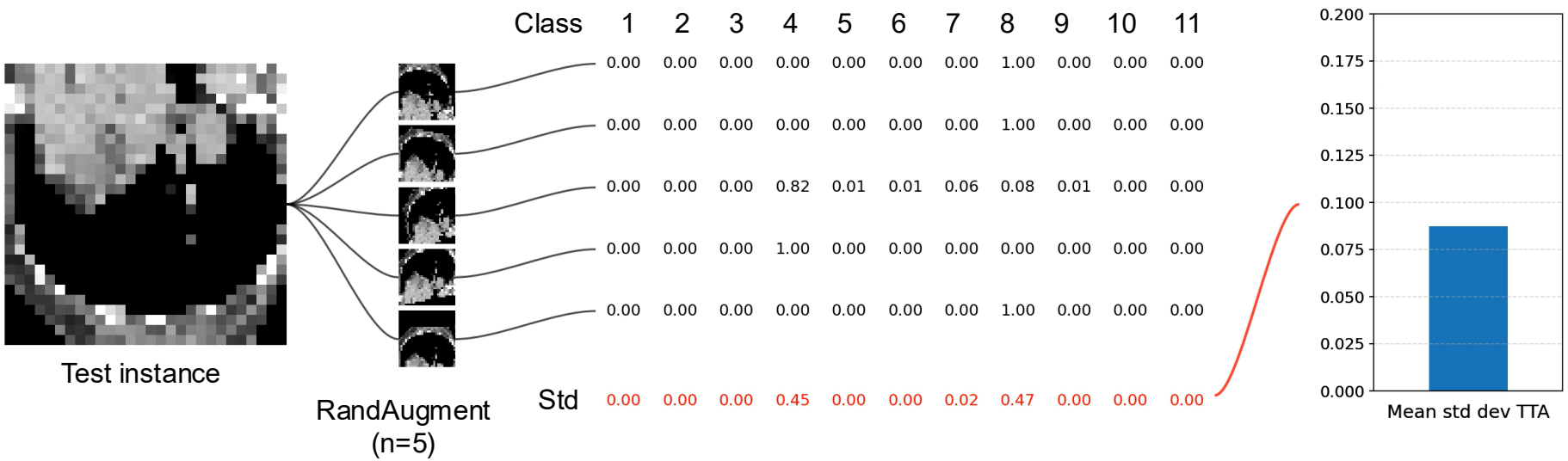
Ensembling std
OrganAMNIST (AUC=0.93)



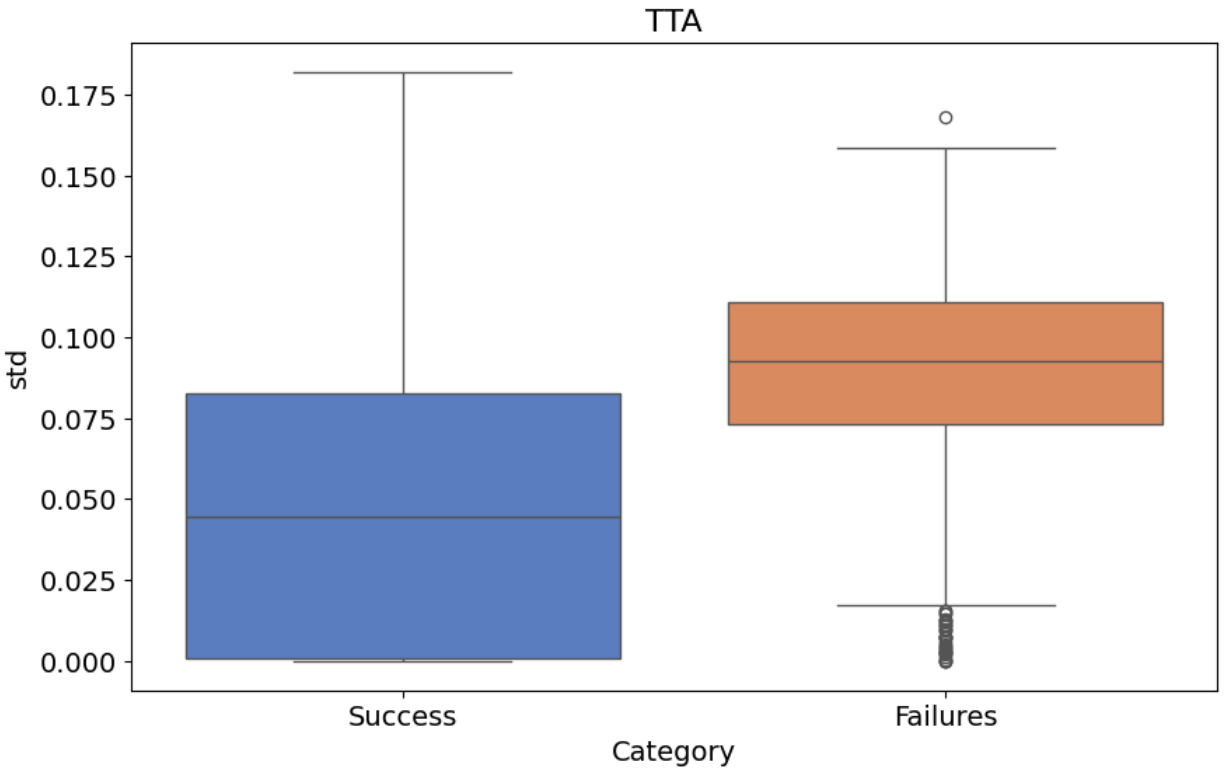
Ensembling std
BreastMNIST (AUC=0.77)

Test Time Augmentation (TTA)

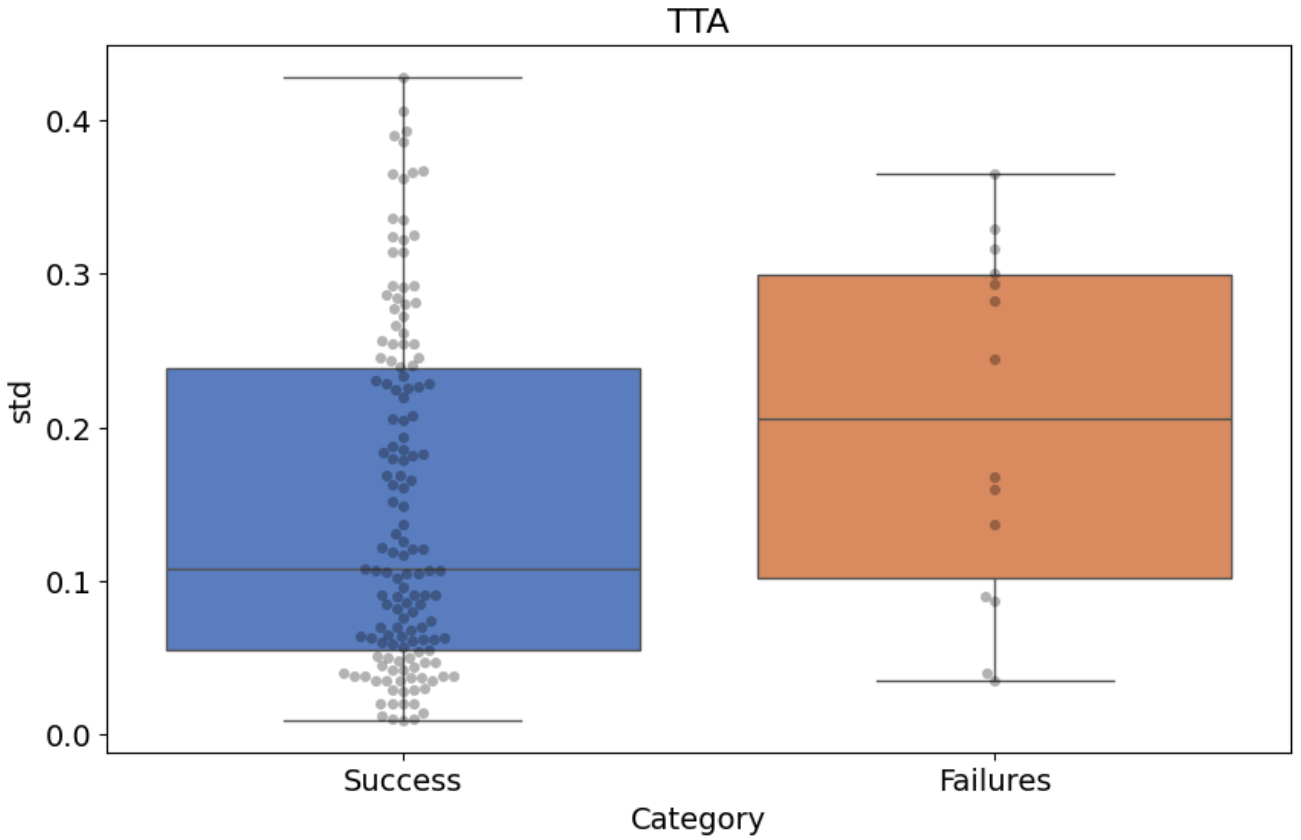
Standard deviation of predictions after random augmentations
(softmax for multiclass – sigmoid for binary classification)



Test Time Augmentation (TTA)



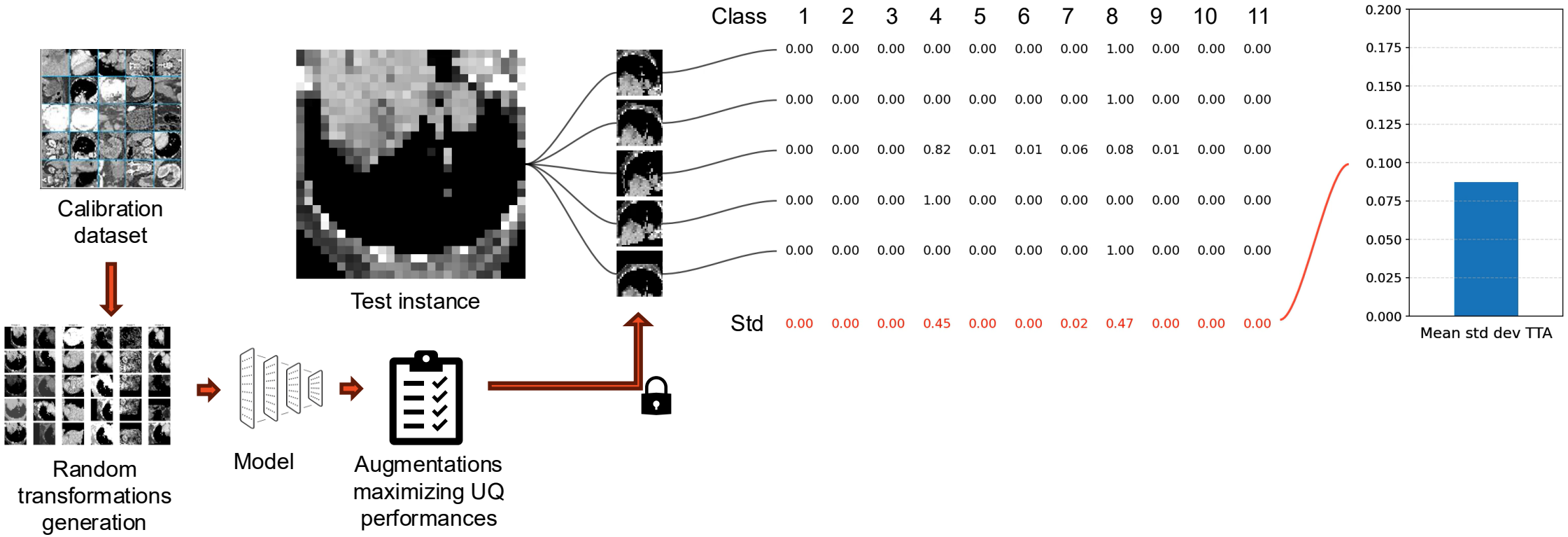
TTA std OrganAMNIST
(AUC=0.76)



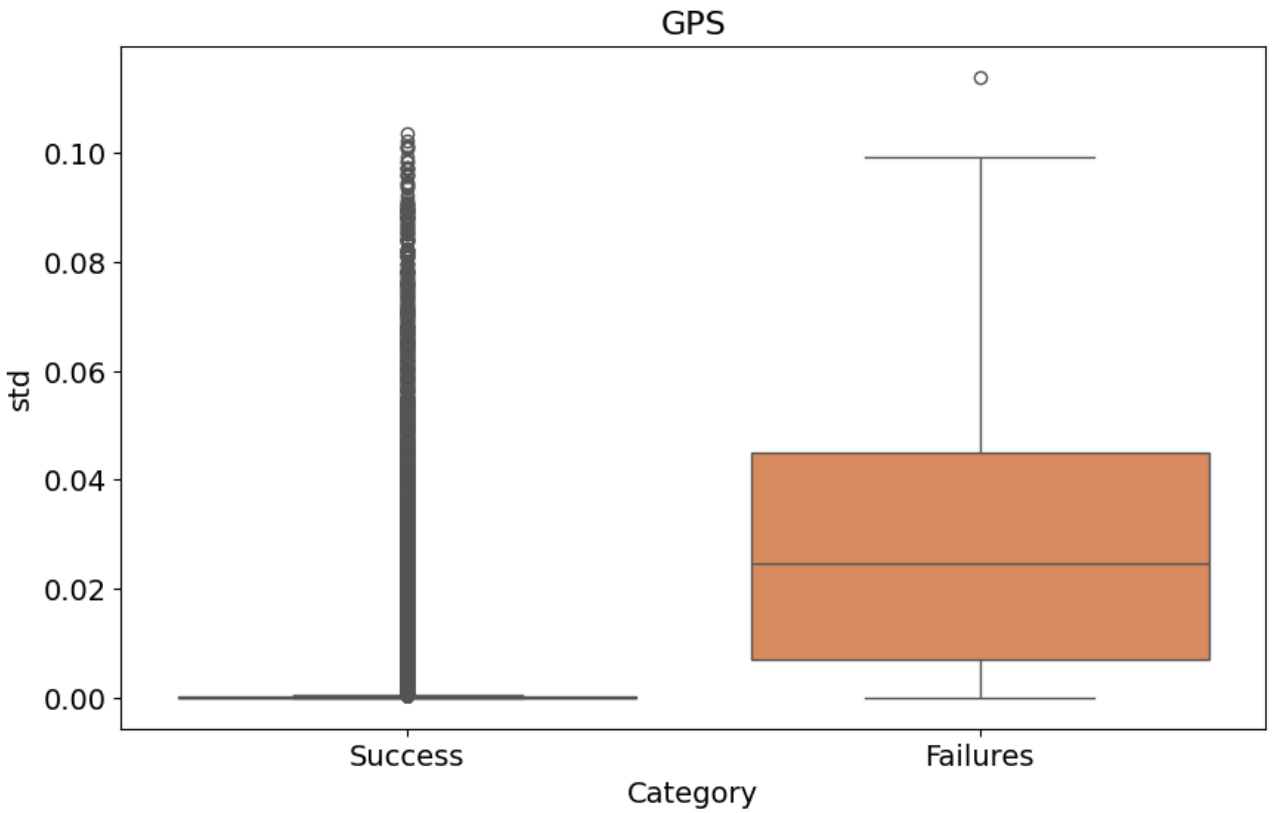
TTA std BreastMNIST
(AUC=0.64)

Greedy Policy Search (GPS)

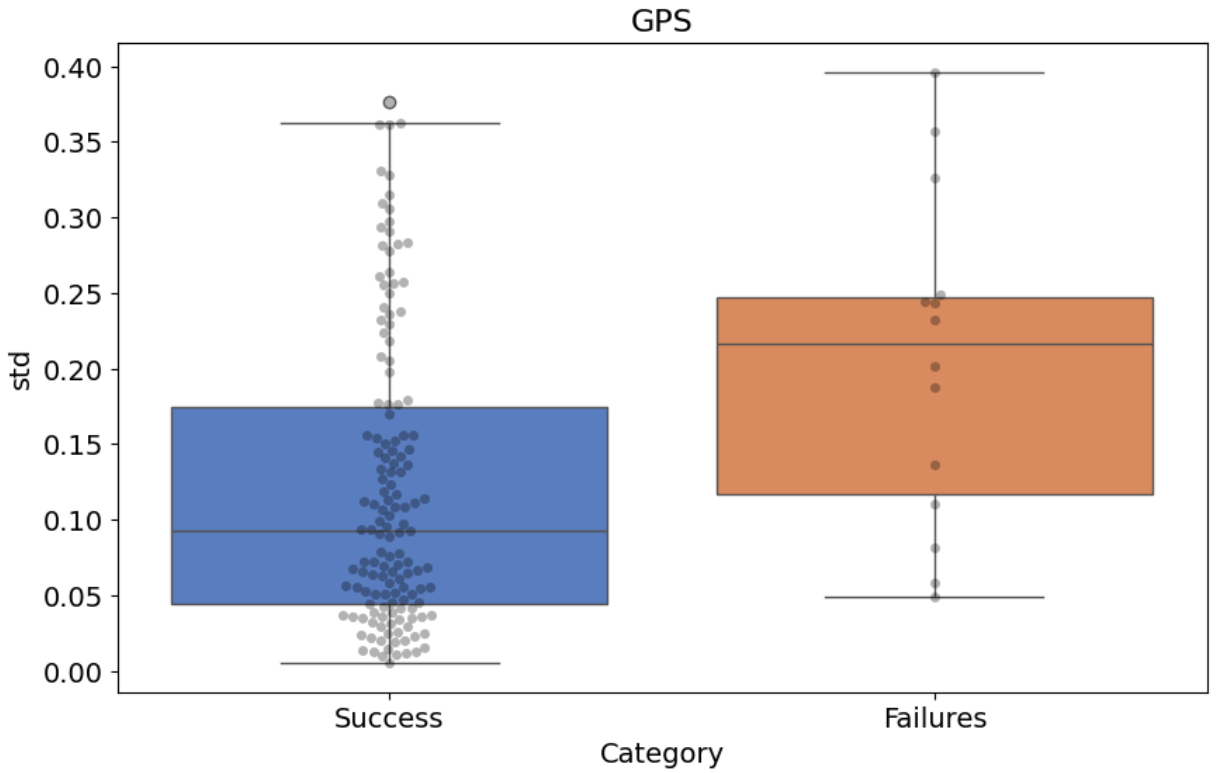
Standard deviation of predictions after learned augmentations using a calibration dataset (softmax for multiclass – sigmoid for binary classification)



Greedy Policy Search (GPS)



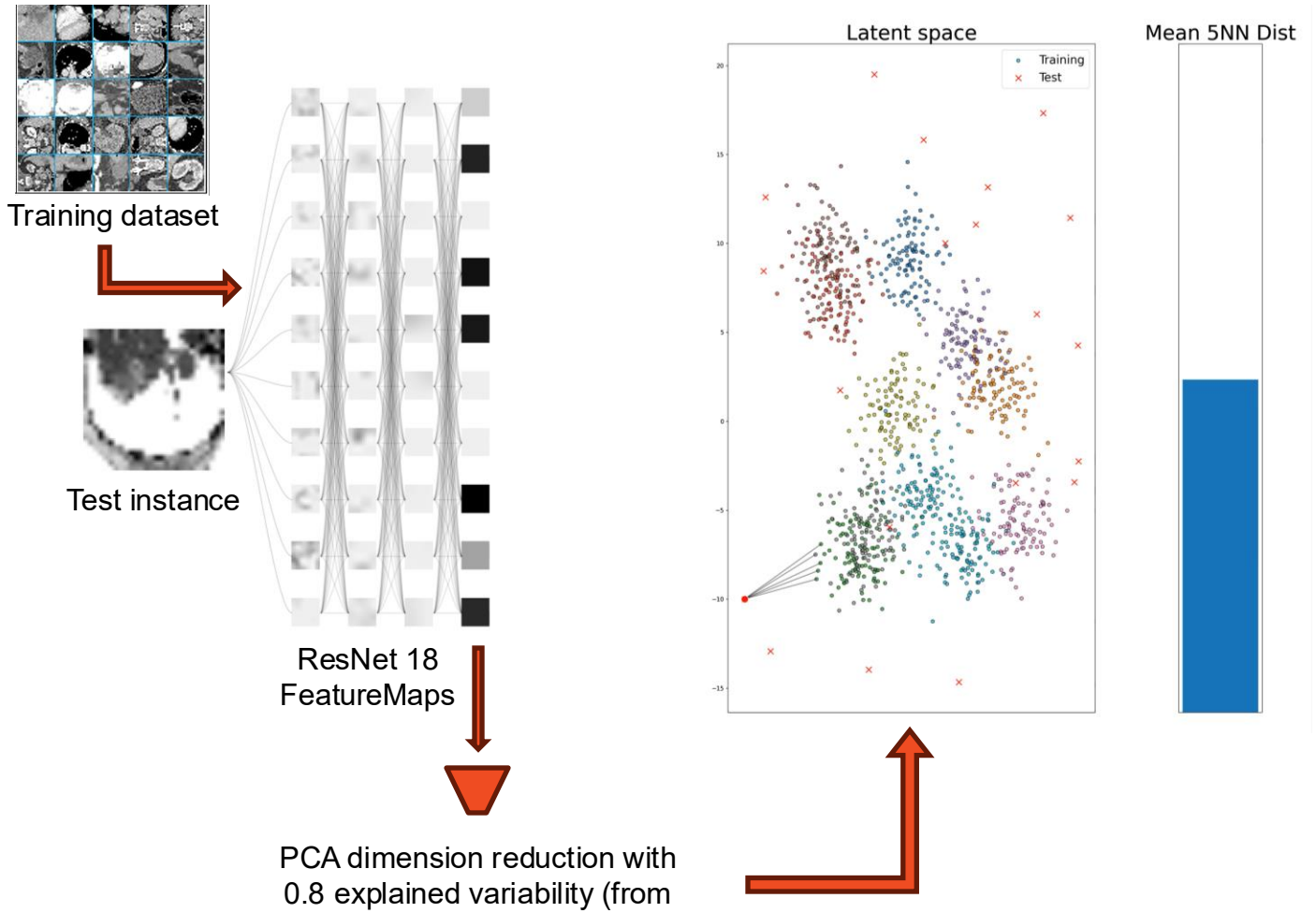
GPS std OrganAMNIST
(AUC=0.93)



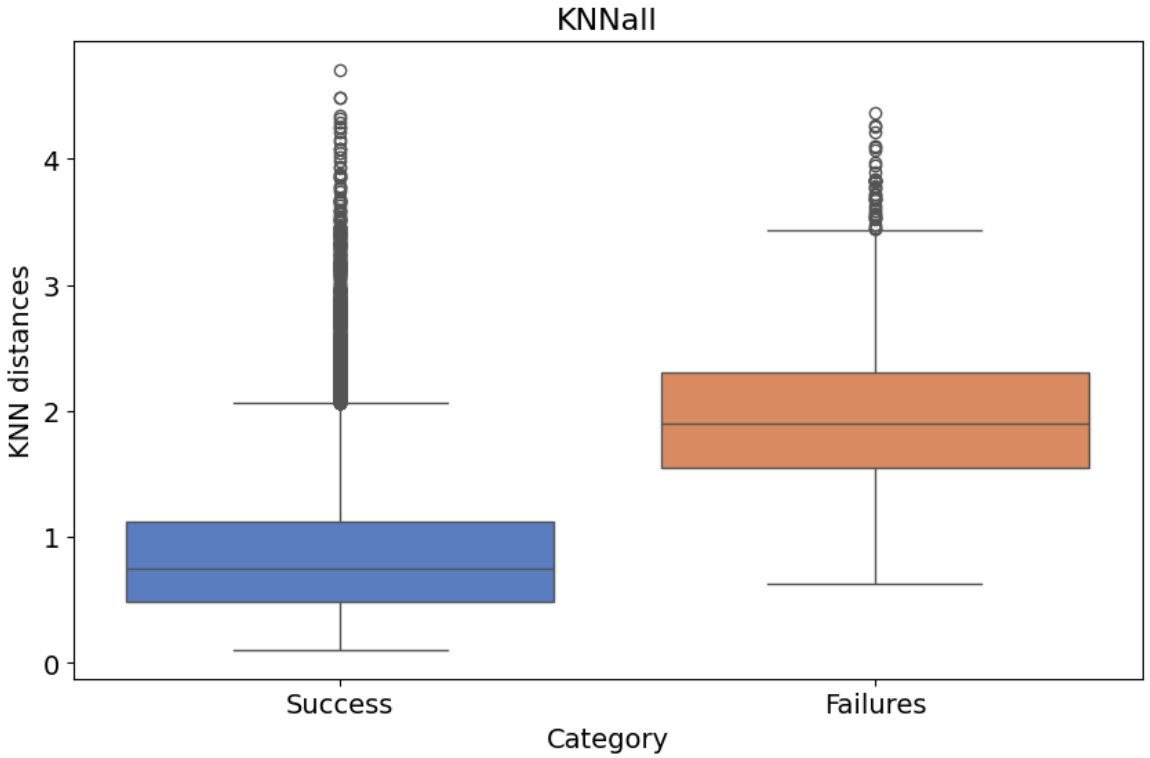
GPS std BreastMNIST
(AUC=0.73)

Distance to KNN in latent space (KNNall)

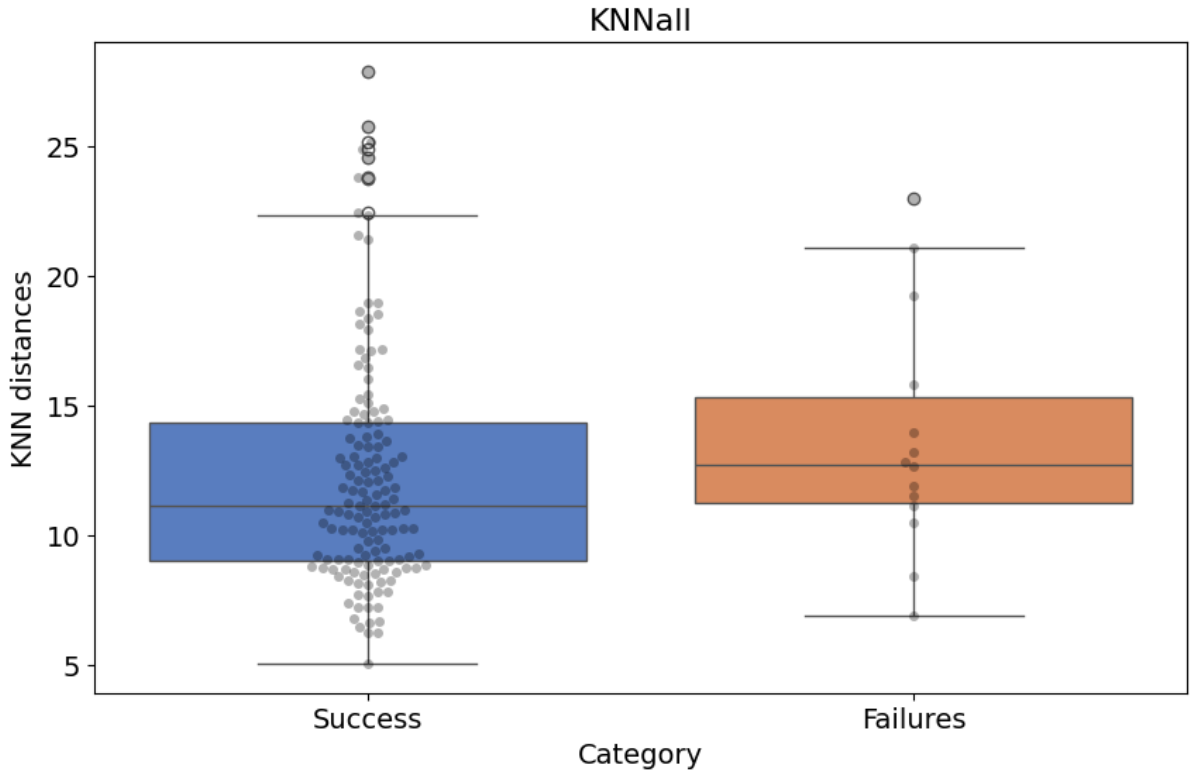
Nearest kneighbors distance with training data in latent space



Distance to KNN in latent space (KNNall)



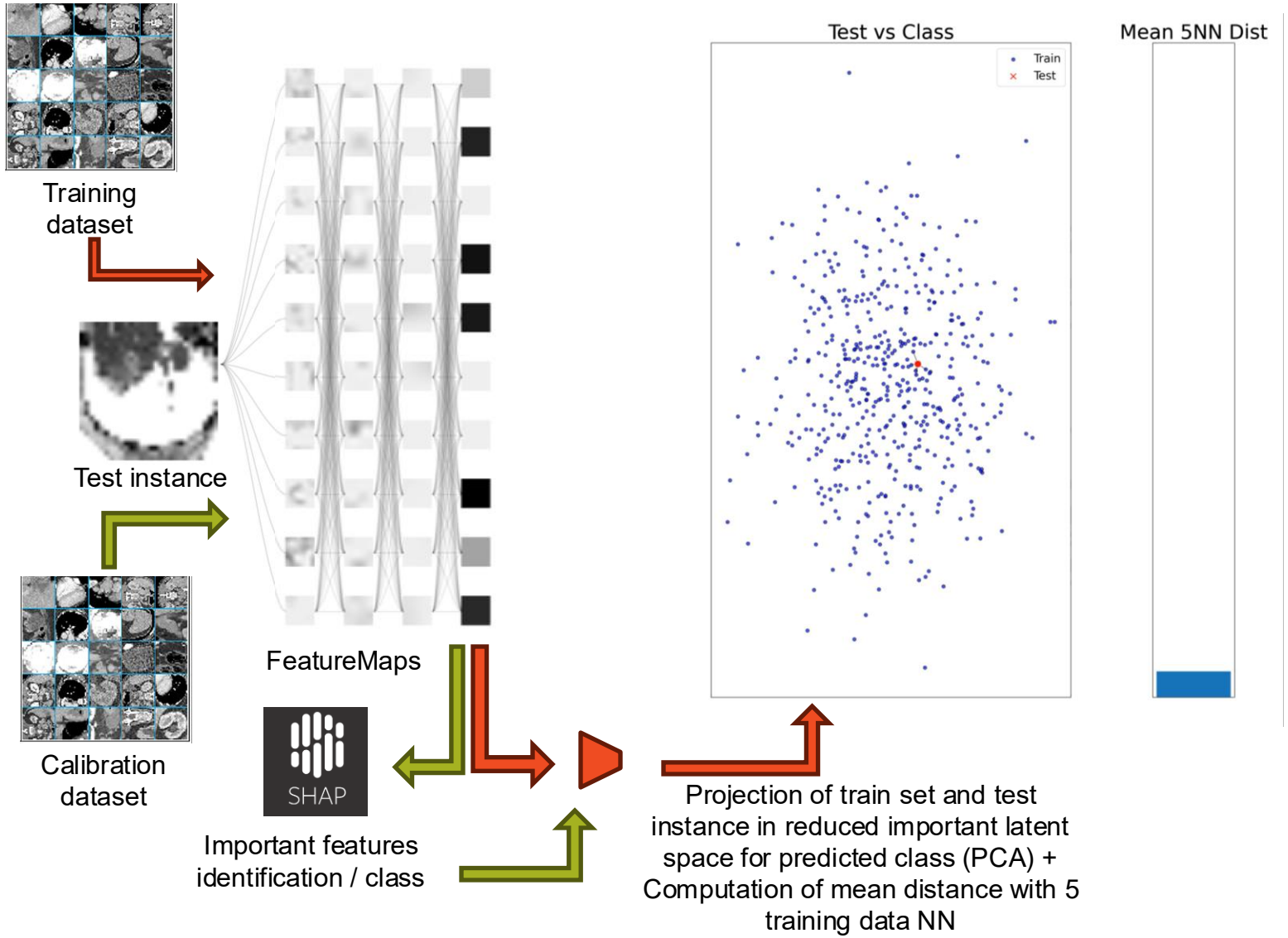
KNNall OrganAMNIST
(AUC=0.91)



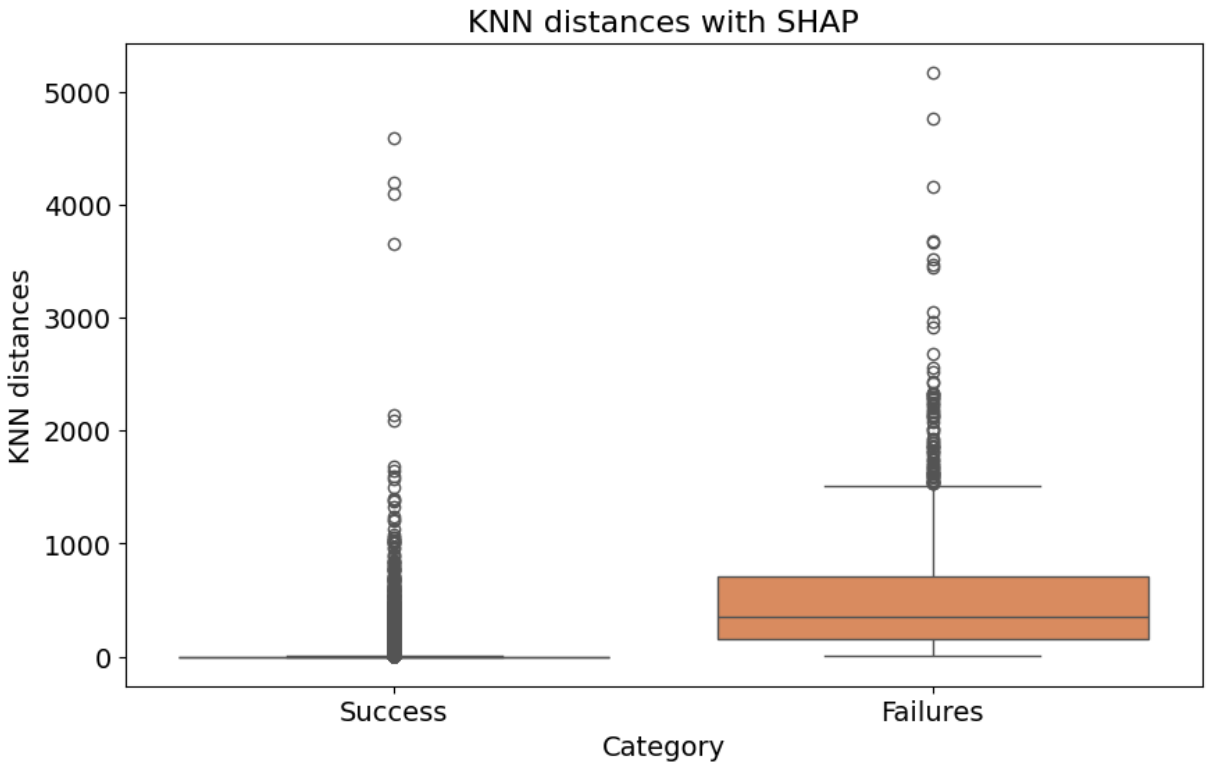
KNNall BreastMNIST
(AUC=0.61)

Distance to KNN in latent space after features selection (KNNshap)

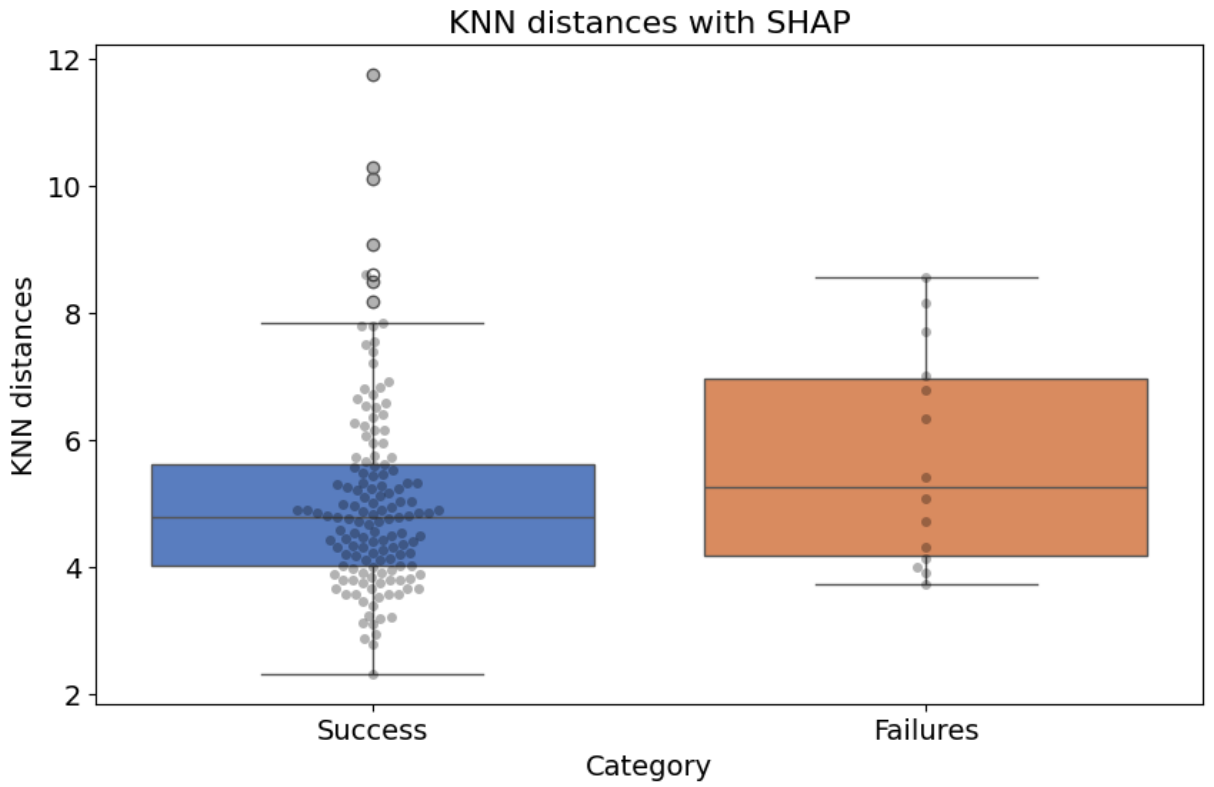
Nearest kneighbors distance with training data in latent space after class-features selection using Shap values



Distance to KNN in latent space after features selection (KNNshap)

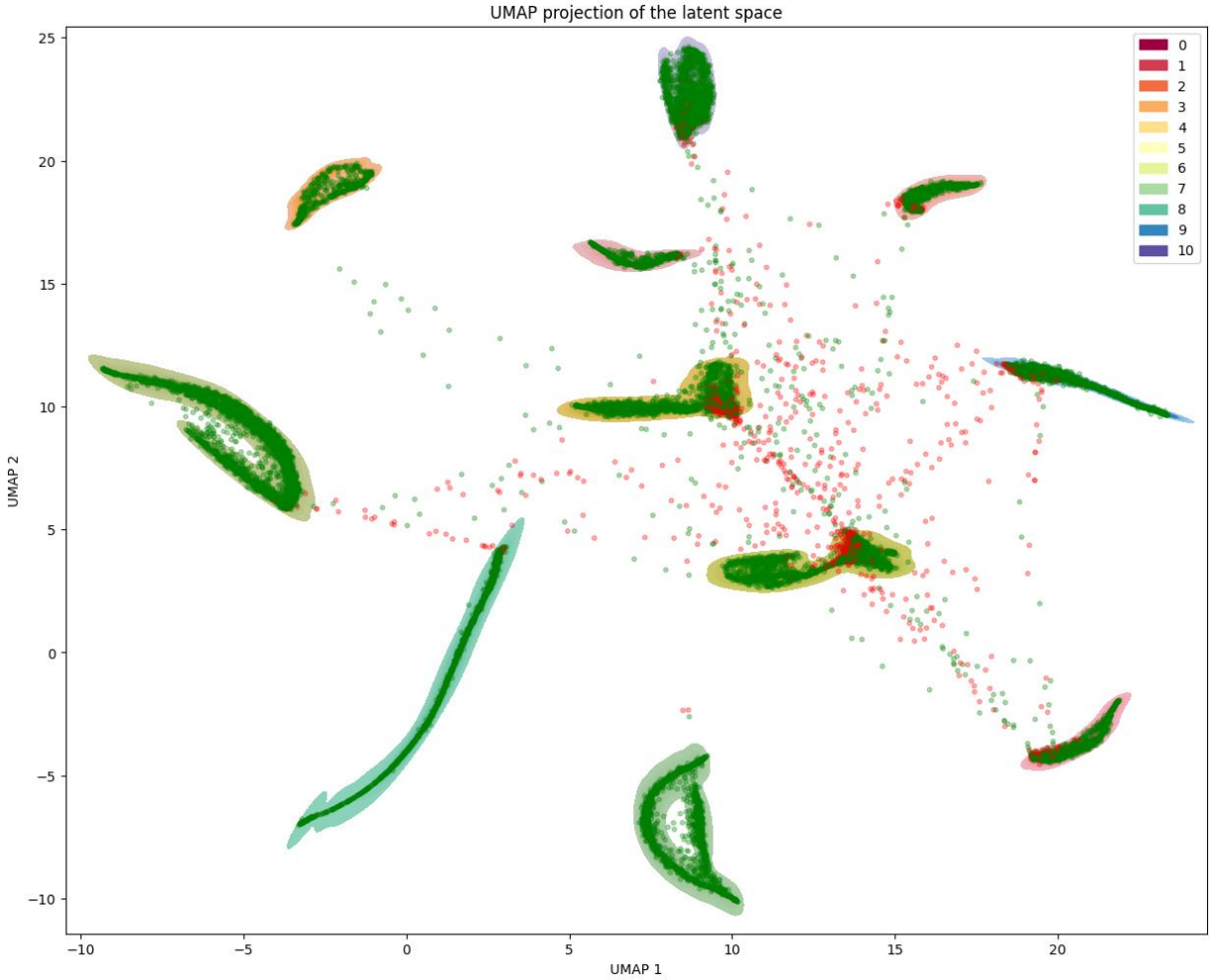


KNNshap OrganAMNIST
(AUC=0.98)

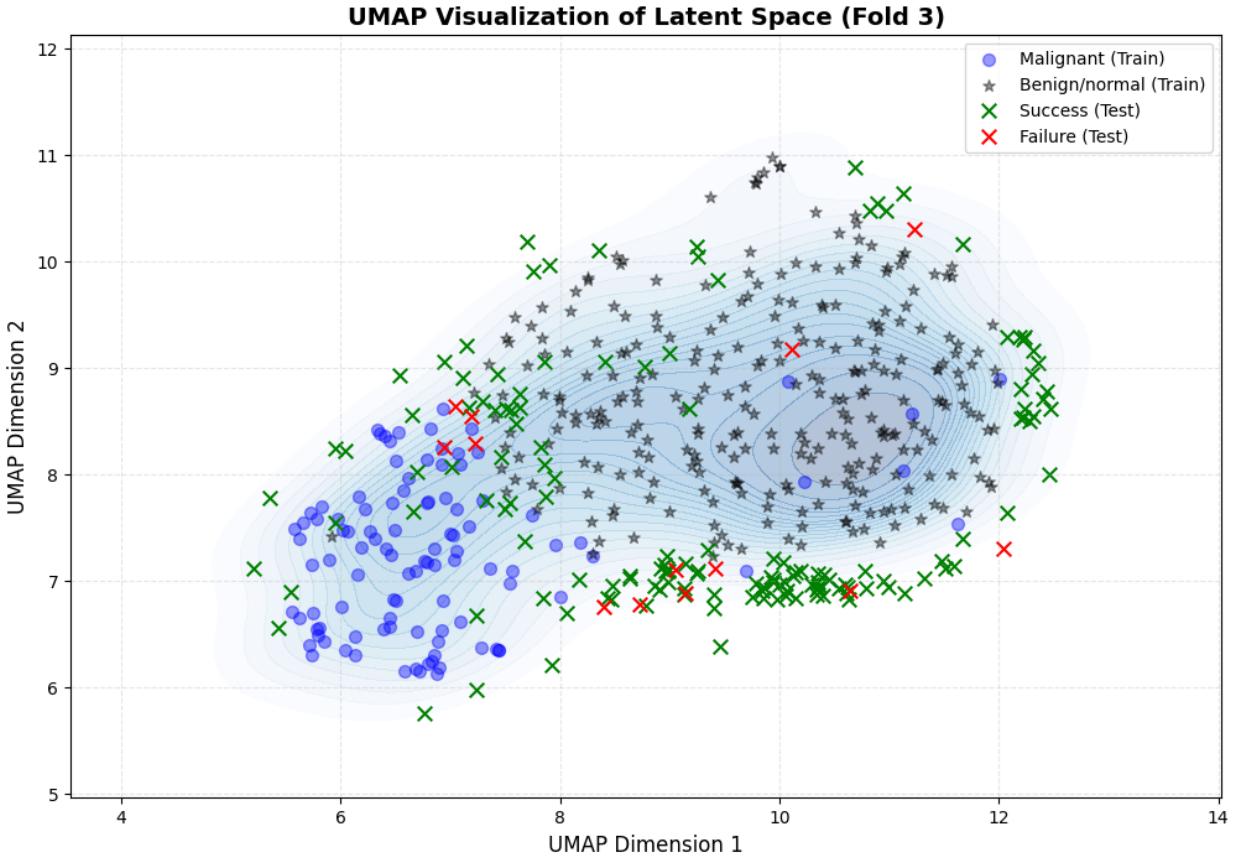


KNNshap BreastMNIST
(AUC=0.6)

UMAP latent space examples

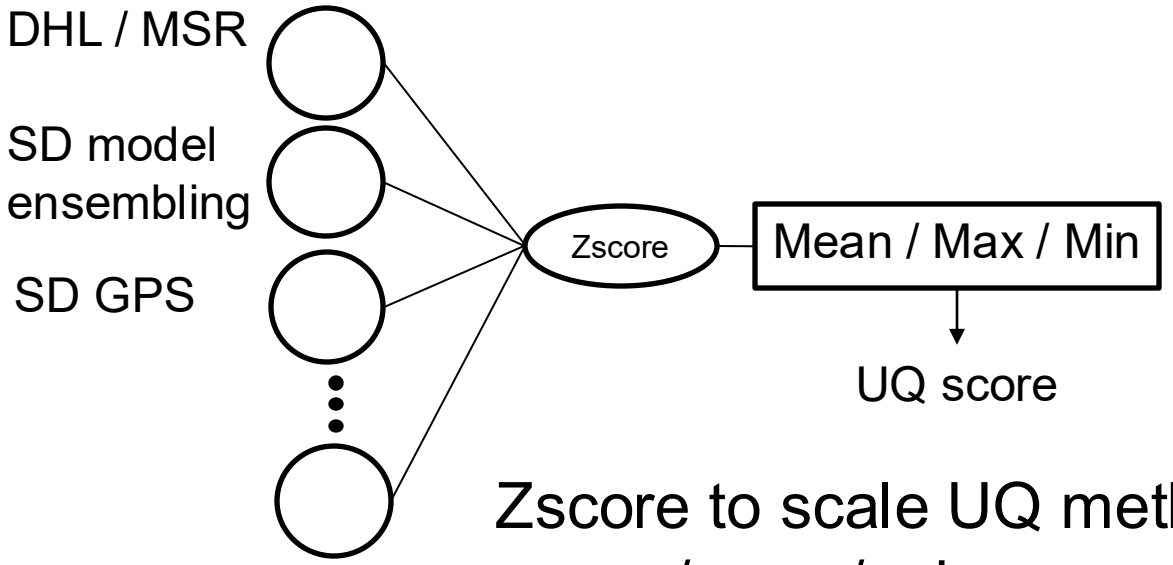


OrganAMNIST

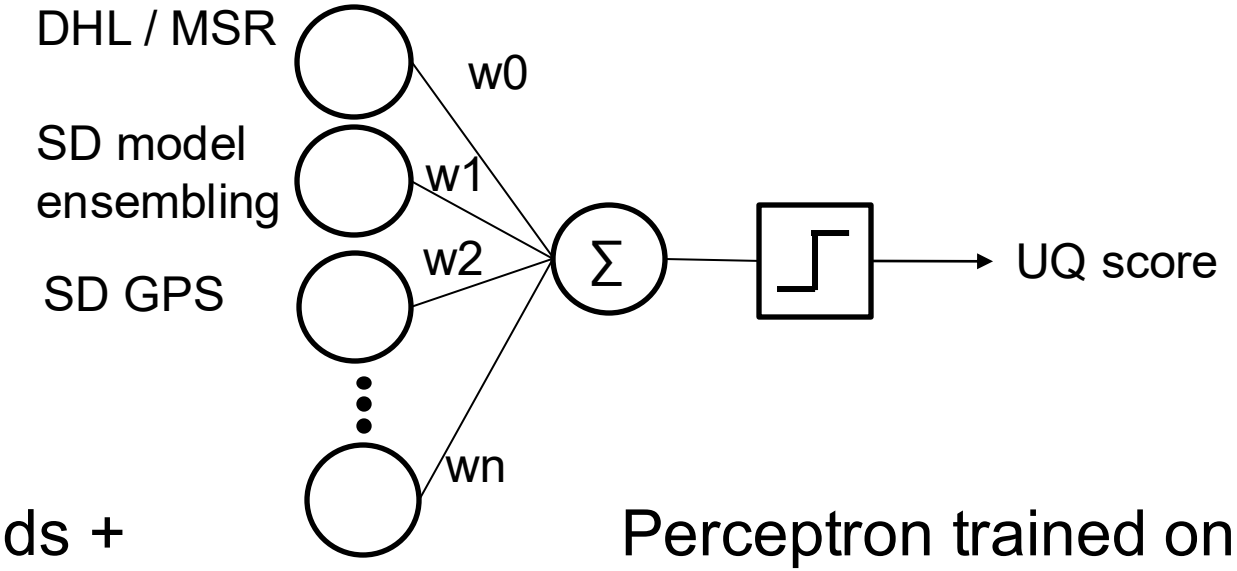


BreastMNIST

Methods combination (Mean, Max, Min, Perceptron)



Zscore to scale UQ methods + mean / max / min ensembling



Perceptron trained on calibration data

	MSR	Temp scaling output	Ensembling	TTA	GPS	KNNall	KNNshap
Softmax/Sigmoid output	Y	Y	Y	Y	Y	Y	Y
Models outputs	N	N	Y	N	N	N	N
Train data	N	N	N	N	N	Y	Y
Calibration data	N	Y	N	N	Y	N	Y
Final layer access	N	N	N	N	N	Y	Y

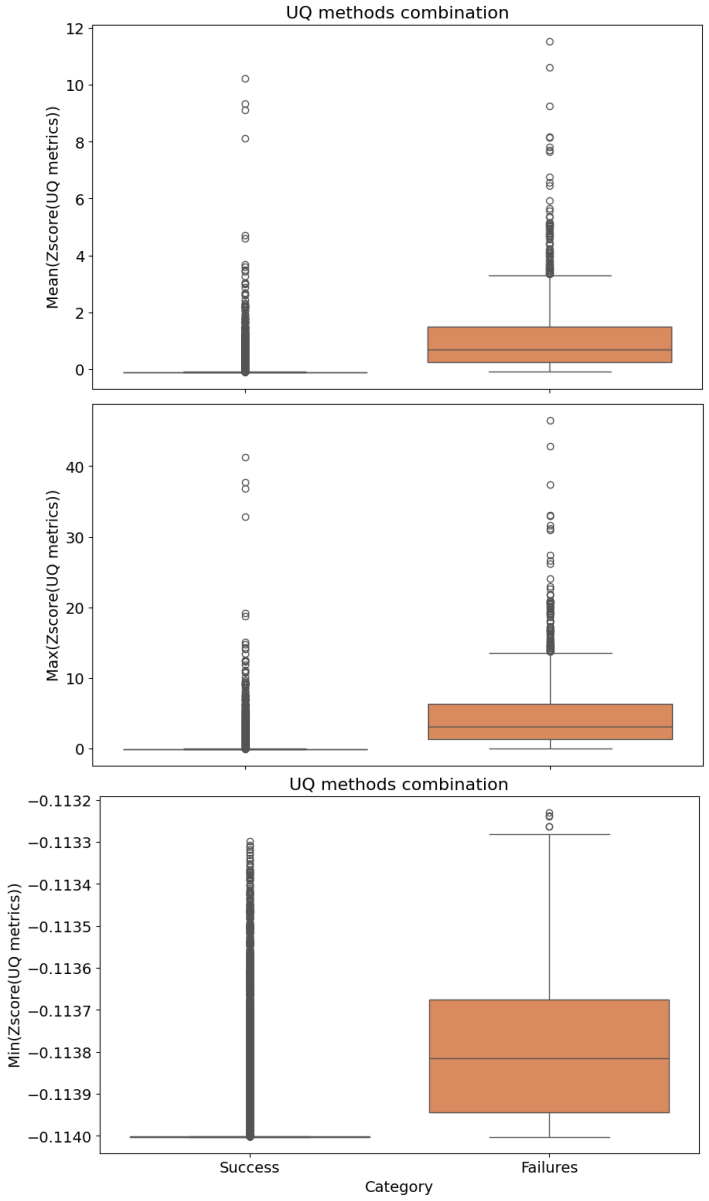
Methods combination (Mean, Max, Min)

OrganAMNIST

Mean Comb
(AUC=0.98)

Max Comb
(AUC=0.98)

Min Comb
(AUC=0.94)

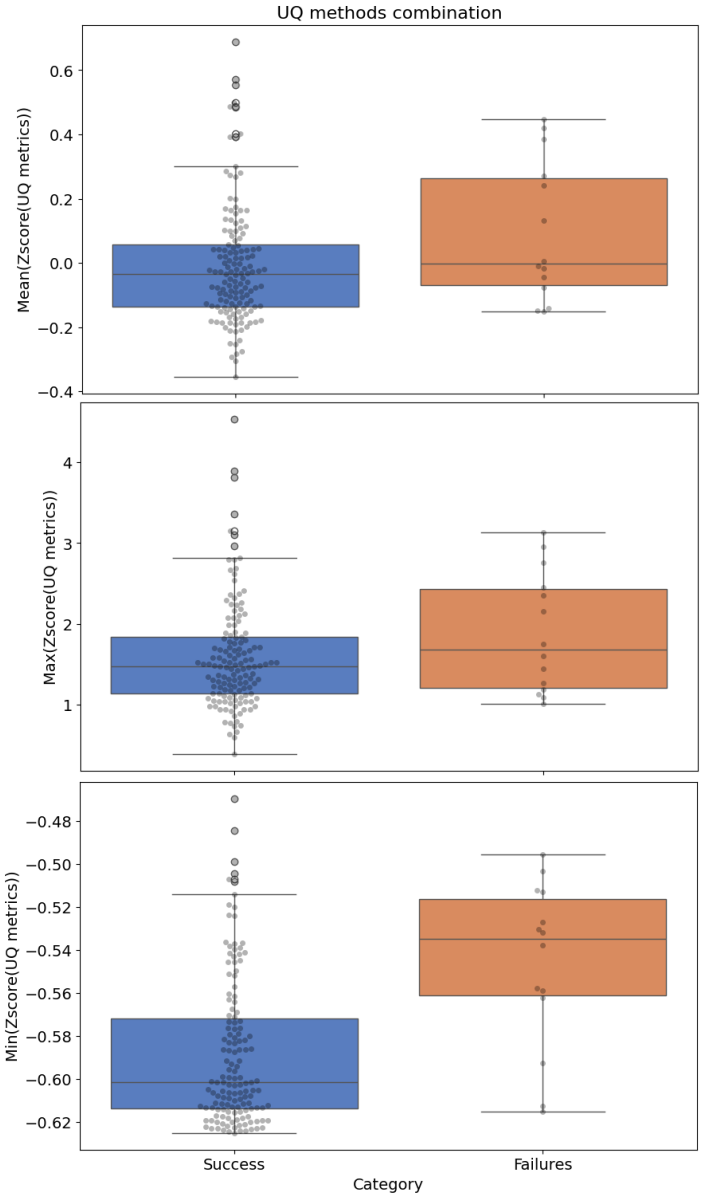


BreastMNIST

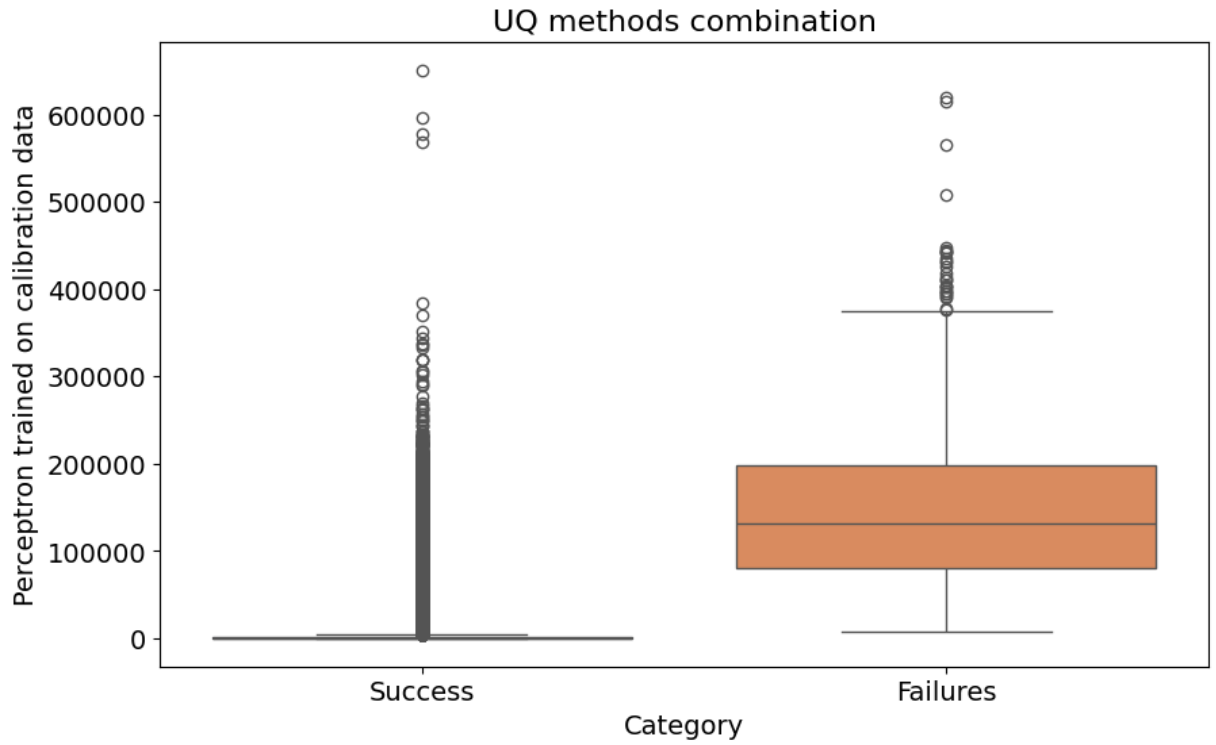
Mean Comb
(AUC=0.63)

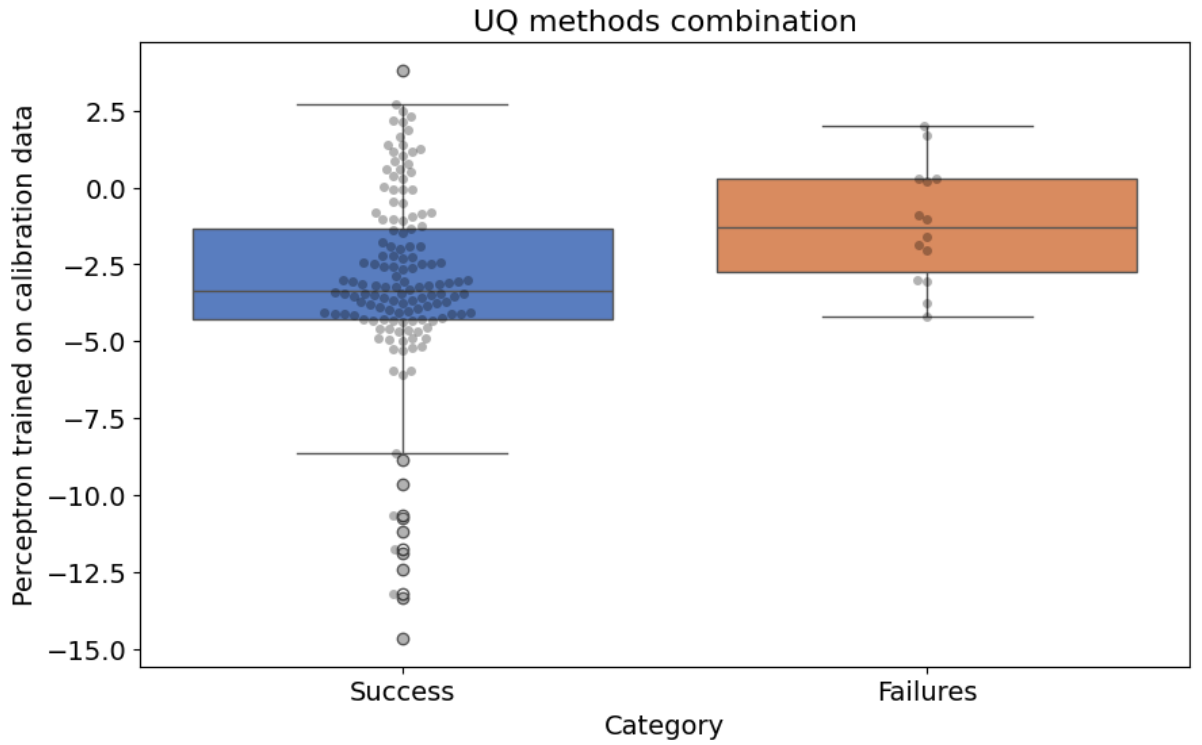
Max Comb
(AUC=0.6)

Min Comb
(AUC=0.79)



Methods combination (Perceptron)





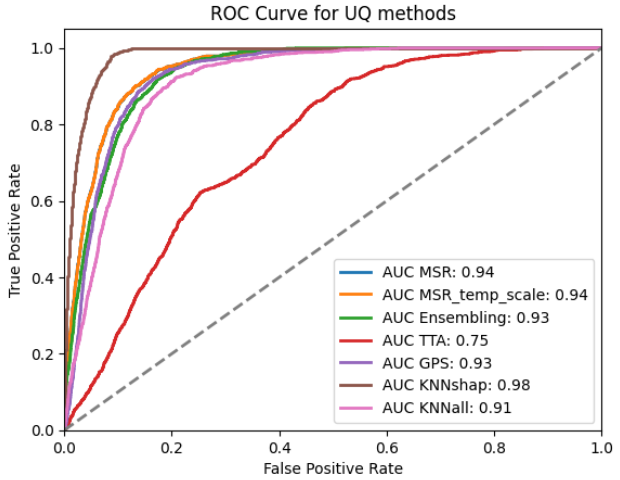
Perceptron Organamnist (AUC=0.97)

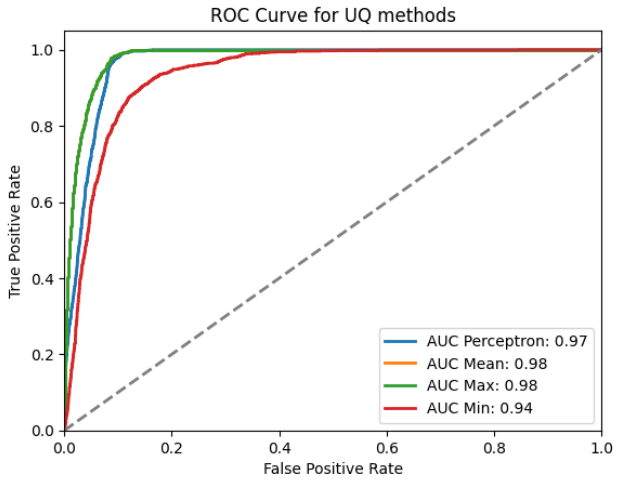
Perceptron Breastmnist (AUC=0.72)

UQ methods comparison and combination

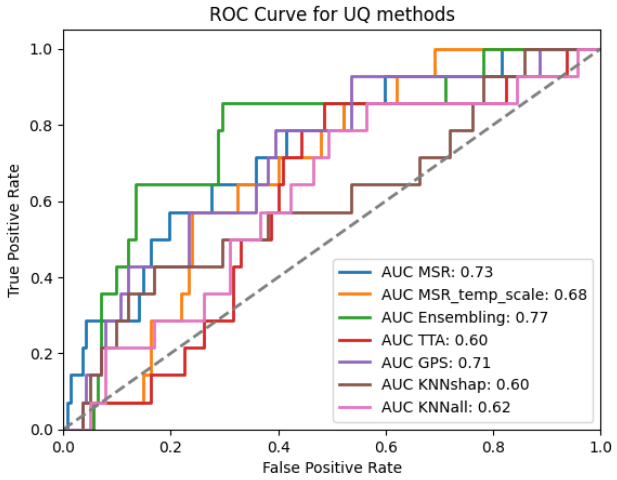
Combination → as good/better than best UQ method. Ensure maximum performances ? → Tbc on other use cases

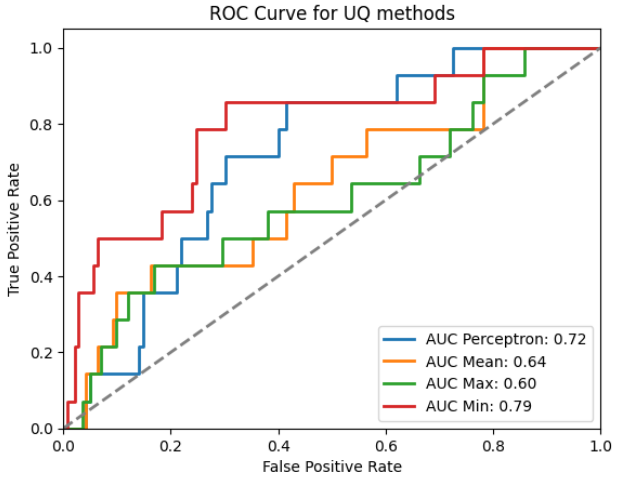
AUC not necessarily best metric to evaluate UQ → if 0% error rate needed, very high AUC can still lead to poor/zero sensitivity





OrganAMNIST





BreastMNIST

UQ method comparison and combination

UQ methods		MSR / DHL	Temp scaling output	Ensembling	TTA	GPS	KNNall	KNNshap	Mean	Max	Min	Perceptron
breastMNIST bAcc	*AUC*	0.73	0.68	0.77	0.64	0.73	0.61	0.6	0.63	0.6	**0.79**	0.72
	bAcc	0.69	0.67	0.78	0.64	0.71	0.64	0.63	0.63	0.63	**0.79**	0.72
	Spec	0.57	**0.86**	**0.86**	0.43	0.64	**0.86**	0.43	0.36	0.43	0.79	**0.86**
	Sens	0.8	0.48	0.7	**0.85**	0.77	0.42	0.83	0.9	0.83	0.79	0.58
breastMNIST spec	*AUC*	0.73	0.68	0.77	0.64	0.73	0.61	0.6	0.63	0.6	**0.79**	0.72
	bAcc	0.59	**0.65**	0.6	0.55	0.64	0.52	0.57	0.61	0.57	0.61	0.64
	Spec	1	1	1	1	1	1	1	1	1	1	1
	Sens	0.18	**0.3**	0.2	0.09	0.27	0.05	0.14	0.21	0.14	0.21	0.27

UQ methods		MSR / DHL	Temp scaling output	Ensembling	TTA	GPS	KNNall	KNNshap	Mean	Max	Min	Perceptron
organAMNIST bAcc	*AUC*	0.94	0.94	0.93	0.76	0.93	0.91	**0.98**	**0.98**	**0.98**	0.94	0.97
	bAcc	0.88	0.88	0.87	0.71	0.87	0.86	**0.95**	**0.95**	**0.95**	0.88	0.94
	Spec	0.9	0.9	0.92	0.91	0.9	0.91	0.98	0.98	0.98	0.91	**0.99**
	Sens	0.87	0.87	0.83	0.51	0.85	0.8	**0.91**	**0.91**	**0.91**	0.85	0.89
organAMNIST spec	*AUC*	0.94	0.94	0.93	0.76	0.93	0.91	**0.98**	**0.98**	**0.98**	0.94	0.97
	bAcc	0.54	0.57	0.5	0.5	0.5	0.69	0.5	0.5	0.5	0.57	**0.92**
	Spec	1	1	1	1	1	1	1	1	1	1	1
	Sens	0.09	0.14	0	0	0	0.38	0	0	0	0.14	**0.83**

New applications and use cases

	Binary classification	Multiclass classification	Segmentation	Regression
MSR / Temp scaling	✓	✓	Needs to be adapted (one distance/voxel) --> heatmap ?	✗
TTA / GPS	✓	✓	Needs to be adapted (one distance/voxel) --> heatmap ?	✓
Ensembling models variability	✓	✓	Needs to be adapted (one distance/voxel) --> heatmap ?	✓
Distance to boundary latent space / KNN	✓	✓	May be difficult because of skip connexions --> use of VAE	Needs to be adapted (no discrete boundaries) --> distance to train distribution ?
Use cases	BI-RADS shape classification MedMNIST	MedMNIST	BI-RADS segmentation MSD decathlon	Use cases to be found