Towards a Rigorous Calibration Assessment Framework: Advancements in Metrics, Methods, and Use

Lorenzo Famiglini[a;*], **Andrea Campagner**[b] **and Federico Cabitza**[a,b]

[a]Universitá degli Studi di Milano-Bicocca, Milan, Italy
[b]IRCCS Istituto Galeazzi Milano, Milan, Italy
ORCiD ID: Lorenzo Famiglini https://orcid.org/0000-0002-1934-5899,
Andrea Campagner https://orcid.org/0000-0002-0027-5157,
Federico Cabitza https://orcid.org/0000-0002-4065-3415

1 Appendix

1.1 Error Comparison Experimental Settings

1.1.1 Descriptive Statistics for Error Metrics and Balanced Accuracy

This section comprehensively summarizes the descriptive statistics for all error metrics and balanced accuracy for training and test sets.

Table 1. Summary statistics simulation binary setting.

Statistic	TE	ECE_{acc}	ECE_{fr}	$1 - \text{ECI}_\text{g}$	Acc_{bal}
Train (mean)	.137	.069	.068	.092	.889
Test (mean)	.160	.075	.074	.099	.805
Train (std)	.030	.060	.062	.090	.086
Test (std)	.039	.044	.045	.065	.081
Train (min)	.071	.004	.003	.001	.486
Test (min)	.090	.007	.005	.006	.480
Train (25%)	.116	.026	.025	.029	.858
Test (25%)	.132	.040	.038	.050	.780
Train (50%)	.131	.044	.038	.048	.893
Test (50%)	.152	.062	.060	.077	.828
Train (75%)	.150	.100	.102	.139	.935
Test (75%)	.179	.103	.104	.138	.858
Train (max)	.310	.270	.269	.438	1.000
Test (max)	.316	.300	.274	.446	.945

The confidence intervals for the true error of the test set, grouped by imbalance ratio and model, reveal that the Logistic Regression model (Logit) consistently outperforms the other models in terms of calibration as shown Table 5.

Across all levels of imbalance, the true error for the Logistic Regression model falls within a lower range compared to the other models. This supports the theoretical expectation that Logistic Regression tends to produce more calibrated probability estimates.

* Corresponding Author. Email: l.famiglini@campus.unimib.it

Table 2. Mean confidence interval at 95% for the difference in means between each measure and the true Error, stratified by different levels of imbalance. Binary task.

Imbalance	ECE_{acc}	ECE_{fp}	$1 - EC_g$
(0.0893, 0.297]	(-0.109, -0.106)	(-0.111, -0.108)	(-0.092, -0.086)
(0.297, 0.428]	(-0.092, -0.089)	(-0.092, -0.089)	(-0.071, -0.067)
(0.428, 0.521]	(-0.087, -0.083)	(-0.088, -0.084)	(-0.065, -0.060)
(0.521, 0.638]	(-0.087, -0.083)	(-0.088, -0.084)	(-0.065, -0.060)

Table 3. Mean confidence interval at 95% for the difference in means between each measure and the true Error, stratified by different levels of imbalance (range of imbalance values for each quartile). Multiclass task.

Imbalance	ECE_{acc}	ECE_{fp}	$1 - EC_g$
(0.036, 0.175]	(-0.068, -0.065)	(-0.069, -0.065)	(-0.058, -0.053)
(0.175, 0.204]	(-0.069, -0.065)	(-0.071, -0.067)	(-0.058, -0.054)
(0.204, 0.226]	(-0.069, -0.066)	(-0.071, -0.068)	(-0.056, -0.052)
(0.226, 0.295]	(-0.071, -0.067)	(-0.072, -0.068)	(-0.059, -0.054)

Table 4. Summary statistics simulation multiclass setting.

Statistic	TE	ECE_{acc}	ECE_{fr}	$1 - \text{ECI}_\text{g}$	Acc_{bal}
Train (mean)	.085	.047	.049	.066	.811
Test (mean)	.097	.041	.042	.057	.674
Train (std)	.034	.043	.051	.073	.137
Test (std)	.039	.030	.035	.050	.121
Train (min)	.046	.004	.006	.008	.370
Test (min)	.046	.010	.008	.010	.260
Train (25%)	.061	.019	.017	.020	.708
Test (25%)	.068	.021	.019	.024	.596
Train (50%)	.071	.027	.025	.029	.836
Test (50%)	.088	.027	.025	.033	.682
Train (75%)	.095	.052	.053	.071	.907
Test (75%)	.105	.049	.050	.067	.772
Train (max)	.189	.161	.200	.284	1.000
Test (max)	.209	.133	.142	.202	.916

Table 5. Confidence intervals (alpha equal to 0.05) for true error in the test set, grouped by imbalance ratio and model. Binary task.

Imbalance Ratio	Logit	XGBoost	SVM	RF	DT	MLP	KNN	NB
(0.0897, 0.2]	(0.136, 0.145)	(0.178, 0.187)	(0.147, 0.155)	(0.165, 0.171)	(0.225, 0.236)	(0.219, 0.232)	(0.190, 0.197)	(0.158, 0.166)
(0.2, 0.309]	(0.125, 0.131)	(0.171, 0.177)	(0.140, 0.144)	(0.173, 0.178)	(0.239, 0.247)	(0.166, 0.183)	(0.191, 0.196)	(0.148, 0.152)
(0.309, 0.419]	(0.123, 0.127)	(0.167, 0.173)	(0.137, 0.141)	(0.178, 0.184)	(0.245, 0.252)	(0.141, 0.150)	(0.189, 0.193)	(0.141, 0.146)
(0.419, 0.528]	(0.121, 0.124)	(0.163, 0.167)	(0.136, 0.139)	(0.182, 0.188)	(0.245, 0.250)	(0.139, 0.147)	(0.189, 0.192)	(0.140, 0.145)
(0.528, 0.638]	(0.121, 0.126)	(0.163, 0.168)	(0.137, 0.141)	(0.180, 0.187)	(0.245, 0.251)	(0.138, 0.146)	(0.189, 0.193)	(0.140, 0.145)

The XGBoost and SVM models also perform relatively well, with their true errors falling within a comparable range to Logistic Regression. However, their confidence intervals are slightly wider, indicating less precision in their estimates.

The Decision Tree (DT), Multilayer Perceptron (MLP), K-Nearest Neighbors (KNN), and Naive Bayes (NB) models show higher levels of true error across all imbalance ratios. This suggests that these models may be less calibrated than Logistic Regression, XGBoost, and SVM, particularly at higher levels of imbalance.

Table 6 extends this analysis to the multiclass setting, offering additional insights for comparison and evaluation:

In the following Figure 1 2,3 4, we present the distribution of balanced accuracy for various machine learning models during the experiments conducted in both binary and multiclass settings. These plots provide insights into the performance of the models on the training and test sets, allowing for a better understanding of the various scenarios.

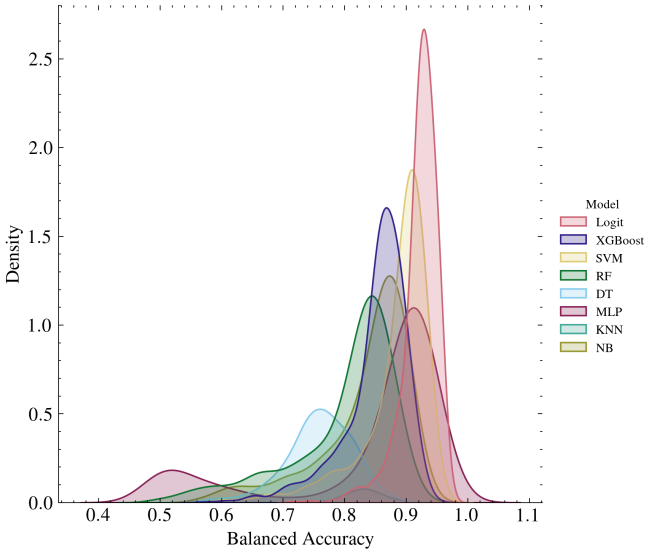

Figure 1. Distribution of Balanced Accuracy for Each Model during the simulation. Binary settings task, Results related to Test Set

1.1.2 Binary Settings, reliability diagrams: True Error $\mathcal{E}$ vs. Estimated Error $\hat{\mathcal{E}}$

Reliability diagrams provide a visual representation of the relationship between the true error, denoted as $\mathcal{E}$, and the estimated error, represented by $\hat{\mathcal{E}}$. We can gain insights into the metrics' behavior by plotting the true error rates against the classifier's estimated error rates. The Figures below are referred to the experimental binary settings simulation.

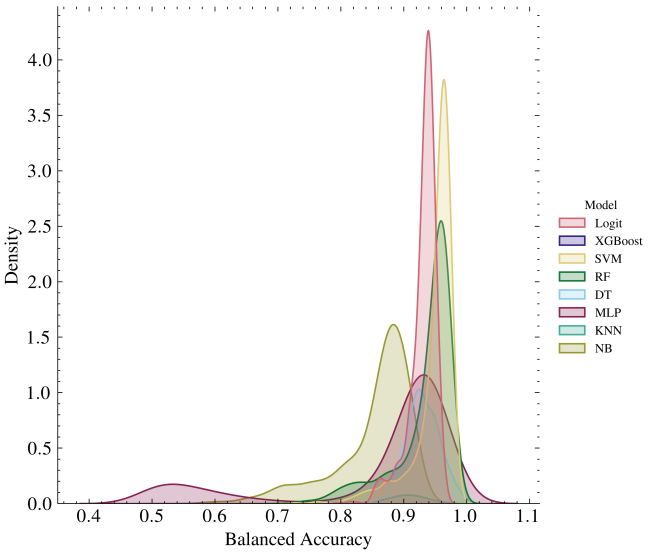

Figure 2. Distribution of Balanced Accuracy for Each Model during the simulation. Binary settings task, Results related to Train Set

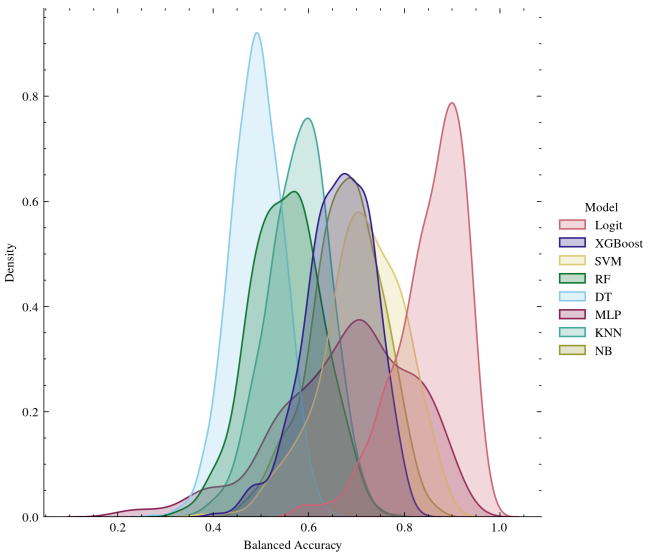

Figure 3. Distribution of Balanced Accuracy for Each Model during the simulation. Multiclass settings task, Results related to Test Set

Table 6. Confidence intervals (alpha equal to 0.05) for true error in the test set, grouped by imbalance ratio and model. Multiclass task.

Imbalance Ratio	Logit	XGBoost	SVM	RF	DT	MLP	KNN	NB
(0.0418, 0.321]	(0.063, 0.065)	(0.099, 0.101)	(0.073, 0.075)	(0.161, 0.164)	(0.166, 0.169)	(0.075, 0.078)	(0.134, 0.136)	(0.098, 0.099)
(0.321, 0.488]	(0.061, 0.062)	(0.1, 0.101)	(0.073, 0.074)	(0.172, 0.174)	(0.177, 0.179)	(0.069, 0.071)	(0.14, 0.142)	(0.1, 0.102)
(0.488, 0.666]	(0.06, 0.062)	(0.099, 0.101)	(0.072, 0.074)	(0.178, 0.18)	(0.181, 0.184)	(0.067, 0.069)	(0.143, 0.145)	(0.101, 0.103)
(0.666, 0.973]	(0.059, 0.06)	(0.098, 0.1)	(0.071, 0.073)	(0.181, 0.184)	(0.184, 0.187)	(0.065, 0.067)	(0.145, 0.146)	(0.101, 0.103)

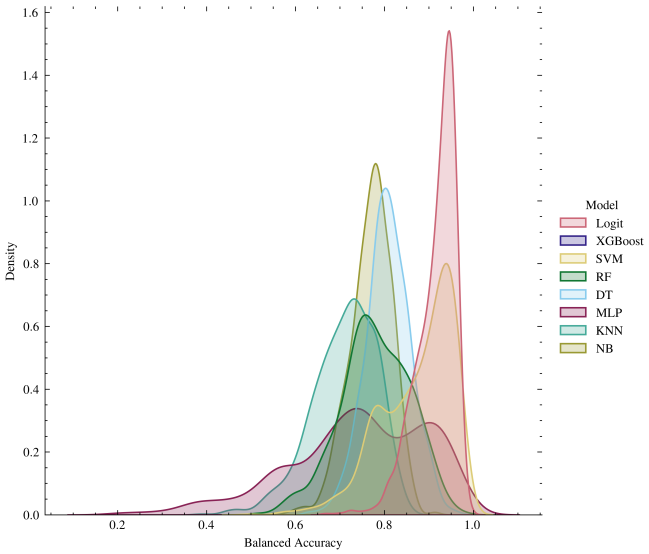

Figure 4. Distribution of Balanced Accuracy for Each Model during the simulation. Multiclass settings task, Results related to Train Set

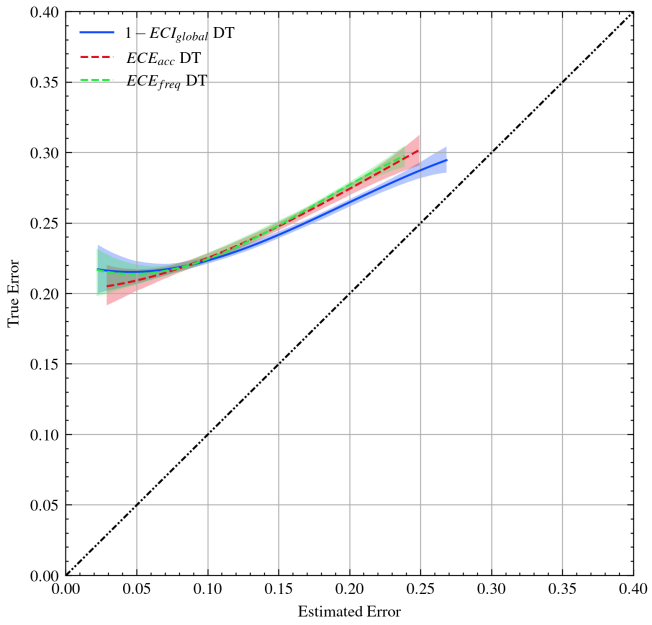

Figure 5. Reliability Diagram: A visual representation of the relationship between the true error (y-axis) and the estimated error (x-axis) for the the binary setting, based on Decision Tree model, with 95% confidence intervals obtained through bootstrap techniques. The Spline Function is employed to smooth the data for enhanced clarity and interpretation

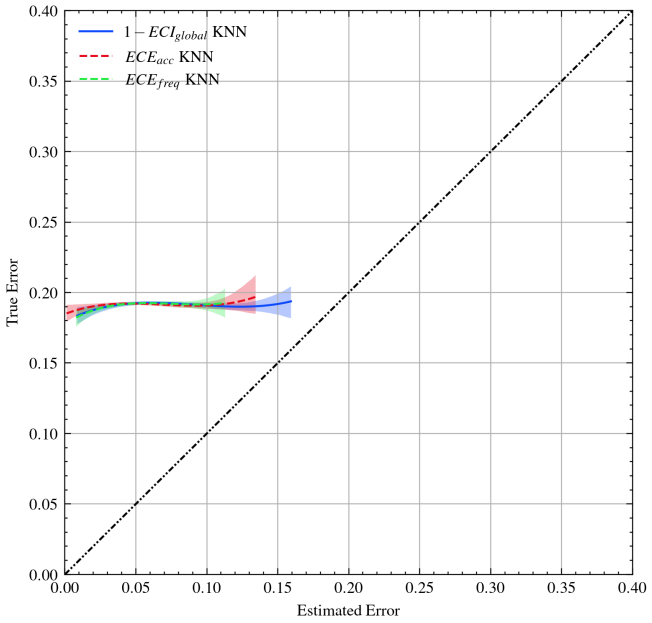

Figure 6. Reliability Diagram: A visual representation of the relationship between the true error (y-axis) and the estimated error (x-axis) for the the binary setting, based on KNN model, with 95% confidence intervals obtained through bootstrap techniques. The Spline Function is employed to smooth the data for enhanced clarity and interpretation

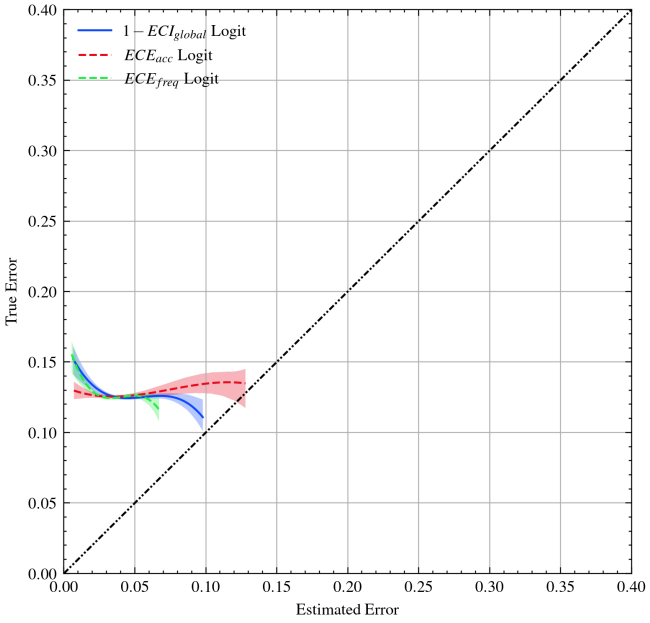

Figure 7. Reliability Diagram: A visual representation of the relationship between the true error (y-axis) and the estimated error (x-axis) for the the binary setting, based on Logistic Regression, with 95% confidence intervals obtained through bootstrap techniques. The Spline Function is employed to smooth the data for enhanced clarity and interpretation

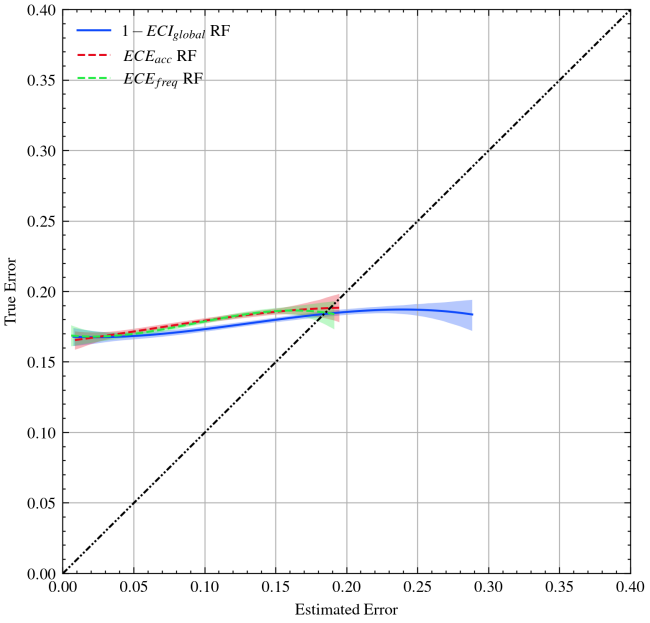

Figure 9. Reliability Diagram: A visual representation of the relationship between the true error (y-axis) and the estimated error (x-axis) for the the binary setting, based on Random Forest, with 95% confidence intervals obtained through bootstrap techniques. The Spline Function is employed to smooth the data for enhanced clarity and interpretation

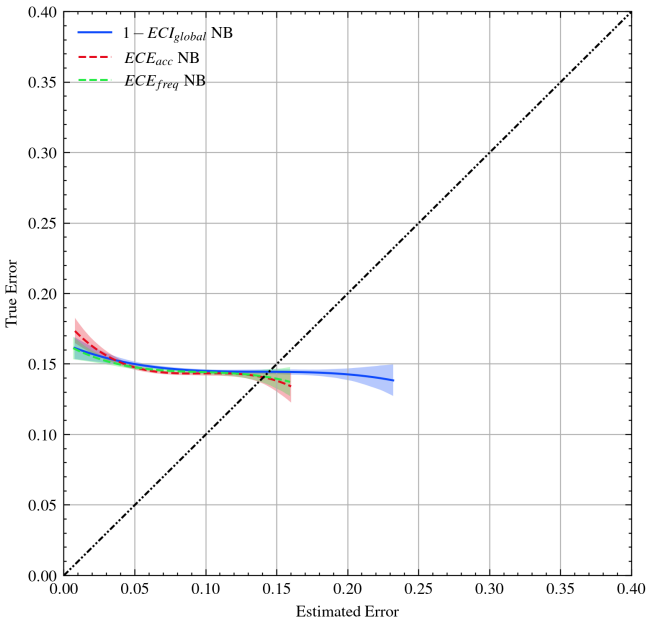

Figure 8. Reliability Diagram: A visual representation of the relationship between the true error (y-axis) and the estimated error (x-axis) for the the binary setting, based on Naive Bayes, with 95% confidence intervals obtained through bootstrap techniques. The Spline Function is employed to smooth the data for enhanced clarity and interpretation

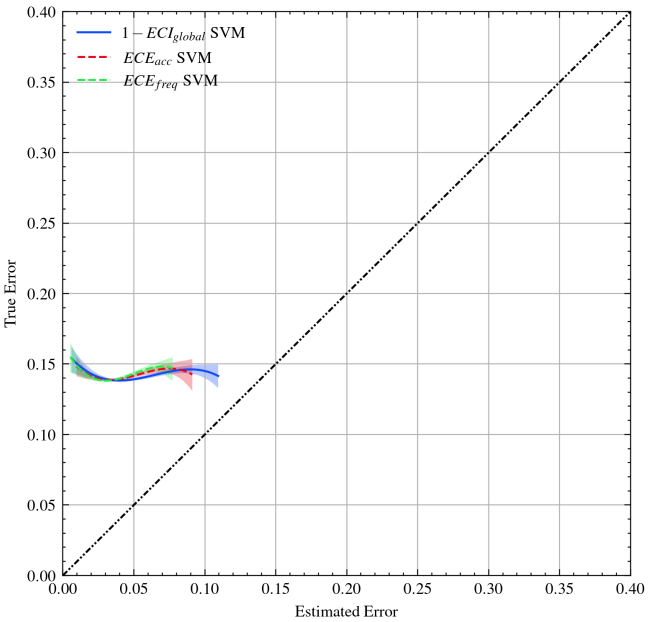

Figure 10. Reliability Diagram: A visual representation of the relationship between the true error (y-axis) and the estimated error (x-axis) for the the binary setting, based on Support Vector Machine, with 95% confidence intervals obtained through bootstrap techniques. The Spline Function is employed to smooth the data for enhanced clarity and interpretation

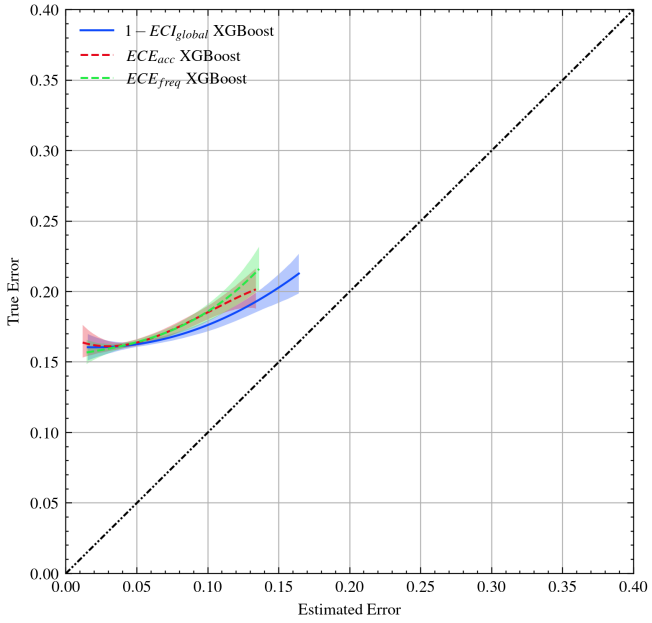

Figure 11. Reliability Diagram: A visual representation of the relationship between the true error (y-axis) and the estimated error (x-axis) for the the binary setting, based on X Gradient Boosting, with 95% confidence intervals obtained through bootstrap techniques. The Spline Function is employed to smooth the data for enhanced clarity and interpretation

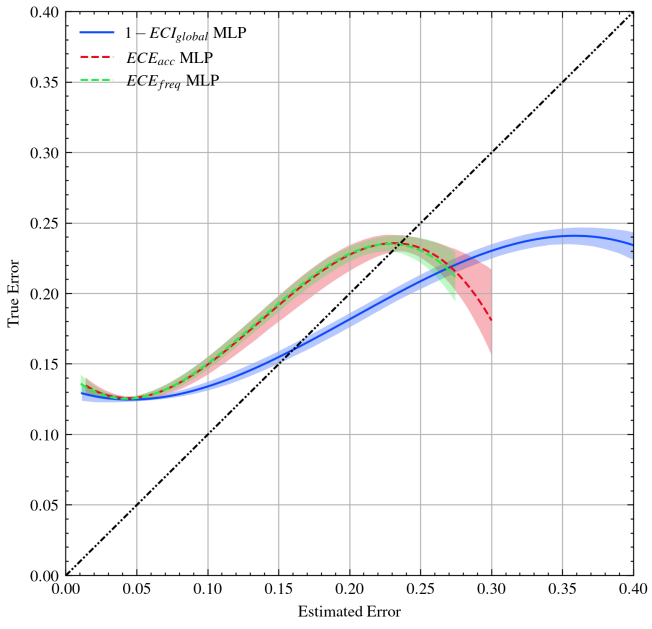

Figure 12. Reliability Diagram: A visual representation of the relationship between the true error (y-axis) and the estimated error (x-axis) for the the binary setting, based on MLP, with 95% confidence intervals obtained through bootstrap techniques. The Spline Function is employed to smooth the data for enhanced clarity and interpretation

1.1.3 Multiclass Settings, reliability diagrams: True Error $\mathcal{E}$ vs. Estimated Error $\hat{\mathcal{E}}$

As previously discussed, we conducted an analysis to examine the relationship between ($\mathcal{E}$) and ($\hat{\mathcal{E}}$) in various experimental scenarios. The subsequent Figure pertain to the multiclass simulations, offering a comprehensive visual representation of the classifier's performance in these contexts.

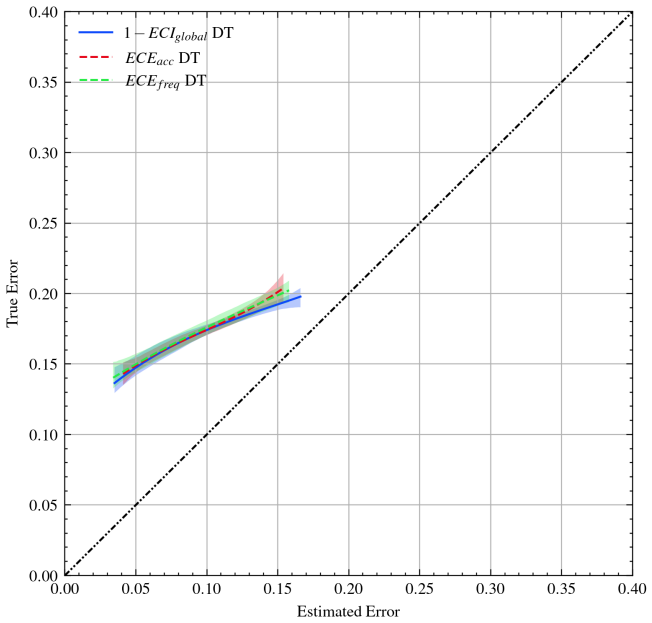

Figure 13. Reliability Diagram: A visual representation of the relationship between the true error (y-axis) and the estimated error (x-axis) for the the multiclass setting, based on Decision Tree model, with 95% confidence intervals obtained through bootstrap techniques. The Spline Function is employed to smooth the data for enhanced clarity and interpretation

1.2 PathMNIST, PneumoniaMNIST

We present the visual representation of our research through the inclusion of sample images from two distinct datasets: Pathmnist and Pneumoniamnist.

In the first set of examples, we illustrate the usage of Pathmnist, a dataset that is widely acclaimed for its pertinence in the domain of Pathology, more specifically for the analysis of histopathological tissue samples. The succeeding 30 figures, each representative of unique instances from the training set, provide an insight into the heterogeneity and complexity of the Pathmnist data. These images underscore the variations encountered in the context of pathological studies, acting as a visual testament to the multitude of challenges that our model has been designed to adeptly handle. In a similar manner, we subsequently delve into the specifics of the Pneumoniamnist training dataset, a critical resource for the examination and identification of pneumonia from medical imagery.

1.3 Optimizing Classification Thresholds through Calibration Information

This experiment aimed to validate the statement presented in the study. To achieve this, the validation set was divided into two subsets. The first subset was used for model weight selection, while the

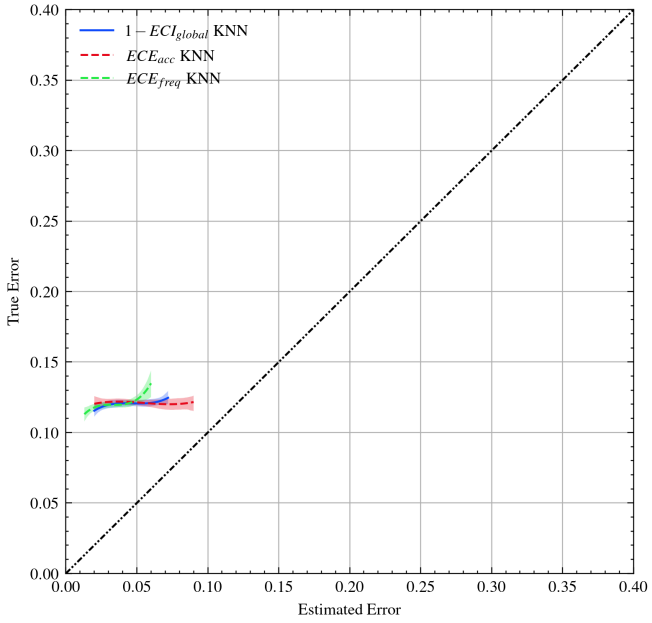

Figure 14. Reliability Diagram: A visual representation of the relationship between the true error (y-axis) and the estimated error (x-axis) for the the multiclass setting, based on KNN model, with 95% confidence intervals obtained through bootstrap techniques. The Spline Function is employed to smooth the data for enhanced clarity and interpretation

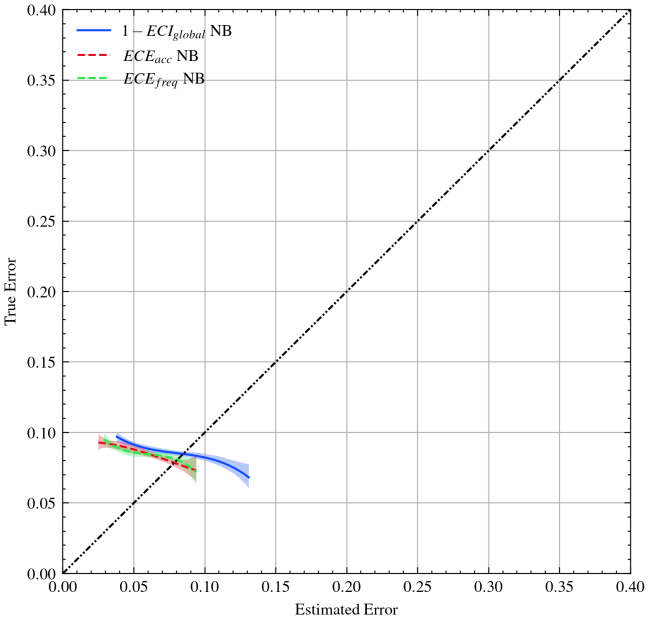

Figure 16. Reliability Diagram: A visual representation of the relationship between the true error (y-axis) and the estimated error (x-axis) for the the multiclass setting, based on Naive Bayes, with 95% confidence intervals obtained through bootstrap techniques. The Spline Function is employed to smooth the data for enhanced clarity and interpretation

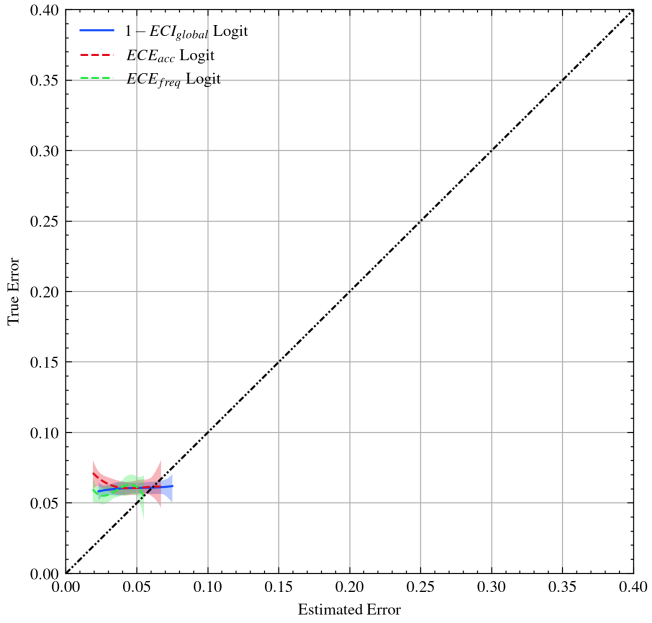

Figure 15. Reliability Diagram: A visual representation of the relationship between the true error (y-axis) and the estimated error (x-axis) for the the multiclass setting, based on Logistic Regression, with 95% confidence intervals obtained through bootstrap techniques. The Spline Function is employed to smooth the data for enhanced clarity and interpretation

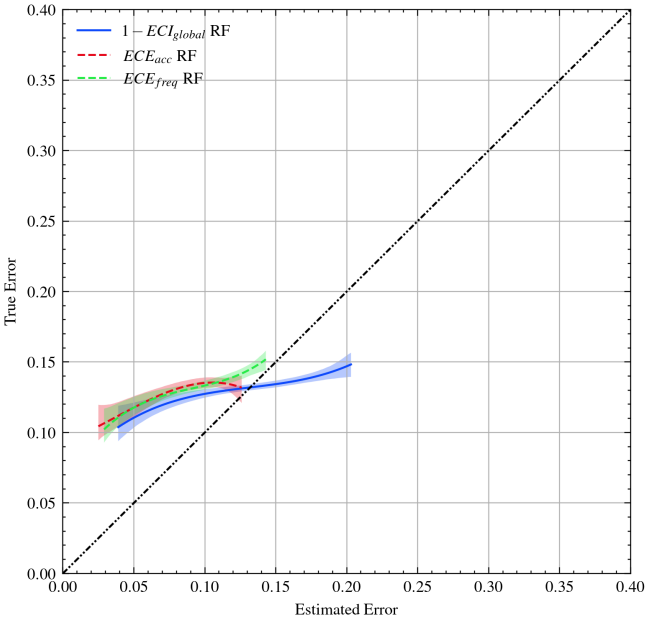

Figure 17. Reliability Diagram: A visual representation of the relationship between the true error (y-axis) and the estimated error (x-axis) for the the multiclass setting, based on Random Forest, with 95% confidence intervals obtained through bootstrap techniques. The Spline Function is employed to smooth the data for enhanced clarity and interpretation

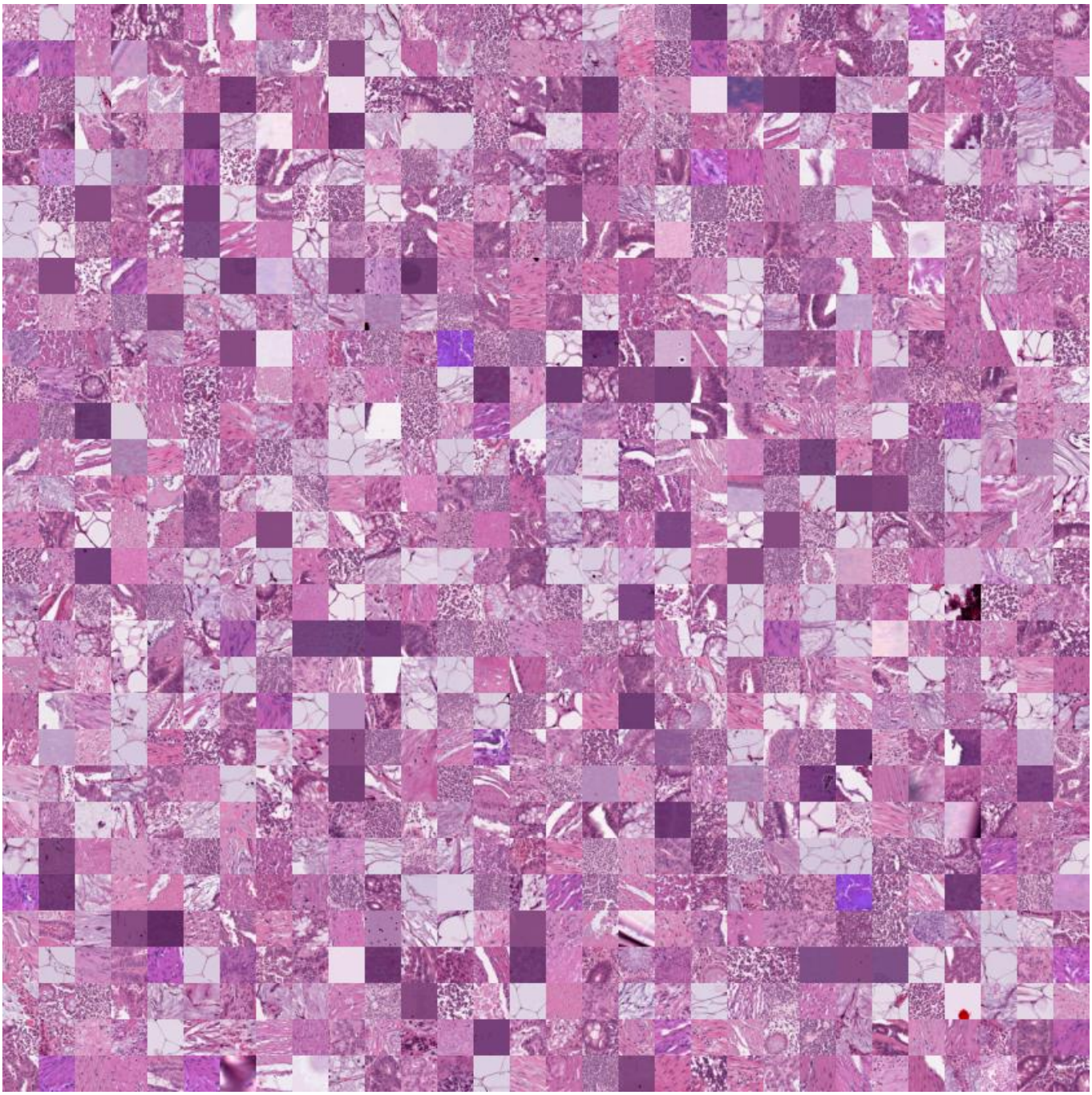

Figure 18. Representative images from the Pathmnist training set.

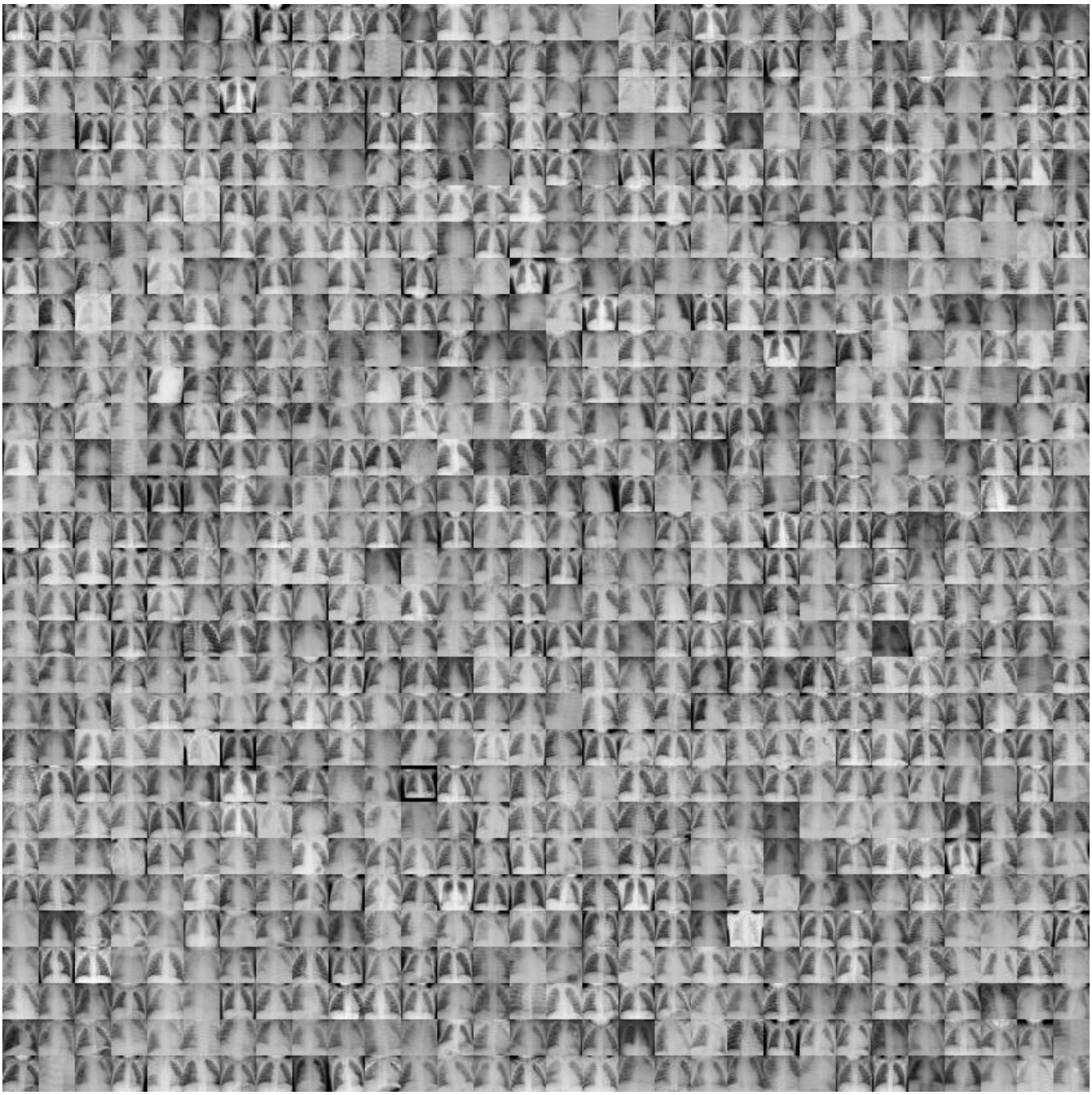

Figure 19. Representative images from the Pneumoniamnist training set.

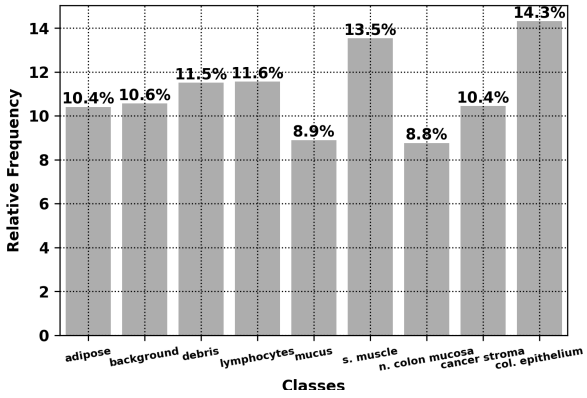

Figure 20. Class distribution PathMNIST dataset (training set)

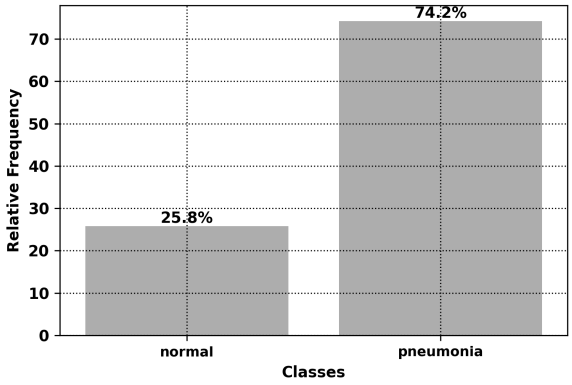

Figure 21. Class distribution PneumoniaMNIST dataset (training set)

second was used as a calibration set for determining the probability threshold based on the Expected Calibration (ECI_l) information.

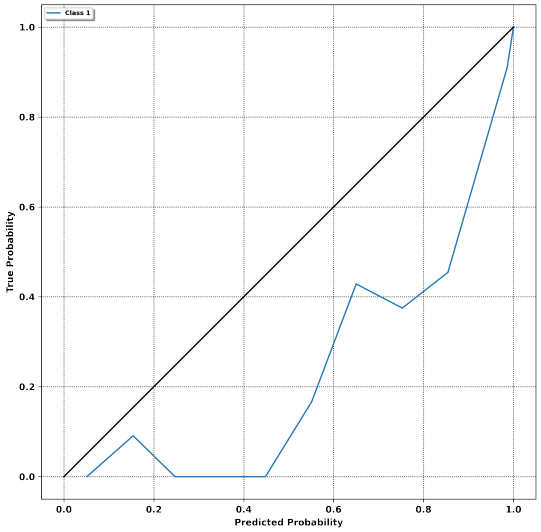

Figure 22. Reliability Diagram Calibration Set DeiT model, PneumoniaMNIST

The first experiment was conducted on the pneumoniaMnist dataset. A specific threshold was selected based on the local information obtained from the confidence scores bin between [0.4, 0.6]. The calibration set revealed that the average ECI_l for the DeiT model in this bin was 0.35, indicating an overconfident region. To improve the model's performance, the threshold was set to 0.65, where the model showed less miscalibration and reduced overconfidence (as depicted in Figure 22). Subsequently, various performance metrics were calculated to evaluate the impact of threshold selection on model performance. This threshold can also be selected based on a validation approach to finding the optimal value.

Table 7. Test set results, DeiT model based on selected thresholds

Threshold	Acc	TPR	PPV	F1
.50	.860	**.896**	.818	.837
.65	**.878**	.891	**.849**	**.863**

1.4 Additional Experimental Results

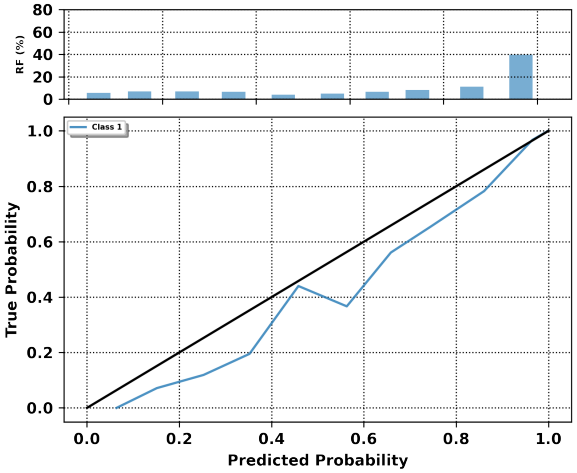

Figure 23. Reliability Diagram, ResNet152 on PneumoniaMNIST test set

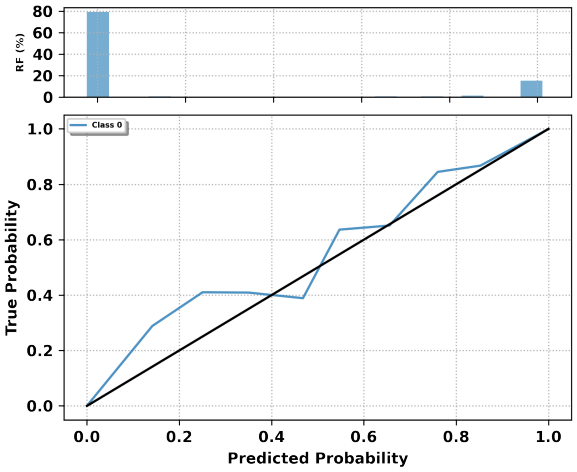

Figure 24. Reliability Diagram, ResNet152 on PathMINST test set, class 0

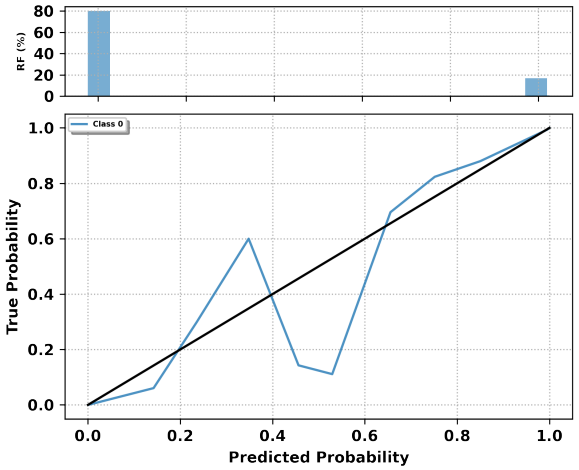

Figure 25. Reliability Diagram, DeiT on PathMINST test set, class 0

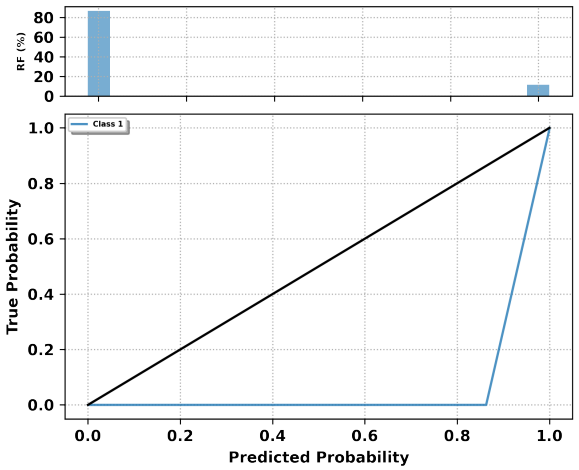

Figure 26. Reliability Diagram, ResNet152 on PathMINST test set, class 1

Table 8. ECI_l ResNet152 PneumoniaMNIST, test set results

$\mathbf{ECI}_l$	Frequency (%)	Bins	Class
.932	5.45	.008-.107	1
.906	6.73	.107-.206	1
.821	6.73	.206-.305	1
.757	6.57	.305-.405	1
.966	4.01	.405-.504	1
.651	4.81	.504-.603	1
.852	6.57	.603-.702	1
.878	8.33	.702-.801	1
.909	11.06	.801-.9	1
.998	39.58	.9-.999	1

Table 9. Multiclass task, Resnet152. The reported ECEs are related to 1-ECE to make an easier comparison with our ECI

Class	$\mathbf{ECE}_{\mathbf{acc}}$	$\mathbf{ECE}_{\mathbf{freq}}$	$\mathbf{ECI}_{\mathbf{global}}$
0	.995	.995	.993
1	.993	.993	.990
2	.988	.987	.984
3	.987	.986	.985
4	.985	.985	.981
5	.981	.982	.978
6	.988	.987	.984
7	.979	.978	.975
8	.993	.995	.993

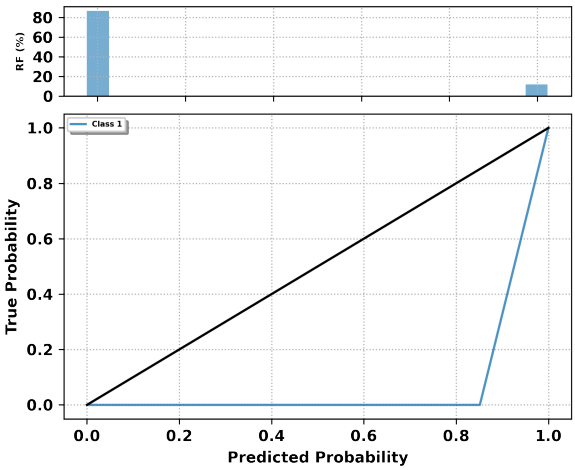

Figure 27. Reliability Diagram, DeiT on PathMINST test set, class 1

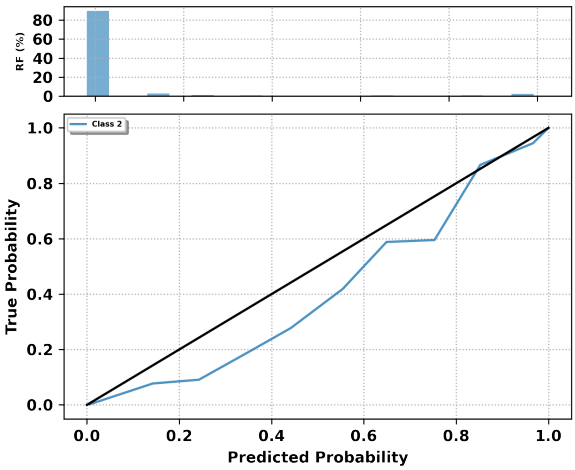

Figure 28. Reliability Diagram, ResNet152 on PathMINST test set, class 2

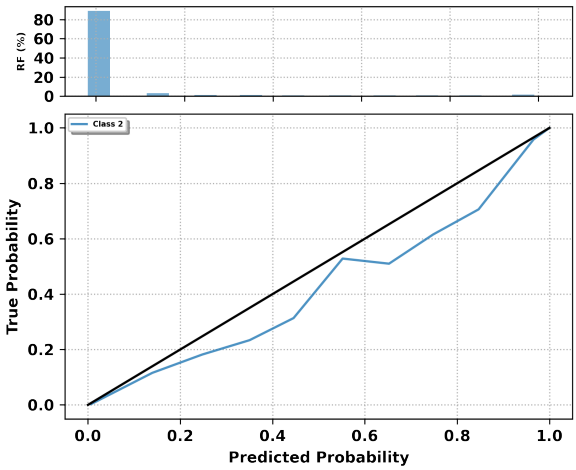

Figure 29. Reliability Diagram, DeiT on PathMINST test set, class 2

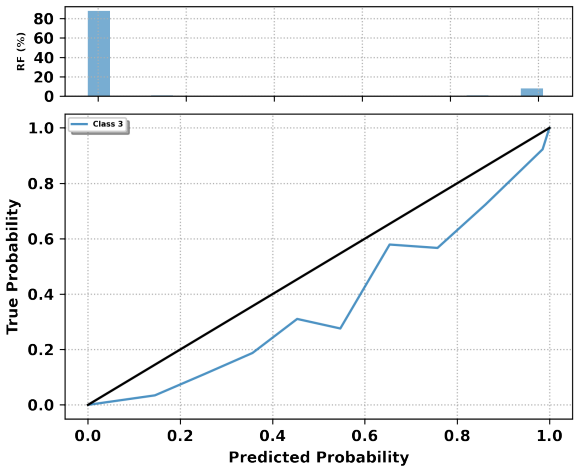

Figure 30. Reliability Diagram, ResNet152 on PathMINST test set, class 3

Table 10. Multiclass task, ResNet152. Model Comparison based on ECI

Class	$\mathbf{ECI}_{balance}$	$\mathbf{ECI}_{over}$	$\mathbf{ECI}_{under}$
0	.0403	.922	.904
1	.01	.416	1
2	-.001	.845	.992
3	-.015	.807	NA
4	-.018	NA	.766
5	-.062	.746	.860
6	-.058	.947	.918
7	.102	.862	.973
8	-.034	.974	.966

Table 11. Multiclass task, DeiT. The reported ECEs are related to 1-ECE to make an easier comparison with our ECI

Class	$\mathbf{ECE}_{\mathbf{acc}}$	$\mathbf{ECE}_{\mathbf{freq}}$	$\mathbf{ECI}_{\mathbf{global}}$
0	.996	.996	.995
1	.994	.993	.991
2	.992	.993	.990
3	.987	.986	.985
4	.992	.992	.991
5	.988	.989	.987
6	.991	.990	.987
7	.983	.982	.976
8	.991	.990	.987

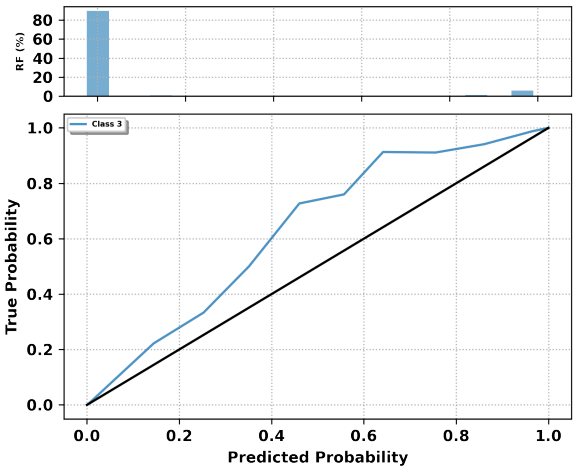

Figure 31. Reliability Diagram, DeiT on PathMINST test set, class 3

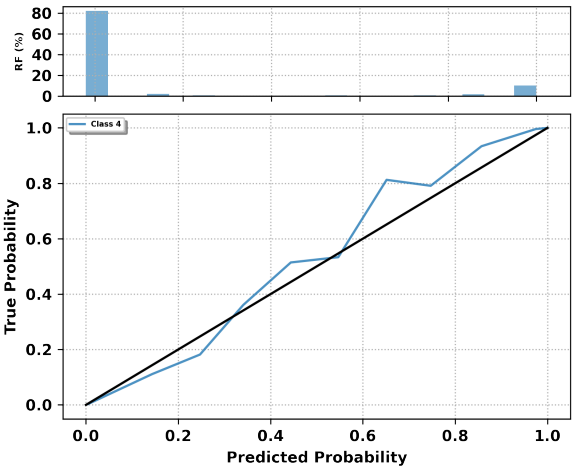

Figure 33. Reliability Diagram, DeiT on PathMINST test set, class 4

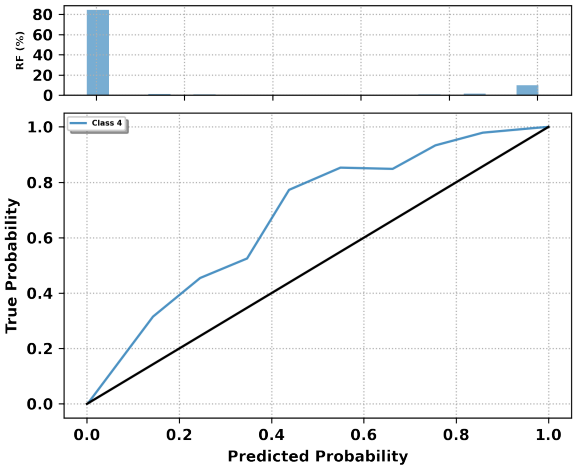

Figure 32. Reliability Diagram, ResNet152 on PathMINST test set, class 4

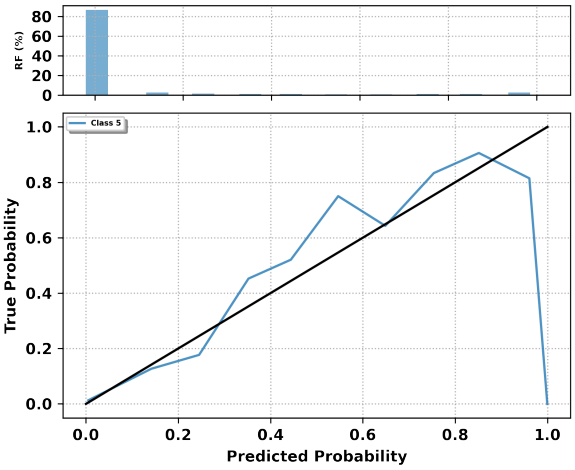

Figure 34. Reliability Diagram, ResNet152 on PathMINST test set, class 5

Table 12. ECI_l ResNet152 PathMNIST, Class 0 results

ECI_l	Frequency (%)	Bins	Class
.998	79.526	.0-.1	0
.829	.724	.1-.2	0
.787	.543	.2-.3	0
.910	.306	.3-.4	0
.851	.251	.4-.5	0
.837	.460	.5-.6	0
.993	.599	.6-.7	0
.889	.808	.7-.8	0
.983	1.365	.8-.9	0
.999	15.404	.9-1	0

Table 13. ECI_l DeiT PathMNIST, test set results, Class 0

ECI_l	Frequency (%)	Bins	Class
.999	80.223	.0-.1	0
.905	.460	.1-.2	0
.916	.279	.2-.3	0
.614	.209	.3-.4	0
.423	.097	.4-.5	0
.210	.125	.5-.6	0
.939	.320	.6-.7	0
.904	.237	.7-.8	0
.965	.348	.8-.9	0
.999	16.783	.9-1	0

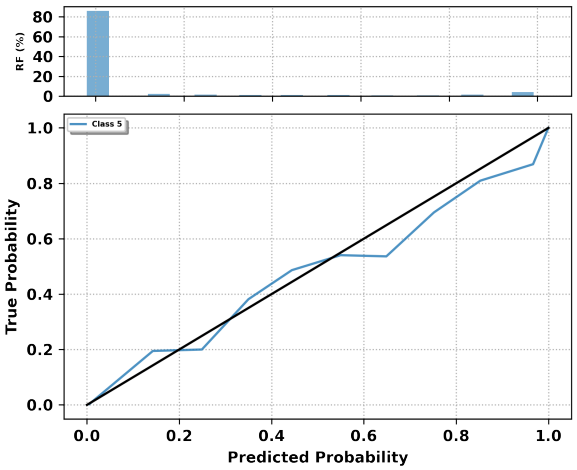

Figure 35. Reliability Diagram, DeiT on PathMINST test set, class 5

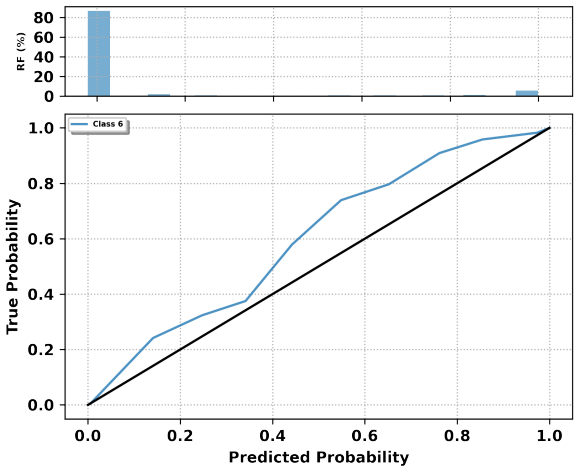

Figure 37. Reliability Diagram, DeiT on PathMINST test set, class 6

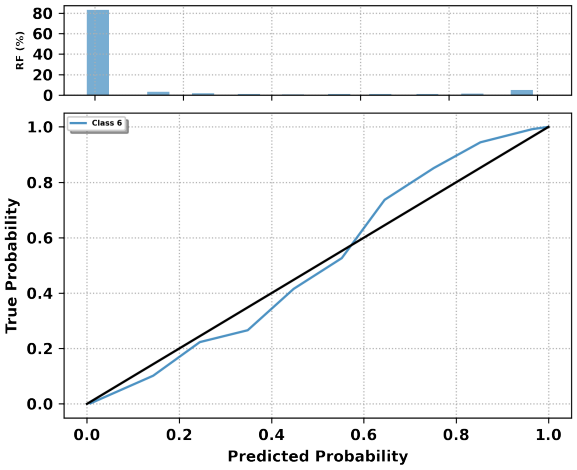

Figure 36. Reliability Diagram, ResNet152 on PathMINST test set, class 6

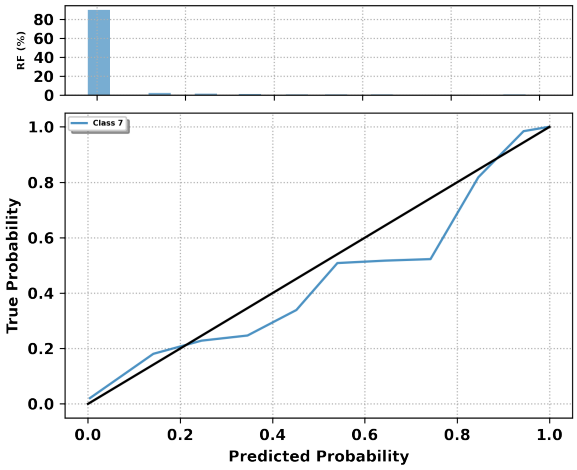

Figure 38. Reliability Diagram, ResNet152 on PathMINST test set, class 7

Table 14. ECI_l ResNet152 PathMNIST, Class 1 results

ECI_l	Frequency (%)	Bins	Class
.999	86.699	.0-.1	1
.827	.446	.1-.2	1
.683	.195	.2-.3	1
.457	.223	.3-.4	1
.204	.153	.4-.5	1
.000	.097	.5-.6	1
.000	.084	.6-.7	1
.000	.070	.7-.8	1
.000	.125	.8-.9	1
.991	11.657	.9-1	1

Table 15. ECI_l ResNet152 PathMNIST, test set results for Class 2

ECI_l	Frequency (%)	Bins	Class
.996	89.763	.0-0.1	2
.924	2.702	.1-.2	2
.800	1.226	.2-.3	2
.754	.794	.3-.4	2
.706	.501	.4-.5	2
.755	.599	.5-.6	2
.906	.710	.6-.7	2
.790	.585	.7-.8	2
.983	.836	.8-.9	2
.977	2.270	.9-1	2

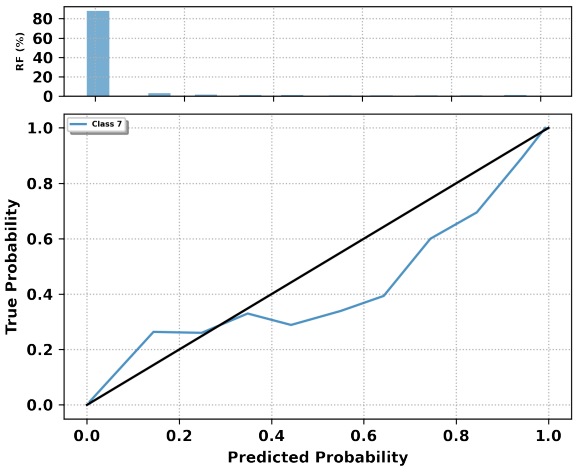

Figure 39. Reliability Diagram, DeiT on PathMINST test set, class 7

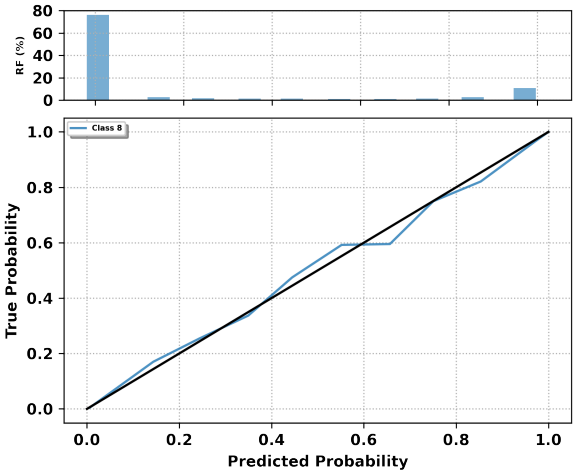

Figure 40. Reliability Diagram, ResNet152 on PathMINST test set, class 8

Table 16. ECI_l ResNet152 PathMNIST, test set results for Class 3

ECI_l	Frequency (%)	Bins	Class
.999	87.967	0-.1	3
.871	.808	.1-.2	3
.811	.501	.2-.3	3
.737	.446	.3-.4	3
.738	.404	.4-.5	3
.504	.404	.5-.6	3
.886	.265	.6-.7	3
.748	.418	.7-.8	3
.840	.655	.8-.9	3
.937	8.120	.9-1	3

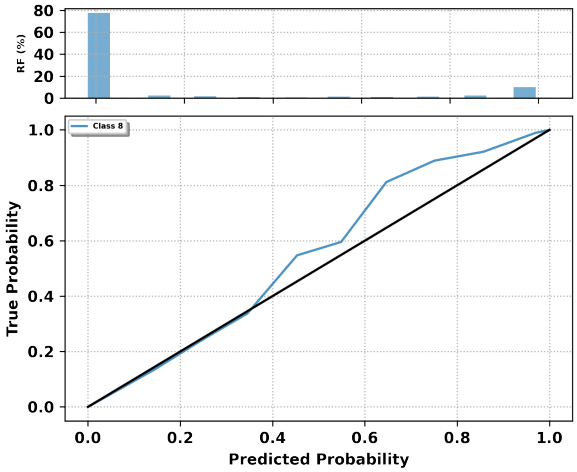

Figure 41. Reliability Diagram, DeiT on PathMINST test set, class 8

Table 17. ECI_l ResNet152 PathMNIST, test set results for Class 4

ECI_l	Frequency (%)	Bins	Class
.998	84.457	0-.1	4
.799	1.017	.1-.2	4
.723	.613	.2-.3	4
.727	.557	.3-.4	4
.404	.306	.4-.5	4
.447	.474	.5-.6	4
.719	.460	.6-.7	4
.764	.836	.7-.8	4
.859	1.337	.8-.9	4
.980	9.930	.9-1	4

Table 18. ECI_l ResNet152 PathMNIST, test set results for Class 5

ECI_l	Frequency (%)	Bins	Class
.992	86.699	0-.1	5
.983	2.744	.1-.2	5
.910	1.574	.2-.3	5
.845	1.170	.3-.4	5
.862	1.017	.4-.5	5
.628	.836	.5-.6	5
.991	.975	.6-.7	5
.895	1.086	.7-.8	5
.936	1.184	.8-.9	5
.848	2.702	.9-1	5

Table 19. ECI_l ResNet152 PathMNIST, test set results for Class 6

$\mathbf{ECI}_l$	Frequency (%)	Bins	Class
.994	83.162	.0-.1	6
.951	3.148	.1-.2	6
.972	1.685	.2-.3	6
.874	1.309	.3-.4	6
.941	0.905	.4-.5	6
.953	1.058	.5-.6	6
.858	1.058	.6-.7	6
.867	1.031	.7-.8	6
.892	1.504	.8-.9	6
.972	5.125	.9-1	6

Table 20. ECI_l ResNet152 PathMNIST, test set results for Class 7

$\mathbf{ECI}_l$	Frequency (%)	Bins	Class
.983	90.042	.0-.1	7
.954	2.465	.1-.2	7
.976	1.769	.2-.299	7
.849	1.128	.299-.399	7
.795	.822	.399-.499	7
.941	.822	.499-.599	7
.803	.808	.599-.699	7
.704	.613	.699-.798	7
.967	.613	.798-.898	7
.957	.905	.898-.998	7

Table 21. ECI_l ResNet152 PathMNIST, test set results, Class 8

$\mathbf{ECI}_l$	Frequency (%)	Bins	Class
.999	76.128	0-.1	8
.969	2.618	.1-.2	8
.986	1.643	.2-.3	8
.981	1.323	.3-.4	8
.946	1.379	.4-.5	8
.927	1.058	.5-.6	8
.907	1.170	.6-.7	8
.999	1.393	.7-.8	8
.962	2.646	.8-.9	8
.995	10.627	.9-1	8

Table 22. ECI_l DeiT PathMNIST, test set results, Class 1

$\mathbf{ECI}_l$	Frequency (%)	Bins	Class
.999	86.880	.0-.1	1
.834	.404	.1-.2	1
.684	.167	.2-.3	1
.452	.111	.3-.4	1
.209	.139	.4-.5	1
.000	.084	.5-.6	1
.000	.084	.6-.7	1
.000	.125	.7-.8	1
.000	.070	.8-.9	1
.990	11.908	0.9-1	1

Table 23. ECI_l DeiT PathMNIST, test set results, Class 2

$\mathbf{ECI}_l$	Frequency (%)	Bins	Class
.996	89.039	.0-.1	2
.972	3.008	.1-.2	2
.913	1.379	.2-.3	2
.820	1.072	.3-.4	2
.762	.933	.4-.5	2
.957	.738	.5-.6	2
.781	.710	.6-.7	2
.823	.724	.7-.8	2
.834	.710	.8-.9	2
.992	1.671	.9-1	2

Table 24. ECI_l DeiT PathMNIST, test set results, Class 3

$\mathbf{ECI}_l$	Frequency (%)	Bins	Class
1.000	89.777	.0-.1	3
.909	.877	.1-.2	3
.892	.418	.2-.3	3
.770	.306	.3-.4	3
.505	.306	.4-.5	3
.636	.348	.5-.6	3
.577	.320	.6-.7	3
.794	.627	.7-.8	3
.907	1.184	.8-.9	3
.977	5.822	.9-1.0	3

Table 25. ECI_l DeiT PathMNIST, test set results, Class 4

$\mathbf{ECI}_l$	Frequency (%)	Bins	Class
.998	82.061	.0-.1	4
.962	2.006	.1-.2	4
.913	.919	.2-.3	4
.970	.501	.3-.4	4
.874	.487	.4-.5	4
.975	.627	.5-.6	4
.753	.446	.6-.7	4
.941	.933	.7-.8	4
.911	1.685	.8-.9	4
.978	10.320	.9-1	4

Table 26. ECI_l DeiT PathMNIST, test set results, Class 5

$\mathbf{ECI}_l$	Frequency (%)	Bins	Class
.998	85.975	.0-.1	5
.939	2.145	.1-.2	5
.935	1.462	.2-.3	5
.951	1.058	.3-.4	5
.924	1.031	.4-.5	5
.983	1.031	.5-.6	5
.827	.961	.6-.7	5
.924	.822	.7-.8	5
.950	1.462	.8-.9	5
.899	4.039	.9-1	5

Table 27. ECI_l DeiT PathMNIST, test set results, Class 6

$\mathbf{ECI}_l$	**Frequency (%)**	**Bins**	**Class**
.999	86.727	.0-.1	6
.883	2.075	.1-.2	6
.899	.947	.2-.3	6
.949	.557	.3-.4	6
.755	.529	.4-.5	6
.652	.641	.5-.6	6
.778	.752	.6-.7	6
.806	.766	.7-.8	6
.879	1.337	.8-.9	6
.991	5.655	.9-1	6

Table 28. ECI_l DeiT PathMNIST, test set results, Class 7

$\mathbf{ECI}_l$	**Frequency (%)**	**Bins**	**Class**
.994	88.106	0.0-.099	7
.860	3.064	.099-.199	7
.985	1.713	.199-.298	7
.972	1.351	.298-.397	7
.725	1.351	.397-.496	7
.617	.780	.496-.596	7
.612	.850	.596-.695	7
.806	.766	.695-.794	7
.823	.822	.794-.894	7
.948	1.184	.894-.993	7

Table 29. ECI_l DeiT PathMNIST, test set results, All Class

$\mathbf{ECI}_l$	**Frequency (%)**	**Bins**	**Class**
.999	77.674	.0-.1	8
.989	2.493	.1-.2	8
.993	1.560	.2-.3	8
.987	1.198	.3-.4	8
.828	.738	.4-.5	8
.914	1.240	.5-.6	8
.745	1.184	.6-.7	8
.816	1.379	.7-.8	8
.926	2.479	.8-.9	8
.980	10.042	.9-1	8