Agree to Disagree: Supplementary Material

Matthew Watson, Bashar Awwad Shiekh Hasan, Noura Al Moubayed
Durham University
Durham, UK
{matthew.s.watson,bashar.awwad-shiekh-hasan,noura.al-moubayed}@durham.ac.uk

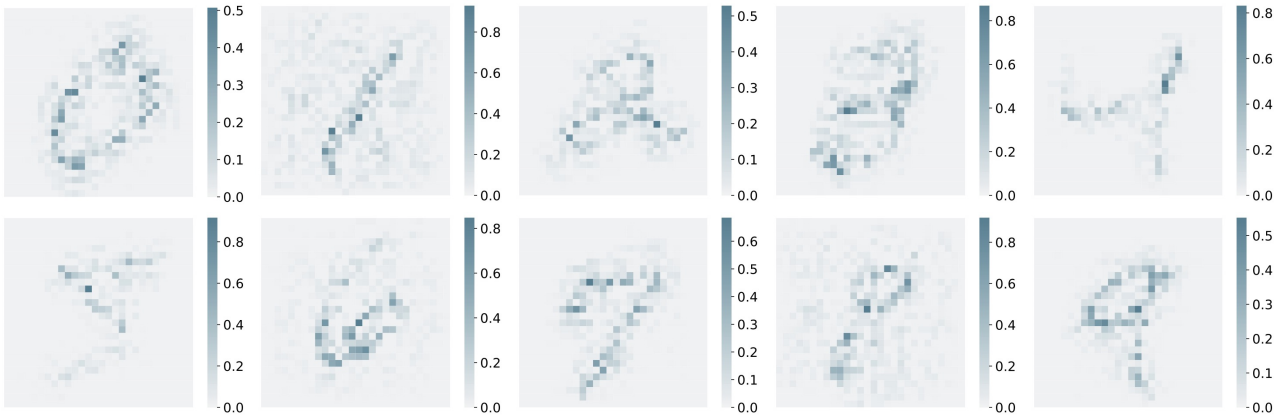

Figure 1: The difference between normalised SHAP values from two CNNs (each trained with different random seeds) for a randomly chosen sample from each MNIST class.

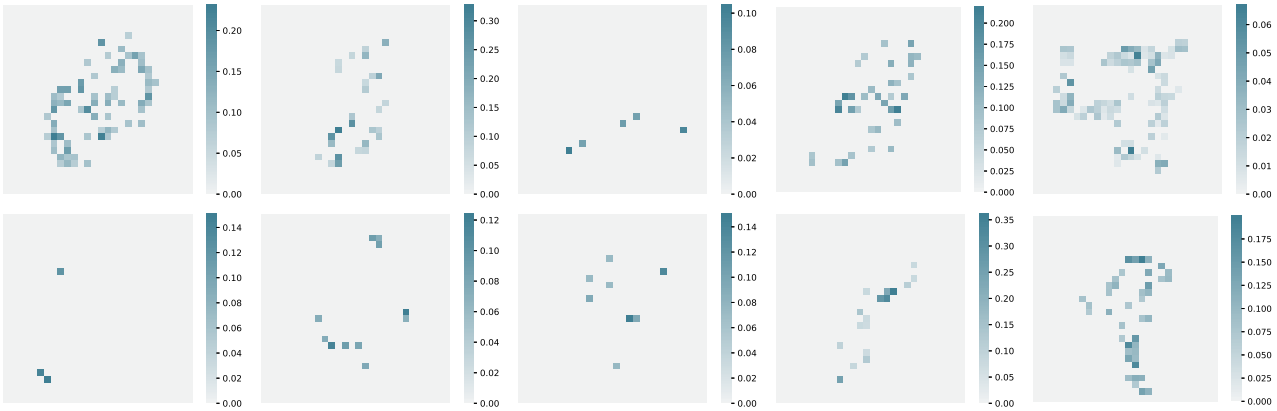

Figure 2: The difference between normalised SHAP values from two SVMs (each trained with different random seeds) for a randomly chosen sample from each MNIST class.

Model Type	Dataset	Dropout	Seed	Shuffle	Explanation Accuracy	Sensitivity	Infidelity
CNN	MNIST	0.0	1	False	96	2.40.	0.0019
		0.1	1	False	97	2.04	0.0018
		0.2	1	False	97	2.02	0.0020
		0.3	1	False	98	1.69	0.0019
		0.4	1	False	98	1.53	0.0016
		0.25	1	False	98	1.84	0.0016
		0.25	12303	False	98	1.70	0.0014
		0.25	15135	False	98	1.58	0.0020
		0.25	16959	False	97	1.67	0.0018
		0.25	20878	False	98	1.61	0.0020
		0.25	79266	True	99	1.51	0.0014
		0.25	79870	True	99	1.67	0.0011
Small-CNN	MNIST	0.0	1	False	99	1.07	0.1810
		0.2	1	False	98	1.00	0.1429
		0.25	1	False	98	1.00	0.1521
		0.25	26417	False	99	1.02	0.1011
		0.25	91110	True	99	1.01	0.1174
		0.25	98281	True	99	1.01	0.1402
GaborNet	MNIST	0.0	0	False	99	1.38	0.2808
		0.1	0	False	99	1.41	0.2256
		0.2	0	False	99	1.42	0.1900
		0.3	0	False	99	1.44	0.1702
		0.4	0	False	99	1.46	0.1523
		0.25	257	False	99	1.17	0.1489
		0.25	6339	False	99	1.34	0.2508
		0.25	29062	False	99	1.40	0.1683
		0.25	51303	False	98	1.45	0.2352
		0.25	17939	True	98	1.34	0.1567
		0.25	23682	True	98	1.28	0.1190
		0.25	27442	True	99	1.31	0.1274
		0.25	53307	True	99	1.27	0.1089
ResNet18	MNIST	0.25	21609	False	99	1.15	0.7214
		0.25	23474	False	99	0.96	0.4426
		0.25	29246	False	99	2.34	0.5284
		0.25	48769	False	98	0.83	0.5007
		0.25	58626	False	99	1.21	0.7121
		0.25	72	True	98	1.21	0.5572
		0.25	1507	True	98	1.42	0.8697
		0.25	4439	True	99	0.97	0.5402
		0.25	10250	True	99	2.10	0.8867
		0.25	21033	True	99	1.01	0.9018
MLP	MNIST	0.0	1	False	99	3.49	0.1748
		0.2	1	False	99	5.56	0.1573
		0.25	1	False	99	4.85	0.1508
		0.25	27833	False	99	3.76	0.1926
		0.25	72	True	99	3.39	0.1427
		0.25	79870	True	99	3.74	0.1399
Densenet121	MIMIC	n/a	2	False	99	1.5966	0.9994
		n/a	3	False	99	1.5031	1.0719
		n/a	4	False	99	1.5987	1.0020
		n/a	5	False	99	1.1431	0.4659
		0.25	6	True	99	1.5122	0.9994
		0.25	7	True	99	1.6078	1.1217
ADP	MNIST	n/a	0	False	99	1.2187	0.9110
		n/a	42	False	99	1.4250	1.4376
		n/a	100	False	98	1.2297	0.9730
		0.25	1	True	99	1.3491	0.9912
		0.25	10	True	98	1.3100	1.266
DNE	MIMIC	n/a	1	False	80	1.5499	1.0357
		n/a	42	False	84	1.3709	0.7340
		0.25	4242	True	81	1.5683	0.6510
		0.25	1000	True	82	1.6932	0.8493
SVM	MNIST	n/a	30828	False	99	1.5763	0.2070
		n/a	31599	False	99	1.1686	0.9074
		n/a	8253	False	99	1.0238	0.6214
		0.25	91244	True	99	1.5439	0.5006
		0.25	79870	True	99	1.5894	0.4823

Table 1: Table reporting explanation quality metrics on SHAP across all model architectures and training variations tested. DNE denotes Densenet-121 Ensemble. Where Shuffle is `True`, Seed refers to the seed used for shuffling the dataset and not the training seed.