Agree to Disagree: When Deep Learning Models With Identical Architectures Produce Distinct Explanations

Matthew Watson, Bashar Awwad Shiekh Hasan, Noura Al Moubayed

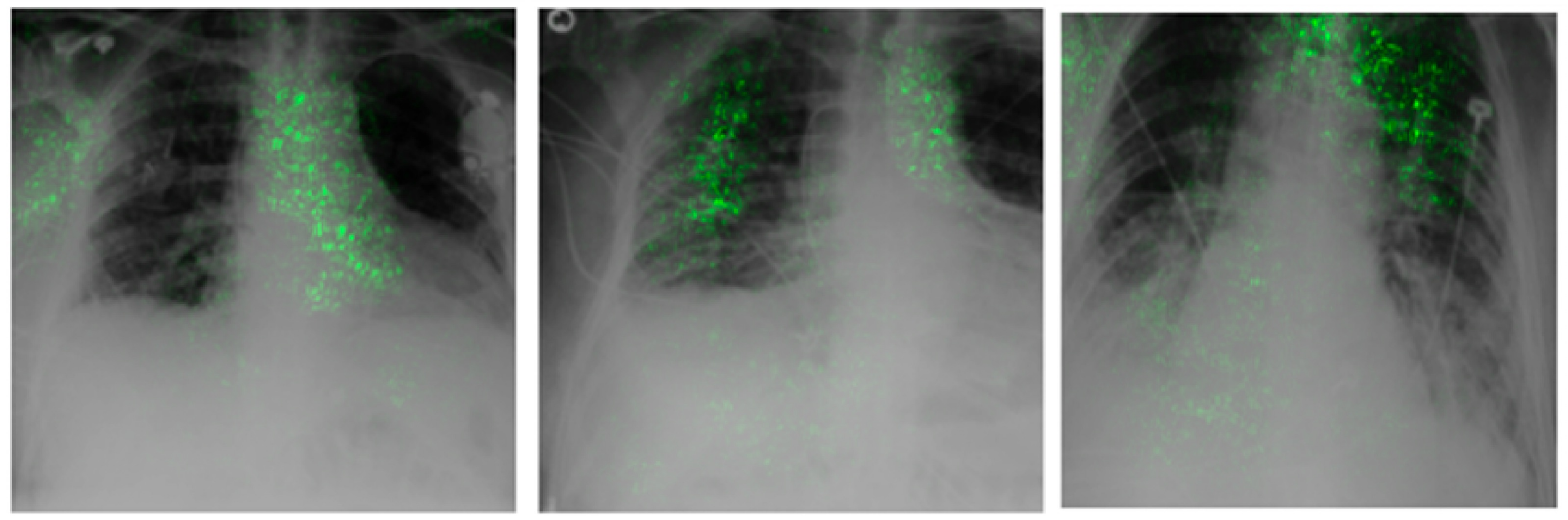
Problem Definition

Recent advances in explainable machine learning have resulted in numerous techniques that can be used to inspect feature importance scores from machine learning (ML) classifiers.

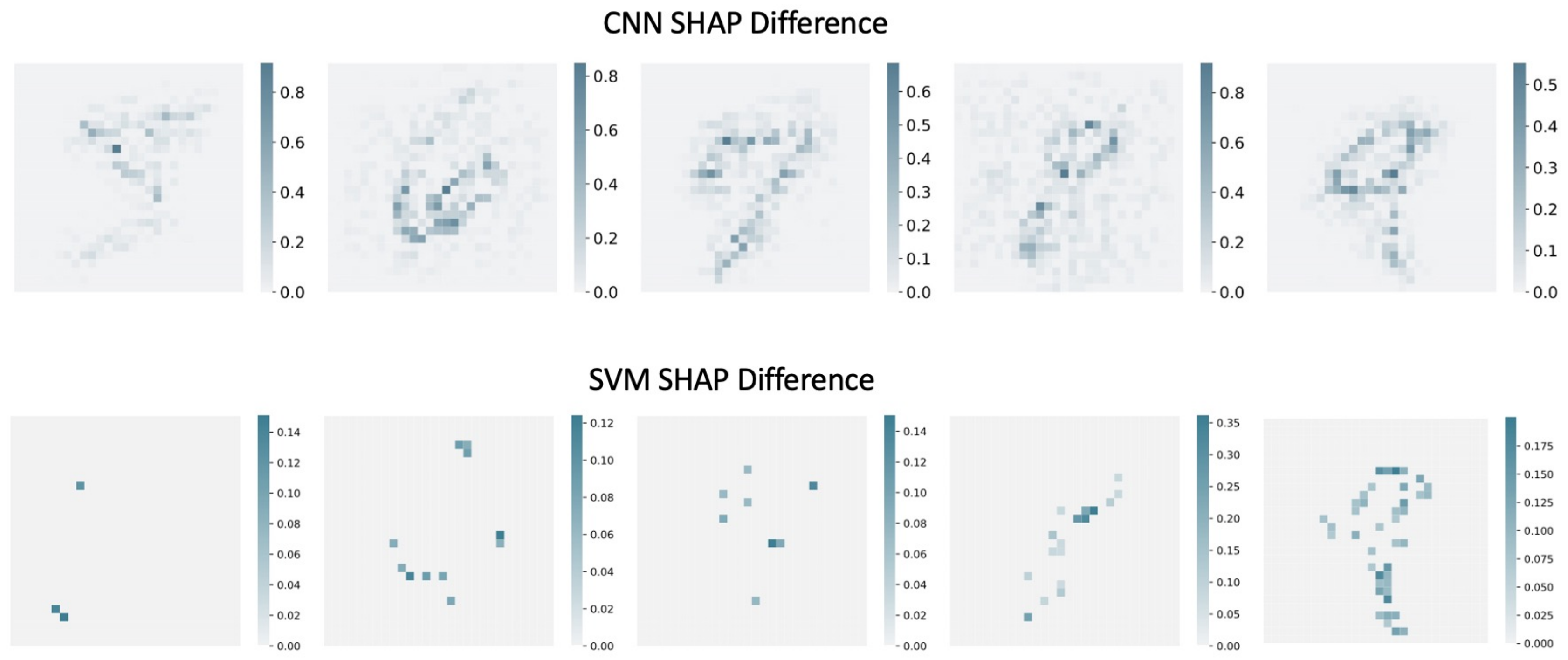
Hyperparameters orthogonal to downstream task can significantly impact model output and associated explanations.

How do model explanations change on identical architectures when training hyperparameters are changed?

We train multiple versions of the same deep learning architecture, changing the hyperparameters each time.



3 random samples from the MIMIC-CXR-JPG dataset overlayed (in green) with the difference between the normalised SHAP values from two Densenet121 training variations.



Absolute difference between the SHAP values of two identical MLPs trained with different random seeds on MNIST (above) vs the absolute different between two SVMs trained with different random seeds on MNIST (below)

Explanation Consistency

We propose a measure for explanation consistency:

$$C = 1 - \frac{\sum_{(a,b)} S_{(a,b)}}{\alpha}$$

Where $S_{(a,b)}$ is a measure of the separability between the explainability values of model a and model b (in the following equation, *D* is any distance measure):

$$S_{(a,b)} = \mathbb{E}_i\Big[D\Big(E(Y^a(x_i)), E(Y^b(x_i))\Big)\Big]$$

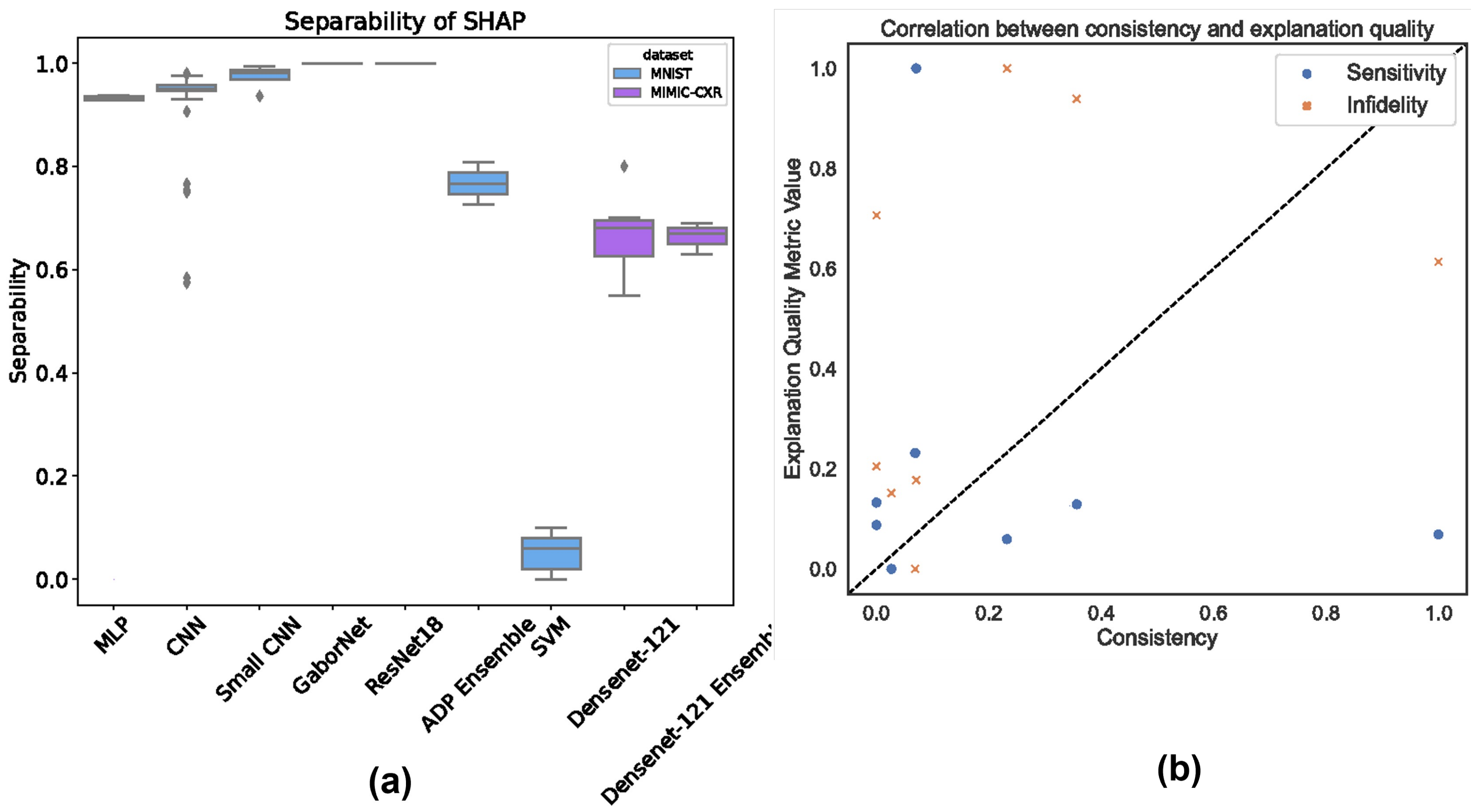
Experiments

Through experimentation, we found a binary LR classifier to be the most suitable choice of *D* in most scenarios. For each pair of models, a binary logistic regression classifier is trained to separate the two sets of SHAP [1] values.

This results in the following definition of explanation consistency using LR accuracy $M_{(a,b)}$ as the separability measure. (a, b) are the training variations and α the number of model pairs tested.

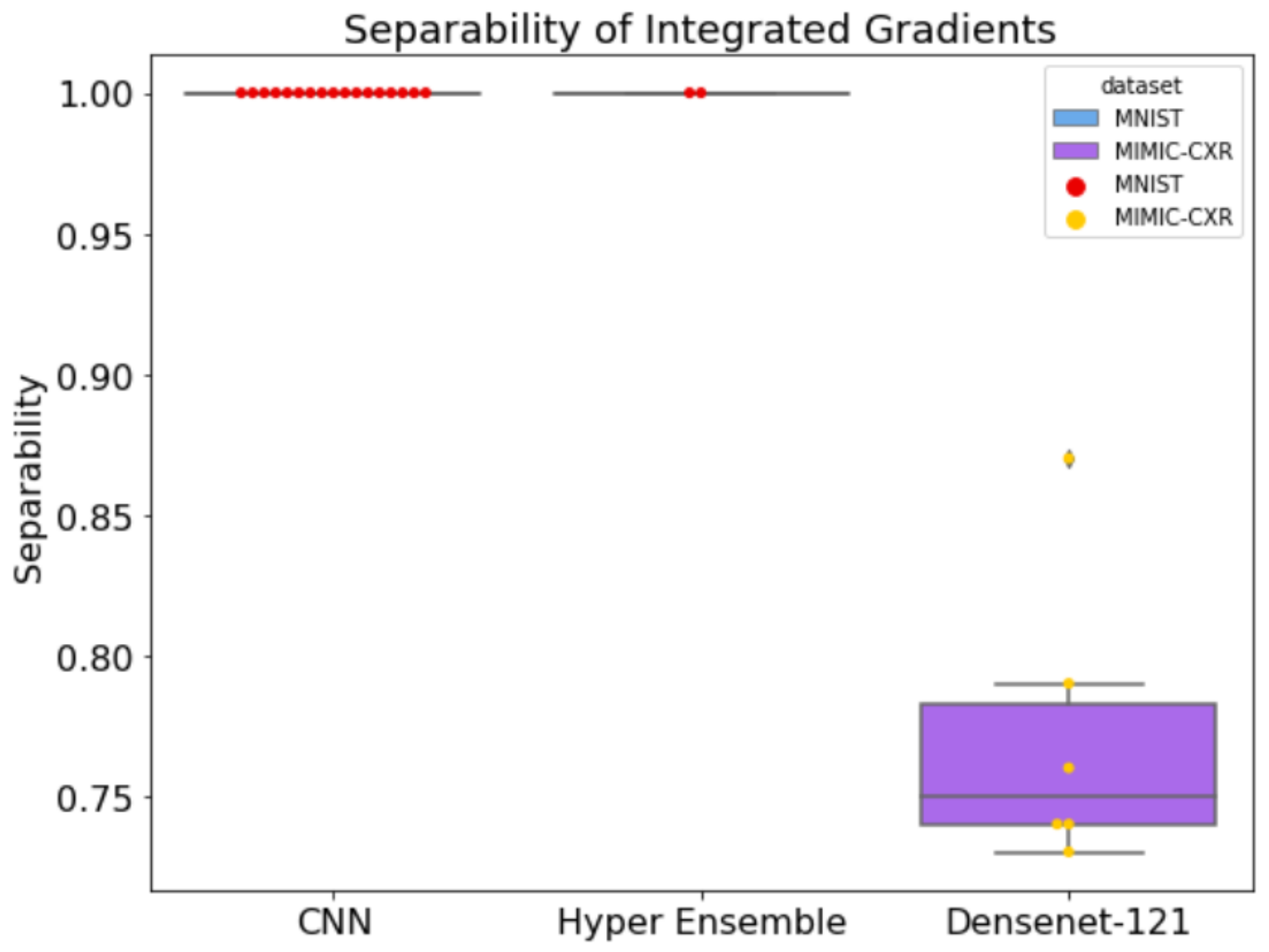
$$C = 1 - \frac{\sum_{(a,b)} 2 * |M_{(a,b)} - 0.5|}{\alpha}$$

All models show high degrees of separability: explanations change significantly when hyperparameters are changed.



(a) Box plot of $S_{(a,b)}$ for SHAP across all training variations (a, b), for all model architectures tested. (b) Plot of SHAP explanation consistency of model architectures vs. SHAP infidelity and sensitivity [2] of the same models across both MNIST and MIMIC-CXR data.

Right: The separability of three model architectures, across two datasets, using Integrated Gradients (IG) [3] instead of SHAP. IG is even less consistent than SHAP; as IG is a gradient-based feature attribution method, this is to be expected as IG's attribution values are calculated based on the network's weights.



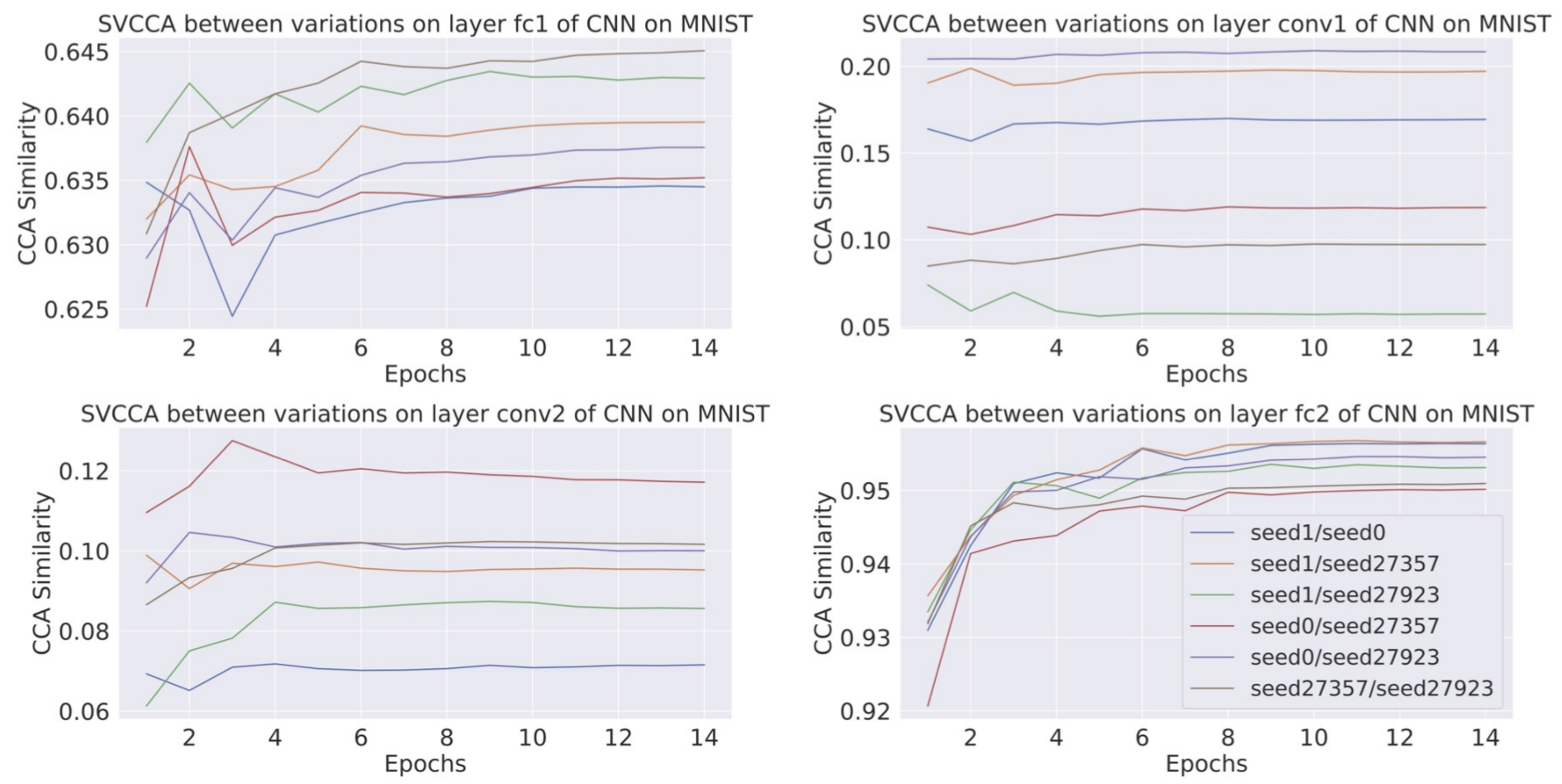
[1] S. M. Lundberg and S.-I. Lee, 'A Unified Approach to Interpreting Model Predictions'
[2] Chih-Kuan Yeh, Cheng-Yu Hsieh, et al. On the (in)fidelity and sensitivity of explanations. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019,* 2019.
[3] Mukund Sundararajan et al.. Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference on Machine Learning, ICML 2017, volume 70 of Proceedings of Machine Learning Research, pages 3319–3328. PMLR, 2017.

GitHub Paper CV News

WACV

BEST PAPER AWARD

Results & Conclusion



Above: CCA similarity as training progresses of layer parameters. a high degree of similarity for the final layer, whereas the middle layer (conv2) shows a significant difference. This corroborates our explainability consistency results; the final layers (fc2) are similar and so the models will produce similar outputs, resulting in similar performance levels.

Explanation consistency results and model accuracy, across all model architectures, datasets and explanation techniques tested.

		Consistency					
Model Architecture	Dataset	α	Overall	Shuffle	Random Seed	Dropout	Accuracy
MLP	MNIST	6	0.0668	0.062	0.066	0.0687	98.125 ± 0.9270
SVM	MNIST	10	0.9444	0.96	0.94	n/a	94.0556 ± 0.6213
Small-CNN	MNIST	6	0.0252	0.018	0.06	0.034	98.3486 ± 0.0360
GaborNet	MNIST	12	0	0	0	0	95.038 ± 0.2824
ResNet18	MNIST	10	0	0	0	n/a	99.425 ± 0.0626
ADP Ensemble	MNIST	6	0.2193	0.192	0.233	n/a	99.083 ± 0.2514
CNN	MNIST	12	0.0652	0.052	0.0564	0.0914	98.9976 ± 0.5756
Densenet-121	MIMIC-CXR	6	0.3329	n/a	0.3329	n/a	75.6723 ± 1.1379
Densenet-121 Ensemble	MIMIC-CXR	4	0.3367	n/a	0.3667	n/a	80.8 ± 0.7483
CNN (IG)	MNIST	12	0	0	0	0	98.9976 ± 0.5756
Hyperensemble (IG)	MNIST	2	0	n/a	0	n/a	99.32 ± 0.0082
Densenet-121 (IG)	MIMIC-CXR	6	0.168	0.115	0.2033	n/a	75.6723 ± 1.1379

Table reporting the consistency between training variations for the models tested and the average accuracy of the model architecture on the base classification task. The Shuffle, Random Seed and Dropout columns report the consistency of models when *only* the respective hyperparameter was changed. The Overall column reports the overall consistency of that architecture, taking an average of the consistency across all hyperparameters. α refers to the number of models tested for the overall architecture consistency.

Our experiments on kernel methods, and explanation quality, demonstrate this is a problem with **DL models and not necessarily the explanation techniques**. This shows DL models are **not robust** - explanations change significantly with changes of the hyperparameters perpendicular to the task at hand.
DL adoption will be hindered in scenarios where transparent models are of paramount importance.