Attack-agnostic Adversarial Detection on Medical Data Using Explainable Machine Learning

Matthew Watson, Noura Al Moubayed
Department of Computer Science

Durham University

Motivation

Medical machine learning models are highly susceptible to adversarial attacks, leading to reduced trust from clinicians [1]

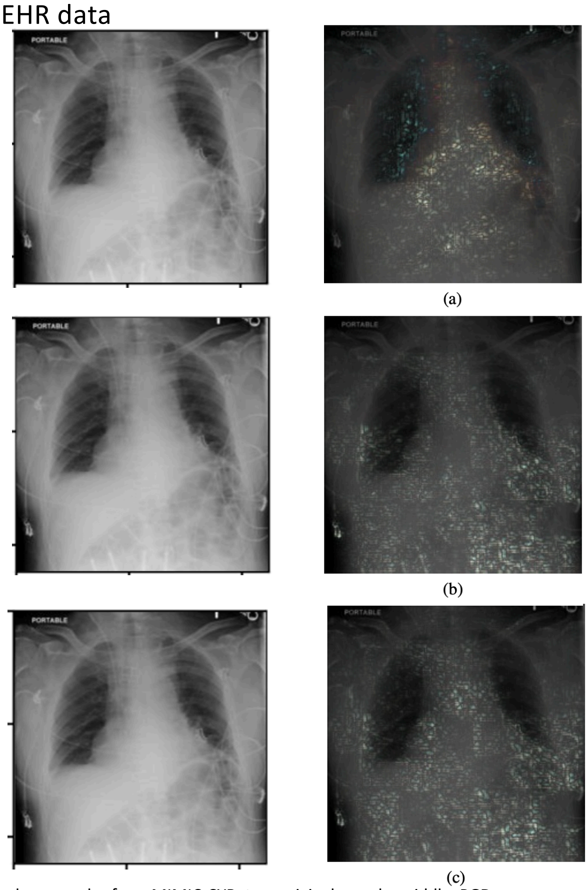
Model	Acc. original data	Acc. adv. data
MIMIC-III RETAIN	81%	43%
Henan-Renmin RETAIN	73%	44%
MIMIC-CXR Densenet121	82%	0%

Accuracy of model trained on original data when tested on genuine data vs. adversarial data

Adversarial Attacks

(1) Projected Gradient Descent (**PGD**) [2] and Carlini & Wagner (**C&W**) [3] attacks on image data

(2) Longitudinal AdVersarial Attack (**LAVA**) [4] on EHR data

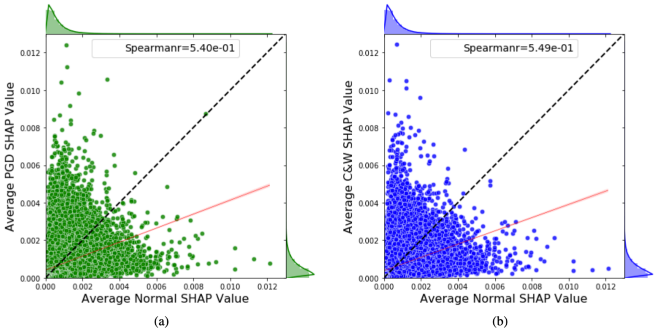


Random samples from MIMIC-CXR; top: original sample, middle: PGD perturbation, bottom: C&W perturbation. The right column shows the images overlayed with SHAP values from a finetuned Densenet121 model

Approach

(1) Use explainability techniques (**SHAP**) to identify sections of the input with high importance

(2) Verify that genuine and adversarial samples have significantly different SHAP values



Correlation between genuine SHAP values and (a) PGD SHAP values; (b) C&W SHAP values

(3) Use **MLP**s, **CNN**s on SHAP values to identify adversarially perturbed samples

(4) Use **VAE**s trained on genuine SHAP values to create a model that can accurately detect adversarial samples from any attack method as anomalies

Conclusions

- Adversarial attacks modify the features of the input that model's place importance on
- SHAP can reliably detect adversarial samples
- Beating current state of the art performance on medical datasets
- MLPs and CNNs are useful in one-attack scenarios
- VAEs are able to detect unseen attacks when the problem is modelled as an anomaly detection scenario

[1] S. G. Finlayson, I. S. Kohane, and A. L. Beam, "Adversarial attacks against medical deep learning systems," *CoRR*, vol. abs/1804.05296, 2018.

[2] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[3] N. Carlini and D. A. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*. IEEE Computer Society, 2017, pp. 39–57.

[4] S. An, C. Xiao, W. F. Stewart, and J. Sun, "Longitudinal adversarial attack on electronic health records data," in *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*. ACM, 2019, pp. 2558–2564.

[5] R. Feinman, R. R. Curtin, S. Shintre, and A. B. Gardner, "Detecting adversarial samples from artifacts," *CoRR*, vol. abs/1703.00410, 2017.

[6] X. Ma, Y. Niu, L. Gu, Y. Wang, Y. Zhao, J. Bailey, and F. Lu, "Understanding adversarial attacks on deep learning based medical image analysis systems," *CoRR*, vol. abs/1907.10456, 2019.

Method	Datasets					
	MIMIC-III	HR	CXR (C&W)	CXR (PGD)	CXR (Train: PGD;Test: C&W)	CXR (Train: C&W;Test: PGD)
SHAP-MLP	**77%**	**81%**	**100%**	99%	58%	46%
SHAP-AE + SVM	65%	53%	79%	79%	77%	79%
SHAP-VAE + SVM	66%	53%	85%	88%	**86%**	**88%**
SHAP-Conv	N/A	N/A	**100%**	**100%**	55%	65%
Kernel Density [5]	67%	67%	84%	83%	72%	66%
ML-LOO [6]	N/A	N/A	71%	78%	71%	71%

cievert

EUROPEAN UNION
European Regional Development Fund