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

Matthew Watson, Noura Al Moubayed

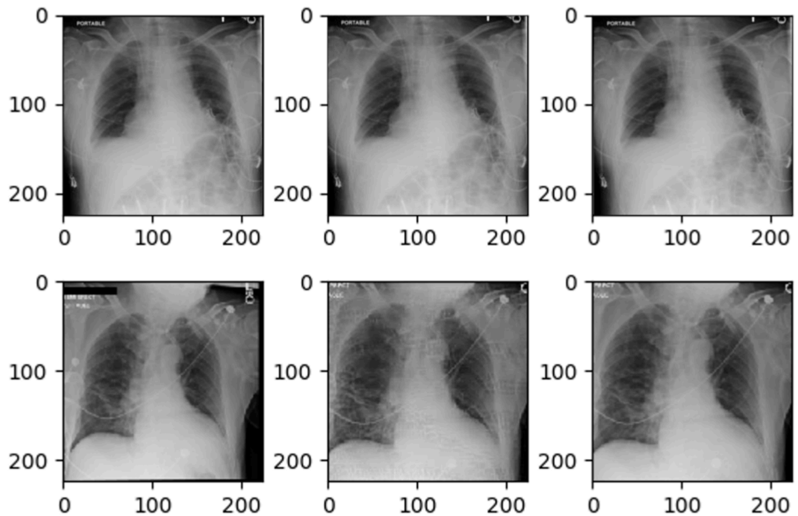
Department of Computer Science

Durham University

EUROPEAN UNION
European Regional Development Fund

Adversarial Attacks

- Making imperceptible changes to the input often changes a model's output [1]: PGD [2], C&W [3]
- We can leverage this to fool a model into making an incorrect prediction
- Even when a human is unable to tell the difference



Two random samples from MIMIC-CXR. Left: original sample, middle: PGD perturbation, right: C&W perturbation

Durham University

Adversarial Attacks in Healthcare

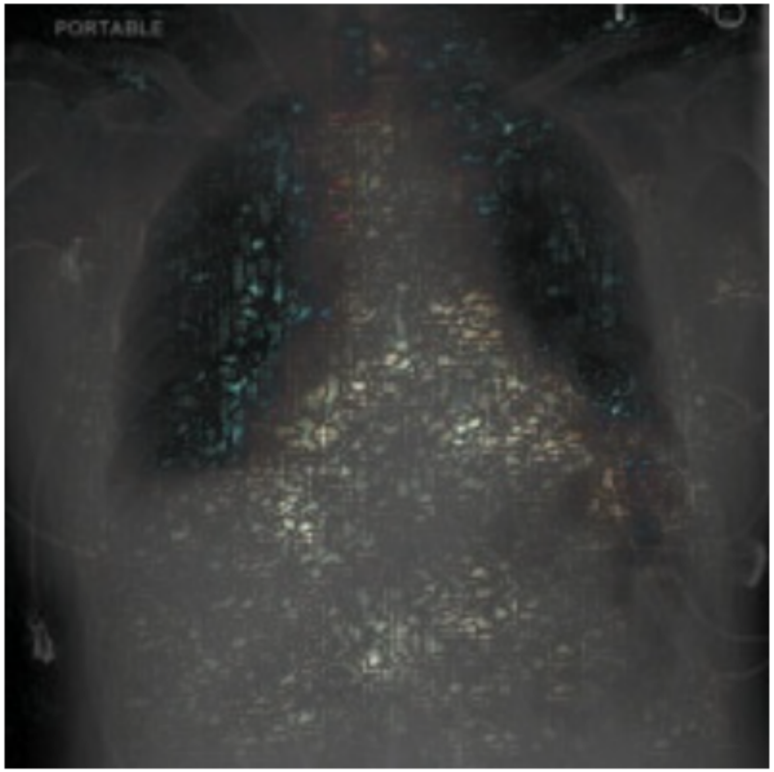
- ML's susceptibility to such attacks reduce trust from clinicians
- Robust decision making is a requirement for ML's deployment in healthcare

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

Durham University

Explainable ML

- Can we trust a classifier?
- How can we check a classifier isn't making spurious correlations?
- Needed for ethical and validated machine learning in healthcare
- SHAP: Current state of the art explainability method
    - Approximates the change in expected model prediction when conditioning on each (combination of) feature(s)

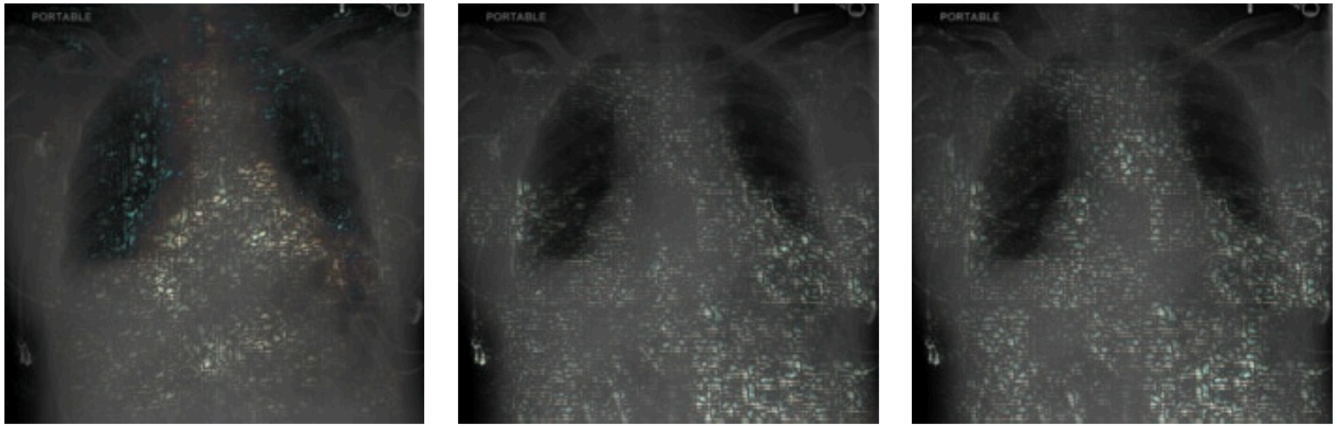


Durham University

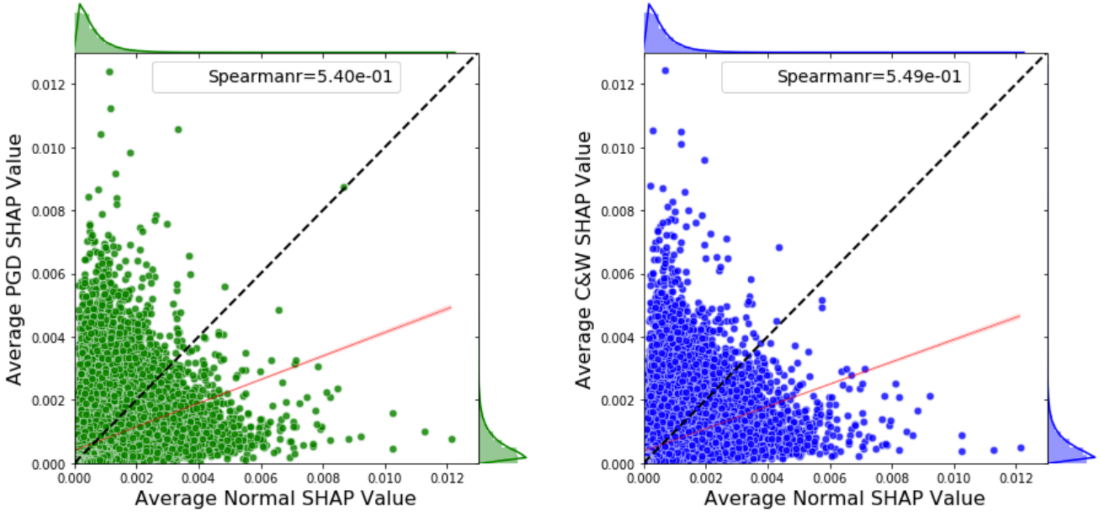
Can explainable ML detect adversarial attacks?

Durham University

Explanations Highlight Attacks



SHAP values on a sample from MIMIC-CXR and a Densenet-121 model trained to detect Cardiomegaly. Left: original sample, middle: PGD attack, right: C&W attack



Figures showing the average absolute importance of each feature in the original MIMIC-CXR dataset, calculated using SHAP values against the adversarial samples.

Durham University

Single-attack Detection

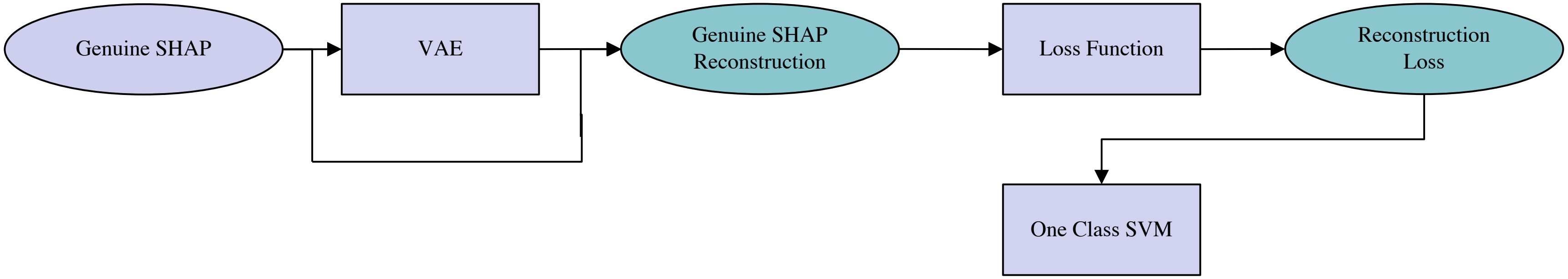
- Using CNNs and MLPs we can accurately classify the origin of explanations:
  - Are the explanations from genuine or adversarial samples?
- We show our methods work on a variety of complex medical datasets
- But what if new adversarial attacks are developed?

Durham
University

Attack-agnostic Detection

- We re-frame the problem as anomaly detection
- VAEs are trained on genuine explanations only
  - One-class SVMs are then used on the reconstruction error

Anomaly Detection Training

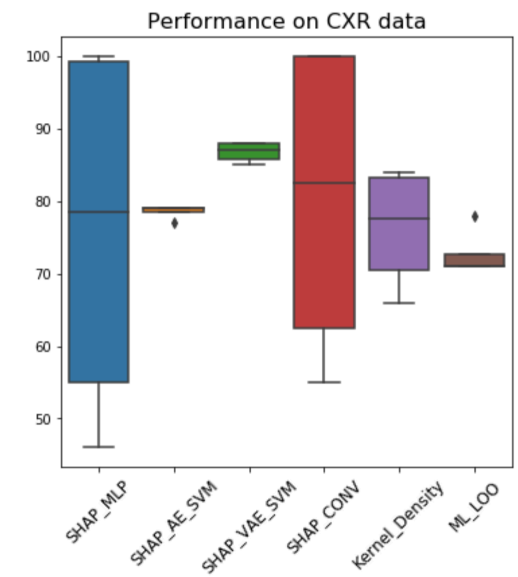


Durham University

Results

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	67%	67%	84%	83%	72%	66%
ML-LOO	N/A	N/A	71%	78%	71%	71%

Results of adversarial sample detection. CXR (C&W) reports the accuracy on C&W generated samples, having been trained on C&W samples, and CXR (PGD) the a curacy od a model trained on PGD samples tested on PGD samples.



Boxplot reporting the performance of adversarial sample detection methods on CXR data.

Durham University

Conclusions

- Adversarial attacks modify the features of the input that model's place importance on.
- We demonstrate explainability techniques can be used to identify adversarial samples.
- This technique works on medical data
  - Despite the challenges that such data poses, such as high-dimensionality and ambiguous ground truths
- MLPs and CNNs can be used in one-attack scenarios.
- Whereas VAEs provide generalisation to unseen attacks.

Durham University

References

[1] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

[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.

Durham University