2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

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

Matthew Watson
*Department of Computer Science*
*Durham University*
Durham, UK
matthew.s.watson@durham.ac.uk

Noura Al Moubayed
*Department of Computer Science*
*Durham University*
Durham, UK
noura.al-moubayed@durham.ac.uk

***Abstract*—Explainable machine learning has become increasingly prevalent, especially in healthcare where explainable models are vital for ethical and trusted automated decision making. Work on the susceptibility of deep learning models to adversarial attacks has shown the ease of designing samples to mislead a model into making incorrect predictions. In this work, we propose a model agnostic explainability-based method for the accurate detection of adversarial samples on two datasets with different complexity and properties: Electronic Health Record (EHR) and chest X-ray (CXR) data. On the MIMIC-III and Henan-Renmin EHR datasets, we report a detection accuracy of $77\%$ against the Longitudinal Adversarial Attack. On the MIMIC-CXR dataset, we achieve an accuracy of $88\%$; significantly improving on the state of the art of adversarial detection in both datasets by over $10\%$ in all settings. We propose an anomaly detection based method using explainability techniques to detect adversarial samples which is able to generalise to different attack methods without a need for retraining.**

***Index Terms*—Adversarial Attacks, Explainability, SHAP, Medical Data**

## I. Introduction

Recently, applications of machine learning in healthcare have shown great success. Machine learning models trained on EHR data are able to predict (with high accuracy) heart failure [1], interpret mammograms [2] and diagnose CXR [3], and in some cases can match the performance of human experts. However, it is now well demonstrated that such models are susceptible to adversarial attacks: attacks that generate samples designed to mislead a machine learning model into making an incorrect prediction [4]. Examples of such attacks are also effective on medical data such as EHR [5] and medical imaging data [6]. The presence of adversarial attacks is of particular concern in the medical domain as it would be unethical to deploy a machine learning model to clinical practice if it is considered vulnerable to such malicious attacks, even if the likelihood of an attack is low [6].

Healthcare ML models are at particular risk of adversarial attacks [6]–[8]. Fraud is already pervasive in the US' healthcare economy, with institutions systematically inflating costs and physicians billing for the largest amount possible [6], [9] and, with machine learning algorithms l ikely to be used for medical decisions in the near future [10], adversarial attacks on ML models will be a new avenue for fraud to occur. The pharmaceutical and medical device markets are also domains where adversarial attacks on medical machine learning systems are a risk. The large amounts of money involved in these markets (the median revenue for a single cancer drug is estimated to be $1.67 billion [11]) combined with the increasing number of drug/device approval decisions being made based on digital surrogates for patient responses (for example, in medical imaging [12]) means that extremely valuable decisions are being made by machine learning algorithms and as such are a likely target for adversarial attacks.

There are also technical vulnerabilities present in many ML models used in healthcare [7], [13]: from low variance in training sets to similar models being used for many different tasks increasing their vulnerability to attacks. Healthcare professionals commonly cite susceptibility to adversarial attacks as a challenge to further adoption of ML in healthcare [8], with the UK's National Health Service (NHS) identifying it as a problem that must be overcome for a machine learning model to be used within the healthcare system [14]. For these reasons, it is prudent to develop methods of defending against, and detecting, adversarial attacks to provide trust in machine learning solution in medical settings [6]–[8], [14].

In parallel, there has recently been an increased effort to improve the explainability of machine learning models. This area of research aims at explaining the decisions made by black-box machine learning models by making the decisions and the processes behind those decisions understandable to a non machine learning expert [15]. This has resulted in a number of methods being developed that allow for post- and ante-hoc explanations of models and their decisions [16].

As adversarial attacks change parts of the input, we hypothesise that ML models place more importance upon these perturbed sections of the input when passed an adversarially perturbed sample. This paper introduces a method that utilises techniques from explainable ML to detect when an adversarial sample is passed to a model by inspecting the areas of the input that the model deems most important. The paper's main contributions are: **I)** The first adversarial sample detection technique that works effectively with EHR data. **II)** We propose a novel and simple method for detecting adversarial attacks using explainable techniques and demonstrate that it beats the state of the art on both medical imaging and EHR data despite the sparse, temporal and high-dimensional nature of the data. **III** The method is model agnostic and will support any machine learning model **IV)** By framing the adversarial detection as an anomaly detection problem this work presents an approach that generalises to any attack type without the need to retrain [1].

## II. Related Work

In this section we provide an overview of current state-of-the-art in explainability and techniques for adversarial generation and detection.

### A. Adversarial Attacks on Medical Data

Adversarial attacks have been developed for numerous data modalities and scenarios [17]. Finlayson *et al.* [6] demonstrate that, despite the challenges that medical imaging data presents, traditional adversarial attack techniques such as Projected Gradient Descent (PGD) [18] and patch attacks [19] can still successfully produce inputs that force a classifier to predict the incorrect label. The authors in [20] demonstrated that adversarial samples can be transferable across models. The authors also introduced a set of three attacks, known as C&W attacks, that are capable of bypassing some of the most robust machine learning models to adversarial attacks.

Longitudinal AdVersarial Attack (LAVA) [5] is designed to generate attacks that are effective on EHR data. LAVA is a saliency score based method that works on discrete and sequential EHR data. By utilising the saliency score it avoids perturbing features that would easily be detected by a human expert whilst maintaining a minimal number of perturbations. The authors showed that it can reduce the accuracy of an attention-based model from $50\%$ to $8\%$.

### B. Adversarial Attack Detection

Metzen *et al.* [21] show that it is possible to detect adversarial samples, despite their imperceptible feature changes, through the training of a simple binary classifier. Feinman *et al.* [22] proposed methods of detecting adversarial samples utilising density estimates of the final hidden layer of the model, and a Bayesian uncertainty estimate. These are designed to complement each other: the density estimate detects adversarial samples as they tend to lie outside the data manifold, and the Bayesian uncertainty detects points in low-confidence regions of the input space. Ma *et al.* [7] show that the methods of [22] can also be successfully applied to medical imaging data. However, no adversarial detection methods have yet been proposed to work on EHR data mainly due to the challenge of dealing with its temporal dependency and high-dimensionality. Significantly, these adversarial detection methods are extremely model-dependent (e.g. Bayesian uncertainty requires dropout networks) and most require retraining for different types of adversarial attacks.

The methods presented in [23] are designed to detect any abnormal sample that is sufficiently far away from training distribution; this includes adversarial samples, but also out-of-distribution samples. The method is based on the probability density of the test sample on the feature space of the neural network using a generative classifier and is able to generalise to unseen attack methods with only a small reduction in accuracy. However, it is only able to be used with classifiers which utilise Softmax.

ML-LOO [24] uses the Leave-One-Out (LOO) explainability technique to detect adversarial samples. LOO is a feature attribution method that uses the reduction in the probability of the selected class when the feature is masked/removed. The authors show that LOO results in the best performance for adversarial detection when compared with other feature attribution methods, however it can be very computationally expensive to compute and as such is impractical when datasets contain a large number of features (i.e. CXR images).

### C. Explainable Machine Learning

The development of explainable machine learning techniques has significantly increased recently. This is mainly driven by regulator's and end user's increased awareness of the impact of machine learning models and the need to understand their decisions. This is of particular interest in healthcare, where interpretable machine learning is seen highly important due to the need for ethical and validated decision making [14]. A comprehensive review of explainable methods can be found in [25]. The most common method to date is SHAP [26]. SHAP approximates the change in expected model prediction when conditioning on each (combination of) feature(s) and is closely tied to Shapley values [27]. Lundberg *et al.* [26] propose a number of both model-agnostic and model-specific approximations that enable the practical computation of SHAP values.

## III. Methodology

We introduce novel solutions that utilise SHAP values to detect adversarial attacks and demonstrate that it works on both medical imaging and EHR data. The proposed solutions consist of both fully- and semi-supervised methods, and exploits the differences between the distribution of SHAP values of genuine and perturbed samples in order to accurately detect adversarial samples. Furthermore, as SHAP values are consistent across the entire genuine dataset, our semi-supervised solution is able to generalise to adversarial attacks generated by alternative methods without the need for retraining.

### A. Datasets and Classification Models

Due to privacy concerns around healthcare data, there has traditionally been few sufficiently large, open datasets available in the literature. We utilise 2 EHR datasets: MIMIC-III [28] and Henan-Renmin[2], and 1 medical imaging dataset: MIMIC-CXR [29].

MIMIC-III [28] is a large EHR dataset collected from the Beth Israel Deaconess Medical Center in Boston. It contains 53,423 records of adult admissions of 38,597 distinct patients

[1] Supporting code can be found at: https://github.com/mattswatson/attack-agnostic-adversarial-attack-detection

[2] http://pinfish.cs.usm.edu/dnn/

to the Intensive Care Unit (ICU) between 2001 and 2012, in addition to 7870 neonatal cases admitted between 2001 and 2008. On average, each admission contains 4579 charted observations and 380 laboratory measurements. All data was collected during routine clinical care and includes bedside monitoring notes, lab and microbiology test results, diagnosis and procedure codes, and demographic information.

The Henan-Renmin dataset contains records from 110,300 patients, with significantly fewer features; 62 features per patient comprised of basic examinations and clinical tests. The class label for each record is a combination of three possible diagnoses: hypertension, diabetes and/or fatty liver.

RETAIN [1] is a state-of-the-art model designed specifically to work with EHR data. The model aims to mimic typical physician practice by inspecting EHR data in reverse-time order, such that more influence is given to more recent visits when making the final classification. In order to provide interpretable results, RETAIN has a two-level neural attention model that first detects key visits and then detects the key diagnoses from these visits.

We train RETAIN on the MIMIC-III dataset [28]. This results in an accuracy of 81% when predicting patient mortality. To ensure that our adversarial attack detection method adapts to different datasets, we also train the RETAIN model on the Henan-Renmin dataset to predict hypertension, with an accuracy of 73%. Hypertension is chosen as it is the most prevalent single label, providing mostly balanced classes. The RETAIN model is not as accurate with this dataset compared to MIMIC-III mainly due to lower number of features.

MIMIC-CXR [29], also collected from the Beth Israel Deaconess Medical Center, is a database of 377,110 chest x-rays from 227,827 studies collected between 2011 and 2016. Each study has an associated free text summary report by a radiologist. The reports are analysed using a label extraction tool such as CheXpert [30] to generate 14 weak labels (of different diagnoses) for the x-ray images. On a stratified test set of 687 manually labelled (by a certified radiologist) images from the MIMIC-CXR dataset, [29] report that the label extraction method has an accuracy of 95% for the Cardiomegaly label.

First, we run CheXpert on the radiologists' reports to extract the diagnosis, resulting in 14 labels, each of which is classified as either a positive mention, a negative mention or an uncertain mention. Following [30] we treat all uncertain labels as positive mentions. We focus on the Cardiomegaly label as this is both a common diagnosis and provides a balance between positive/negative labels with a low number of uncertain mentions.

Furthermore, we ignore any chest x-rays whose reports do not contain any mention of Cardiomegaly. If these were included, it would be difficult to apply a label to them without making further assumptions. This process generates a set of weak labels for each chest x-ray. An image may have multiple labels (e.g. a patient may have both Cardiomegaly and Pneumonia). We fine-tune Densenet-121 [31] (pre-trained on ImageNet [32]) on MIMIC-CXR, based on the method presented by Rajpurkar *et al.* [33], to predict a diagnosis of Cardiomegaly, achieving an accuracy of 82%.

### B. Adversarial Sample Generation

We use state-of-the-art adversarial sample generation techniques that are known to be successful on medical data. LAVA [5] is used for the two EHR datasets. Both RETAIN trained on MIMIC-CXR and RETAIN trained on Henan-Renmin see a significant reduction in accuracy, as shown in Table I. The reduction in accuracy is similar to that reported in [5].

TABLE I
TABLE SHOWING ACCURACY OF THE MODELS ON THE ORIGINAL AND ADVERSARIAL ATTACK DATASETS. AS PGD NECESSARILY PERFORMS PERTURBATIONS UNTIL THE SAMPLE IS CLASSIFIED INCORRECTLY, THE MIMIC-CXR MODEL MUST ACHIEVE AN ACCURACY OF 0% ON THE ADVERSARIAL SET.

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

Projected Gradient Descent (PGD) [18] is used to generate the CXR adversarial samples. PGD produces adversarial images to mislead the machine learning model whilst keeping the perturbations small enough that they are not easily detected via traditional methods or even the human eye. Fig. 1 shows examples generated from random samples in the MIMIC-CXR dataset. As shown in Table I, PGD successfully produces adversarial samples that are able to mislead the model into making an incorrect classification.

In order to test our method's ability to generalise to different attack types, we use the attack method proposed by Carlini & Wagner [20] (C&W). Unlike PGD which uses $L_\infty$ norm, C&W uses the $L_2$ distance metric to produce a second set of adversarial samples for the MIMIC-CXR dataset. These two approaches are chosen as they perturb the images differently and hence we can better test the generalisation of our approach.

### C. Adversarial Attack Classification

As adversarial attacks subtly change small parts of the input, we hypothesize the SHAP values for an adversarial sample will be different than those for a genuine sample. This is illustrated by Fig. 2, which shows how PGD and C&W affect the distribution of SHAP values compared to the SHAP values of genuine data (correlation is low between the two with most values away from the ideal linear line). This demonstrates that although adversarial attacks methods aim to make the minimal feature perturbations possible, they still greatly impact the distribution of the explanation of the model predictions. Fig. 2 also demonstrates that the PGD and C&W attacks perturb the samples differently.

In order to quantify the importance that our models place on different parts of their respective inputs, we utilise SHAP values as calculated by GradientSHAP [26]. SHAP values reflect the contribution of each individual feature to a model's prediction, which is important when only a small number of

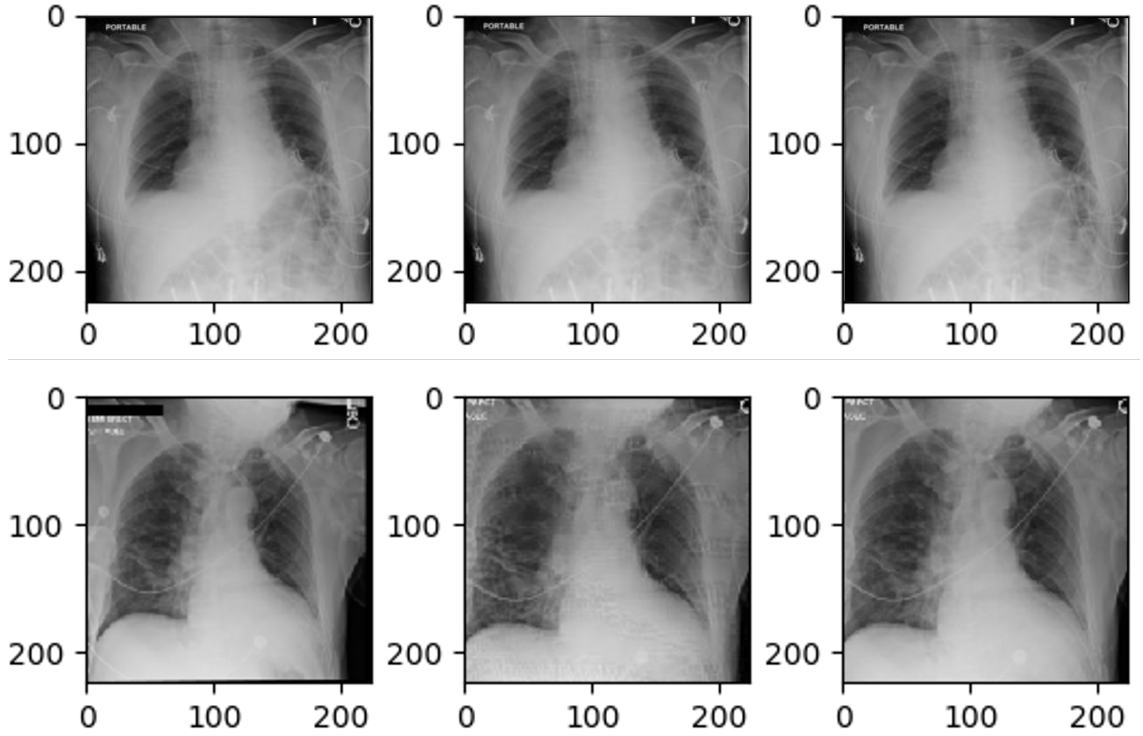

Fig. 1. Random adversarial examples generated on the MIMIC-CXR. Images on the left are the original images, the middle have been generated via PGD, and the right via C&W.

features are changed under perturbation during the adversarial attack.

SHAP values are the (approximate) solution to Eq. 1, where $\phi_i(f, x)$ is the importance of feature $i$ of input $x$ to model $f$, $M$ is the number of features, $|z'|$ is the number of non-zero entries in $z'$, $z' \subseteq x'$ represents all $z'$ whose non-zero entries are a subset of the non-zero entries in $x'$ and $S$ is the set of non-zero indices in $z'$.

$$\phi_i(f,x) = \sum_{z'\subseteq x'} \frac{|z'|!(M-|z'|-1)!}{M!}\left[E[f(x)|z_S] - E[f(x)|z_{S\setminus i}]\right] \tag{1}$$

We calculate SHAP values for the unperturbed (genuine) dataset and the set of perturbed samples to generate the data for the negative and positive class respectively. Fig. 3 demonstrates how the SHAP values for a sample change when the model is looking at a perturbed sample, illustrating how a model focuses on different parts of the input when presented with an adversarial sample. The model seems to utilise clusters of pixels in the chest area in the original picture while the important pixels are scatter across the attack images.

We propose both fully- and semi-supervised methods using SHAP values to detect adversarial samples utilising this information.

**SHAP-MLP:** We train a simple multi-layer perceptron (SHAP-MLP) on the set of SHAP values from both genuine and adversarial samples of the dataset. The model consists of an input layer, output layer and a single hidden layer. More details about the model are in Section IV.

**SHAP-Conv:** We train a convolutional neural network (CNN) on the set of SHAP values from both genuine and adversarial samples. The CNN consists of two convolutional layers, the first going from 3 channels to 16 with a kernel of size 5 and the second going from 16 channels to 32 with a kernel size of 5. We use max pooling with a kernel size and stride of 2, and the ReLU activation function throughout. Following the convolutional layers is a series of 3 fully connected layers of sizes $89888 \times 256$, $256 \times 84$ and $256 \times 1$. We apply dropout with a probability of $0.4$ after the second convolutional layer and again after the second fully connected layer.

**SHAP-AE & SHAP-VAE:** Typically, an adversarial attack can be seen as any sample which a model classifies incorrectly; this can include genuine images which the model misclassifies. SHAP-MLP and SHAP-Conv both attempt to classify these images as adversarial. However, it is often more useful to only detect samples which have been specifically perturbed to be adversarial [22]. This results in a smaller number of samples being present in the adversarial set. Therefore we propose the use of anomaly detection methods to detect the adversarial samples.

We experiment with two semi-supervised models: autoencoders (SHAP-AE) and variational autoencoders (SHAP-VAE) [34] trained to reproduce SHAP values of genuine samples. The reconstruction error of the autoencoder, i.e. the error

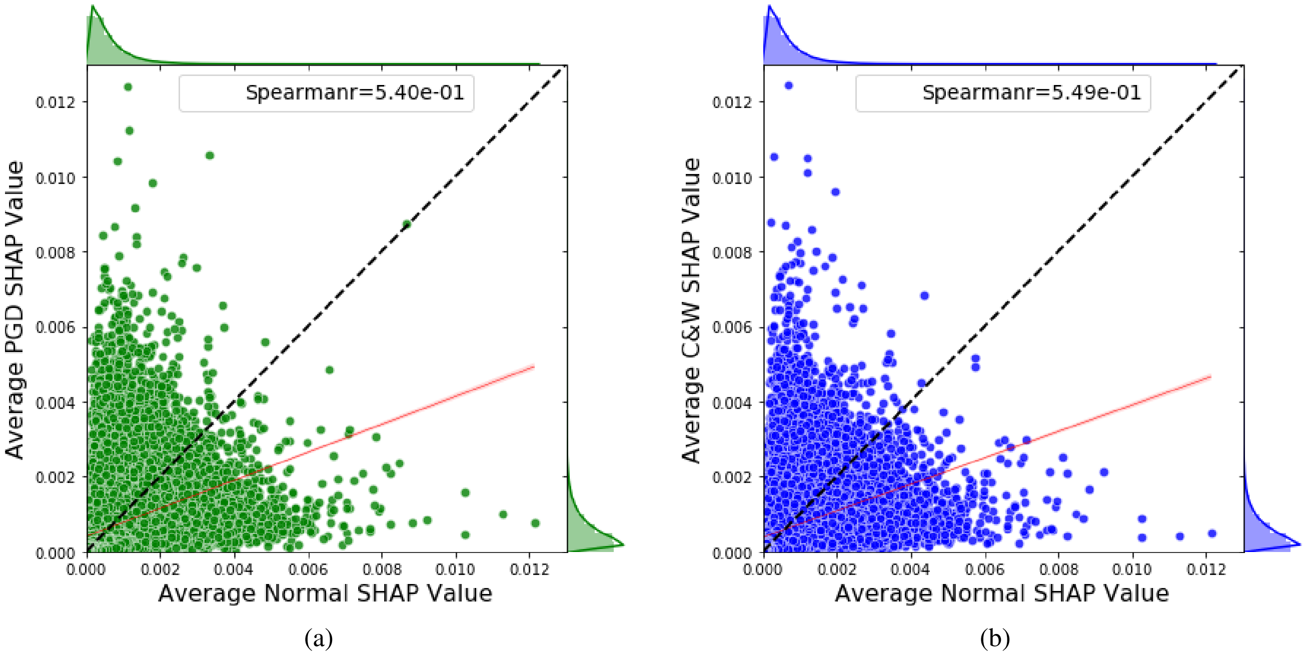

Fig. 2. Figures showing the average absolute importance of each feature in the original MIMIC-CXR dataset, calculated using SHAP values against the adversarial samples. (a) Scatter plot of the SHAP values of PGD adversarial samples on the Y axis against the SHAP values of original sample on the X axis, the dashed line represents the ideal line while the red line is the linear fit. The histogram of each axis is plotted. The Spearman Rank correlation value is reported.(b) Scatter plot of the SHAP values of C&W adversarial samples on the Y axis against that of the original set on the X axis.

between the original and reconstructed value, is then used as a measure to detect an adversarial sample. For SHAP-AE, mean squared error (MSE) is used as the loss function. For SHAP-VAE, MSE plus the Kullback-Liebler divergence is used. As the autoencoder is trained only on genuine SHAP values, the reconstruction error from adversarial SHAP values are expected to be higher - the (V)AE has not learned how to reproduce the adversarial values. We thus train an SVM to classify reconstruction error into two classes (adversarial and genuine). The performance of both methods are reported in Section IV. As both of these methods are semi-supervised approaches, they are able to generalise to different attack types; they learn to reproduce the SHAP values of a genuine dataset, so anything that deviates from that is labelled adversarial. This would be a useful property, as it enables the model to detect novel, unseen attacks.

## IV. Experiments and Results

### A. Experiments on EHR data

We first report the results of experiments on EHR data. Throughout all experiments, we normalise the SHAP values so they have a mean of 0 and variance of 1, and have a train/test split of 80/20. We train SHAP-MLP on the genuine and adversarial SHAP values from the MIMIC-III dataset. A grid-based cross validation search method is used to find the optimal hyperparameters for SHAP-MLP, resulting in a hidden layer of dimension 160 and a learning rate of 0.01 with the Adam optimiser. This leads to an accuracy of 77%. Similarly, on the Henan-Renmin dataset, a hidden layer dimension of 140 and learning rate of 0.01 are optimal, achieving an accuracy of 81%.

A similar approach is used for testing the autoencoder-based methods. SHAP-AE and SHAP-VAE are both trained on the set of genuine SHAP values from MIMIC-III and Henan-Renmin. After performing the same hyperparameter optimisation method described above, we find that an autoencoder with 2 hidden layers (in both the encoder and decoder), a code size of 20 and a learning rate of 0.01 with an Adam optimiser provides optimal results. Experiments find that an SVM with an RBF kernel with $C = 1$ and $\gamma = \frac{1}{M}$ (where $M$ is the number of features) gives the best results compared to logistic regression, and SVMs with other parameters, that are validated using grid-based cross validation search. Similarly, SHAP-VAE has a code size of 5 and a learning rate of 0.01 with an Adam optimiser. For the loss function, the MSE is added to the Kullback-Leibler divergence. An SVM using an RBF kernel with $C = 1$ and $\gamma = \frac{1}{M}$ (where $M$ is the number of features) gives the optimal results.

### B. Experiments on Imaging Data

To test the proposed solutions ability to work on different data modalities, we run the same set of experiments on MIMIC-CXR data. CNNs are shown to achieve superior performance when compared to other model structures [35], hence the use of convolutions in SHAP-Conv allows the model to work well on imaging data. This is highlighted by the fact that they outperform all other methods on all medical imaging experiments carried out with a 100% accuracy on both attack types (Table II). Class imbalances in the dataset

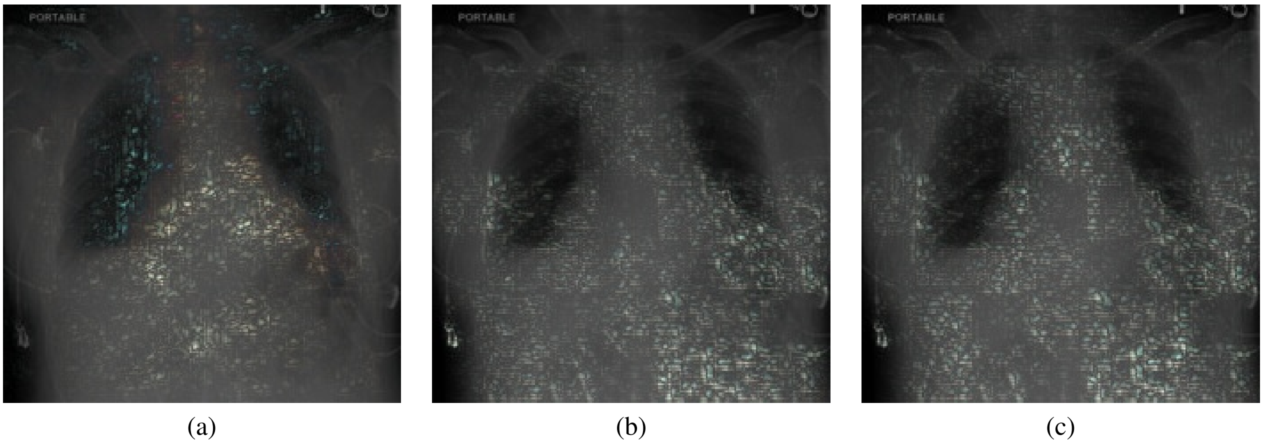

Fig. 3. (a) The heatmap of SHAP values overlayed on a genuine sample from the MIMIC-CXR dataset, (b) The heatmap of SHAP values overlayed on the same image after being perturbed via PGD, (c) The heatmap of SHAP values overlayed on the same image after being perturbed by C&W.

TABLE II
RESULTS OF ADVERSARIAL SAMPLE DETECTION. HR COLUMN REPORTS THE ACCURACY ON THE HENAN-RENMIN. CXR (C&W) REPORTS THE ACCURACY ON C&W GENERATED SAMPLES, HAVING BEEN TRAINED ON C&W SAMPLES AND CXR (PGD) THE ACCURACY OF A MODEL TRAINED ON PGD SAMPLES TESTED ON PGD SAMPLES.

| **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 [22] | 67% | 67% | 84% | 83% | 72% | 66% |
| ML-LOO [7] | N/A | N/A | 71% | 78% | 71% | 71% |

do not affect our results as our adversarial attack detectors work on balanced classes (non-perturbed images and perturbed images), and we have chosen to focus on the Cardiomegaly label within MIMIC-CXR as it itself provides a balance of positive/negative classes.

To test the semi-supervised models' ability to generalise to different attack types, we test the models trained on the MIMIC-CXR PGD data on MIMIC-CXR data perturbed by the C&W attack and vice versa. Table II shows that both SHAP-AE and SHAP-VAE are able to generalise to different attack types, achieving identical accuracy when C&W-perturbed examples are added to the test set, confirming that our model can generalise to different attack methods without the need for retraining. This is extremely useful, as it means our model is able to detect unseen attacks. However, as SHAP-MLP and SHAP-Conv are both fully-supervised and are trained on both the genuine and adversarial samples, they are unable to generalise to different attack types. Interestingly, while neither model are able to generalise, SHAP-Conv performs better when trained on PGD images whereas SHAP-MLP achieves a better performance when trained on the C&W samples. This could indicate that PGD perturbs images in such a way that higher-level features are affected (which will be more difficult for SHAP-MLP to detect), whereas C&W changes features on a lower level which SHAP-MLP has more success in recognising.

The ability of SHAP-AE and SHAP-VAE (both with SVMs) to generalise to different adversarial attack techniques is further demonstrated through Fig. 4; both of these techniques have a significantly smaller inter-quartile range than the other techniques tested, showing that the performance of these models is not affected by the type of attack that they are attempting to detect. SHAP-VAE is the clear best performer on CXR data with a stable high performance in all settings.

### C. Comparison to existing methods

The adversarial sample detection method outlined in [7] is used to run the kernel density based adversarial detection method presented in [22] on the MIMIC-CXR and MIMIC-III datasets. We estimate the kernel density of the final hidden layer of Densenet-121 and RETAIN respectively, performing grid-based cross validation search to find the optimal bandwidths, and fitting a logistic regression classifier on the estimated densities to detect adversarial samples. A bandwidth of 0.1 produces optimal results; the results are reported in Table II. This method is unable to generalise to different attack types without retraining, as the accuracy drops to $66\%$ when C&W attacks are introduced into the test dataset.

We also compare our methods against the state-of-the-art explainability-based adversarial detection method ML-LOO [24]. We follow the experiments of the authors on Densenet-121, extracting the LOO features from the same layers and utilising the inter-quartile range of these feature attribution

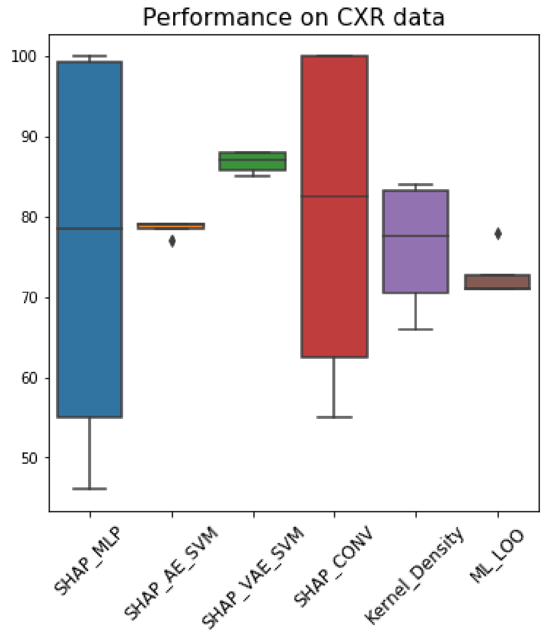

Fig. 4. Box plot reporting the performance of adversarial sample detection methods on CXR data.

maps. We test ML-LOO's ability to generalise in the same way as SHAP-AE and SHAP-VAE. ML-LOO is able to maintain comparable accuracy on the unseen attack type with a $> 10\%$ lower detection accuracy compared to SHAP-VAE. The Leave-One-Out (LOO) feature attribution method is also extremely computationally intensive, and is impractical for datasets with large feature spaces. Our method, however, does not suffer from the same issue as we are able utilise one of many possible approximations when calculating SHAP values (for example, throughout this paper the GradientSHAP approximation [26] is used).

Our proposed methods outperform the state of the art on all data modalities, as reported in Table II. Additionally, SHAP-AE and SHAP-VAE are both able to generalise to different attack types without retraining. In contrast, Kernel Density suffers a significant drop in accuracy when tested on unseen attack types in the test set, showing it is unable to accurately classify attacks it has not been trained on, while ML-LOO maintains it is performance but at a significant computational cost. Our results are compatible with those of [6], [7] in terms of EHR being a more difficult data to address with SHAP-MLP beating Kernel Density's performance by over $10\%$ in accuracy.

## V. Discussion

The presented results demonstrate the difficulty to detect adversarial attacks on EHR data. This is due to both the challenges associated with the data, and how LAVA generates adversarial samples; unlike the PGD and C&W attacks on medical imaging data, LAVA is a saliency-based attack method. This results in smaller changes being made to the SHAP values of adversarial samples, and so they are naturally more difficult to detect.

The MIMIC-CXR data is easier to work with. However, through inspection of the distribution of original labels of the adversarial examples that our model fails to detect, we find that for all labels apart from Cardiomegaly (the label our model is trying to predict) the distribution of positive/negative labels is the same as in the original dataset. However, upon investigation of the distribution of Cardiomegaly labels, we find that our semi-supervised adversarial detection methods incorrectly classifies a higher proportion of positive samples as adversarial than negative samples ($40\%$ of the incorrectly classified samples are CXRs with the Cardiomegaly diagnosis, whereas in the dataset only $29\%$ of images have the label). This shows that class imbalance in the dataset leads to difficult-to-detect adversarial samples. As the original model will most likely have an inherent difficulty to classify one of the classes (due to the class imbalance in the training data), the adversarial sample classifier needs to learn to classify *both* perturbed samples and misclassified-genuine samples as adversarial. As the SHAP values of misclassified-genuine samples will be much closer to that of the genuine training set, this is difficult to do.

The ability of all the proposed models to work on different datatsets is useful in medical scenarios where multi-modal data [36] and non-standardised data formats [6] are common. Additionally, the ability to detect adversarial samples from unseen adversarial attacks is invaluable, as it reduces the need for bespoke detection techniques to be developed when new attack methods are discovered.

## VI. Conclusion

We present a novel method of detecting adversarial samples using SHAP values that is able to adapt to different attack types and data modalities. Our method is the first such method designed specifically to work on both EHR and medical imaging data, despite the challenges of high-dimensionality, sparsity and temporality that it presents, and as such beats the current state of the art adversarial attack detection techniques on these data modalities. It is also able to generalise to different attack methods without any additional training. By using SHAP values we are able to explain how different attack methods work on different datasets, and use this information to detect samples which have been adversarially perturbed.

Further work will investigate the possibility of modifying current attack methods such as PGD and C&W to minimise the perturbation of SHAP values rather than features, and explore the effectiveness of such an attack against our detection methods. Additionally, it will explore how explainability, and SHAP in particular, can be used to inspect the distributions of different types of generated data (for example, synthetic data) and utilise these findings to evaluate the usefulness of such data.

## VII. Acknowledgements

The authors would like to thank Cievert Ltd and the European Regional Development Fund for sponsoring this work.

## References

[1] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. F. Stewart, "RETAIN: an interpretable predictive model for healthcare using reverse time attention mechanism," in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 2016, pp. 3504–3512.

[2] N. Wu, J. Phang, J. Park, Y. Shen, Z. Huang, M. Zorin, S. Jastrzębski, T. Févry, J. Katsnelson, E. Kim, S. Wolfson, U. Parikh, S. Gaddam, L. L. Y. Lin, K. Ho, J. D. Weinstein, B. Reig, Y. Gao, H. T. K. Pysarenko, A. Lewin, J. Lee, K. Airola, E. Mema, S. Chung, E. Hwang, N. Samreen, S. G. Kim, L. Heacock, L. Moy, K. Cho, and K. J. Geras, "Deep neural networks improve radiologists' performance in breast cancer screening," *IEEE Transactions on Medical Imaging*, pp. 1–1, 2019.

[3] A. Majkowska, S. Mittal, D. F. Steiner, J. J. Reicher, S. M. McKinney, G. E. Duggan, K. Eswaran, P.-H. Cameron Chen, Y. Liu, S. R. Kalidindi, A. Ding, G. S. Corrado, D. Tse, and S. Shetty, "Chest radiograph interpretation with deep learning models: Assessment with radiologist-adjudicated reference standards and population-adjusted evaluation," *Radiology*, vol. 294, no. 2, pp. 421–431, 2020, pMID: 31793848.

[4] 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.

[5] 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.

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

[7] 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.

[8] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Medicine*, vol. 17, no. 1, p. 195, Oct 2019.

[9] P. E. Kalb, "Health Care Fraud and Abuse," *JAMA*, vol. 282, no. 12, pp. 1163–1168, 09 1999.

[10] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, no. 1, pp. 44–56, Jan 2019.

[11] V. Prasad and S. Mailankody, "Research and Development Spending to Bring a Single Cancer Drug to Market and Revenues After Approval," *JAMA Intern Med*, vol. 177, no. 11, pp. 1569–1575, 11 2017.

[12] "Using imaging biomarkers to accelerate drug development and clinical trials," *Drug Discovery Today*, vol. 10, no. 4, pp. 259 – 266, 2005.

[13] J. Lu, H. Sibai, E. Fabry, and D. A. Forsyth, "NO need to worry about adversarial examples in object detection in autonomous vehicles," *CoRR*, vol. abs/1707.03501, 2017.

[14] J. Morley and I. Joshi, "Artificial intelligence: How to get it right. putting policy into practice for safe data-driven innovation in health and care." *NHS*, 2019.

[15] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv: Machine Learning*, 2017.

[16] F. K. Došilović, M. Brčić, and N. Hlupić, "Explainable artificial intelligence: A survey," in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, May 2018, pp. 0210–0215.

[17] S. Qiu, Q. Liu, S. Zhou, and C. Wu, "Review of artificial intelligence adversarial attack and defense technologies," *Applied Sciences*, vol. 9, no. 5, p. 909, Mar 2019.

[18] 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.

[19] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," *CoRR*, vol. abs/1712.09665, 2017. [Online]. Available: http://arxiv.org/abs/1712.09665

[20] 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.

[21] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff, "On detecting adversarial perturbations," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

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

[23] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., 2018, pp. 7167–7177.

[24] P. Yang, J. Chen, C. Hsieh, J. Wang, and M. I. Jordan, "ML-LOO: detecting adversarial examples with feature attribution," *CoRR*, vol. abs/1906.03499, 2019.

[25] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, Jul 2019.

[26] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 4765–4774.

[27] S. Lipovetsky and M. Conklin, "Analysis of regression in game theory approach," *Applied Stochastic Models in Business and Industry*, vol. 17, no. 4, pp. 319–330, 2001.

[28] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, no. 1, p. 160035, May 2016.

[29] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific Data*, vol. 6, no. 1, p. 317, Dec. 2019.

[30] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. L. Ball, K. S. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," *CoRR*, vol. abs/1901.07031, 2019.

[31] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017.* IEEE Computer Society, 2017, pp. 2261–2269.

[32] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[33] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Y. Ding, A. Bagul, C. Langlotz, K. S. Shpanskaya, M. P. Lungren, and A. Y. Ng, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *CoRR*, vol. abs/1711.05225, 2017.

[34] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2014.

[35] N. Sharma, V. Jain, and A. Mishra, "An analysis of convolutional neural networks for image classification," *Procedia Computer Science*, vol. 132, pp. 377 – 384, 2018, international Conference on Computational Intelligence and Data Science.

[36] K. Xu, M. Lam, J. Pang, X. Gao, C. Band, P. Mathur, F. Papay, A. K. Khanna, J. B. Cywinski, K. Maheshwari, P. Xie, and E. P. Xing, "Multimodal machine learning for automated icd coding," in *Proceedings of the 4th Machine Learning for Healthcare Conference*, ser. Proceedings of Machine Learning Research, F. Doshi-Velez, J. Fackler, K. Jung, D. Kale, R. Ranganath, B. Wallace, and J. Wiens, Eds., vol. 106. Ann Arbor, Michigan: PMLR, 09–10 Aug 2019, pp. 197–215.