Interpretation Attacks and Defenses on Predictive Models using Electronic Health Records *

Fereshteh Razmi[1], Jian Lou[2], Yuan Hong[3], and Li Xiong[1] ✉

[1] Emory University, Atlanta GA 30322, USA
[2] Zhejiang University, Hangzhou, Zhejiang 310027, China
[3] University of Connecticut, Storrs CT 06269, USA
{frazmim,lxiong}@emory.edu,
jian.lou@zju.edu.cn, yuan.hong@uconn.edu

Abstract. The emergence of complex deep neural networks made it crucial to employ interpretation methods for gaining insight into the rationale behind model predictions. However, recent studies have revealed attacks on these interpretations, which aim to deceive users and subvert the trustworthiness of the models. It is especially critical in medical systems, where interpretations are essential in explaining outcomes. This paper presents the first interpretation attack on predictive models using sequential electronic health records (EHRs). Prior attempts in image interpretation mainly utilized gradient-based methods, yet our research shows that our attack can attain significant success on EHR interpretations that do not rely on model gradients. We introduce metrics compatible with EHR data to evaluate the attack's success. Moreover, our findings demonstrate that detection methods that have successfully identified conventional adversarial examples are ineffective against our attack. We then propose a defense method utilizing auto-encoders to denoise the data and improve the interpretations' robustness. Our results indicate that this de-noising method outperforms the widely used defense method, SmoothGrad, which is based on adding noise to the data.

Keywords: Interpretation Models · Electornic Health Records (EHR) · Adversarial Attack · Robustness · Autoencoder

1 Introduction

Machine learning algorithms, particularly deep neural networks, are widely used in various real-world tasks. However, their inner workings are often seen as a black box. Thus, interpretation methods are essential for explaining an algorithm's output, allowing users to understand how and why an algorithm arrived at a particular decision. Especially in sensitive applications such as medicine,

* This work was funded by National Science Foundation (NSF) IIS-2302968, CNS-2124104, CNS-2302689 and CNS-2308730, National Institute of Health (NIH) R01ES033241, R01LM013712, and NSFC (62206207).

interpretations improve the system's reliability and enable the discovery of new biomarkers and important features for future decision-making processes. For instance, Quellec et al. [15] use heatmaps to identify local patterns and demonstrate which pixels in retinal fundus photographs are involved in the early signs of retinal disease.

Adversarial examples [23] have been extensively studied in recent years as a potential vulnerability of deep neural networks. Traditionally, they aim to add a small perturbation to the input at inference time, causing the model to classify it differently. With the increasing use of interpretation methods, a new type of attack has emerged. These attacks focus on generating misleading interpretations that deviate significantly from the true classifier interpretations, leading to inaccurate conclusions about the importance of certain features or rendering the interpretations unreliable [8].

Sequential electronic health records (EHR) are crucial data sources in the medical field, containing discrete data of patients' vital values and lab values collected over time and across hospital visits. Due to the importance of these data and their use in many classification based predictive models, recent efforts have been made to enhance the interpretability of models trained on EHR data. Despite the prevalence of interpretation attacks in image classification, to the best of our knowledge, no interpretation attacks have been studied targeting EHR-based models.

Conducting interpretation attacks on EHR data presents significant challenges due to the unique characteristics of the data. Firstly, for building interpretable models using EHR data, models are designed to produce predictions and interpretations simultaneously. In contrast, image interpretations are mostly gradient-based and created via post-hoc approaches. Thus, manipulating the EHR interpretations can easily alter the patient phenotype, consequently affecting the predicted class.

Secondly, the structure of EHR data is vastly different from images. As a result, the widely used L_∞ norm based attacks in image domain are less meaningful in the EHR domain since L_∞ does not capture the distance between the sequential data well (e.g., the temporal trends). Also, unlike images, EHR data consist of multiple attributes, such as heart rate or temperature, whose values are sequential and time-dependent. Therefore, moving across time and attributes significantly influences the interpretations. Consequently, the criteria used for assessing the image interpretation's robustness on previous works cannot be directly applied in the EHR domain.

This work proposes an interpretation attack on EHR data, utilizing specific metrics suitable for this data type. We evaluate our attack against a powerful existing detection technique designed for conventional adversarial examples on EHR data and demonstrate that the attack is not detectable. Furthermore, we aim to make the EHR interpretations robust against the proposed attack. We show that using an auto-encoder to de-noise the input is significantly more effective than using noisy input, as in the state-of-the-art method SmoothGrad.

The source code of our implementation is publicly available on GitHub[4]. We summarize our contributions as follows:

- We propose an interpretation attack on EHR data. This attack is created on top of an interpretable model, so the interpretations are closely tied to the model's predictions. It differs from previous attacks in the image domain, which rely on gradient-based and post-hoc interpretation methods.
- We propose three metrics to assess the EHR interpretation attack. In the previous works, top-K salient explanations between the clean and adversarial images were used for evaluation. However, it is not suitable for EHR data. Two of our evaluation metrics are alternatives to the top-K criteria, and the third metric is based on the Wasserstein distance which better captures the similarity between temporal data.
- We conduct experiments showing that the state-of-the-art detector RADAR, which was designed to detect conventional EHR adversarial examples, are not successful in detecting the proposed attack. We then explore the factors that contribute to this attack evasion.
- Finally, we present a method to enhance the interpretations' robustness and reduce the attack strength. We employ an auto-encoder to boost the robustness of our interpretations through a de-noising process. We show that out approach outperforms SmoothGrad, which is commonly used in gradient-based methods by averaging noisy data.

2 Related Work and Preliminaries

2.1 Attacks on Image Model's Gradients

Post-hoc interpretability are a set of interpretation methods that seek to explain the predictions of models without relying on their underlying mechanisms [11]. Gradient-based approaches are commonly used in image classification to extract these explanations [17,19,20]. They result in a saliency map that explains the output of the model (usually a convolutional neural network (CNN)) by visualizing the areas of the input image that contribute the most to the network's output. However, saliency maps are less common in Recurrent Neural Networks (RNN) since RNNs are typically used for sequential data such as time-series.

Recent research has shown that these methods are vulnerable to interpretation attacks, where small perturbations are deliberately crafted and added to input images to distort the explanations [8]. These attacks primarily focus on images as they rely on gradient-based techniques and face significant challenges in other domains. Several techniques have been proposed to address this issue, including adding randomness to the input called SmoothGrad [21,26], modification of the model architecture [6], or altering the training process using regularization or integrated gradients [3,7]. These approaches are highly dependent on the architecture of image models and their gradients. Interpretation attacks

[4] https://github.com/Emory-AIMS/EHR-Interpretation-Attack

in other domains including EHR have been relatively overlooked due to the difficulty in attacking against complex saliency maps and the lack of a definitive interpretation benchmark.

2.2 Medical Attention-based Models

Recent research in the medical field has focused on using the attention mechanism to improve the interpretability and accuracy of predictions made using EHR data [4,5]. The attention mechanism is an approach used in machine learning models that assigns a weight to each input feature, indicating its relative importance to the model's final decision. They generally use BERT models [10,16,18] or multi-layer RNNs [9,12,13,25] as the baseline to obtain the attentions. BERT models are mostly focused on binary medical codes and their pre-trained models are often not publicly available due to the sensitive nature of the medical data used for their training. In this work, we use RETAIN [5] as a well-known EHR attention-based RNN model. RETAIN can give interpretation on both visit (temporal point) and attribute levels, and in contrast to other works, it does not need access to extra meta data [9]. We then propose interpretation attacks considering the structure of EHR data and also the intrinsic nature of their non-post-hoc interpretable models.

3 Our Approach

In this section, we first describe the problem setting, then present our approach to the interpretation attack on EHR models and elaborate the rationale behind each objective loss term. We then improve the attack by incorporating dynamic weighing to penalize the attack optimization process and reduce the detectability by modifying the penalty term. We propose new metrics as the current evaluation metrics are unsuitable for EHR data. Finally, we explore methods for defending against the attack and demonstrate that de-noising is more effective than the state-of-the-art method for improving the robustness of interpretations.

3.1 Problem Setting

EHR dataset is a set of clinical trajectories for patients where each trajectory is a sequence of hospital or clinic visits, each visit corresponding to a set of attributes/measurements [1]. For a given dataset with longitudinal EHR data from N patients, we represent the clinical trajectory of patient n as $X^{(n)}$. This trajectory is characterized by a sequence of t_n hospital visits and can be expressed as:

$$X^{(n)} = [X_1, X_2, ..., X_{t_n}], \tag{1}$$

where $X_i \in R^d$ denotes the variables from d vital sign measurements and lab events of the *i-th* visit made by patient n. Each $x_{i,j}$ shows *j-th* attribute in

the *i-th* visit. We will exclude the superscript (n) in the subsequent sections to simplify the presentation.

Given a neural network model $f : R^{(t,d)} \rightarrow R^c$ where c is the number of possible classes, we denote the interpretation that is associated with the parameters of function f as $\Phi_f : R^{(t,d)} \rightarrow R^{(t,d)}$ in which every attribute in a specific visit gets a score that shows its importance on the predicted outcome. Given a test input X, the class and explanations of this input is determined by $c^* = \arg\max_c f(X)$ and $\omega = \Phi_f(X)$, respectively. In RETAIN [5], the impact of each input $x_{i,k}$ on the final classification result is calculated using the two-level attention weights:

$$\omega_{i,k} = \alpha_i W(\beta \odot W_{emb}[:,k])\ x_{i,k}, \tag{2}$$

where α_i is the attention weight assigned to the *i-th* visit, β_i is an attention weight vector for all attributes and measurements $x_{i,k}$ of the *i-th* visit, W is the output weight matrix, W_{emb} is the weight matrix at the embedding layer, and the symbol $\odot$ represents element-wise multiplication. $\omega_{i,k}$ is the corresponding contribution to the input $x_{i,k}$. Therefore, we can obtain the contribution matrix ω using all $\omega_{i,k}$.

3.2 Interpretation Attack Formulation

Given a patient record X, the goal is to find a new perturbed record $\widetilde{X}$ that is similar to the original record X both in input space and class predictions but with distorted interpretations. The attack can either be targeted, where we try to make the interpretations of $\widetilde{X}$ closer to a new explanation $\omega^\dagger$, or untargeted, where we attempt to change the interpretations to be far from those of X. Here we aim for a targeted one and formulate the interpretation adversarial attack by

$$\min_{\widetilde{X}} \alpha\|\Phi_f(\widetilde{X}) - \omega^\dagger\| + \gamma\|\widetilde{X} - X\|_1 + \beta(\max\{Logit(\widetilde{X})_i : i \neq c^*\} - Logit(\widetilde{X})_{c^*})^+ \tag{3}$$

where $(r)^+$ represents $max(r, 0)$, c^* is the predicted class of X, *Logit* is the outcome of the neural network before the Softmax layer and $\widetilde{X}$ is the adversarial example resulting in misleading interpretations. α, β and γ are the coefficients to balance the impact of the loss function terms. We will discuss each term one by one:

1. Interpretation Loss: The first term ensures that the interpretations of $\widetilde{X}$ resemble the targeted interpretation $\omega^\dagger$. This attack can be reformulated as an untargeted attack by replacing the current term with $-\|\Phi_f(\widetilde{X}) - \Phi_f(X)\|$. In the case of the targeted attack, $\omega^\dagger$ can come from another set of interpretations with different but still realistic phenotypes, such as the interpretations of a randomly-selected patient, or patients' average interpretations of a different class than the X's class c^*. Since this leads to a more realistic scenario we proceed with targeted attacks.

2. Perturbation Loss: The second term aims to keep the adversarial perturbations small. We optimize the perturbations using L_1 norm rather than widely used L_2-norm or L_∞-norm for images. L_1 norm for adversarial attacks on EHR data are more meaningful for several reasons. First, EHR data are sparse, where

Algorithm 1: Interpretation Attack on EHR

Function : MINIMIZE-ATTACK-LOSS(.) : returns X and the corresponding Y by minimizing Eq. 3

Input : initial clean sample (X_{clean}, Y_{clean}), initial coefficients $(\alpha_{init}, \beta_{init})$ in Eq. 3, number of iterations T, the maximum possible β value $\beta_{treshold}$ and the number of extra steps for penalizing $steps_{extra}$

Initialize : α, $\beta = \alpha_{init}$, β_{init}; X_0, $Y_0 = X_{clean}$, Y_{clean}

for $t \in \{1, ..., T\}$ **do**
  X_t, Y_t = MINIMIZE-ATTACK-LOSS(X_{t-1},α,β)
  if $Y_t \neq Y_{clean}$ **then** `// Dynamically penalize the optimization`
    while $Y_t \neq Y_{clean}$ **do**
      α, $\beta = \alpha/2$, $\beta \times 2$
      X_t, Y_t = MINIMIZE-ATTACK-LOSS(X_t,α,β)
      if $\beta > \beta_{threshold}$ **then return** Attack-failure;
    end
    for $s_e \in \{1, ..., steps_{extra}\}$ **do**
      X_t, Y_t = MINIMIZE-ATTACK-LOSS(X_t,α,β)
    end
    α, $\beta = \alpha_{init}$, β_{init}
  end
end
Return X_i from $\{X_1, ..., X_T\}$ with $Y_i = Y_{clean}$ and its interpretations have the least distance to the target interpretations (i.e. min $\|\Phi_f(X_i) - \omega^\dagger\|$)

many of the values are either zero or imputed and hence do not carry much information. Second, unlike images, different medical attributes carry different influences and weights on the output. Consequently, L_1 norm is suitable to meet both sparsity and heterogeneity of the EHR data [1,22].

3. Classification Loss: The third term aims to keep the class prediction unchanged. Our interpretation method is non-post-hoc, so the predictions are highly tied to the interpretations. Thus we need a more powerful function to keep the class of $\widetilde{X}$ unchanged. We employ the logits based function for this purpose since it can be well optimized for manipulating the class predictions, especially for non-linear objective $f(\widetilde{x}) = c^*$ [2]. We will show in Sec. 3.4 that it can be improved so that the output space $Logit(\widetilde{X})$ resembles $Logit(X)$ and hence helps the adversarial example remain undetectable.

3.3 Optimization with Dynamic Penalty

Equation 2 denotes how the parameters of the model, including weights and attributions, are directly involved in the explanations of the input. We observed that in some cases, the objective to change in interpretations might lead to a different class label. Given that the interpretation attack is conducted using a gradient descent algorithm, we use dynamic penalty for the interpretation and classification loss terms for preventing the prediction change.

Concretely, it involves adjusting the coefficient in Equation 3 to prioritize the objective of keeping the prediction label unchanged, i.e., incur a higher penalty whenever encountering a label change in any iteration. We can achieve this by decreasing α and increasing β by a factor (e.g., the factor is set to 2 in our implementation) until the original class label is attained. We can then continue using these coefficients for a few more steps to move away from the classification boundaries. If this penalization process continues without successfully restoring the original class, the algorithm is considered to have failed. Algorithm 1 outlines the different components of the attack.

3.4 Minimizing Detectability

To carry out a stealthy attack, two aspects must be considered. The first is to keep the perturbations in the input space minimum, while the second is to maintain the integrity of the output space which includes the final class predictions and their associated logits. The reason is that many state-of-the-art defense methods for adversarial examples check changes both in the input feature space and the output logits space [14,24]. So in order to minimize the detectability, it is necessary to ensure that the logits do not change drastically during the attack. We observed that as we repeatedly apply and remove the penalty according to algorithm 1, it causes the output space of the adversarial example to oscillate near the classification boundaries. Consequently, while the final label is the same as the original class, the logits do not resemble the original logits, nor does the confidence level of the adversarial prediction. This difference in logits, which we will refer to as output space, can be used to detect the attack.

To address this issue, we propose enhancing (3) by replacing the classification loss with two different alternatives. First we use the Kullback-Leibler divergence to directly compare the distribution of the original sample and adversarial example logits in order to keep them similar. We denote this divergence by $KL\,(\,Logit(X)\,||\,Logit(\widetilde{X})\,)$ (KL **attack**). Second, similar to the idea of C&W conventional adversarial attacks [2], we use $\max\,(\,\max\{Logit(\widetilde{X})_i : i \neq c^*\} - Logit(\widetilde{X})_{c^*}, -\kappa\,)$ where κ is a positive adjustable value and maintains a margin between the predicted logit and the second largest logit value to ensure high confidence in the predicted class (**Confident attack** parameterized by κ). Since the classifier is trained based on the clean examples' manifold, it can classify them with high confidence. So by ensuring high confidence predictions for the adversarial examples, we can keep their logits similar to their original counterparts.

3.5 Metrics for Evaluation

For conventional adversarial examples, attack success rate (ASR) is measured as percentage of examples with flipped class labels. However, interpretation attacks aim to alter the multi-dimensional interpretation vector, making it difficult to establish a clear binary metric for measuring the success of the attack. In the

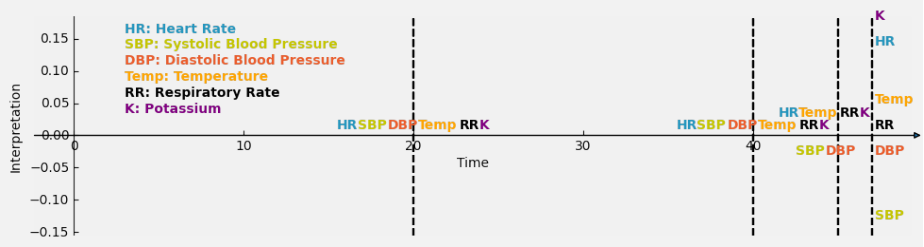


Fig. 1. Interpretations of a patient's EHR data for six attributes *(RR, HR, K, SBP, DBP, Temp)* with heart failure at the final time-stamp. Interpretations of different attributes can be compared with each other in each specific time stamp. Also each attribute separately can be explored for its changes across time. The interpretations for EHR data generally gain more importance as the time of disease onset approaches.

subsequent discussion, we will outline two particular aspects of EHR data that must be taken into account when defining evaluation metrics.

In many application of EHR data, the interpretations may carry either positive or negative connotations, each with its unique significance. For example, when predicting the likelihood of a specific disease, the use of a particular medication may negatively affect the prognosis and decrease the chance of disease onset. For a clinician, the classifier's explanation of such a drug is no less important than the factors that indicate positive interpretations towards the prediction.

Another characteristic of EHR data is the heterogeneity and time sensitivity. Unlike pixels in images, the diverse attributes in EHR data hold distinct meanings, and clinician's interpretation may differ for each attribute. Additionally, the value of interpretations for clinicians is affected by the timing of attribute collection. Clinicians attach more significance to the data points that are closer to the disease onset. Fig. 1 displays interpretations of some attributes calculated by RETAIN for predicting heart failure in a patient. Given these factors, we propose three metrics to evaluate the sucess of the interpretation attack, which consider the connotations of the interpretations, the attribute-level heterogeneity, and the visit-level time awareness.

Signed top-K intersection size: According to Ghorbani et al. [8], in many cases, when interpreting a model, the explanations of the most important features are often of interest. In a gradient-based saliency map, the top-K features are determined by their magnitudes. Here we involve the connotation of the interpretations and assess the success of the attack by comparing the proportion of top-K features with consistent signs before and after the attack. So if $A = \{a_1, ..., a_k\}$ and $B = \{b_1, ..., b_k\}$ are the sets of the K largest absolute-value dimensions of $\Phi(\widetilde{X})$ and $\Phi(X)$ respectively, and $C = A \cap B$, then we have

$$topK(C) = |\{c_i \in C : \Phi(\widetilde{X})_{c_i} * \Phi(X)_{c_i} > 0\}|. \tag{4}$$

Asymmetrical signed top-K intersection size: Since the EHR is sequential and time-sensitive, the importance of different attributes are comparable in each timestamp that they are collected. To reflect that, we suggest a new metric that

measures the top-K salient features in corresponding multivariate time series at each time point and then aggregate them.

Also, we assign weight ϕ_i to each time to better attain the perspectives of clinicians who may place greater emphasis on certain times. These weights can be achieved by background knowledge (e.g., higher weight on certain time points before the disease onset) or approximated by how the interpretable model weight different times, e.g., by taking 100 random samples from the clean data and summing up their interpretation values of all attributes at any given time. The resulting values are averaged over all samples to derive the weight that should be assigned to that specific time. For time $t_i \in \{1, \ldots, t\}$, we denote $A_{t_i} = \{a_j^{t_i}\}_{j=1}^k$ and $B_{t_i} = \{b_j^{t_i}\}_{j=1}^k$ as the sets of the K largest absolute-value dimensions of $\Phi(\widetilde{X}_{t_i})$ and $\Phi(X_{t_i})$, respectively, and their intersectino as $C_{t_i} = A_{t_i} \cap B_{t_i}$.

$$topK\_asym = \sum_{i=1}^{t} \phi_i * topK(C_i). \tag{5}$$

Wasserstein distance: The Wasserstein distance measures the cost of moving a variable mass and is well-suited for comparing changes in time series. Its ability to capture perturbations has made it increasingly popular in the context of adversarial examples. We use the Wasserstein distance to measure the changes of contribution by each attribute as time series - since the modality of data is different across different attributes as discussed before. The resulting distances are then summed to obtain the final Wasserstein distance. Given attribute index $d_j \in [d]$, we denote $X_{[t]}^{d_j}$ as the sequential values of a specific attribute, and $Wass$ as the Wasserstein distance. Then, we calculate the final distance as:

$$Wass\_dist = \sum_{j=1}^{d} W_1(\Phi(\widetilde{X}_{[t]}^j), \Phi(X_{[t]}^j)). \tag{6}$$

where W_1 denotes 1-Wasserstein distance for one dimensional data.

To make equations 4, 5 and 6 consistent with our targeted attack, we calculate these relative metrics:

$$topK^{targeted} = topK(\Phi(\widetilde{X}_i), \omega_i^\dagger)/topK(\Phi(\widetilde{X}_i), \Phi(X_i)); \tag{7}$$

$$topK\_asym^{targeted} = topK\_asym(\Phi(\widetilde{X}_i), \omega_i^\dagger)/topK\_asym(\Phi(\widetilde{X}_i), \Phi(X_i)); \tag{8}$$

$$Wass\_dist^{targeted} = Wass\_dist(\Phi(\widetilde{X}_i), \omega_i^\dagger)/Wass\_dist(\Phi(\widetilde{X}_i), \Phi(X_i)). \tag{9}$$

These three new metrics not only measure how the adversarial interpretations are distant from the original ones, but also reflect how they resemble the target interpretations $\omega^\dagger$. The attacks with larger $topK^{targeted}$ and $topK\_asym^{targeted}$, and with smaller $Wass\_dist^{targeted}$ are more powerful. From now on, when we mention these metrics, we are specifically referring to their targeted version.

3.6 Robustness

To provide robustness, we propose using a sequential auto-encoder to de-noise the input data at inference time and recover the original information. A typical

auto-encoder comprises an encoder that compresses the data into a smaller intermediate representation and a decoder that attempts to reconstruct the input data from those embeddings. As the encoder and decoder process the data, the output becomes de-noised. We train the auto-encoder on clean data so it learns the normal manifold. As a result, at inference time, it can remove the noise that caused the input data to become far from this manifold. We then utilize the interpretations of the decoder's output instead of those of the input. Our results show that this approach leads to robust interpretations.

There are two reasons for this. Firstly, the EHR attack perturbations are sparse and have a greater magnitude wherever the features have notable interpretations. Therefore, the de-noiser can restore the original interpretations by reducing the large sparse perturbations on the salient features. Secondly, interpretation attacks differ from traditional adversarial examples in that they aim to modify smoothly distributed, high-dimensional interpretations, especially in EHR data. Once the de-noiser eliminates sudden, sparse perturbations, the interpretations can be regained by relying on the information present in the surrounding neighborhood.

We compare our method with SmoothGrad, a known and strong defense against interpretation attacks [21]. Although the attack in our case is gradient-free, the idea of SmoothGrad is still applicable. It involves adding noise to the data multiple times (usually 10 to 50) and averaging their contributions. However, this method is neither computationally efficient nor effectively provides robustness against EHR attacks as we will show empirically.

4 Experiments

In this section, we will address these questions: 1) What is the effectiveness of the attack in altering the interpretations while maintaining the classification outcomes? 2) Can existing defense methods against adversarial examples detect the interpretation attack? 3) How does the proposed de-noiser approach help with the robustness?

Dataset. The MIMIC-III dataset is a collection of electronic health records from thousands of patients in intensive care units. We use a dataset that was processed by [22] for the binary task of mortality prediction, resulting in 3177 positive samples and 30344 negative samples, each comprising 19 attributes across 48 timestamps including vital signs and lab events. Missing features were filled using the average value across all timestamps, and outliers were removed and imputed according to interquartile range criteria. Finally, each sequence was truncated or padded to 48 hours, and each feature was normalized using min-max normalization. We use 80% of the data for training and the rest for testing.

Model Architecture and Parameters. Adversarial examples were generated against RETAIN [5] as our target model, which includes an embedding layer of size 128 and two GRU layers with 128 hidden units. The evaluation results of the test data on the final trained model are $AUROC = 0.92$, $AUPRC = 0.73$, $F1Score = 0.57$ and $Accuracy = 0.86$. We evaluate the detectability of the

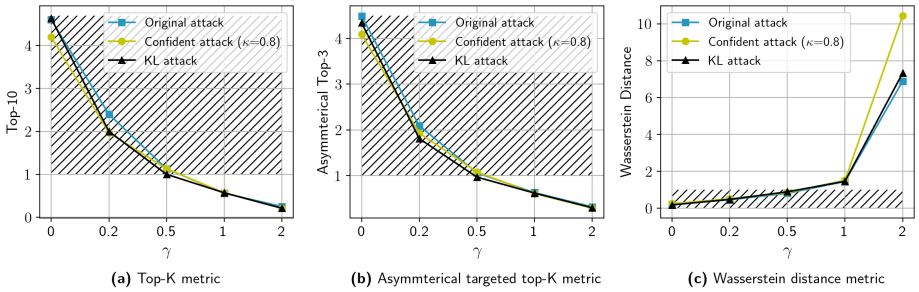


Fig. 2. The comparison of three interpretation attacks, which differ in their penalty term, shown using three metrics. The desirable results are located in the hatched area. A lower perturbation achieved by a smaller γ leads to better attack success, but may also result in a higher detection rate.

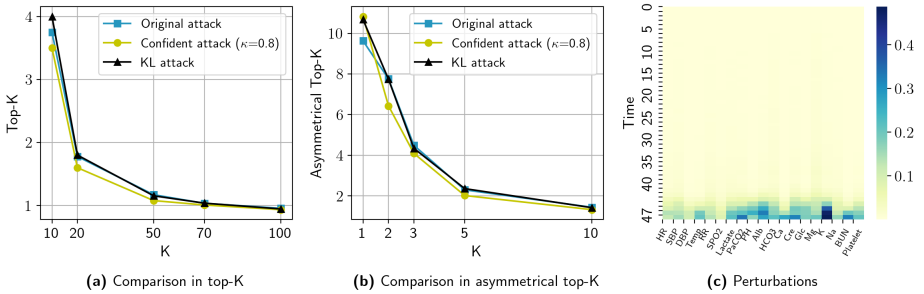
interpretation attack using RADAR [24]. It is a robust detection method, specifically developed for traditional EHR adversarial examples where the objective is to change the class. This detector identifies adversarial examples through both changes in input space and also output space relative to the normal manifold, making it well-suited for our purposes. Finally to enhance the data robustness, we de-noised data by the same auto-encoder architecture as that used in RADAR.

4.1 Attack Performance

Comparison of Attacks. We evaluate the attack performance based on three different metrics introduced in Sec. 3.5. We compare the original attack (equation 3) with two alternatives, the KL attack and the Confident attack, proposed in Sec. 3.4. In our experiments with the Confident attack, we set $\kappa = 0.8$, as it provides a high level of undetectability. Our comparison is based on different values of the coefficient γ in equation 3, which constrains the perturbation size. The higher the value of γ, the more restricted the attack is in terms of its distance from the original sample. Since the parameters α and β are dynamically adjusted by algorithm 1, we simply select their initial values as 1. Also based on a grid-search we set $T = 1000$ and $steps_{extra} = 10$.

Fig. 2 illustrates the results based on the three metrics (equation 7, 8, 9) from left to right, respectively. The hatched area in each figure demonstrates the most desirable results. For Fig. 2.a and b, a ratio of over 1 implies that the interpretations are more similar to the targeted interpretations than the original ones, and the larger the ratio, the better. Conversely, in Fig. 2.c, the opposite is true, as this measurement employs a distance metric rather than the intersection of salient features. Although the attacks are very similar, in the next section, we will show the main difference lies in the stealthiness of each of these attacks.

Selection of K. Fig. 3 demonstrates that how the selection of K in top-K metrics (7 and 8) impacts our evaluation of the attack's success when $\gamma = 0$. In



(a) Comparison in top-K **(b)** Comparison in asymmetrical top-K **(c)** Perturbations

Fig. 3. The comparison of different values of K in two metrics, top-K (a) and asymmetrical top-K (b). The concentration of perturbations on the latest time-stamps (c) confirms that small values of K are sufficient for evaluation.

metric 4 since K is calculated in each time and over a lower dimension than the entire EHR data, we set the value of K to a lower number than in metric 8. As expected, the value of K affects the degree of overlap between interpretations before and after attack. Fig. 3.c shows the average perturbations of all the adversarial examples when γ is zero and there is no constraint on the input space. The perturbations are concentrated on the latest time-stamps which hold the most significant interpretations in the model and clinical environments. Therefore, selecting a large K does not yield significant interpretations, particularly since many interpretations that are distant from these timestamps have close to zero. Consequently, considering large K results in overlapping interpretations that do not offer meaningful insights into the attack's success.

4.2 Attack Detectability

Fig. 4 illustrates an example before and after the attack and their difference for the confident attack with $\gamma = 0.5$. The attack causes sparse but strong perturbations, which lead the interpretations to shift from the original to the target interpretations. As previously discussed, the low number of perturbations and their sparsity make them undetectable in EHR data. By decreasing γ, the magnitude and density of the perturbations become more flexible. Fig. 5 illustrates the interpretations of the original sample and its adversarial counterpart from Fig. 4 as well as the target interpretations across the latest timestamps. Due to space limitations, only three attributes are included in the figure. It reveals that the sparse perturbation attack caused the adversarial interpretations to deviate from their original values and align more closely with the target interpretations.

We evaluated RADAR to demonstrate whether our proposed interpretation attacks can be detected by existing defense methods against conventional adversarial examples. RADAR exhibits a 100% detection rate for conventional adversarial examples on RETAIN. Fig. 6 presents the detection percentage of different interpretation attacks by RADAR which are significantly lower. As γ increases,

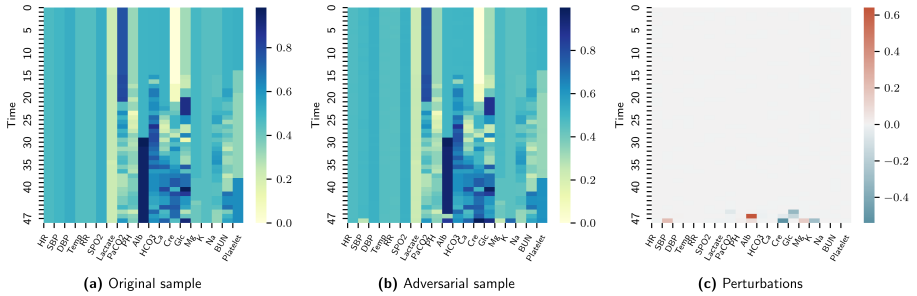


Fig. 4. An input space of a patient's EHRs before (a), and after attack with γ=5 (b), and its additive perturbation (c). The perturbation is minimal and sparse.

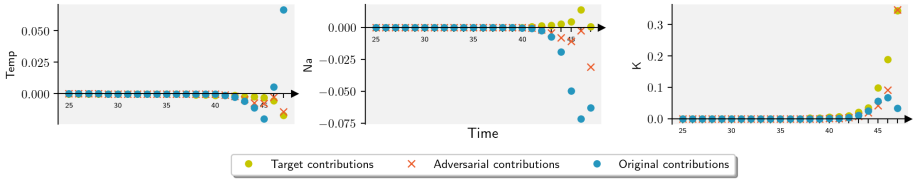


Fig. 5. Comparison of Adversarial, original and target contributions (interpretations) of three attributes of a patient's EHR over time.

the perturbations become smaller, resulting in a decrease in the detection rate in input space. Additionally, considering the detection in output space,, when γ is small, the adversarial example has more flexibility during optimization, allowing it to approach the classification decision boundary more closely and activate the penalty process in algorithm 1 more frequently. In Sec. 3.4, we discussed how KL and confident attacks better maintain similarity between the original and output space in such cases. However, for larger values of γ, the original attack is less likely to trigger the penalty process and remains more stealthy than the KL attack. Generally, the confident attack keeps the output space less detectable and maintains a greater distance from the class boundary.

4.3 Robustness

In this part, we evaluate the effectiveness of the proposed auto-encoder (AE) denoiser based defense method. We report the attack success rate of the attack under the proposed defense method, and compare it with the attack without defense, and the attack with the SmoothGrad defense. We select the confident attack with $\kappa = 0.8$ and $\gamma = 0.5$ as the representative of successful attacks with reasonably high success rate and low detection rate. For SmoothGrad, the best results are reported based on selecting a noise level of 0.1 and calculating the average over 50 samples, which is consistent with the result in paper [21]. Fig. 7 displays a comparison of the median and quartile charts of the attack versus

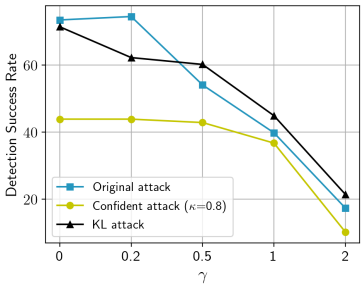


Fig. 6. Detection ratio of interpretation attacks using RADAR.

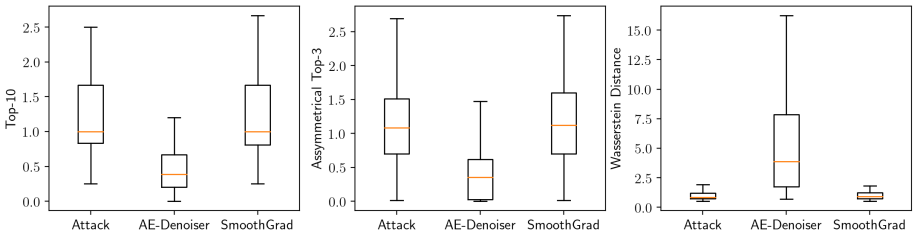


Fig. 7. Robustness of de-noising method vs. SmoothGrad based on three metrics. All figures show the de-noising method outperforms SmoothGrad.

the robustness achieved through the de-noising method and SmoothGrad for 100 samples. Smaller values for top-K and asymmetric top-K indicate better robustness, whereas higher values for Wasserstein distance indicate better robustness. As depicted, the de-noising method outperforms SmoothGrad in all metrics.

5 Conclusion

This paper is the first study to develop and adapt interpretation attacks for EHR models. We investigated various aspects of EHR data as well as interpretable models designed specifically for EHR data. We presented interpretation attacks on EHR models optimizing both attack success and detectability and evaluated the attack using customized metrics that address EHR specifications. Our results show that the attack not only can successfully alter the interpretations of the model, but also can evade the detector RADAR, which is capable of detecting 100% of conventional adversarial examples. To counteract the attack, we proposed a de-noiser defense and demonstrated that it improved the robustness and outperformed existing method SmoothGrad. Future research can focus on modifying EHR interpretable models to make them more robust, as well as exploring data preprocessing, data augmentation, and adversarial training to enhance the robustness of EHR models.

References

1. An, S., Xiao, C., Stewart, W.F., Sun, J.: Longitudinal adversarial attack on electronic health records data. In: The world wide web conference. pp. 2558–2564 (2019)
2. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 ieee symposium on security and privacy (sp). pp. 39–57. Ieee (2017)
3. Chen, J., Wu, X., Rastogi, V., Liang, Y., Jha, S.: Robust attribution regularization. Advances in Neural Information Processing Systems **32** (2019)
4. Chen, P., Dong, W., Wang, J., Lu, X., Kaymak, U., Huang, Z.: Interpretable clinical prediction via attention-based neural network. BMC Medical Informatics and Decision Making **20**(3), 1–9 (2020)
5. Choi, E., Bahadori, M.T., Sun, J., Kulas, J., Schuetz, A., Stewart, W.: Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. Advances in neural information processing systems **29** (2016)
6. Dombrowski, A.K., Alber, M., Anders, C., Ackermann, M., Müller, K.R., Kessel, P.: Explanations can be manipulated and geometry is to blame. Advances in neural information processing systems **32** (2019)
7. Dombrowski, A.K., Anders, C.J., Müller, K.R., Kessel, P.: Towards robust explanations for deep neural networks. Pattern Recognition **121**, 108194 (2022)
8. Ghorbani, A., Abid, A., Zou, J.: Interpretation of neural networks is fragile. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 3681–3688 (2019)
9. Kwon, B.C., Choi, M.J., Kim, J.T., Choi, E., Kim, Y.B., Kwon, S., Sun, J., Choo, J.: Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records. IEEE transactions on visualization and computer graphics **25**(1), 299–309 (2018)
10. Li, Y., Rao, S., Solares, J.R.A., Hassaine, A., Ramakrishnan, R., Canoy, D., Zhu, Y., Rahimi, K., Salimi-Khorshidi, G.: Behrt: transformer for electronic health records. Scientific reports **10**(1), 1–12 (2020)
11. Lipton, Z.C.: The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue **16**(3), 31–57 (2018)
12. Luo, J., Ye, M., Xiao, C., Ma, F.: Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 647–656 (2020)
13. Ma, F., Chitta, R., Zhou, J., You, Q., Sun, T., Gao, J.: Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1903–1911 (2017)
14. Meng, D., Chen, H.: Magnet: a two-pronged defense against adversarial examples. In: Proceedings of the 2017 ACM SIGSAC conference on computer and communications security. pp. 135–147 (2017)
15. Quellec, G., Charriere, K., Boudi, Y., Cochener, B., Lamard, M.: Deep image mining for diabetic retinopathy screening. Medical image analysis **39**, 178–193 (2017)
16. Rasmy, L., Xiang, Y., Xie, Z., Tao, C., Zhi, D.: Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ digital medicine **4**(1), 1–13 (2021)

17. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
18. Shang, J., Ma, T., Xiao, C., Sun, J.: Pre-training of graph augmented transformers for medication recommendation. In: Kraus, S. (ed.) Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI 2019. pp. 5953–5959. IJCAI International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence (2019). https://doi.org/10.24963/ijcai.2019/825
19. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: International conference on machine learning. pp. 3145–3153. PMLR (2017)
20. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
21. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825 (2017)
22. Sun, M., Tang, F., Yi, J., Wang, F., Zhou, J.: Identify susceptible locations in medical records via adversarial attacks on deep predictive models. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 793–801 (2018)
23. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
24. Wang, W., Tang, P., Xiong, L., Jiang, X.: Radar: Recurrent autoencoder based detector for adversarial examples on temporal ehr. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 105–121. Springer (2020)
25. Xu, Y., Biswal, S., Deshpande, S.R., Maher, K.O., Sun, J.: Raim: Recurrent attentive and intensive model of multimodal patient monitoring data. In: Proceedings of the 24th ACM SIGKDD international conference on Knowledge Discovery & Data Mining. pp. 2565–2573 (2018)
26. Yeh, C.K., Hsieh, C.Y., Suggala, A., Inouye, D.I., Ravikumar, P.K.: On the (in) fidelity and sensitivity of explanations. Advances in Neural Information Processing Systems **32** (2019)