Generative Inpainting for Shapley-Value-Based Anomaly Explanation

Julian Tritscher[1,2(✉)], Philip Lissmann[1,2], Maximilian Wolf[3], Anna Krause[1,2], Andreas Hotho[1,2], and Daniel Schlör[1,2]

[1] CAIDAS Center for Artificial Intelligence and Data Science, Würzburg, Germany
[2] Julius-Maximilians-University of Würzburg, Würzburg, Germany
[3] Coburg University of Applied Sciences, Coburg, Germany
{tritscher, m.wolf, anna.krause, hotho, schloer}@informatik.uni-wuerzburg.de
philip.lissmann@studmail.uni-wuerzburg.de

Abstract. Feature relevance explanations currently constitute the most used type of explanation in anomaly detection related tasks such as cyber security and fraud detection. Recent works have underscored the importance of optimizing hyperparameters of post-hoc explainers which show a large impact on the resulting explanation quality. In this work, we propose a new method to set the hyperparameter of replacement values within Shapley-value-based post-hoc explainers. Our method leverages ideas from the domain of generative image inpainting, where generative machine learning models are used to replace parts of a given input image. We show that these generative models can also be applied to tabular replacement value generation for Shapley-value-based feature relevance explainers. Experimentally, we train a denoising diffusion probabilistic model for generative inpainting on two tabular anomaly detection datasets from the domains of network intrusion detection and occupational fraud detection, and integrate the generative inpainting model into the SHAP explanation framework. We empirically show that generative inpainting may be used to achieve consistently strong explanation quality when explaining different anomaly detectors on tabular data.

Keywords: Feature Relevance · XAI · SHAP · Diffusion · Perturbation.

1 Introduction

Explainable Artificial Intelligence (XAI) is currently an ever-increasing research topic in the domain of anomaly detection [30], with most attention being devoted to feature relevance explanations [30]. While many works simply apply existing post-hoc explainers with default configurations to obtain feature relevance explanations in anomaly detection [26,28], recent work shows that setting post-hoc explainer hyperparameters can have large impacts on explanation quality [26,28]. Popular Shapley-value-based feature relevance explainers for instance

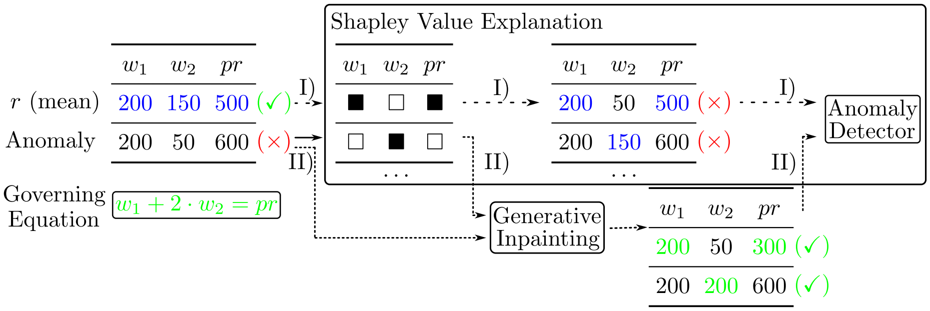

Fig. 1: Illustration of Shapley value explanations with I) exemplary mean replacements r and II) our alternative generative inpainting process with diffusion models on a fictional dataset with two weight attributes w_1, w_2 and a price pr that holds the normal behavior of $w_1 + 2 \cdot w_2 = pr$. Note that replacing perturbed features (marked ■) with r (mean) causes resulting data to not adhere to the governing equation, while inpainting respects kept values (marked □) to find suitable replacements.

require a heuristic for replacing perturbed feature values with other values as many anomaly detectors are not able to handle missing input features [26]. To handle this perturbation process, common choices are the specification of static reference data that is accessed for replacement values [2,15] or the marginalization of perturbed features, i.e., through using values from nearest neighbors in training data [15,21]. As replacing perturbed values without considering the kept values may result in artificial anomalies that introduce an erroneous signal into the perturbation process, Takeishi and Kawahara [24] introduce an optimization-based process for generating replacement values in anomaly detection and provide multiple relaxations for computational feasibility.

In this work, we explore an alternative option to obtain replacement values within Shapley-value-based explainers in anomaly detection. Our approach uses methods from inpainting with generative machine learning models [4], which we refer to as generative inpainting from hereon. This task, which has longstanding use in computer vision [4], as well increasing use in other domains such as sensor processing [11], human-computer music co-creation [29], and video processing [32], uses generative models to replace marked features within a given input with suitable values according to the remaining original feature values. In contrast to popular replacement value techniques for Shapley value explanations, generative inpainting respects dependencies to kept values while replacing marked features, and additionally enables efficient creation of perturbed datapoints during the explanation phase without the need for further datapoint optimizations as previous work in anomaly detection. We show that the functioning of generative inpainting models closely corresponds to the task of finding replacement values for a given datapoint under perturbations, and demonstrate how to leverage generative inpainting to supply Shapley-value-based explainers with replacement values for given datapoint perturbations. We illustrate the process of generative inpainting within Shapley value explanations in Figure 1. To achieve generative inpainting on tabular data, we train a Tabular Denoising Diffusion Probabilistic Model (TabDDPM) [31] and use the RePaint inpainting procedure [14] to generate replacement values for given perturbations. We integrate generative inpainting within the SHapley Additive exPlanations (SHAP) [15] explanation framework and conduct experiments on two tabular security datasets from the domains of network intrusion detection and occupational fraud detection. Evaluations on three different anomaly detectors per dataset reveal that generative inpainting can effectively serve as a source of replacement values for perturbation-based explainers, achieving good explanation performance across all detectors.

In summary, our contributions are as follows: 1) We show the compatibility between replacement values in Shapley value explanations and inpainting through generative models. 2) We demonstrate this compatibility on tabular data by training a TabDDPM [31] diffusion model with the RePaint inpainting procedure [14] and integrating the resulting model into the SHAP [15] explanation framework. 3) We quantitatively evaluate these SHAP explanations with generative inpainting replacements on two datasets and three anomaly detectors. 4) We provide code for our experiments that integrates generative inpainting with TabDDPM into the SHAP explanation framework.[4]

2 Related Work

As Shapley-value-based explainers are commonly used to obtain feature relevance explanations, multiple approaches exist for obtaining replacement values for features under perturbation. Reference-based approaches using for example the zero vector or the mean of the training data are considered as reasonable but arbitrary first choices in this area [2]. Marginalization of features e.g. through nearest neighbor search constitutes an additional option that might however influence explanations through its sensitivity to the data distribution [22]. Further, some model-specific choices of reference values exist in literature. For instance, Takeishi [23] uses the structure of their principle component analysis-based anomaly detection model to forgo the need of replacement values, but obtain an approach that is limited to principle component analysis-based anomaly detection. Beyond these simple heuristic approaches, Takeishi and Kawahara [24] provide a method to generate replacement values conditional to the values to keep. Their proposed method uses gradient-based input optimization and can therefore be applied to all fully differentiable anomaly detectors. In contrast to these works, we are the first to explore generative inpainting for the task of obtaining replacement values for any anomaly detector, regardless of architecture.

As we intend to leverage generative inpainting for model explanations, we require a generative machine learning model for our experiments. Here, diffusion models are a recent state-of-the-art architecture for data generation with multiple recent applications to tabular data. TabDDPM [10] combines a Gaussian

[4] Code and data are available at: https://professor-x.de/xai-diffusion.

diffusion model and a multinomial diffusion model to handle both numerical and categorical input data respectively. TabADM [31] apply a Gaussian diffusion model for tabular anomaly detection and add a datapoint rejection step to handle anomalies in the training data. DTPM [13] specialize a diffusion model for tabular anomaly detection by removing some generative abilities and focusing on the anomaly detection task. While this primes the model for anomaly detection, it removes its generative abilities, making it incompatible with our approach. From the available architectures we select TabDDPM [10] as it provides generative capabilities in contrast to DTPM and, in contrast to TabADM, is specifically constructed to handle the large amounts of categorical data inherent in our occupational fraud detection and network intrusion detection data.

Upon the training of a generative machine learning model, we further require a strategy to conduct inpainting using this model. While generative diffusion models may be used directly for inpainting by repeatedly feeding a noised version of the kept inputs into the model during the sampling process, this approach is known to create edge artifacts as the model can not build smooth transitions between replaced and kept inputs [16]. To mitigate this effect, Nichol et al. [16] showcase how to fine-tune a generative diffusion model to the task of inpainting. In contrast, RePaint [14] is an inpainting strategy for image-based diffusion models that does not require any re-training of the underlying diffusion model. Instead, RePaint introduces additional loops into the generative diffusion process to ensure homogeneity between kept and newly generated areas of an image. Since RePaint does not require training and can be applied to a given diffusion model post-hoc, therefore reducing the computational complexity by omitting the training phase, we utilize it as inpainting strategy in this work.

3 Methodology

In this section, we first give a brief overview of Shapley value explanations and their inherent need for replacement values. We then introduce the task of generative inpainting and demonstrate how to adapt this task to provide replacement values for Shapley value explanations. Finally, we introduce the generative inpainting process based on the generative TabDDPM model [10] and the RePaint inpainting process [14] used throughout our experiments.

3.1 Shapley Value Explanations and Replacement Values

To explain the decision process of an anomaly detector, feature relevance explanations are a commonly used technique that ranks the relevance of each input feature with respect to the output of the detector [30]. One popular way of obtaining these explanations is based on Shapley values [15,23]. Shapley values [20] constitute a popular result from cooperative game theory that evenly distribute a jointly generated gain to a group of participants by iteratively assessing a fictional gain that would be achieved by different subgroups. By viewing the input features of a data point as participants and setting the machine learning model output as the generated gain, the framework of Shapley values may be used to obtain a relevance score ϕ_i for each feature i through

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S)), \tag{1}$$

with a set of features $N = \{1, 2, \ldots, n\}$ and a function $v(S)$ that calculates the gain achieved by the subgroup $S \subseteq N$. Since the gain that we wish to calculate is the anomaly score of the model to explain, calculating Shapley values requires a way to calculate the anomaly score obtained by using only a subset of features and removing all features that are not present in the subgroup.

This poses a significant challenge for a variety of machine learning models, as most models are not able to handle missing inputs. To remedy this issue, common implementations of Shapley value explanations such as SHapley Additive exPlanation (SHAP) [15] use reference data to replace features missing in the subgroup with default values from the reference data instead of deleting the feature. Thus for a given datapoint $x = [x_1, x_2, \ldots, x_n]$ of dimensionality n to explain and a subgroup S of features to investigate, a reference point $r \in \mathbb{R}^n$ is used to replace all features not in the subgroup S through $h(x, r, S) = \hat{x}$ with

$$\hat{x}_s = \begin{cases} x_s, & \text{if } s \in S \\ r_s, & \text{else} \end{cases}, \quad \forall s \in N. \tag{2}$$

The resulting perturbed datapoint may then be scored by the original model, providing a new anomaly score that serves as the subgroup gain $v(S)$.

As demonstrated in [26], this perturbation procedure can produce problematic inputs in the domain of anomaly detection, as the combination of replaced and kept features may introduce unwanted anomalous signals. As previously explored in [24], this issue needs to be remedied by choosing the replacing values for missing features conditional such that the kept feature values do not introduce new unwanted anomalies into the data. This leads to obtaining references $r \sim k(x, S)$ through a function k that produces replacement vectors for a given datapoint x and a defined set S of features to keep.

3.2 Perturbation with Generative Inpainting

Following the notation of [14], generative inpainting takes an input $x \in \mathcal{X}$ and an associated binary mask $m \in \mathcal{X}_{bin}$ of same dimensionality, that determines which inputs are to be kept (1) or replaced (0) through a generative inpainting model. Generative inpainting $i : \mathcal{X} \times \mathcal{X}_{bin} \rightarrow \mathcal{X}$ then aims to produce a new data point that combines the old inputs that are to be kept $m \odot x$ with matching new inputs $(1 - m) \odot g(m \odot x, m)$ from a generative model $g : \mathcal{X} \times \mathcal{X}_{bin} \rightarrow \mathcal{X}$ that is conditioned on the kept inputs and the binary mask.

$$i(x, m) = m \odot x + (1 - m) \odot g(m \odot x, m) \tag{3}$$

Here we note that the inputs of the generative process, namely the binary mask of features to keep m and the feature values x, contain the same information as the inputs of the conditional replacement value generation process $r \sim k(x, S)$. To make the replacements compatible with generative inpainting, we convert the Shapley-value-based set representation of kept features S to a binary mask

$$m_s = \begin{cases} 1, & \text{if } s \in S \\ 0, & \text{else} \end{cases}, \quad \forall s \in N. \tag{4}$$

Through this mapping we are able to directly use generative inpainting models to provide perturbed datapoints for the calculation of Shapley values.

3.3 Tabular Diffusion with TabDDPM

In the following sections, we describe our process of obtaining a generative model for tabular data and the inpainting strategy used to achieve generative inpatining on tabular anomaly detection data.

TabDDPM [10] is a generative denoising diffusion probabilistic model (DDPM) [6] designed specifically to generate both numerical and categorical tabular data. DDPM models are trained by reversing a so called diffusion process that iteratively converts a given input to random noise. The diffusion process is defined as a Markov chain

$$q(x_{1:T}|x_0) := \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{5}$$

that gradually adds a small amount of noise $q(x_t|x_{t-1})$ to a datapoint x_0 over T steps, such that the final point x_T consists of entirely random noise. The reverse process

$$p_\theta(x_{0:T}) := \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \tag{6}$$

also describes a Markov chain that gradually reconstructs data from noise and may be approximated through a neural network that parameterizes the function $p_\theta(x_{t-1}|x_t)$ and removes one step of noise with parameters θ. Training a network with a known noising function q on a given dataset by learning the reverse function p_θ through optimizing a variational lower bound then allows to iteratively apply the learned function p_θ to random noise to generate new datapoints.

To model tabular data through this diffusion process, TabDDPM joins both a Gaussian diffusion model that uses Gaussian noise within q for numerical data and a multinomial diffusion model that applies noise to a multinomial distribution in q for categorical data. Gaussian diffusion models parameterize both q and p as Gaussian distributions, using neural networks to approximate both means and variances of the Gaussian distribution p. Gaussian DDPM models [6] further ease the learning process by fixing the mean and learning a diagonal variance matrix for p, which enables them to replace the variational lower bound with a simple mean squared error between true and predicted added noise when training the neural network. Since Gaussian noise is not readily applicable to categorical attributes, multinomial diffusion models [8] instead add noise to the probability distribution of all possible values for a categorical feature and then re-draw the categorical feature according to the noised distribution. Training the neural network for p then focuses on predicting the one-step de-noised probability distribution for each categorical feature while observing only the re-drawn feature values and the current diffusion step, using the Kullback-Leibler divergence between true and predicted distribution as loss function. TabDDPM simply combines both models by jointly training the contained neural networks with a combined loss function.

3.4 Generative Inpainting for Diffusion Models

While diffusion-based generative models are directly applicable to inpainting by generating only the desired area of the input while feeding a noised version of the input to keep into the model during each denoising step, this process may cause artifacts due to the model's inability to adapt newly generated inputs to kept input [16]. The recent RePaint [14] approach proposes an adaptation of this denoising process for inpainting that prevents artifacts without any fine-tuning.

Formally, RePaint generates the unknown (i.e. removed) feature values through the neural network-based reverse diffusion process from random noise

$$x_{t-1}^{unknown} = p_\theta(x_{t-1:T}) \tag{7}$$

while taking the values for kept features from the original input x_0 by adding noise according to the current timestep t through

$$x_{t-1}^{known} = q(x_{1:t-1}|x_0). \tag{8}$$

The resulting latent input in each timestep for a given mask m is then joint together through

$$x_{t-1} = m \odot x_{t-1}^{known} + (1-m) \odot x_{t-1}^{unknown}. \tag{9}$$

To harmonize x such that x^{known} and $x^{unknown}$ do not diverge and create a combined x that does not match the data distribution, RePaint adds an additional noising step to the joint x_{t-1} such that

$$x_t' = q(x_t|x_{t-1}). \tag{10}$$

On subsequent steps, this x_t' is then integrated in the calculation of $x^{unknown}$ by replacing Equation (7) with

$$x_{t-1}^{unknown} = p_\theta(x_{t-1:T}'), \tag{11}$$

effectively creating a loop that provides the denoising component which generates $x^{unknown}$ with contextual knowledge of x^{known} and therefore allowing the diffusion model to create smooth boundaries between the kept known and newly generated unknown segments. This noising-denoising-loop is repeatedly used throughout the data generation process. For the full iterative procedure with pseudo code, we refer to the original paper [14].

Table 1: Data splits with number of anomalies and labeled ground truth explanations for ERP and CIDDS-001 data.

(a) ERP

split	samples	anomalies	explanations
train	32, 337	0	0
eval	36, 778	50	0
test	37, 407	86	86

(b) CIDDS-001

split	samples	anomalies	explanations
train	12, 525, 224	0	0
eval	10, 310, 540	910, 375	0
test	8, 451, 274	746, 230	80

4 Experiments

We quantitatively evaluate the suitability of generative inpainting for anomaly explanations on two security datasets from the domains of network intrusion detection and occupational fraud detection.

4.1 Data, Anomaly Detectors, and Metrics

In order to evaluate our proposed replacement generation procedure, we use two established anomaly detection datasets with ground truth explanations. The ERP dataset [25] contains the enterprise resource planning data of a production company, where multiple occupational fraud cases are included next to normal business behavior. The CIDDS-001 dataset [18] consists of a simulated computer network of virtual machines, which interact within the network based on normal actions or attack scenarios. Dataset statistics are listed in Table 1. For both datasets, fully trained anomaly detectors and a binary explanation ground truth for selected anomalies within the unseen test set are openly available [26,28] and may be used to evaluate feature relevance XAI approaches.

The fully trained anomaly detectors include an autoencoder neural network (AE) [3], a one-class support vector machine (OC-SVM) [19], and an isolation forest (IF) [12] with hyperparameters and preprocessing strategies optimized on the validation splits. The optimized models all use one-hot encoding for categorical features and quantization for numerical features and achieve high mean average precision (PR) scores as seen in Table 2.

Table 2: PR scores of the used fully trained anomaly detectors on the test data of CIDDS-001 and ERP after the hyperparameter tuning by [27,28]. Higher is better.

Detector	CIDDS-001	ERP
OC-SVM [19]	0.995	0.73
AE [3]	0.992	0.69
IF [12]	0.993	0.49

To obtain quantitative scores of explanation quality, explainers are applied to existing, fully trained anomaly detectors on the anomalies with available ground truth. The resulting explanations are compared to the binary explanation ground truth through area-under-the-receiver-operator-characteristic (ROC) and cosine similarity (COS), which are established metrics for the evaluation of feature relevance explanations with binary ground truth [5,9] that can be applied to directly compare binary ground truth and continuous feature relevance scores in a single datapoint. While the ROC score describes the sensitivity to anomalous features in relation to false positives for multiple thresholds [5], COS denotes the direct similarity between given feature relevance scores and the entire binary ground truth [9]. To obtain a quantitative score of all explanations, these metrics are then aggregated across all labeled anomalies [5,9].

4.2 Generative Models and Inpainting

In order to perform generative inpainting on both security datasets, we first train two separate TabDDPM models in a generative setting using the official PyTorch [17] implementation. To remain compatible with the fully trained anomaly detectors described in Section 4.1, we follow the train-validation-test split and the preprocessing steps of the fully trained anomaly detectors during TabDDPM training, while only removing the anomalies from the validation data to assess the generative performance of the models on purely normal data. For model optimization, we conduct a hyperparameter search using the optuna framework [1] with 100 trials over the search space listed in Table 3, following the recommendations of the TabDDPM authors [10], while limiting batch size to 4096, adding an intermediate diffusion step option of 250, and adding a training step option of 30.000 steps to ensure convergence. Training is carried out on the purely normal training splits and validation optimizes the TabDDPM loss on the unseen validation set with removed anomalies. As inpainting process we use the RePaint framework as described in Section 3.4 with its default parameters from the original paper without optimization.

Having obtained both anomaly detectors and generative diffusion models, we follow the experimental setup of previous work on feature relevance explanations

Table 3: Hyperparameter space for TabDDPM within the optuna optimization framework [1] over 100 steps, following the authors' recommendations in [10].

Hyperparameter	Search space
Learning rate	LogUniform{[0.00001, 0.003]}
Diffusion steps train	Cat{100, 250, 1000}
Training steps	Cat{5000, 20000, 30000}
MLP Layer count	Cat{2, 4, 6, 8}
MLP layer dimension	Cat{128, 256, 512, 1024}
Batch size	4096

on the two datasets [26,28] by applying SHAP on the detectors for the anomalies with available ground truth. We integrate repaint into the official implementation of the SHAP explanation framework [15] by retrieving both the datapoint and the desired perturbation masks from SHAP and supplying the inpainted data back into the framework.

Lastly, we compare the inpainted replacement values with other established choices for replacement values by observing the quality of the obtained explanations through the ROC and COS scores as described in Section 4.1. We compare against commonly used replacement strategies in anomaly detection [26], namely using cluster centers of k-means clustering on the training data (kmeans), the mean of the training data (mean), the zero vector (zeros), the nearest neighbor of the explained datapoint within the training data (NN), as well as the gradient-based optimization strategy of [24] (lopt).

4.3 Results

We report the explanation scores of SHAP with varying references on three anomaly detectors per dataset and evaluate random explanations drawn from uniform noise for comparison. We observe that random explanations achieve a mean ROC score of around 50% and a mean COS score of roughly 0%, matching the intuition of the metrics. On the occupational fraud detection dataset ERP in Table 4 we see that our generative inpatining approach with TabDDPM manages to outperform all other references for both the OC-SVM and the IF detector. Note that the interpretation of the standard deviation here is non-trivial due to the complex effects of averaging across multiple ROC scores [7], but high values can be seen as indicators of the varying complexity of ground truth explanations across datapoints, with the data containing both datapoints with simple and challenging explanations simultaneously. On the AE detector, the gradient-based optimization procedure, which is only applicable to differentiable detectors, achieves highest results, suggesting that it may be beneficial to rely on this technique when applicable. Nevertheless, our diffusion-based approach also achieves decent explanation scores on this detector, making it a method that successfully obtains high quality explanations throughout all datasets. This is especially relevant, as other reference approaches showcase a more erratic behavior throughout the different anomaly detectors. While these approaches may

Table 4: Mean and standard deviation of explainer performance in % for SHAP on multiple detectors using different reference values on the ERP data.

	AE		OC-SVM		IF		avg. over detectors	
Explainer	ROC_{XAI}	COS_{XAI}	ROC_{XAI}	COS_{XAI}	ROC_{XAI}	COS_{XAI}	ROC_{XAI}	COS_{XAI}
random explanation	50.7 (15.8)	0.3 (17.0)	50.7 (15.8)	0.3 (17.0)	50.7 (15.8)	0.3 (17.0)	50.7 (15.8)	0.3 (17.0)
SHAP + kmeans	75.4 (14.2)	44.8 (12.2)	57.2 (12.4)	17.7 (14.2)	70.3 (23.6)	39.8 (28.7)	67.6 (16.7)	34.1 (18.4)
SHAP + mean	74.4 (17.1)	32.3 (25.1)	54.6 (13.2)	13.7 (16.7)	64.2 (26.4)	26.2 (39.8)	64.4 (18.9)	24.1 (27.2)
SHAP + zeros	82.3 (14.5)	58.2 (16.3)	64.2 (15.5)	−8.4 (12.7)	66.7 (27.8)	25.8 (47.0)	71.1 (19.3)	25.2 (25.3)
SHAP + NN	56.0 (15.1)	16.8 (38.0)	60.5 (12.5)	32.1 (28.2)	53.7 (13.3)	12.3 (33.9)	56.7 (13.6)	20.4 (33.4)
SHAP + lopt	**88.6 (11.2)**	**66.1 (20.5)**	N/A	N/A	N/A	N/A	N/A	N/A
SHAP + TabDDPM	71.9 (15.2)	47.8 (19.8)	**72.4 (20.3)**	**43.2 (28.8)**	**73.3 (17.6)**	**45.1 (23.4)**	**72.5 (17.7)**	**45.4 (24.0)**

at times achieve very high explanation quality on one detector, they may also obtain scores that are close to entirely random explanations on other detectors, as observable for instance with kmeans and mean on the OC-SVM (57.2% and 54.6% ROC respectively), or the NN references on the AE detector (56.0% ROC). The proposed inpatining method using TabDDPM manages to maintain explanation quality considerably above random noise even on the least well performing AE explanations (71.9% ROC), making it a stable and high-performing option for obtaining replacement values.

On the network intrusion detection dataset CIDDS-001 in Table 5 we observe consistently higher explanation scores for all references compared to the ERP dataset. On this dataset, our generative inpainting method falls short of surpassing competing reference methods, which exhibit performance spikes with certain detectors. For instance kmeans achieves 91.3% ROC on the AE detector or zeros yield 87.2% ROC on the IF detector. Nevertheless, our generative inpatining approach manages to provide the second highest explanation scores across all detectors (82.1% ROC), while other methods show considerable decreases in explanation quality across detectors, e.g. with kmeans on IF (73.5% ROC) or zeros on AE (51.9% ROC). While the findings from CIDDS-001 imply that the generative model might not consistently yield the most suitable references via inpainting for every data instance, we note that it still maintains high explanation scores across all anomaly detectors by using a standard generative model and an inpainting procedure without any hyperparameter optimization, clearly indicating potential for further enhancement and refinement in future work.

Overall, using a standard generative model for tabular data in combination with a recent inpainting technique, we successfully demonstrate an initial proof of concept for utilizing generative inpainting to determine reference values within perturbation-based explanations. Our proposed generative inpainting strategy manages to perform exceptionally well on the ERP data, while also maintaining acceptable explanation performance on the CIDDS-001 data throughout all anomaly detectors without providing poor explanations as is observable for other replacement options. This shows the feasibility of obtaining reference values through generative inpainting, while leaving further room for improvement through adaptations of generative models and inpainting strategies to the underlying domain.

Table 5: Mean and standard deviation of explainer performance in % for SHAP on multiple detectors using different reference values on the CIDDS-001 data.

	AE		OC-SVM		IF		avg. over detectors	
Explainer	ROC$_{XAI}$	COS$_{XAI}$	ROC$_{XAI}$	COS$_{XAI}$	ROC$_{XAI}$	COS$_{XAI}$	ROC$_{XAI}$	COS$_{XAI}$
random explanation	50.6 (21.7)	1.7 (38.1)	50.6 (21.7)	1.7 (38.1)	50.6 (21.7)	1.7 (38.1)	50.6 (21.7)	1.7 (38.1)
SHAP + kmeans	**91.3 (8.7)**	**71.8 (15.9)**	88.1 (8.7)	68.6 (12.6)	73.5 (19.3)	52.8 (23.1)	**84.3 (12.2)**	**64.4 (17.2)**
SHAP + mean	82.6 (12.7)	57.4 (19.6)	**90.1 (7.5)**	**71.0 (12.9)**	65.8 (18.7)	33.4 (25.7)	79.5 (13.0)	53.9 (19.4)
SHAP + zeros	51.9 (11.7)	−20.7 (16.6)	85.2 (9.2)	48.1 (13.9)	**87.2 (8.7)**	**65.9 (13.2)**	74.8 (10.0)	31.1 (14.6)
SHAP + NN	75.4 (11.6)	61.1 (15.7)	71.3 (11.9)	56.4 (18.1)	68.8 (11.7)	43.8 (22.7)	71.8 (11.7)	53.8 (18.8)
SHAP + lopt	88.5 (8.7)	68.9 (14.3)	N/A	N/A	N/A	N/A	N/A	N/A
SHAP + TabDDPM	83.9 (13.9)	65.9 (15.5)	82.9 (10.6)	59.7 (12.9)	79.6 (12.6)	57.9 (15.2)	82.1 (12.4)	61.2 (14.5)

5 Conclusion

In this paper, we introduced a novel way to find suitable replacement data for perturbation-based feature relevance explainers in anomaly detection. We explored the connection between these replacement values and the task of image inpainting from the domain of computer vision, showing a compatibility that allows the use of generative inpainting models for replacement value generation. To demonstrate the feasibility of generative inpainting for replacement data generation, we integrated a generative inpainting strategy into the popular SHAP explanation framework. We conducted experiments on two security datasets from occupational fraud detection and network intrusion detection, training a generative diffusion model for each dataset and applying SHAP with the proposed generative inpainting procedure to explain multiple existing anomaly detectors. Our experiments show that SHAP achieves stable explanation quality using reference values from generative inpainting on both datasets and all evaluated anomaly detectors, making generative inpainting a reliable choice for reference values in anomaly detection. The proposed work is a two stage approach that leverages generative machine learning models and inpainting procedures from computer vision. While we demonstrated the viability of this approach in first experiments using an established generative model and an existing inpainting strategy, we note that the approach necessitates a model with good generative capabilities. On the used network intrusion detection data, for instance, the use of different data preprocessing strategies for anomaly detectors and generative models may further enhance generative performance, as choosing preprocessing schemes based on the anomaly detector may result in datasets that sufficiently encode anomalous behavior, but impede the generative model's ability to learn all data characteristics. Additionally, while we illustrated our approach on occupational fraud detection and network intrusion detection, it may be applied to further domains where complex data dependencies need to be respected during perturbations.

Disclosure of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.

References

1. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: 25th ACM SIGKDD (2019)
2. Ancona, M., Oztireli, C., Gross, M.: Explaining Deep Neural Networks with a Polynomial Time Algorithm for Shapley Value Approximation. In: 36th International Conference on Machine Learning. pp. 272–281. PMLR (May 2019)
3. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016), http://www.deeplearningbook.org

4. Guillemot, C., Le Meur, O.: Image inpainting: Overview and recent advances. IEEE signal processing magazine **31**(1), 127–144 (2013)
5. Hägele, M., Seegerer, P., Lapuschkin, S., Bockmayr, M., Samek, W., Klauschen, F., Müller, K.R., Binder, A.: Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. Scientific reports **10**(1), 1–12 (2020)
6. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. In: Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851. Curran Associates, Inc. (2020)
7. Hogan, J., Adams, N.M.: On Averaging ROC Curves. Transactions on Machine Learning Research (Feb 2023)
8. Hoogeboom, E., Nielsen, D., Jaini, P., Forré, P., Welling, M.: Argmax Flows and Multinomial Diffusion: Learning Categorical Distributions. In: Advances in Neural Information Processing Systems. vol. 34, pp. 12454–12465 (2021)
9. Kauffmann, J., Ruff, L., Montavon, G., Müller, K.R.: The clever hans effect in anomaly detection. arXiv preprint arXiv:2006.10609 (2020)
10. Kotelnikov, A., Baranchuk, D., Rubachev, I., Babenko, A.: Tabddpm: Modelling tabular data with diffusion models. In: ICML 2023. pp. 17564–17579. PMLR (2023)
11. Li, Y., Song, L., Hu, Y., Lee, H., Wu, D., Rehm, P., Lu, N.: Load profile inpainting for missing load data restoration and baseline estimation. IEEE Transactions on Smart Grid (2023)
12. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation Forest. In: ICDM. pp. 413–422. IEEE Computer Society (2008)
13. Livernoche, V., Jain, V., Hezaveh, Y., Ravanbakhsh, S.: On Diffusion Modeling for Anomaly Detection (May 2023)
14. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: RePaint: Inpainting using Denoising Diffusion Probabilistic Models. In: CVPR 2022. pp. 11451–11461. IEEE (Jun 2022)
15. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems. pp. 4765–4774 (2017)
16. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In: ICML 2022. PMLR (2022)
17. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32. pp. 8024–8035. Curran Associates, Inc. (2019), http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
18. Ring, M., Wunderlich, S., Grüdl, D., Landes, D., Hotho, A.: Flow-based benchmark data sets for intrusion detection. In: 16th European Conference on Cyber Warfare and Security (ECCWS), pp. 361–369. ACPI (2017)
19. Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. Neural Computation **13**(7), 1443–1471 (Jul 2001)
20. Shapley, L.S.: A value for n-person games. Classics in game theory **69** (1997)
21. Štrumbelj, E., Kononenko, I.: Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems **41**(3), 647–665 (Dec 2014)

22. Sundararajan, M., Najmi, A.: The Many Shapley Values for Model Explanation. In: ICML 2020. pp. 9269–9278. PMLR (Nov 2020)
23. Takeishi, N.: Shapley values of reconstruction errors of pca for explaining anomaly detection. In: int. conf. on data mining workshops (icdmw). pp. 793–798 (2019)
24. Takeishi, N., Kawahara, Y.: A characteristic function for shapley-value-based attribution of anomaly scores. Transactions on Machine Learning Research (2023)
25. Tritscher, J., Gwinner, F., Schlör, D., Krause, A., Hotho, A.: Open ERP System Data For Occupational Fraud Detection (Jul 2022)
26. Tritscher, J., Krause, A., Hotho, A.: Feature relevance XAI in anomaly detection: Reviewing approaches and challenges. Frontiers in Artificial Intelligence **6** (2023)
27. Tritscher, J., Schlör, D., Gwinner, F., Krause, A., Hotho, A.: Towards Explainable Occupational Fraud Detection. In: Machine Learning and Principles and Practice of Knowledge Discovery in Databases. pp. 79–96. Communications in Computer and Information Science, Springer Nature Switzerland, Cham (2023)
28. Tritscher, J., Wolf, M., Hotho, A., Schlor, D.: Evaluating feature relevance XAI in network intrusion detection. In: The World Conference on eXplainable Artificial Intelligence (2023)
29. Wei, S., Xia, G., Zhang, Y., Lin, L., Gao, W.: Music phrase inpainting using long-term representation and contrastive loss. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 186–190. IEEE (2022)
30. Yepmo, V., Smits, G., Pivert, O.: Anomaly explanation: A review. Data & Knowledge Engineering **137**, 101946 (2022)
31. Zamberg, G., Salhov, M., Lindenbaum, O., Averbuch, A.: TabADM: Unsupervised Tabular Anomaly Detection with Diffusion Models (Jul 2023)
32. Zeng, Y., Fu, J., Chao, H.: Learning joint spatial-temporal transformations for video inpainting. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16. pp. 528–543. Springer (2020)