Data Generation for Explainable Occupational Fraud Detection

Julian Tritscher[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

Abstract. Occupational fraud, the deliberate misuse of company assets by employees, causes damages of around 5% of yearly company revenue. Recent work therefore focuses on automatically detecting occupational fraud through machine learning on the company data contained within enterprise resource planning systems. Since interpretability of these machine learning approaches is considered a relevant aspect of occupational fraud detection, first works have already integrated post-hoc explainable artificial intelligence approaches into their fraud detectors. While these explainers show promising first results, systematic advancement of explainable fraud detection methods is currently hindered by the general lack of ground truth explanations to evaluate explanation quality and choose suitable explainers. To avoid expensive expert annotations, we propose a data generation scheme based on multi-agent systems to obtain company data with labeled occupational fraud cases and ground truth explanations. Using this data generator, we design a framework that enables the optimization of post-hoc explainers for unlabeled company data. On two datasets, we experimentally show that our framework is able to successfully differentiate between explainers of high and low explanation quality, showcasing the potential of multi-agent-simulations to ensure proper performance of post-hoc explainers.

Keywords: Feature Relevance · XAI · Fraud Detection · Simulation.

1 Introduction

Occupational fraud describes the misuse of company assets by an internal employee, for instance through theft or bribery. This type of fraud costs companies around 5% of their annual revenue [1], making the detection of occupational fraud relevant for many companies. With the increasing company digitization in enterprise resource planning (ERP) systems, large amounts of company ERP data enable an automated detection of occupational fraud through machine learning [15]. Next to the development of new methods for automated occupational

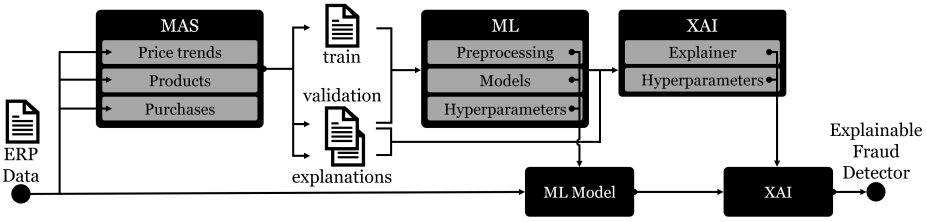

Fig. 1. Our proposed framework for automated and explainable fraud detection on ERP system data in the absence of labeled data. The system uses an MAS that is set up with market information from an unlabeled dataset to generate data with anomaly detection and explanation labels. This labeled data is then used to determine design choices for anomaly detectors and explainers to use on the original unlabeled data.

fraud detection, economics research has also identified explainability as especially relevant in this domain [7]. In the past, post-hoc feature relevance explainers [21] have already been applied in occupational fraud detection [21,23].

One challenge in this area of research is the lack of ground truth explanations that are needed to differentiate between well- and poorly-performing explainers. When new ERP system data is made available for auditing within the company, the presence and location of occupational frauds are usually not known within the data. Therefore, obtaining explanation ground truth through human experts would first require experts to manually identify occupational frauds in the large amounts of unlabeled company data, before ground truth explanations may be created for found fraud cases. To avoid this expensive ground truth generation through expert auditors, we propose to use a synthetic data generator based on Multi Agent Systems (MAS) to provide the necessary ground truth explanations. In prior work, we proposed to use MAS-based simulations to mirror the normal business processes of a company after a given unlabeled dataset and integrated known fraud scenarios into the simulation to obtain labeled data for occupational fraud detection [22]. While this simulation allows for automated occupational fraud detection with machine learning approaches, it does not provide any ground truth explanations for the identification of suitable explainers. As the common practice of arbitrarily choosing explainers in the domains of anomaly and fraud detection can result in near random explanations [21,24], ground truth explanations are highly relevant for designing explainable occupational fraud detectors.

In this paper, we therefore extend our MAS-based simulation [22] to generate ground truth explanations. As the underlying processes that govern the data generation are known within the simulation, generated anomalous entries can be identified during generation time. Using this extended MAS-based simulation, we propose a framework that enables the optimization of post-hoc feature relevance explainers for occupational fraud detection in the absence of real ground truth explanations.

Our framework is illustrated in Figure 1. When given new and unlabeled data from a company's ERP system, we extract general environment information such as price and purchasing trends from the company data and feed them into the MAS-based simulation. Within the MAS, the normal business process of the company is then simulated through established best practices from economics research and known fraud cases are modeled and added during the generation process. Building upon this simulation framework, we generalize this synthetic generative labeling approach of fraud versus non-fraud to generate ground truth explanation labels along the simulation process that reflect the actual feature-combinations that are indicative of the generated fraud case. With this labeled data with ground truth explanations, both machine learning approaches and explainers may be selected according to established performance metrics on the synthetically created simulation data. Choices for machine learning and explainer approaches are then directly applied on the original company data, enabling automated explainable occupational fraud detection. In our experiments, we use occupational fraud data with manually labeled ground truth explanations [20] to evaluate whether well-performing explainers on MAS-generated data transfer their high explanation scores to the underlying company dataset. Experiments on two datasets show that explanation quality indeed transfers from the MAS data to the underlying data, successfully enabling the use of MAS-based simulation to determine suitable post-hoc explainers on unlabeled company data for occupational fraud detection.

In summary, our contributions are as follows: (1) We extend an existing MAS-based simulation for ERP data to automatically provide ground truth explanations for occupational fraud cases created during data generation. (2) We provide a framework that identifies well-performing feature relevance explainers for occupational fraud detection in company data in the absence of real ground truth explanations. (3) Experimentally, we show on two datasets that the well-performing explainers identified by our MAS-based simulation framework successfully generalize to unlabeled company data. (4) We provide code for our Java-based implementation of the MAS simulation, as well as the Python-based optimization framework.[4]

To our knowledge, we are the first to use MAS-based simulation to aid explainable AI through automated ground truth generation. This is especially relevant to the XAI community, as ground truth explanations remain scarce and arbitrarily choosing explainers can result in near random explanations [21].

2 Related Work

Feature relevance explanations are the most common explanations in anomaly detection and assign a score to each feature of an individual datapoint that expresses its relevance to the model prediction [25]. While many works use existing feature relevance explainers in the domains of anomaly and fraud detection

[4] Code and data are available under: https://professor-x.de/xai-fraud-mas

[21,25], established unsupervised explainer evaluation schemes are not readily applicable to anomaly and fraud detection as they rely on perturbation schemes that would introduce additional anomalous signals into the data [23]. While binary ground truth can be used to evaluate feature relevance explainers, the low availability of ground truth leads to only few works using ground truth evaluations in anomaly detection [11,23,24]. In our work, we follow these evaluations by using the ground truth datasets from [23,24] in our experiments.

Beyond the evaluation of explainers, multiple works have already explored the combination of explainable artificial intelligence (XAI) and MAS. One line of research focuses on explaining single agents within MAS systems. Here, [10] focus on explaining agents within MAS that have been trained through reinforcement learning. [13] provides a framework for obtaining explanations regarding the actions and interactions between agents within financial MAS. In contrast to these works, our work does not aim to explain the MAS itself but merely uses it as a tool for generating ground truth explanations that enables the choice of well-performing XAI methods for the task of fraud detection in unlabeled company data. An additional line of research regarding MAS and XAI uses MAS as a theoretical framework to model general properties of explanations such as the interaction between different types of explanation methods [4] or the theoretical modeling of interesting explanations [5]. Our work does not focus on theoretical aspects of XAI and instead uses an MAS as a data generator that provides ground truth explanations.

Overall, while previous work on MAS and XAI exist, we are the first to use MAS-based simulation to generate synthetic data with ground truth explanations and enable explainable occupational fraud detection on fully unlabeled data while ensuring good explainer performance.

3 Methodology

This section gives an overview of the methodological aspects of our work. Firstly, we describe our MAS-based simulation, as well as our proposed additions that allow the generation of ground truth explanations. Secondly, we integrate our MAS-based simulation into the proposed framework for explainable fraud detection on unlabeled company data. Finally, we introduce the XAI models and XAI hyperparameters that are integrated into the proposed framework and explain the metrics used for quantitative evaluation of the XAI models within the framework.

3.1 MAS-based Company Simulation

Our goal in this work is the selection of well-performing feature relevance explainers for explainable occupational fraud detection on given unlabeled company data, for which we require ground truth explanations to quantitatively evaluate different explainers. However, manual generation of ground truth explanations for fraud cases requires expensive expert annotations, where occupational fraud

scenarios need to first be identified within the data and afterwards explanations may be derived from identified frauds. Since even the identification of frauds in the data is costly, in previous work we proposed to use agent-based simulation to obtain fraud detection labels for unlabeled company data [22]. Here, existing economics research, which provides established ways to model the business process of a production company through an MAS, formulates a theoretical backbone for a data-driven simulation [6]. Instead of learning the policies of the individual agents within the MAS from data, best practices for business and market strategies are used by individual planning, purchasing, production, and delivery agents that make up the company, resulting in a fully functional company simulation that uses established economics strategies, e.g., for purchasing goods or scheduling production. Using the agents, individual actions can be translated directly into ERP transactions and written into a relational database. Although this process can simulate an entire ERP system, we design the generator specifically to optimize existing explainable fraud detection methods that use a condensed table view of ERP data [23] and therefore configure the simulation to directly generate transactions for the aggregated table. This results in a single table that includes accounting information in 42 categorical and 10 numerical columns [23]. To generate normal business data that closely mimics the economic situation of a given company within the MAS-based simulation, prior knowledge of an underlying company regarding price trends and product information may be extracted from unlabeled company data and integrated into the simulation environment. Finally, the agent functionality within the MAS is extended to allow four different types of occupational fraud, namely Invoice Kickback, Selling Kickback, Non-Cash Larceny, and Corporate Injury [1]. These frauds include both indirect kickback schemes where employees are bribed externally to accept company trades at harmful prices during material purchases (Invoice Kickback) and product sales (Selling Kickback), as well as direct schemes where materials are stolen during production (Non-Cash Larceny) or employee wages are erroneously increased to damage the company (Corporate Injury). Frauds are committed directly during the data generation, where occurrences of these cases are tracked to obtain labeled normal and fraudulent data for occupational fraud detection. The resulting MAS-based data generator provides labeled training data to optimize fraud detection models, but requires further extensions to be capable of generating ground truth explanations for generated fraud cases.

3.2 MAS-based Ground Truth Explanations

While the previously introduced MAS-based simulation enables the optimization of fraud detection models for completely unlabeled data, we aim to extend this simulation to further enable the optimization of feature relevance explainers. To achieve this, we note that the known causal relationships and well-defined processes within the simulation can also be used to obtain full explanations for each fraud case.

To generate ground truth explanations, we extend the MAS-based simulation. The simulation resembles a full description of the normal business process

of the simulated company while also integrating known deviations from this normal process in the form of occupational fraud cases. We therefore extend processes and thus agents modeling non-normal behavior by defining rules and heuristics that highlight features and feature-combinations that deviate from the normal behavior and are indicative for each given fraud case. For instance, in the non-cash larceny case implemented in the MAS-based simulation, a delivery of raw materials is partially stolen during delivery and covered up by altering the already payed purchase order to book only the remaining materials into storage. The fraudulent change of the purchase covers up differences in true and internally tracked materials that would be revealed during the regular stocktaking processes within company warehouses, but leaves a transaction within the company ERP system that has substantially higher average prices for the ordered materials compared to normal orders. This anomalous entry can be identified through the recorded paid price and the amount of delivered materials, which show a much higher price for a single material than normal entries. In this instance, we mark all recorded prices that deviate from the normal non-fraudulent prices and additionally add all order amounts that are needed to identify the prices as too high for the ordered amounts. While kickback frauds show similar characteristics with too high purchase prices or too low sales prices per material for Invoice and Selling Kickbacks respectively, the relevant prices and amounts are recorded in different columns within the ERP system. Lastly, in Corporate Injury, reoccurring employee wages are higher than normal, which is only noticeable in prices which do not match the normal behavior observed in the remaining data. Overall, the identification of indicative ERP system entries allows the individual agents and processes to automatically provide an explanation for generated fraudulent data in the form of a binary ground truth explanation, where 1 to 5 features are highlighted per fraud case.

By adding an additional explanation logging module into the MAS-based simulation, we are able to provide a full mapping of fraud-related and unrelated features for each case of occupational fraud generated during the simulation run, and output a resulting explanation file that includes a binary ground truth for anomalous features for each generated fraud case.

3.3 Explainer Optimization Framework

Having extended the MAS-based simulation with ground truth explanations, we now propose a framework that enables the optimization of explainable fraud detectors for an unlabeled company dataset by leveraging the simulation to generate auxiliary data with both fraud detection labels and ground truth explanations. The entire framework is also visualized in Figure 1. As using an MAS in our setting requires adapting the simulation to the actual data provided by a company to follow data characteristics such as price trends, products, and purchase quantities, we follow our previous work [22] to extract those characteristics and integrate them into the environment of the MAS-based simulation. Note that more complex processes used within the underlying company might require modeling additional business processes and process derivations within

the simulation as well, which can be incorporated as additional agent behavior to further improve the simulation's ability to mimic the underlying company. This priming on real data allows the simulation to closely mimic the business scenario of a company, while additionally providing fraud detection labels and ground truth explanations that are missing in the original company data.

Using this labeled data, we can identify well-performing preprocessing methods, fraud detection models, and detector hyperparameters on the synthetic MAS data. Additionally, our proposed extension with ground truth explanations enables us to expand the evaluation to the explainer stage of our framework. After the best fraud detector is identified on the MAS-generated data, we apply multiple post-hoc feature relevance explainers and hyperparameter configurations on the previously identified best fraud detection model and select the most suitable explainer on the MAS-generated data by comparing the explainer outputs to the ground truth explanations on the feature level. Having identified the best fraud detector and explainer on the synthetic MAS data, we apply them directly to the original unlabeled company data in a completely unsupervised fashion.

3.4 Integrated XAI Methods

Our framework allows the integration of existing post-hoc feature relevance explainers to provide reasoning for the decision of the used fraud detector. Formally, for a given d-dimensional datapoint $x \in \mathcal{X} \subseteq \mathbb{R}^d$, a feature relevance explainer f provides a relevance score to each input feature that expresses the feature's relevance to the overall model decision through $f : \mathcal{X} \to \mathbb{R}^d$. As these explanations may be obtained through multiple explainers that exhibit varying performance dependent on the underlying data [21], we design the framework as open as possible to integrate many post-hoc explainers.

[21] provides an overview of post-hoc feature relevance explainers currently used in anomaly detection, structuring them into four categories based on their reliance on the underlying data and the internals of the underlying model to explain. Data-specific explainers, which only rely on training data and internally train a new anomaly detector, are omitted in our implementation as they produce explanations that do not reflect the decision process of the framework's anomaly detector. We, however, implement all perturbation-based explainers, which only access a detectors input and output interface and are therefore fully model-agnostic, as well as all gradient-based explainers, that leverage detector gradients of the output with respect to the input and therefore enforce differentiability of the detector as only requirement. Lastly, we omit model-specific explainers, as they require a specific fraud detection architecture. We, however, note that these explainers can also be integrated in our modular framework as long as the choice of the fraud detector is limited to the corresponding architecture needed for the explainer. In total, we implement the following six post-hoc feature relevance explainers into our framework: The perturbation-based explainers LIME [14] and SHAP [12] are entirely model-agnostic explainers that repeatedly perturb a datapoint by replacing regions with alternate values and reason about relevant features by monitoring the model's output when feeding in the perturbed data.

The gradient-based explainers Saliency [17], Gradient×Input [16], Integrated Gradients (IG) [18], and Layer-wise Relevance Propagation (LRP) [2] all use slight variations of the gradient of the output with respect to the input to identify features where small changes strongly affect the model output.

As [21] showed that some explainers have hyperparameters that strongly effect the resulting explanations, we also include an option to test different explainer hyperparameters in our framework. SHAP allows the user to define the reference values that are used to replace feature values in the datapoint during perturbations. Tested approaches to obtain these reference values in anomaly detection are using cluster centroids of k-means clustering (k-means) or the mean of training data (mean), as well as using a zero vector (zeros) or using a gradient-based reference value generation approach designed by [19] (lopt) [21]. Integrated Gradients also allows for choosing a reference value that is used to take gradients at multiple points between the point to explain and the reference, thus adding additional information regarding the vicinity around the point itself. We use the mean of training data (mean), the zero vector (zeros), as well as the optimized variant by [19] (lopt) in our experiments. For more in-depth information on specific explainers and explainer hyperparameters we refer to [21].

3.5 Optimization of XAI Methods

The MAS-based simulation enables the generation of binary ground truth explanations, that indicate for each feature of a fraudulent datapoint whether the feature is indicative of the underlying fraudulent activity or can be seen as normal data entry. Using this binary ground truth, established metrics may be used to evaluate feature relevance explanations from different explainers. Area-under-the-receiver-operator-characteristic (ROC) scores can be used to evaluate how highly the indicative features of the binary ground truth rank within the feature relevance scores of a single datapoint explanation [9], giving a metric that rates how well the highest scoring features according to the explanation correspond to truly indicative features. Cosine similarity (COS) is another established metric for comparing feature relevance scores of a single datapoint to ground truth [11], and provides information on how well the relevance scores match the entire ground truth. Both metrics have been used to evaluate feature relevance explanations against binary ground truth by individually applying them to all available datapoints to explain and afterwards aggregating the results to obtain a single score of explanation quality [9,11]. Following previous experiments with binary explanation ground truth [21,24], we use the ROC score as main metric to determine the best explainers according to the synthetic MAS-generated data.

4 Experiments

In this section, we apply our framework for explainable fraud detection and evaluate the transferability of explanation quality from synthetic MAS-generated data to unlabeled data. We generate data and ground truth explanations with

Table 1. Number of transactions and fraud cases of ERPSim and generated MAS data. Note that ERPSim frauds are not labeled during MAS data generation and anomaly detector training and detection and explanation labels are only used for final evaluation.

Dataset	Transactions	Frauds with Explanations	Invoice Kickback	Selling Kickback	Non-cash Larceny	Corporate Injury
ERPSim(1)	36778	50	24	0	22	4
MAS(1)	92985	0	0	0	0	0
MAS(1)^{fraud}	93356	223	51	104	66	2
ERPSim(2)	37407	86	30	0	48	8
MAS(2)	59378	0	0	0	0	0
MAS(2)^{fraud}	64858	187	51	102	34	0

the extended MAS simulation for two occupational fraud detection datasets, and evaluate whether optimizing explainer and hyperparameter choices transfer from the generated data to the underlying unlabeled company data.

4.1 Experimental Setup

Our goal is to evaluate whether strong explainer performance on MAS-generated data translates to the underlying ERP data used to adapt the simulation. While the underlying company data is treated as entirely unlabeled during the adjustment of the MAS-based simulation, existing fraud labels and ground truth explanations for the data may be used after the fraud detectors and explainers are adjusted by the simulation, in order to test whether the simulation was able to identify detectors and explainers that indeed perform well on the underlying company data. As this evaluation requires company data with labeled fraud cases and explanations, we choose the ERPSim data from [20] that contains both known and labeled fraud cases and ground truth explanations.

We use the two datasets ERPSim(1) and ERPSim(2) as used in [22] that constitute two individual fiscal years of a make-to-stock cereal production company to undertake two complete evaluation runs of our framework. The MAS-based simulation, implemented in the Java Agent DEvelopment framework (JADE) [3], is used to generate both clean training data without frauds for training detectors

Table 2. Anomaly detection performance of the trained autoencoder for the two runs on MAS(1) (MAS(2)) and ERPSim(1) (ERPSim(2)) respectively. Reporting mean average precision (PR) score on synthetic data, which was used for parameter selection, and on ERPSim where selected parameters are simply applied. Results show competitive detection performance as also discussed in [22].

run	PR_{MAS}	PR_{ERPSim}
run(1)	17.5 ± 2.6	26.5 ± 4.3
run(2)	16.6 ± 1.0	53.7 ± 5.7

and contaminated validation data with ground truth explanations for optimizing anomaly detectors and explainers. The resulting datasets are illustrated in Table 1, with data generation on a laptop taking less then one hour per dataset. As we focus our evaluation on the newly introduced XAI component of the framework, we limit the choice for the anomaly detector to the autoencoder neural network [8] and only optimize its data preprocessing and hyperparameters in these experiments. We specifically select the autoencoder, since it already showed high detection performance in occupational fraud detection [15,22] and allows the use of gradient-based explainers due to its differentiability. The anomaly detection performance of the optimized autoencoders on both underlying company data and the MAS-simulated datasets is shown in Table 2, with the performance of MAS-optimized models transferring well to the underlying unlabeled company data as previously observed in [22].

Using the fully trained autoencoders, we apply multiple post-hoc feature relevance explainers with varying explainer hyperparameters as described in Section 3.4 to explain the fraud cases contained in the synthetic MAS data. After obtaining explanations from all explainers with associated hyperparameter configurations, we identify the most suited explainer by comparing explanations to the ground truth explanations generated by the MAS through the XAI metrics described in Section 3.5. Having identified the best fraud detector and explainer, we apply them to the original ERPSim data in an unsupervised fashion. Finally, we run the explainers on the applied detector and evaluate them using the ground truth explanations of the ERPSim data to assess whether the proposed MAS-based optimization provides good explainability on a given unlabeled dataset.

5 Results

Explainers are applied to both the optimized autoencoder trained on the MAS data, as well as the applied autoencoder on the ERPSim data to obtain expla-

Table 3. Mean and standard variation of explainer performance on ERPSim(1) and the simulated MAS(1) data. Best and second-best results highlighted in bold and underline.

Explainer	reference	$\text{ROC}_{\text{XAI}}^{MAS(1)}$	$\text{COS}_{\text{XAI}}^{MAS(1)}$	$\text{ROC}_{\text{XAI}}^{ERPSim(1)}$	$\text{COS}_{\text{XAI}}^{ERPSim(1)}$
Saliency		51.3 ± 25.2	-1.7 ± 18.5	62.3 ± 19.6	13.4 ± 21.5
Gradient×Input		71.1 ± 26.2	33.6 ± 39.3	$\mathbf{84.7 \pm 12.8}$	$\mathbf{59.9 \pm 16.9}$
LRP		55.2 ± 28.3	2.0 ± 39.9	59.7 ± 15.0	20.2 ± 16.7
IG	mean	71.8 ± 16.9	$\underline{37.6 \pm 25.4}$	65.1 ± 12.6	18.1 ± 22.8
IG	zeros	57.1 ± 24.8	-14.6 ± 24.7	62.4 ± 14.5	-15.9 ± 17.3
IG	lopt	$\underline{73.1 \pm 19.2}$	30.6 ± 26.3	80.7 ± 15.6	47.4 ± 19.6
LIME	k-means	69.6 ± 7.4	16.8 ± 4.2	63.1 ± 10.5	22.6 ± 14.1
SHAP	k-means	55.8 ± 23.1	11.7 ± 35.7	63.6 ± 15.8	35.7 ± 14.6
SHAP	mean	68.3 ± 20.5	31.5 ± 30.4	57.3 ± 18.0	18.1 ± 29.0
SHAP	zeros	59.9 ± 20.1	-13.8 ± 20.0	63.8 ± 14.8	-18.9 ± 12.7
SHAP	lopt	$\mathbf{82.6 \pm 13.9}$	$\mathbf{49.4 \pm 29.1}$	$\underline{83.7 \pm 12.8}$	$\underline{57.7 \pm 21.7}$

nations for each explainer and explainer hyperparameter configuration. Then, we evaluate the quality of the resulting explanations using the respective ground truth, observing whether the MAS data is sufficient to identify highly performing explainers that transfer well to the ERPSim data.

At first, we examine the explanation performance on the MAS(1) data and the underlying ERPSim(1) company data in Table 3. When observing the general explanation scores across all configurations, our results show that a high performance on MAS(1) translates well to the underlying ERPSim(1) data. Especially for the SHAP and IG explainers different reference values can cause both high and low explanation quality, matching existing results that explanations are highly sensitive to reference values [21]. The optimized (lopt) references obtain the highest scores on the synthetic MAS(1) data for both SHAP and IG and would therefore be selected by our framework. These explainer choices transfer well to the ERPSim(1) data, with both choices achieving very high explanation scores. While our framework is clearly able to identify well-performing explainers, it does not find the explainer with the highest overall performance on the ERPSim(1) data, which is the Gradient×Input method in this instance, indicating that this method does not transfer from the simulated fraud cases to the cases contained in ERPSim. Nevertheless, we observe that explainers that perform highly on the synthetic MAS(1) data, and would therefore be selected for use on the underlying ERPSim(1) data in our framework, also consistently achieve high explanation scores on the ERPSim(1) data.

Even more importantly, our analysis reveals the absence of configurations that perform well on MAS(1) yet exhibit complete failure on ERPSim(1). This mitigates concerns regarding the framework producing misleading explanations when applied to actual company data. Our analysis indicates that, while pinpointing the single most effective explainer on the company data remains challenging, the ability of our framework to identify highly effective explainers nonetheless represents a significant advancement. This outcome is particularly valuable

Table 4. Mean and standard variation of explainer performance on ERPSim(2) and the simulated MAS(2) data. Best and second-best results highlighted in bold and underline.

Explainer	reference	$\text{ROC}_{\text{XAI}}^{MAS(2)}$	$\text{COS}_{\text{XAI}}^{MAS(2)}$	$\text{ROC}_{\text{XAI}}^{ERPSim(2)}$	$\text{COS}_{\text{XAI}}^{ERPSim(2)}$
Saliency		52.4 ± 18.6	-1.8 ± 16.9	61.4 ± 11.9	6.9 ± 10.5
Gradient×Input		61.1 ± 30.4	17.9 ± 42.1	86.7 ± 15.4	$\underline{65.2 \pm 18.0}$
LRP		66.3 ± 30.9	28.7 ± 51.1	66.3 ± 19.9	28.0 ± 22.8
IG	mean	65.9 ± 32.8	30.5 ± 44.9	74.4 ± 19.4	34.0 ± 28.7
IG	zeros	67.6 ± 21.4	0.7 ± 18.1	71.1 ± 12.9	-5.2 ± 10.6
IG	lopt	73.1 ± 21.1	37.2 ± 24.3	$\mathbf{91.1 \pm 6.1}$	54.7 ± 15.2
LIME	k-means	55.3 ± 9.8	15.7 ± 6.2	75.0 ± 6.6	31.3 ± 9.0
SHAP	k-means	60.8 ± 34.6	1.2 ± 49.3	76.9 ± 13.0	46.3 ± 11.2
SHAP	mean	75.8 ± 25.6	$\underline{38.4 \pm 36.9}$	69.2 ± 24.6	34.7 ± 38.8
SHAP	zeros	$\underline{77.4 \pm 25.7}$	-0.0 ± 20.0	75.2 ± 12.1	-8.0 ± 12.9
SHAP	lopt	$\mathbf{86.5 \pm 16.3}$	$\mathbf{61.7 \pm 28.1}$	$\underline{89.3 \pm 10.0}$	$\mathbf{66.9 \pm 19.1}$

given that, in the absence of our framework, selecting explainers and reference values for company data in real-world applications would be largely arbitrary, owing to the absence of ground truth. Observing our evaluation results, such an arbitrary choice of explainers and their hyperparameters could prove detrimental in this scenario, as our results on ERPSim(1) include explainers that perform close to entirely random noise (50.0 ROC and 0.0 COS) e.g. when using the LRP explainer or the SHAP explainer with mean reference values.

In the analysis of the second dataset, ERPSim(2), and the correspondingly generated MAS(2) data, as detailed in Table 4, we observe consistent overall patterns in the performance of explainers and the transferability of explanation scores between the two datasets. While the best explainer choice according to the MAS(2) data again does not achieve the best total scores on the ERPSim(2) data on all metrics, it still manages to perform very well compared to all other explainer choices. Other high performing explainer choices on the MAS(2) data also manage to score well on the ERPSim(2) data, avoiding particularly low explanation scores on ERPSim(2), as seen with the Saliency explainer.

Overall, our framework is able to successfully identify explainer and hyperparameter choices that provide high quality explanations on unlabeled occupational fraud detection datasets through the use of MAS-based data generation.

6 Conclusion

In this study, we introduced a solution to address the complexities of selecting optimal feature relevance explainers and their corresponding hyperparameters within unlabeled occupational fraud detection datasets. Our methodology extends prior research on MAS-based data generation for modeling company data by reflecting the economic state and business processes of a given company. Specifically, we advanced the simulation capabilities to generate ground truth explanations for fraud instances addressing the lack of validation of post-hoc feature relevance explainers particularly in the setting of anomaly detection with imbalanced, rare and unlabeled instances such as fraud. The efficacy of our framework was validated on two fraud detection datasets, where it consistently showed its capabilities to identify well performing explanation methods. At the same time, our findings underscore the importance of a methodological selection process such as our proposed framework for choosing suitable explainers, as arbitrary explainer choices may result in near random explanations. The application of MAS-based data generators has proven to be a useful approach in deriving ground truth explanations, thereby contributing to the improvement and validation of explainable fraud detection methods. While our proposed work relies on economics research to construct the simulation and is therefore currently limited to the task of occupational fraud detection, future research might be directed towards further application domains of XAI where foundations on data generation are available to fuel simulations. In addition, the positive results from our experiments justify further evaluation of our framework for explainable occupational fraud detection directly within company operation in future research.

References

1. ACFE: Occupational Fraud 2022: A Report to the nations (2022), https://legacy.acfe.com/report-to-the-nations/2022/, [Online; accessed 23. Jan. 2024]
2. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one **10**(7), e0130140 (2015)
3. Bellifemine, F., Bergenti, F., Caire, G., Poggi, A.: Jade—a java agent development framework. Multi-agent programming: Languages, platforms and applications pp. 125–147 (2005)
4. Ciatto, G., Calegari, R., Omicini, A., Calvaresi, D.: Towards XMAS: eXplainability through Multi-Agent Systems. In: AI&IoT@AI*IA (2019)
5. Ciatto, G., Schumacher, M.I., Omicini, A., Calvaresi, D.: Agent-Based Explanations in AI: Towards an Abstract Framework. In: Explainable, Transparent Autonomous Agents and Multi-Agent Systems. pp. 3–20. Lecture Notes in Computer Science, Springer International Publishing, Cham (2020)
6. Domínguez, R., Cannella, S., Framinan, J.M.: SCOPE: A Multi-Agent system tool for supply chain network analysis. In: IEEE EUROCON 2015 - International Conference on Computer as a Tool (EUROCON). pp. 1–5 (Sep 2015)
7. Fuchs, A., Fuchs, K., Gwinner, F., Winkelmann, A.: A Meta-Model for Real-Time Fraud Detection in ERP Systems. In: HICSS (2021)
8. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016)
9. 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)
10. Heuillet, A., Couthouis, F., Díaz-Rodríguez, N.: Collective eXplainable AI: Explaining Cooperative Strategies and Agent Contribution in Multiagent Reinforcement Learning With Shapley Values. IEEE Computational Intelligence Magazine **17**(1), 59–71 (2022)
11. Kauffmann, J., Ruff, L., Montavon, G., Müller, K.R.: The clever hans effect in anomaly detection. arXiv preprint arXiv:2006.10609 (2020)
12. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems. pp. 4765–4774 (2017)
13. Ohana, J.J., Ohana, S., Benhamou, E., Saltiel, D., Guez, B.: Explainable AI (XAI) Models Applied to the Multi-agent Environment of Financial Markets. In: Explainable and Transparent AI and Multi-Agent Systems. pp. 189–207. Springer (2021)
14. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?: Explaining the predictions of any classifier. In: 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data mining. pp. 1135–1144. ACM (2016)
15. Schreyer, M., Sattarov, T., Schulze, C., Reimer, B., Borth, D.: Detection of accounting anomalies in the latent space using adversarial autoencoder neural networks. In: 2nd KDD Workshop on Anomaly Detection in Finance. ACM (2019)
16. Shrikumar, A., Greenside, P., Shcherbina, A., Kundaje, A.: Not just a black box: Learning important features through propagating activation differences. arXiv preprint arXiv:1605.01713 (2016)
17. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. In: ICLR (2014)
18. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: 34th Int. Conf. on Machine Learning-Volume 70. pp. 3319–3328. JMLR. org (2017)

19. Takeishi, N., Kawahara, Y.: A characteristic function for shapley-value-based attribution of anomaly scores. Transactions on Machine Learning Research (2023)
20. Tritscher, J., Gwinner, F., Schlör, D., Krause, A., Hotho, A.: Open ERP System Data For Occupational Fraud Detection (Jul 2022)
21. Tritscher, J., Krause, A., Hotho, A.: Feature relevance XAI in anomaly detection: Reviewing approaches and challenges. Frontiers in Artificial Intelligence **6** (2023)
22. Tritscher, J., Roos, A., Schlör, D., Hotho, A.: Occupational Fraud Detection through Agent-based Data Generation. In: Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Communications in Computer and Information Science (2023)
23. 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)
24. Tritscher, J., Wolf, M., Hotho, A., Schlör, D.: Evaluating feature relevance xai in network intrusion detection. In: World Conference on Explainable Artificial Intelligence. pp. 483–497. Springer (2023)
25. Yepmo, V., Smits, G., Pivert, O.: Anomaly explanation: A review. Data & Knowledge Engineering **137**, 101946 (2022)