Towards Explainable Occupational Fraud Detection

Julian Tritscher[1], Daniel Schlör[1], Fabian Gwinner[2], Anna Krause[1], and Andreas Hotho[1]

[1] University of Würzburg, Am Hubland, 97074 Würzburg, Germany
{tritscher, schloer, anna.krause, hotho}@informatik.uni-wuerzburg.de
[2] fabian.gwinner@uni-wuerzburg.de

Abstract. Occupational fraud within companies currently causes losses of around 5% of company revenue each year. While enterprise resource planning systems can enable automated detection of occupational fraud through recording large amounts of company data, the use of state-of-the-art machine learning approaches in this domain is limited by their untraceable decision process. In this study, we evaluate whether machine learning combined with explainable artificial intelligence can provide both strong performance and decision traceability in occupational fraud detection. We construct an evaluation setting that assesses the comprehensibility of machine learning-based occupational fraud detection approaches, and evaluate both performance and comprehensibility of multiple approaches with explainable artificial intelligence. Our study finds that high detection performance does not necessarily indicate good explanation quality, but specific approaches provide both satisfactory performance and decision traceability, highlighting the suitability of machine learning for practical application in occupational fraud detection and the importance of research evaluating both performance and comprehensibility together.

Keywords: Fraud detection · Anomaly detection · XAI · ERP.

1 Introduction

As a study by the Association of Certified Fraud Examiners shows, occupational fraud, such as theft of materials or abuse of permissions by employees, is estimated to cause average losses of around 5% of an organization's revenue each year [1]. Digitization of business operation, for example in Enterprise Resource Planning (ERP) systems, unifies business processes and provides a standardized data base, which also opens up new possibilities for automated occupational fraud detection with Machine Learning (ML) [38, 37, 47].

To qualify for practical use, fraud detection systems must, on the one hand, accurately detect fraud and, on the other hand, provide comprehensible suggestions and decisions [19]. Consequently, prior studies on ML-based fraud detection name explainability explicitly as requirement [10, 14] and future research [20].

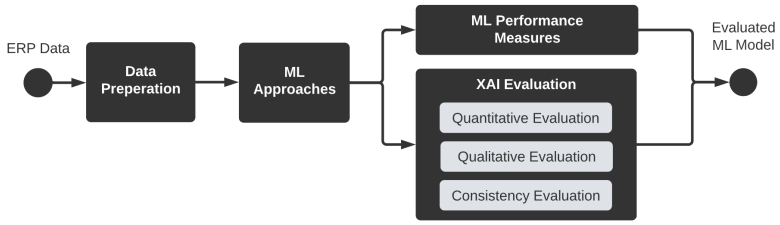


Fig. 1: Experimental setup for the evaluation of explainable occupational fraud detection in ERP system data.

However, for many state-of-the-art ML techniques, especially newly popular deep learning approaches, their high precision is attributed to a non-linear decision function that makes explaining their decision process non-trivial and turns them into a non-transparent black-box. This black-box nature is problematic for detecting fraud in ERP systems, where applying algorithms that do not act on reasonable fraudulent characteristics can introduce major consequences for potential wrongly suspected persons, in addition to ethical and legal requirements regarding privacy, transparency, and antidiscrimination [13, 18].

The research discipline of eXplainable Artificial Intelligence (XAI) has developed several approaches towards explaining a model's decision and finally gaining insight into the decision process of black box models [4]. The question of whether ML approaches provide strong performance in detecting occupational fraud in ERP systems while maintaining a comprehensible decision process when used in combination with XAI, however, has not yet been answered in current research to the best of our knowledge.

We therefore construct a three-fold evaluation setting to assess the explainability of ML based occupational fraud detection approaches on ERP data based on quantitative, qualitative and consistency criteria. We then conduct extensive experiments on the detection performance and explainability of different fraud detection approaches, as outlined in our experimental setup depicted in Figure 1: We combine five different ML-based fraud detection approaches, two classical anomaly detection approaches (One-Class SVM [36] and Isolation Forest [21]), a linear approach (Principle Component Analysis based Anomaly Detection [40]), and two task specific Deep Learning based approaches (Autoencoder [15] and improved Neural Arithmetic Logic Units [34]), with a state-of-the-art XAI component (Shapley Additive exPlanations [22]) providing explanations for the underlying decision process of each ML model. We evaluate the ML performance and the explanations generated from each fraud detection approach based on our three-tier XAI evaluation on ERP system data that includes fraudulent transactions with labeled explanations. As datasets with labeled explanations are rare and costly to obtain, we conduct an additional experiment investigating whether our measured detection performance and explanation quality are transferable to other datasets without the need for repeating hyperparameter studies that may require expensive labeling procedures in practice.

Our experiments identify ML-based fraud detection approaches that are capable of strong fraud detection performance on ERP system data, while also acting upon reasonable fraudulent characteristics. However, our experiments also show that strong detection performance does not automatically result in good explanation quality, suggesting that joint evaluation of model performance and explainability is essential in applications that require comprehensible decisions.

Our contributions can be summarized as follows:

- We construct a three-fold evaluation setting for the explainability of occupational fraud detection.
- We evaluate multiple ML approaches with respect to their detection performance and explainability through our evaluation scheme.[3]
- We assess whether our results obtained transfer to new datasets without the need for expensive labeling procedures.

The remaining paper is structured as follows: Section 2 gives an overview on occupational fraud detection and ML explainability. Section 3 introduces data preparation, ML approaches, and ML evaluation measures used in this study. Section 4 describes the constructed XAI evaluation setting. Section 5 presents our experiments and discusses results, while Section 6 concludes the paper.

2 Related Work

2.1 Fraud Detection

In the field of financial fraud detection, traditional ML, deep learning, and anomaly detection methods, as well as approaches relying on expert knowledge and auditors' views, have been subject to studies on both synthetic and real-world datasets. There are several surveys summarizing findings specifically from the financial domain [44] or from the broader field of anomaly detection, including financial fraud detection as an application domain [8, 9]. While many of these works utilize machine learning to detect frauds such as credit card misuse [23, 32], there are also works that focus on detecting occupational fraud within ERP system data.

Earlier work on fraud detection in ERP systems leverages statistical, visual, or clustering based approaches to detect frauds in database logs and transactional data [26, 30, 41, 43, 42]. Process mining, sometimes extended with filtering and rule based approaches, is another promising approach [5, 33, 24, 12]. Instead of transactional data, process mining uses event logs as main data source, which introduces further layers of abstraction with the creation of process graphs.

Since the rise of neural networks, an increasing number of publications directly utilize ERP data. Multiple works proposed autoencoder neural network architectures for fraud detection directly on transactional data [38, 37, 47], or conducted case studies that provide empirical evidence for the practical application of autoencoder architectures on real data [39, 25]. These approaches either

[3] Our code is available under https://professor-x.de/xai-erp-fraud

limit the ERP tables to few discriminative attributes used in their approach, or directly rely on feature engineering to create audit-relevant aspects of entries through domain knowledge. In contrast, our approach does not rely on explicit modeling of domain knowledge or process mining. Instead, we focus on raw ERP data without extensive feature selection or feature engineering.

2.2 Explainable Artificial Intelligence

With the rise of deep learning, assessing the decision process of non-transparent black-box models has become a core field of research in the ML community, which has been categorized by Arrieta et al. [4].

In the domain of fraud detection, multiple works have focused on detecting and explaining fraud caused by malicious credit card transactions. In addition to work proposing inherently explainable network architectures [48], multiple works investigate the use of popular post hoc feature relevance XAI approaches [22, 29] in detecting credit card fraud [3, 28]. Post hoc feature relevance XAI explains an already trained ML model by finding the impact of each input feature toward the model's final decision. This allows insight into the reasoning behind single model decisions, which has been identified in prior studies as desirable property for ML-based fraud detection in general [10] and in occupational fraud detection on ERP system data [14]. Due to their promising results in the related domain of credit card fraud detection, we follow Antwarg et al. [3] and Psychoula et al. [28] in utilizing a post hoc feature relevance approach to obtain explanations, but propose a different XAI evaluation scheme: Rather than performing XAI evaluations on artificial data [3] and comparing them to simple linear models [28], we construct an evaluation on expert-labeled ground truth and derive requirements for consistent XAI decisions in ERP system data. To our knowledge, we are the first to investigate the performance of feature relevance XAI in the domain of occupational fraud detection in ERP system data.

3 Machine Learning Methodology

This section introduces the data preprocessing schemes, ML approaches, and ML performance metrics used in our study.

3.1 Data Preprocessing

With transactions in ERP systems that contain sparse information in many columns, manual feature extraction in combination with the feedback of business experts may seem like a promising approach for detecting common and known fraud cases. However, we argue that in a live setting, attackers can continuously create new and previously unseen frauds, which can only be detectable through additional information contained in sparse columns of the ERP system. This makes the ability of monitoring all available data appealing for a fraud detection system in a realistic setting. Therefore, we utilize established preprocessing

techniques that largely retain the information contained in the ERP dataset and do not require vast amounts of manual feature engineering.

Categorical Columns are transformed by one-hot encoding. We further add a column for empty values, retaining the information of a column within the ERP data being left empty and allowing us to distinguish between empty columns and column entries that have not been observed during training time.

Numerical Columns can cause problems for many ML approaches due to large value ranges, which is problematic in a domain where monetary amounts or quantities vary from single digits to figures in millions. We test multiple established scaling techniques on numerical ERP system data, implementing z-score and minmax scaling [27], as well as quantization which transforms numerical values into categorical buckets. To highlight outliers within the data, we adapt the quantization technique to first choose two buckets that include the 1% highest and lowest numerical values and then choose buckets that equally distribute the remaining data. This allows the data representation to highlight unusually high or low values that may indicate fraudulent abuse of the system.

In our experiments, all preprocessing schemes are fitted purely on the training data and applied without fitting to both evaluation and test data. For the quantization scheme, we use 5 buckets since we observed decreasing performance with larger bucket sizes in preliminary testing.

3.2 ML Approaches

Common difficulties in the training and application of anomaly detection algorithms are the unavailability of anomalies during training time, and the diverse characteristics of potential anomalies [9]. These issues are further fortified in occupational fraud detection by a very high ratio of normal to anomalous datapoints and the motivation of fraudsters to create frauds that are highly diverse, novel, and difficult to detect. In this work, we therefore employ ML algorithms that exclusively train with normal data and are designed to detect anomalous datapoints that show deviating behavior. For our study, we investigate established deep learning approaches, classical anomaly detection approaches, and a linear model, which we introduce with the abbreviations and references used in Table 1. To unify the approaches, we make the following adjustments:

iNALU is used in an autoencoder setup with linear layers at the beginning and end and intermediate mixed layers that contain an even number of ReLU

Table 1: Utilized ML algorithms.

Approach	Description	Source
AE	Autoencoder neural network architecture with ReLU activation	[15]
iNALU	AE with ReLU and improved neural arithmetic logic unit activation	[34]
IF	Isolation Forest	[21]
OC-SVM	One-Class Support Vector Machine using rbf kernel	[36]
PCA	Anomaly detection using Principle Component Analysis	[40]

and improved neural arithmetic logic unit activations [35]. While IF and PCA have a direct anomaly scoring function, we use reconstruction loss to detect anomalies with AE and iNALU. For the OC-SVM, we utilize the signed distance from the datapoint to the hyperplane in feature space as our anomaly score.

3.3 ML Evaluation

Classification metrics such as precision, recall, and f-score are widely used in ML applications to assess model performance [6] but require a direct classification of transactions into normal or fraudulent datapoints. This, in turn, requires a fixed threshold value on the anomaly scores of our ML approaches. Since the optimal threshold choice depends on task, data, use case, ML approach, and possibly even the ML model's parameters used during training, setting this threshold is non-trivial and requires striking a balance between detection rate and the number of anomalies detected [7]. Area-Under-the-Curve (AUC) scores omit a threshold by calculating scores over varying threshold values. A popular choice for AUC scores, the well-known AUC Receiver-Operating-Characteristic (ROC) score is sensitive to class imbalance, which skews its results in highly unbalanced settings such as fraud detection. Therefore, we base our evaluation on the AUC Precision-Recall (PR) score which addresses this issue [11].

Furthermore, we report the rank of the least suspicious fraud r_{min}, which corresponds to the practical question of how many transactions would have to be inspected until all frauds are found. Mathematically, the rank of the least suspicious fraud of all frauds $F \subseteq X$ of dataset X is given as

$$r_{min} = |\{x \in X : score(x) \geq \min_{f \in F} score(f)\}| \tag{1}$$

where *score* denotes the anomaly scoring function of the detection approach that yields high values for anomalous samples.

4 XAI Methodology

To assess the decision process of different ML approaches during occupational fraud detection, we construct a three-fold evaluation process based on quantitative evaluation, qualitative inspection, and consistency testing.

4.1 XAI Approach

In previous work, the post hoc feature relevance approach Shapley Additive exPlanations (SHAP) [22] has been identified as XAI approach providing good comprehensibility both in the related area of credit card fraud detection [3, 28] and on categorical tabular data [45], which encompass a large number of columns within ERP system data. We therefore employ SHAP on individual model predictions to find which features are most relevant for the model's decision. SHAP utilizes game theory to find feature relevance by switching feature combinations

with background data and assessing the resulting behavior of the model. SHAP is model-agnostic and can be employed on any ML model. In this work, we employ SHAP's KernelSHAP method that generates background data through the centroids of k-means clustering with $k = 20$ clusters.

4.2 Quantitative XAI Evaluation

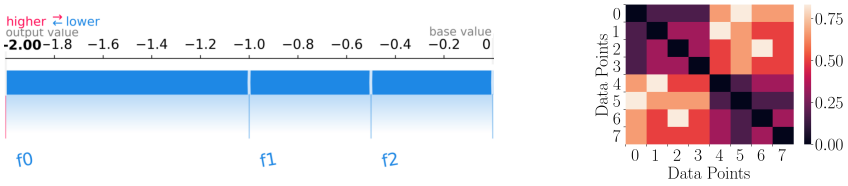
To measure the quality of the feature relevance explanations generated by SHAP, a quantitative evaluation measure is required. Samek et al. [31] propose a quantitative evaluation procedure which assumes that perturbing relevant feature entries leads to different model decisions. This evaluation may prove problematic in heavily unbalanced domains such as anomaly or fraud detection, as finding replacement values that form valid, non-anomalous datapoints is non-trivial. While Hooker et al. [17] extend this approach by proposing to retrain the entire model from scratch after perturbing the training data, their resulting scheme requires repeating extensive training steps and potential repetition of entire hyperparameter studies, which limits its use in practice.

To quantitatively evaluate XAI explanations in our occupational fraud detection scenario, we therefore propose an evaluation scheme based on suspicious data entries of fraud cases: In the evaluation, we focus on fraudulent datapoints, where feature entries that deviate from the normal business process can be seen as indicative of the fraud case. For a given fraud, this requires identification of indicative feature entries by auditing experts, which then serve as ground truth for quantitatively evaluating feature relevance explanations, where indicative feature entries should be rated as more relevant than normal features.

For the evaluation of explanation heatmaps through ground truth on image data, Hägele et al. [16] proposes an evaluation scheme based on ROC scores. We adopt this evaluation scheme for our tabular data, ranking the quality of single datapoint explanations through ROC score against our ground truth. For one datapoint, this results in a ranking score that increases when deviating data entries are given higher feature relevance within the explanation. Additionally, the ROC score is scale-invariant, focusing only on whether deviating entries are found before normal entries. For all fraud cases, we aggregate the individual ROC scores for each datapoint into an average ROC score and use this metric as a quantitative measure that represents how highly features with information concerning the fraud case rank in the given explanation.

4.3 Qualitative XAI Evaluation

Beyond examining individual metrics, we use qualitative inspection of explanations as a more in-depth assessment of XAI explanation quality. To qualitatively assess explanations, feature relevance of single datapoints may be visualized using SHAP's force plots, as seen in Figure 2a. Here, feature names are listed at the bottom of their corresponding bar, with a larger bar width corresponding to a feature's greater impact on the (negative) anomaly score shown above the bars (e.g. feature f0 showing a greater impact than the features f1 or f2).



(a) Exemplary SHAP force plot [22] (b) Exemplary explanation heatmap

Fig. 2: Demonstration plots of explanations for single datapoints (2a) and heatmaps to compare datapoint explanations (2b).

4.4 XAI Consistency Evaluation

Cirqueira, Helfert, and Bezbradica [10] discover that auditors check previously detected frauds and their explanations to assess new frauds, and define a requirement for XAI in fraud detection to allow comparison of frauds based on previous explanation patterns. To evaluate new cases with explanations based on historical patterns, similar frauds need to be consistent in their explanations.

Approaches that assess the consistency of explanations are currently limited to the robustness of XAI methods against adversarial attacks [2] and analyze whether minimal changes to a *single* datapoint maintain a similar explanation. In contrast, we propose an evaluation that focuses on the consistency of explanations of *similar* frauds. We construct a heatmap for fraud datapoints that can be used both to evaluate explanation similarity between different fraud cases and to find similar anomalies for currently evaluated fraud cases during the application.

Consider a feature relevance explanation $y \in \mathbb{R}^d$ for a fraudulent datapoint $x \in F$ within the fraudulent subset $F \subseteq X$ of dataset X with data dimensionality d. We first binarize y by applying a threshold of 25% of the highest relevance value, to focus our evaluation on highly relevant features and reduce noise.

$$\hat{y}_i = \begin{cases} 1\ , & y_i > 0.25 \cdot \max(y) \\ 0\ , & y_i \leq 0.25 \cdot \max(y) \end{cases} \tag{2}$$

We apply this transformation to the explanations of all fraudulent datapoints and compute the Manhattan distance pairwise as a measure of explanation similarity through $dist(\hat{y}, \hat{y}') = \|\hat{y} - \hat{y}'\|_1$ for a pair of binarized explanations $\hat{y}$ and $\hat{y}'$. We then arrange explanations so that similar frauds are grouped together and visualize the pairwise similarity of explanations in a heatmap. Therefore, consistent explanations for similar frauds are expected to form a block-like structure around the diagonal, as shown in Figure 2b for the datapoints 0-3 and 4-7, which features all similar fraud cases within one block of low pairwise distances.

5 Experiments

Using our XAI evaluation setting introduced in Section 4, we now evaluate multiple ML approaches introduced in Section 3 with respect to performance and comprehensibility when detecting occupational fraud in ERP system data.

Table 2: Total datapoints and fraud cases within the used ERP system data [46].

Group	Dataset	Transactions	Frauds	IK1	IK2	L1	L2	L3	L4	CI
group 1	normal 1	54677	0	0	0	0	0	0	0	0
group 1	fraud 1	39430	24	4	0	2	4	0	0	14
group 2	normal 2	32337	0	0	0	0	0	0	0	0
group 2	fraud 2	36778	50	6	18	2	4	10	6	4
group 2	fraud 3	37407	86	24	6	8	10	26	4	8

L=Larceny, IK=Invoice Kickback, CI=Corporate Injury

5.1 ERP system data

For our experiments, we use publicly available ERP system data that contains both normal business documents and fraudulent activities [46]. The data contains five distinct datasets obtained from data generation of a simulated production company with two participant groups, with two datasets consisting of completely normal operation and three datasets including different fraud cases within the normal business process. Frauds include different scenarios of material theft (larceny), bribery in material procurement (invoice kickback), and cause of malicious damage to the company (corporate injury), with details on specific fraud cases introduced in the original paper [46]. Table 2 gives a brief overview of the distribution of normal and fraudulent transactions within the data. Beyond the separation of transactions into normal or fraudulent behavior, these datasets also contain expert annotations of individual fraud cases. As all fraudulent transactions have marked column entries that correspond to the entries that are indicative of the underlying fraud, these annotations are used as ground truth for our quantitative XAI evaluation of Section 4.2.

5.2 Experiment 1: Explainable Occupational Fraud Detection

Prior research into fraud detection has found both high performance and comprehensibility to be desirable properties of detection approaches [14, 10, 19]. In our first experiment, we therefore evaluate multiple established ML approaches on occupational fraud detection in ERP system data. We conduct a hyperparameter study encompassing more than 1500 cpu core hours to assess the detection performance of the algorithms studied. We further generate explanations for these approaches through SHAP as described in Section 4.1, and analyze explanation quality through our XAI evaluation setting from Section 4 to discover approaches that deliver both high performance and satisfactory explanations.

Experimental Setup In this experiment, we focus on the ERP datasets generated by the second participant group (normal 2, fraud 2, fraud 3), as introduced in Section 5.1. We choose these datasets as they contain a larger amount and broader spectrum of fraud cases, and additionally offer two fraudulent datasets

Table 3: Best results of each approach on evaluation (1) and test (2) set.

approach	$PR^{(1)}$	$PR^{(2)}$	$ROC^{(1)}$	$ROC^{(2)}$	$r_{min}^{(1)}$	$r_{min}^{(2)}$
OC-SVM	0.34	0.73	0.99	1.00	1201.0	740.0
iNALU†	0.34	0.52	0.99	1.00	1769.0	1022.4
AE†	0.31	0.69	0.99	1.00	1615.0	825.0
IF†	0.19	0.49	0.99	0.99	2232.0	1046.0
PCA	0.08	0.12	0.82	0.91	36778.0	37407.0

†Non-deterministic: averaged over 5 seeds to mitigate statistical fluctuation

that can be used as separate validation and test datasets. As training data, we use the dataset normal 2, which only contains normal data and simulates training on records that have previously been audited. While training of our ML algorithms only requires normal data, all algorithms have additional hyperparameters that influence the detection rate. Therefore, an audited dataset containing fraudulent samples is required as evaluation dataset, which is potentially not available in practice. The necessity of this dataset will be assessed in the subsequent experiment in Section 5.3. In Experiment 1, the partially fraudulent dataset fraud 2 is used as evaluation set to select hyperparameters and the overall performance is evaluated on dataset fraud 3 as test set. This separation allows for tuning ML hyperparameters on an evaluation dataset with fraudulent transactions, while retaining an unseen test dataset for the evaluation of the resulting algorithms.

Parameter Search and Performance Results To assess detection performance, we utilize the metrics introduced in Section 3.3. To select the best performing hyperparameters, we rank architectures by PR score on the evaluation set. Table 3 shows the best results of each approach for both evaluation set (1) and test set (2), where we also report ROC and r_{min} denoting how many transactions would have to be audited to find all frauds using the detectors. We make the tested hyperparameters and results of individual runs available online for reproducibility[4]. Our findings can be summarized as follows:

For the linear PCA we find no parameter setting capable of reliably detecting fraudulent transactions, with even the best hyperparameters yielding poor detection results on all metrics. Although IF is capable of detecting fraud cases, it performs considerably worse than the remaining approaches in PR score. AE and OC-SVM both show very strong detection performance, with OC-SVM highlighting all fraud cases within 1201 and 740 suspected datapoints for the evaluation and test set, respectively. iNALU performs on par with AE and OC-SVM on the evaluation data, but detects fraud cases considerably later on the test set. Upon closer inspection, all well-performing approaches highlight the Larceny 4 and Corporate Injury frauds within the first anomalous transactions. Lowered scores are caused mainly by Larceny 3 and Invoice Kickback 1 frauds. This may be explained by the subtle and well-hidden nature of the two frauds. For Larceny

[4] Supplementary material under https://professor-x.de/xai-erp-fraud-supplementary

Table 4: Quantitative explanation evaluation for Experiment 1 (see Section 4.2).

approach	$\text{ROC}_{XAI}^{(1)}$	$\text{ROC}_{XAI}^{(2)}$
OC-SVM	0.542	0.579
iNALU	**0.642**	**0.794**
AE	0.603	0.658

3 only a small portion of materials is stolen and for Invoice Kickback 1 prices are increased only by a small percentage that may well be within the range of normal price fluctuations. As a result, while the approaches manage to find the frauds, detection occurs later than on cases with clearly identifiable characteristics such as items that have never been purchased before in Larceny 4.

Overall, we observe high performance for OC-SVM, iNALU and AE when detecting occupational fraud in ERP data.

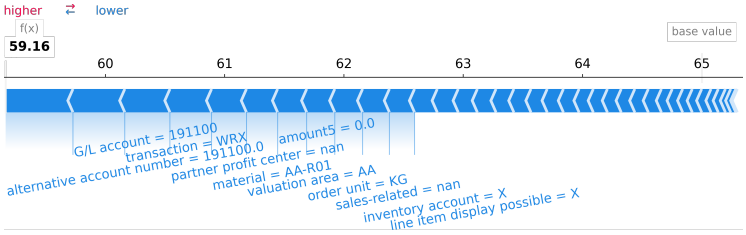
Model Explanation Results To evaluate, whether well performing detection systems can also provide a satisfactory decision process, we generate post hoc feature relevance explanations for the best performing OC-SVM, iNALU and AE approaches through the XAI approach SHAP as outlined in Section 3.3 and evaluate the resulting explanations through our XAI evaluation setting.

Quantitative Evaluation. To quantitatively assess the explanation quality of our trained models, we evaluate the explanations of fraudulent datapoints with ground truth as described in Section 4.2. Table 4 shows the quality of the explanation measured in the evaluation set (1) and the test set (2). The explanations for OC-SVM show the smallest similarity to the ground truth, in spite of its strong detection rate in our performance evaluation. While AE displays higher explanation quality, iNALU explanations produce the highest ROC scores.

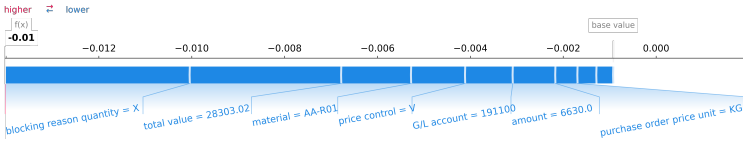
Qualitative Evaluation. To discover the reasons for this behavior, we qualitatively evaluate SHAP plots across all frauds from the test data. To illustrate the fraud visualization process, we show a non-cherry-picked explanation visualization of a fraud case from the test set fraud 3 in Figure 3.

In the Larceny 1 case shown in Figure 3, only iNALU focuses on the anomalous entry that marks the transaction as blocked by the ERP system (blocking reason quantity). While AE focuses on suspiciously small quantities ordered and in stock, OC-SVM highlights many features that are not related to fraud. Approaches also show sensitivity towards columns, such as G/L account, valuation class, or transaction, that describe the general transaction type (e.g. material entry, withdrawal). This may be caused by value combinations that are anomalous for the given transaction, but characteristic of another transaction type, causing the transaction type to be seen as anomalous. OC-SVM is particularly sensitive to this behavior and highlights many transaction-type features that are not indicative of fraud. This pattern is also noticeable in other larceny frauds.

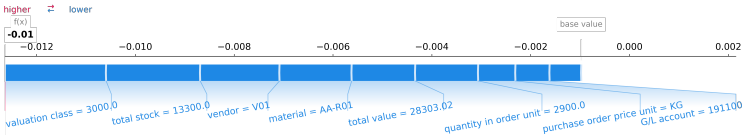
On invoice kickback frauds, where the fraudster's activity causes atypically high unit prices, both iNALU and AE highlight the amounts and quantities



(a) OC-SVM on Larceny 1 G/L account posting



(b) iNALU on Larceny 1 G/L account posting



(c) AE on Larceny 1 G/L account posting

Fig. 3: SHAP explanations on a fraudulent Larceny 1 transaction, showing feature influence through bar width. iNALU and AE focus on anomalous quantities and amounts, while OC-SVM reacts to a variety of features.

required for inference. While OC-SVM is sensitive to some amount columns, they carry only small influence over columns that are not related to fraud.

In the corporate injury scenario, fraudulent purchase activities result in high purchase amounts and purchase quantities. Here, iNALU is strongly sensitive to anomalous quantities and amounts, while AE additionally focuses on some not directly relevant columns such as vendor or material entries, and OC-SVM focuses on many columns that do not directly indicate fraudulent activities.

Overall, the qualitative observations are consistent with the quantitative results, indicating that AE and iNALU consistently show sensitivity to columns that are sufficient to explain and detect fraudulent transactions, with iNALU providing the best explanations. OC-SVM, despite its slightly stronger detection performance observed in Table 3, produces explanations that are noisy and difficult to interpret, potentially limiting its use in practice, when insights into the decision process or justifications are required.

Explanation Consistency. To evaluate the consistency of trained approaches when explaining similar anomalies, we create heatmap plots as described in Section 4.4. For both iNALU and AE, Figure 4 shows clear similarities between the explanation of transactions from the same fraud scenarios, indicating that both approaches react to fraud cases in a consistent way. While iNALU is capable of producing slightly sharper contrasts between similar and dissimilar fraud cases

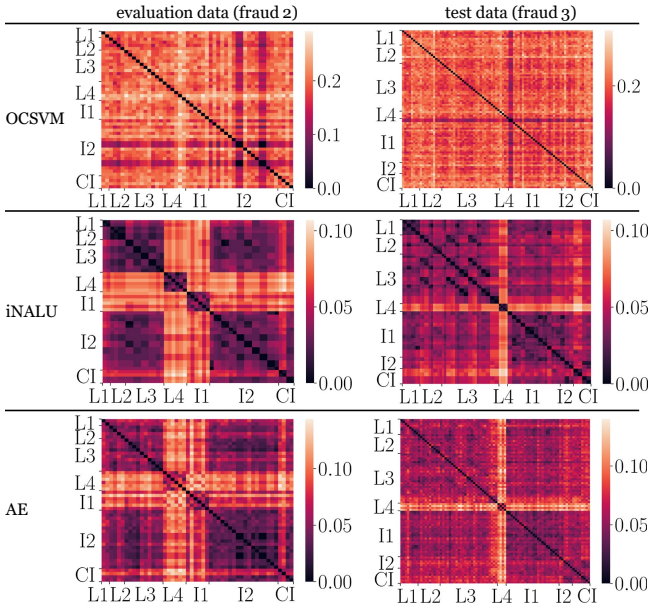


Fig. 4: Manhattan distance of SHAP explanations for larceny frauds (L1, L2, L3, L4), invoice kickback (I1, I2), and corporate injury (CI), ordered by type and time of occurrence. Plots show consistent explanations for iNALU and AE.

in comparison to AE, OC-SVM focuses on very different features even when considering very similar fraudulent data samples. Overall, both iNALU and AE show consistency in their decisions, which may be used to compare explanations with historical patterns of fraud cases.

5.3 Experiment 2: Performance After Retraining

Through the conducted hyperparameter study and evaluation of explanation quality, Experiment 1 has produced models and hyperparameter configurations that can explain fraud cases precisely and comprehensibly on the used dataset. For practical applications, however, annotated data for selecting models and parameters is usually not readily available. In particular, if the normal behavior within a company shifts (for example, due to changing employees or operational changes), the model must also be retrained on the new normal data to adopt for the changes. If an additional annotated dataset is necessary in these cases to re-evaluate the model and parameter selection, this would be associated with considerable costs that may prove prohibitive in practical applications. In this experiment, we therefore evaluate to what extent previously found optimal model parameters can also be transferred to a new dataset with changed business circumstances and different employees, investigating how stable the respective models are with regard to their hyperparameter selection on different datasets.

Experimental Setup We use the normal (normal 1) and fraudulent (fraud 1) game runs introduced in Section 5.1 that originate from different ERP system users than data from Experiment 1 and retrain our models with the hyperparameters that showed the best performance during our first experiment in Section 5.2. To gain further insights into whether re-evaluation of the approaches would have allowed for choosing a better performing set of hyperparameters for each approach, we also repeat our first experiment on the new dataset entirely. This allows us to rank our previously best performing parameter configurations in comparison to all other parameter settings that could have been chosen through reevaluation with regard to each of our evaluation metrics.

Stability Results For all the ML approaches evaluated, we report both the performance metrics, as well as the ranking of the hyperparameters of Experiment 1 compared to the optimal hyperparameter configurations in Table 5. We further use the feature-level fraud annotations within the dataset to repeat our quantitative explanation evaluation on the previously best performing detection systems. Although the linear PCA initially shows strong performance through a high PR score, the r_{min} metric reveals that the approach detected few frauds very early at the expense of completely failing to detect other fraud cases, resulting in a high PR score that is sensitive to this behavior. As all other approaches are detecting all fraud cases within a small number of audited datapoints, as shown in their r_{min} scores, PCA remains a non-desirable detection approach for this scenario. IF remains behind the three approaches that have been best performing previously on the r_{min} metric. The previously best performing AE and OC-SVM drop considerably in performance, with multiple better performing hyperparameter configurations as seen in both r_{min}_rank and PR_$rank$, which would require costly re-evaluation with an annotated evaluation dataset. iNALU, on the other hand, is capable of maintaining its performance on the dataset with a highly stable PR score and a very strong r_{min} score. Furthermore, there are only a few parameter configurations for iNALU that improve over the previous best parameter set on the r_{min} metric, with the lower rank on the PR score being caused by the sensitive PR score behavior discussed above.

Table 5: Results for retraining the best models found in Experiment 1. The rankings compare performance with other hyperparameter settings.

approach	PR	ROC	r_{min}	PR_$rank$	ROC_$rank$	r_{min}_rank	ROC_{XAI}
OC-SVM	0.16	1.00	228.00	50	10	17	0.582
iNALU†	0.34	1.00	146.60	46	5	6	0.859
AE†	0.21	1.00	253.20	36	34	19	0.820
IF†	0.22	1.00	366.40	187	66	31	0.783
PCA	0.44	0.95	19432.00	16	1	18	0.517

†Non-deterministic: averaged over 5 seeds to mitigate statistical fluctuation

Overall, this study highlights iNALU as an approach that, while showing slightly lower performance compared to AE and OC-SVM in Experiment 1, is capable of providing satisfactory decision traceability and additionally proving stable towards model retraining, making it a strong model choice within the domain of occupational fraud detection.

6 Conclusion

In this study, we investigated whether different ML approaches could provide strong performance and a satisfactory decision process in occupational fraud detection. We first constructed an evaluation setting for the explanations of occupational fraud detection approaches based on quantitative, qualitative, and consistency criteria. We then conducted extensive experiments on multiple fraud detection approaches combined with post hoc XAI, finding highly performing detection approaches through a hyperparameter study and assessing the quality of their decision process through the XAI evaluation setting. Further, we assessed whether ML approaches are capable of maintaining their performance and explanation quality on company data with changed underlying characteristics, by retraining and re-evaluating the approaches on an additional ERP dataset.

Our results indicate that high detection performance does not necessarily come with good explanation quality, as the OC-SVM approach displays a strong detection rate with poorly performing explanations. However, the AE and iNALU approaches provide satisfactory performance and decision traceability. Despite its lower detection performance compared to the AE, our second experiment reveals that iNALU is the more stable detection approach, managing to best retain its performance after retraining. Our findings demonstrate a possible strong performance and explanation quality of ML-based occupational fraud detection approaches and motivate the use of the investigated deep learning approaches for detecting occupational fraud in ERP system data.

In this work, we conducted a first broad evaluation on established ML-based detection approaches covering deep learning, anomaly detection, and linear models. With the promising results of our experiments, we plan to systematically extend our research to further detection architectures. Similarly, our explanation experiments conducted with an established and proven XAI algorithm could be extended to other types of explanations to provide additional comprehensibility in occupational fraud detection. With this study, we took a first step towards explainable ML-based occupational fraud detection systems on ERP system data, and encourage future research by highlighting the need to investigate detection performance and explainability in a joint fashion.

Acknowledgement

The authors acknowledge the financial support from the German Federal Ministry of Education and Research as part of the DeepScan project (01IS18045A).

References

[1] ACFE. "Occupational Fraud 2022: A Report to the nations". In: *Report To the Nations* (2022). [Online; accessed 01. Jun. 2022]. URL: https://legacy.acfe.com/report-to-the-nations/2022/.

[2] David Alvarez Melis and Tommi Jaakkola. "Towards robust interpretability with self-explaining neural networks". In: *Advances in neural information processing systems* 31 (2018).

[3] Liat Antwarg et al. "Explaining anomalies detected by autoencoders using SHAP". In: *arXiv preprint arXiv:1903.02407* (2019).

[4] Alejandro Barredo Arrieta et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information Fusion* 58 (2020), pp. 82–115.

[5] Galina Baader and Helmut Krcmar. "Reducing false positives in fraud detection: Combining the red flag approach with process mining". In: *Int. Journal of Accounting Information Systems* 31 (2018), pp. 1–16.

[6] Anna L Buczak and Erhan Guven. "A survey of data mining and machine learning methods for cyber security intrusion detection". In: *IEEE Communications surveys & tutorials* 18.2 (2015), pp. 1153–1176.

[7] Christian Callegari et al. "When randomness improves the anomaly detection performance". In: *2010 3rd International Symposium on Applied Sciences in Biomedical and Communication Technologies (ISABEL 2010)*. IEEE. 2010, pp. 1–5.

[8] Raghavendra Chalapathy and Sanjay Chawla. "Deep Learning for Anomaly Detection: A Survey". In: *arXiv preprint arXiv:1901.03407* (2019).

[9] Varun Chandola, Arindam Banerjee, and Vipin Kumar. "Anomaly Detection: A Survey". In: *ACM Computing Surveys (CSUR)* 41.3 (2009), pp. 1–58.

[10] Douglas Cirqueira, Markus Helfert, and Marija Bezbradica. "Towards Design Principles for User-Centric Explainable AI in Fraud Detection". In: *Artificial Intelligence in HCI*. Springer International Publishing, 2021, pp. 21–40. ISBN: 978-3-030-77772-2.

[11] Jesse Davis and Mark Goadrich. "The relationship between Precision-Recall and ROC curves". In: *Proceedings of the 23rd int. conf. on Machine learning*. 2006, pp. 233–240.

[12] Kadek Dwi Febriyanti, Riyanarto Sarno, and Yutika Amelia Effendi. "Fraud detection on event logs using fuzzy association rule learning". In: *2017 11th International Conference on Information & Communication Technology and System (ICTS)*. IEEE, Oct. 2017, pp. 149–154. DOI: 10.1109/ICTS.2017.8265661.

[13] Heike Felzmann et al. "Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns". In: *Big Data & Society* 6.1 (2019), p. 2053951719860542. DOI: 10.1177/2053951719860542. eprint: https://doi.org/10.1177/2053951719860542. URL: https://doi.org/10.1177/2053951719860542.

[14] Anna Fuchs et al. "A Meta-Model for Real-Time Fraud Detection in ERP Systems". In: *Proceedings of the 54th Hawaii International Conference on System Sciences*. 2021, p. 7112.

[15] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[16] Miriam Hägele et al. "Resolving challenges in deep learning-based analyses of histopathological images using explanation methods". In: *Scientific reports* 10.1 (2020), pp. 1–12.

[17] Sara Hooker et al. "A benchmark for interpretability methods in deep neural networks". In: *Advances in neural information processing systems* 32 (2019).

[18] Dimitra Kamarinou, Christopher Millard, and Jatinder Singh. "Machine learning with personal data". In: *Queen Mary School of Law Legal Studies Research Paper* 247 (2016).

[19] Felix Krieger, Paul Drews, and Patrick Velte. "Explaining the (non-) adoption of advanced data analytics in auditing: A process theory". In: *International Journal of Accounting Information Systems* 41 (June 2021), p. 100511. ISSN: 1467-0895. DOI: 10.1016/j.accinf.2021.100511.

[20] Johannes Lahann, Martin Scheid, and Peter Fettke. "Utilizing machine learning techniques to reveal vat compliance violations in accounting data". In: *2019 IEEE 21st Conference on Business Informatics (CBI)*. Vol. 1. IEEE. 2019, pp. 1–10.

[21] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. "Isolation forest". In: *2008 Eighth IEEE Int. Conf. on Data Mining*. IEEE. 2008, pp. 413–422.

[22] Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. 2017.

[23] Smita Prava Mishra and Priyanka Kumari. "Analysis of techniques for credit card fraud detection: a data mining perspective". In: *New Paradigm in Decision Science and Management*. Springer, 2020, pp. 89–98.

[24] Mohammad Farid Naufal. "Fraud detection using Process mining and analytical hierarchy process with verification rules on ERP business process". In: *International Conference on Informatics, Technology, and Engineering (InCITE)-2nd*. 2019.

[25] Jakob Nonnenmacher et al. "Using Autoencoders for Data-Driven Analysis in Internal Auditing". In: *Proceedings of the 54th Hawaii Int. Conf. on System Sciences*. 2021.

[26] William Ferreira Moreno Oliverio et al. "A Hybrid Model for Fraud Detection on Purchase Orders". In: *Intelligent Data Engineering and Automated Learning – IDEAL 2019*. Cham, Switzerland: Springer, Oct. 2019, pp. 110–120. ISBN: 978-3-030-33606-6. DOI: 10.1007/978-3-030-33607-3_13.

[27] S Patro and Kishore Kumar Sahu. "Normalization: A preprocessing stage". In: *arXiv preprint arXiv:1503.06462* (2015).

[28] Ismini Psychoula et al. "Explainable machine learning for fraud detection". In: *Computer* 54.10 (2021), pp. 49–59.

[29] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should i trust you?: Explaining the predictions of any classifier". In: *22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data mining*. ACM. 2016, pp. 1135–1144.

[30] Andrei Sorin Sabau. "Survey of clustering based financial fraud detection research". In: *Informatica Economica* 16.1 (2012), p. 110.

[31] Wojciech Samek et al. "Evaluating the visualization of what a deep neural network has learned". In: *IEEE transactions on neural networks and learning systems* 28.11 (2016), pp. 2660–2673.

[32] Marco Sánchez-Aguayo, Luis Urquiza-Aguiar, and José Estrada-Jiménez. "Fraud Detection Using the Fraud Triangle Theory and Data Mining Techniques: A Literature Review". In: *Computers* 10.10 (2021), p. 121.

[33] Riyanarto Sarno et al. "Hybrid Association Rule Learning and Process Mining for Fraud Detection." In: *IAENG International Journal of Computer Science* 42.2 (2015).

[34] Daniel Schlör, Markus Ring, and Andreas Hotho. "iNALU: Improved Neural Arithmetic Logic Unit". In: *Frontiers in Artificial Intelligence* 3 (2020), p. 71. ISSN: 2624-8212. DOI: 10.3389/frai.2020.00071.

[35] Daniel Schlör et al. "Financial Fraud Detection with Improved Neural Arithmetic Logic Units". In: vol. Fifth Workshop on mining data for financial applications. 2020.

[36] Bernhard Schölkopf et al. "Estimating the Support of a High-Dimensional Distribution". In: *Neural Computation* 13.7 (July 2001), pp. 1443–1471.

[37] Marco Schreyer et al. "Detection of accounting anomalies in the latent space using adversarial autoencoder neural networks". In: *2nd KDD Workshop on Anomaly Detection in Finance*. ACM. 2019.

[38] Marco Schreyer et al. "Detection of anomalies in large scale accounting data using deep autoencoder networks". In: *arXiv preprint arXiv:1709.05254* (2017).

[39] Martin Schultz and Marina Tropmann-Frick. "Autoencoder Neural Networks versus External Auditors: Detecting Unusual Journal Entries in Financial Statement Audits". In: *Proceedings of the 53rd Hawaii Int. Conf. on System Sciences*. 2020.

[40] Mei-Ling Shyu et al. *A novel anomaly detection scheme based on principal component classifier*. Tech. rep. Coral Gables, Florida: Miami univ. dept. of electrical and computer engineering, 2003.

[41] Kishore Singh and Peter Best. "Interactive visual analysis of anomalous accounts payable transactions in SAP enterprise systems". In: *Managerial Auditing Journal* (2016).

[42] Kishore Singh, Peter Best, and Joseph Mula. "Automating vendor fraud detection in enterprise systems". In: *Journal of Digital Forensics, Security and Law* 8.2 (2013), p. 1.

[43] Kishore Singh, Peter Best, and Joseph M Mula. "Proactive fraud detection in enterprise systems". In: *Proceedings of the 2nd International Conference on Business and Information: Steering Excellence of Business Knowledge*

(ICBI 2011). University of Kelaniya, Faculty of Commerce and Management Studies. 2011.

[44] Jimmy Singla et al. "A Survey of Deep Learning based Online Transactions Fraud Detection Systems". In: *2020 Int. Conf. on Intelligent Engineering and Management (ICIEM)*. IEEE. 2020, pp. 130–136.

[45] Julian Tritscher et al. "Evaluation of post-hoc XAI approaches through synthetic tabular data". In: *25th Int. Symposium on Methodologies for Intelligent Systems ISMIS*. 2020.

[46] Julian Tritscher et al. "Open ERP System Data For Occupational Fraud Detection". In: *arXiv preprint arXiv:2206.04460* (2022).

[47] Jongmin Yu et al. "Unusual Insider Behaviour Detection Framework on Enterprise Resource Planning Systems using Adversarial Recurrent Autoencoder". In: *IEEE Transactions on Industrial Informatics* (2021).

[48] Yongchun Zhu et al. "Modeling users' behavior sequences with hierarchical explainable network for cross-domain fraud detection". In: *Proceedings of The Web Conference 2020*. 2020, pp. 928–938.