Detection of Anomalies in Large Scale Accounting Data using Deep Autoencoder Networks

Marco Schreyer[a], **Timur Sattarov**[b], **Damian Borth**[a], **Andreas Dengel**[a] **and Bernd Reimer**[b]

[a]German Research Center for Artificial Intelligence (DFKI)
Kaiserslautern, Germany
firstname.lastname@dfki.de

[b]PricewaterhouseCoopers GmbH WPG
Frankfurt am Main, Germany
lastname.firstname@pwc.com

Abstract

Learning to detect fraud in large-scale accounting data is one of the long-standing challenges in financial statement audits or forensic investigations. Nowadays, the majority of applied techniques refer to handcrafted rules derived from known fraud scenarios. While fairly successful, these rules exhibit the drawback that fraudsters gradually adapt and find ways to circumvent them. In addition, these rigid rules often fail to generalize beyond known fraud scenarios. To overcome this challenge we propose a novel method of detecting anomalous journal entries using deep autoencoder networks. We demonstrate that the trained networks' reconstruction error regularized by the individual attribute probabilities of a journal entry can be interpreted as a highly adaptive anomaly assessment. Our empirical study, based on two datasets of real-world journal entries, demonstrates the effectiveness of the approach and outperforms several baseline anomaly detection methods. Resulting in a fraction of less than 0.15% (0.7%) of detected anomalous entries while achieving a high detection precision of 19.71% (9.26%). Initial feedback received by accountants underpinned the quality of our approach capturing highly relevant anomalies in the data. We envision this method as an important supplement to the forensic examiners' toolbox.

1 Introduction

The Association of Certified Fraud Examiners estimates in its Global Fraud Study 2016 [ACFE, 2016] that the typical organization loses 5% of its annual revenues due to fraud. The term "fraud" refers to "the abuse of one's occupation for personal enrichment through the deliberate misuse of an organization's resources or assets" [Wells, 2017]. A similar study, conducted by the auditors of PwC, revealed that nearly a quarter (22%) of respondents experienced losses of between $100,000 and $1 million [PwC, 2016]. The study also showed that financial statement fraud caused by far the greatest median loss of the surveyed fraud schemes[1].

At the same time organizations accelerate the digitization and reconfiguration of business processes [Markovitch and Willmott, 2014] affecting in particular Accounting Information Systems (AIS) or more generally Enterprise Resource Planning (ERP) systems. Steadily, these systems collect vast

[1]The ACFE study encompasses an analysis of 2,410 cases of occupational fraud investigated between January 2014 and October 2015 that occurred in 114 countries. The PwC study encompasses over 6,000 correspondents that experienced economic crime in the last 24 months.

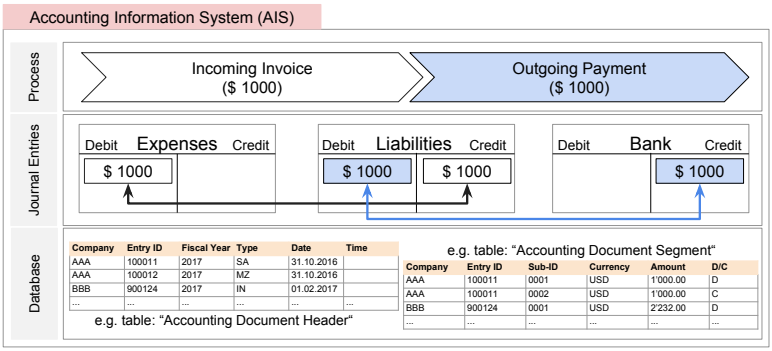

Figure 1: Hierarchical view of an Accounting Information System (AIS) recording process and journal entry information in designated database tables.

quantities of electronic evidence at an almost atomic level. This holds in particular for the journal entries of an organization recorded in its general ledger and sub-ledger accounts. SAP, one of the most prominent enterprise software providers, estimates that approx. 76% of the world's transaction revenue touches one of their ERP systems [SAP, 2017]. Figure 2 depicts a hierarchical view of an Accounting Information System (AIS) recording process and journal entry information in designated database tables.

To detect potentially fraudulent activities international auditing standards require the direct assessment of journal entries [AICPA, 2002, IFAC, 2009]. However, the detection of traces of fraud in up to several hundred million of journal entries is a labor-intensive task requiring a significant time effort and resources. Therefore, the described technological advances and vast amounts of data poses an enormous challenge for accountants during financial statement audits.

Nowadays, the majority of applied techniques to examine journal entries refer to rules defined by experienced chartered accountants or forensic examiners that are handcrafted and executed manually. The techniques are usually based on known fraud scenarios or markers "red-flags" (e.g. postings late at night, multiple vendor bank account changes, backdated expense account adjustments) or statistical analyses (e.g. Benford's Law [Benford, 1938], time series evaluation). Unfortunately, these techniques don't generalize beyond known historical fraud cases and therefore fail to detect novel schemes of fraud corresponding to the highly innovative nature of perpetrators. In addition, they exhibit the drawback that they become rapidly outdated while fraudsters adaptively find ways to circumvent them. Recent advances in deep learning [LeCun et al., 2015] made it possible to extract complex nonlinear features from raw sensory data leading to breakthroughs across many domains e.g. computer vision [Krizhevsky et al., 2012] and speech recognition [Mikolov et al., 2013]. Inspired by those developments we propose the application of deep autoencoder neural networks to detect anomalous journal entries in large volumes of accounting data. We envision that this automated and deep learning based examination of journal entries as an important supplementation of the accountants and forensic examiners toolbox [Pedrosa and Costa, 2014].

In order to conduct fraud, perpetrators need to deviate from regular system usage or posting pattern. This deviation will be "weakly" recorded and reflected accordingly by a very limited number of "anomalous" journal entries of an organization. In this work, we show that deep autoencoder neural networks can be trained to learn a compressed but lossy model of regular journal entries and their underlying posting pattern. Imposing a strong regularization onto the networks hidden layers limits the networks' ability to memorize the characteristics of anomalous journal entries. Once the training process is completed, the network will be able to reconstruct regular journal entries, while failing to do so for the anomalous ones. The autoencoder network is trained end-to-end and detects accounting anomalies without the need for traditional handcrafted rules. We demonstrate that the magnitude of a trained autoencoders' reconstruction error can be interpreted as an anomaly assessment of individual journal entries. A journal entry that exceeds a predefined reconstruction error threshold is flagged

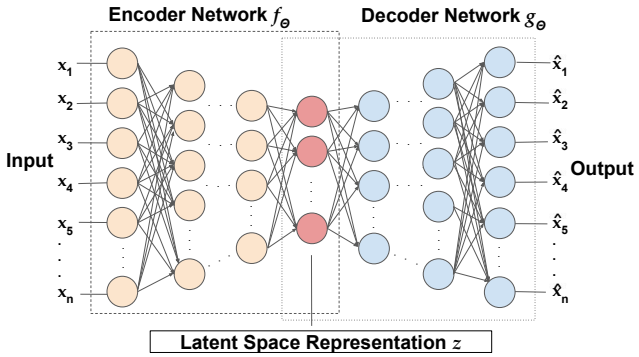

Figure 2: Schematic view of an autoencoder network comprised of two non-linear mappings (feed forward neural networks) referred to as encoder $f_\theta : \mathbb{R}^{d_x} \mapsto \mathbb{R}^{d_z}$ and decoder $g_\theta : \mathbb{R}^{d_z} \mapsto \mathbb{R}^{d_y}$.

as "anomalous". We evaluate the proposed approach based on two anonymized real-world datasets of journal entries extracted from large-scaled SAP ERP systems. Finally, the effectiveness of the autoencoder based method is underpinned by a baseline evaluation against several anomaly detection algorithms.

Section 2 of this work follows with an introduction of the general autoencoder network architecture and underlying principles. In section 3 we provide an overview of related work in the field. Next, in section 4 we describe the proposed methodology to detect accounting anomalies. The experimental setup and results are presented in section 5. In section 6 the paper concludes with a brief summary of the current work and future directions.

2 Background

We consider the task of training an autoencoder neural network using a set of N journal entries $X = \{x_1, x_2, ..., x_N\}$, where each entry $x_i \in \mathbb{R}^K$ consists of a K-dimensional vector of attributes e.g. the posting type, the affected general ledger or the posting date and time. An autoencoder or replicator neural network defines a special type of feed-forward multilayer neural network that can be trained to reconstruct its input. The difference between its reconstruction and the original input is referred to as reconstruction error. Figure 2 illustrates a schematic view of an autoencoder neural network.

Autoencoder networks are usually comprised of two nonlinear mappings referred to as encoder and decoder [Rumelhart et al., 1986]. Most commonly the encoder and the decoder are of symmetrical architecture consisting of several layers of neurons followed by a nonlinear function and shared parameters θ. The encoder mapping f_θ maps an input vector x to a hidden representation z referred to as compressed "latent space" representation. This representation z is then mapped back by the decoder g_θ to a re-constructed vector $\hat{x}$ in the original input space. Formally, the nonlinear mappings of the encoder and the decoder can be defined by:

$$f_\theta(x) = s(Wx + b), \text{and } g_\theta(z) = s'(W'z + d), \tag{1}$$

where s and s' denote the non-linear activations with model parameters $\theta = \{W, b, W', d\}$, $W \in \mathbb{R}^{d_x \times d_z}, W' \in \mathbb{R}^{d_z \times d_y}$ are weight matrices and $b \in \mathbb{R}^{d_x}, d \in \mathbb{R}^{d_z}$ are the offset bias vectors. In an attempt to achieve $x \approx \hat{x}$ the autoencoder is trained to minimize the dissimilarity between the input vector x and its reconstruction $\hat{x} = g_\theta(f_\theta(x))$ as faithfully as possible. Thereby, the objective of the autoencoder training is to optimize:

$$\arg\min_\theta \|x - g_\theta(f_\theta(x))\|, \tag{2}$$

over the shared encoder and decoder model parameters θ. During training, one typically minimizes a loss function $\mathcal{L}_{\mathcal{R}}$ defined by the squared reconstruction loss or, as used in our experiments, the cross-entropy loss given by:

$$\mathcal{L}_{\mathcal{R}}(\theta; x_{ij}) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} x_{ij} ln(\hat{x}_{ij}) + (1 - x_{ij}) ln(1 - \hat{x}_{ij}), \tag{3}$$

for each individual journal entry x_{ij}, $i = 1, ..., N$ and its attributes $j = 1, ..., K$. For one-hot encoded features, $\mathcal{L}_{\mathcal{R}}(\theta; x_{ij})$ measures the cross-entropy between two independent multivariate Bernoulli distributions, with mean x and mean $\hat{x}$ respectively [Bengio et al., 2013b].

To prevent the autoencoder from just learning the identity function a constraint is imposed onto the network's layers by limiting the number of the hidden layers neurons. The adoption of a bottleneck structure, $\mathbb{R}^{d_x} < \mathbb{R}^{d_z}$, and reduction of the number of hidden neurons will restrict the learning capability of the autoencoder. Thereby, the so called "undercomplete" autoencoder will be enforced to learn an optimal set of parameters θ^* resulting a compressed representation of the journal entries most prevalent feature distributions and dependencies.

The learned compressed feature representation corresponds to nonlinear manifolds in the latent (or hidden) space z which represent the variability and structural dependencies of the journal entries. It can be shown that the representations learned by a linear autoencoder can be viewed as a simple linear form of manifold learning, i.e., characterizing a lower-dimensional region in input space. In the case of linear autoencoders, the latent space representation is similar to the one learned by Principal Component Analysis (PCA) [Bengio et al., 2013a].

3 Related work

The task of detecting fraud and accounting anomalies has been studied both by practitioners [Wells, 2017] and academia [Amani and Fadlalla, 2017]. Several references exist that describe different fraud schemes and ways to detect unusual and "creative" accounting practices [Singleton and Singleton, 2006]. Our literature survey presented below focuses on (1) anomaly detection in Enterprise Resource Planning (ERP) data and (2) anomaly detection using autoencoder networks.

3.1 Anomaly Detection in Enterprise Resource Planning (ERP) Data

The forensic analysis of journal entries emerged with the advent of Enterprise Resource Planning (ERP) systems and the increased volume of data recorded in such systems. Bay et al. in [Bay et al., 2006] used Naive Bayes methods to identify suspicious general ledger accounts, by evaluating features derived from journal entries measuring any unusual general ledger account activity. Their approach was enhanced by McGlohon et al. applying link analysis to identify (sub-) groups of high-risk general ledger accounts [McGlohon et al., 2009].

Kahn et al. in [Khan and Corney, 2009] and [Khan et al., 2010] created transaction profiles of SAP ERP users. The profiles are derived from journal entry based user activity pattern recorded in two SAP R/3 ERP system in order to detect suspicious user behavior and segregation of duties violations. Similarly, Islam et al. also used SAP R/3 system audit logs to detect known fraud scenarios and collusion fraud via "red-flag" based scenario matching [Islam et al., 2010].

Debreceny and Gray in [Debreceny and Gray, 2010] analyzed dollar amounts of journal entries obtained from 29 US organizations. In their work, they searched for violations of Benford's Law [Benford, 1938], anomalous digit combinations as well as unusual temporal pattern such as end-of-year postings. More recently, Poh-Sun et al. in [Seow, Poh-Sun; Sun, Gary; Themin, 2016] confirmed the generalization of the approach by applying it to journal entries obtained from 12 non-US organizations.

Jans et al. in [Jans et al., 2010] used latent class clustering to conduct an uni- and multivariate clustering of SAP ERP purchase order transactions. Transactions significantly deviating from the cluster centroids are flagged as anomalous and are proposed for a detailed review by auditors. The approach was enhanced in [Jans et al., 2011] by a means of process mining to detect deviating process flows in an organization procure to pay process.

Argyrou et al. in [Argyrou, 2012] evaluated self-organizing maps to identify "suspicious" journal entries of a shipping company. In their work, they calculated the Euclidean distance of a journal entry and the code-vector of a self-organizing maps best matching unit. In subsequent work, they estimated optimal sampling thresholds of journal entry attributes derived from extreme value theory [Argyrou, 2013].

Concluding from the reviewed literature, the majority of references draw on historical accounting and forensic knowledge about various "red-flags" and fraud schemes. In agreement with [Wang, 2010], we see a demand for unsupervised approaches capable to detect so far unknown pattern of fraudulent journal entries.

3.2 Deep Learning based Anomaly Detection using Autoencoder Networks

Nowadays, autoencoder networks have been widely used in image classification and understanding [Hinton and Salakhutdinov, 2006], machine translation [Lauly et al., 2014] and speech processing [Tóth and Gosztolya, 2004] for their unsupervised data compression capabilities. To the best of our knowledge Hawkins et al. and Williams et al. were the first who proposed autoencoder networks for anomaly detection [Hawkins et al., 2002, Williams and Baxter, 2002].

Since then the ability of autoencoder networks to detect anomalous records was demonstrated in different domains such as X-ray images of freight containers [Andrews et al., 2016], the KDD99, MNIST, CIFAR-10 as well as several other datasets obtained from the UCI Machine Learning Repository[2] [Dau et al., 2014, An, 2015, Zhai et al., 2016]. In [Zhou, 2017] Zhou and Paffenroth enhanced the standard autoencoder architecture by an additional filter layer and regularization penalty to detect anomalies.

More recently autoencoder networks have been also applied in the domain of forensic data analysis. Cozzolino and Verdoliva used the autoencoder reconstruction error to detect pixel manipulations of images [Cozzolino and Verdoliva, 2017]. In [D'Avino et al., 2017] the method was enhanced by recurrent neural networks to detect forged video sequences. Lately, Paula et al. in [Paula et al., 2017] used autoencoder networks in export controls to detect traces of money laundry and fraud by analyzing volumes of exported goods.

To the best of our knowledge, this work presents the first deep learning inspired approach for the detection of anomalous journal entries in large-scaled and real-world accounting data.

4 Proposed Method

To detect anomalous journal entries we first define "normality" with respect to accounting data. We assume that the majority of journal entries recorded within an organizations' ERP system relate to regular day-to-day business activities. In order to conduct fraud, perpetrators need to deviate from the "regular". Such deviation will be recorded accordingly by a very limited number of journal entries we refer to as accounting anomalies. Based on this assumption we can learn a model of regular journal entries with minimal "harm" caused by the potential anomalous ones. It also implies that during an audit accountants or forensic examiners will, if at all, only observe a few anomalous journal entries reflecting such deviation.

4.1 Classification of Accounting Anomalies

When conducting a detailed examination of real-world journal entries recorded in large-scaled ERP systems two prevalent characteristics can be observed: First, individual journal entry attributes exhibit a high variety of distinct attribute values and second, journal entries exhibit strong dependencies between certain attribute values. Derived from this observation and similarly to Breunig et al. in [Breunig et al., 2000] we distinguish two classes of anomalous journal entries, namely *global* and *local anomalies*:

Global anomalies, are journal entries that exhibit unusual or rare individual attribute values. These anomalies usually relate to highly skewed attributes e.g. seldom posting users, rarely used ledgers, or unusual posting times. Traditionally "red-flag" tests, performed by auditors during annual audits, are

[2] https://archive.ics.uci.edu/ml/datasets.html

designed to capture those types of anomalies. However, such tests might result in a high volume of false positive alerts due to e.g. regular reverse postings, provisions and year-end adjustments usually associated with a low fraud risk.

Local anomalies, are journal entries that exhibit an unusual or rare combination of attribute values while the individual attribute values occur quite frequently e.g. unusual accounting records. This type of anomalies is significantly more difficult to detect since perpetrators intend to disguise their activities trying to imitate a regular behavior. As a result, such anomalies usually pose a high fraud risk since they might correspond to e.g. misused user accounts, irregular combinations of general ledger accounts and posting keys that don't follow usual activity pattern of an ERP system.

In regular audits, accountants and forensic examiners desire to detect journal entries corresponding to both anomaly classes that are "suspicious" enough to be followed-up. However, the concept of "suspicious" is difficult to quantify. In this work, we interpret this concept as the detection of (1) any unusual individual attribute value or (2) any unusual combination of attribute values observed. This interpretation is also inspired by earlier work of Das and Schneider [Das and Schneider, 2007] on the detection of anomalous records in categorical datasets.

4.2 Scoring of Accounting Anomalies

To account for the observation of irregular attribute values and to target global anomalies, we determine the probability of occurrence of each attribute in the dataset. A journal entry that exhibits a significant number of rarely occurring attribute values causes the entry to look unusual. Therefore, it might not have been created by a regular business activity. Formally, we define the attribute probability score AP_i as the sum of normalized attribute value log-probabilities of a journal entry x_i, given by:

$$AP(x_i) = \frac{\sum_{j=i}^{K} \log p(x_{ij}) - \log p_{min}(x_j)}{\log p_{max}(x_j) - \log p_{min}(x_j)}, \tag{4}$$

for each individual journal entry $x_{ij}, i = 1, ..., N$ and its attributes $j = 1, ..., K$.

However, when consulting with accountants and forensic examiners journal entries exhibiting a rare attribute value might not be of ultimate interest during an audit. This holds in particular for skewed attributes exhibiting a high variety of attribute values.

Therefore, to account for the observation of irregular attribute value combinations and to target local anomalies, we train autoencoder neural networks. During training, the autoencoder learns to model the linear and nonlinear attribute value dependencies of journal entries. Journal entries comprised of regular attribute values conform the learned model and are reconstructed with a small error. In contrast, a journal entry that violates the regular attribute value dependencies results in a high reconstruction error and potentially wasn't created by a regular business activity. Formally, we define the autoencoder network reconstruction error RE_i as the squared- or $L2$-difference of a journal entry x_i and its reconstruction $\hat{x}_i$, given by:

$$RE(x_i; \theta^*) = \frac{1}{K} \sum_{j=1}^{K} (x_{ij} - \hat{x}_{ij})^2, \tag{5}$$

for each individual journal entry $x_{ij}, i = 1, ..., N$ and its attributes $j = 1, ..., K$ under optimal model parameters θ^*.

Ultimately, we are interested in the detection of anomalous journal entries that correspond to a high reconstruction error and a high individual attribute probability. To detect such anomalies in real world accounting data we propose a novel anomaly score AS_i. Our score considers both observed characteristics, namely (1) any co-occurrence of attribute values regularized by (2) the occurrence of individual attribute values. Formally, we score each journal entry x_i by the weighted sum of its reconstruction error RE_i and its normalized attribute probability score AP_i, given by:

$$AS(x_i; \theta^*) = \alpha \times RE(x_i; \theta^*) + (1 - \alpha) \times AP(x_i), \tag{6}$$

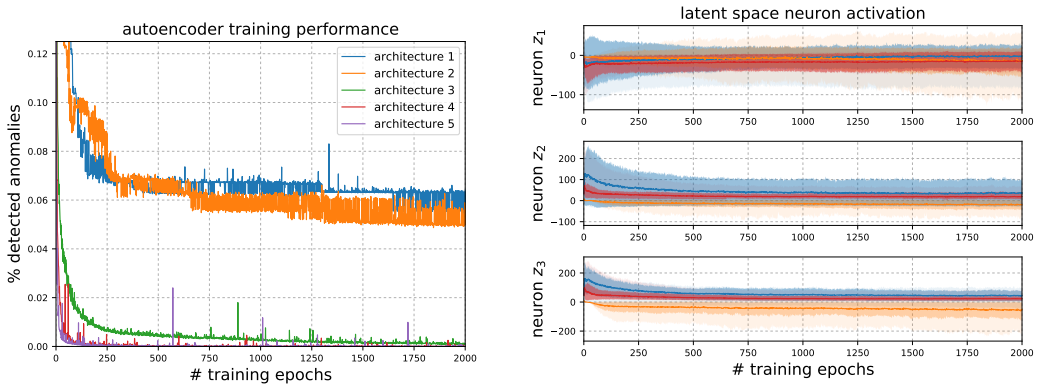

Figure 3: Training performance of distinct network architectures 1-5 using dataset A (left). Latent space neuron activation of the deep autoencoder architecture (Autoencoder 5) with progressing training epochs (right).

for each individual journal entry x_i, $i = 1, ..., N$ under optimal model parameters θ^*. Observing both characteristics for a single journal entry, we can reasonably conclude if an entry is anomalous or not and prevents journal entries to look suspicious due to rarely occurring attribute values. It also implies that we have seen enough evidence to support our judgment. We introduce α as a regularization factor that denotes the weight assigned to both characteristics. In addition, an individual journal entry is classified as anomalous if its reconstruction error RE_i exceeds a threshold parameter β, as defined by:

$$RE(x_i; \theta^*) = \begin{cases} RE(x_i; \theta^*), & RE(x_i; \theta^*) \geq \beta \\ 0, & \text{otherwise} \end{cases}, \tag{7}$$

for each individual journal entry x_i, $i = 1, ..., N$ under optimal model parameters θ^*.

5 Experimental Setup and Evaluation

We evaluated the anomaly detection performance of five distinct autoencoder architectures using the same learning algorithm and hyper-parameters settings in all our experiments. The result obtained for two real-world benchmark datasets demonstrate that the approach is robust enough to work on a variety of accounting datasets.

5.1 Datasets and Data Preparation

Both datasets have been extracted from SAP ERP systems for a single fiscal year respectively, denoted SAP ERP dataset A and dataset B in the following. In compliance with strict data privacy regulations, all journal entry attributes have been anonymized using a one-way hash function as part of the extraction process. Upon successful data extraction, the journal entries were reconciled against the general ledger trial balances of the respective companies to ensure data completeness.

SAP ERP systems record a variety of attributes of each journal entry predominantly in two tables named "BKPF" and "BSEG". The table "BKPF" - "Accounting Document Header" contains the meta information of a journal entry e.g., document id, type, date, time, currency. The table "BSEG" - "Accounting Document Segment", also referred to as line-item, contains the journal entry details e.g., posting key, general ledger account, debit and credit information, amount. We extracted a subset of 6 (dataset A) and 10 (dataset B) most distinctive attributes of the "BKPF" and "BSEG" tables.

The majority of attributes recorded in ERP systems correspond to categorical (discrete) variables, e.g. posting date, account, posting type, currency. In order to train autoencoder neural networks, we preprocessed the journal entry attributes to obtain a one-hot encoded representation of each categorical attribute. This preprocessing resulted in a total of 401 encoded dimensions for dataset A and 576 encoded dimensions for dataset B.

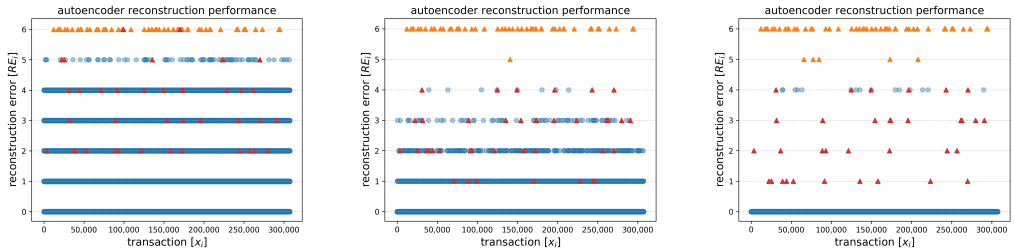

Figure 4: Journal entry reconstruction error RE obtained for each of the 307.457 journal entries x_i of dataset A after 10 (left), 100 (middle) and 400 (right) training epochs. The deep autoencoder (Autoencoder 4) learns to distinguish global anomalies (orange) and local anomalies (red) from original journal entries (blue) with progressing training epochs.

Architecture	Architecture [fully connected layers and neurons]
Autoencoder 1	[401; 576]-3-[401; 576]
Autoencoder 2	[401; 576]-4-3-4-[401; 576]
Autoencoder 3	[401; 576]-32-16-8-4-3-4-8-16-32-[401; 576]
Autoencoder 4	[401; 576]-256-128-64-32-16-8-4-3-4-8-16-32-64-128-256-[401; 576]
Autoencoder 5	[401; 576]-512-256-128-64-32-16-8-4-3-4-8-16-32-64-128-256-512-[401; 576]

Table 1: Autoencoder architecture applied in the experiments ranging from a shallow architecture (Autoencoder 1) encompassing a single hidden layer to a deep architecture (Autoencoder 5) encompassing multiple hidden layers.

To allow for an extensive analysis and quantitative evaluation of the experiments we induced synthetic global and local anomalies into both datasets. The induced global anomalies are comprised of attribute values not evident in the original data while the local anomalies exhibit combinations of attribute subsets not occurring in the original data. Ground truth labels are available in all datasets. Each journal entry was labeled as either synthetic *global anomaly*, synthetic *local anomaly* or non-synthetic *regular entry*. Both datasets highly unbalanced datasets are summarized by the following descriptive statistics:

- **Dataset A** contains 307'457 journal entry line items comprised of 6 categorical attributes. In total 95 (0.03%) synthetic anomalous journal entries have been induced the into dataset. These entries encompass 55 (0.016%) global anomalies and 40 (0.015%) local anomalies.
- **Dataset B** contains 172'990 journal entry line items comprised of 10 categorical attributes. In total 100 (0.06%) synthetic anomalous journal entries have been induced into the dataset. These entries encompass 50 (0.03%) global anomalies and 50 (0.03%) local anomalies.

5.2 Autoencoder Neural Network Training

In annual audits auditors aim to limit the number of journal entries subject to substantive testing while not missing any error or fraud related entry. Derived from this desire we formulated three evaluation objectives guiding our training procedure: (1) minimize the overall autoencoder reconstruction error, (2) maintain a recall of 100% to guarantee none of the synthetic anomalous journal entry will remain undetected, and (3) maximize the autoencoder detection precision to reduce the number of false-positive alerts.

In our experiments, we trained five distinct autoencoder architectures ranging from shallow (Autoencoder 1 architecture) to deep autoencoder networks (Autoencoder 5 architecture). Table 1 shows an overview of the evaluated architectures. The depth was increased by continuously adding fully-connected hidden layers of size 2^k neurons, where $k = 2, 3, ..., 9$. Increasing the autoencoder depth also increases its non-linearity but might yield the learned model to overfit. To prevent overfitting and learn more robust features we set the drop-out $p = 0.2$ [Hinton et al., 2012]. Furthermore, to reduce training time and prevent saturation of the nonlinearities we choose leaky rectified linear units (LReLU) [Xu et al., 2015] as the nonlinear activations given by $s(t) = \max(t, at)$, where $a = 0.4$.

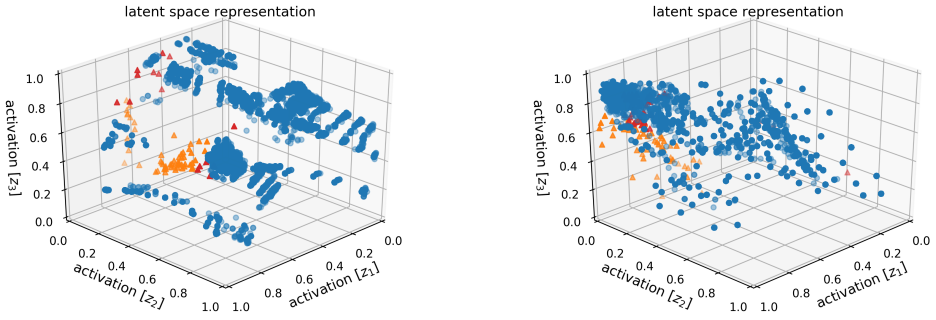

Figure 5: Latent space activations z_1, z_2 and z_3 of dataset A (left) and dataset B (right) learned by a deep autoencoder (Autoencoder 5). An increased manifold complexity can be observed for dataset B due to the high variety of journal entries contained in the dataset.

Each autoencoder architecture was trained applying an equal learning rate of $\eta = 10^{-4}$ for all layers, using a mini-batch size of 100 journal entries. Furthermore, we used adaptive moment estimation [Kingma and Ba, 2014] and initialized the activations of each network layer as proposed in [Glorot and Bengio, 2010]. The training was conducted via standard back propagation until convergence (max. 2'000 training epochs) while maintaining a recall of 100% of the synthetically induced anomalies. For each architecture, we run the experiments five times and noticed that the results stay consistent.

Figure 3 (left) illustrates the training performance of dataset A over distinct network topologies. Increasing the number of hidden units decreases the number of detected anomalies and converges significantly faster in terms of training epochs. Figure 3 (right) shows the mean neuron activation $[z_1, z_2, z_3]$ of the deep autoencoders (Autoencoder 5) three latent LReLU neurons z_1, z_2 and z_3 . With progressing training epochs the network learns a distinct activation pattern for each journal entry class. After 2'000 epochs a mean activation of $[-12.998, -20.576, -50.552]$ can be observed for global anomalies (orange), $[-15.091, 18.992, 23.403]$ for local anomalies (red) and $[-1.674, 37.277, 44.758]$ for regular journal entries (blue) is observed.

Once the training converged the learned models are used to derive an anomaly score AS for each journal entry. Figure 4 illustrates the reconstruction performance obtained for all 307.457 journal entries of dataset A after 10 (left), 100 (middle) and 400 (right) training epochs. The autoencoder (Autoencoder 4) learns to reconstruct the majority of original journal entries (blue) with progressing training epochs while failing to do so for global anomalies (orange) and local anomalies (red).

In real world audit scenarios accountants tend to handle fraudulent journal entries in a conservative manner to mitigate risks and not miss a potential true positive. In compliance with this guiding principle, we set the anomaly threshold $\beta = 1$ in experiments implying that a journal entry is labeled "anomalous" if at least one of its attributes isn't reconstructed correctly.

5.3 Experimental Results

We evaluate our hypothesis according to two criteria: (1) Are the trained autoencoder architectures capable of learning a model of the regular journal entries and thereby detect the induced anomalies (quantitative evaluation)? (2) Are the detected and non-induced anomalies "suspicious" enough to be followed up by accountants or forensic examiners (qualitative evaluation)?

Quantitative Evaluation: To quantitatively evaluate the effectiveness of the proposed approach, a range of evaluation metrics including precision, F1-score, absolute and relative number of detected anomalies are reported. The choice of F1-score is to account for the highly unbalanced class distribution of the datasets. Similarly, to visual recognition tasks, we also report the top-k accuracy. We set $k = 95$ (dataset A) or $k = 100$ (dataset B) reflecting the number of synthetic anomalies contained in both benchmark datasets.

Architecture	Dataset	Precision	F1-Score	Top-k	Accuracy	# Anomalies	% Anomalies
Autoencoder 1	A	0.0049	0.0098	0.0049	0.9378	19'233	6.26
Autoencoder 2	A	0.0063	0.0126	0.0063	0.9516	14'966	4.86
Autoencoder 3	A	0.0641	0.1204	0.6632	0.9955	1'483	0.48
Autoencoder 4	A	0.1201	0.2144	0.5684	0.9977	791	0.25
Autoencoder 5	A	**0.1971**	**0.3293**	**0.6947**	**0.9987**	**482**	**0.15**
Autoencoder 1	B	0.0020	0.0040	0.0020	0.7121	49'897	28.84
Autoencoder 2	B	0.0030	0.0059	0.0030	0.8054	33'762	19.52
Autoencoder 3	B	0.0087	0.0173	0.7400	0.9344	11'444	6.62
Autoencoder 4	B	0.0187	0.0368	0.7100	0.9697	5'335	3.08
Autoencoder 5	B	**0.0926**	**0.1695**	**0.4200**	**0.9943**	**1'080**	**0.62**

Table 2: Results for SAP ERP accounting dataset A evaluated on different network topologies, training was constrained to a recall of 100% using standard back propagation and stops at 2'000 epochs. Best detection performances are obtained by a deep autoencoder of 17 hidden layers and LReLU activations.

Model	Dataset	Precision	F1-Score	Recall	Accuracy	# Anomalies	% Anomalies
PCA	A	0.0008	0.0016	1.0000	0.6083	20'541	39.20
One Class SVM	A	0.0017	0.0034	0.9900	0.8458	47'485	15.44
LOF	A	0.0031	0.0062	1.0000	0.9001	30'795	10.04
HDBSCAN	A	0.0256	0.0499	1.0000	0.9882	3'714	1.21
Autoencoder 5	A	**0.1971**	**0.3293**	**1.0000**	**0.9987**	**482**	**0.15**
PCA	B	0.0009	0.0019	1.0000	0.3807	107'231	61.99
One Class SVM	B	0.0019	0.0038	0.9900	0.6999	52'018	30.06
LOF	B	0.0058	0.0115	1.0000	0.9006	17'300	10.00
HDBSCAN	B	0.0085	0.0168	0.9800	0.9336	11'574	6.69
Autoencoder 5	B	**0.0926**	**0.1695**	**1.0000**	**0.9943**	**1'080**	**0.62**

Table 3: Comparative evaluation of the autoencoder based approach to several anomaly detection techniques. For each method the best detection performance is reported that results in a recall of 100% of the synthetic anomalies.

Table 2 shows the results obtained for both benchmark datasets using distinct network architectures. Increasing the number of hidden layers significantly reduces the number of detected anomalies. While preserving recall of 100%, the deepest trained autoencoder architecture (Autoencoder 5) results in a low fraction of 0.15% detected anomalies in dataset A and 0.62% detected anomalies in dataset B. The observed results show that the autoencoder depth substantially affects its ability to model the inherent manifold structure within each dataset. Figure 6 (left) illustrates the anomaly score distributions obtained for the distinct journal entry classes.

To understand the observed difference in detection performance obtained for both datasets we investigated the learned latent space representations as illustrated in figure 5. The manifold structure learned for dataset B shows a significantly higher complexity compared to dataset A. We hypothesize that the limited capacity of the autoencoder architecture (Autoencoder 5) falls short to model the variety of journal entries contained in dataset B compared to dataset A.

Qualitative Evaluation: To qualitatively evaluate the character of the detected anomalies contained in both datasets we reviewed all regular (non-synthetic) journal entries corresponding to a reconstruction error $RE \geq 1.0$ detected by the Autoencoder 5 architecture. To distinguish local from global anomalies we empirically choose to set $\alpha = 0.5$ and flagged the journal entries resulting in an $AS \geq 0.6$ as local anomalies and $AS < 0.6$ as global anomalies. Figure 6 (right) shows the distribution of anomaly scores AS for distinct α values obtained for dataset A.

The review of the *global anomaly* journal entry population revealed that the majority of the anomalies refer to journal entries that exhibit one or two rare attribute values e.g. journal entries that correspond to seldom vendors or seldom currencies. However, it also turned out that these journal entries correspond to (1) posting errors due to wrongly used general ledger accounts, (2) journal entries of unusual document types containing extremely infrequent tax codes, or (3) incomplete journal entries exhibiting missing currency information. Especially, the latter observations indicated a weak control environment around certain business processes.

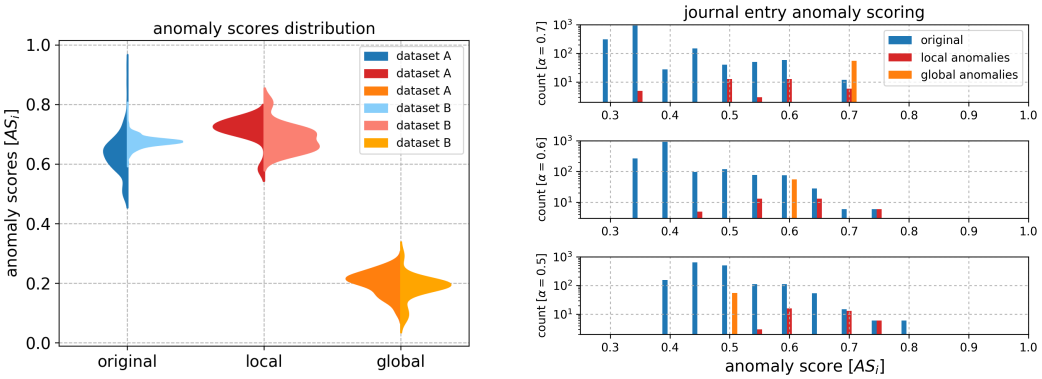

Figure 6: Distribution of anomaly scores AS_i obtained for the distinct journal entry classes in both datasets using a trained deep autoencoder (Autoencoder 5) (left). Anomaly scores AS_i of distinct α values obtained for a trained deep autoencoder (Autoencoder 5) and dataset A (right).

The review of the *local anomaly* journal entry population revealed that, as intended, these anomalies refer to journal entries exhibiting frequently observable attributes that rarely occur in combination e.g. changes of business process or rarely applied accounting practices. A more detailed investigation of the detected instances uncovered (1) shipments to customers that are invoiced in different than the usual currency; (2) products send to a regular client but were surprisingly booked to another company code; or (3) document types that are not supposed to be used in combination with the particular client. The initial feedback received by accountants regarding the detected anomalies underpinned not only their relevance from an audit but also from a forensic perspective.

Baseline Evaluation: We compared the autoencoder network based approach against unsupervised and state-of-the art anomaly detection techniques, namely (1) reconstruction error-based: Principal Component Analysis (PCA) [Pearson, 1901], (2) kernel-based: One Class Support Vector Machine (SVM) [Scholkopf et al., 2000], (3) density based: Local-Outlier Factor (LOF) [Breunig et al., 2000], and (4) hierarchical nearest-neighbor based: Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [Ester et al., 1996] anomaly detection. For all methods, besides DBSCAN, the performance was assessed using the implementations available in the sci-kit machine learning library [Pedregosa and Varoquaux, 2011]. For DBSCAN we used the optimized HDBSCAN implementation by Campello et al. in [Campello et al., 2013]. An exhaustive grid search over the parameter space was conducted to determine the best performing hyper-parameters of each technique. Table 3 shows the best results obtained for both dataset that yield to a recall of 100% of the synthetically induced anomalies. In instances where we couldn't determine hyper-parameters that yield to recall of 100% we report the scores with highest obtained recall values.

The evaluation results show, that for both benchmark datasets, HDBSCAN and the deepest trained autoencoder architecture (Autoencoder 5) yield to a low number of detected anomalies while preserving a high recall. In terms of anomaly detection precision, the autoencoder based approach outperforms the other benchmark techniques. In comparison to HDBSCAN the trained deep autoencoder networks results in 309 less detected anomalies in dataset A and 4'255 less detected anomalies in dataset B. This is of high relevance in the context of real audit scenarios where the substantiative evaluation of a single detected anomaly results in considerable time and effort.

6 Conclusion and Future Work

We introduced a novel method for the detection of anomalous journal entries in large scaled accounting data. The key to the performance of the approach is the end-to-end training of deep autoencoder neural networks using the native one-hot encoded journal entries as an input. Our empirical evaluation on two real-world accounting datasets supports the hypothesis that the regularized magnitude of a trained deep autoencoders reconstruction error can be used as an anomaly assessment of individual journal entries. Leaving only a small but highly accurate fraction of 0.15% journal entries in dataset A and 0.62% journal entries in dataset B for a further manual inspection. Qualitative feedback on the

detected anomalies received by accountants and fraud examiners revealed that our method captures journal entries highly relevant for a detailed audit while outperforming baseline anomaly detection methods.

Future work should encompass a more detailed investigation of the journal entries latent space representations learned by deep autoencoders. We believe that investigating the learned manifolds will provide additional insights into the fundamental structures of accounting data and underlying business processes. Furthermore, we aim to evaluate the anomaly detection ability of more recently proposed autoencoder architectures e.g. variational or adversarial autoencoder. Given the tremendous amount of journal entries recorded by organizations annually, an automated and high precisions detection of accounting anomalies can save auditors considerable time and decrease the risk of fraudulent financial statements.

Acknowledgments

We thank NVIDIA for its generous DGX-1 and GPU donations. We also thank the developers of TensorFlow [Abadi et al., 2015], a machine learning framework we used in our experiments.

References

M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, L. Kaiser, M. Kudlur, J. Levenberg, D. Man, R. Monga, S. Moore, D. Murray, J. Shlens, B. Steiner, I. Sutskever, P. Tucker, V. Vanhoucke, V. Vasudevan, O. Vinyals, P. Warden, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, 2015. URL `https://arxiv.org/abs/1603.04467`.

ACFE. *Report to the Nations on Occupational Fraud and Abuse, The 2016 Global Fraud Study*. Association of Certified Fraud Examiners (ACFE), 2016. URL `https://s3-us-west-2.amazonaws.com/acfepublic/2016-report-to-the-nations.pdf`.

AICPA. *Consideration of Fraud in a Financial Statement Audit*. American Institute of Certified Public Accountants (AICPA), 2002. URL `https://www.aicpa.org/Research/Standards/AuditAttest/DownloadableDocuments/AU-00316.pdf`.

F. A. Amani and A. M. Fadlalla. Data mining applications in accounting: A review of the literature and organizing framework. *International Journal of Accounting Information Systems*, 24:32–58, 2017.

J. An. Variational Autoencoder based Anomaly Detection using Reconstruction Probability. Technical report, 2015.

J. T. A. Andrews, E. J. Morton, and L. D. Griffin. Detecting Anomalous Data Using Auto-Encoders. *International Journal of Machine Learning and Computing*, 6(1):21–26, 2016.

A. Argyrou. Auditing Journal Entries Using Self-Organizing Map. In *Proceedings of the Eighteenth Americas Conference on Information Systems (AMCIS)*, number 16, pages 1–10, Seattle, Washington, 2012.

A. Argyrou. Auditing Journal Entries Using Extreme Vale Theory. *Proceedings of the 21st European Conference on Information Systems*, 1(2013), 2013.

S. Bay, K. Kumaraswamy, M. G. Anderle, R. Kumar, D. M. Steier, A. Blvd, and S. Jose. Large Scale Detection of Irregularities in Accounting Data. In *Data Mining, 2006. ICDM'06. Sixth International Conference on*, pages 75–86. IEEE, 2006.

F. Benford. The Law of Anomalous Numbers. *Proceedings of the American Philosophical Society*, 78(4):551–572, 1938.

Y. Bengio, A. Courville, and P. Vincent. Representation Learning: A Review and New Perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1798–1828, 2013a.

Y. Bengio, L. Yao, G. Alain, and P. Vincent. Generalized denoising auto-encoders as generative models. In *Advances in Neural Information Processing Systems*, pages 899–907. 2013b.

M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. LOF: Identifying Density-Based Local Outliers. In *Proceedings of the 2000 Acm Sigmod International Conference on Management of Data*, pages 1–12, 2000.

R. J. G. B. Campello, D. Moulavi, and J. Sander. Density-Based Clustering Based on Hierarchical Density Estimates. Technical report, 2013.

D. Cozzolino and L. Verdoliva. Single-image splicing localization through autoencoder-based anomaly detection. In *8th IEEE International Workshop on Information Forensics and Security, WIFS 2016*, pages 1–6, 2017.

K. Das and J. Schneider. Detecting anomalous records in categorical datasets. *ACM SIGKDD International conference on Knowledge discovery and data mining*, pages 220–229, 2007.

H. A. Dau, V. Ciesielski, and A. Song. Anomaly Detection Using Replicator Neural Networks Trained on Examples of One Class. In *Asia-Pacific Conference on Simulated Evolution and Learning*, pages 311–322, 2014.

D. D'Avino, D. Cozzolino, G. Poggi, and L. Verdoliva. Autoencoder with recurrent neural networks for video forgery detection. *arXiv preprint*, (March):1–8, 2017. URL `https://arxiv.org/abs/1708.08754`.

R. S. Debreceny and G. L. Gray. Data mining journal entries for fraud detection: An exploratory study. *International Journal of Accounting Information Systems*, 11(3):157–181, 2010.

M. Ester, H. P. Kriegel, J. Sander, and X. Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, pages 226–231, 1996.

X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 9:249–256, 2010.

S. Hawkins, H. He, G. Williams, and R. Baxter. Outlier Detection Using Replicator Neural Networks. In *International Conference on Data Warehousing and Knowledge Discovery*, number September, pages 170–180. Springer Berlin Heidelberg, 2002.

G. E. Hinton and R. R. Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786):504–507, 2006.

G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. Technical report, 2012.

IFAC. *International Standards on Auditing 240, The Auditor's Responsibilities Relating to Fraud in an Audit of Financial Statements*. International Federation of Accountants (IFAC), 2009. URL `http://www.ifac.org/system/files/downloads/a012-2010-iaasb-handbook-isa-240.pdf`.

A. K. Islam, M. Corney, G. Mohay, A. Clark, S. Bracher, T. Raub, and U. Flegel. Fraud detection in ERP systems using Scenario matching. *IFIP Advances in Information and Communication Technology*, 330:112–123, 2010.

M. Jans, N. Lybaert, and K. Vanhoof. Internal fraud risk reduction: Results of a data mining case study. *International Journal of Accounting Information Systems*, 11(1):17–41, 2010.

M. Jans, J. M. Van Der Werf, N. Lybaert, and K. Vanhoof. A business process mining application for internal transaction fraud mitigation. *Expert Systems with Applications*, 38(10):13351–13359, 2011.

R. Khan and M. Corney. A role mining inspired approach to representing user behaviour in ERP systems. In *Proceedings of The 10th Asia Pacific Industrial Engineering and Management Systems Conference*, number December, pages 2541–2552, 2009.

R. Khan, M. Corney, A. Clark, and G. Mohay. Transaction Mining for Fraud Detection in ERP Systems. *Industrial Engineering and Management Systems*, 9(2):141–156, 2010.

D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint*, pages 1–15, 2014. URL `http://arxiv.org/abs/1412.6980`.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Advances In Neural Information Processing Systems*, pages 1–9, 2012.

S. Lauly, H. Larochelle, and M. Khapra. An Autoencoder Approach to Learning Bilingual Word Representations. In *Advances in Neural Information Processing Systems*, pages 1853–1861, 2014.

Y. LeCun, Y. Bengio, G. Hinton, L. Y., B. Y., and H. G. Deep learning. *Nature*, 521(7553):436–444, 2015.

S. Markovitch and P. Willmott. Accelerating the digitization of business processes. *McKinsey & Company*, pages 1–5, 2014.

M. McGlohon, S. Bay, M. G. M. Anderle, D. M. Steier, and C. Faloutsos. SNARE: A Link Analytic System for Graph Labeling and Risk Detection. *Kdd-09: 15Th Acm Sigkdd Conference on Knowledge Discovery and Data Mining*, pages 1265–1273, 2009.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, pages 1–12, 2013.

E. L. Paula, M. Ladeira, R. N. Carvalho, and T. Marzagão. Deep learning anomaly detection as suppor fraud investigation in Brazilian exports and anti-money laundering. In *Proceedings - 2016 15th IEEE International Conference on Machine Learning and Applications, ICMLA 2016*, pages 954–960, 2017.

K. Pearson. LIII. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*, 2(11):559–572, 1901. ISSN 1941-5982. doi: 10.1080/14786440109462720.

F. Pedregosa and G. Varoquaux. *Scikit-learn: Machine learning in Python*, volume 12. 2011. ISBN 9781783281930. doi: 10.1007/s13398-014-0173-7.2. URL `http://dl.acm.org/citation.cfm?id=2078195`.

I. Pedrosa and C. J. Costa. New trends on CAATTs: what are the Chartered Accountants' new challenges? *ISDOC '14 Proceedings of the International Conference on Information Systems and Design of Communication, May 16–17, 2014, Lisbon, Portugal*, pages 138–142, 2014.

PwC. *Adjusting the Lens on Economic Crime, The Global Economic Crime Survey 2016*. PricewaterhouseCoopers LLP, 2016. URL `https://www.pwc.com/gx/en/economic-crime-survey/pdf/GlobalEconomicCrimeSurvey2016.pdf`.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representation by error propagation. In *Parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations*, pages 318–362. 1986.

SAP. *SAP Global Corporate Affairs, Corporate Factsheet 2017*. 2017. URL `https://www.sap.com/corporate/de/documents/2017/04/4666ecdd-b67c-0010-82c7-eda71af511fa.html`.

B. Scholkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt. Support Vector Method for Novelty Detection. In *Advances in neural information processing systems*, volume 12, pages 582–588, 2000.

S. Seow, Poh-Sun; Sun, Gary; Themin. Data Mining Journal Entries for Fraud Detection : a Replication of Debreceny and Gray ' S (2010). *Journal of Forensic Investigative Accounting*, 3 (8):501–514, 2016.

T. Singleton and A. J. Singleton. *Fraud auditing and forensic accounting*. John Wiley & Sons, 2006. ISBN 047087791X.

L. Tóth and G. Gosztolya. Replicator Neural Networks for Outlier Modeling in Segmental Speech Recognition. *Advances in Neural Networks–ISNN 2004*, pages 996–1001, 2004.

S. Wang. A comprehensive survey of data mining-based accounting-fraud detection research. In *2010 International Conference on Intelligent Computation Technology and Automation, ICICTA 2010*, volume 1, pages 50–53, 2010.

J. T. Wells. *Corporate Fraud Handbook: Prevention and Detection*. John Wiley & Sons, 2017. ISBN 9781118728574.

G. Williams and R. Baxter. A comparative study of RNN for outlier detection in data mining. *IEEE International Conference on Data Mining*, (December 2002):1–16, 2002.

B. Xu, N. Wang, T. Chen, and M. Li. Empirical Evaluation of Rectified Activations in Convolution Network. *ICML Deep Learning Workshop*, pages 1–5, 2015.

S. Zhai, Y. Cheng, W. Lu, and Z. Zhang. Deep Structured Energy Based Models for Anomaly Detection. In *International Conference on Machine Learning*, volume 48, pages 1100–1109, 2016.

C. Zhou. Anomaly Detection with Robust Deep Auto-encoders. In *KDD '17 Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 665–674, 2017.