Proceedings of the 53rd Hawaii International Conference on System Sciences | 2020

Autoencoder Neural Networks versus External Auditors: Detecting Unusual Journal Entries in Financial Statement Audits

Martin Schultz
University of Applied Sciences Hamburg
martin.schultz@haw-hamburg.de

Marina Tropmann-Frick
University of Applied Sciences Hamburg
marina.tropmann-frick@haw-hamburg.de

Abstract

With the increasing complexity of business processes in today's organizations and the ever-growing amount of structured accounting data, identifying erroneous or fraudulent business transactions and corresponding journal entries poses a major challenge for public accountants at annual audits. In current audit practice, mainly static rules are applied which check only a few attributes of a journal entry for suspicious values. Encouraged by numerous successful adoptions of deep learning in various domains we suggest an approach for applying autoencoder neural networks to detect unusual journal entries within individual financial accounts. The identified journal entries are compared to a list of entries that were manually tagged by two experienced auditors. The comparison shows high f-scores and high recall for all analyzed financial accounts. Additionally, the autoencoder identifies anomalous journal entries that have been overlooked by the auditors. The results underpin the applicability and usefulness of deep learning techniques in financial statement audits.

1. Introduction and Motivation

Not only after the financial scandals and fraud cases in recent years [1], most companies are required by law to have their financial statements audited annually by external auditors. The purpose of an annual audit is to strengthen the confidence of prospective users in the issued financial statements. The auditor has to express an opinion on whether the financial statements have been prepared in accordance with applicable accounting standards, are free from material misstatements (e.g. due to errors or fraud) and faithfully represent the financial performance of the audited entity [2, 3]. With this aim in mind, several auditing standards are issued by national and international bodies (e.g. International Auditing and Assurance Standards Board (IAASB)) providing binding guidelines for all audit assignments. Thus, in today's audit practice usually a risk-oriented audit approach is applied which encompasses an assessment of risks for material misstatements, audit procedures related to significant business processes and internal controls embedded in them (test of control) as well as substantive audit procedures (test of details).

This audit approach is designed to cope with the ever-increasing complexity of the companies to be audited. Todays' organizations strive for a high degree of standardization and automation of their business processes extensively relying on information systems (IS). This leads to a continuously growing amount of audit-relevant data. Enterprise resource planning systems (ERP) handle millions of business transactions every day, the financial impact of which is recorded as journal entries in the financial accounts. For auditing major classes of usual business transactions, auditors try to gain an in-depth understanding of the underlying business processes. The auditor traces all steps of a small sample of business transactions from origination through completion along with the financial impact and enacted internal controls (denoted as walk-throughs) [4]. By focusing on a small sample of major process variants, audit efficiency is increased but auditors take the risk of overlooking rare and anomalous business transactions that do not originate from well-controlled processes. In particular, such unusual business transactions may involve an inherent risk of material misstatements.

In this regard, audit standards require an analysis of accounting data on a detailed level, namely on journal entry level [5]. Using computer-assisted audit techniques (CAAT), the full population of journal entries is scanned for suspicious attribute values. This audit procedure is often referred to as *Journal Entry Testing*. Although enforced by auditing standards, there is no generally accepted set of rules or indicators that should be used for identifying anomalous journal entries. In current audit practice, mostly static rules are applied that check only a few attributes of a journal entry at a time (e.g. postings with large amounts, postings created outside normal business hours or on weekends, postings close to fiscal year-end, users with

URI: https://hdl.handle.net/10125/64408
978-0-9981331-3-3

few postings, etc.) which results in a high false positive rate. Besides, apart from the Big 4 audit firms, many auditors lack the necessary technical skills to perform such data-driven analyses on large datasets [6].

In contrast, with the substantive progress in the last years with the adoption of deep learning in various domains (e.g. image and speech recognition) also auditing research intensively discusses application scenarios for deep learning techniques and projects fundamental changes for several audit tasks or the audit approach as a whole [6, 7]. At the same time, it is outlined that the audit domain has lagged behind the business in technology adoption in the past as a large number of regulations and standards causing a delaying effect. [6]. Regarding deep learning, it is also discussed whether common populations in audits are large enough to apply such techniques [6].

Against this background, this paper proposes an approach for applying a deep learning technique - namely autoencoder neural networks - to detect unusual journal entries for audit purposes. With the ability to extract nonlinear features and more complex patterns from a dataset, this technique may support auditors in purposefully selecting journal entries for further inspection. Our approach applies autoencoder to individual financial accounts each with a small population of journal entries. This approach respects that individual financial accounts are the main structuring element in annual audits. Furthermore, we incorporate domain knowledge during the preparation of the used real-world dataset by adding several computed attributes that emulate audit relevant aspects of journal entries. The overarching goal is to investigate the applicability and usefulness of this approach in a realistic audit scenario. By comparing the results of the autoencoder with the professional judgment of two experienced external auditors, the paper presents valuable insights regarding the extent autoencoder can contribute to increasing audit effectiveness in financial statement audits.

The research presented in this study follows a design science research approach [8, 9]. Accordingly, the remainder of this paper is structured as follows: The next section elaborates on related research regarding anomaly detection with autoencoder neuronal networks and the detection of unusual journal entries within annual audits. Section 3 presents our designed artifact for detecting unusual journal entries with autoencoder neuronal networks. The findings gathered in the evaluation step carried out with the assistance of domain experts are described in Section 4. The paper concludes in section 5 with a summary followed by a conclusion along with implications for future research work.

2. Related work

2.1. Anomaly detection with autoencoder neural networks

An autoencoder neural network is an unsupervised deep learning technique that is used in various domains, such as image classification, natural language processing, anomaly detection, etc. The general autoencoder network applies *backpropagation* and consists of two components, an *encoder* and a *decoder*. The main concept is that both encoder and decoder are trained together minimizing the discrepancy between the original data and its reconstruction. The encoder $e(x)$ represents a mapping of an input $x \in \mathbb{R}^{dim}$ to a hidden compressed representation, and the decoder $d(x)$ maps this compressed representation back to a reconstructed version of x, such that $d(e(x)) \approx x$. The goal is to find such model parameters Θ which minimize the reconstruction error, measured by some loss function $L_\Theta(x, d(e(x)))$.

Various forms of autoencoders have been developed for different purposes. The first introduction was made in the 1980s by Hinton and the PDP research group [10] to address the problem of *backpropagation* without a supervisor, followed by [11, 12, 13].

Autoencoders can be considered successful also in the area of anomaly or outlier detection. In [14] an application of an autoencoder network was presented finding outliers in large multivariate databases. The authors in [15] compare autoencoders and linear principal component analysis (PCA) with the result of better accuracy of autoencoder without complex computation required by PCA. In [16] the authors presented a comprehensive survey of deep learning-based network anomaly detection approaches describing various methods for intrusion detection including autoencoders as one of the most promising techniques.

2.2. Anomaly detection in accounting data from an audit perspective

In the past, the audit domain has lagged behind the business in technology adoption not least because of the fact that it is highly regulated and standard-driven with a noticeable delaying effect [6]. However, the extent impact of technology support on audit effectiveness and efficiency is widely acknowledged [17, 18]. In academia and practice, information technology support for audit tasks is discussed under the term computer-assisted audit technique (CAAT) which is broadly defined as any use of technology to assist in the completion of an audit [19]. In comprehensive literature reviews,

many different analytical techniques are listed. They are applied in all phases of an audit [20, 21]. The identified techniques range from ratio analysis, sampling and descriptive statistics over statistical regression models, Benford's Law, analytical hierarchy process and Monte Carlo simulation to clustering, classification models, text mining, process mining, and artificial neural networks. For high-risk areas, audit literature often refers to statistical sampling techniques for testing accounting data [22]. A common application of CAATs in current audit practice that examines the full population of journal entries is the afore mention Journal Entry Testing.

One stream of literature discusses anomaly detection in accounting data mainly from a fraud perspective. *Bay et al.* applied and compared classification methods (Naive Bayes, expectation-maximization, logistic regression) for identifying suspicious financial accounts by evaluating a set of features derived from journal entry attributes that indicate different types of unusual activities on the respective accounts [23]. *McGlohon et al.* used link analysis to identify suspicious accounts not only based on irregularities in a single account but also in accounts that are related due to shared business transactions [24]. *Argyrou* proposed a model based on self-organizing maps that can detect suspicious journal entries [25]. A subsequent paper proposed a bipartite model based on extreme value theory and Bayesian analysis to detect journal entries with a low probability of causing a material misstatement [26]. Another stream of literature focusses on business processes and the application of process mining to identify unusual or not well-controlled process instances from ERP system data [27, 28].

Schreyer et al. described one of the few attempts to apply deep learning in the context of auditing [29]. They also used autoencoder neural networks to detect fraudulent journal entries in large-scale accounting data extracted from ERP systems. By comparing the autoencoder results with other unsupervised anomaly detection techniques (Principal Component Analysis, One Class Support Vector Machine, Local-Outlier Factor, and Density-Based Spatial Clustering of Applications with Noise (DBSCAN)), they demonstrated that the autoencoder based approach provides superior results which leads to the conclusion that autoencoder can be used as an adaptive anomaly assessment of journal entries. However, for their quantitative evaluation, they injected a small fraction of synthetic anomalous journal entries into their datasets. Furthermore, the autoencoder network is applied to large datasets (more than 150.000 journal entry line items) comprising all journal entries of a complete fiscal year. In this paper, no synthetic anomalous journal entries are used and the chosen approach considers that individual financial accounts are an essential structuring element in financial audits. The autoencoder network is applied to subsets of journal entries on an account by account basis. Doing so, this paper addresses also the research question of whether an autoencoder based approach is also applicable to small populations of journal entries.

From the review of related work, it can be concluded that the application of deep learning techniques in auditing is a promising research field with several open questions to be addressed.

3. Analysis setup and autoencoder network training

3.1. Dataset and data preparation

The real-world dataset for our analysis has been extracted from an SAP ERP instance of a food production and trading company. The dataset comprises the entire population of journal entries of one legal entity (SAP: company code) and a single fiscal year. In total, the dataset consists of 72.917 journal entries (accounting documents) with 302.365 line items. This original dataset has been used by the company's external auditor within the annual audit for a journal entry testing. Therefore, long descriptions for technical abbreviations (e.g. account name, transaction code name, document origin, etc.) were added to make the data more readable for a human auditor. This original dataset forms the basis for the evaluation step in this paper. Due to the strict data privacy regulations in the audit domain, all journal entry attribute values (except the attribute amount) of the original dataset have been anonymized using a one-way hash function. This anonymized dataset is used for the analysis with the autoencoder.

In line with the common procedures for an audit of financial statements, we conduct the analysis on a per-account basis rather than across all journal entries. This approach considers, that individual accounts constitute the main structuring element for annual audits in terms of related inherent risks, expected accounting schemes, and to be applied audit procedures. In cooperation with the auditors involved, three suitable accounts are selected: 1) Revenue Domestic, 2) Revenue Foreign, 3) Expenses. This carefully considered choice is based on the auditors' expectations that these accounts are subject to a uniform accounting scheme and are addressed essentially by standardized and highly automated business processes. In addition, the revenue cycle is one of the main audit areas in all financial

statement audits as auditing standards denominate it as a high-risk area [30]. This selection is also guided by the research question of whether deep learning techniques can provide useful results also for a small population of data items. Therefore, one large account and two accounts with a relatively low number of journal entries are included in the sample (Revenue Domestic: 6.643 journal entries, Revenue Foreign: 651 journal entries, Expenses: 778 journal entries).

Technically, the data is extracted from mainly two tables (BKPF – Accounting Document Header, BSEG – Accounting Document Segment) which contain the most relevant attributes for all accounting documents in an SAP ERP system. By reconciling the extracted data with the trial balance report of the SAP ERP system, the completeness of the dataset is confirmed. Most of the stored attributes for an accounting document are categorical. In the dataset used in this paper, only the amount attribute is an exception. Based on the insights gather from a discussion with the involved external auditors, three additional attributes are computed: 1) *day of week* – derived from the creation date of the accounting document, 2) *dates equal* - a Boolean attribute indicating whether all three date fields (posting date, document date, creation date) are equal or not, and 3) *dates in accounting year* - a Boolean attribute representing whether the values of all three date fields lie within the date range of the fiscal year under review. To take the structure of the posting document into account in the analysis, sorted lists of all account numbers on the debit and credit side of the respective accounting document are supplemented for each journal entry line item (group_concat SQL function, attribute names: *debit/ credit account list*). The same approach is used for the customer and vendor attributes (attribute names: *customer/ vendor number list*, *customer name list*, *vendor name list*, *customer/ vendor country list*).

In total, the following 25 attributes form the basis for the subsequent analysis: *doc. number, doc. type, posting key, dates equal, dates in accounting year, day of week, currency, debit/ credit indicator, tax code, revenue indicator, user id, user group, debit account list, credit account list, doc. notes, doc. line item notes, customer/ vendor number list, customer name list, vendor name list, customer/ vendor country list, transaction code, recurring doc. number, reversal doc. number, doc. origin and amount (local currency).*

3.2. Autoencoder network training

In this work, we use an autoencoder network for anomaly detection. The general structure of such a network consists of two components - an *encoder* and a *decoder*. Figure 1 shows a schematic representation of a three-layer autoencoder network. The encoder $e(x)$ transforms the input x to a hidden representation h. In this step we use *leaky-ReLU* as an activation function. The decoder transforms the hidden representation h to an approximation of the original input $y \approx x$. As the output activation function a sigmoid function is used here. The goal of the autoencoder is to find optimal model parameters Θ for the minimization of a loss function. For our purpose we use the *Mean Squared Error* loss function (MSE)

$$L(x, y) = argmin_{\Theta} ||x - y||_2.$$

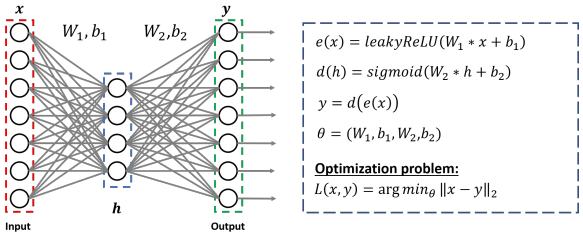

Figure 1. Autoencoder network

For our setting, we choose a deep autoencoder network with multiple fully-connected hidden layers, see Figure 2. However, our tests showed that a maximal depth of 9 hidden layers with [128 (input layer), 64, 32, 16, 8, 4, 8, 16, 32, 64, 128 (output layer)] neurons respectively is sufficient and deeper networks don't bring any additional improvement. Our implementation is done in Python using *keras* and *tensorflow* and doesn't require many resources. The execution takes at most several seconds on common hardware.

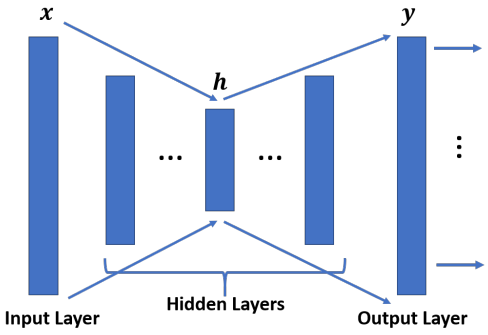

Figure 2. Deep autoencoder network

After loading, the input data is divided into two sets - one for categorical attributes and one for numerical attribute 'amount'. The separate treatment of the amount-attribute is done due to its special importance in this application domain. The outlier detection for this attribute is done as follows:

1. Transformation of the attribute values $x = (x_0, ..., x_n)$ to a zero-mean and unit standard deviation: $\frac{x-\overline{x}}{\sigma(x)}$, whereby $\overline{x}$ is the mean value and $\sigma(x)$ the standard deviation of x.
2. x_i is considered as outlier if $|x_i| > k * \sigma(x)$. Throughout this paper k was fixed to $k = 3$.

All other attribute values of the input data are categorical and cannot be directly used for the autoencoder. For proper handling, we propose the following approach (similar to [29]):

1. Transformation of the attribute values to one-hot-encoded vectors.
2. Training of the autoencoder network with the one-hot representation of the categorical values. We use a fixed set of parameters for all executions:
   - learning rate $= 0.0001$,
   - batch size - equal to the number of samples,
   - number of epochs - 400 to 500,
   - activation function for the hidden layers - *LeakyReLU* function with $\alpha = 0.4$,
   - activation function for the output layer - *Sigmoid* function.
3. Samples are considered as outliers if the reconstruction error of the autoencoder

   $$L(x, y) = argmin_{\Theta}||x - y||_2$$

   is above a specific threshold. For each training set the threshold is set dynamically, so that 10% of samples with the highest error are considered as outliers.

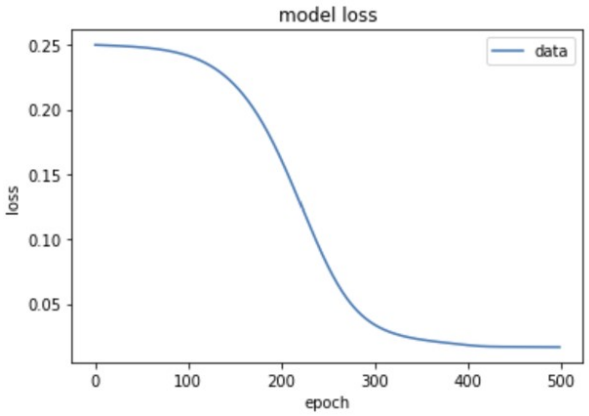

Figure 3. Learning for account revenue domestic

Figure 3 exemplarily depicts the learning (or loss) curve for one financial account we evaluated. The other learning curves are very similar. It shows that the autoencoder network learns the pattern of 'normal values' in a proper way.

Finally, a journal entry is classified as an outlier by our autoencoder network if one of the approaches (numerical or categorical) described above marks it as such. With this implementation, the main goal is to identify all unusual journal entries respectively reduce the number of false negatives to the lowest possible level. This goal is derived from the requirements of the audit domain. Even a single journal entry can have an inherent risk for a material misstatement. This goal relates to audit effectiveness. Since auditing standards require that all identified exceptions should be examined, the second goal of the implementation is to produce as few false positives as possible [30]. This goal is related to audit efficiency.

4. Evaluation

4.1. Evaluation design

In a design science research project, the evaluation step tries to observe and measure how well the designed artifact supports a solution for the addressed problem [31]. For evaluating our autoencoder approach, a qualitative analysis is conducted with the help of domain experts. Two experienced auditors (both having more than 10 years working experience in the audit domain, one of the auditors (Auditor 1) is the responsible external auditor of the company that provided the data for this paper) reviewed a complete list of journal entries that have been posted to the three selected accounts. For this task, an excel file containing the journal entry details and a worksheet with the frequencies of all attribute values per account is provided. The auditors are asked to identify unusual journal entries and to give a short explanation of the conspicuous characteristics for the selected ones. There is a common understanding among auditors on what constitutes an unusual journal entry. The irregularities fall into three categories [23]:

- Irregularities due to financial statement fraud (e.g. improper revenue recognition, asset misappropriation),
- Unintentional errors (e.g. erroneous data input, postings to a wrong account) that have a material effect on the financial statements,
- Unusual entries in the general ledger (neither fraud nor error) that an auditor would consider worth investigating.

This task is divided into two consecutive steps: 1) The two external auditors initially code the journal

entries independently of each other. 2) In a joined meeting coding differences are discussed and combined to a final list of unusual journal entries (agreed result). The results of this preparatory evaluation step are depicted in table 1.

Table 1. Tagging results of the auditors

		No. of tagged journal entries			
Account	Items	Aud. 1	Aud. 2	Both	Agreed res.
Rev. Domestic	6.643	200	32	25	207
Rev. Foreign	651	12	11	2	21
Expenses	778	10	10	6	14

As expected, only a small portion of the journal entries are tagged as unusual (Revenue Domestic: 3.1 pct., Revenue Foreign: 3.2 pct., Expenses: 1.8 pct.) which in turn results in a high intercoder reliability [32] (Revenue Domestic: 97.3 pct., Revenue Foreign: 97.1 pct., Expenses: 99.0 pct.). However, a closer look at the results reveals that only a relatively small portion of journal entries was tagged as unusual by both auditors. When discussing each journal entry with tagging differences, the auditors mutually agreed in tagging it as unusual as it has been overlooked. Therefore, all journal entries with tagging differences were added to the agreed list of unusual journal entries.

In this regard, both auditors mention that due to the large number of journal entries and attributes tagging was a complex task which made it challenging to stay fully focused throughout the whole process (approx. two hours for all three accounts). Therefore, the tagging differences were not surprising to them. This is in line with accounting and auditing literature which well documents that exposure to large amounts of information can lead to difficulties in identifying relevant information and patterns, with adverse effects on the quality of audit judgment [6, 33, 34]. They also point out that the provided overviews of attribute frequencies were useful for applying a systematic tagging approach especially with regard to the high number of entries on the Revenue Domestic account. It was much more the case that the overviews and the uniform accounting scheme for the reviewed accounts (which results in many uniform journal entries) enabled the auditors to tag a large number of journal entries in a reasonable time.

Despite these difficulties mentioned before, the agreed list of unusual journal entries forms the basis for evaluating the results of the autoencoder network for each analyzed financial account.

4.2. Evaluation results

The evaluation results presented in this section demonstrate to what extent the autoencoder network is capable to learn a model of regular journal entries for each analyzed account based on the prepared dataset consisting of 25 journal entry attributes. At first, outliers detected by the autoencoder network are compared with tagged journal entries (agreed list of unusual journal entries). As ground truth for our evaluation, we choose the results of manual classification by the involved external auditors. Corresponding figures display the initial results of the comparison. In these figures, the journal entries are plotted with the attribute posting date on the x-axis and the $L(x, y)$-error on the y-axis. The black horizontal line shows the threshold for the respective account. All points above this line are classified as outliers by the autoencoder network. The red-colored dots highlight those journal entries that are classified as outliers by external auditors. In this way, we can visualize the number of matches between our classification and that of external auditors. For a more in-depth analysis common statistical measures for evaluating classification models are calculated.

In a second step, the results of the autoencoder were discussed with the external auditors. As our focus lies on audit effectiveness, during this discussion only the false positives and false negatives were considered. This final step should highlight the extent to which the autoencoder network is able to identify unusual entries that may have been overlooked by the auditors. In this step, the auditors had the opportunity to re-evaluate the discussed journal entries. For comparison, figures are also created for the adjusted results and corresponding statistical measures are calculated.

1. Account revenue domestic With 6.643 journal entries posted to this account, the transformation of categorical attributes to the one-hot-encoded representation increases the number of attributes for this account to 357. After training the autoencoder for 500 epochs and combining the result with the outlier detection based on the numerical attribute (amount) our autoencoder network detected 702 outliers. Figure 4 depicts the results for this account. The corresponding statistical measures are listed in table 2.

For the largest account in our sample, the autoencoder achieves good results in terms of F-score and recall. However, 16 tagged journal entries are not identified by the autoencoder (false negatives). Unusual characteristics of the journal entries only identified by the auditors are: 1) customers that have only been used once, 2) users with a low number of postings on this

account, 3) postings close to year-end, and 4) postings with a creation date before the posting date. In the initial results, with the chosen threshold a relatively large number of false positives are marked (511 journal entries). During the discussion, this number was reduced as the auditors re-evaluated 35 journal entries. These entries are re-evaluated as for the respective customers the posted amounts are unusually high. In some instances, also a more complex accounting scheme was applied, which leads to an unusual structure of the journal entry. With this re-evaluation, the F-score slightly increases to 0.48. The remaining false positives are mainly identified as outliers because of the amount whereas the auditors would apply another amount threshold for this account.

Table 2. Statistics for account revenue domestic

	Initial result	After discussion
True negatives	5.925	5.925
True positives	191	226
False negatives	16	16
False positives	511	476
F-score	0.42	0.48
Accuracy	0.92	0.93
Precision	0.27	0.32
Recall	0.92	0.93

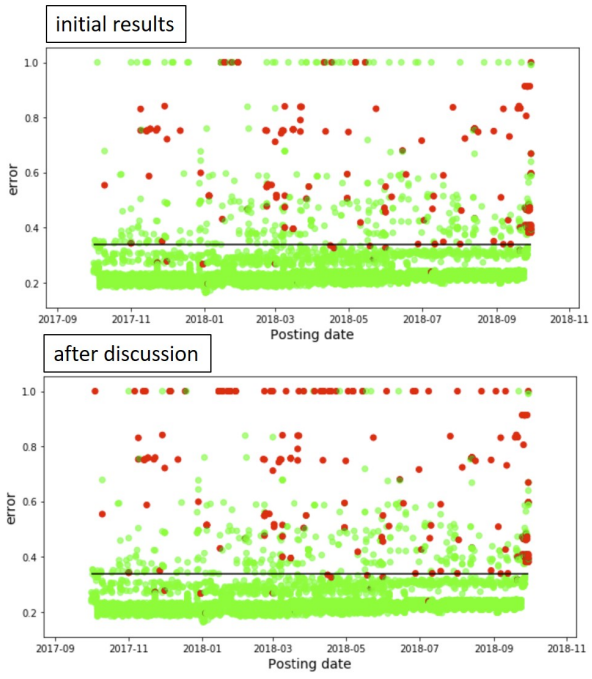

Figure 4. Comparison for account revenue domestic

2. Revenue foreign With 651 journal entries posted to this account, the transformation of categorical attributes to the one-hot-encoded representation increases the number of attributes for this account to 125. After training the autoencoder for 400 epochs and combining the result with the outlier detection based on the numerical attribute (amount) our autoencoder network detected 77 outliers. Figure 5 depicts the results for this account. The corresponding statistical measures are listed in table 3.

For the account with the smallest population size in our sample, the autoencoder achieves better results in terms of F-score and recall compared to the large account. All manually tagged journal entries are identified (no false negatives). Again, in the initial results with the chosen threshold, a relatively large number of false positives are marked (56 journal entries). During the discussion, this number was reduced as the auditors re-evaluated 5 journal entries. These entries were created with a normal system function (transaction code), but posted to the unusual debit side of this account. Due to their rather small amount, these were initially ignored by the auditors. With this re-evaluation the F-score increases to 0.51. Again, the remaining false positives are mainly identified as outliers because of the amount.

Table 3. Statistics for account revenue foreign

	Initial result	After discussion
True negatives	574	574
True positives	21	26
False negatives	0	0
False positives	56	51
F-score	0.43	0.51
Accuracy	0.91	0.92
Precision	0.27	0.34
Recall	1.0	1.0

3. Expenses With 778 journal entries posted to this account, the transformation of categorical attributes to the one-hot-encoded representation increases the number of attributes for this account to 62. After training the autoencoder for 400 epochs and combining the result with the outlier detection based on the numerical attribute (amount) our autoencoder network detected 54 outliers. Figure 6 depicts the results for this account. The corresponding statistical measures are listed in table 4.

For the second account with small population size in our sample, the autoencoder achieves the best results in terms of F-score and recall. All manually tagged journal entries are identified (no false negatives). With

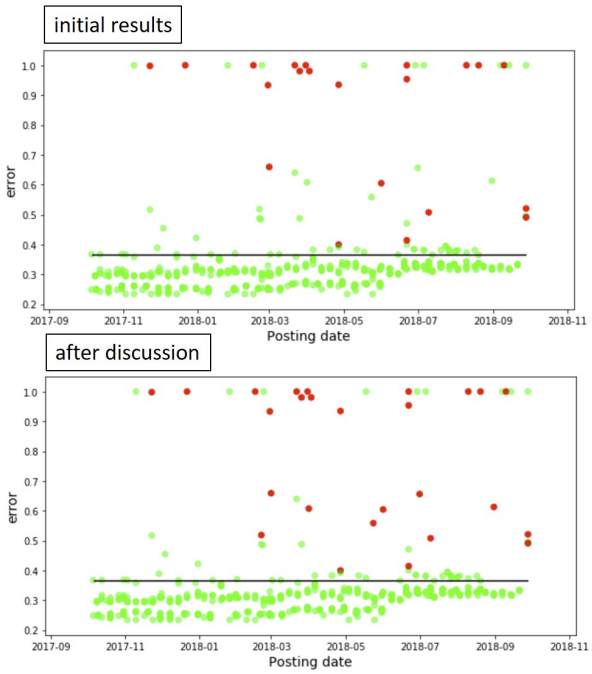

Figure 5. Comparison for account revenue foreign

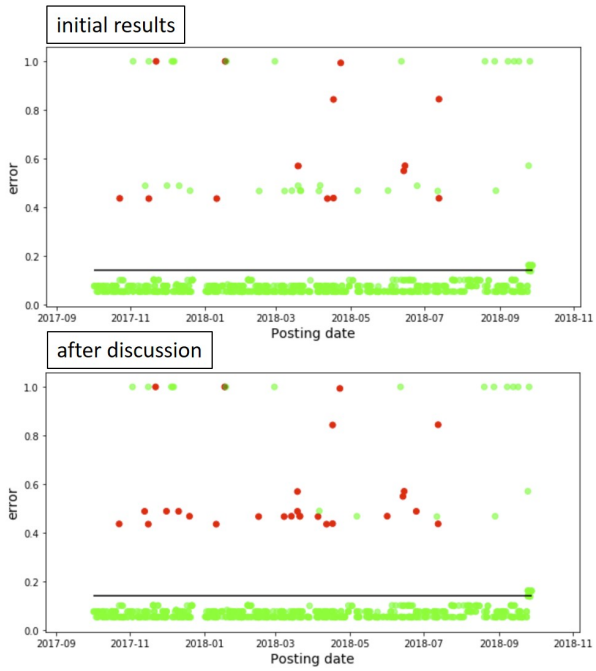

Figure 6. Comparison for account expenses

the chosen threshold a reasonable number of false positives are identified (40 journal entries). During the discussion, this number was reduced as the auditors re-evaluated 13 journal entries. For these journal entries, an unusual accounting scheme was applied that has been initially overlooked by the external auditors. With this re-evaluation, the F-score increases to 0.67 and the number of false positives is reduced to 27 journal entries. Again, the remaining false positives are mainly identified as outliers because of the amount.

Table 4. Statistics for account expenses

	Initial result	After discussion
True negatives	724	724
True positives	14	27
False negatives	0	0
False positives	40	27
F-score	0.41	0.67
Accuracy	0.95	0.97
Precision	0.26	0.5
Recall	1.0	1.0

In summary, it can be stated that the use of autoencoder networks can provide valuable support for the identification of unusual journal entries in an audit context. This is also confirmed by the involved external auditors. This especially holds true for datasets with a highly unbalanced distribution of anomalous and non-anomalous journal entries. In such a case, it can be expected that autoencoder networks achieve superior results compared to statistical sampling methods as they are used in the current audit practice. Furthermore, the evaluation results reveal that the autoencoder also performs on small population sizes, in our case with even better results in terms of F-score and recall. Conversely, further investigations are required to reduce the number of false positives. In this context, numeric attributes such as the amount must be taken into account.

5. Conclusion and future work

The identification of unusual business transactions and related journal entries steaming either from fraud or unintentional errors is one of the main tasks for public accountants when auditing financial statements to give a profound opinion on whether the financial statements are free from material misstatements. On the one hand, the increasing amount of electronically available accounting data poses a major challenge for auditors when considering that conventional audit procedures comprise a sample-based manual review of business transactions. On the other hand, the data allow for applying various analytical techniques to identify irregularities and complex pattern within the complete population of journal entries. However, in the current audit practice, the full potential of such techniques is not yet realized. Mainly less complex techniques like static rules are applied that check only a few journal

entry attributes at a time.

To broaden the range of more complex data analytics for audit purposes, we presented a deep learning-based approach for detecting unusual journal entries. An autoencoder network is applied to the population of journal entries from three financial accounts extracted from a real-world dataset. An account-based data selection is used to investigate whether an autoencoder network also generates applicable results on small populations.

The evaluation results demonstrate that the reconstruction error of autoencoder networks can be used as an indicator for unusual journal entries. When comparing the outliers identified by the autoencoder with the list of manually tagged journal entries, our approach achieved high F-scores (Revenue Domestic: 0.48, Revenue Foreign: 0.51, Expenses: 0.67) and a high recall (Revenue Domestic: 0.93, Revenue Foreign: 1.0, Expenses: 1.0) for all three analyzed accounts. These evaluation results demonstrate the applicability and usefulness of autoencoder networks in practical audit scenarios.

The insights from the implementation of our approach and the evaluation results also point to several opportunities for further research. Especially, the domain-driven handling of numerical attributes like the amount needs to be further investigated as especially the amount is of high importance for external auditors. In particular, domain-specific concepts like materiality are worth to be considered [35, 36]. The same applies to guidelines for the weighting of individual journal entry attributes and the selection of appropriate thresholds to reduce false positives.

From a more technical perspective, a comparison with other anomaly detection techniques (like it is done in [29] for large datasets) on small populations is one promising research direction. In addition, comparing the performance of autoencoder networks that are applied to a complete population of journal entries (e.g. a full fiscal year) with networks that are trained on purposively selected subsets (e.g. individual accounts like in our approach) would further address the research question regarding required population sizes for deep learning application scenarios in the audit domain [6]. Besides population size, also qualitative aspects could be of interest. For our paper, accounts are used for which a uniform accounting scheme and a high degree of automation are assumed. It remains an open research question whether the results presented in this paper are reproducible for accounts that are to a higher degree subject to professional judgment. By addressing the mentioned research questions, also valuable insights could be gained for an improved transfer of deep learning techniques into audit practice.

References

[1] PricewaterhouseCoopers, "PwC's 2018 Global Economic Crime and Fraud Survey," 2018.

[2] International Auditing and Assurance Standards Board (IAASB), "International Standard on Auditing (ISA) 200. Overall Objectives of the Independent Auditor and the Conduct of an Audit in Accordance with International Standards on Auditing," 2009.

[3] International Auditing and Assurance Standards Board (IAASB), "International Standard on Auditing (ISA) 315 (Revised). Identifying and Assessing the Risks of Material Misstatement," 2018.

[4] W. R. Knechel and S. E. Salterio, *Auditing: Assurance and risk*, p. 215. Routledge, 2016.

[5] International Federation of Accountants (IFAC), "International Standard on Auditing (ISA) 240. The Auditor's Responsibilities Relating to Fraud in an Audit of Financial Statements," 2018.

[6] H. Issa, T. Sun, and M. A. Vasarhelyi, "Research ideas for artificial intelligence in auditing: The formalization of audit and workforce supplementation," *Journal of Emerging Technologies in Accounting*, vol. 13, no. 2, pp. 1–20, 2016.

[7] J. Kokina and T. H. Davenport, "The emergence of artificial intelligence: How automation is changing auditing," *Journal of Emerging Technologies in Accounting*, vol. 14, no. 1, pp. 115–122, 2017.

[8] R. H. Von Alan, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS Quarterly*, vol. 28, no. 1, pp. 75–105, 2004.

[9] V. K. Vaishnavi and W. Kuechler, *Design science research methods and patterns: innovating information and communication technology*. Crc Press, 2015.

[10] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1," ch. Learning Internal Representations by Error Propagation, pp. 318–362, Cambridge, MA, USA: MIT Press, 1986.

[11] P. Baldi and K. Hornik, "Neural networks and principal component analysis: Learning from examples without local minima," *Neural Networks*, vol. 2, no. 1, pp. 53 – 58, 1989.

[12] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, (New York, NY, USA), pp. 1096–1103, ACM, 2008.

[13] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[14] S. Hawkins, H. He, G. Williams, and R. Baxter, "Outlier detection using replicator neural networks," in *International Conference on Data Warehousing and Knowledge Discovery*, pp. 170–180, Springer, 2002.

[15] M. Sakurada and T. Yairi, "Anomaly detection using autoencoders with nonlinear dimensionality reduction," in *Proceedings of the MLSDA 2014 2Nd Workshop on Machine Learning for Sensory Data Analysis*, MLSDA'14, (New York, NY, USA), pp. 4:4–4:11, ACM, 2014.

[16] D. Kwon, H. Kim, J. Kim, S. C. Suh, I. Kim, and K. J. Kim, "A survey of deep learning-based network anomaly detection," *Cluster Computing*, pp. 1–13, 2017.

[17] A. Ahmi and S. Kent, "The utilisation of generalized audit software (gas) by external auditors," *Managerial Auditing Journal*, vol. 28, no. 2, pp. 88–113, 2012.

[18] T. Sun and M. A. Vasarhelyi, "Deep learning and the future of auditing: How an evolving technology could transform analysis and improve judgment.," *CPA Journal*, vol. 87, no. 6, 2017.

[19] R. L. Braun and H. E. Davis, "Computer-assisted audit tools and techniques: analysis and perspectives," *Managerial Auditing Journal*, vol. 18, no. 9, pp. 725–731, 2003.

[20] S. Wang, "A comprehensive survey of data mining-based accounting-fraud detection research," in *2010 International Conference on Intelligent Computation Technology and Automation*, vol. 1, pp. 50–53, IEEE, 2010.

[21] D. A. Appelbaum, A. Kogan, and M. A. Vasarhelyi, "Analytical procedures in external auditing: A comprehensive literature survey and framework for external audit analytics," *Journal of Accounting Literature*, vol. 40, pp. 83–101, June 2018.

[22] N. B. Hitzig, "Statistical sampling revisited," *The CPA Journal*, vol. 74, no. 5, pp. 30–35, 2004.

[23] S. Bay, K. Kumaraswamy, M. G. Anderle, R. Kumar, and D. M. Steier, "Large scale detection of irregularities in accounting data," in *Sixth International Conference on Data Mining (ICDM'06)*, pp. 75–86, IEEE, 2006.

[24] M. McGlohon, S. Bay, M. G. Anderle, D. M. Steier, and C. Faloutsos, "Snare: a link analytic system for graph labeling and risk detection," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1265–1274, ACM, 2009.

[25] A. Argyrou, "Auditing journal entries using self-organizing map," in *18th Americas Conference on Information Systems 2012, AMCIS 2012, 9 August 2012 through 12 August 2012, Seattle, WA, USA*, pp. 986–995, 2012.

[26] A. Argyrou, "Auditing journal entries using extreme value theory," *Auditing*, vol. 7, pp. 1–2013, 2013.

[27] M. Jans, M. G. Alles, and M. A. Vasarhelyi, "A field study on the use of process mining of event logs as an analytical procedure in auditing," *The Accounting Review*, vol. 89, no. 5, pp. 1751–1773, 2014.

[28] M. Werner, "Materiality maps–process mining data visualization for financial audits," in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[29] M. Schreyer, T. Sattarov, D. Borth, A. Dengel, and B. Reimer, "Detection of Anomalies in Large Scale Accounting Data using Deep Autoencoder Networks," *arXiv e-prints*, Sept. 2017.

[30] D. Appelbaum, A. Kogan, and M. A. Vasarhelyi, "Big data and analytics in the modern audit engagement: Research needs," *AUDITING: A Journal of Practice & Theory*, vol. 36, no. 4, pp. 1–27, 2017.

[31] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A Design Science Research Methodology for Information Systems Research," *Journal of Management Information Systems*, vol. 24, pp. 45–77, Dec. 2007.

[32] M. Lombard, J. Snyder-Duch, and C. C. Bracken, "Content analysis in mass communication: Assessment and reporting of intercoder reliability," *Human communication research*, vol. 28, no. 4, pp. 587–604, 2002.

[33] H. Brown-Liburd, H. Issa, and D. Lombardi, "Behavioral implications of big data's impact on audit judgment and decision making and future research directions," *Accounting Horizons*, vol. 29, no. 2, pp. 451–468, 2015.

[34] M. G. Alles, A. Kogan, and M. A. Vasarhelyi, "Putting continuous auditing theory into practice: Lessons from two pilot implementations," *Journal of Information Systems*, vol. 22, no. 2, pp. 195–214, 2008.

[35] N. Martinov and P. Roebuck, "The assessment and integration of materiality and inherent risk: An analysis of major firms' audit practices," *International Journal of Auditing*, vol. 2, no. 2, pp. 103–126, 1998.

[36] D. V. Budescu, M. E. Peecher, and I. Solomon, "The joint influence of the extent and nature of audit evidence, materiality thresholds, and misstatement type on achieved audit risk," *Auditing: A Journal of Practice & Theory*, vol. 31, no. 2, pp. 19–41, 2012.