Using Autoencoders for Data-Driven Analysis in Internal Auditing

Jakob Nonnenmacher
University of Oldenburg
jakob.nonnenmacher@uol.de

Felix Kruse
University of Oldenburg
felix.kruse@uol.de

Gerrit Schumann
University of Oldenburg
gerrit.schumann@uol.de

Jorge Marx Gómez
University of Oldenburg
jorge.marx.gomez@uol.de

Abstract

New challenges in internal auditing are created as all areas of companies are digitalized. These challenges are forcing internal auditing to implement more and more data-driven procedures. Auditing is increasingly using artificial intelligence methods such as neural networks to overcome these challenges. Since in internal auditing labels are usually not available at the beginning of an audit engagement, unsupervised methods have to be used. We used autoencoders as an unsupervised method, which we evaluated for its use in auditing in a practical case study with an international automobile manufacturer. For the case study, two real-world, non-financial data sets from production-related processes were provided. The results of the case study show that the use of autoencoders can support auditors in the audit execution and in the audit planning process step to improve the quality of the internal audit engagement.

1. Introduction

Multiple researchers believe that the auditing profession will be transformed by data analytic technologies and artificial intelligence (AI) [1–3]. They predict that new technologies will have a huge impact on the auditing profession through automation, a larger audit scope, shortened processing times and as a result, an improved auditing quality [2]. These new technologies are also used to react to changes that companies made in their business processes [4]. These changed business processes generate large amounts of data which makes some manual auditing methods obsolete, or even impossible [4]. This makes auditing an ideal domain for AI with all big four accounting firms investing in it [5] and already using some AI functionalities for the external audit [5, 6].

Research also points to the relevance of these technologies for the internal audit [2, 7]. Internal auditing is an independent and objective assurance and consulting activity whose purpose is to improve an organization's operations [8]. For this purpose, audits are conducted through defined audit engagements. During an internal audit engagement, drawing sample transactions and comparing those against guidelines is still a large component of the collection and evaluation of information [9]. The main disadvantage of this approach is that relevant information could be in those transactions which have not been picked as a sample. This is especially problematic when considering that auditors often choose a sample that is smaller than one percent of the original data population [10]. One attempt to mitigate this disadvantage is to utilize data analytic technologies to perform full population testing [11] in which the complete dataset is checked against programmed, rule-based tests. These tests are created based on the kind of problems auditors anticipate. Full population testing against these hand-crafted rules fails when the data contains deviations which cannot be found by just checking against guidelines or process descriptions. These deviations could for example be novel fraud attempts. Fraudsters can and will find new ways that are not anticipated but which show up as deviations in the data [12]. Additionally, it can be difficult for auditors to derive rules in the first place or decide what to test since they often deal with complex topics which they have not experienced before [13].

These disadvantages could be addressed by using AI in the form of neural networks (NNs) to perform unsupervised anomaly detection. It can be used to find suspicious patterns in the data without checking against the process guidelines and thus without requiring process knowledge. Furthermore, it can be used to preselect a subset of the data as potentially problematic which can then be investigated in more detail by the auditor. Through this, unsupervised anomaly detection can supplement rule-based full population testing,

At the beginning of an internal audit engagement, there is usually no knowledge of which data points might be correct or incorrect. This is due to how internal auditing functions by auditing systems of one company which have not been audited before or that

URI: https://hdl.handle.net/10125/71317
978-0-9981331-4-0

have greatly changed from previous audits. The reason for this is the risk-based approach for selecting areas, departments or processes to audit [8]. One criteria in the selection process is how long an area has not been audited [14]. This means areas that have not been audited recently have a higher risk-score and thus a higher likelihood of being selected for an audit. Due to this, a longer period of time passes before an area and, with it, a system is audited again. With approaches like "continuous delivery" gaining prevalence, new software features are released and used at a highly accelerated pace [15]. Due to this, systems go through multiple new versions before another audit, making any possibly collected labels obsolete. This makes the obtaining of labels impractical [16, 17] and thus supervised methods in internal auditing virtually unusable. This leads to unsupervised machine learning being the only feasible kind of machine learning for internal auditing in most cases. Unsupervised learning can find interesting transformations of the data without the need for labels [18] and these transformations can reveal anomalies in the data. Therefore, we propose the use of unsupervised NNs to detect anomalies during the audit of processes.

To assess unsupervised NNs' usefulness for internal auditing, we implemented them on data collected for a completed internal audit engagement which was conducted on an industrial process and already has a finalized audit report. Due to the existence of an audit report, the results achieved with the NNs can be compared against a baseline to evaluate the technology's suitability for internal auditing. The question guiding our research is: "Is there potential for unsupervised neural networks in internal auditing and how can they be integrated into the audit process?".

The remainder of the paper is structured as follows: in the next section, related research regarding unsupervised NNs and the utilization of NNs in auditing is presented and research gaps are highlighted. After that, our case study conducted in the internal auditing department of an international automobile manufacturer is presented followed by an evaluation. Our paper closes with a conclusion and a presentation of further research opportunities.

2. Related research

Internal audit functions within companies are considering or have already started to implement data analytic and artificial intelligence applications. Nonetheless, empirical research on the technologies used by internal auditing is mostly confined to the use of continuous auditing and generalized audit tools. This narrow focus creates a considerable gap in the literature, presenting plenty of opportunities for future research. There are already some studies which examine the effectiveness of specific tools for auditing, such as process mining, but there are many other tools that still have to be examined [19]. Researchers studying emerging technologies can thus provide important insights to help internal auditing in identifying which technologies have potential. This is especially important since internal audit functions "that use technology-based auditing techniques are more efficient and effective and stakeholders rely more on their work" [19]. This is what this study addresses by examining the potential of unsupervised anomaly detection, specifically NNs, for internal auditing.

2.1. Neural networks in auditing

There are a number of different methods for unsupervised anomaly detection, one of which are NNs [20]. First studies on how NNs could assist auditors in their work, for example to detect management fraud [21, 22], have been conducted in the 1990s. Considering that the research on this topic started almost 30 years ago, it is surprising that the use of technologies like NNs is still not widespread in the auditing profession [1].

Since the 1990s, there have been few studies on possible NN use in auditing. The areas covered in the studies range from auditors' opinion prediction [23] to financial distress prediction [24] and for the largest part, fraud prediction and detection [25–29].

The majority of those studies use publicly available financial data. The studies by Gaganis, Ravisankar et al., Omar, and Hajek and Henriques all use financial variables or ratios to detect financial statement fraud [26–29]. Fernández-Gámez et al. use financial and corporate governance variables to predict the opinion of auditors using a NN [23]. Chakraborty and Sharma use financial ratios to predict a corporation's financial health [24]. Most of these studies use supervised NNs.

Despite the wide variety of studies of NNs in auditing, there seems to be a lack of studies on their use in internal auditing as well as on their use on non-financial data and with unsupervised approaches. This is why in this paper we utilize unsupervised NNs in the form of Autoencoders (AEs) on real-world non-financial data to explore their potential for internal auditing.

2.2. Autoencoders for anomaly detection

AEs as a kind of NN can be used for unsupervised anomaly detection. The goal of unsupervised anomaly detection is to find deviations in the data without the use of labels. An anomaly can be defined as "an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism." [20]. In the context of an internal audit, this different mechanism could be a deviation from an intended process or even fraud.

An AE is a NN which is trained to replicate its input into its output. It possesses a hidden middle layer which is smaller, i.e. consists of less neurons, than the outer layers and represents an encoded version of the input [12]. Since an AE is trained in such a way that it replicates its input as close as possible with its output [30], it is forced to focus on the most important aspects of the data for replicating during training [31]. This means it must adjust its weights in such a way that it learns a compressed representation of its input in its middle layer. This is shown in Figure 1.

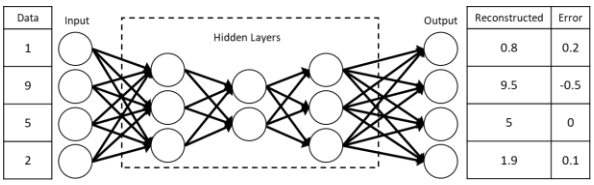

Figure 1. Autoencoder network

An indicator to measure the ability of the AE to replicate a certain input is the reconstruction error (RE). The RE is obtained by calculating the difference between the input values of a sample and the output the AE is generating based on it [30]. The basic idea of using an AE for anomaly detection is that after training, it will reconstruct normal entries better than anomalous entries. Anomalous entries can thus be identified based on their higher RE.

One of the challenges which comes with using NNs is that their results are highly dependent on several parameters which must be set and are the so-called hyperparameters. Hyperparameters are for example the architecture of the NN, i.e. the number of layers and neurons, the size of the training batches and the number of epochs the NN is trained for. With supervised learning, these parameters can be fine-tuned based on already known labels [20]. With unsupervised learning, this possibility does not exist. That is why different methods must be used to determine the right hyperparameters or to select the model with a good hyperparameter configuration when using unsupervised NNs. For generating a number of different models with different configurations, gridsearch can be used [32]. After creating the models, one model must be selected as the model to be used for identifying anomalies.

There are three methods for selecting an AE model in the literature and we evaluate them as part of our study. These three methods are: using the model with the lowest average RE [33], a method using the RE histogram [34] and the "Incremental Training Set Refinement" method which is a method that works without gridsearch and uses a single model per dataset [35].

AEs have already been successfully applied for outlier or anomaly detection in a variety of domains outside of auditing. Chen et al. utilize AE ensembles to detect outliers in various public outlier detection datasets. They compare their methods to multiple other state-of-the art outlier detection methods and achieve superior results on most of the datasets [30]. There have also been multiple successful applications of AEs for detecting fraud and irregularities in government spending [33, 36] as well as for fraud and anti-money laundering investigations [31].

Only recently, studies have explored the use of unsupervised NNs for auditing. Schreyer et al. use AEs to detect anomalous journal entries to support financial statement audits and fraud investigations. In their approach, they utilize injected anomalies to guide the training of the AE which means the approach is not entirely unsupervised [12]. Schultz and Tropmann-Frick build on the approach but work without the injected anomalies achieving similar results [37]. Both studies show promise for the use of AEs in auditing.

3. Case study

A case study has been used to evaluate the value of AE NNs for detecting anomalies in data of audited processes to support internal auditing in a real-world setting. With this setting, non-financial and not publicly available data can be used addressing the identified research gaps. For conducting the case study, an approach outlined by Benbasat et al. has been used [38]. With this approach, first the unit of analysis is defined which is the use of AEs within the internal auditing department. Then the design is determined which in this paper is a single-case study design. Afterwards, the data sources, which are primarily the process data and the AE models, are selected. Finally, the data is analyzed and the results are presented. The different parts of the case study are described in more detail in the following sections.

3.1. Environment

The case study has been conducted in the internal auditing department of an international automobile manufacturer. To evaluate the developed AE, data from an already conducted audit has been used. Due to this, the results of the AE can be compared directly with the labeled data from the audit report highlighting which benefits the use of an AE could offer.

In the internal auditing department, the AE would be utilized during an audit engagement whose general process is illustrated in Figure 2.

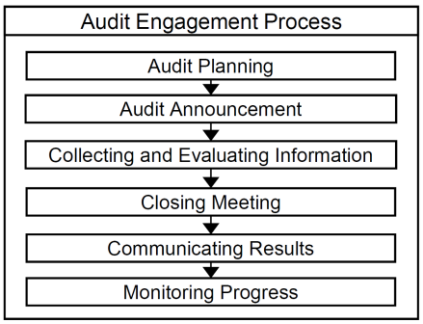

Figure 2. Internal audit engagement process

In the audit planning step, a plan for the audit including its objectives and scope is developed. Afterwards, the auditee is informed of the audit. During the audit itself, the auditors collect and analyze data and check this evidence against the relevant documentation to derive findings. These findings are reconciled with the auditee in a closing meeting [39]. Afterwards, a report which contains the findings and appropriate actions is compiled and distributed to the responsible managers. Finally, the progress on the implementation of these actions is monitored [40].

3.2. Data

The data used in this case study is data from a production-related permit process with each row referring to an individual process entity which possessed multiple, numerical as well as categorical attributes. To prepare the data for the use in the NN, all date values have been converted to epoch. Then all NA values in the dataset have been replaced with 0. After that, all categorical attributes have been one-hot encoded [18] and then all numerical values have been scaled to a range between 0 and 1.

For selecting the attributes to use with the NNs, two different approaches have been chosen. For one approach, no knowledge of the process has been assumed and all attributes have been used with the NN. This approach is named the *all attributes* approach. For the other approach, some knowledge of the process has been assumed and the attributes have been selected based on the audit report. This approach is named the *process knowledge* approach.

Based on these approaches two different datasets have been created. These two datasets have then been filtered and split again as to have two distinct datasets for two subsidiaries becoming subsidiary 1 (S1) and subsidiary 2 (S2). This has been done to reduce the training time and to avoid skewed results due to process differences between the subsidiaries. In the end, this led to four different datasets. For a general evaluation of the approach, three public anomaly detection datasets, the Lympho, the Seismic and the E.coli dataset[1] used in the AE paper of Chen et al. [30], have been used as well. The Lympho, the Seismic and the E.coli dataset are prepared as described in the paper of Chen et al. [30]. A description of all used datasets can be found in Table 1.

Table 1. Used datasets

Dataset	Number of Entries	Number of Attributes	Confirmed Anomalies
S1 *All attributes*	7882	33	25.59%
S1 *Process knowledge*	7882	5	25.59%
S2 *All attributes*	709	33	15.09%
S2 *Process knowledge*	709	5	15.09%
Lympho Dataset	148	18	4.05%
Seismic Dataset	2548	14	6.58%
E.coli Dataset	336	7	2.68%

3.3. Creation of models

For developing the NNs, the framework keras with a tensorflow backend [18] has been used. To create the NN models, gridsearch has been used. The values to use for the gridsearch have been chosen based on other AE studies and previous papers offering recommendations for how to configure NNs. This has been done since in an unsupervised setting, which internal auditing is at the beginning of a new engagement, model parameters cannot be tuned based on existing labels. Unlike for supervised approaches, where cross-validation can be used to approximate the optimal parameters of a model, this is not possible for unsupervised approaches [20].

The learning rate range has been chosen based on the recommendation by Bengio [41], who suggests the learning rate should be smaller than 1 and greater than 10^{-6}. The batch size has been chosen in accordance to Bengio as well who suggests 32 as a good default size with typical values between one and a few hundred

[1] available at https://archive.ics.uci.edu/ml/datasets

[41]. The number of epochs have been used in accordance with Bergstra and Bengio [32]. The two optimizers have been picked because they are the ones mentioned in the reviewed papers on unsupervised NNs for anomaly detection [30, 42]. The architecture range is chosen similar to Schreyer et al. [12] by creating deeper architectures by adding hidden layers of size 2^k neurons where k=2,3,..9. To prevent the dying Rectified Linear Unit (ReLU) problem, leaky ReLUs have been used [12]. To prevent exploding gradients for deeper architectures, gradient clipping has been added to the grid as well [43]. For the *all attributes* approach, the values presented in Table 2 and for the *process knowledge* approach, the values presented in Table 3 have been chosen. The architecture for the *process knowledge* approach has been designed shallower, since it uses less attributes than the *all attributes* approach and therefore has a much smaller input size. This is also done to prevent the model from just learning to replicate the process dataset [12] which could happen with the high amount of parameters compared to attributes fed into it.

Table 2. Gridsearch hyperparameter grid – all attributes approach

Hyperparameter	Value Range
learning_rate	[0.00001, 0.1]
batch_size	[32,512]
optimizer	['RMSprop', 'Adam']
architecture	[Inputsize-3-Inputsize; Inputsize-512-256-128-64-32-16-8-4-3-4-8-16-32-64-128-256-512-Inputsize]

Table 3. Gridsearch hyperparameter grid – process knowledge approach

Hyperparameter	Value Range
learning_rate	[0.00001, 0.1]
batch_size	[32,512]
optimizer	['RMSprop', 'Adam']
architecture	[Inputsize-3-Inputsize; Inputsize-8-4-3-4-8-Inputsize]

Based on these values, 64 models have been created for every dataset. To select the model to use for identifying possible anomalies in the dataset, different methods have been attempted. Since no labels are available at the beginning of a new audit engagement, the goal was to find an approach that could be used for selecting a good model without the need for labels. These are:

1. Using the model with the lowest average RE (LARE) [33].
2. Using the RE histogram (REH) [34].
3. The Incremental Training Set Refinement (ITSR), a method which works without gridsearch [35].

3.4. Lowest average reconstruction error

The first method was to select the model with the LARE which is the best model according to this method. To evaluate the method of selecting the model with the LARE as the best model, the top-k-precision as described in [12] has been calculated for each of the 448 models created in the gridsearch. For the calculation, k has been set to the number of known anomalies in the dataset based on the existing audit report. After the training of the models, the respective datasets are one more time fed through the models for determining the REs for all entries of all datasets. This RE has been averaged over the individual entries for each model. Based on the average RE, the model with the LARE for each dataset is selected. The top-k-precision calculation works by sorting the entries for each model according to their RE, selecting the subset of k entries with the highest RE, with k=2017 for the S1 datasets, k=107 for the S2 datasets, k=6 for the Lympho, k=170 for the Seismic and k=9 for the E.coli dataset, and then determining the amount of confirmed anomalies in that subset.

For all seven datasets, the model with the LARE was not the model with the best top-k-precision. The model with the lowest RE is actually the model with the lowest top-k-precision for the E.coli dataset which can be seen in Figure 3 and only with an increasing RE does the top-k-precision increase.

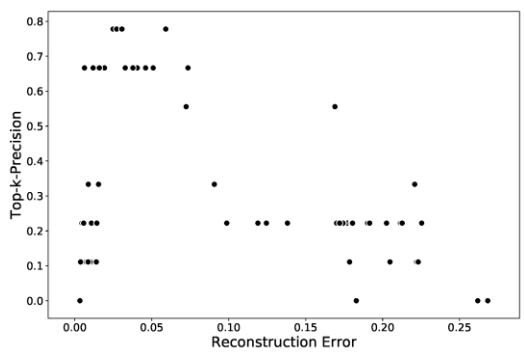

Figure 3. Top-k-precision plotted against RE for the models of the E.coli dataset

This is an indicator that the method might not be reliable for selecting the best or even a good model. A possible explanation for this is that in the models with the lowest RE, the AE most likely has also learned to reconstruct the anomalous entries well. This leads to a smaller difference or even overlap in RE values between normal and anomalous entries, which can make it impossible to distinguish between normal and anomalous entries [35]. This can explain why for the

E.coli dataset, the model with the LARE is actually the model with the lowest top-k-precision.

3.5. Reconstruction error histogram

Another method for selecting a good model is described by Ordway-West et al. and uses the REH of each model. The method is based on creating a histogram of each model's RE and calculating its Full Width Half Max (FWHM) [34].

Ordway-West et al. show how the RE increases with increasing FWHM and the number of anomalies fluctuates with a low FWHM but then stabilizes with an increasing FWHM. They do not provide a reasoning for why this stabilization in the number of found anomalies is occurring. The idea when utilizing this method was that with a stabilizing number of anomalies the quality of the model would also stabilize.

To replicate this method, first the models created in the gridsearch for the E.coli dataset were used. Since their paper provides no information in regards to how the histogram has been created, the methods for choosing the bin width included in the scipy library's "numpy.histogram" function [44] have been tested with the method leading to the clearest plot being Sturges rule [45]. The percentage for selecting a RE to define the number of anomalies using the cumulative distribution function has been set to 95% since the one of 99.99% used in the paper lead to zero found anomalies for each histogram width. This is most likely due to the fact that in the case of Ordway-West et al.[34], the number of entries was probably more than a million [34] while the E.coli dataset only has 336 entries. With the 336 entries of the E.coli dataset, less than one entry would be selected with the percentage of 99.99%.

After following the steps, a result where the number of anomalies fluctuates for a very low FWHM but starts to stabilize with an increasing FWHM was achieved with the E.coli dataset. To see whether this stabilization in the number of anomalies also transfers to a stabilization in the quality of the model the top-k-precision has been calculated and a stabilization can be observed, when the number of anomalies stabilizes. When attempting the method with the S2 *process knowledge* dataset, no stabilization at all could be seen neither in the number of anomalies nor in the quality of the models. Since the method does not work for the S2 process dataset it has not been attempted with the remaining datasets because there would be no way to reliably know whether it has worked in an unsupervised setting at the beginning of a new audit.

3.6. Incremental training set refinement

To address the problem of AE learning to replicate the anomalous entries just as well as the normal entries, Beggel et al. suggest a method which involves the stepwise removal of possible anomalies from the training set to prevent the AE from learning to reconstruct them. In addition to that, the ITSR method uses a one class support vector machine for the stepwise removal of potential anomalies. At the core of the ITSR method is an adversarial AE. An adversarial AE is an AE with an added discriminator network like the one shown in Figure 4.

The first part of this discriminator network is also the encoder part of the AE, which is still trained to replicate its input in its output. In addition, the discriminator network is trained to differentiate samples coming from a prior distribution and from the middle layer. In the training step of the AE, the weights of the discriminator are fixed and the loss of the discriminator is used to adjust the weights of the encoder in such a way that the latent space in the middle layer starts to resemble the prior distribution.

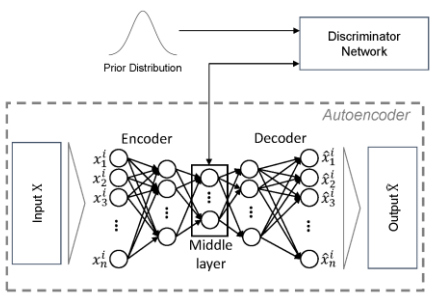

Figure 4. Adversarial autoencoder

With the latent layer, an adversarial AE can provide another indicator for anomalies. Normal entries are more likely to be in the dense area of the approximated distribution in the middle layer and anomalous entries are more likely to be in the sparse area of the approximated distribution in the middle layer. This likelihood of belonging to the distribution can be calculated for each entry and can be used as an additional anomaly score. The specific steps of the ITSR method are described in the paper of Beggel et al. [35].

The method has been implemented in accordance with that description. Deviating from the paper, leaky ReLU is used as an activation function instead of ReLU, to avoid the "dying ReLU" problem.

The method has then been executed for all datasets. The top-k-precision has initially been calculated using both the RE and the likelihood. For all datasets, using the RE lead to the higher top-k-

precision which is why the RE should be used for selecting a subset of anomalies when using this method.

For both the S2 process and the S1 process dataset, the method achieved a top-k-precision of 100%. This means that this method works very well when the anomalies are obvious from the data or when the auditor has some idea about which attributes might point towards anomalies. The clear results that can be achieved are shown as an example for the S1 *process knowledge* dataset in Figure 5 in which the confirmed anomalies are marked in grey and normal entries in black.

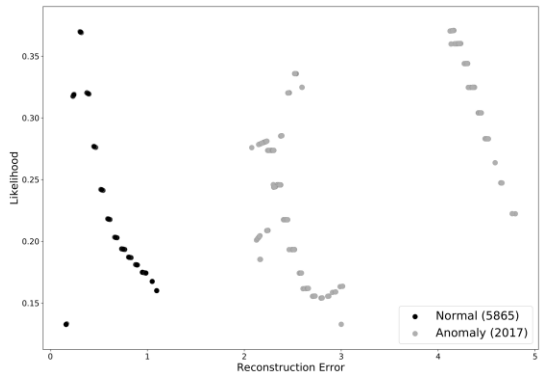

Figure 5. Likelihood and RE of individual entries for the S1 process dataset using the ITSR method

3.7. Evaluation

The results of the case study have been evaluated in a quantitative and qualitative evaluation.

3.7.1. Quantitative Evaluation. Our case study shows that NN models can successfully support the selection of a subset of entries from the dataset which contain a high percentage of confirmed anomalies. When comparing the different methods that have been attempted, which are the LARE method, the REH method and the ITSR method, only the ITSR method is suitable for utilization in a completely unsupervised setting. This is because the LARE method returns the worst model for some datasets whereas the REH method fails to identify a specific model at all for certain datasets. The top-k-precision of the methods for which a result could be calculated, which are the LARE method and the ITSR method, is shown in Figure 6. Averaged over all datasets, the ITSR method offers about a 20% better performance than the LARE method. The ITSR method most likely provides the best results, since it removes potential anomalies during the training, preventing the AE from overfitting on these. Another possible explanation is, that the combination of different anomaly detection methods, can improve on the performance of the individual methods by combining their assumptions about the "normal" behavior of the data [20].

The ITSR method has been evaluated with 7 different datasets: the Lympho, the Seismic and the E.coli datasets as well as the *process knowledge* and *all attributes* datasets of S1 and S2. Even in a completely unsupervised setting where no previous knowledge about the dataset exists, like at the beginning of an audit, the ITSR method can be used to select a subset which contained nearly 50% of all confirmed anomalies, as shown with the *all attributes* approach. A possible explanation for cases in which the number of confirmed anomalies in the selected subset is lower is that what is an anomaly from the pure data perspective is not necessarily an anomaly from an auditor's perspective. It could also be that certain actual anomalies have been uncovered by the AE which have not been found during the conducted audit.

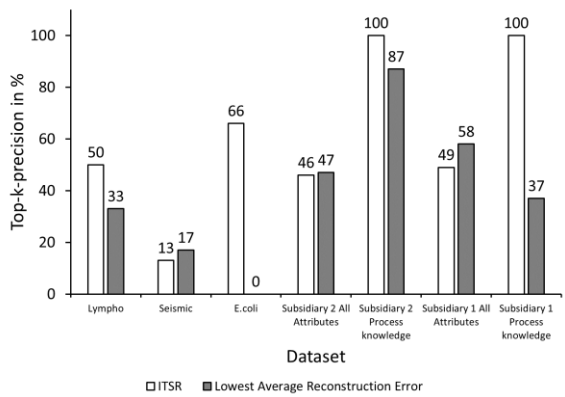

Figure 6. Top-k-precision in % for the LARE model and the ITSR model per dataset (REH method is not included because models could not be selected)

If an auditor already has some ideas about which attributes might reveal anomalies in the process, this knowledge can be used to generate even better results. This is shown in the *process knowledge* approach where those attributes have been preselected which would most likely return the confirmed anomalies. Attempting this approach with the ITSR method, subsets could be selected which contained 100% of the confirmed anomalies for the S2 and the S1 *process knowledge* dataset. This approach could, for example, be utilized in the audit execution step when the auditor already has specified the objectives in more detail and narrowed down the number of interesting attributes

due to an initial screening of documents in the audit planning step.

3.7.2. Qualitative evaluation. For the qualitative evaluation of the case study results, a workshop has been conducted using an approach outlined by Ørngreen and Levinsen [46]. The approach consists of a presentation of the research goal. In the next step, the findings and what the analysis shows is presented. In the last step, room is given to the workshop's participants to contribute their interpretations and ideas. Afterwards, the results are documented based on event recall and analyzed in regard to the research goal. As part of the workshop, it was also explored how unsupervised anomaly detection in the form of NNs could best be integrated into the existing auditing process.

The workshop has been conducted with both auditors as well as data scientists from the internal auditing department the study has been conducted in. During it, the motivation for our research, the approach, the case study, and the results of the case study have been presented. Afterwards the results were discussed, and feedback was given from both auditors and data scientists. The following feedback was received during the workshop:

Feasibility Auditors as well as data scientists agreed that the presented approach can be beneficially utilized in an audit and could help reduce the time to gather findings. One of the voiced concerns was how this approach could be integrated considering an already defined audit plan with little room or time to deviate from previously defined objectives. That means that if the AE results would generate more possible leads, it could be difficult to pursue them all. They furthermore pointed out the ultimate necessity of process knowledge to validate findings and to write the report.

Integration into the audit process The approach could be used to amend other analytical techniques used during the audit. Some auditors suggested that the approach could be the most beneficial if it is not just used in the audit execution phase but in the audit planning phase to help set objectives or to generate ideas.

The previously discussed concern that there is little room to deviate from an existing audit plan with defined objectives that was raised by an auditor could be addressed by integrating the approach not only into the execution part of the audit but also into the planning phase. This way, the objectives could be shaped or partially defined based on the results of the AE. This would be a new audit approach in which the data is already gathered before the planning phase of an audit engagement and analyzed with the unsupervised NN. The anomalies that are found based on the data could then inform the creation of the plan for the audit engagement and could help the auditors in defining objectives and focus areas of the audit. The affected process steps and the potential impacts of this approach are shown in Figure 7.

The found anomalies could support the auditors in narrowing down the number of data points they have to investigate. This would give the auditor more time to focus on those fewer points. Through this, the approach might even enable the identification of more findings. Due to how the AE functions, it could help to identify errors and their causes which would not be found when just checking against a process documentation or a set of predefined rules.

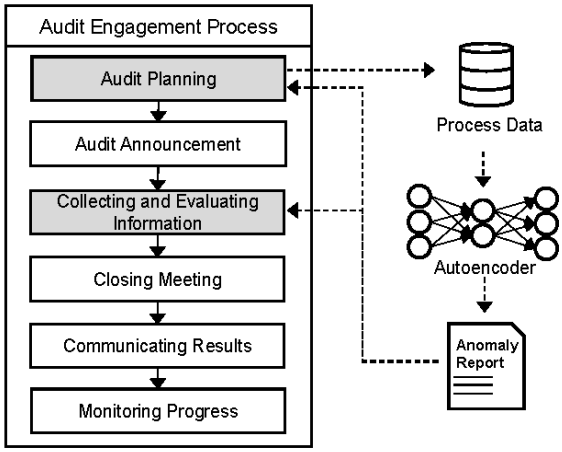

Figure 7. Extended audit engagement process with AE

What is important as well for the integration is that the company who wants to employ the AE as part of their internal auditing, plans for its utilization and makes sure that the flexibility that might be necessary because of its results can be accommodated for.

The point regarding how necessary process knowledge is, which was raised in the workshop, highlights the importance of this kind of understanding for internal auditing. To be able to write up results into an audit report and to define appropriate countermeasures, process knowledge is required. AEs and process knowledge could be utilized together by an auditor with the use of visualizations of the individual REs of attributes.

4. Conclusion and further research

This paper describes the use of unsupervised NNs in internal auditing. At the beginning of an audit engagement, labels for which data points are correct or

not correct are usually not available thus unsupervised NNs such as AEs have to be used. An AE was used in our case study in the internal auditing department of an international automobile manufacturer. For the case study, two real-world data sets from production processes were used. In addition, three public datasets to benchmark the methods were also used. The main problem with using the AE for anomaly detection was to determine the model with the best quality in an unsupervised setting. For this purpose, the methods "Lowest Average RE", "RE Histogram" and "Incremental Training Set Refinement" (ITSR) were implemented. The ITSR procedure provided the best results in the case study. The qualitative evaluation of the results shows that AEs can support audit execution as well as audit planning within internal auditing.

Through this, the study provides additional insights into the application of data analytic methods within auditing, addressing an important research gap [19]. Future research should focus on conducting several practical case studies to further validate the use of the AE with the ITSR method. The use of unsupervised methods in the internal auditing process should be further explored as well by utilizing different methods and different process data.

Because of the difficulty of obtaining labels in internal auditing which would be valid over the course of multiple audits, the study has focused on the evaluation of the unsupervised NNs within internal auditing. Nonetheless, the method could also be used to detect anomalies within an external audit context. This should be investigated in further studies.

One aspect that was not in the scope of our study but could be an important avenue for further research is how the auditor is able to get to findings from the detected anomalies. Since many modern unsupervised anomaly detection methods are “black boxes”, it is not always obvious why a specific entry has been selected as an anomaly. To support the auditor in discussing detected anomalies with the auditee and ultimately in deriving findings to write a report, so-called anomaly explanation methods could be used. In the case of AEs, “Shapley values” have already been successfully utilized as a method to provide explanations for why an entry has been detected as an anomaly [47]. Explanation and interpretability methods could be an important aspect of making unsupervised approaches useful for auditing. Thus, it should be investigated how anomaly explanations could support auditors when utilizing unsupervised methods.

References

[1] Gepp, A., M.K. Linnenluecke, T.J. O’Neill, and T. Smith, "Big data techniques in auditing research and practice: Current trends and future opportunities", Journal of Accounting Literature, 40, 2018, pp. 102–115.

[2] Dai, J. and M.A. Vasarhelyi, "Imagineering Audit 4.0", Journal of Emerging Technologies in Accounting, 13(1), 2016, pp. 1–15.

[3] Issa, H., 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, 13(2), 2016, pp. 1–20.

[4] Chiu, V., Q. Liu, and M.A. Vasarhelyi, "The Development and Intellectual Structure of Continuous Auditing Research", in Continuous Auditing, D.Y. Chan, V. Chiu, and M.A. Vasarhelyi, Editors. 2018. Emerald Publishing Limited.

[5] Kokina, J. and T.H. Davenport, "The Emergence of Artificial Intelligence: How Automation is Changing Auditing", Journal of Emerging Technologies in Accounting, 14(1), 2017, pp. 115–122.

[6] Sun, T. and M. Vasarhelyi, "Embracing Textual Data Analytics in Auditing with Deep Learning", The International Journal of Digital Accounting Research, 2018, pp. 49–67.

[7] Salijeni, G., A. Samsonova-Taddei, and S. Turley, "Big Data and changes in audit technology: Contemplating a research agenda", Accounting and Business Research, 2018, pp. 1–25.

[8] Institute of Internal Auditors (IIA), "Die Internationalen Grundlagen für die berufliche Praxis der Internen Revision (IPPF)", DIIR - Deutsches Institut für Interne Revision e.V., 2017.

[9] Eric, B.P., "Evolution of Auditing: From the Traditional Approach to the Future Audit", in Continuous Auditing, A.-A. Abdullah, D.Y. Chan, V. Chiu, and M.A. Vasarhelyi, Editors. 2018. Emerald Publishing Limited.

[10] No, W.G., K. Lee, F. Huang, and Q. Li, "Multidimensional Audit Data Selection (MADS): A Framework for Using Data Analytics in Audit Data Selection Process", Accounting Horizons, 2019.

[11] Appelbaum, D., A. Kogan, and M.A. Vasarhelyi, "Big Data and Analytics in the Modern Audit Engagement: Research Needs", Auditing: A Journal of Practice & Theory, 36(4), 2017, pp. 1–27.

[12] Schreyer, M., T. Sattarov, D. Borth, A. Dengel, and B. Reimer, "Detection of Anomalies in Large Scale Accounting Data using Deep Autoencoder Networks", arXiv:1709.05254, 2017.

[13] Nguyen, L. and Y. Kohda, "Toward a model of wisdom determinants in the auditing profession", Proceedings of the 50th Hawaii International Conference on System Sciences, 2017.

[14] Krüger, H.A. and J.M. Hattingh, "A combined AHP-GP model to allocate internal auditing time to projects", ORiON, 22(1), 2006, pp. 59–76.

[15] Shahin, M., M.A. Babar, and L. Zhu, "Continuous Integration, Delivery and Deployment: A Systematic Review on Approaches, Tools, Challenges and Practices", IEEE Access, 5, 2017, pp. 3909–3943.

[16] Jans, M., N. Lybaert, and K. Vanhoof, "Data mining for fraud detection: Toward an improvement on internal control systems?", 2007.
[17] Kim, Y. and A. Kogan, "Development of an Anomaly Detection Model for a Bank's Transitory Account System", Journal of Information Systems, 28(1), 2014, pp. 145–165.
[18] Chollet, F., Deep learning with Python, Manning Publications Co, Shelter Island, New York, 2018.
[19] Christ, M.H., M. Eulerich, R. Krane, and D.A. Wood, "New Frontiers for Internal Audit Research", Available at SSRN 3622148, 2020.
[20] Aggarwal, C.C., ed., Outlier analysis, Springer, 2017.
[21] Coakley, J.R. and C.E. Brown, "Artificial Neural Networks Applied to Ratio Analysis in the Analytical Review Process", Intelligent Systems in Accounting, Finance and Management, 2(1), 1993, pp. 19–39.
[22] Fanning, K., K.O. Cogger, and R. Srivastava, "Detection of Management Fraud: A Neural Network Approach", Intelligent Systems in Accounting, Finance and Management, 4(2), 1995, pp. 113–126.
[23] Fernández-Gámez, M.A., F. García-Lagos, and J.R. Sánchez-Serrano, "Integrating corporate governance and financial variables for the identification of qualified audit opinions with neural networks", Neural Computing and Applications, 27(5), 2016, pp. 1427–1444.
[24] Chakraborty, S. and S.K. Sharma, "Prediction of corporate financial health by Artificial Neural Network", International Journal of Electronic Finance, 1(4), 2007, p. 442.
[25] Chen, H.-J., S.-Y. Huang, and C.-L. Kuo, "Using the artificial neural network to predict fraud litigation: Some empirical evidence from emerging markets", Expert Systems with Applications, 36(2), 2009, pp. 1478–1484.
[26] Gaganis, C., "Classification techniques for the identification of falsified financial statements: A comparative analysis", International Journal of Intelligent Systems in Accounting, Finance & Management, 16(3), 2009, pp. 207–229.
[27] Ravisankar, P., V. Ravi, G. Raghava Rao, and I. Bose, "Detection of financial statement fraud and feature selection using data mining techniques", Decision Support Systems, 50(2), 2011, pp. 491–500.
[28] Omar, N., Z.A. Johari, and M. Smith, "Predicting fraudulent financial reporting using artificial neural network", Journal of Financial Crime, 24(2), 2017, pp. 362–387.
[29] Hajek, P. and R. Henriques, "Mining corporate annual reports for intelligent detection of financial statement fraud – A comparative study of machine learning methods", Knowledge-Based Systems, 128, 2017, pp. 139–152.
[30] Chen, J., S. Sathe, C. Aggarwal, and D. Turaga, "Outlier Detection with Autoencoder Ensembles", in Proceedings of the 2017 SIAM International Conference on Data Mining. 2017.
[31] Paula, E.L., M. Ladeira, R.N. Carvalho, and T. Marzagao, "Deep Learning Anomaly Detection as Support Fraud Investigation in Brazilian Exports and Anti-Money Laundering", in 15th IEEE International Conference on Machine Learning and Applications. 2016.
[32] Bergstra, J. and Y. Bengio, "Random Search for Hyper-parameter Optimization", J. Mach. Learn. Res., 13, 2012, pp. 281–305.
[33] Gomes, T.A., R.N. Carvalho, and R.S. Carvalho, "Identifying Anomalies in Parliamentary Expenditures of Brazilian Chamber of Deputies with Deep Autoencoders", in 16th IEEE International Conference on Machine Learning and Applications. 2017.
[34] Ordway-West, E., P. Parveen, and A. Henslee, "Autoencoder Evaluation and Hyper-Parameter Tuning in an Unsupervised Setting", in IEEE International Congress on Big Data. 2018.
[35] Beggel, L., M. Pfeiffer, and B. Bischl, "Robust Anomaly Detection in Images using Adversarial Autoencoders", arXiv:1901.06355, 2019.
[36] Domingos, S.L., R.N. Carvalho, R.S. Carvalho, and G.N. Ramos, "Identifying IT Purchases Anomalies in the Brazilian Government Procurement System Using Deep Learning", in 16th IEEE International Conference on Machine Learning and Applications. 2017.
[37] Schultz, M. and M. Tropmann-Frick, "Autoencoder Neural Networks versus External Auditors: Detecting Unusual Journal Entries in Financial Statement Audits", in Hawaii International Conference on System Sciences. 2020.
[38] Benbasat, I., D.K. Goldstein, and M. Mead, "The Case Research Strategy in Studies of Information Systems", MIS Quarterly, 11(3), 1987, pp. 369–386.
[39] Deutsches Institut für Interne Revision e.V., "Criteria Catalogue for the Assessment of the Internal Audit System: Annex 1 from DIIR Revisionsstandard No. 3", 2018.
[40] The Institute of Internal Auditors, "International Standards for the Professional Practice of Internal Auditing (Standards)", 2017.
[41] Bengio, Y., "Practical Recommendations for Gradient-Based Training of Deep Architectures", in Neural Networks: Tricks of the Trade: Second Edition, G. Montavon, G.B. Orr, and K.-R. Müller, Editors. 2012. Springer Berlin Heidelberg.
[42] Nolle, T., S. Luettgen, A. Seeliger, and M. Mühlhäuser, "Analyzing business process anomalies using autoencoders", Machine Learning, 107(11), 2018, pp. 1875–1893.
[43] Goldberg, Y., "A Primer on Neural Network Models for Natural Language Processing", arXiv:1510.00726, 2015.
[44] http://www.scipy.org/, accessed 4-1-2019.
[45] Scott, D.W., "Sturges' rule", Wiley Interdisciplinary Reviews: Computational Statistics, 1(3), 2009, pp. 303–306.
[46] Ørngreen, R. and K. Levinsen, "Workshops as a Research Methodology", The Electronic Journal of e-Learning, 15(1), 2017, pp. 70–81.
[47] Antwarg, L., B. Shapira, and L. Rokach, "Explaining Anomalies Detected by Autoencoders Using SHAP", arXiv:1903.02407, 2019.