The International Journal of Digital Accounting Research
Vol. 21, 2021, pp. 1-22
ISSN: 2340-5058

Unsupervised anomaly detection for internal auditing: Literature review and research agenda

Jakob Nonnenmacher. University of Oldenburg. Germany jakob.nonnenmacher@uni-oldenburg.de

Jorge Marx Gómez. University of Oldenburg. Germany jorge.marx.gomez@uni-oldenburg.de

Abstract. Auditing has to adapt to the growing amounts of data caused by digital transformation. One approach to address this and to test the full audit data population is to apply rules to the data. A disadvantage of this is that rules most likely only find errors, mistakes or deviations which were already anticipated by the auditor. Unsupervised anomaly detection can go beyond those capabilities and detect novel process deviations or new fraud attempts. We conducted a systematic review of existing studies which apply unsupervised anomaly detection in an auditing context. The results reveal that most of the studies develop an approach for only one specific dataset and do not address the integration into the audit process or how the results should be best presented to the auditor. We therefore develop a research agenda addressing both the generalizability of unsupervised anomaly detection in auditing and the preparation of results for auditors.

Keywords: Auditing, anomaly detection, unsupervised, outlier detection.

1. INTRODUCTION

Machine learning and data analytic approaches are changing the auditing field (Kokina & Davenport, 2017) and are becoming an important part of auditing's toolbox (No et al., 2019). At the same time, existing approaches are becoming unfeasible (Chiu et al., 2018). The main reason for this is digital transformation.

Submitted October 2020
Accepted January 2021

DOI: 10.4192/1577-8517-v21_1

Digital transformation is affecting the business environment and influences companies in every industry. As more and more processes are transitioning from paper based to digital, more data is generated within companies (Reinsel et al., 2018). A study by the IDC predicts that the global amount of data will grow to 175 Zetabytes by 2025, more than fivefold the amount of data existent in 2018 (Reinsel et al., 2018).

The growing amount of data and the automation of processes presents a unique challenge for auditing (Chiu et al., 2018). Today, auditing still mainly relies on drawing samples of transactions to audit processes during audit engagements (Byrnes, 2018). The main drawback of this approach is that relevant information could be in the transactions which have not been picked to be audited. This is referred to as sampling-risk. When sampling is used, auditors often choose a sample that is smaller than one percent of the data population (No et al., 2019). With the growing amount of data, this approach becomes both obsolete and the sampling risk is aggravated (Chiu et al., 2018).

On the other hand, new opportunities open up due to more and more of a company's information and data being available in digital form. All Big 4 accounting firms are undertaking efforts to utilize data analytics for auditing (Appelbaum et al., 2017). These could enable faster processing of the drawn samples. On top of that, they could also enable a transition from sampling to full population testing (No et al., 2019). Full population testing means that all of the available data instances of a process are tested (No et al., 2019). One of the already existing kinds of full population testing works by applying pre-defined tests to the data based on specific audit objectives (Appelbaum et al., 2017). Other approaches such as continuous auditing try to automate the auditing of processes and systems. This is done by developing systems which automatically apply rules and red flag checks to data of a system or multiple systems on a regular basis to identify exceptions (American Institute of Certified Public Accountants, 2015). A key disadvantage with this way of full-population testing is that it will likely only find errors, mistakes or deviations which were already expected or anticipated. Novel ways of circumventing the intended process within the guidelines or new fraud attempts cannot be found with these hand-crafted rules (Schreyer et al., 2017). This is a problem because unidentified mistakes could open up the door for possible fraud (Jans et al., 2010). On top of that auditors often deal with complex topics which they have not experienced before (Nguyen & Kohda, 2017), which would make it difficult for them to derive rules or decide what to test a priori. And even if they are familiar with the topic, biases in human cognition can interfere with auditors' ability to give credibility to unexpected events (Liu, 2014).

Machine learning approaches can go beyond the capabilities of hand-crafted rules (Hajek & Henriques, 2017). They can be used to find anomalies in large amounts of data (Kokina & Davenport, 2017) which can then be used to identify errors, mistakes, circumvented processes or even fraud. A large part of the machine learning approaches that have been attempted in auditing are using supervised learning (Hajek & Henriques, 2017; Sharma & Panigrahi, 2013). In supervised learning, an algorithm learns to map input data to known targets which are given as labels based on a set of examples (Chollet, 2018). In external auditing, these examples could be cases of fraudulent and non-fraudulent financial statements which would be labeled accordingly (Dutta et al., 2017). This is possible in external auditing since financial statements possess a similar structure across multiple companies due to reporting regulations. This makes it easier to obtain data labels for training supervised machine learning models.

In internal auditing, on the other hand, labels for processes do not exist most of the time. This is because internal auditing audits systems of one company which might have never been audited before or which have changed considerably since the last time they were audited. This is due to the risk-based approach of selecting areas, departments or processes to audit (Institute of Internal Auditors [IIA], 2017). A criteria in this is for example the turnover in a business area but another is how long an area has not been audited (Krüger & Hattingh, 2006). The longer an area has not been audited the higher its risk score grows which at the same time increases the likelihood that it will be selected for an audit. This is why the obtaining of labels (Jans et al., 2007; Kim & Kogan, 2014) and thus the use of supervised methods in internal auditing is in most cases not realistic. Therefore, unsupervised machine learning is often the only feasible option when using machine learning in internal auditing.

Despite a few literature reviews on analytic techniques in auditing (Appelbaum et al., 2018; Gepp et al., 2018) there has been, to the best of our knowledge, no review which addresses the specific situation of internal auditing and the application of unsupervised anomaly detection to enable full population testing within it.

Therefore, this review aims to explore which kind of techniques of unsupervised anomaly detection have been previously applied in auditing to guide further research on how efficient and effective full population testing can be enabled in internal auditing. As part of that advantages and disadvantages of specific techniques are considered and areas for subsequent studies are highlighted.

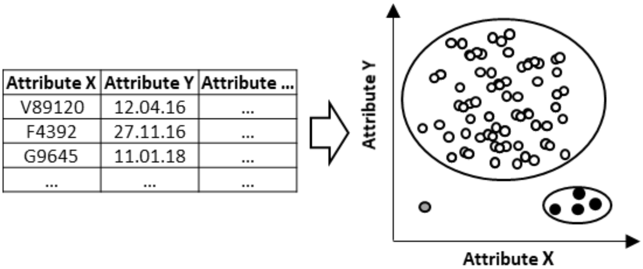
2. THEORETICAL FOUNDATION

The goal of internal auditing is to ensure the correctness and effectiveness of a company's processes and to reduce risks (The Institute of Internal Auditors, 2018). To achieve this goal, internal auditing conducts audits through defined audit engagements (IIA, 2017). During these the auditors identify, evaluate and document adequate information to achieve the objectives of the engagement (IIA, 2017). Through this, they can uncover waste, fraud and wrongdoing (Lipman & Lipman, 2012).

The general idea behind machine learning is that data-processing rules are not programmed but instead learned by a computer based on data. Instead of providing the computer with data and rules to process to get results, the computer is presented with data and in some cases previous results. Based on these, the computer derives rules which can then be applied to before unseen data to obtain results. Machine Learning is especially useful in cases where it would be difficult for humans to define those data processing rules (Chollet, 2018).

Machine learning algorithms in general can be categorized into different types of learning (Chollet, 2018). One of these types is supervised learning which learns to map input data to known targets which are given as labels, based on a set of examples. Multiple applications for deep learning such as image classifications, speech recognition and language translations belong to this group (Chollet, 2018).

Unsupervised learning’s goal on the other hand is to find interesting transformations of the input data without utilizing any targets. These are used for visualizing, compressing or to better understand the correlations in the data used. Clustering is an example of unsupervised learning (Chollet, 2018) and illustrated in Figure 1.



Attribute X	Attribute Y	Attribute …
V89120	12.04.16	…
F4392	27.11.16	…
G9645	11.01.18	…
…	…	…

Figure 1. Illustration of clustering with an anomaly shown in grey

The purpose of using unsupervised machine learning in internal auditing, is the detection of outliers and anomalies to then derive further findings based on these. Aggarwal sees outliers as all data points that could be considered as abnormalities or noise from a pure data perspective whereas an anomaly is an outlier that is of interest to an analyst (Aggarwal, 2017). Anomalies is what internal auditing wants to identify. Both the terms outlier and anomaly are used interchangeably in the literature (Chandola et al., 2009) and there are multiple definitions for them. Hawkins et al. see an anomaly as “an observation, which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism” (2002, p. 170). In another definition given by Barnett and Lewis (1994) and quoted in Hodge and Austin (2004, p. 86) an anomaly is an “observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data”. All these definitions describe anomalies from a slightly different angle but have in common that an anomaly is a data point that differs in a certain way and to a certain degree from all other data points.

To denote whether a data point is normal or anomalous, labels can be used as part of anomaly detection. Retrieving this information for all possible manifestations of the variables of a data point can be difficult. Sometimes it is easier to only get the labels for normal manifestations than it is to get labels for anomalous ones, especially since anomalies do not necessarily demonstrate a fixed behavior. Based on the availability of data labels for the training data, there are different kinds of anomaly detection which can be distinguished (Chandola et al., 2009). When labels for neither normal nor anomalous data are available, which would most likely apply in internal auditing, it is a case of unsupervised anomaly detection. It can be applied in most cases since it does not require the collection of labels at all. The general assumption that the technique works with is that only a small amount of data points in the dataset is anomalous and that the majority is normal (Chandola et al., 2009).

3. METHODOLOGY

Based on the recommendations by Kitchenham and Charters (2007) as well as Okoli (2015), a protocol has been developed to document the exact steps taken during this literature review to reduce the chance of researcher bias.

The main research question which is addressed by this literature review is: “What kind of unsupervised anomaly detection approaches have already been implemented for auditing?”’. Based on this research question two sub-questions have been derived: “What are the limitations of current unsupervised anomaly detection approaches in auditing?” “How can unsupervised anomaly detection approaches in auditing be developed further?”

The search process is a manual search using the databases and the search terms listed in Table 1. The terms hypothesis-free and data-driven have been used since they refer to data analytic techniques which, based on the data, generate new insights (Jung et al., 2000) which is what unsupervised anomaly detection does as well. Google Scholar was chosen as a database to get a broad overview. Emerald Insight has been selected since it contains several relevant accounting publications. The remaining databases have been chosen based on an existing review on artificial intelligence research in accounting by Sutton et al. (2016) which mentions important databases for research at the cross-section of information systems and accounting. In each database, the first 110 results have been taken into consideration and scanned for relevance. If a search revealed duplicates, which was automatically detected by the used reference manager in almost every case, they were not accounted for in the number of results for a certain search.

In the initial search, 168 results were returned. Of these, 81 were found in Google Scholar, 11 in ScienceDirect, 2 in EBSCOhost - Business Source Premier, 9 in IEEEXplore Digital Library, 11 in Wiley Online Library, 43 in the American Accounting Association's Journal Database and 11 in Emerald Insight. The criteria used to filter the retrieved results are presented in Table 2.

Databases	***Search Terms***
• Google Scholar • ScienceDirect • IEEE Xplore Digital Library • Wiley Online Library • American Accounting Association • Emerald Insight • EBSOChost - Business Source Premium	• ("unsupervised learning" OR "unsupervised machine learning") AND (auditing OR "internal auditing" OR audit OR "internal audit") • ("hypothesis-free" OR "data-driven") AND (auditing OR "internal auditing" OR audit OR "internal audit") • ("outlier detection" OR "anomaly detection") AND (auditing OR "internal auditing" OR audit OR "internal audit")

Table 1. Used databases and search terms

Included in the review are all papers which describe practical unsupervised anomaly detection approaches which can be applied during an audit engagement and which mention their possible utilization by auditors. The focus has been on practical implementations since they can highlight what worked or did not work in practice and which potential obstacles still exist. On top of that they can gather valuable feedback from auditors. To get a broad overview of the utilization despite the limited number of studies, studies from both internal and external auditing have been included.

Inclusion Criteria	***Exclusion Criteria***
• Article or dissertation • Describing a practical implementation • Unsupervised machine learning method • Auditing or internal auditing	• Book • Paper which is not in English • Theoretical paper • Review paper • Paper which covers bank fraud (credit card fraud and money laundering) or tax fraud • Paper which contains the term audit but covers topics relating to intrusion detection, cyber- and network security • Paper which does not utilize unsupervised machine learning

Table 2. Used filtering criteria

Excluded are all purely theoretical papers, review papers or surveys, books, and patents. Since this review wants to explore the utilization of unsupervised machine learning in auditing all papers which did not utilize unsupervised machine learning were excluded. Another category of papers which were excluded were those papers which covered bank fraud, such as credit card fraud and money laundering as well as those which covered tax fraud to delimit the scope of this review.

After applying the inclusion and exclusion criteria, 11 relevant papers remained. The main part of excluded results were review or survey papers (43) which did not describe applications of unsupervised machine learning for auditing (Richhariya & Singh, 2012; Sadgali et al., 2019; Sharma & Panigrahi, 2013). Further results which were excluded were theoretical frameworks (12), books (7), did not address auditing (42), addressed auditing but did not meet the inclusion criteria (18), were entirely rule based (2), used supervised approaches (22), were not a paper or dissertation (5), not accessible online (3), not in English (2) or duplicates (1).

Based on the filtered results, forward and backward searches were conducted which revealed 5 further papers. The forward search has been conducted using Google Scholar. To reveal relevant papers, the forward search has been restricted to documents which contained both the terms audit and auditor.

In the next step, we analyzed the selected articles to show relationships, draw conclusions and identify possible directions for further research.

4. RESULTS

In total, 16 papers have been retrieved after all criteria were applied. The studies were published between 2006 and 2019, as can be seen in Figure 2, with an almost even number of publications over the years.

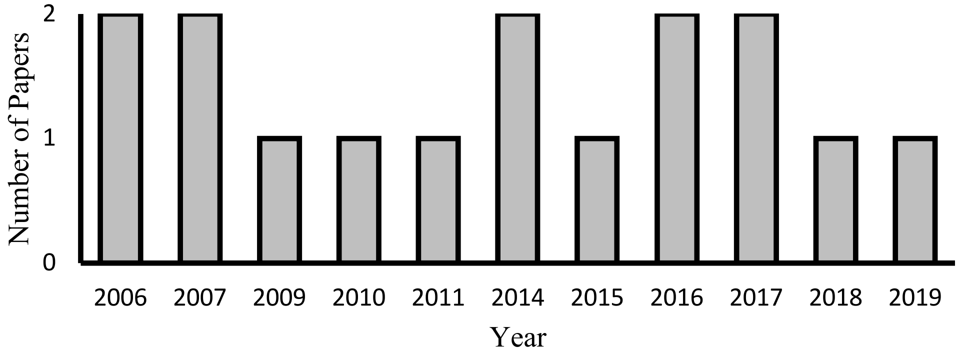


Figure 2. Number of unsupervised anomaly detection papers in auditing from 2006 to 2019

The main reason for the number of results is most likely that external auditing does not require unsupervised approaches and that the main focus of research on applying data analytics in internal auditing is currently on continuous auditing. With its mainly rule based approaches, continuous auditing is gaining prevalence and being adopted in industry (Rikhardsson et al., 2019). Approaches which go beyond those predefined rules will become more important because they can find those problems which are not covered by the predefined rules or can support in defining new rules. This once again highlights the importance of understanding the literature of unsupervised approaches within internal auditing and to draw researcher's attention to this topic and to provide directions for further developing this area.

The topics extracted from the papers and analyzed in this review are the reasons for using unsupervised approaches and the studies' relation to auditing. On top of that, the domain the studies were conducted in and the data that was used for them as well as how the data was prepared, and which methods were used are described. Finally, the advantages, disadvantages and potentials for further research named in the studies themselves are presented.

4.1. Reasons for unsupervised approaches

The goal of using an unsupervised approach was in almost all cases to perform outlier or anomaly detection (Bay et al., 2006; Hagstrom et al., 2018; Jans et al., 2007; Kuna et al., 2014; Lu et al., 2006; Lu, 2007). Different reasons for attempting an unsupervised approach are named by the studies with the most prominent reason being the absence of labels or the impossibility of obtaining them (Jans et al., 2007). A different reason is that with testing beyond the violation of key controls, the auditor can see whether there are anomalies which do not violate any established controls but which may hint towards potential problems (Kogan et al., 2014). Another advantage that comes with using unsupervised methods is that new, previously unknown fraud attempts can be identified (Lu, 2007). This stands in contrast to hand crafted rules which are still used for the most part in fraud detection and which do not generalize to new fraud attempts (Schreyer et al., 2017). On top of that, unusual entries and unintentional errors can be identified (Bay et al., 2006). Another reason for using anomaly detection in the first place is that with a small number of identified anomalies, the internal auditor might have more time for additional checks and an in depth investigation of those anomalies (Thiprungsri & Vasarhelyi, 2011).

4.2. Relation to auditing

A large number of studies mentioned the term auditing only focused on fraud or irregularity detection in general (Gomes et al., 2017; Hagstrom et al., 2018; Jans et al., 2010; Lu et al., 2006; Lu, 2007; Paula et al., 2016) or in a systems audit context (Kuna et al., 2014). Some of the studies explicitly addressed auditing. Of those studies which directly addressed auditing, a majority addressed external or financial auditing (Bay et al., 2006; Byrnes, 2015; Deng & Mei, 2009; Schreyer et al., 2017; Schreyer et al., 2019). Only a few of the studies mentioned internal auditing (Jans et al., 2007; Thiprungsri & Vasarhelyi, 2011) with just the study by Kogan et al. (2014) addressing how anomaly detection could be integrated into internal auditing, in this case through a continuous monitoring approach. With this focus on external auditing and the studies not explicitly addressing the auditing process there exists little guidance on how unsupervised anomaly detection could be integrated into the internal auditing process.

4.3. Domains and Data

The studies were conducted within a variety of different domains such as health insurance (Lu et al., 2006), retail marketing (Lu, 2007), external or financial auditing (Bay et al., 2006; Deng & Mei, 2009; Schreyer et al., 2017; Schreyer et al., 2019), procurement or purchasing (Domingos et al., 2016; Jans et al., 2007, 2010; Kogan et al., 2014; Kuna et al., 2014), life insurance (Thiprungsri & Vasarhelyi, 2011), university student management (Kuna et al., 2014), banking (Byrnes, 2015), export (Paula et al., 2016), parliamentary expenditures (Gomes et al., 2017) and telecommunications (Hagstrom et al., 2018).

The studies used different kinds of data such as insurance claims data (Lu et al., 2006; Thiprungsri & Vasarhelyi, 2011) retail data (Lu, 2007), companies' general ledgers (Bay et al., 2006), purchasing orders and procurement transactions (Domingos et al., 2016; Jans et al., 2010; Kogan et al., 2014; Kuna et al., 2014), goods' receipts and invoices (Jans et al., 2007), financial statements (Deng & Mei, 2009), students’ exam, course and enrollment data (Kuna et al., 2014) credit card customer data (Byrnes, 2015), export operations data (Paula et al., 2016), expenditure data (Gomes et al., 2017), accounting data (Schreyer et al., 2017; Schreyer et al., 2019) and mobile phone invoices (Hagstrom et al., 2018).

4.4. Data Preparation

When using this data, different degrees of data preparation were necessary. Of the studies in which the data preparation and selection is described, most only used a small subset of the initially provided data, for example only a subset of the attributes (Lu et al., 2006). In some cases this is done to mitigate data quality issues (Byrnes, 2015; Thiprungsri & Vasarhelyi, 2011). An example of this is to remove lines with missing values (Domingos et al., 2016).

In most of the studies, other reasons than data quality for selecting a subset of the data were present. In the case of the study by Lu et al. a selection of a subset of the data was necessary since their method could only be applied to attributes which follow specific rules. They identified the attributes that follow the rules by using domain knowledge about the data (Lu et al., 2006).

Other studies, such as the one by Bay et al. (2006) or Jans et al. (2007) used the initial data to engineer new features to use in the anomaly detection (Bay et al., 2006; Byrnes, 2015; Hagstrom et al., 2018; Jans et al., 2007). They did this using information from prior research (Deng & Mei, 2009) or domain knowledge about the data (Bay et al., 2006). In some cases, they even selected the attributes based on the types of anomalies they were expecting (Jans et al., 2007). Some studies consult with experts to decide which features to engineer or select for the anomaly detection (Thiprungsri & Vasarhelyi, 2011). In the study by Jans et al., new features are engineered and a subset of the data is selected based on both domain knowledge and using descriptive statistics (Jans et al., 2010).

Another reason that is given for selecting a subset of the data is that for the particular method, in this case clustering, only attributes which are considered to be relevant should be used (Byrnes, 2015). Some studies take the selection of data one step further by using injected anomalies to guide the parameter setting for their method (Schreyer et al., 2017).

Especially in the case in which the attributes are selected based on the expected anomalies, the argument could be made that this much data engineering diminishes some of the advantages which unsupervised approaches can provide. In this case, the advantage of being able to detect new previously unknown anomalies is impeded.

4.5. Methods

Different unsupervised methods have been utilized to detect anomalies. The type of methods which are used in the majority of the papers are either unsupervised neural networks (NNs) or clustering, with one paper combining the two (Deng & Mei, 2009). When clustering is used, the most prevalent method is k-means (Byrnes, 2015; Hagstrom et al., 2018; Jans et al., 2007; Thiprungsri & Vasarhelyi, 2011). For unsupervised NNs, the method which is used the most is the autoencoder (Domingos et al., 2016; Gomes et al., 2017; Paula et al., 2016; Schreyer et al., 2017; Schreyer et al., 2019). As presented in Table 3, the unsupervised NN has been the most prevalent method in recent years.

Overall, the clustering methods use less attributes than the unsupervised NNs. In all cases the k-means clustering has only been performed on either two (Jans et al., 2007; Thiprungsri & Vasarhelyi, 2011) or four (Byrnes, 2015; Hagstrom et al., 2018) attributes. A possible explanation for this is given by Byrnes (2015) who mentions that if the number of attributes grows over a certain amount, some data mining algorithms cannot produce meaningful results anymore. In the studies which use the autoencoder approach on the other hand, the number of different attributes used was 21 (Gomes et al., 2017), 18 (Domingos et al., 2016; Paula et al., 2016) and 10 (Schreyer et al., 2017). A reason for this could be that autoencoders can handle thousands of different dimensions (Murphree, 2016).

One study combines Benford’s law with reinforcement learning (Lu et al., 2006; Lu, 2007). Benford's law is a statistical analysis method (Lu et al., 2006; Lu, 2007) and can be used to detect irregularities in numerical attributes. Although, its application is limited since it can only be applied when it is used on data which has been recorded from a single naturally growing phenomenon (Lu, 2007). There can also be no built-in minimum or maximum values, the numbers cannot be assigned numbers such as social security numbers and the attribute in the dataset has to have more small value than large value entries (Lu et al., 2006). The reinforcement learning is used to connect multiple suspicious entries from different database tables to one audit case to facilitate the work of the auditor (Lu et al., 2006).

Another study used Naive Bayes (Bay et al., 2006) and one study used continuity equations (Kogan et al., 2014). Continuity equations were used to supervise specific, predefined business process metrics. For this the value for a certain metric, such as the dollar amounts, was predicted and then compared to the actual value.

Year	*Author*	*Method Category*	*Method*	*Attribute #*
2006	Bay et al.	Naive Bayes	Positive Naive Bayes; Naive Bayes using the EM algorithm	50
2006	Lu et al.	Hybrid	Benford's Law and Reinforcement Leaning	1
2007	Jans et al.	Clustering	K-means	2
2007	Lu	Hybrid	Benford's Law and Reinforcement Leaning	na
2009	Deng and Mei	Ensemble	SOM and k-means	47
2010	Jans et al.	Clustering	Latent class clustering algorithm	4
2011	Thiprungsri and Vasarhelyi	Clustering	K-means	2
2014	Kuna et al.	Ensemble	LOF, DBSCAN, C4.5, Bayesian Network and PART	23
2014	Kogan et al.	Continuity Equations	Continuity Equations, Vector Autoregressive Model, Linear Regression Model	4
2015	Byrnes	Clustering	Complete, K-Means, Ward, EM, PAM	4
2016	Domingos et al.	Unsupervised NN	Autoencoder	18
2016	Paula et al.	Unsupervised NN	Autoencoder	18
2017	Gomes et al.	Unsupervised NN	Autoencoder	21
2017	Schreyer et al.	Unsupervised NN	Autoencoder	10
2018	Hagstrom et al.	Clustering	K-means	4
2019	Schreyer et al.	Unsupervised NN	Adversarial Autoencoder	na

Table 3. Methods which have been used in the studies

In only two of the studies, researchers used ensemble methods, which is the combination of multiple unsupervised machine learning methods. One is based on Local Outlier Factor (LOF), DBSCAN, C4.5, a bayesian network and Projective Adaptive Resonance Theory (PART) while still relying on a number of handcrafted rules to combine the methods (Kuna et al., 2014). They combine the methods in an effort to reduce the number of false-positives (Kuna et al., 2014). One of the studies combines Self-organizing maps (SOM) and K-means clustering. This is done to first reduce the number of attributes for the clustering as well as to use the silhouette index of the resulting clusters to select the parameters for both the SOM and k-means (Deng & Mei, 2009). This way the silhouette index is used as validity measure for both the SOM as well as k-means, which is useful for setting the parameters of the method when there is no knowledge about the used data.

4.6. Advantages, disadvantages, and potential for further research

Almost all methods achieve good results according to the authors of the studies (Jans et al., 2010; Kuna et al., 2014; Lu et al., 2006; Schreyer et al., 2019; Thiprungsri & Vasarhelyi, 2011) and some mention their approaches can enable a more effective audit (Kogan et al., 2014) as well as support auditors in adding value (Byrnes, 2015). They nonetheless name certain gaps or disadvantages that exist with the current results as well as certain requirements for making the unsupervised methods usable. One of these is that the expertise of people with domain knowledge is necessary to evaluate the detected anomalies (Thiprungsri & Vasarhelyi, 2011). Another one is that pure anomaly detection, without further information for the auditor, can return too many results which are not actually interesting to an auditor (Bay et al., 2006). This can lead to a heightened cost for the audit function since the auditor has to spend time investigating them (Bay et al., 2006). Also, if there are a lot of false positives and the auditor has no way of identifying why an entry has been selected as an anomaly, the auditor might deem the system to be unreliable. One way of addressing this problem is to establish a ranking for the found anomalies so that the entries which the method identifies as most anomalous are investigated first (Byrnes, 2015). The problem with this is that those entries which the method identifies as very anomalous could still be false positives. Only one study describes an approach that is used to reduce the number of false positives. The study achieves this reduction in the number of false positives by combining multiple outlier detection methods (Kuna et al., 2014).

To be able to utilize unsupervised methods in auditing, it is important that their results are interpretable and understandable and also can directly guide the auditor towards specific findings for further review (Bay et al., 2006). Ideally, the auditor should be able to understand why a method selected an entry as anomalous. Bay et al. (2006) address this problem by using a supervised classifier on top of their unsupervised approach to then identify individual attributes which contribute the most to the anomaly score but they do not manage to provide a solution in an entirely unsupervised scenario. Knowing why an entry has been selected as an anomaly can make the results more actionable for an auditor a well as strengthen their confidence in a particular method. It could also make it easier for an auditor to screen out false positives before starting a time intensive investigation.

Multiple studies raise the point that outlier detection is only the first step (Domingos et al., 2016; Gomes et al., 2017; Jans et al., 2007; Paula et al., 2016; Thiprungsri & Vasarhelyi, 2011) but do not address how they could make their method more understandable and their results easier to interpret for facilitating the next step of investigating the anomalies and converting them into audit findings. Even though the point of needing understandable results is made (Bay et al., 2006) none of the studies addressed how this can be achieved in an unsupervised setting.

One point which is almost exclusively addressed by the studies which use clustering are validity measures. Validity measures can be used to confirm the correct parameter selection for a machine learning algorithm even in cases where no labels for the data are available. They use for example the elbow method (Hagstrom et al., 2018; Thiprungsri & Vasarhelyi, 2011) to decide how many clusters to use. Apart from the approach by Deng and Mei (2009), none the unsupervised NN studies discuss how they want to assure validity in an unsupervised setting. The study by Schreyer et al. (2017) has a method to select the parameters for their algorithm but it is dependent on known injected anomalies which diminishes the advantages of using an unsupervised approach.

5. DISCUSSION

The literature review presents the unsupervised anomaly detection approaches which have been attempted in auditing so far. A majority of the studies discuss auditing only peripherally. Only one study directly addresses internal auditing but entirely in a continuous auditing context. None of the studies address how the unsupervised anomaly detection approaches can be best integrated into the regular internal auditing process.

Another aspect which is not addressed sufficiently by the studies is how the results can be made more understandable and actionable for the auditors and how the unsupervised models can be made more interpretable to achieve that.

Some of the studies use a very high degree of domain knowledge, even anticipating which kind of anomalies to expect. This stands in stark contrast to two points. One of them is that a key advantage of unsupervised anomaly detection methods is the ability to identify previously unknown and novel errors or fraud attempts. This ability is diminished if the attributes are mainly preselected based on the expected errors. It also raises the question of how applicable the method would still be in a case where the auditor only has a limited understanding of the underlying process. In these cases, the unsupervised methods which require a large degree of data preprocessing based on expected anomalies would not be applicable. The potential of unsupervised anomaly detection lies in a broad, undirected search to identify fraud and process weaknesses.

Another aspect that can be pointed out in regard to the studies is that most of them develop their method for one specific kind of data if not dataset. Apart from the fact that k-means clustering seems to be used on smaller datasets, there is also no general guidance on which method should be used for which kind of datasets. This means that there is little to no guidance for an auditor on which method to select for which kind of data. Since most of the methods use a high degree of domain knowledge to prepare the algorithm for one kind of data the question of practicality in cases of new audits is also raised. This would be especially important if the auditor had to have a considerable time and effort investment for each new audit in which they wanted to apply the method.

Findings in literature review	***Research questions to address***
Most of the studies only use one method on one kind of data	What kind of methods should be applied to which kind of auditing data? What kind of methods can be applied across multiple kinds of data?
Only one of the studies describes how the unsupervised approach could be integrated in internal auditing and only within a continuous auditing context	How can unsupervised machine learning approaches be integrated within internal auditing?
Most of the studies only use one kind of method	Can methods be combined to provide better results?
Almost all studies address the fact that detecting anomalies based on the unsupervised method is only the first step	What kind of further processing can be used to help the auditor in working with the results of the unsupervised method?

Table 4. Unsupervised anomaly detection in internal auditing: Findings and research opportunities

Multiple conclusions for further developing unsupervised machine learning approaches for internal auditing can be taken from the review. One of them is that an approach with no domain knowledge at all is most likely not realistic. At least some degree of domain knowledge on the data level must exist in any case to first select the data to be used and then to pre-process it for the specific method. Based on the insights of the literature review we derived multiple findings and research opportunities, which are presented in Table 4.

What the literature review further shows, is that the two most prevalent methods for unsupervised anomaly detection in auditing are k-means clustering and autoencoder NNs. Of the two, the autoencoder has been used in the more recent publications. The applicability of Benford's law is limited to very specific areas and requires a high degree of domain knowledge but the approach of combining multiple detected anomalies to an audit case could be further explored, maybe in conjunction with different unsupervised learning methods.

For being able to apply a method to different kinds of data, validity measures seem also important. They could be used to guide the parameter selection of the model even in cases in which new data is used and no labels exist. For this, as well as to reduce the number of false positives in cases with little domain knowledge for the auditor, the use of ensemble methods should be further explored as well.

Our review shows that there are already promising attempts of utilizing unsupervised anomaly detection in auditing. For enabling unsupervised anomaly detections integration into internal auditing further research is necessary both from a technological as well as an organizational perspective.

6. REFERENCES

Aggarwal, C. C. (2017). *Outlier analysis*. Springer. https://doi.org/10.1007/978-3-319-47578-3

American Institute of Certified Public Accountants. (2015). *Audit Analytics and Continuous Audit: Looking Toward the Future*. https://www.aicpa.org/content/dam/aicpa/interestareas/frc/assuranceadvisoryservices/downloadabledocuments/auditanalytics-lookingtowardfuture.pdf* Accessed 20 October 2020.

Appelbaum, D. A., Kogan, A., & Vasarhelyi, M. A. (2017). Big Data and Analytics in the Modern Audit Engagement: Research Needs. *AUDITING: A Journal of Practice & Theory*, *36*(4), 1–27. https://doi.org/10.2308/ajpt-51684

Appelbaum, D. A., Kogan, A., & Vasarhelyi, M. A. (2018). Analytical procedures in external auditing: A comprehensive literature survey and framework for external audit analytics. *Journal of Accounting Literature*, *40*, 83–101. https://doi.org/10.1016/j.acclit.2018.01.001

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (3. ed. 1994, repr. with corr. Oct. 1994). *Wiley series in probability and mathematical statistics : Applied probability and statistics.*

Bay, S., Kumaraswamy, K., Anderle, M., Kumar, R., & Steier, D. (2006). Large Scale Detection of Irregularities in Accounting Data. *Sixth International Conference on Data Mining (ICDM'06)*, 75–86. https://doi.org/10.1109/ICDM.2006.93

Byrnes, P. E. (2015). Developing automated applications for clustering and outlier detection. Advance online publication. https://doi.org/10.7282/T3R78H7Q

Byrnes, P. E. (2018). Evolution of Auditing: From the Traditional Approach to the Future Audit. In A.-A. Abdullah, D. Y. Chan, V. Chiu, & M. A. Vasarhelyi (Eds.), *Rutgers Studies in Accounting Analytics. Continuous Auditing* (pp. 285–297). Emerald Publishing Limited. https://doi.org/10.1108/978-1-78743-413-420181014

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Comput. Surv.*, *41*(3), 1–58. https://doi.org/10.1145/1541880.1541882

Chiu, V., Liu, Q., & Vasarhelyi, M. A. (2018). The Development and Intellectual Structure of Continuous Auditing Research. In D. Y. Chan, V. Chiu, & M. A. Vasarhelyi (Eds.), *Continuous Auditing* (pp. 53–85). Emerald Publishing Limited. https://doi.org/10.1108/978-1-78743-413-420181003

Chollet, F. (2018). *Deep learning with Python*. Manning Publications Co.

Deng, Q., & Mei, G. (2009). Combining self-organizing map and K-means clustering for detecting fraudulent financial statements. *2009 IEEE International Conference on Granular Computing*, 126–131. https://doi.org/10.1109/GRC.2009.5255148

Domingos, S. L., Carvalho, R. N., Carvalho, R. S., & Ramos, G. N. (2016). Identifying IT Purchases Anomalies in the Brazilian Government Procurement System Using Deep Learning. *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 722–727. https://doi.org/10.1109/ICMLA.2016.0129

Dutta, I., Dutta, S., & Raahemi, B. (2017). Detecting financial restatements using data mining techniques. *Expert Systems with Applications*, *90*, 374–393. https://doi.org/10.1016/j.eswa.2017.08.030

Gepp, A., Linnenluecke, M. K., O'Neill, T. J., & Smith, T. (2018). Big data techniques in auditing research and practice: Current trends and future opportunities. *Journal of Accounting Literature*, *40*, 102–115. https://doi.org/10.1016/j.acclit.2017.05.003

Gomes, T. A., Carvalho, R. N., & Carvalho, R. S. (2017). Identifying Anomalies in Parliamentary Expenditures of Brazilian Chamber of Deputies with Deep Autoencoders, 940–943. https://doi.org/10.1109/ICMLA.2017.00-33

Hagstrom, R., Carvalho, R., Faria, H., Melo, T., Luz, W., & Solis, P. (2018). Brazilian Republic Presidency Mobile Telephony Consumption Cost Reduction with Outliers Detection. *2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC)*, 760–766. https://doi.org/10.1109/ICIVC.2018.8492738

Hajek, P., & Henriques, R. (2017). Mining corporate annual reports for intelligent detection of financial statement fraud – A comparative study of machine learning methods. *Knowledge-Based Systems*, *128*, 139–152. https://doi.org/10.1016/j.knosys.2017.05.001

Hawkins, S., He, H., Williams, G., & Baxter, R. (2002). Outlier Detection Using Replicator Neural Networks. In Y. Kambayashi, M. Arikawa, & W. Winiwarter (Eds.), *Lecture Notes in Computer Science: Vol. 2454. Data Warehousing and Knowledge Discovery: 4th International Conference, DaWaK 2002 Aix-en-Provence, France, September 4-6, 2002 Proceedings* (Vol. 2454, pp. 170–180). Springer-Verlag Berlin Heidelberg. https://doi.org/10.1007/3-540-46145-0_17

Hodge, V., & Austin, J. (2004). A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review*, *22*(2), 85–126. https://doi.org/10.1023/B:AIRE.0000045502.10941.a9

Institute of Internal Auditors. (2017). *Die Internationalen Grundlagen für die berufliche Praxis der Internen Revision (IPPF).* Institute of Internal Auditors (IIA). https://www.diir.de/fileadmin/fachwissen/standards/downloads/IPPF_2017_Standards__Version_6.1___20180110.pdf* Accessed 20 October 2020.

The Institute of Internal Auditors. (2018). *Definition of Internal Auditing.* https://na.theiia.org/standards-guidance/mandatory-guidance/pages/definition-of-internal-auditing.aspx* Accessed 20 October 2020.

Jans, M., Lybaert, N., & Vanhoof, K. (2007). Data mining for fraud detection: Toward an improvement on internal control systems? *Proceedings of the 30th Annual Congress European Accounting Association (EAA2007).*

Jans, M., Lybaert, N., & Vanhoof, K. (2010). Internal fraud risk reduction: Results of a data mining case study. *International Journal of Accounting Information Systems*, *11*(1), 17–41. https://doi.org/10.1016/j.accinf.2009.12.004

Jung, R., Winter, R., Knobloch, B., & Weidner, J. (2000). Eine kritische Betrachtung von Data Mining-Prozessen — Ablauf, Effizienz und Unterstützungspotenziale. *Data Warehousing 2000.* Advance online publication. https://doi.org/10.1007/978-3-642-57681-2_19

Kim, Y., & Kogan, A. (2014). Development of an Anomaly Detection Model for a Bank's Transitory Account System. *Journal of Information Systems*, *28*(1), 145–165. https://doi.org/10.2308/isys-50699

Kitchenham, B. A., & Charters, S. (2007). *Guidelines for performing Systematic Literature Reviews in Software Engineering*. http://cdn.elsevier.com/promis_misc/525444systematicreviewsguide.pdf Accessed 20 October 2020.

Kogan, A., Alles, M. G., Vasarhelyi, M. A., & Wu, J. (2014). Design and Evaluation of a Continuous Data Level Auditing System. *AUDITING: A Journal of Practice & Theory*, *33*(4), 221–245. https://doi.org/10.2308/ajpt-50844

Kokina, J., & Davenport, T. H. (2017). The Emergence of Artificial Intelligence: How Automation is Changing Auditing. *Journal of Emerging Technologies in Accounting*, *14*(1), 115–122. https://doi.org/10.2308/jeta-51730

Krüger, H. A., & Hattingh, J. M. (2006). A combined AHP-GP model to allocate internal auditing time to projects. *ORiON*, *22*(1), 59–76. https://doi.org/10.5784/22-1-33

Kuna, H. D., García-Martinez, R., & Villatoro, F. R. (2014). Outlier detection in audit logs for application systems. *Information Systems*, *44*, 22–33. https://doi.org/10.1016/j.is.2014.03.001

Lipman, F. D., & Lipman, L. K. (2012). *The Internal Audit Function. Wiley Online Books.* https://doi.org/10.1002/9781119197195.ch5

Liu, Q. (2014). *The application of exploratory data analysis in auditing* [Ph.D.]. Rutgers, The State University of New Jersey. https://doi.org/10.7282/T3CC129J

Lu, F. (2007). Uncovering Fraud in Direct Marketing Data with a Fraud Auditing Case Builder. *Knowledge Discovery in Databases: PKDD 2007*, 540–547. https://doi.org/10.1007/978-3-540-74976-9_56

Lu, F., Boritz, J. E., & Covvey, D. (2006). Adaptive fraud detection using Benford's law. *Canadian AI 2006*, 347–358. https://doi.org/10.1007/11766247_30

Murphree, J. (2016). Machine learning anomaly detection in large systems. In *IEEE AUTOTESTCON 2016: Anaheim, California, USA, September 12-15, 2016 : proceedings* (pp. 1–9). IEEE. https://doi.org/10.1109/AUTEST.2016.7589589

Nguyen, L., & Kohda, Y. (2017). Toward a model of wisdom determinants in the auditing profession. *Proceedings of the 50th Hawaii International Conference on System Sciences*, 4604–4613. https://doi.org/10.24251/HICSS.2017.560

No, W. G., Lee, K., Huang, F., & Li, Q. (2019). Multidimensional Audit Data Selection (MADS): A Framework for Using Data Analytics in Audit Data Selection Process. *Accounting Horizons*, *33*(3), 127–140. https://doi.org/10.2308/acch-52453

Okoli, C. (2015). A Guide to Conducting a Standalone Systematic Literature Review. *Communications of the Association for Information Systems*, *37*, 879–910. https://doi.org/10.17705/1CAIS.03743

Paula, E. L., Ladeira, M., Carvalho, R. N., & Marzagão, T. (2016). Deep Learning Anomaly Detection as Support Fraud Investigation in Brazilian Exports and Anti-Money Laundering. *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 954–960. https://doi.org/10.1109/ICMLA.2016.0172

Reinsel, D., Gantz, J., & Rydning, J. (2018). *Data age 2025: the digitization of the world from edge to core*. https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf* Accessed 20 October 2020.

Richhariya, P., & Singh, P. K. (2012). A survey on financial fraud detection methodologies. *International Journal of Computer Applications*, *45*(22), 15–22. https://doi.org/10.5120/7080-9373

Rikhardsson, P., Singh, K., & Best, P. (2019). Exploring continuous auditing solutions and internal auditing: A research note. *Accounting & Management Information Systems/Contabilitate Si Informatica De Gestiune*, *18*(4), 614–639. https://doi.org/10.24818/jamis.2019.04006

Sadgali, I., Sael, N., & Benabbou, F. (2019). Performance of machine learning techniques in the detection of financial frauds. *THE SECOND INTERNATIONAL CONFERENCE on INTELLIGENT COMPUTING in DATA SCIENCES, ICDS2018*, *148*, 45–54. https://doi.org/10.1016/j.procs.2019.01.007

Schreyer, M., Sattarov, T., Borth, D., Dengel, A., & Reimer, B. (2017). *Detection of Anomalies in Large Scale Accounting Data using Deep Autoencoder Networks*. https://arxiv.org/abs/1709.05254

Schreyer, M., Sattarov, T., Schulze, C., Reimer, B., & Borth, D. (2019). *Detection of Accounting Anomalies in the Latent Space using Adversarial Autoencoder Neural Networks*. https://arxiv.org/abs/1908.00734

Sharma, A., & Panigrahi, P. K. (2013). *A Review of Financial Accounting Fraud Detection based on Data Mining Techniques*. https://arxiv.org/abs/1309.3944

Sutton, S. G., Holt, M., & Arnold, V. (2016). “The reports of my death are greatly exaggerated”—Artificial intelligence research in accounting. *2015 Research Symposium on Information Integrity & Information Systems Assurance*, *22*, 60–73. https://doi.org/10.1016/j.accinf.2016.07.005

Thiprungsri, S., & Vasarhelyi, M. A. (2011). Cluster Analysis for Anomaly Detection in Accounting Data: An Audit Approach. *International Journal of Digital Accounting Research*, *11*. https://doi.org/10.4192/1577-8517-v11_4