Continual Learning for Unsupervised Anomaly Detection in Continuous Auditing of Financial Accounting Data

Hamed Hemati[*1], **Marco Schreyer**[*1], **Damian Borth** [1]

[1]University of St. Gallen (HSG), St. Gallen, Switzerland

Abstract

International audit standards require the direct assessment of a financial statement's underlying accounting journal entries. Driven by advances in artificial intelligence, deep-learning inspired audit techniques emerged to examine vast quantities of journal entry data. However, in regular audits, most of the proposed methods are applied to learn from a comparably stationary journal entry population, e.g., of a financial quarter or year. Ignoring situations where audit relevant distribution changes are not evident in the training data or become incrementally available over time. In contrast, in continuous auditing, deep-learning models are continually trained on a stream of recorded journal entries, e.g., of the last hour. Resulting in situations where previous knowledge interferes with new information and will be entirely overwritten. This work proposes a continual anomaly detection framework to overcome both challenges and designed to learn from a stream of journal entry data experiences. The framework is evaluated based on deliberately designed audit scenarios and two real-world datasets. Our experimental results provide initial evidence that such a learning scheme offers the ability to reduce false-positive alerts and false-negative decisions.

Introduction

According to the *International Standards in Auditing (ISA)*, auditors are obliged to collect reasonable assurance that an issued financial statement is free from fraud that causes a material misstatement (AICPA 2002; IFAC 2009). At the same time, the *Association of Certified Fraud Examiners (ACFE)* estimates that organizations lose 5% of their annual revenues due to fraud (ACFE 2020). The term *fraud* refers to *'the abuse of one's occupation for personal enrichment through the deliberate misuse of an organization's resources or assets'* (Wells 2017). A similar study revealed that approx. 30% of the respondents experienced losses between $100,000 and $5 million due to fraud (PwC 2020). The study also showed that financial statement fraud caused the highest median loss of the surveyed fraud schemes.[1] To detect

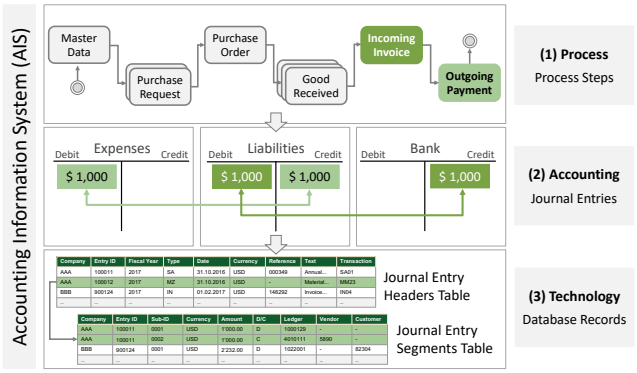


Figure 1: Exemplary view of an Accounting Information System (AIS) that records distinct layer of abstractions, namely (1) the business process, (2) the accounting and (3) technical journal entry information in designated tables.

such fraud, the ISA demands auditors to examine a statement's underlying accounting transactions, usually referred to as *journal entries* (CAQ 2008; IFAC 2009). Nowadays, organizations record vast quantities of journal entries in *Accounting Information Systems (AIS)* or more general *Enterprise Resource Planning (ERP)* systems (Grabski, Leech, and Schmidt 2011). Figure 1 depicts an exemplary hierarchical view of the journal entry recording process in designated database tables of an AIS system. In general, such recorded entries originate from non-stationary activities due to the dynamics of organizational processes, e.g., the (i) introduction, (ii) redesign, or (iii) discontinuation of business processes. Therefore, the detection of fraud in non-stationary journal entry distributions remains a challenging task requiring significant efforts and resources.

In order to conduct fraud, perpetrators need to deviate from regular journal entry posting patterns. Such a divergence is then recorded in a small fraction of journal entries that exhibit *'anomalous'* attribute values. Nowadays, auditors apply a wide range of *Computer Assisted Audit Techniques (CAATs)* when examining journal entries. These techniques often encompass rule-based analyses, such as the detection of unusual posting times (Fay and Negangard 2017) or multiple vendor bank account changes (Coderre 2009). Furthermore, statistical analyses are applied, such as the *Benford's Law* (Benford 1938) or extreme value analysis

[*]These authors contributed equally.

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1]The ACFE study encompasses an analysis of 2,504 cases of occupational fraud surveyed between July 2019 and September 2019 in 125 countries. The PwC study encompasses over 5,000+ respondents that experienced economic crime in the last 24 months.

(Bolton and Hand 2002). Driven by the rapid technological advances of artificial intelligence, deep-learning (LeCun, Bengio, and Hinton 2015) enabled audit techniques have recently emerged in financial audits (Sun 2019; Nonnenmacher and Gómez 2021).

At the same time, the application of deep-learning in modern audit engagements results, among others, in the challenge of (i) *stationary training setups* and (ii) *catastrophic forgetting*. The first challenge originates from the situation that in *regular auditing* deep-learning models are learned from a limited and therefore *stationary population* of journal entries. Typically the training only encompasses entries posted throughout the in-scope audit period, e.g., the current financial quarter or year. Such a learning setup exhibits the drawback that relevant distribution changes might not be evident in the training data. The second challenge originates from the fact that deep-learning based audit models are prone to *Catastrophic Forgetting* (CF) (McCloskey and Cohen 1989), i.e., training a model with new information interferes with previously learned knowledge. Typically, in *continuous auditing* (Vasarhelyi and Halper 1991; Vasarhelyi, Alles, and Kogan 2004) the model training encompasses a constant stream of entries recorded within incremental time intervals, e.g., the recent hour or day. This phenomenon leads, in the worst-case, to a situation in which previous knowledge will be entirely overwritten by new one.

In order to mitigate false-positive alerts or false-negative decisions in real-world audits, it is paramount to learn important audit relevant information that becomes incrementally available over time. For example, an organizational change in previous financial years might be relevant when concluding on a recently detected audit finding. This holds in particular for applying unsupervised anomaly detection techniques to audit large-scale accounting data. Recently, the idea of *Continual Learning* (CL), also referred to as *lifelong learning*, regained attention (Chen and Liu 2018). CL denotes a learning paradigm in which the assumption is made that data will become available over time. We think accommodating new knowledge over time while retaining previously learned information is essential for future deep-learning enabled audit techniques. In this work, inspired by the successes of CL (Sun 2019; Zhai, Liu, and Sun 2019; Lee and Lee 2020), we investigate whether such a CL setup can be utilized in financial audits to detect accounting anomalies? In summary, we present the following contributions:

- *Learnability* - We propose and evaluate a novel learning framework to continuously learn from real-world journal entry for unsupervised anomaly detection.
- *Transferability* - We demonstrate that the continual learning techniques outperform the re-training on an exhaustive dataset and the sequential fine-tuning.

The remainder of this work is structured as follows: First, we provide an overview of related work. Following, we present the proposed continual learning framework for unsupervised anomaly detection in financial audits. Afterwards, the experimental setup is described and results are outlined. The work concludes with a discussion and summary.

Related Work

Lately, techniques that build on (deep) machine learning have been gradually applied to different audit tasks (Sun 2019; Cho et al. 2020). In parallel, the idea of continual learning triggered a sizable body of machine learning research (Parisi et al. 2019; Delange et al. 2021). In the following, we focus our literature review on (i) CAATs applied in financial audits and (ii) continual deep-learning techniques.

Computer Assisted Audit Techniques: With the evolution of AIS systems, the development of CAATs has been of central relevance in auditing. Various techniques have been proposed to audit structured accounting data, such as (un-) supervised machine-learning approaches (Bay et al. 2006; McGlohon et al. 2009; Jans, Lybaert, and Vanhoof 2010; Thiprungsri and Vasarhelyi 2011; Argyrou 2012, 2013; Hu et al. 2020), statistical analyzes (Debreceny and Gray 2010; SEOW, Pan, and Suwardy 2016), rule-based procedures (Khan et al. 2009, 2010, 2014; Islam et al. 2010), business process mining (Jans et al. 2011; Jans, Alles, and Vasarhelyi 2014; Werner and Gehrke 2015; Werner 2016), and the combination thereof (Baader and Krcmar 2018; Stephan, Lahann, and Fettke 2021). Recently, deep-learning techniques have emerged in auditing. Such techniques encompass, autoencoder neural networks (Schreyer et al. 2017; Schultz and Tropmann-Frick 2020; Nonnenmacher et al. 2021; Zupan et al. 2020; Schreyer et al. 2020) to detect accounting anomalies and learn representative audit samples. Lately, self-supervised learning has been proposed to learn multi-task audit models (Schreyer, Sattarov, and Borth 2021).

Continual Deep-Learning Techniques: The challenge of CF in training neural networks, has been studied extensively (French 1999; Díaz-Rodríguez et al. 2018; Toneva et al. 2018; Lesort, Caccia, and Rish 2021). To mitigate CF, several approaches have been proposed, broadly classified as (i) *rehearsal*, (ii) *regularization*, and (iii) *dynamic architecture* techniques. *Rehearsal* based techniques (Rolnick et al. 2018; Isele and Cosgun 2018; Chaudhry et al. 2019) replay samples from past experiences. Thereby, samples are saved in a buffer and replayed to conduct knowledge distillation (Rebuffi et al. 2017). Other methods use gradients of previous tasks to mitigate gradient interference (Lopez-Paz and Ranzato 2017; Chaudhry et al. 2018). Furthermore, generative models are used to generate synthesized samples for rehearsal (Shin et al. 2017). *Regularization* based techniques consolidate previously acquired knowledge when optimizing the model parameters. *Learning without Forgetting (LwF)* (Li and Hoiem 2017), use knowledge distillation to regularize parameter updates. *Elastic Weight Consolidation (EWC)* (Kirkpatrick et al. 2017) directly regularizes the parameters based on their importance concerning previous data. Similarly to EWC, in *Synaptic Intelligence* (Zenke, Poole, and Ganguli 2017) the model parameter importance is regularized but in a separate parameter optimization step. *Dynamic architecture* methods prevent forgetting by the deliberate reuse of (Mendez and Eaton 2021) or increase in (Rusu et al. 2016) model parameters.

Concluding from the reviewed literature, this work presents, to the best of our knowledge, the first analysis on

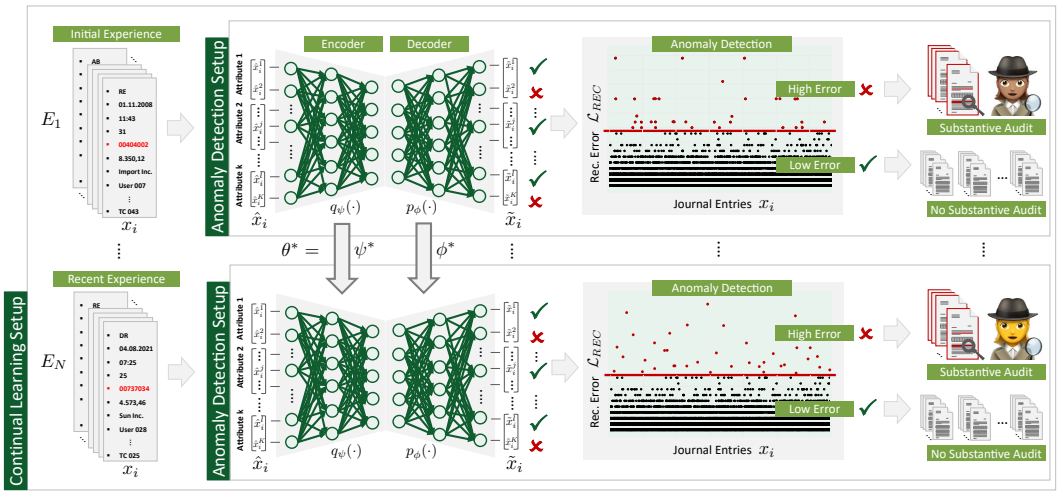


Figure 2: The proposed continual anomaly detection framework that continuously learns from a stream of accounting data experiences $\{E_i\}_{i=1}^{N}$ (e.g., financial quarters or years) by accommodating new information while preserving previous knowledge.

the application of continual deep-learning methods for the purpose of real-world financial statement audits.

Methodology

Let $X = \{x_1, x_2, x_3, ..., x_N\}$ formally define a population of $i = 1, 2, 3, ..., N$ journal entries. Each individual entry, denoted by $x_i = \{x_i^1, x_i^2, ..., x_i^M; x_i^1, x_i^2, ..., x_i^K\}$, consists of $j = 1, 2, 3, ..., M$ categorical accounting attributes and $l = 1, 2, 3, ..., K$ numerical accounting attributes. The individual attributes describe the journal entries details, e.g., the entries' posting type, posting date, amount, general-ledger. When examining journal entries, recorded in large-scaled ERP systems, two characteristics can be observed: First, the journal entry attributes exhibit a high variety of distinct attribute values, e.g., due to the high number of vendors or distinct posting amounts. Second, journal entries exhibit a plethora of attribute value correlations, e.g., a document type usually posted in combination with a certain general ledger account. Derived from this observation we distinguish two classes of anomalous entries (Breunig et al. 2000), namely *global* and *local* accounting anomalies:

- **Global Anomalies** correspond to journal entries that exhibit unusual or rare individual attribute values, e.g., rarely used ledgers, or unusual posting times. Such anomalies are often simple to detect since they are created unintentionally in the context of regular posting activities. Hence, global anomalies are usually associated with a high *error* risk.
- **Local Anomalies** correspond to journal entries that exhibit unusual or rare attribute value correlations, e.g., unusual co-occurrences of general ledgers, posting types and user accounts. Such anomalies are often difficult to detect since perpetrators aim to imitate regular posting activities. Hence, local anomalies are usually associated with a high *fraud* risk.

Ultimately, auditors aim to learn a model that detects both classes of anomalies using a continual anomaly detection framework. In the following, we describe the main components of our proposed framework. A schematic view of the framework is illustrated in Fig. 2, encompassing an *anomaly detection* and a *continual* learning setup.

Anomaly Detection Setup: Hawkins et al. (2002) proposed the application of *Autoencoder Networks (AEN's)*, a special type of feed-forward multi-layer neural network for unsupervised anomaly detection. Formally, AEN's are comprised of two nonlinear functions referred to as *encoder* and *decoder* network (Hinton and Salakhutdinov 2006) that are jointly trained to reconstruct a given input. The encoder network $q_\psi(\cdot)$ maps an input $x_i \in \mathcal{R}^k$ to a code vector $z_i \in \mathcal{R}^m$ referred to as *latent* representation, where usually $k > m$. Subsequently, this representation is mapped back by the decoder network $p_\phi(\cdot)$ to a reconstruction $\tilde{x}_i \in \mathcal{R}^k$ of the original input space. In an attempt to achieve $x \approx \tilde{x}$ the AEN is trained to minimize the dissimilarity of a given journal entry x_i and its reconstruction $\tilde{x}_i = p_\phi(q_\psi(x_i))$ as faithfully as possible. Thereby, the training objective is to learn a set of optimal model parameters $\theta^* = \{\psi^* \cup \phi^*\}$ by minimizing the AEN's reconstruction loss ($\mathcal{L}_{Rec}$), formally denoted as:

$$\arg\min_{\phi,\psi} \|x_i - p_\phi(q_\psi(x_i))\|, \tag{1}$$

where $i = 1, 2, 3, ..., N$. Upon successful AEN training, the model is used to obtain the individual reconstruction error of each journal entry. In a real-world audit scenario, entries that correspond to a high reconstruction error will then be selected for subsequent substantial examination.

Continual Learning Setup: In *Continual Learning (CL)* a model f_θ, e.g., an autoencoder $f_\theta : q_\psi \oplus p_\phi$, is trained in a setup where data $D = \bigcup_{i=1}^{M} D_i$ is observed as a stream of M disjoint experiences $\{E_i\}_{i=1}^{M}$. In the context of auditing, the journal entry data D_i of a particular experience E_i originates from N organizational posting activities A_j as defined by $D_i = \bigcup_{j=1}^{N} A_j$. A model trained on the data of the i-th experience is denoted as $f_\theta^{(i)}$, where θ defines the model parameters. The models reconstruction loss can then be derived by:

$$\mathcal{L}_{Rec}^{(i)} = \sum_{j=1}^{N} \mathcal{L}_{Rec}(f_\theta^{(i)}, A_j), \quad (2)$$

where A_j denotes the organizational posting activities of experience E_i.

In this work three CL setups are utilized to audit financial accounting data, namely (i) *Sequential Fine-tuning*, (ii) *EWC*, and (iii) *Experience Replay*. In each setup an anomaly detection model is incrementally trained on a stream of $M-1$ past experiences $\{E_i\}_{i=1}^{M-1}$. For each experience E_i, the model parameters θ are initialized with the optimal model parameter learned at the previous experience E_{i-1}. In the *Sequential Fine-Tuning (SFT)* setup, the parameters θ are optimized, as defined by:

$$\theta^{*(i)} = \arg\min_\theta \mathcal{L}(f(\theta), D_i | \theta_{init} = \theta^{*(i-1)}), \quad (3)$$

where $\theta^{*(i)}$ denotes the optimal parameters obtained for the previous experience E_i. In the regularization based *Elastic Weight Consolidation (EWC)* setup, the parameters θ are optimized, as defined by:

$$\theta^{*(i)} = \arg\min_\theta \widehat{\mathcal{L}}(f(\theta), D_i | \theta_{init} = \theta^{*(i-1)}), \quad (4)$$

where $\theta^{*(i)}$ denotes the optimal parameters obtained for the previous experience E_i. To update the model's parameters in a given experience, EWC (Kirkpatrick et al. 2017) is applied by determining the Fisher information matrix F according to $\hat{\mathcal{L}}(\theta) = \mathcal{L}_{Rec}(\theta) + \sum_k \frac{\lambda}{2} F_k(\theta_i - \tilde{\theta}_k)^2$, where k denotes the model parameters of the current model and $\tilde{\theta}$ the optimal parameters of the previous experience. In the rehearsal based *Experience Replay (ER)* setup, the parameters θ are optimized, as defined by:

$$\theta^{*(i)} = \arg\min_\theta \mathcal{L}(f(\theta), \widehat{D}_i | \theta_{init} = \theta^{*(i-1)}), \quad (5)$$

where $\theta^{*(i)}$ denotes the optimal parameters obtained for the previous experience E_i. To update the model's parameters in a given experience, samples of past experiences are replayed. Therefore, a sample buffer $\mathcal{B}$ of capacity N_B is employed. The buffer comprises an equal number of $\lfloor \frac{N_B}{M} \rfloor$ samples for each observed experience.

We evaluate the continual learning setups against two baselines, namely (i) *single experience* and (ii) *joint experience* learning. In *Single Experience Learning (SEL)* only the journal entries of the current in-scope audit experience E_M are used to train f_θ. Such a learning setup is commonly applied in contemporary audits. The model parameters θ are randomly initialized and optimized, as defined by:

$$\theta^{*(i)} = \arg\min_\theta \mathcal{L}(f(\theta), D_i | \theta_{init} = Rand), \quad (6)$$

where $\theta^{*(i)}$ denotes the optimal parameters obtained after training on E_i. We view this setup as performance upper-bound in our experiments. In *Joint Experience Learning (JEL)* the journal entries of all previous experiences $\{E_i\}_{i=1}^{M-1}$ and the current in-scope experience E_M are used to train f_θ from scratch. However, such a learning setup is not desirable in contemporary audit engagements due to time-budget constraints. The model parameters θ are randomly initialized and optimized, as defined by:

$$\theta^{*(i)} = \arg\min_\theta \mathcal{L}(f(\theta), \{D_1, ..., D_i\} | \theta_{init} = Rand), \quad (7)$$

where $\theta^{*(i)}$ denotes the optimal parameters obtained after training on E_i. Nowadays, this setup is commonly rejected due to time-budget constraints of audit engagements. We view this setup as performance lower-bound in our experiments.

Experimental Setup

In this section, we describe the dataset, data pre-processing, audit scenarios and experimental details to continually learn from real-world accounting data.

Datasets and Data Preparation: To evaluate the learning capabilities of the proposed continual learning framework we use two publicly available datasets of real-world financial payment data.[2] The datasets exhibit high similarity to real-world ERP accounting data, e.g., typical manual payments or payment run records.

- The *City of Philadelphia* payments[3] denoted as D^A encompass a total of 238,894 payments generated by 58 distinct city departments and comprises 10 categorical and 1 numerical attribute(s).
- The *City of Chicago* payments[4] denoted as D^B encompass a total of 399,158 payments generated by 54 distinct city departments and comprises of 6 categorical and 1 numerical attribute(s).

For both datasets, let $\{A_j\}_{j=1}^N$ denote the set of payments generating city departments. For each dataset we conduct the following pre-processing steps: First, the τ departments exhibiting the highest number of payments $\{A_k^*\}_{k=1}^\tau$ are determined. Second, for each city department A_k^* a set of η payments is randomly sampled $x_i \sim \{A^*\}_{k=1}^\tau$, where $i = 1, .., \eta$. Third, the sampled payment's attribute values are transformed according to:

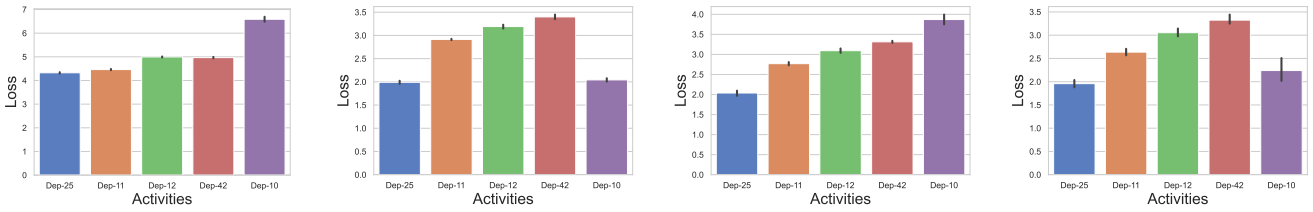
- The categorical attribute values of x_i^j are converted into *one-hot* numerical tuples of bits $\hat{x}_i^j \in \{0,1\}^\upsilon$, where υ denotes the number of unique attribute values in x^j.
- The numerical attribute values of x_i^l are scaled, according to $\hat{x}_i^l = (x_i^l - \min(x^l)/(\max(x^l) - \min(x^l))$, where the min and max are obtained over all attribute values in x^l.

We set the number of departments $\tau = 10$ and number of samples $\eta = 10,000$ in all our experiments to derive the pre-processed journal entries $\hat{x} \in \hat{X}$, where

[2]Due to the general confidentiality of journal entry data, we evaluate the proposed methodology based on two public available real-world datasets to allow for the reproducibility of our results.

[3]https://www.phila.gov/2019-03-29-philadelphias-initial-release-of-city-payments-data/

[4]https://data.cityofchicago.org/Administration-Finance/Payments/s4vu-giwb/



(a) Single Experience Learning (SEL) (b) Joint Experience Learning (JEL) (c) Sequential Fine-Tuning (SFT) (d) Experience Replay (ER)

Figure 3: Average reconstruction losses $\mathcal{L}_{Rec}$ in the in-scope audit period E_M of dataset D^A for the decaying target department (purple bar) and the four departments exhibiting the highest loss (non-purple bars). It can be observed that in CL the linearly decaying target department A' results in a significantly reduced number of loss magnitude based false-positive alerts.

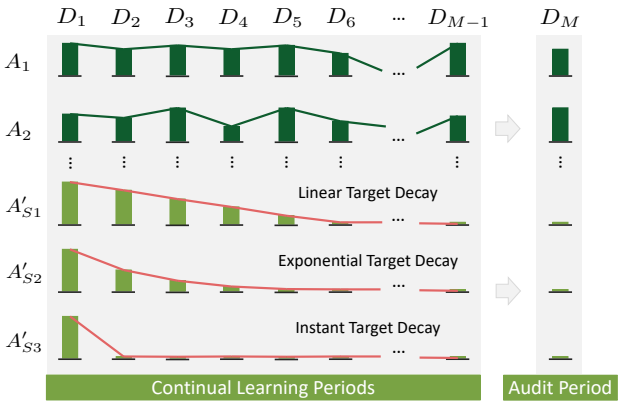


Figure 4: Journal entry distribution for distinct activities $A_1, A_2, ..., A_\tau$ over progressing experiences (top) and target department decaying scenarios $A'_{S1}, A'_{S2}, A'_{S3}$ (bottom).

$\hat{x}_i = \{\hat{x}_i^1, \hat{x}_i^2, ..., \hat{x}_i^M; \hat{x}_i^1, \hat{x}_i^2, ..., \hat{x}_i^K\}$. Ultimately, each payment $\hat{x}_i \in D^A$ resulted in 5,588 encoded dimensions, while each payment $\hat{x}_i \in D^B$ in 7,486 encoded dimensions.

Continual Audit Scenarios: We create a stream of N experiences denoted by $\{E\}_{i=1}^M$ where each experience exhibits its own dataset D_i. For each D_i a number of $\rho \cdot \frac{\eta}{N}$ pre-processed payments is randomly sampled from each $\hat{x}_i \sim A_k^*$, where $\rho \in [0.9; 1.0]$ and $k = 1, .., \tau$. In a real-world audit setup the experiences may correspond to distinct time periods, e.g., a financial quarter or year. Furthermore, we assume that the most recent experience E_M defines the in-scope period of the audit with journal entry data D_N. In this initial work, we evaluate two scenarios observable when examining journal entry data in financial audits, namely (i) *discontinued operations* and (ii) *anomalous postings*. Figure 4 illustrates the distribution of sampled journal entries for distinct activities over progressing experiences.

The first *discontinued operations* scenario, refers to a situation in which an organizational activity or department is discontinued, e.g., due to a shut down a divestiture. Thereby, the number of corresponding journal entries vanishes over time. Using stationary deep-learning enabled audit techniques such a scenario regularly results in a high number of **false-positive alerts** within the in-scope audit period. To establish such a scenario a to be discontinued department $A' \in \{A_k^*\}_{k=1}^\tau$ is selected. Subsequently, a decaying number of journal entries is randomly sampled from A' with progressing experiences. We investigated three discontinuation scenarios, resulting in a (i) *linear* (A'_{S1}), (ii) *exponential* (A'_{S2}), and (iii) *instant* (A'_{S3}) reduction of journal entries.

The second *anomalous postings* scenario, refers to a situation in which *'anomalous'* journal entries are posted within the in-scope period of an audit, e.g., due to error or fraudulent activities. Using stationary deep-learning enabled audit techniques such a scenario regularly results in an increased number of undetected **false-negative observations** within the in-scope audit period. To establish such a scenario, a small number of α_1 local (A_{P1}) and α_2 global (A_{P2}) accounting anomalies is injected into the dataset D_N of the last experience.[5] For both classes of of injected anomalies we set $\alpha_1 = 10$ and $\alpha_2 = 10$ in all our experiments.

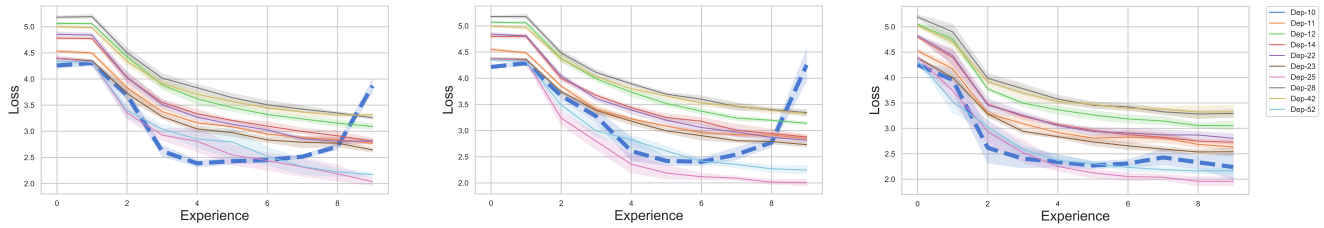
Autoencoder Learning Setup: The encoder q_ψ and a decoder p_θ network use Leaky-ReLU activation functions (Xu et al. 2015) with scaling factor $\alpha = 0.4$ except in the bottleneck (last) layer where tanh linearities (non-linearities) are applied. Table 1 depicts the architectural details of the models used in our experiments. Both AEN parameter sets ψ and ϕ are initialized as described by Glorot and Bengio (2010). The models are trained for 500 epochs, using a batch size of $b = 128$ journal entries, and early stopping once the loss converges. We use Adam optimization (Kingma and Ba 2014) with $\beta_1 = 0.9$, $\beta_2 = 0.999$. Furthermore, we compute the binary cross-entropy error of a given encoded journal entry $\hat{x}_i$, as defined by $\mathcal{L}_{Rec}(\tilde{x}_i, \hat{x}_i) = \frac{1}{N}\sum_{i=1}^N \tilde{x}_i \cdot \log(\hat{x}_i) + (1 - \tilde{x}_i) \cdot \log(1 - \hat{x}_i)$, where $\tilde{x}_i = p_\phi(q_\psi(\hat{x}_i))$ denotes the i-*th* reconstructed entry.

Table 1: Number of neurons per layer η of the encoder q_ψ and decoder p_ϕ network that constitute the distinct AEN architectures used in our experiments.

Net	$\eta = 1$	2	3	4	5	6	7
$q_\psi(z\|\hat{x})$	128	64	32	16	8	4	2
$p_\phi(\hat{x}'\|z)$	2	4	8	16	32	64	128

Continual Learning Setup: We set the number of experiences $N = 10$ in all our experiments. To find optimal parameters in EWC and ER we conducted a grid search of the hyper-parameter space. For EWC we found that $\lambda = 50$ preserves an optimal degree of plasticity. In ER learning we set $N_B = 500$ which seemed sufficient for both datasets given $N = 10$ experiences. We reset the optimizer parame-

[5]To create both classes of anomalies, we built upon the *Faker* project, which is publicly available via the following URL: https://github.com/joke2k/faker.



(a) Sequential Fine-Tuning (SFT) (b) Elastic Weight Consolidation (EWC) (c) Experience Replay (ER)

Figure 5: Reconstruction losses $\mathcal{L}_{Rec}$ of the $\tau = 10$ departments in dataset D_A over a stream of $\{E\}_{i=1}^{M}$ experiences for distinct CL setups. The dashed blue line corresponds to reconstruction loss of the linearly decaying target department A'. It can be observed that ER results in a significantly reduced number of false-positives over the progressing experiences.

ters upon each experience to avoid past experience information transfer through the optimizer state.

Experimental Results and Discussion

In this section, we present the experimental continual learning results for both of the previously introduced continual audit scenarios, namely (i) *discontinued operations* and (ii) *anomalous postings*.

The experimental results of the *discontinued operations* scenario are presented in Tab. 2. For both datasets D^A and D^B it can be observed that the average reconstruction loss difference of the target department A' and the department exhibiting the highest loss $\tilde{A}$, as defined by $\Delta_{FP} := \mathcal{L}_{Rec}(A'_S) - \mathcal{L}_{Rec}(\tilde{A})$. A negative (positive) Δ_{FP} indicates a reduction (an increase) of false-positive alerts originating from the decaying target department within the in-scope audit period E_M. Thereby, JEL and SEL provide a lower- and upper-bound on the false-positive alerts. The obtained results demonstrate that CL mitigates the risk of false-positive alerts while saving data and computational resources. When comparing SFT, EWC, and ER for the decaying setups, as also illustrated in Fig. 5, the best performance is achieved by ER. Thereby, the exponential A'_{S2} and instant A'_{S3} decay results in a more challenging CL setup.

Table 2: Average reconstruction loss $\mathcal{L}_{Rec}$ difference Δ_{FP} at E_M between the decaying target department and the department with the highest loss in both datasets D^A and D^B.

Discontinued Operations - Philadelphia Payments D^A

Method	$\Delta_{FP}(A'_{S1})$	$\Delta_{FP}(A'_{S2})$	$\Delta_{FP}(A'_{S3})$
SEL	1.45 ± 0.115	1.40 ± 0.064	1.45 ± 0.062
JEL	-1.36 ± 0.047	-1.36 ± 0.047	-1.36 ± 0.047
SFT	0.56 ± 0.112	1.39 ± 0.401	1.82 ± 0.284
EWC	0.91 ± 0.400	1.31 ± 0.313	2.03 ± 0.218
ER	$\mathbf{-1.08 \pm 0.371}$	$\mathbf{0.02 \pm 0.539}$	$\mathbf{0.02 \pm 0.089}$

Stdv. $\pm$ originates from parameter initialization using five random seeds.

Discontinued Operations - Chicago Payments D^B

Method	$\Delta_{FP}(A'_{S1})$	$\Delta_{FP}(A'_{S2})$	$\Delta_{FP}(A'_{S3})$
SEL	0.66 ± 0.101	0.66 ± 0.053	0.61 ± 0.099
JEL	-0.96 ± 0.153	-0.96 ± 0.153	-0.96 ± 0.153
SFT	0.78 ± 0.268	1.71 ± 1.207	1.35 ± 0.213
EWC	1.18 ± 0.168	1.12 ± 0.184	1.10 ± 0.230
ER	$\mathbf{-0.05 \pm 1.514}$	$\mathbf{-0.50 \pm 0.192}$	$\mathbf{-0.19 \pm 0.243}$

Stdv. $\pm$ originates from parameter initialization using five random seeds.

The experimental results of the *anomalous postings* scenario and a linear decaying target department are presented in Tab. 3. For both datasets D^A and D^B it can be observed that SEL, the average reconstruction error $\mathcal{L}_{Rec}$ of the decaying target department A' and the local anomalies A_{P1} exhibit a similar error magnitude within the in-scope audit period E_M. Such a scenario increases the risk of not detecting local anomalies and might result in false-negative decisions. The obtained results demonstrate that CL mitigates this risk by forward transfer of knowledge from previous experiences. This observation is also reflected by the loss difference $\Delta_{FN} := \mathcal{L}_{Rec}(A') - \mathcal{L}_{Rec}(A_{P1})$ of the target department A' and the local anomalies A_{P1}. When comparing SFT, EWC, and ER, ER achieves the best performance.

Table 3: Average reconstruction loss $\mathcal{L}_{Rec}$ at E_M of the decaying target department A' and the injected local A_{P1} and global A_{P2} anomalies in both datasets D^A and D^B.

Anomalous Postings - Philadelphia Payments D^A

Method	$\mathcal{L}_{Rec}(A') \downarrow$	$\mathcal{L}_{Rec}(A_{P1}) \uparrow$	$\mathcal{L}_{Rec}(A_{P2}) \uparrow$
SEL	6.59 ± 0.125	7.03 ± 0.130	12.21 ± 0.208
JEL	2.04 ± 0.042	36.71 ± 5.948	18.12 ± 7.399
SFT	3.87 ± 0.141	34.10 ± 6.025	12.99 ± 2.286
EWC	4.25 ± 0.339	29.97 ± 4.761	13.22 ± 1.145
ER	$\mathbf{2.24 \pm 0.292}$	$\mathbf{47.69 \pm 6.482}$	$\mathbf{16.64 \pm 5.176}$

Stdv. $\pm$ originates from parameter initialization using five random seeds.

Anomalous Postings - Chicago Payments D^B

Method	$\mathcal{L}_{Rec}(A') \downarrow$	$\mathcal{L}_{Rec}(A_{P1}) \uparrow$	$\mathcal{L}_{Rec}(A_{P2}) \uparrow$
SEL	4.27 ± 0.097	3.55 ± 0.171	4.53 ± 0.053
JEL	1.30 ± 0.145	16.75 ± 4.036	4.71 ± 0.454
SFT	3.01 ± 0.288	13.03 ± 3.845	3.93 ± 0.169
EWC	3.42 ± 0.185	12.20 ± 1.958	5.18 ± 1.709
ER	$\mathbf{2.26 \pm 1.154}$	$\mathbf{22.23 \pm 5.439}$	$\mathbf{6.25 \pm 1.857}$

Stdv. $\pm$ originates from parameter initialization using five random seeds.

Conclusion

In this work, we proposed the first CL inspired framework to detect accounting anomalies in the context of financial statement audits. Based on two deliberately designed audit scenarios, we demonstrated the ability of our method to continuously learn data distribution shifts observable in datasets of real-world city payments. Our experimental results provide initial evidence that CL enabled anomaly detection improve stationary baseline detection techniques in terms of false-positive alerts and false-negative decisions.

References

ACFE. 2020. *Report to the Nations - Global Study on Occupational Fraud and Abuse*. Association of Certified Fraud Examiners (ACFE).

AICPA. 2002. *Statement on Auditing Standards No. 99.: Consideration of Fraud in a Financial Statement Audit*.

Argyrou, A. 2012. Auditing Journal Entries Using Self-Organizing Map. In *Proceedings of the Eighteenth Americas Conference on Information Systems*.

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

Baader, G.; and Krcmar, H. 2018. Reducing False Positives in Fraud Detection: Combining the Red Flag Approach with Process Mining. *International Journal of Accounting Information Systems*, 31: 1–16.

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

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

Bolton, R. J.; and Hand, D. J. 2002. Statistical fraud detection: A review. *Statistical science*, 235–249.

Breunig, M. M.; Kriegel, H.-P.; Ng, R. T.; and Sander, J. 2000. LOF: Identifying Density-based Local Outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 93–104.

CAQ. 2008. *Practice Aid for Testing Journal Entries and Other Adjustments Pursuant to AU Section 316*. Center for Audit Quality (CAQ).

Chaudhry, A.; Ranzato, M.; Rohrbach, M.; and Elhoseiny, M. 2018. Efficient Lifelong Learning with A-GEM. *arXiv preprint arXiv:1812.00420*.

Chaudhry, A.; Rohrbach, M.; Elhoseiny, M.; Ajanthan, T.; Dokania, P. K.; Torr, P. H.; and Ranzato, M. 2019. Continual learning with tiny episodic memories.

Chen, Z.; and Liu, B. 2018. Lifelong machine learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 12(3): 1–207.

Cho, S.; Vasarhelyi, M. A.; Sun, T.; and Zhang, C. 2020. Learning from Machine Learning in Accounting and Assurance. *Journal of Emerging Technologies in Accounting*, 17(1): 1–10.

Coderre, D. 2009. *Computer-aided Fraud Prevention and Detection: A Step-by-step Guide*. John Wiley & Sons.

Debreceny, R. S.; and Gray, G. L. 2010. Data Mining Journal Entries for Fraud Detection: An Exploratory Study. *International Journal of Accounting Information Systems*, 11(3): 157–181.

Delange, M.; Aljundi, R.; Masana, M.; Parisot, S.; Jia, X.; Leonardis, A.; Slabaugh, G.; and Tuytelaars, T. 2021. A Continual Learning Survey: Defying Forgetting in Classification Tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Díaz-Rodríguez, N.; Lomonaco, V.; Filliat, D.; and Maltoni, D. 2018. Don't Forget, There is More than Forgetting: New Metrics for Continual Learning. *arXiv preprint arXiv:1810.13166*.

Fay, R.; and Negangard, E. M. 2017. Manual Journal Entry Testing: Data Analytics and the Risk of Fraud. *Journal of Accounting Education*, 38: 37–49.

French, R. M. 1999. Catastrophic Forgetting in Connectionist Networks. *Trends in Cognitive Sciences*, 3(4): 128–135.

Glorot, X.; and Bengio, Y. 2010. Understanding the Difficulty of Training Deep Feedforward Neural Networks. In *Proceedings of the thirteenth International Conference on Artificial Intelligence and Statistics*, 249–256.

Grabski, S. V.; Leech, S. A.; and Schmidt, P. J. 2011. A Review of ERP Research: A Future Agenda for Accounting Information Systems. *Journal of Information Systems*, 25(1): 37–78.

Hawkins, S.; He, H.; Williams, G.; and Baxter, R. 2002. Outlier Detection using Replicator Neural Networks. In *International Conference on Data Warehousing and Knowledge Discovery*, 170–180. Springer.

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

Hu, H.; Sun, T.; Vasarhelyi, M. A.; and Zhang, M. 2020. A Machine Learning Approach of Measuring Audit Quality: Evidence From China. *Available at SSRN 3732563*.

IFAC. 2009. *International Standards on Auditing 240, The Auditor's Responsibilities Relating to Fraud in an Audit of Financial Statements*.

Isele, D.; and Cosgun, A. 2018. Selective Experience Replay for Lifelong Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

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

Jans, M.; Alles, M. G.; and Vasarhelyi, M. A. 2014. A Field Study on the Use of Process Mining of Event Logs as an Analytical Procedure in Auditing. *The Accounting Review*, 89(5): 1751–1773.

Jans, M.; Lybaert, N.; and Vanhoof, K. 2010. Internal Fraud Risk Reduction: Results of a Data Mining Case Study. *International Journal of Accounting Information Systems*, 11(1): 17–41.

Jans, M.; Van Der Werf, J. M.; Lybaert, N.; and Vanhoof, K. 2011. A Business Process Mining Application for Internal Transaction Fraud Mitigation. *Expert Systems with Applications*, 38(10): 13351–13359.

Khan, R.; Clark, A.; Mohay, G.; and Suriadi, S. 2014. Detecting Fraud Using Transaction Frequency Data. *Information Technology in Industry*, 2(3): 79–88.

Khan, R.; Corney, M.; Clark, A.; and Mohay, G. 2009. A Role Mining inspired Approach to Representing user Behaviour in ERP Systems. In *Proceedings of Asia Pacific*

Industrial Engineering & Management Systems Conference 2009, 2541–2552. APIEMS Society.

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

Kingma, D. P.; and Ba, J. 2014. ADAM: A Method for Stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming Catastrophic Forgetting in Neural Networks. *Proceedings of the National Academy of Sciences*, 114(13): 3521–3526.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *nature*, 521(7553): 436.

Lee, C. S.; and Lee, A. Y. 2020. Clinical Applications of Continual Learning Machine Learning. *The Lancet Digital Health*, 2(6): e279–e281.

Lesort, T.; Caccia, M.; and Rish, I. 2021. Understanding Continual Learning Settings with Data Distribution Drift Analysis. *arXiv preprint arXiv:2104.01678*.

Li, Z.; and Hoiem, D. 2017. Learning Without Forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12): 2935–2947.

Lopez-Paz, D.; and Ranzato, M. 2017. Gradient Episodic Memory for Continual Learning. *Advances in Neural Information Processing Systems*, 30: 6467–6476.

McCloskey, M.; and Cohen, N. J. 1989. Catastrophic Interference in connectionist Networks: The Sequential Learning Problem. In *Psychology of Learning and Motivation*, volume 24, 109–165. Elsevier.

McGlohon, M.; Bay, S.; Anderle, M. G. M.; Steier, D. M.; and Faloutsos, C. 2009. SNARE: A Link Analytic System for Graph Labeling and Risk Detection. In *KDD'09: 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1265–1273.

Mendez, J. A.; and Eaton, E. 2021. Lifelong Learning of Compositional Structures. In *International Conference on Learning Representations*.

Nonnenmacher, J.; and Gómez, J. M. 2021. Unsupervised Anomaly Detection for Internal Auditing: Literature Review and Research Agenda. *The International Journal of Digital Accounting Research*, 21(27): 1–22.

Nonnenmacher, J.; Kruse, F.; Schumann, G.; and Marx Gómez, J. 2021. Using Autoencoders for Data-Driven Analysis in Internal Auditing. In *Proceedings of the 54th Hawaii International Conference on System Sciences*, 5748.

Parisi, G. I.; Kemker, R.; Part, J. L.; Kanan, C.; and Wermter, S. 2019. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113: 54–71.

PwC. 2020. *Fighting Fraud: A never-ending Battle, The Global Economic Crime Survey 2020*. PricewaterhouseCoopers LLP.

Rebuffi, S.-A.; Kolesnikov, A.; Sperl, G.; and Lampert, C. H. 2017. ICARL: Incremental Classifier and Representation Learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2001–2010.

Rolnick, D.; Ahuja, A.; Schwarz, J.; Lillicrap, T. P.; and Wayne, G. 2018. Experience Replay for Continual Learning. *arXiv preprint arXiv:1811.11682*.

Rusu, A. A.; Rabinowitz, N. C.; Desjardins, G.; Soyer, H.; Kirkpatrick, J.; Kavukcuoglu, K.; Pascanu, R.; and Hadsell, R. 2016. Progressive Neural Networks. *arXiv preprint arXiv:1606.04671*.

Schreyer, M.; Sattarov, T.; and Borth, D. 2021. Multi-view Contrastive Self-Supervised Learning of Accounting Data Representations for Downstream Audit Tasks. In *Proceedings of the International Conference on Artificial Intelligence (ICAIF)'21*. Association of Computing Machinery (ACM).

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

Schreyer, M.; Sattarov, T.; Gierbl, A. S.; Reimer, B.; and Borth, D. 2020. Learning Sampling in Financial Statement Audits using Vector Quantised Variational Autoencoder Neural Networks. In *Proceedings of the International Conference on Artificial Intelligence (ICAIF)'20*. Association of Computing Machinery (ACM).

Schultz, M.; and Tropmann-Frick, M. 2020. Autoencoder Neural Networks versus External Auditors: Detecting Unusual Journal Entries in Financial Statement Audits. In *Proceedings of the 53rd Hawaii International Conference on System Sciences*.

SEOW, P. S.; Pan, G.; and Suwardy, T. 2016. Data Mining Journal Entries for Fraud Detection: A Replication of Debreceny and Gray's (2010) Techniques. *Journal of Forensic and Investigative Accounting*, 8(3): 501.

Shin, H.; Lee, J. K.; Kim, J.; and Kim, J. 2017. Continual Learning with Deep Generative Replay. *arXiv preprint arXiv:1705.08690*.

Stephan, S.; Lahann, J.; and Fettke, P. 2021. A Case Study on the Application of Process Mining in Combination with Journal Entry Tests for Financial Auditing. In *Proceedings of the 54th Hawaii International Conference on System Sciences*, 5718.

Sun, T. 2019. Applying Deep Learning to Audit Procedures: An Illustrative Framework. *Accounting Horizons*, 33(3): 89–109.

Thiprungsri, S.; and Vasarhelyi, M. A. 2011. Cluster Analysis for Anomaly Detection in Accounting Data: An Audit Approach. *International Journal of Digital Accounting Research*, 11.

Toneva, M.; Sordoni, A.; Combes, R. T. d.; Trischler, A.; Bengio, Y.; and Gordon, G. J. 2018. An Empirical Study of Example Forgetting During Deep Neural Network Learning. *arXiv preprint arXiv:1812.05159*.

Vasarhelyi, M. A.; Alles, M. G.; and Kogan, A. 2004. Principles of Analytic Monitoring for Continuous Assurance. *Journal of Emerging Technologies in Accounting*, 1(1): 1–21.

Vasarhelyi, M. A.; and Halper, F. B. 1991. The Continuous Audit of Online Systems. In *Auditing: A Journal of Practice and Theory*. Citeseer.

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

Werner, M. 2016. Process Model Representation Layers for Financial Audits. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*, 5338–5347. IEEE.

Werner, M.; and Gehrke, N. 2015. Multilevel Process Mining for Financial Audits. *IEEE Transactions on Services Computing*, 8(6): 820–832.

Xu, B.; Wang, N.; Chen, T.; and Li, M. 2015. Empirical Evaluation of Rectified Activations in Convolutional Network. *arXiv preprint arXiv:1505.00853*.

Zenke, F.; Poole, B.; and Ganguli, S. 2017. Continual Learning through Synaptic Intelligence. In *International Conference on Machine Learning*, 3987–3995. PMLR.

Zhai, M.; Liu, H.; and Sun, F. 2019. Lifelong Learning for Scene Recognition in Remote Sensing Images. *IEEE Geoscience and Remote Sensing Letters*, 16(9): 1472–1476.

Zupan, M.; Budimir, V.; Letinic, S.; et al. 2020. Journal Entry Anomaly Detection Model. *Intelligent Systems in Accounting, Finance and Management*, 27(4): 197–209.