LEUPHANA
UNIVERSITÄT LÜNEBURG

Toward the Digitalization of Auditing: Applying Machine Learning for Information Extraction from Invoices

Der Fakultät Management und Technologie
der Leuphana Universität Lüneburg zur Erlangung des Grades

Doktor der Wirtschaftswissenschaften
- Dr. rer. pol. -

vorgelegte Dissertation von Felix Friedrich Anton Krieger

geboren am 20.01.1989 in Bamberg

eingereicht am	2.5.2023

Erstbetreuer:	Prof. Dr. Paul Drews, Leuphana Universität Lüneburg
Erstgutachter:	Prof. Dr. Paul Drews, Leuphana Universität Lüneburg
Zweitgutachter:	Prof. Dr. Burkhardt Funk, Leuphana Universität Lüneburg
Drittgutachter:	Prof. Dr. habil. Ansgar Scherp, Universität Ulm

Abstract

Motivation: Artificial intelligence, most prominently in the form of machine learning, is shaping up to be one of the most transformational technologies of the 21st century. Auditors are among the professions forecasted to be the most affected by artificial intelligence, as the profession encompasses many highly structured and repetitive tasks. Automating such tasks would naturally increase the efficiency of financial statement audits. By allowing auditors to focus on higher value-added tasks, and the capability to analyze large volumes of data at a fracture of the time a human would need, artificial intelligence would also benefit the effectiveness of auditing. Despite these benefits, to this day, the actual adoption of artificial intelligence in the audit domain remains rather limited. The audit profession is highly regulated and has to consider requirements regarding, e.g. the application of professional standards, codes of conduct, and data protection obligations. Hence, the question arises of how audit firms can be supported in their efforts to adopt artificial intelligence and how machine learning systems can be designed to comply with the specific demands of the audit domain.
Research Approach: The goal of this dissertation is to better understand the adoption of artificial intelligence in the audit domain and to actively support the adoption of artificial intelligence in auditing based on this understanding. To this end, we employ a mixture of research methods. On the one hand, the research presented here adopts a qualitative approach, examining the adoption of artificial intelligence and other advanced analytical technologies of the audit domain through taxonomy development and grounded theory. The findings of these studies inspire the second stream of work within this dissertation, which adopts a quantitative and design-oriented approach: It focuses on using machine learning to extract information from invoices for tests of details. Tests of details are essential substantive audit procedures used in nearly every audit. This dissertation proposes a new machine learning model architecture for information extraction from invoices, compares different machine learning models, and proposes design principles for machine learning pipelines for an audit application addressing the test of details through action design research.
Contribution: This dissertation presents several contributions to the research on the adoption of artificial intelligence in auditing. To form an initial understanding of the problem environment around the application of artificial intelligence to auditing, we developed a taxonomy. The taxonomy integrates the audit and technology perspective in a structured manner and supports the description of use cases in the audit domain. The dissertation further presents a process theory that illustrates how audit firms adopt artificial intelligence and other advanced data analytics technologies. The study uses a previously unused theoretical perspective, which allows for contextualizing known technology adoption factors in the audit domain. Based on the understanding of the problem environment obtained through the taxonomy and process theory, we engaged in developing artifacts and methods for applying information extraction from invoices. Here, we offer the first contribution by developing a novel graph-based neural network architecture and showing its ability to extract information accurately from invoice data sets with a significant layout variance. The second contribution deepens the understanding of the effects of layout distributions on the generalization ability of neural networks: We compared different model types and disaggregated the evaluation into in-sample and out-of-sample layouts. We show that the gap in accuracy between in- and out-of-sample layouts varies across models. To arrive at these results, we developed an end-to-end machine

learning pipeline. As part of the last contribution of this dissertation, we automatically orchestrated this pipeline which serves as a structured approach to evaluate and deploy machine learning models for information extraction from invoices. We designed it such that new models from the continuously flowing stream of research are easily integrated. By reflecting on the genesis of the pipeline and the design choices that guided its emergence, we also propose a set of design principles for information extraction pipelines in audit tools.

Limitations: The results presented in this dissertation must be seen in the light of some limitations. First, we obtained the taxonomy's dimensions and characteristics to describe use cases from the scientific literature. Use cases only identified in practice might not be characterized in their entirety by the taxonomy. The presented process theory is grounded in data obtained from expert interviews. Hence, the sampling of interview partners can affect its generalizability. For instance, most of our interview partners are located in Germany and take on roles in the upper management of their respective organizations. The results presented in the design-oriented studies are limited by the characteristics of the available data sets. These characteristics include the languages of the documents, which is primarily English, their quantity, and the recurring vendor layouts. Finally, we conducted the action design-oriented research within a large multinational audit firm. Hence, the requirements for the developed artifact and the proposed design principles might not be transferable to smaller firms.

Future Research: Several threads are laid out in the presented body of work that may be picked up in future research endeavors. The taxonomy could be updated to the most recent developments in artificial intelligence, such as generative and conversational systems. In the process theory, the nature of the relationship between the contextual factors and the adoption process could be explored in more detail. Concerning information extraction for the test of details, future research could explore how the extraction results could be parsed into standardized formats or how they could be internally validated. Larger audit firms have clients from a variety of countries, which begs the question of whether language-specific models or multilingual models are better. In this context, the need for labeled training data poses a challenge for adapting models to different languages. Therefore, future inquiries could explore how the utilization of training paradigms such as active learning to reduce the need for labeled training data.

Keywords

Audit Digitization, Natural Language Processing, Document Analysis

"We are stuck with technology when what we really want is just stuff that works."
- Douglas Adams

Acknowledgement

Writing these lines fills me with joy. They represent the conclusion of a journey that lasted over five years and was filled with setbacks and achievements. Finishing this journey would have been unimaginably more difficult without the support and encouragement of my academic supervisors, colleagues, friends, and family.

First, I want to thank Professor Paul Drews for his outstanding guidance throughout the years, his differentiated criticism, and for helping me to untangle my thoughts over and over again. I also would like to thank Professor Burkhardt Funk and Professor Patrick Velte, who have helped me to navigate this highly interdisciplinary topic with their valuable input regarding machine learning and auditing. I highly appreciate Professor Ansgar Scherp's interest in my work and his review of my dissertation. Additionally, I want to thank the other doctoral students at the institute, particularly those from C4.320 and C4.308, for their advice and for teaching me the ropes of academia.

This journey could probably have never started without the scholarship I received from EY. But it was the freedom, trust, and support I received from my counselors and colleagues at EY that helped me succeed. For this, I want to thank Dr. Michael Wiese, Nikola Bubalo, Till Heckschen, and Dr. Till Blume.

Further, I want to express my sincerest gratitude for the words of encouragement and understanding from my family throughout the years, especially my mother Eva. The same applies to my friends that let me ramble about my research and helped me divert my thoughts.

Last, but definitely not least, I want to thank my wife Carolin. You have lifted me up when I struggled, reminded me - sometimes with a bit of emphasis - of life outside of work, and cheered for me all the way. I'm lucky to share my life with you.

.....

Author's signature

Contents

Preamble

Introduction **10**
1.1 Motivation . . . 10
1.2 Research Questions . . . 11
1.3 Research Approach . . . 14
Background **17**
2.1 Applying Emerging Technologies to Audit Procedures . . . 17
2.1.1 The Audit Risk Model . . . 18
2.1.2 Audit Procedures and Audit Evidence . . . 19
2.1.3 Emerging Technologies in Auditing . . . 21
2.2 Information Extraction from Documents for Tests of Details . . . 23
2.2.1 Deep Neural Networks for Information Extraction . . . 25
2.2.2 Information Extraction from Invoices . . . 27
2.2.3 Evaluation of Models for Information Extraction from Invoices . . . 29
Contributions **33**
3.1 Understanding the Adoption of ADA in Auditing . . . 33
3.2 Assessing the Generalization Ability of Neural Networks for Information Extraction . . . 34
3.3 Designing Information Extraction Pipelines for Audit Tools . . . 36
Limitations **38**
Future Research **40**
Conclusion **42**
Bibliography **44**

Publications

Leveraging Big Data and Analytics for Auditing ... **55**
Explaining the (Non-) Adoption of Advanced ... **69**
Information Extraction from Invoices: A Graph Neural ... **107**
Automated Invoice Processing: Machine Learning-Based ... **125**
Benchmarking Machine Learning Models in Auditing: ... **154**

Appendix

Appendix to chapter II **176**
Appendix to chapter IV **182**
Author contributions **183**
Complete list of publications **187**
Curriculum Vitæ **188**

List of Abbreviations

ADA Advanced Data Analytics

AR Audit Risk

ARM Audit Risk Model

AI Artificial Intelligence

CAATT Computer-assissted audit techniques and tools

CR Control Risk

CV Computer Vision

DKCF Design Knowledge Contribution Framework

DL Deep Learning

DR Detection Risk

ERP Enterprise Resource Planning

EU European Union

FN False Negative

FP False Positive

GAT Graph Attention

IAASB International Auditing and Assurance Standards Board

IE Information Extraction

IR Inherent Risk

ISA International Standard on Auditing

LM Language Models

ML Machine Learning

MM Material Misstatement

NLP Natural Language Processing

NN Neural Networks

OCR Optical Character Recognition

PDF Portable Document Format

RMM Risk of Material Misstatement

RPA Robotic Process Automation

RQ Research Question

TN True Negative

TOD Test of Details

TP True Positive

VRD Visually Rich Document

Preamble

Chapter 1

Introduction

1.1 Motivation

The ability to process and analyze large quantities of heterogeneous data has become a differentiating factor in the marketplace. It has shown the potential to transform entire industries, such as marketing, translation services, or manufacturing. This transformational effect spills over into the audit industry: By analyzing the data captured by their clients and external data, auditors could better understand their clients' businesses and the associated risks. The analysis of data could also be used to increase the efficiency of audits: Repeating patterns in data can be utilized to automate routine tasks within an audit engagement. In digitalizing their business, auditors are also experiencing a pull effect: To be further perceived as a trustworthy and knowledgeable business partner by their clients, auditors need to stay on top of technological developments which affect the financial domain. Analyzing their client's data may also help auditors generate helpful insights for their clients, leading to advantages in the highly competitive market for external audit services. One cluster of technologies is considered to enable such analyses: Data analytics, artificial intelligence (AI), and big data (Alles and Gray, 2016; Cao et al., 2015; Chan et al., 2018; Kokina and Davenport, 2017). As they represent an evolutionary step beyond the well-established data visualization and interrogation techniques currently used in auditing, they are referred to as advanced data analytics (ADA). Especially AI, most prominently in the form of its subfields machine learning (ML) and deep learning (DL), is attributed to be potentially disruptive to the profession: As auditing encompasses many highly repetitive and structured tasks (Abdolmohammadi, 1999), auditors have been forecasted to be amongst the professions most affected by computerization through ML (Frey and Osborne, 2017). Apt examples for such tasks are tests of details (TOD); audit procedures that substantiate the transactions recorded in the client's accounting system with their source documents (ISA 330, 2009). TODs are central audit procedures and, as such, are performed on nearly every audit engagement. By making them more efficient, automating TODs also drastically increases the number of transactions that may be substantiated, ultimately leading to higher audit effectiveness.

Audit firms are known to lag in adopting new technologies (Alles and Gray, 2016). Despite the strong arguments for using ADA in auditing, scientific evidence points towards a rather slow adoption in practice (Eilifsen et al., 2019; Salijeni et al., 2019). The potential reasons for this are manifold: Audit firms operate within a complex institutional environment governed by both international and national professional standards and codes

of conduct, government regulations and data protection obligations. The institutional environment may limit the extent to which technologies can be applied. As they provide a professional service, the relationship between audit firms and their clients must also be considered. The clients exhibit varying organizational characteristics that technology applications must account for (Dagilienė and Klovienė, 2019). Furthermore, audit clients have expressed concerns regarding the security and governance around the data provided to auditors (Salijeni et al., 2019). Nevertheless, there are also differences in the audit firms' characteristics which may influence their use of technology, such as their respective size and structure (Dagilienė and Klovienė, 2019).

The complexity induced by the institutional environment and the characteristics of both audit firms and their clients is met by the complexity of ADA: The range of methods to draw from is ample; a plethora of techniques can be utilized toward a multitude of analysis goals, such as regression, classification, clustering, graph analytics, or process mining. This complexity is amplified by the versatile nature of the underlying data. Sources of structured data relevant to auditors are (inter alia) the general and subsidiary ledgers, trial balances, and charts of accounts. However, only a small percentage of data found in companies are structured (Cukier, 2010). A significant proportion of data accumulated by companies is in textual format (Chen et al., 2012), such as business documents. And even within the domain of textual data, structural differences can be found; Many types of business documents which are commonly used for TODs, such as invoices, delivery notes, or receipts, are characterized by sparsity and a two-dimensional layout, as opposed to the sequential text commonly found in contracts (Katti et al., 2018). This applies another layer of complexity to the methods in data analytics and ML, as these structural characteristics required specialized approaches. Consequently, they have drawn the attention of the research community, leading to a continuously growing stream of research on specialized ML models and document representations (Denk and Reisswig, 2019; Katti et al., 2018; Liu et al., 2019; Lohani et al., 2019; Xu, Li, et al., 2020; Yu et al., 2020; Zhang et al., 2020; Zhao et al., 2019).

The goal of this dissertation is to understand and actively advance the adoption of ADA in the audit domain, to which it attempts to master this complexity.

1.2 Research Questions

As pointed out above, two matters of rich complexity need to be aligned for the application of ADA to auditing. At the beginning of this research, therefore, stood the challenge of understanding how this alignment could be achieved to create a benefit. Different examples of the application of ADA had been proposed in the scientific literature, including the first literature reviews. However, the reviews were primarily concerned with the ADA techniques' nature and integration into the audit process, mainly ignoring the nature of the underlying data. Hence, we aimed for a way to structure the literature holistically. Additionally, this structure should support us in identifying use cases for the design-oriented phase of the dissertation through expert interviews. We concluded that a taxonomy would enable us to achieve both goals simultaneously. This led us to pose the first research question (RQ):

Which dimensions and characteristics should a taxonomy of ADA use cases comprise?

Research Question 1

Audit firms are rather slow in adopting ADA, especially AI. The factors affecting technology adoption in audit firms are manifold, such as the institutional environment, the client-auditor relationship, and the characteristics of audit firms. They have been explored in previous research and are generally understood. However, previous research has treated the *process* by which audit firms adopt ADA as a black box. Uncovering this process allows for contextualizing known adoption factors, deepening the understanding of ADA adoption. It can further inspire and guide research aiming at constructing artifacts by allowing their design to be optimized for successful adoption. To shed light into the black box of ADA adoption in auditing, the following RQ is addressed:

Which process do audit firms utilize to adopt ADA technologies and how is this process affected by contextual factors?

Research Question 2

As the overarching goal of this dissertation is to advance the adoption of ADA in the audit domain actively, it also aims to construct practice-relevant artifacts and design knowledge. From the empirical study conducted toward answering RQ2, we were aware that the considerable heterogeneity of audit clients and their data posed a challenge to the application of ADA. This also concerns unstructured data: Audit clients receive invoices from a usually large number of different vendors. Similar to other business documents, such as receipts or delivery notes, invoices are characterized by a vendor-specific layout. A cooperating audit firm had identified the need to automatically extract structured representations of the documents' contents, as they could be used to automate TODs. In this context, an ADA application should be able to generate these representations from a significant heterogeneity of document layouts. Based on our examination of the literature, we proposed DL-based information extraction (IE) to meet this requirement. The peculiar characteristics of invoices inspired researchers to propose different layout-preserving structural representations of invoices that DL models could consume. A promising approach was representing documents as graphs and applying graph neural networks for IE. However, the previous empirical evaluation of the proposed graph models had left doubt as to whether they would be capable of effective IE when trained and evaluated on invoices exhibiting a large variety of layouts. Hence, this dissertation further seeks to answer the following research question:

How can graph-based neural networks be applied to extract key items from invoices with a high variety of layouts?

Research Question 3

When our research toward RQ3 had concluded, the list of publications proposing DL

models for IE from invoices and similar documents had grown. Different classes of approaches to representing documents and correspondent models had started to emerge: Graph-based models, grid-based models, and transformer models. As most of these works relied on proprietary data sets and used different evaluation metrics, a direct comparison from the literature was not feasible. We, therefore, chose to evaluate several different approaches on an experimental data set provided by the corresponding audit firm. The distribution of layouts in the data set was characterized by a minority of highly recurring vendors and a long tail of rarely occurring vendors. In the context of the bias-variance trade-off (James et al., 2021, pp. 33-36), we were especially interested in the models' accuracy over invoices exhibiting layouts not present in the training data. Since this vendor distribution is representative of many companies (Koch, 2019), and the effects of layout distributions in the training and evaluation data for ML-based IE are generally not well understood, or respectively underreported, in the literature, we posed the following RQ:

> *How do ML-based approaches to IE from invoices respond to skewed vendor distributions?*

Research Question 4

The body of work around IE models from invoices and similar visually rich documents is continuously growing, giving rise to a plethora of models. The accuracy of the IE step is crucial for automated TODs, as it reduces the human effort to inspect the results. Therefore, audit firms have a strong incentive to evaluate new models continuously. In addition, especially large audit firms face a considerable heterogeneity of clients from a multitude of countries. This heterogeneity poses the challenge of evaluating IE models for different contextual settings: The *"no-free-lunch"* theorem for supervised learning states that no model can be assumed to be better than all other models across a multitude of different settings. Therefore, the model evaluation should be designed to be performed continuously. Upon evaluation, the trained models need to be deployed into an audit tool to be used in practice. Both challenges can be addressed using automated ML pipelines. To accommodate the requirements of the audit domain in the design of such pipelines, we formulate the last RQ in this dissertation:

> *How can an information extraction pipeline for the TOD be designed and which design principles could guide the development of ML benchmarking pipelines?*

Research Question 5

The individual RQs outlined in this section are interrelated in the sense that the associated research inspires and draws from one another. However, addressing them requires different approaches to research, we outlined in the following subsection.

1.3 Research Approach

The individual RQs are addressed in chapters I through V. Each chapter represents one self-contained piece of research, which addresses one of the RQs put forward in the previous chapter. They employ different research methodologies and offer self-contained research contributions. Table 1.1 lists individual chapters, along with the respective RQ they address, the utilized research methods, and their contributions.

When the work on this dissertation started, the research on the application of ADA to auditing was still in its infancy. It lacked the coherence and structure to systematically identify interesting areas of application, that would allow us to contribute to both research and practice. This led us to formulate RQ1 and develop the use case taxonomy presented in chapter I. Taxonomies are theories for analysis and useful for understanding emerging phenomena (Gregor, 2006). For developing the taxonomy, we drew upon the method proposed by Nickerson et al. (2013). The method involves 'build/test' cycles, in which the dimensions and characteristics of the taxonomy can be either deducted ('conceptual-to-empirical') or inducted ('empirical-to-conceptual'). In the first cycle, we developed dimensions and characteristics from the conceptual literature on ADA in auditing and ADA in general. To evaluate the results of the first cycle, we applied the taxonomy to the ADA use cases which had been published in scientific outlets. We initiated a second cycle from this evaluation, where we inductively added additional characteristics.

Based on our understanding of the field gained through the taxonomy, we began an empirical qualitative study to identify use cases that were still under-addressed in practice.

Table 1.1: Methodology and contribution of the papers presented in this dissertation.

Chapter	Title	RQ	Methodology	Classification according to Gregor (2006)	Contribution
I	"Leveraging Big Data and Analytics for Auditing: Towards a Taxonomy"	RQ1	Taxonomy development (Nickerson et al., 2013)	Theory for analysis	Integrated perspective on auditing, data analysis, and data management; taxonomy to support the analysis and ideation of use cases for ADA in auditing.
II	"Explaining the (non-) adoption of advanced data analytics in auditing"	RQ2	Grounded theory (Corbin and Strauss, 1990)	Theory for explanation	Process theory depicting the adoption process of ADA in audit firms; contextualization of known factors of technology adoption in the audit domain.
III	"Information Extraction from Invoices: A Graph Neural Network Approach for Datasets with High Layout Variety"	RQ3	Machine learning model development	Theory for design and action	Empirical evaluation of the applicability of graph-based ML for IE from invoice data sets with a large layout variability; development of new model architecture.
IV	"Automated Invoice Processing: Machine Learning-Based Information Extraction for Long Tail Suppliers"	RQ4	Machine learning model benchmarking	Theory for design and action	Empirical evaluation of different ML-based approaches for IE from invoices sourced from long-tail suppliers; presentation and implementation of model evaluation methodology which accounts for the distribution of layouts in a data set.
V	"Benchmarking Machine Learning Models in Auditing: Toward an Information Extraction Pipeline for the Test of Details"	RQ5	Action design research (Sein et al., 2011)	Theory for design and action	Design and implementation of machine learning pipelines for a TOD application; development of design principles for IE in audit applications.

From the first results of the empirical study, we concluded that the adoption of ADA in practice was progressing slowly, and only a few ADA applications had found an initial

adoption. Consequently, the focus of the study shifted toward RQ2. In chapter II, we investigated by which process audit firms adopt ADA techniques and which contextual factors affect it. We proposed a process theory that was derived from expert interviews using the grounded theory methodology (Corbin and Strauss, 1990). According to Gregor (2006), this type of theory corresponds to theories of explanation that form an understanding of *how* and *why* phenomena occur. They can be further used to inform theories for design and action, which propose "*[...] how to do something*" (ibid.).

Kogan et al. (2019) call for more design-oriented research on adopting ADA in the audit domain. Following this call and the results of our empirical study, we aimed to generate relevant design knowledge in the second part of the dissertation. Design knowledge is inherently prescriptive and developed through building and evaluating artifacts (Hevner et al., 2004; Sein et al., 2011). To achieve this goal, we adopted the action design research methodology (Sein et al., 2011), in which the researcher is placed within an organization to solve "[...] *immediate organizational problems*" (ibid.). To this end, we cooperated with a large international audit firm.

Based on examining the scientific literature and the results of the empirical study, we discussed promising areas of research for the design-oriented part of the dissertation with the cooperating audit firm. They had identified the need to analyze unstructured data systematically, most importantly, business documents. Business documents obtained from clients are essential sources of audit evidence. They are used in audit procedures designed to assess the client's business risk and to respond to the assessed risks. For the scope of the joint research between us and the audit firm, we decided to focus on automated information extraction (IE) for tests of details (TOD). TODs encompass reconciling business documents to accounting records, such as general or subsidiary ledgers. IE requires cognitive capabilities, and as such, is best addressed through AI resp. DL. A solution for the IE from contracts was already under active development at the audit firm. Consequently, it was decided to focus on invoices, which are structurally different from contracts, as they are characterized by sparse use of text and vendor-dependent layouts.

The work described in chapters III through IV was conducted within the cooperating audit firm. I took on the role of a data scientist within an interdisciplinary team composed of a statistician, two computer scientists, and a frontend designer. This specific team operated within the audit firm's digital innovation unit and was concerned with developing a software tool to automate the abovementioned reconciliation step for invoices. My specific role within the team was developing and deploying ML models for IE from invoices. While my role within the team was very specific, the size of the team allowed me to obtain extensive insights into all technical and business-related dependencies.

Initially, I wanted to prove that the problem of IE from invoices was solvable through DL. To better understand the development, training, and evaluation of DL models, I started experimenting with DL for text classification. Starting with recurrent neural networks, I worked my way up to graph neural networks. The first graph-based models for IE from invoices had already been proposed in the literature. However, the experimental setups described in the respective contributions had left doubt whether they were suitable in application contexts where the invoices exhibit many different layouts. This led to the formulation of RQ3. To shed some light on this issue, chapter III evaluates how well

graph-based ML models are capable of dealing with a large variety of different invoice layouts, and introduces a novel model architecture. Following up on the insights of chapter III, we wanted to explore how different types of DL models (grid-based, graph-based, transformer) compared to each other. We were specifically interested in the effects of layout distributions on the training and evaluation of the models, leading us to pose RQ4. Consequently, in chapter IV, multiple ML models are benchmarked against each other on a common set of invoices. We present an evaluation strategy that explicitly considers the distribution of different invoice layouts within the data set. To this end, I reimplemented most of the models considered in our study from scratch, as openly available implementations were unavailable. I developed a research pipeline for end-to-end data preparation, hyperparameter tuning, model training and model evaluation to automate the required experiment runs using Microsoft's Azure Machine Learning platform. Through the platform, the execution of processing steps in a pipeline can be orchestrated across GPU-enabled nodes in a cluster. Chapter V reflects on the genesis of the research pipeline and its continued development into the IE backend for the TOD tool. It contextualizes the work presented in the chapters III and IV within the *Building, Intervention and Evaluation* cycles of ADR, and reflects the design choices against the theory proposed in chapter II in the *Reflection and Learning* step. Based on this reflection, we propose design principles for IE pipelines for audit applications.

Chapter 2

Background

Most of the research presented in this dissertation is trans- and interdisciplinary. It draws from areas such as technology adoption and machine learning-based information extraction and applies it to the audit domain, contributing to technology adoption in auditing and information extraction from invoices. This chapter provides the necessary background to contextualize the research and its contributions. The remainder of this chapter is structured as follows: In section 2.1, I give an overview of auditing and audit procedures and the previous research on applying ADA to audit procedures. Section 2.2 introduces using deep neural networks for IE in the context of TODs. Toward the end of both sections, I delineate the research gaps addressed by the abovementioned RQs.

2.1 Applying Emerging Technologies to Audit Procedures

Financial statements fulfill a purpose that is central to the orderly functioning of financial markets: They reduce the information asymmetry between a company's management and external parties, such as stake- and shareholders (Healy and Palepu, 2001; Jensen and Meckling, 1976; Knechel and Salterio, 2016, p. 8). Information asymmetries arise in the case of uneven distribution of information between two parties of an economic transaction, i.e., one party is better informed about the quality of the transaction's subject than the other (Knechel and Salterio, 2016, p. 8). In the case of financial statements, this asymmetry relates to information about the company's financial situation. To its addressees, it is therefore vital that the information in these financial statements is reliable to avoid unfavorable or even harmful decisions. The purpose of an external audit in this context is to augment the information's reliability (ISA 200, para. 3), and thus reduce the risks associated with their utilization for decision making (Knechel and Salterio, 2016, pp. 8-9). To this end, the external auditor[1] expresses an opinion on the accuracy of the company's accounts, in all material aspects, with respect to an applicable financial reporting framework (ISA 200, para. 3).

Financial reporting frameworks govern the preparation and presentation of financial statements. Their applicability depends on a company's country of incorporation and its

[1]Different types of information exchanged between two parties can be subject to an external audit. For the remainder of this dissertation, the term 'auditor' strictly refers to a certified accounting professional performing independent external audits of financial statements.

legal form. For instance, publically listed companies in member states of the European Union (EU) are obligated to prepare and publish their financial statements in accordance with the International Financial Reporting Standards (IFRS) (EU, European Parliament, 2002). Corresponding to the statutory application of financial reporting frameworks, companies are also obligated to submit their financial statements to an external audit: In the EU, external audits are required for all listed companies, banks, and insurances (EU, European Parliament, 2014). The U.S. regulations foresee statutory audits for all public companies (IFAC, 2016). Companies that do not fall under these categories may also opt for a voluntary audit, as the augmented reliability of their financial statements can facilitate raising equity or debt capital (Knechel and Salterio, 2016, p. 7). In analogy to the financial reporting frameworks, auditors adhere to statutory professional standards and professional codes of conduct in the provision of external audits. Efforts to harmonize different international standards have led to the creation of the international standards on auditing (ISA), which are established by the International Auditing and Assurance Standards Board (IAASB) (ibid., p. 30). The ISA are mandatory in the EU; however, individual member states may adopt them into their respective local standards (Accountancy Europe, 2015).

2.1.1 The Audit Risk Model

As per the ISA 200, auditors are required to base their expressed audit opinion on a reasonable level of assurance whether the client's financial statements are free from any material misstatements MM (ISA 200, para. 5). In formal terms, a misstatement is a deviation of a financial statement item from its correct amount, classification, presentation, or disclosure with respect to the applicable reporting framework. A deviation is considered material if it would affect the decision of the financial statement's user. Conceptually, it is essential to highlight that *reasonable* assurance is a high -but not absolute- level of assurance. The standards acknowledge the limitations of an audit, through which the collected evidence is of a rather persuasive than conclusive nature (ISA 200). Reasonable assurance is achieved by reducing the *audit risk* (AR) to an acceptable low level through the collection of sufficient appropriate audit evidence (ISA 200). The AR refers to the auditor's risk of expressing a wrong audit opinion in the presence of MM (ISA 200, para. 5). In the ISA, the AR is understood as a function of the risk of material misstatement (RMM) and the detection risk (DR):

$$AR = RMM \times DR = (IR \times CR) \times DR \tag{2.1}$$

Equation 2.1 reveals the so-called audit risk model (ARM) - the methodological core of auditing. In the ARM, the RMM is the risk that the financial statements are materially misstated *prior* to the audit. The RMM can be further decomposed into the inherent risk (IR) and the control risk (CR) (ISA 200, para. 13). IR is the risk that a MM may occur without considering the existence of an internal control system (ISA 200, para. 13). The CR describes the risk of the audit client's internal control system being unable to detect or prevent MMs (ISA 200, para. 13). To determine the RMM, the auditor performs a combined assessment of IR and CR, applying his or her professional judgement[2] (ISA 200; ISA 315). The RMM is generally considered to be exogenous and can only be assessed,

[2]Professional judgment is the application of the auditor's relevant knowledge and experience to make informed decisions (ISA 200, 2009).

not influenced. Consequently, the auditor must reduce the DR to reduce the AR to an acceptably low level. The DR is defined as the risk that the *audit procedures* performed in response to the assessed RMM fail to detect (potentially material) misstatements (ISA 200, 2009, paras. 13, A42-A44).

According to ISA 200, the auditor is neither expected nor able to reduce the AR to zero (ibid., para. A45). This is due to the abovementioned inherent limitations to an audit, which arise from the nature of financial reporting (ISA 200, para. A46), the nature of audit procedures (ISA 200, para. A47), and the requirement to perform the audit within a reasonable time and cost budget (ISA 200, para. A48). As pointed out by Knechel and Salterio (2016, p. 66), the usage of language like "*reasonable*" in the ISA shows that subjectivity is an inherent element in auditing, as a consequence of the many forms of uncertainty faced by auditors. This also applies to the nature of the abovementioned professional judgment conducted by an auditor, which in consequence, means that the risk terms in the ARM are usually not expressed as precise measurements (ISA 200, para. A32).

2.1.2 Audit Procedures and Audit Evidence

The previous section introduced the methodological basis for auditing, the ARM. This risk-based approach to auditing entails a process that is encoded in the standards, depicted in Figure 2.1: Auditors must assess the RMM (ISA 315) and design appropriate responses (ISA 330). For the assessment of the RMM, ISA 315 requires the auditor to first develop an understanding of the client's business and its environment, along with its internal control system (ISA 315, paras. A48-A183). Based on this understanding, the auditor identifies significant risks and determines the nature, timing, and extent of further audit procedures (ISA 315, paras. A184-A236). This process is to be understood as an iterative one; the understanding of the risks and appropriate consequent responses may evolve during the audit (ISA 315, para. 7).

Audit procedures are activities performed by the auditor to gather audit evidence (ISA 500), on which the audit opinion is based (ISA 700). Depending on the context of their application, they may be used for risk assessment, as tests of controls or as substantive audit procedures (ISA 500, para. A10). Tests of controls are employed to assess the effectiveness of the controls embedded in the internal control system in preventing or detecting MM (ISA 500, para. A29). They support the auditor's assessment of the CR but do not provide evidence that the financial statements are in accordance with the respective financial reporting framework. To reduce the DR, the auditor must employ substantive procedures. Substantive procedures detect MM by testing individual transactions, balances, and disclosures (ibid., pp. 317-318). Figure 2.1 contextualizes the use of audit procedures within the ARM.

The ISA mention two types of substantive procedures: *Analytical procedures* and *tests of details*. *Analytical procedures* are concerned with the analysis of both financial and non-financial data to identify inconsistencies between different sources of information or deviations from expected values (ISA 500, paras. A21). *Tests of details* entail activities such as the *inspection* of documents, the *observation* of activities as others are performing them, obtaining *confirmation* (e.g. over outstanding liabilities) from a third party, and

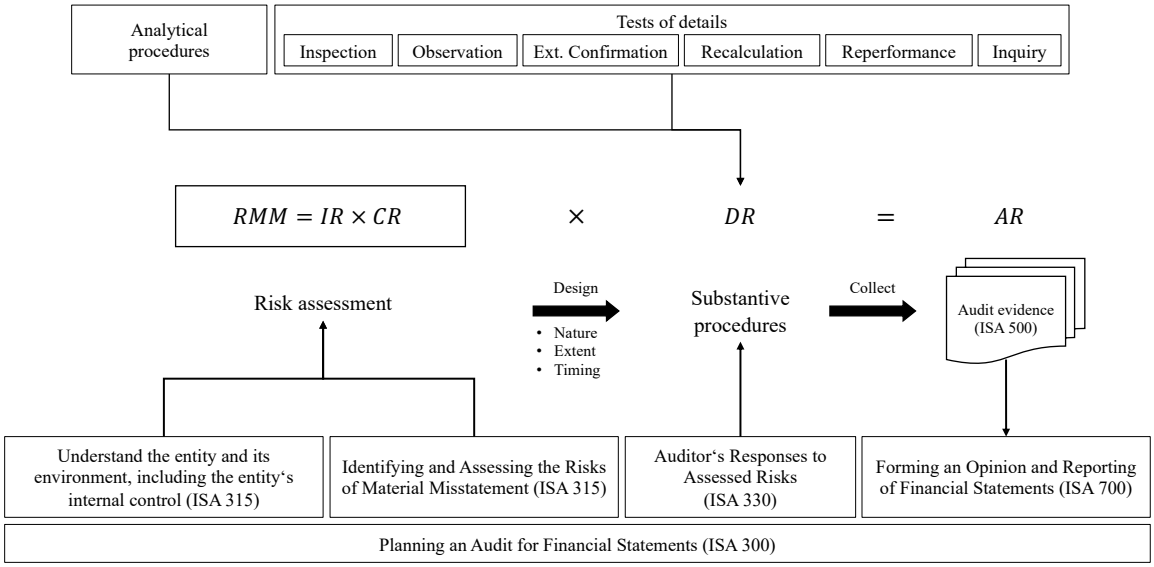

Figure 2.1: Contexualization of the application of substantive audit procedures within the ARM.

evaluating the mathematical correctness of documents or accounting records (*recalculation*) (ISA 500, paras. A2, A14-A25, A30). The client's accounting records are a central source of audit evidence to be collected via substantive procedures. Accounting records comprise both the accounting entries (e.g., general and subsidiary ledgers, journal entries) and any supporting documents (e.g., invoices, contracts, records of electronic fund transactions) (ISA 500, paras. 5, A1). Through their reconciliation, the auditor may gather evidence that the accounting records are internally consistent (ISA 500, para. A7), e.g., by reconciling recorded transactions with their supporting documents (Werner et al., 2021).

Audit evidence comprises information collected by the auditor that corroborates and supports, or contradicts, the client management's assertions regarding the accordance of the financial statements with the applicable framework (ISA 500, para. A1). The ISA 500 introduces two criteria for the evaluation of audit evidence: *Appropiateness* and *sufficiency*. *Appropiateness* is a measure of the evidence's quality in terms of relevance and reliability. According to the ISA, the nature, source, and circumstances under which information was obtained affect its reliability (ISA 500, para. A31): Information obtained from independent, external sources is assumed to be generally more reliable. The *sufficiency* is a measure of quantity. The quantity of evidence required to reduce the audit risk to an acceptable low level depends on the assessed RMM and the quality of the information. Appropriateness and sufficiency are interconnected (ISA 500, para. A5): The quantity of evidence required to be qualified as sufficient decreases as its quality increases. However, the quantity cannot compensate for poor quality; irrelevant evidence in large quantities remains irrelevant.

As mentioned above, the auditor decides on the timing and extent of substantive procedures based on the previously assessed RMM. The timing of the procedure relates to the point in time in which the procedure is performed, e.g., interim, at the end or after the

financial reporting period (ISA 330, paras. A11-14). The extent refers to the quantity of substantive testing and increases with the RMM (ISA 330, paras. A15): If the assessed RMM is low, and tests of controls reveal effective controls, the auditor may choose to perform only limited substantive procedures. Vice versa, if the RMM is high and controls ineffective, more extensive testing is required to reduce the AR to an acceptably low level (Knechel and Salterio, p. 318). However, some extent of substantive procedures is to be conducted irrespectively of the assessed RMM (ISA 330, para. 18). Naturally, the extent of substantive testing also relates to the sampling of data being used for testing: More extensive testing requires bigger sample sizes.

The ISA allow for different types of sampling, based on either quantitative or qualitative criteria (ISA 500, paras. A52-A55; ISA 530). The alternative to sampling is to perform full population testing (ISA 500, paras. A52-53). However, full population sampling is generally not considered to be cost- and time-efficient (Marten et al., 2020, p. 448).

2.1.3 Emerging Technologies in Auditing

The previous section has introduced some elemental methodological concepts of financial statement audits. Among them, it has introduced the requirement of audits to be time- and cost-efficient, reflected in the ARM, the concept of materiality, and sampling practices. Those efficiency considerations have to be harmonized with the effectiveness of an audit - the expression of an audit opinion based on a reasonable assurance (ibid., p. 309). Balancing these two objectives two is a challenging endeavor. The economics of the market for audit services incentivize efficiency: The market is generally characterized as highly saturated with only limited room for growth (ibid., p. 82) and an intense price competition (ibid., p. 86). On the other side, issuing a wrong audit opinion bears the risk of reputational damage, litigation, or - in rare cases - even criminal prosecution for the auditor (Francis, 2011).

Audit firms utilizing technology to increase both the efficiency and effectiveness of audits is not a new phenomenon: The ubiquity of enterprise resource planning (ERP) systems on the client side accelerated the development of computer-assisted audit techniques and tools (CAATT), which assists auditors with accessing and analyzing their client's data (Alles, 2015; Braun and Davis, 2003). Larger audit firms have invested in developing their own CAATTs, such as PWC's Halo (PWC, 2023) or EY's Helix (EY, 2023). Beyond that, generealised audit software (GAS) (Braun and Davis, 2003) like IDEA (IDEA, 2023) or Audicon (Audicon, 2023) is commercially available to all firms.

Currently, a new set of technologies is perceived as means to further increase the efficiency and effectiveness of auditing - big data, data analytics, AI, and robotic process automation (RPA) (Alles and Gray, 2016; Cao et al., 2015; IAASB, 2020; Kokina and Davenport, 2017; Moffitt et al., 2018; Salijeni et al., 2019). The term 'big data' relates to the dimensions that can be used to describe the nature of data: Volume, variety, and velocity. The volume describes the quantity of the data, variety its format, and velocity its rate of generation (Gandomi and Haider, 2015). 'Big' data assets are therefore characterized as being large in volume, exhibiting heterogeneous formats, or being generated at a high rate. 'Data analytics' refers to analyzing data assets to derive valuable insights from them (ibid.). Big data and AI are complementary: AI, most prominently deep learning, enables the

processing of unstructured data in the form of text-, image-, video-, and audio files through natural language processing (NLP), computer vision (CV) and audio signal processing. Especially deep learning profits from large amounts of such data to train effective models: Modern language models (LM) such as BERT (Devlin et al., 2019) or GPT3 (Brown et al., 2020) have been trained on text corpora encompassing billions (Wikipedia) up to trillions of words (CommonCrawl). In contrast, RPA does not rely on data: It leverages business rules and activity choreographies to mimic the human interaction with user interfaces (Moffitt et al., 2018).

The application of these technologies to increase the efficiency and effectiveness of auditing has drawn substantial attention from the research community and, of course, from the profession itself. In the revised version of the ISA 315 of 2019, the IAASB introduced the possibility of employing "automated tools and techniques" to perform audit procedures. The standard provides a high-level definition of the term and focuses on how such automated tools and techniques may be used in the context of risk assessment. The supporting documentation states that the definition is intentionally broad to accommodate future emerging technologies (IAASB, 2020). The IAASB list data analytics, artificial intelligence, and robotic process automation as examples of current emerging technologies.

While the standard focuses on using these technologies for risk assessment procedures, research has shown their applicability beyond risk assessment. Appelbaum et al. (2018) have conducted an extensive literature review on using data analytics for analytical audit procedures. The authors found that data analytics could be applied through all stages of an audit engagement, including substantive procedures. One of the methods they found applicable throughout the audit process is process mining. This technique allows for discovering processes from the total populations of transactions stored in an ERP system. This is supported by the findings of Werner et al. (2021), who present a framework to embed process mining into the audit process, showing that it can also be used as a substantive analytical procedure. A similar conclusion for RPA is drawn by Moffitt et al. (2018). The authors present a framework for the application of RPA, showing that it can be used to automate analytical procedures, tests of controls, and tests of details.

Generally, substantive audit procedures offer a vast potential for automation. As shown by Abdolmohammadi (1999), substantive testing encompasses many structured and semi-structured tasks. The authors characterize structured tasks as well-defined, with few alternatives. This plays to the strength of RPA, which requires well-defined tasks of low complexity (Moffitt et al., 2018). However, RPA, being based on business rules and choreographies, is not well suited to handle inputs from unstructured data sources. This constitutes a severe limitation, considering that documents such as contracts, invoices, delivery notes, bank statements, or receipts form part of the client's accounting records and, therefore are a central source of evidence for tests of details. Fortunately, AI in the form of DL can be used to process textual data, allowing for an automated *inspection* of document contents (Issa et al., 2016; Kokina and Davenport, 2017; Sun, 2019). Automating the inspection of documents increases the efficiency of TODs by cutting expensive person hours - a DL-enabled system can inspect documents within a fraction of the time an auditor would need. This ability also enables the processing of more significant amounts of documents, thus increasing the test's effectiveness by collecting more audit evidence. Automating document inspection further increases the effectiveness

of the audit as a whole, as it frees personnel capacities to focus on higher value-added tasks.

The vast potential for applying the above emerging technologies in the audit domain bears the question of why their adoption is progressing at such a slow pace. The adoption of technologies such as CAATT and GAS has been studied extensively, both on the individual and audit firm level (Ahmi and Kent, 2012; Curtis and Payne, 2014; Janvrin, Bierstaker, et al., 2009; Janvrin, Lowe, et al., 2008; Li et al., 2018; Pedrosa, Costa, and Aparicio, 2020; Pedrosa, Costa, and Laureano, 2015; Rosli et al., 2012; Siew et al., 2020; Widuri et al., 2016). However, these technologies are to be seen as a "new breed" of technology, as their adoption is considered potentially disruptive to the audit profession (Alles and Gray, 2016; Alles, 2015; Dagilienė and Klovienė, 2019; Haddara et al., 2018; Salijeni et al., 2019). Hence, several empirical studies have explored the factors affecting ADA adoption by audit firms. The factors are mostly related to the organizational characteristics of audit firms, the characteristics of their clients and the auditor-client relationship, and the institutional framework around auditing (Dagilienė and Klovienė, 2019; Eilifsen et al., 2019; Haddara et al., 2018; Salijeni et al., 2019). While the factors of adoption for emerging technologies have been explored by previous research, the process by which audit firms adopt them has been largely neglected. The research on technology adoption in auditing is primarily concerned with variance studies, which model the outcome (adoption or non-adoption) as a binary variable (Markus and Robey, 1988). Hence, they treat the process underlying the adoption as a black box (Langley, 1999; Markus and Robey, 1988). Exploring the process of ADA adoption would help to deepen the understanding of technology adoption at the audit firm level, and could also be leveraged to inform design-oriented research that actively advances the adoption in practice.

Moving beyond theories for understanding, Kogan et al. (2019) call for more design-oriented research on the adoption of ADA. Several research contributions introduce conceptual frameworks on how ADA can be used in auditing (Appelbaum et al., 2018; Moffitt et al., 2018; Sun, 2019; Zhaokai and Moffitt, 2019). Conceptually, the individual contributions focus on either single technologies (Moffitt et al., 2018; Sun, 2019) or individual audit procedures (Appelbaum et al., 2018; Zhaokai and Moffitt, 2019). A more holistic view could enable researchers and practitioners to identify interesting use cases for applying ADA. Furthermore, scientific accounts of ADA-based tools or applications in the audit domain are rare. Werner et al. (2021) demonstrates a process mining tool developed within an audit firm, Tecuci et al. (2020) introduce an ML-based platform for the inspection of contracts. However, neither reflect the design of the resulting tool with respect to the requirements of the audit domain nor propose any design principles that can help guide the development of future solutions.

2.2 Information Extraction from Documents for Tests of Details

As mentioned in section 2.1.2, tests of details are quintessential audit procedures used to lower the detection risk and, consequently, the overall audit risk. The primary source of audit evidence for tests of details are the client's accounting records, which encompass structured (e.g., general and subsidiary ledger) and unstructured data (e.g., invoices, contracts, receipts). By reconciling accounting records, the auditor collects evidence

regarding their internal consistency. Especially by reconciling accounting transactions to *external* documents, which are considered highly reliable in the context of ISA 500, audit evidence of high appropriateness is generated. As explained in the previous section 2.1.3, automating this reconciliation would benefit the efficiency of TODs by cutting person-hours, as well as their effectiveness by collecting more evidence.

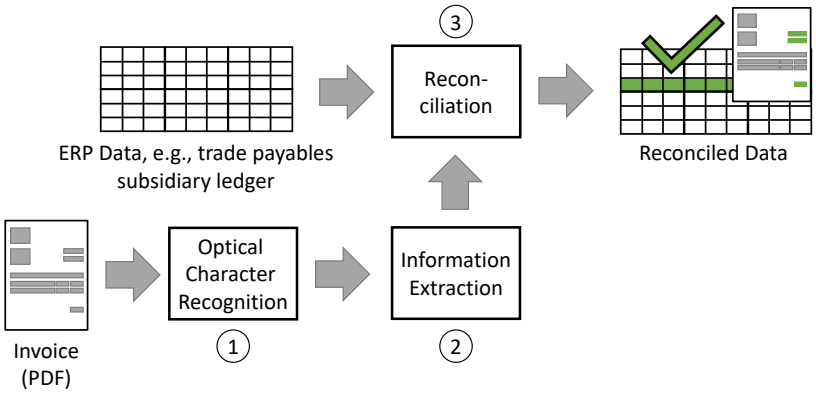

Figure 2.2: Processing steps required to reconcile structured and unstructured sources of accounting records.

The automated substantiation of transactions to their supporting documents requires multiple steps, depicted in Figure 2.2 (Zhaokai and Moffitt, 2019). The supporting documents must be imported from their respective source system as a preliminary step. These could be the client's ERP system, local or remote file systems, or physical copies. A standard file format for documents is the portable document file PDF, which represents the documents in their formatted form. PDFs can hold multiple media formats such as text, images, or even video and audio (Hardy et al., 2017). Therefore, a document formatted as PDF may contain only images (e.g., scanned physical copies of documents) or images containing text (e.g., specifically designed letterheads). Consequently, any further electronic text processing requires optical character recognition (OCR), which recognizes text in images. As pictured in Figure 2.2, applying OCR is the first step (1) in the reconciliation process. OCR is a well-studied technique and is available through many open-source and commercial software offerings. It parses documents into a set of bounding boxes described by their coordinates on the two-dimensional plane along with their text content. Figure 2.6 provides an example of a document along with recognized text boxes. Using the machine-readable text, the document's contents can be extracted in the second step (2). The contents of interest depend of course, on the structured data source to be reconciled against. For instance, an accounts payable subsidiary ledger typically contains detailed information about the client's procurement transactions. Relevant details include

- the monetary amount to be paid,
- the payment due date,
- the supplier's name and (value added) tax identification number,
- the identification number of the document,
- the date of the recording,

- and the date (range) on which the related services have been performed or the products have been delivered.

As a last step (3), the contents extracted from the document are reconciled against the structured data utilizing, e.g., string-distance comparison or more sophisticated entity matching methods (Cohen et al., 2003; Köpcke and Rahm, 2010). The IE step is hereby crucial: Wrongly extracted information can lead to missed reconciliations or, worse, to false reconciliations, which is the reason for focusing our research on the IE step. Some pieces of information may be captured via regular expressions to a high degree of accuracy. For instance, the value-added tax identification numbers in the EU follow a standardized pattern. However, other pieces of information depend heavily on their context. E.g., dates can appear in various contexts on a business document, such as delivery date, payment date, or document date. The challenge hereby is that the semantics of the language used in business documents is complex in the sense that different terms relate to the same fact or vice versa. For instance, the necessary context to identify a document date on an invoice can be given through the terms "document date," "invoice date," and "issue date," among many others. This can be addressed through NN-based LM, which are able to model the semantic (dis-) similarity between terms based on their context.

This section will cover the utilization of NN for IE and highlight the specific challenges of applying IE to visually-rich business documents on the example of invoices.

2.2.1 Deep Neural Networks for Information Extraction

In the scientific discourse around the adoption of AI in auditing, the term of AI is used almost synonymously for ML resp. DL. DL is a subfield of ML, and concerns itself with *deep* NN, i.e., NNs exhibiting several hidden layers. As the name suggests, the research on NN has been inspired by the search for computational models of neurons in the human brain (Han et al., 2011, p. 398).

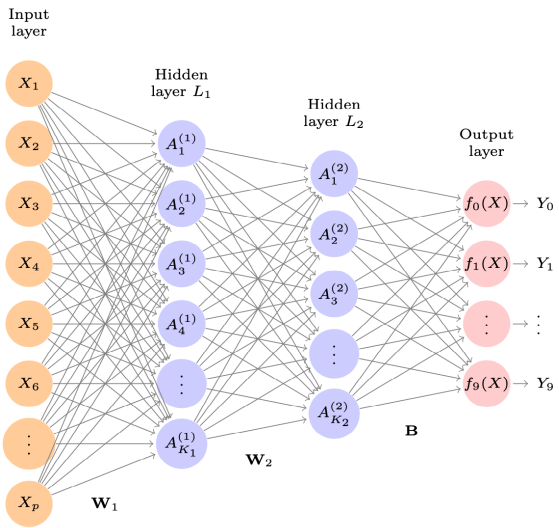

Figure 2.3: Schematic depiction of a NN with two hidden layer predicting a categorical variable with $n = 10$ distinct values (James et al., 2021, p. 409).

NNs take a vector of p variables $X = (X_1, X_2, \ldots, X_p)$ as input, and learn a non-linear function $f(X)$ to predict a output variable Y (James et al., 2021, p. 404). They are composed of multiple connected input- and output units, in which each connection has a weight associated with it (Han et al., 2011, p. 398). In the terminology of NNs, the units are organized in so-called "hidden layers." Figure 2.3 schematically depicts a simple feed-forward NN with two hidden layers L_1 and L_2. The network performs a classification task: It predicts a categorical variable that can take on n distinct values $Y = (Y_1, Y_2, \ldots, Y_n)$. The variables $X_1, \ldots, X_p$ form the "input layer" of the NN. In the forward pass, the inputs are fed into the K_1 units of the first hidden layer L_1, as indicated by the edges denoted with $\mathbf{W}_1$ in Figure 2.3. Each of the K_1 hidden units computes an *activation* from its weighted inputs (which may include a bias that is omitted in this example), using an activation function g. In this example, the weights $w^{(1)}$ applied to the p inputs for a hidden unit k shall be denoted as $w_{kp}^{(1)}$. The activation of unit k is therefore computed as $A_k^{(1)} = g(\sum_{j=1}^{p} w_{kj}^{(1)} X_j)$. The activation function g, which is specified in advance, introduces the non-linearity into the model. Without it, the NN would collapse into a linear model. Most modern NN use the rectified linear unit ($ReLu$) as an activation function, as it is efficiently computed and stored (James et al., 2021, p. 406):

$$g(x) = ReLu(x) = \begin{cases} 0 & \text{if } x < 0 \\ x, & \text{otherwise} \end{cases} \tag{2.2}$$

The activations $A^{(1)}$ become the inputs for the K_2 units of L_2, from which the activations $A^{(2)}$ are computed. The "output layer" of the model in Figure 2.3 computes the responses Z_m for each class m, treating the activations $A^{(2)}$ as input *features*: $Z_m = f_m(X) = \sum_{l=1}^{K_2} \beta_{ml} A_l^{(2)}$, where β_m are the coefficients of each unit in the output layer. The responses can be converted into class probabilities $Pr(Y = m|X)$ by applying the softmax function. The model's final prediction $\hat{Y}$ for the class associated with an input vector is obtained by selecting the class with the highest probability score (ibid., p. 410):

$$\hat{Y} = \arg\max(Pr(Y = m|X)) = \arg\max\left(\frac{e^{Z_m}}{\sum_{l=0}^{m} e^{Z_l}}\right) \tag{2.3}$$

In Figure 2.3, $\mathbf{W}_1$, $\mathbf{W}_2$, and $\mathbf{B}$ denote the entirety of model parameters that are learned from the data. In NNs, this is achieved through the *backpropagation* algorithm. Backpropagation fits the parameters by iteratively minimizing the error between the network's prediction $\hat{Y}$ and the ground truth Y for the data tuples in the training data set. It modifies the parameters in a backward direction throughout the network, starting with the output layer (Han et al., 2011, pp. 400-406). The backpropagation algorithm can be applied to train virtually any combination of input, hidden, and output layers, provided the structure of the model can be represented as a computational graph (Lecun et al., 1998).

A particular strength of NNs often mentioned in the literature is their ability to learn features from "raw" unstructured data (Sun, 2019): Discrete representations of objects, such as images or text, are translated into a continuous, real-valued vector space. The vector representations of the objects are learned such that they yield good results for the task to be solved by the model (Grohe, 2020). This usually requires the model to capture semantic relationships between objects. To deal with the structural heterogeneity of "raw" data representations, the research community has proposed several types of NN model layers. Sequences, such as text, can be modeled using recurrent layers like gated

recurrent units (Cho et al., 2014) or long-term short memory (LSTM) units (Hochreiter and Schmidhuber, 1997). The transformer layers are a newer addition to this family, which are based on the self-attention mechanism (Vaswani et al., 2017). In computer vision, convolutional layers enable the extraction of feature maps through learnable filters (Lecun et al., 1998). Convolutional layers may also be used to model sequences (Gehring et al., 2017). Graph convolutional layers apply the idea of convolution to graph-structured data: They aggregate the activations of objects which are connected in a graph via edges (Kipf and Welling, 2017).

Especially NNs utilizing sequential layers have been used for IE from contracts (Chalkidis and Androutsopoulos, 2017; Elwany et al., 2019; Hu and Su, 2021; Tecuci et al., 2020). However, contracts are characterized by rich, sequential text. Many other types of business documents, such as invoices, receipts, and delivery notes, are sparse in their use of text and exhibit a two-dimensional layout. These characteristics render sequential approaches ineffective.

2.2.2 Information Extraction from Invoices

IE is usually formulated as a classification task. The model classifies the individual text boxes on a document into n predefined classes, which reflect the extracted information entities. The n classes include a *background* class, into which all text is classified that is not considered relevant information. As already pointed out, many business documents exhibit a layout and are rather sparse in their use of text. They are, therefore, often referred to as *visually rich documents* (VRD). To distinguish between the background class and the different information entities, machine learning models may leverage different types of signals from VRDs. For example, Figure 2.4 shows an invoice cut-out containing multiple monetary amounts. While most are equal, only the total amount, highlighted in red, shall be extracted.

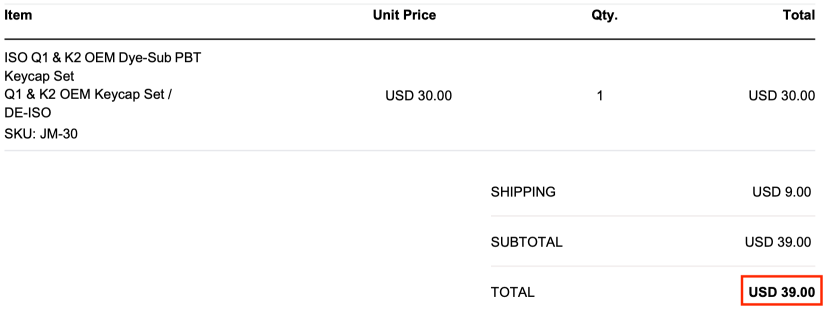

Figure 2.4: Different monetary amounts on an invoice. The text box around the total amount to be extracted is highlighted in red.

The following signals can be leveraged to identify the total amount in the example:
Character string: A strong signal is given through the string of characters pertaining to a text box. For instance, monetary amounts can be expected to be a sequence of numbers with two decimals separated by a point or a comma.
Semantics: Monetary amounts, or dates, can appear in different forms on a VRD, such as gross- and net amounts or issue-and due dates. Here, the context can help to

differentiate them from one another. In the example of 2.4, the monetary amounts may be disambiguated through their context ("Total", "Unit Price", etc.).
Spatial information: The position of a text box on the two-dimensional plane of the document can inform its class membership. E.g., the total amount is expected to be in the lower right area of an invoice (as shown in Figure 2.5).
Visual features: VRD may exhibit visual features that can help to identify information entities. For instance, invoice line items are usually in table-like formats. In the example of Figure 2.4, the horizontal bar may inform the classification of a line item as such, thus allowing for further disambiguation between the line item's total amount and the overall total amount.

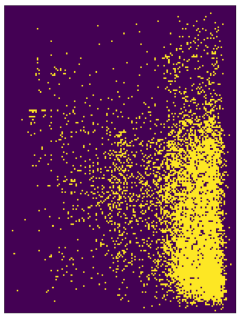

Figure 2.5: The spatial distribution of total amounts across a set of invoices (Katti et al., 2018).

The example in Figure 2.4 further highlights the challenge of capturing contextual dependencies in VRD. The context for the shipping costs, the subtotal and total amounts are found directly left to the amount. In contrast, for the unit price, the discriminating context is situated *above* the amount. The arrangement of visual and textual components on VRDs, the layout, is usually freely decided by the issuing entity. The resulting variety in layouts makes IE from VRD a particularly interesting and challenging endeavor, which continues to draw a lot of attention from the research community.

Different approaches have been proposed to tackle the challenge of representing VRDs such that the different signals can be effectively leveraged for IE. They can be broadly classified into graph- and grid-based approaches. In analogy to the representation of images in CV, grid-based approaches (Denk and Reisswig, 2019; Katti et al., 2018; Zhao et al., 2019) represent the document D as a matrix M characterized by its dimensions (H, W, C): $D \in M_{H \times W \times C}$. H, W are the height resp. width of the document grid. The dimension C denotes the channels of the input matrix, which is used to attach the semantic or string-related input variables pertaining to a grid cell. The corresponding models use convolutional layers to embed this information jointly. Graph-based approaches (Krieger, Drews, Funk, et al., 2021; Liu et al., 2019; Lohani et al., 2019; Yu et al., 2020; Zhang et al., 2020) model the document as a graph $D \in G = (V, E)$, where the nodes V represent the text elements of the VRD, and the edges E represent the relative spatial relationships between them. Both the nodes and the edges may hold input variables. These representations are leveraged through graph convolutional layers to jointly embed the input variables of nodes and edges connected to each other. Figure 2.6 gives an example of a document being represented as a graph and a grid: The graph in the middle panel is obtained via the algorithm presented by Krieger, Drews, Funk, et al. (2021). The right panel depicts a Chargrid (Katti et al., 2018); the colors on the document grid represent

the one-hot-encoded characters. Transformer models can also be broadly classified as graph-based models: Through the self-attention mechanism, they connect all the text elements on D to one another. The transformer models proposed specifically for VRD (Garncarek et al., 2021; Huang, Lv, et al., 2022; Kim et al., 2021; Wang et al., 2022; Xu, Xu, et al., 2022; Xu, Li, et al., 2020) extend the one dimensionals positional embeddings of the original transformer architecture (Vaswani et al., 2017) to four dimensions that represent the coordinates of the text box. They are pretrained on massive corpora of unlabeled VRDs and can be finetuned for IE. The LayoutLM model family (Huang, Lv, et al., 2022; Xu, Xu, et al., 2022; Xu, Li, et al., 2020) further employs a convolutional neural network to embed the visual features of VRDs in addition to the textual and spatial information.

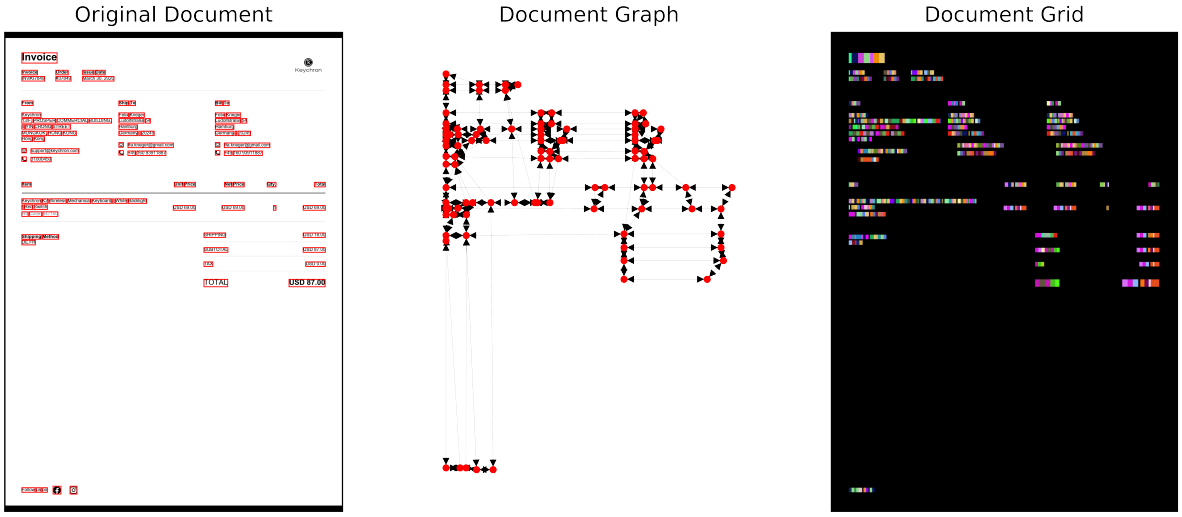

Figure 2.6: Graph and document representations obtained from an OCR-ed invoice.

2.2.3 Evaluation of Models for Information Extraction from Invoices

The body of research on IE from VRD has grown substantially in the last years. This poses the challenge of selecting the most suitable model given the demands of the audit domain. It is further foreseeable that new approaches will continue to be proposed. Hence, it is reasonable to continuously evaluate newly proposed models to keep up with the most recent developments to attain the highest achievable accuracy for IE.

It is generally safe to assume that documents sent out by one company to its clients or other third parties follow a single document type-wise layout or a minimal variation thereof. Testing the client's outgoing documents could be very well addressed using IE systems that employ rules- or template-based processing (Dengel and Bertin, 2002; Esser et al., 2012; Schuster et al., 2013). The advantage of ML-based over rule- or template-based IE for the TOD is the ability to generalize to previously unseen document layouts. This advantage makes ML especially feasible to test the client's incoming documents, i.e., documents the client receives from suppliers or other third parties. However, this requires a large enough variety of layouts in the training data to avoid overfitting. Overfitting occurs when the model too closely resembles the data it was used to train on, hence hurting its

generalization ability to unseen data (Han et al., 2011, p. 330; James et al., 2021, p.22). Therefore, evaluating the models appropriately before using them in an audit is paramount.

Table 2.1: Exemplary confusion matrix for a class m.

	$Y \neq m$	$Y = m$
$\hat{Y} \neq m$	TN_m	FN_m
$\hat{Y} = m$	FP_m	TP_m

An appropriate evaluation of ML models requires a methodology and metrics. In supervised learning settings, such as classification, it is well established to randomly split the data available for model training into two disjoint sets of data: The majority of the data is used for training the model (training set), and a smaller set of data is set aside to test the model (test set) (Han et al., 2011, p. 370). The model is applied to the test set, and the model's predictions $\hat{Y}$ are compared against the ground truth Y. The stochastic influences in the data splitting step can be accounted for through k-fold cross-validation: The data is divided into k approximately equally sized sets, and the training/validation procedure is repeated k times. In each iteration, the k-th set is used as a test set, and the other sets are used to train the model. By stratifying the sampling process to divide the data, it can be ensured that each set exhibits the same statistical properties (ibid., pp. 370-371).

The comparison of the predictions to the ground truth is usually represented in a confusion matrix (see Table 2.1). Given a *positive* class m, all instances where the prediction $\hat{Y}$ matches the ground-truth Y are considered true positives for the respective class, TP_m. Conversely, any instaces wrongly predicted to pertain class m are counted as false positives FP_m. If the model correctly predicts the data instance to be of another class $\hat{Y} \neq m$, they are true negatives TN_m. In analogy, any misclassifications of $Y \neq m$ are considered false negatives FN_m. Higher-level metrics can be derived from the confusion matrix to evaluate a model. Standard metrics in IE are precision, recall, and the F_1 score (Christopher D Manning et al., 2008). Precision is a measure of exactness, i.e., it determines whether all the predictions pertain to the target class. Recall is a measure of completeness and is used to evaluate whether all instances of the target class are classified as such. The class-wise precision and recall are given through

$$Precision_m = \frac{TP_m}{TP_m + FP_m} \tag{2.4}$$

$$Recall_m = \frac{TP_m}{TP_m + FN_m} \tag{2.5}$$

The (class-wise) F_1 score is the harmonic mean of precision and call:

$$F_{1_m} = \frac{2 \times Precision_m \times Recall_m}{Precision_m + Recall_m} \tag{2.6}$$

The F_1 score is especially well suited for classification tasks with a high class imbalance (Han et al., 2011, pp. 367-368). As documents contain a large amount of background text that is not extracted, the F_1 score is frequently used to evaluate IE models for VRD

(Garncarek et al., 2021; Liu et al., 2019; Lohani et al., 2019; Majumder et al., 2020; Xu, Li, et al., 2020; Zhang et al., 2020). As IE models typically extract multiple information entities, the class-wise measures are commonly aggregated into a global F_1 score. The most common methods for averaging found in the literature are micro-averaging (e.g., Lohani et al. (2019)) and macro-averaging (e.g., Majumder et al. (2020)). For micro-averaging, the class-wise $\mathrm{TP}_m \ldots \mathrm{FN}_m$ are first summed into global TP, ..., FN, e.g., $TP = \sum_{m=1}^{n} TP_m$, from which the aggregated F_1 is computed. The macro-averaged F_1 score is the mean of the class-wise measures: $F_1 = (\sum_{m=1}^{n} F_{1_m})/n$. Other than the F_1 score, some studies have employed metrics such as the word-error rate (Denk and Reisswig, 2019; Katti et al., 2018), mean entity precision (Yu et al., 2020) or the average precision (Zhao et al., 2019).

The research on IE from VRDs is contingent on the data available for the training and evaluation of models. For some document types, such as receipts, non-disclosure agreements, and financial statements, open data sets are available (Huang, Chen, et al., 2019; Park et al., 2019; Stanisławek et al., 2021). Other document types lack such data sets, for instance, invoices. This can be attributed to the fact that companies and individuals often perceive the contained information as highly sensitive. Therefore, researchers resort to using proprietary invoice data sets, which exhibit varying characteristics in terms of size and distribution of vendors, resp. layouts (Denk and Reisswig, 2019; Katti et al., 2018; Liu et al., 2019; Lohani et al., 2019; Majumder et al., 2020). The use of proprietary data severely impedes the ability to compare different models with each other from studying the literature, particularly if crucial details on the data set characteristics are underreported. Especially early works on graph-based models for IE (Liu et al., 2019; Lohani et al., 2019) left to wonder whether they would generalize well to data sets exhibiting a considerable variety of layouts. The already limited comparability between models is further exacerbated by the use of different evaluation metrics or varying methods of averaging them across classes.

The effects of the distribution of layouts on the generalization ability of IE models are generally not well understood. Existing studies either resort to data sets containing only very few instances of repeating layouts (Denk and Reisswig, 2019; Katti et al., 2018; Majumder et al., 2020), or do not examine the effects of frequently occurring layouts. In practice, recurring layouts pose a challenge as in most companies, the majority of invoices are received from only a few suppliers (Koch, 2019; Tanner and Richter, 2018).

To audit firms, the effects of layout distributions on IE models are relevant for operationalizing IE. A trustful relationship between auditors and their clients is fundamental to their business, rendering client data highly sensitive. In addition, audit firms are subject to data protection obligations, such as the EU *General Data Protection Regulation*, or the *Gramm Leach Bliley Act* and the *California Consumer Privacy Act* in the United States (Schreyer et al., 2022). Further restrictions are given through the audit-specific professional codes of conduct: "*[...] a member in public practice shall not disclose any confidential client information without the specific consent of the client*" (AICPA, 2014, para. 1.700.001.01). From this perspective, the feasibility of pooling data from different audit engagements to create an extensive and comprehensive set of data to train on is questionable. An obvious solution is leveraging the audit firm's own received invoices for training models. However, the large variety of clients faced by audit firms requires the IE models to generalize well to unseen layouts to automate the TOD effectively.

The heterogeneity of audit clients further especially requires large audit firms to adapt their IE models for a multitude of countries. The resulting diversity of languages and the availability of labeled training data for the respective language should therefore be taken into account when selecting the most appropriate IE models: As formulated in the *"no-free-lunch" theorem for supervised learning* (Wolpert, 2002), no single model is universally better than all models across a multitude of settings. Model selection is, therefore, a non-trivial endeavor in the audit domain. The continuous stream of newly proposed models for IE from VRD further exacerbates the complexity of model selection.

Chapter 3

Contributions

The first chapter of the dissertation motivates using ADA to increase the efficiency and effectiveness of financial statement audits. Chapter 2 provides the background to contextualize the research contributions made in the chapters I through V. In this chapter, I outline the contributions of the respective chapters vis-a-vis the previous research and contextualize them within the design knowledge contribution framework (DKCF) proposed by Hevner et al. (2004). A summary of the individual chapter's contributions, the addressed RQ, and the utilized research methods is given in Table 1.1.

3.1 Understanding the Adoption of ADA in Auditing

In the DKCF, the application of ADA to auditing fits best into the *improvement* category. Improvements are characterized by a high maturity of the application domain but a low maturity of the application. In other terms, the application context is known, but useful solution artifacts do not yet exist. The researcher's goal is consequently to create "[...] *better solutions in the form of more efficient or effective products, processes, technologies, or ideas.*" (ibid.). To this end, the research must draw from a thorough understanding of the problem environment. Forming this understanding was the objective of the dissertation's first part, i.e., the chapters I and II.

Chapter I addresses RQ1. As stated, the research on ADA in auditing was still in its beginnings. Some studies had explored the potential benefits of analyzing (big) data on a conceptual level, such as increasing the efficiency and effectiveness (Byrnes et al., 2015; Stewart, 2015), or even completely changing the audit process (Bumgarner and Vasarhelyi, 2018; Issa et al., 2016). The first literature reviews on the application of ADA to accounting and auditing had already been published (Amani and Fadlalla, 2017; Appelbaum et al., 2018). Other publications had discussed the nature of the data that could be analyzed and how such data was to be evaluated from an audit perspective (Vasarhelyi et al., 2015; Yoon et al., 2015). Yet, an integration of these vital aspects was missing in the literature. To support discussions around the direction for future research, we aimed for a structuring of the literature that allowed for a comprehensive description of ADA use cases. Therefore, we devised a taxonomy that maps the solution space for ADA use cases in auditing. It integrates dimensions pertaining to three broader aspects; the auditing dimensions, the data management dimensions, and the analytics dimensions.

The research presented in chapter II answers the second research question. It proposes a theory that suggests that the activities performed within audit firms to adopt ADA can be broadly arranged into a process. The activities within the process are affected by several contextual factors, which relate to the characteristics of the underlying technology, the characteristics of the audit firm, and the characteristics of the audit domain. Through the activities, the audit firm develops an ADA-based solution, starting with the ideation of a use case and ending with the solution's diffusion into practice. By focusing on single solutions, the theory adopts a different perspective than previous research on ADA adoption, which has seen the underlying technologies as single entities (Dagilienė and Klovienė, 2019; Haddara et al., 2018; Salijeni et al., 2019). It further stresses the active role of the audit firm in achieving adoption: For instance, variance studies have used the *task-technology fit* (Rosli et al., 2012; Widuri et al., 2016) and the closely related *performance expectancy* (Curtis and Payne, 2014; Pedrosa, Costa, and Laureano, 2015; Rosli et al., 2012) as exogenous variables to explain technology adoption. In our theory, this *task-technology fit* is achieved in the *ideation* activity, in which the technology characteristics have to be aligned with the requirements of the audit domain. The complexity of the underlying technology can be absorbed throughout the development of the solution, such that the usability of the solution for the auditor is maximized, affecting the *auditor technology acceptance*. The *auditor technology acceptance* from our theory can be related to the research on technology adoption in auditing at the individual level (Curtis and Payne, 2014; Janvrin, Bierstaker, et al., 2009; Janvrin, Lowe, et al., 2008; Li et al., 2018; Payne and Curtis, 2006; Pedrosa, Costa, and Laureano, 2015). One interesting finding in our study, which has not been addressed in previous research, is that the *organizational context* through which ADA solutions are made available to audit teams is of relevance for the adoption. Further, several studies have referred to the technological competence and resources within the adopting audit firm as factors for technology adoption (Haddara et al., 2018; Hampton and Stratopoulos, 2016; Li et al., 2018; Rosli et al., 2012; Salijeni et al., 2019; Siew et al., 2020; Widuri et al., 2016). This is supported by our results, which indicate that the Big Four take the lead in ADA adoption. The 'deep pockets' of the Big Four enable them to build technological capabilities and develop their own solutions based on ADA. According to our results, the size of the firm also correlates with its ability to attract talent, which is crucial to building technological capabilities. Hence, our results contradict the suggestion of Lowe et al., 2018, which state that mid-sized audit firms have caught up to the Big Four in their use of technology. Apart from the audit firm characteristics, the characteristics of their clients also affect the adoption process. Any solution developed must account for the complexity of clients, which is characterized by the variance of different industries they pertain to, the varying organizational complexity, and their respective IT systems.

3.2 Assessing the Generalization Ability of Neural Networks for Information Extraction

The declared goal of the chapters III through V was to create design knowledge and implement artifacts. The work presented in these chapters is guided by the comprehensive understanding of the problem environment around applying ADA to auditing achieved

through the taxonomy and the process theory. As stated in chapter 1.2, a challenge for the application of IE from VRD in auditing is the large variety of vendor-dependent document layouts due to the significant heterogeneity of audit clients. We, therefore, wanted to better understand and assess the ability of DL-based IE models to generalize when presented with data sets exhibiting many different layouts. To this end, chapter III develops and evaluates a graph-based model for IE from invoices. Chapter IV continues this endeavor by comparing different model types and presents a method for model evaluation that allows narrowing the evaluation results down into in-sample and out-of-sample layouts. The nature of the generated contributions in these chapters can be best classified as methods (including algorithms), and in the case of chapter IV, also as *instantiation*.

Chapter III is the first chapter to employ a design-based research approach and addresses the RQ3. It builds upon previous research on IE from invoices, which employed graph-based NNs (Liu et al., 2019; Lohani et al., 2019), but left doubt as to whether they could appropriately generalize to data sets with a wide variety of different invoice layouts. At the chapter's core lies the development and evaluation of a novel ML model architecture for information extraction (IE) from invoices based on the graph attention (GAT) mechanism (Veličković et al., 2018). As the chapter proposes an algorithm, its contribution lies in proposing a method, according to the DKCF. We used a comprehensive data set of 1,129 English invoices from 277 different vendors for its evaluation. The results show that the developed model is able to extract the information entities *invoice number*, *total amount*, and *invoice date* with satisfying accuracy[1]. Given that the model's designated area of application are TODs, the results need to be discussed in the context of generating audit evidence. FP and FN incur different costs to auditors. FP need to be corrected by an auditor, whereas FN can be offset by further testing until sufficient audit evidence is collected. In that regard, the achieved macro-averaged F_1 score of 0.8753 still leaves room for improvement. It would not enable full automation of the TODs but can undoubtedly lead to efficiency gains, as long as the effort of correcting FP is smaller than the effort to extract the information by hand. With respect to the works that inspired our study (Liu et al., 2019; Lohani et al., 2019), our results did not match their respective results. While our model achieves F_1 scores of 0.8963, 0.8200, and 0.9095 for the abovementioned entities, Liu et al. (2019) and (Lohani et al., 2019) report 0.961, 0.910 and 0.963 and 0.90, 0.99 and 0.95 respectively. We attribute this gap to be rooted mainly in the difference in the size of the data sets used and, of course, the distribution of layouts. This partly inspired the research presented in chapter IV.

The study presented in chapter IV is concerned with RQ4. It benchmarks different types of models for IE from invoices; the computer vision-based models Chargrid (Katti et al., 2018) and BERTgrid (Denk and Reisswig, 2019), the graph-based models proposed by (Lohani et al., 2019) and (Liu et al., 2019), the transformer model LayoutLM (Xu, Li, et al., 2020), and a random forest model. The study addresses a vital characteristic of invoice data sets, which has been largely ignored in previous studies: The distribution of layouts. A large amount of invoices typically received by a company stems from only a few vendors, whereas most vendors only send a few invoices (Koch, 2019; Tanner and Richter, 2018). Hence, the distribution of vendors - and therefore invoice layouts - is characterized by a long tail of infrequent layouts.

[1]Accuracy here refers to the predictive quality of the model, measured using the macro-averaged F_1 score

Our results show that ML models for IE are less accurate on invoice layouts not represented in the training set (out-of-sample layouts). While this can generally be expected, our study shows that this gap in accuracy varies between the models tested; models employing semantic text embeddings from pretrained language models such as word2vec or BERT are more accurate over out-of-sample layouts than models which do not utilize such features. This, however, comes with a cost - the models not employing semantic input features are more accurate over in-sample layouts. Interestingly, in this case, the random forest achieves the second-best macro averaged F_1 score - directly after the transformer model LayoutLM. LayoutLM yields the best results out of all models in our study. It achieves macro-averaged F_1 scores of 0.8761 and 0.7019 for in-sample resp. out-of-sample invoice layouts. For out-of-sample layouts, it also achieves the best results for each extracted information entity. Consequently, it is the most appropriate model choice for the TOD. However, it only achieves a mere F_1 score of 0.2826 for line item-level tax amounts on out-of-sample layouts. It should therefore be evaluated if this entity can be ommitted, as the low accuracy of the model will most likely lead to high amounts of FPs and FNs.

The study's results further show that, depending on the distribution of vendors in the population of invoices available for training, the abovementioned gap in accuracy between in-sample and out-of-sample can go undetected, leading to suboptimal model choices. By disaggregating the evaluation into in-sample and out-of-sample layouts, our results are more transparent than other benchmarking studies or ladders for IE models from VRD (Huang, Chen, et al., 2019; Stanisławek et al., 2021).

To arrive at the presented results, we implemented an end-to-end ML pipeline (Google, 2022; Hapke and Nelson, 2020), which we describe in the chapter. In the DKCF, this corresponds to an artifact's *instantiation*. The pipeline follows a specifically designed evaluation methodology that factors in the distribution of vendors during data splitting and evaluation. The methodology further accounts for differences in the granularity of model predictions by aggregating the predictions. Hence, the study also offers a methodological contribution, according to the DKCF. As the topic of layout distribution has been under-addressed in the research on IE from layout-rich documents, the paper calls both researchers and practitioners to attention when training and evaluating their models.

3.3 Designing Information Extraction Pipelines for Audit Tools

By addressing RQ5, chapter V ties together the chapters II, III and IV. Chapter V offers two types of contributions. The first type is the *instantiation* of an artifact, its thorough description, and its embedding in the audit context. The chapter recounts the genesis of the research pipeline from chapter IV, which started with developing the model presented in chapter III. Embedded within its application context - both in terms of the TODs as well as the tool under development in the audit firm - the pipeline is further developed into fully automated model training and document scoring pipelines, which serve as the IE backend for the audit tool developed to automate TODs. In the logic of the process theory, this audit tool is an AI-based solution. While scientific accounts

of ADA-based audit tools are generally rare (Tecuci et al., 2020; Werner et al., 2021), the combination of the level of detail provided on the pipelines' implementation and the contextualization within the audit domain is unprecedented in the research on ADA adoption in auditing. It also addresses a timely issue with the research on applied ML beyond auditing: As researchers strive for better results on the ever-same benchmarking data sets, more and more models are being proposed. At the same time, the context of their real-world application is dismissed. Through the artifact's design, newly proposed models from the growing stream of research on IE from layout-rich documents can be quickly integrated and evaluated. Following the idea of "no free lunch"-theorem for supervised learning (Wolpert, 2002), the automated training and evaluation accounts for the multitude of application contexts of the TOD in auditing in terms of language and extracted key items, as models can be easily evaluated in the case of changing contexts.

The second type of contribution are *design principles*. The study adopts the ADR methodology (Sein et al., 2011), which is a previously unused methodology to study ADA adoption in auditing. In the reflection and learning step within the ADR, we reflected the design choices that led to the artifact's emergence against the process theory from chapter II. The design choices which are most relevant to the artifact's compliance with the demands of the auditing domain are elevated to design principles for IE pipelines in audit tools. Through the artifact's construction within the firm, the firm's adoption of ADA is *actively* supported. At the same time, design knowledge is generated that is applicable beyond the case of the individual firm.

Chapter 4

Limitations

The previous chapter summarized each study's core results and highlighted their respective contribution to the research fields. However, it has to be acknowledged that each study exhibits aspects that limit the generalization of its results.

The taxonomy in chapter I was developed based on ADA use cases published in the scientific literature. Consequently, it may be blind to use cases that have only been identified in practice and have not been published in academic outlets. Taxonomies are theories for analysis, useful for understanding emerging phenomena. As such, they face the challenge of mapping a moving target, due to which optimal solutions are not realistically achievable (Hevner et al., 2004; Nickerson et al., 2013). This also applies here: Since the taxonomy's creation, AI has seen tremendous development in the form of, e.g. generative and conversational AI. These could enable a range of novel use cases that the taxonomy does not account for in its current state.

The limitations of the theory depicted in chapter II are primarily rooted in the sampling of interview partners. Many interviewees hold positions in the upper management of their respective firms and are involved with digitalizing the audit. Consequently, they may hold overly optimistic views toward the benefits of technologies and their importance for audit practice. However, we accounted for this bias by including interviewees involved with the regular audit fieldwork. Another source for bias in our data are the interviewee's respective countries of employment. Most of them live and work in Germany, so the respective results might generalize inadequately to the audit digitalization in other countries. However, we expected this effect to be relatively weak, as most interviewees work for large multinational audit firms, and Germany has global economic ties. Furthermore, the German institutional environment around auditing is highly internationalized, e.g., through the adaption of the European audit regulations to the regulation in the United States. However, the role of external audits in the German two-tier corporate governance setting might lead to a higher importance of delivering additional insights to the clients as compared to other countries.

The studies in the chapters III and IV are both limited by the use of their respective data sets. As no large sets of annotated invoices were openly available, we were restricted to the experimental data sets supplied by the audit firm. Compared with the data sets used in other studies (Denk and Reisswig, 2019; Katti et al., 2018; Majumder et al., 2020), ours were relatively small. Since DL models benefit from large quantities of data, the

relatively small data sets might affect the accuracy of the model developed in chapter III and models benchmarked in chapter IV. However, we performed extensive hyperparameter searches in both studies to ensure the models' optimal fit to the data set.

Both studies evaluate the models very strictly on the text box level. High F_1 were only achieved if a model recognized all instances of an information item on an invoice. In practice, some items may appear multiple times. For instance, total amounts may appear in the letter head as well as in the body of an invoice. The results could therefore understate the models' ability to extract information. Also, the evaluation does not consider any additional rules-based processing to create composite results from the classified text boxes or to enhance the classification results.

In addition, most models benchmarked in chapter IV had to be implemented from scratch due to a lack of publicly available implementations. Our implementations may not be accurate in all aspects of the respective authors' original implementation. However, we provided extensive details of their implementations.

Chapter V provides the implementation details for the ML pipelines used in the TOD tool. One important limitation of the presented artifact is its implementation using the Azure Machine Learning platform. The implementation employs classes and methods from the Azure Machine Learning software development kit, so it may not be directly transferable to other cloud platforms. However, the proposed design principles are universally applicable, independent of the underlying platform or programming language. Another limitation of the results in chapter V is the narrow focus on IE. The artifact's design does not directly consider interdependencies with up- or downstream components of the TOD application, e.g., the OCR step or entity matching. Concerning technology adoption in auditing, we also do not consider other relevant aspects for successful adoption, such as the application's user interface, its organizational embedding, user training, and the use of trusted or explainable AI methods to ensure the application's trustworthiness for end users.

Chapter 5

Future Research

The body of work presented here does not only aim to make contributions to the research field but also to inspire further research.

Following up on its current state, the taxonomy should be further developed. This could include its adaption to use cases outside of only academic publications. But also, within the scientific literature, it may be updated to more recent developments in AI, NLP, and their application to auditing. For instance, it may include generational tasks, such as text summarization. Conceptually, the dependencies between the characteristics and dimensions should be explored. Concerning data management, further inquiries could explore the topic of data governance, i.e., whether the respective data is proprietary or public domain. Use cases based on the utilization of data which are public domain might be easier implemented.

The process theory describes relationships between the contextual factors and the process steps. However, it does not further specify the nature of these relationships. Further qualitative research could explore their characteristics and uncover through which mechanisms the factors affect the steps. Another interesting path would be to deepen the understanding of the effects of the audit firm's ownership structure on the adoption of innovations, as the audit equity partner's firms are directly participating in the firm's financial results - both positively and negatively. As innovation, and more so disruptive innovation, requires both substantial investments and a positive disposition toward risk, future research might explore relationship between audit firms' ownership structure on audit firm risk disposition and capital dedicated toward innovation. We also see potential for quantitative research to follow up on our study: As the study's research approach is oriented towards hypothesis generation, future research could apply Occam's razor through a quantitative-reductionist approach.

As chapters III and IV were affected by the size of the data sets, it would be interesting to perform these studies with larger data sets, especially using invoices in other languages, possibly even performing cross-language experiments. A big challenge for scaling IE from invoices to different languages and country-wise characteristics is the acquisition of *labeled* training data. Especially with respect to the results of chapter IV, it would be interesting to compare the models' accuracy for changing training data set sizes. In this context, further research could explore the applicability of methods to reduce the amount of labeled training data for IE from invoices within the audit domain, such as

active or semi-supervised learning. The results of chapter IV imply that designers of applications using IE have to consider a trade-off when composing the training data set: Attain a greater generalization to address long-tail vendors or to attain a high accuracy for highly recurring vendors. Here, a sensitivity analysis could be performed that explores how many examples of each invoice layout are required for the model to perform highly accurate IE. The results could be leveraged to fine-tune models for certain vendors efficiently.

As previously stated, the design-oriented research in this dissertation focuses very narrowly on the IE task within the TOD. Future research could broaden this perspective. For instance, it could assess whether there are better IE evaluation metrics to optimize for than the F_1 score: For the TOD, false positives may lead to wrongfully reconciled information, whereas false negatives could be compensated for by analyzing more invoices. As stated above, rules-based processing could be employed to logically verify the classification results, thus enhancing the classification results. E.g., a rule could verify whether the amount recognized as the total gross amount is the highest amount on the invoice. Further, parsing invoice contents into standardized formats could facilitate downstream entity matching. But also within the IE module within the TOD application, further ML-based capabilities are required, especially for document management. Language and document classification can help identify the right model for an incoming document, and layout clustering would support the data splitting and model evaluation methods proposed in chapter IV even without vendor metadata. Future research could also elaborate on the operationalization of ML models. As pointed out, the protection of client data hampers the ability of audit firms to pool data from different engagements. This could be addressed through federated learning, a decentralized training paradigm specifically designed to protect confidential data.

To summarize, the works in this dissertation lay out many threads that can be continued in order to advance the adoption of ADA in auditing.

Chapter 6

Conclusion

Frey and Osborne (2017) argue that auditors are amongst the professions most susceptible to computerization, invoking an image in which the automation of audit work is inevitable. The results of this dissertation suggest that quite the opposite is true. The conduct of an audit is subject to professional standards, codes of conduct, and data protection obligations, placing auditors who deviate therefrom under scrutiny from regulators, professional bodies, and peer reviewers. Therefore, any application of ADA must comply with the regulatory framework around auditing. The taxonomy from chapter I highlights the complex nature of applying ADA to auditing, in which many different perspectives - the audit process, data management, and analytical methods - must be considered. In the taxonomy, the literal and metaphorical bottom line is the "improvement potential," i.e., efficiency and effectiveness gains for audit firms if they manage to employ ADA accordingly.

Audit firms engage with ADA in the pursuit of their own benefit. Chapter II stresses the active role audit firms have to take to realize these benefits: Adoption is realized on the use case level by developing solutions or sourcing them from third parties. Either way, resources - time and money - must be allocated towards this effort. In the process theory, the provision of resources represents a phase gate: If the required resources are not provided, naturally, no further action can be taken. In making this decision, the management has to consider trade-offs. Given that auditing is a professional service, the primary resource of audit firms are people. This leads - in addition to direct costs - to opportunity costs incurred by the development of ADA solution: Any personnel devoted to solution development cannot generate billable hours, which directly affects the cash flow of the firm's proprietors. As most audit firms are organized in partnerships, their ownership and management are in the same hands. Therefore, the investment of resources into solution development underlies potential conflicts of interest. The conclusion to be drawn here is the following: For ADA to become the transformative force observed in other industries, audit firms must strategically prioritize the development of ADA solutions and ensure long-term commitment to this goal.

Apart from the complexity arising from the institutional environment around auditing, the development of solutions needs to cope with the complexities of the underlying technology. Especially the field of AI has seen leap developments in the past years. An unceasingly flowing stream of research has proposed new or improved neural network architectures and training paradigms. This phenomenon also applies to IE from VRD, in which we have partaken in chapter III by introducing a novel graph neural network-based model

architecture. Large international audit firms face an overwhelming heterogeneity of clients. Hence, the VRDs of these clients can exhibit different languages and country-specific characteristics, resulting in a multitude of different contextual settings for IE models. The theorem developed by Wolpert (2002) suggests that the suitability of IE models should be assessed for each setting. Our research presented in chapters IV and V therefore proposes a structured approach to benchmarking models in the form of an ML pipeline. The pipeline has been designed to work as the IE engine for a tool automating TODs. The results from chapter IV further suggest that the evaluation of IE models should pay special attention to the distribution of VRD layouts, for which we propose a methodology. For the application context given through the experimental setting of chapter IV, the transformer model LayoutLM would be the model of choice, with a macro-averaged F_1 score of 0.7019. This score is too low to perform full population testing of the client's document. If all invoices held by a company were tested, the model would potentially yield hundreds of FP and FN. Nevertheless, it allows for generating large amounts of audit evidence in the fracture of the time a human would need. To apply the principle of continuous evaluation to a broader class of problems, chapter V proposes design principles for IE pipelines in auditing.

The developed pipelines and implemented models now form the IE backbone of the TOD tool. The pipelines are undergoing continued development to abstract the code base further and enhance stability. The tool itself is currently undergoing integration testing and is awaiting to go into piloting.

In summary, the dissertation opens the black box of ADA adoption in audit firms. It actively advances the adoption of AI through a practical artifact and the generation of design knowledge. The work offers several contributions to the research on ADA adoption in auditing and the application of AI to IE from invoices. It also highlights open questions that will hopefully be answered by future research to further advance the understanding of ADA and AI adoption in this complex, highly regulated application domain.

Bibliography

Abdolmohammadi, Mohammad J (1999). "A Comprehensive Taxonomy of Audit Task Structure, Professional Rank and Decision Aids for Behavioral Research". In: *Behavioral Research in Accounting* 11. Publisher: American Accounting Association, pp. 51–92.

Accountancy Europe, Fédération des Experts Comptables Européens (2015). *Overview of ISA Adoption in the European Union*. URL: https://www.accountancyeurope.eu/wp-content/uploads/MA_ISA_in_Europe_overview_150908_update.pdf.

Ahmi, Aidi and Simon Kent (2012). "The utilisation of generalized audit software (GAS) by external auditors". In: *Managerial Auditing Journal* 28.2, pp. 88–113. ISSN: 02686902. DOI: 10.1108/02686901311284522.

AICPA (2014). *AICPA Code of Professional Conduct*. URL: https://us.aicpa.org/content/dam/aicpa/research/standards/codeofconduct/downloadabledocuments/2014-december-15-content-asof-2020-June-20-code-of-conduct.pdf (visited on 03/08/2023).

Alles, Michael and Glen L. Gray (2016). "Incorporating big data in audits: Identifying inhibitors and a research agenda to address those inhibitors". In: *International Journal of Accounting Information Systems* 22, pp. 44–59. ISSN: 1467-0895. DOI: https://doi.org/10.1016/j.accinf.2016.07.004. URL: https://www.sciencedirect.com/science/article/pii/S1467089516300811.

Alles, Michael G. (2015). "Drivers of the Use and Facilitators and Obstacles of the Evolution of Big Data by the Audit Profession". In: *Accounting Horizons* 29.2. _eprint: https://meridian.allenpress.com/accounting-horizons/article-pdf/29/2/439/1592722/acch-51067.pdf, pp. 439–449. ISSN: 0888-7993. DOI: 10.2308/acch-51067. URL: https://doi.org/10.2308/acch-51067.

Amani, Farzaneh A. and Adam M. Fadlalla (2017). "Data mining applications in accounting: A review of the literature and organizing framework". In: *International Journal of Accounting Information Systems* 24, pp. 32–58. ISSN: 14670895. DOI: 10.1016/j.accinf.2016.12.004. URL: http://dx.doi.org/10.1016/j.accinf.2016.12.004.

Appelbaum, Deniz A. et al. (2018). "Analytical procedures in external auditing: A comprehensive literature survey and framework for external audit analytics". In: *Journal of Accounting Literature* 40, pp. 83–101. ISSN: 07374607. DOI: 10.1016/j.acclit.2018.01.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S0737460716300611 (visited on 08/23/2021).

Audicon (2023). *Audicon Software*. URL: https://audicon.net/software/ (visited on 02/22/2023).

Braun, Robert L. and Harold E. Davis (2003). "Computer-assisted audit tools and techniques: Analysis and perspectives". In: *Managerial Auditing Journal* 18.9, pp. 725–731. ISSN: 02686902. DOI: 10.1108/02686900310500488.

Brown, Tom et al. (2020). "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran

Associates, Inc., pp. 1877–1901. URL: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Bumgarner, Nancy and Miklos A. Vasarhelyi (2018). "Continuous Auditing—A New View". In: *Continuous Auditing*. Ed. by David Y. Chan et al. Emerald Publishing Limited, pp. 7–51. ISBN: 978-1-78743-413-4 978-1-78743-414-1. DOI: 10.1108/978-1-78743-413-420181002. URL: https://doi.org/10.1108/978-1-78743-413-420181002 (visited on 09/16/2022).

Byrnes, Paul et al. (2015). "Reimagining Auditing in a Wired World". In: *Audit Analytics and Continuous Audit: Looking Toward the Future*. New York: American Institute of Certified Public Accountants, Inc. ISBN: 978-1-943546-08-4. URL: https://us.aicpa.org/content/dam/aicpa/interestareas/frc/assuranceadvisoryservices/downloadabledocuments/auditanalytics_lookingtowardfuture.pdf (visited on 09/16/2022).

Cao, Min et al. (2015). "Big data analytics in financial statement audits". In: *Accounting Horizons* 29.2, pp. 423–429. ISSN: 15587975. DOI: 10.2308/acch-51068.

Chalkidis, Ilias and Ion Androutsopoulos (2017). "A Deep Learning Approach to Contract Element Extraction". In: *Legal Knowledge and Information Systems*, p. 10.

Chan, David Y. et al. (2018). "New Perspective: Data Analytics as a Precursor to Audit Automation". In: *Continuous Auditing*, pp. 315–322. DOI: 10.1108/978-1-78743-413-420181016.

Chen, Hsinchun et al. (2012). "Business Intelligence And Analytics: From Big Data to Big Impact". In: *MIS Quarterly* 36.4, pp. 1165–1188.

Cho, Kyunghyun et al. (2014). "Learning phrase representations using RNN encoder-decoder for statistical machine translation". In: *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 1724–1734. DOI: 10.3115/v1/d14-1179. arXiv: 1406.1078.

Christopher D Manning et al. (2008). *Introduction to Information Retrieval*. New York, United States of America: Cambridge University Press Cambridge. 440 pp. ISBN: 978 0 521 86571 5.

Cohen, William W. et al. (2003). "A Comparison of String Distance Metrics for Name-Matching Tasks". In: *Proceedings of the 2003 International Conference on Information Integration on the Web*. IIWEB'03. event-place: Acapulco, Mexico. AAAI Press, pp. 73–78.

Corbin, Juliet M. and Anselm Strauss (1990). "Grounded theory research: Procedures, canons, and evaluative criteria". In: *Qualitative Sociology* 13.1, pp. 3–21. ISSN: 01620436. DOI: 10.1007/BF00988593.

Cukier, Kenneth (2010). "Data, data everywhere: A special report on manag- ing information". In: *The Economist*. ISSN: 0013-0613. URL: https://www.economist.com/special-report/2010/02/27/data-data-everywhere (visited on 01/27/2023).

Curtis, Mary B. and Elizabeth A. Payne (2014). "Modeling voluntary CAAT utilization decisions in auditing". In: *Managerial Auditing Journal* 29.4, pp. 304–326. ISSN: 02686902. DOI: 10.1108/MAJ-07-2013-0903.

Dagilienė, Lina and Lina Klovienė (2019). "Motivation to use big data and big data analytics in external auditing". In: *Managerial Auditing Journal* 34.7, pp. 750–782. ISSN: 02686902. DOI: 10.1108/MAJ-01-2018-1773.

Dengel, Andreas and Klein Bertin (2002). "smartFIX: A Requirements-Driven System for Document Analysis and Understanding". In: *Lecture Notes in Computer Science* 2423.August 2002, pp. 272–282. DOI: 10.1007/3-540-45869-7. URL: http://www.

springerlink.com/index/C4WW94M0BTJQJ4L6.pdf%7B%5C%%7D5Cnhttp://www.springerlink.com/index/10.1007/3-540-45869-7.

Denk, Timo I. and Christian Reisswig (2019). "BERTgrid: Contextualized Embedding for 2D Document Representation and Understanding". In: *arXiv:1909.04948 [cs]*. arXiv: 1909.04948. URL: http://arxiv.org/abs/1909.04948 (visited on 02/16/2021).

Devlin, Jacob et al. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://www.aclweb.org/anthology/N19-1423.

Eilifsen, Aasmund et al. (2019). "An Exploratory Study into the Use of Audit Data Analytics on Audit Engagements". In: *SSRN Electronic Journal* 1, pp. 1–18. ISSN: 1556-5068. DOI: 10.2139/ssrn.3458485. URL: https://www.ssrn.com/abstract=3458485.

Elwany, Emad et al. (2019). *BERT Goes to Law School: Quantifying the Competitive Advantage of Access to Large Legal Corpora in Contract Understanding*. _eprint: 1911.00473.

Esser, Daniel et al. (2012). "Automatic indexing of scanned documents: a layout-based approach". In: *Document Recognition and Retrieval XIX* 8297.May 2014, 82970H. ISSN: 0277786X. DOI: 10.1117/12.908542.

EU, European Parliament (2002). *Regulation (EC) No 1606/2002 of the European Parliament and of the Council of 19 July 2002 on the application of international accounting standards*. URL: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex%3A32002R1606 (visited on 02/09/2023).

— (2014). *Regulation (EU) No 537/2014 of the European Parliament and of the Council of 16 April 2014 on Specific Requirements Regarding Statutory Audit of Public-Interest Entities and Repealing Commission Decision 2005/909/EC*. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32014R0537.

EY (2023). *EY Helix*. URL: https://www.ey.com/en_gl/audit/technology/helix (visited on 02/22/2023).

Francis, Jere R. (2011). "A Framework for Understanding and Researching Audit Quality". In: *Auditing: A Journal of Practice & Theory* 30.2, pp. 125–152. ISSN: 0278-0380, 1558-7991. DOI: 10.2308/ajpt-50006. URL: https://meridian.allenpress.com/ajpt/article/30/2/125/54555/A-Framework-for-Understanding-and-Researching (visited on 02/21/2023).

Frey, Carl Benedikt and Michael A. Osborne (2017). "The future of employment: How susceptible are jobs to computerisation?" In: *Technological Forecasting and Social Change* 114, pp. 254–280. ISSN: 00401625. DOI: 10.1016/j.techfore.2016.08.019. URL: https://linkinghub.elsevier.com/retrieve/pii/S0040162516302244 (visited on 05/04/2021).

Gandomi, Amir and Murtaza Haider (2015). "Beyond the hype: Big data concepts, methods, and analytics". In: *International Journal of Information Management* 35.2, pp. 137–144. ISSN: 0268-4012. DOI: https://doi.org/10.1016/j.ijinfomgt.2014.10.007. URL: https://www.sciencedirect.com/science/article/pii/S0268401214001066.

Garncarek, Lukasz et al. (2021). "LAMBERT: Layout-Aware Language Modeling for Information Extraction". In: *Document Analysis and Recognition – ICDAR 2021*. Ed. by Josep Lladós et al. Cham: Springer International Publishing, pp. 532–547. ISBN: 978-3-030-86549-8.

Gehring, Jonas et al. (2017). "Convolutional Sequence to Sequence Learning". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 1243–1252. URL: https://proceedings.mlr.press/v70/gehring17a.html.

Google (2022). *MLOps: Continuous delivery and automation pipelines in machine learning*. URL: https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning?hl=de (visited on 11/08/2022).

Gregor, Shirley (2006). "The Nature of Theory in Information Systems". In: *MIS Quarterly* 30.3. Publisher: Management Information Systems Research Center, University of Minnesota, pp. 611–642. DOI: https://doi.org/10.2307/25148742. URL: http://www.jstor.org/stable/25148742 (visited on 01/28/2023).

Grohe, Martin (2020). "word2vec, node2vec, graph2vec, X2vec: Towards a Theory of Vector Embeddings of Structured Data". In: *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*. SIGMOD/PODS '20: International Conference on Management of Data. Portland OR USA: ACM, pp. 1–16. ISBN: 978-1-4503-7108-7. DOI: 10.1145/3375395.3387641. URL: https://dl.acm.org/doi/10.1145/3375395.3387641 (visited on 03/01/2023).

Haddara, Moutaz et al. (2018). "Applications of Big Data Analytics in Financial Auditing-A Study on The Big Four". In: *Twenty-fourth Americas Conference on Information Systems, New Orleans* 2018, pp. 1–10. DOI: 10.1201/b18737-189.

Hampton, Clark and Theophanis C. Stratopoulos (2016). "Audit Data Analytics Use: An Exploratory Analysis". In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.2877358.

Han, Jiawei et al. (2011). *Data Mining: Concepts and Techniques*. 3rd. The Morgan Kaufmann Series in Data Management Systems. Waltham, USA: Elsevier, Morgan Kauffman. ISBN: 978-0-12-381479-1.

Hapke, Hannes Max and Catherine Nelson (2020). *Building machine learning pipelines: automating model life cycles with TensorFlow*. First edition. OCLC: on1138611607. Sebastopol, California: O'Reilly Media, Inc. 337 pp. ISBN: 978-1-4920-5319-4.

Hardy, M. et al. (2017). *The application/pdf Media Type*. RFC8118. RFC Editor, RFC8118. DOI: 10.17487/RFC8118. URL: https://www.rfc-editor.org/info/rfc8118 (visited on 02/27/2023).

Healy, Paul M and Krishna G Palepu (2001). "Information asymmetry, corporate disclosure, and the capital markets: A review of the empirical disclosure literature". In: *Journal of Accounting and Economics* 31.1, pp. 405–440. ISSN: 0165-4101. DOI: 10.1016/S0165-4101(01)00018-0. URL: https://www.sciencedirect.com/science/article/pii/S0165410101000180 (visited on 02/08/2023).

Hevner, Alan R. et al. (2004). "Design Science in Information Systems Research". In: *MIS Q.* 28.1. Place: USA Publisher: Society for Information Management and The Management Information Systems Research Center, pp. 75–105. ISSN: 0276-7783.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long Short-Term Memory". In: *Neural Computation* 9.8, pp. 1735–1780. ISSN: 0899-7667, 1530-888X. DOI: 10.1162/neco.1997.9.8.1735. URL: https://direct.mit.edu/neco/article/9/8/1735-1780/6109 (visited on 02/28/2023).

Hu, Xiang and Wenwei Su (2021). "Information Extraction from Contract Based on BERT-BiLSTM-CRF". In: *Advancements in Mechatronics and Intelligent Robotics*. Ed. by Zhengtao Yu et al. Singapore: Springer Singapore, pp. 109–115. ISBN: 978-981-16-1843-7.

Huang, Yupan, Tengchao Lv, et al. (2022). *LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking*. DOI: 10.48550/ARXIV.2204.08387. URL: https://arxiv.org/abs/2204.08387.

Huang, Zheng, Kai Chen, et al. (2019). "ICDAR2019 Competition on Scanned Receipt OCR and Information Extraction". In: *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1516–1520. DOI: 10.1109/ICDAR.2019.00244.

IAASB (2020). *Non-Authoritative Support Material Related to Technology: Frequently Asked Questions (FAQ)—The Use of Automated Tools and Techniques when Identifying and Assessing Risks of Material Misstatement in Accordance with ISA 315 (Revised 2019)*. URL: https://www.ifac.org/system/files/publications/files/IAASB-Technology-FAQ-Automated-Tools-Techniques.pdf.

IDEA (2023). *IDEA Products*. URL: https://idea.caseware.com/products/idea (visited on 02/22/2023).

IFAC, International Federation of Accountants (2016). *United States of America: Legal and Regulatory Environment, Adoption of International Standards*. IFAC. URL: https://www.ifac.org/about-ifac/membership/profile/united-states-america (visited on 02/11/2023).

ISA 200, IAASB (2009). *International Standard On Auditing 200 Overall Objective of the Independent Auditor, and the Conduct of an Audit in Accordance with International Standards on Auditing*. URL: https://www.ifac.org/system/files/meetings/files/3393.pdf.

ISA 315, IAASB (2019). *International Standard On Auditing 315 (Revised 2019) Identifying and Assessing the Risks of Material Misstatement*. URL: https://www.ifac.org/system/files/downloads/a016-2010-iaasb-handbook-isa-300.pdf.

ISA 330, IAASB (2009). *International Standard On Auditing 330 The Auditor's Responses To Assessed Risks*. URL: https://www.ifac.org/system/files/downloads/a019-2010-iaasb-handbook-isa-330.pdf.

ISA 500, IAASB (2009). *International Standard On Auditing 500 Audit Evidence*. URL: https://www.ifac.org/system/files/downloads/a022-2010-iaasb-handbook-isa-500.pdf (visited on 09/16/2022).

ISA 530, IAASB (2009). *International Standard On Auditing 530 Audit Sampling*. URL: https://www.ifac.org/system/files/downloads/a027-2010-iaasb-handbook-isa-530.pdf (visited on 02/20/2023).

ISA 700, IAASB (2016). *International Standard On Auditing 700 Forming an Opinion and Reporting on Financial Statements*. URL: https://www.ifac.org/system/files/publications/files/ISA-700-Revised_8.pdf.

Issa, Hussein et al. (2016). "Research Ideas for Artificial Intelligence in Auditing: The Formalization of Audit and Workforce Supplementation". In: *Journal of Emerging Technologies in Accounting* 13.2, pp. 1–20. ISSN: 1558-7940, 1554-1908. DOI: 10.2308/jeta-10511. URL: https://meridian.allenpress.com/jeta/article/13/2/1/115980/Research-Ideas-for-Artificial-Intelligence-in (visited on 10/05/2021).

James, Gareth et al. (2021). "Statistical Learning". In: *An Introduction to Statistical Learning*. Series Title: Springer Texts in Statistics. New York, NY: Springer US, pp. 15–57. ISBN: 978-1-07-161417-4 978-1-07-161418-1. DOI: 10.1007/978-1-0716-1418-1_2. URL: https://link.springer.com/10.1007/978-1-0716-1418-1_2 (visited on 02/28/2023).

Janvrin, Diane, James Bierstaker, et al. (2009). "An Investigation of Factors Influencing the Use of Computer-Related Audit Procedures". In: *Journal of Information Systems*

23.1, pp. 97–118. ISSN: 0888-7985. DOI: 10.2308/jis.2009.23.1.97. URL: https://meridian.allenpress.com/jis/article/23/1/97/75367/An-Investigation-of-Factors-Influencing-the-Use-of.

Janvrin, Diane, D. Jordan Lowe, et al. (2008). "Auditor Acceptance of Computer-Assisted Audit Techniques". In: *American Accounting Association Auditing Mid Year Meeting AAA* April, pp. 1–26.

Jensen, Michael C. and William H. Meckling (1976). "Theory of the firm: Managerial behavior, agency costs and ownership structure". In: *Journal of Financial Economics* 3.4, pp. 305–360. ISSN: 0304-405X. DOI: https://doi.org/10.1016/0304-405X(76)90026-X. URL: https://www.sciencedirect.com/science/article/pii/0304405X7690026X.

Katti, Anoop Raveendra et al. (2018). "Chargrid: Towards Understanding 2D Documents". In: *arXiv:1809.08799 [cs]*. arXiv: 1809.08799. URL: http://arxiv.org/abs/1809.08799 (visited on 02/16/2021).

Kim, Geewook et al. (2021). *OCR-free Document Understanding Transformer*. DOI: 10.48550/ARXIV.2111.15664. URL: https://arxiv.org/abs/2111.15664.

Kipf, Thomas N. and Max Welling (2017). "Semi-supervised classification with graph convolutional networks". In: *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*. _eprint: 1609.02907, pp. 1–14.

Knechel, W. Robert and Steven E. Salterio (2016). *Auditing : Assurance and Risk*. London, UNITED KINGDOM: Taylor & Francis Group. ISBN: 978-1-315-53172-4. URL: http://ebookcentral.proquest.com/lib/leuphana/detail.action?docID=4709862.

Koch, Bruno (2019). *The E-Invoicing Journey 2019-2025*. URL: https://www.billentis.com/The_einvoicing_journey_2019-2025.pdf (visited on 03/25/2022).

Kogan, Alexander et al. (2019). "Audit data analytics research—an application of design science methodology". In: *Accounting Horizons* 33.3, pp. 69–73. ISSN: 15587975. DOI: 10.2308/acch-52459.

Kokina, Julia and Thomas H. Davenport (2017). "The Emergence of Artificial Intelligence: How Automation is Changing Auditing". In: *Journal of Emerging Technologies in Accounting* 14.1, pp. 115–122. ISSN: 1558-7940, 1554-1908. DOI: 10.2308/jeta-51730. URL: https://meridian.allenpress.com/jeta/article/14/1/115/116001/The-Emergence-of-Artificial-Intelligence-How (visited on 04/19/2021).

Köpcke, Hanna and Erhard Rahm (2010). "Frameworks for entity matching: A comparison". In: *Data & Knowledge Engineering* 69.2, pp. 197–210. ISSN: 0169023X. DOI: 10.1016/j.datak.2009.10.003. URL: https://linkinghub.elsevier.com/retrieve/pii/S0169023X09001451 (visited on 02/27/2023).

Krieger, Felix and Paul Drews (2018). "Leveraging Big Data and Analytics for Auditing: Towards a Taxonomy". In: *Proceedings of the 39th International Conference on Information Systems, San Francisco*. International Conference on Information Systems. San Francisco.

Krieger, Felix, Paul Drews, Burkhardt Funk, et al. (2021). "Information Extraction from Invoices: A Graph Neural Network Approach for Datasets with High Layout Variety". In: *Wirtschaftsinformatik 2021 Proceedings*. International Conference on Wirtschaftsinformatik.

Krieger, Felix, Paul Drews, and Patrick Velte (2021). "Explaining the (non-) adoption of advanced data analytics in auditing: A process theory". In: *International Journal of Accounting Information Systems* 41, p. 100511. ISSN: 14670895. DOI: 10.1016/j.

accinf.2021.100511. URL: https://linkinghub.elsevier.com/retrieve/pii/S1467089521000130 (visited on 08/23/2021).

Langley, Ann (1999). "Strategies for theorizing from process data". In: *Academy of Management Review* 24.4, pp. 691–710. ISSN: 03637425. DOI: 10.5465/AMR.1999.2553248. URL: http://www.jstor.org/stable/259349.

Lecun, Y. et al. (1998). "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11, pp. 2278–2324. ISSN: 00189219. DOI: 10.1109/5.726791. URL: http://ieeexplore.ieee.org/document/726791/ (visited on 02/28/2023).

Li, He et al. (2018). "Understanding usage and value of audit analytics for internal auditors: An organizational approach". In: *International Journal of Accounting Information Systems* 28.November 2017, pp. 59–76. ISSN: 14670895. DOI: 10.1016/j.accinf.2017.12.005.

Liu, Xiaojing et al. (2019). "Graph Convolution for Multimodal Information Extraction from Visually Rich Documents". In: *arXiv:1903.11279 [cs]*. arXiv: 1903.11279. URL: http://arxiv.org/abs/1903.11279 (visited on 04/19/2021).

Lohani, D. et al. (2019). "An Invoice Reading System Using a Graph Convolutional Network". In: *Computer Vision – ACCV 2018 Workshops*. Lecture Notes in Computer Science 11367. Ed. by Gustavo Carneiro and Shaodi You, pp. 144–158. DOI: 10.1007/978-3-030-21074-8_12. URL: http://link.springer.com/10.1007/978-3-030-21074-8_12 (visited on 04/19/2021).

Lowe, D. Jordan et al. (2018). "Information technology in an audit context: Have the big 4 lost their advantage?" In: *Journal of Information Systems* 32.1, pp. 87–107. ISSN: 15587959. DOI: 10.2308/isys-51794.

Majumder, Bodhisattwa Prasad et al. (2020). "Representation Learning for Information Extraction from Form-like Documents". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, pp. 6495–6504. DOI: 10.18653/v1/2020.acl-main.580. URL: https://www.aclweb.org/anthology/2020.acl-main.580 (visited on 04/19/2021).

Markus, M Lynne and Daniel Robey (1988). "Information Technology and Organizational Change: Causal Structure in Theory and Research". In: *Management Science* 34.5, pp. 583–598. ISSN: 0025-1909. DOI: 10.1287/mnsc.34.5.583. URL: http://pubsonline.informs.org/doi/abs/10.1287/mnsc.34.5.583.

Marten, Kai-Uwe et al. (2020). *Wirtschaftsprüfung: Grundlagen des betriebswirtschaftlichen Prüfungswesens nach nationalen und internationalen Normen*. 6., überarbeitete Auflage. Lehrbuch. Stuttgart: Schäffer-Poeschel Verlag. 1055 pp. ISBN: 978-3-7910-4385-2 978-3-7910-4384-5.

Moffitt, Kevin C. et al. (2018). "Robotic Process Automation for Auditing". In: *Journal of Emerging Technologies in Accounting* 15.1, pp. 1–10. ISSN: 1558-7940, 1554-1908. DOI: 10.2308/jeta-10589. URL: https://meridian.allenpress.com/jeta/article/15/1/1/9413/Robotic-Process-Automation-for-Auditing (visited on 10/06/2021).

Nickerson, Robert C. et al. (2013). "A method for taxonomy development and its application in information systems". In: *European Journal of Information Systems* 22.3, pp. 336–359.

Park, Seunghyun et al. (2019). "CORD: A Consolidated Receipt Dataset for Post-OCR Parsing". In: *Workshop on Document Intelligence at NeurIPS 2019*. URL: https://openreview.net/forum?id=SJl3z659UH (visited on 12/09/2022).

Payne, Elizabeth A. and Shirley Curtis (2006). "An Examination of Contextual Factors and Individual Characteristics Affecting Technology Implementation Decisions in Auditing". In: *Journal of Chemical Information and Modeling* 53.9, pp. 1689–1699. ISSN: 1098-6596. DOI: 10.1017/CBO9781107415324.004. arXiv: arXiv:1011.1669v3.

Pedrosa, Isabel, Carlos J. Costa, and Manuela Aparicio (2020). "Determinants adoption of computer-assisted auditing tools (CAATs)". In: *Cognition, Technology and Work* 22.3, pp. 565–583. ISSN: 14355566. DOI: 10.1007/s10111-019-00581-4. URL: https://doi.org/10.1007/s10111-019-00581-4.

Pedrosa, Isabel, Carlos J. Costa, and Raul M.S. Laureano (2015). "Motivations and limitations on the use of information technology on statutory auditors' work: An exploratory study". In: *2015 10th Iberian Conference on Information Systems and Technologies, CISTI 2015* June. DOI: 10.1109/CISTI.2015.7170623.

PWC (2023). *PWC Halo*. URL: https://www.pwc.de/de/im-fokus/abschlusspruefung/unsere-tools.html#halo (visited on 02/22/2023).

Rosli, Khairina et al. (2012). "Factors Influencing Audit Technology Acceptance by Audit Firms: A New I-TOE Adoption Framework". In: *Journal of Accounting and Auditing: Research & Practice* 2012, pp. 1–11. ISSN: 2165-9532. DOI: 10.5171/2012.876814.

Salijeni, George et al. (2019). "Big Data and changes in audit technology: contemplating a research agenda". In: *Accounting and Business Research* 49.1, pp. 95–119. ISSN: 21594260. DOI: 10.1080/00014788.2018.1459458.

Schreyer, Marco et al. (2022). *Federated and Privacy-Preserving Learning of Accounting Data in Financial Statement Audits*. arXiv: 2208.12708[cs]. URL: http://arxiv.org/abs/2208.12708 (visited on 03/31/2023).

Schuster, Daniel et al. (2013). "Intellix - End-user trained information extraction for document archiving". In: *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, pp. 101–105. ISSN: 15205363. DOI: 10.1109/ICDAR.2013.28.

Sein et al. (2011). "Action Design Research". In: *MIS Quarterly* 35.1, p. 37. ISSN: 02767783. DOI: 10.2307/23043488. URL: https://www.jstor.org/stable/10.2307/23043488 (visited on 10/05/2021).

Siew, Eu Gene et al. (2020). "Organizational and environmental influences in the adoption of computer-assisted audit tools and techniques (CAATTs) by audit firms in Malaysia". In: *International Journal of Accounting Information Systems* 36, p. 100445. ISSN: 14670895. DOI: 10.1016/j.accinf.2019.100445. URL: https://doi.org/10.1016/j.accinf.2019.100445.

Stanisławek, Tomasz et al. (2021). "Kleister: Key Information Extraction Datasets Involving Long Documents with Complex Layouts". In: *Document Analysis and Recognition – ICDAR 2021*. Ed. by Josep Lladós et al. Cham: Springer International Publishing, pp. 564–579. ISBN: 978-3-030-86549-8.

Stewart, Trevor R. (2015). "Data analytics for financial statement audits". In: *AICPA Audit Analytics and Continuous Audit* 105, pp. 105–128.

Sun, Ting Sophia (2019). "Applying deep learning to audit procedures: An illustrative framework". In: *Accounting Horizons* 33.3, pp. 89–109. ISSN: 15587975. DOI: 10.2308/acch-52455.

Tanner, Christian and Sarah-Louise Richter (2018). "Digitalizing B2B Business Processes—The Learnings from E-Invoicing". In: *Business Information Systems and Technology 4.0: New Trends in the Age of Digital Change*. Ed. by Rolf Dornberger. Cham: Springer International Publishing, pp. 103–116. ISBN: 978-3-319-74322-6. DOI: 10.1007/978-3-319-74322-6_7. URL: https://doi.org/10.1007/978-3-319-74322-6_7.

Tecuci, Dan G. et al. (2020). "DICR: AI Assisted, Adaptive Platform for Contract Review". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.9, pp. 13638–13639. ISSN: 2374-3468, 2159-5399. DOI: 10.1609/aaai.v34i09.7106. URL: https://aaai.org/ojs/index.php/AAAI/article/view/7106 (visited on 10/05/2021).

Vasarhelyi, Miklos A. et al. (2015). "Big Data in Accounting: An Overview". In: *Accounting Horizons* 29.2, pp. 381–396. ISSN: 0888-7993. DOI: 10.2308/acch-51071.

Vaswani, Ashish et al. (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Veličković, Petar et al. (2018). "Graph attention networks". In: *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, pp. 1–12. arXiv: 1710.10903.

Wang, Jiapeng et al. (2022). *LiLT: A Simple yet Effective Language-Independent Layout Transformer for Structured Document Understanding*. DOI: 10.48550/ARXIV.2202.13669. URL: https://arxiv.org/abs/2202.13669.

Werner, Michael et al. (2021). "Embedding process mining into financial statement audits". In: *International Journal of Accounting Information Systems* 41, p. 100514. ISSN: 1467-0895. DOI: 10.1016/j.accinf.2021.100514. URL: https://www.sciencedirect.com/science/article/pii/S1467089521000166 (visited on 02/23/2023).

Widuri, Rindang et al. (2016). "Adopting generalized audit software: an Indonesian perspective". In: *Managerial Auditing Journal* 31.8-9, pp. 821–847. ISSN: 02686902. DOI: 10.1108/MAJ-10-2015-1247.

Wolpert, David H. (2002). "The Supervised Learning No-Free-Lunch Theorems". In: *Soft Computing and Industry*. Ed. by Rajkumar Roy et al. London: Springer London, pp. 25–42. ISBN: 978-1-4471-1101-6 978-1-4471-0123-9. DOI: 10.1007/978-1-4471-0123-9_3. URL: http://link.springer.com/10.1007/978-1-4471-0123-9_3 (visited on 11/08/2022).

Xu, Yang, Yiheng Xu, et al. (2022). "LayoutLMv2: Multi-Model Pre-Training For Visually-Rich Document Understanding". In: *arXiv:2012.14740 [cs.CL]*, p. 17. DOI: 10.48550/arXiv.2012.14740. URL: https://arxiv.org/abs/2012.14740 (visited on 12/09/2022).

Xu, Yiheng, Minghao Li, et al. (2020). "LayoutLM: Pre-training of Text and Layout for Document Image Understanding". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1192–1200. DOI: 10.1145/3394486.3403172. arXiv: 1912.13318. URL: http://arxiv.org/abs/1912.13318 (visited on 05/09/2021).

Yoon, Kyunghee et al. (2015). "Big Data as Complementary Audit Evidence". In: *Accounting Horizons* 29.2, pp. 431–438. ISSN: 0888-7993. DOI: 10.2308/acch-51076.

Yu, Wenwen et al. (2020). "PICK: Processing Key Information Extraction from Documents using Improved Graph Learning-Convolutional Networks". In: *arXiv:2004.07464 [cs]*.

DOI: 10.48550/arXiv.2004.07464. arXiv: 2004.07464. URL: http://arxiv.org/abs/2004.07464 (visited on 02/16/2021).

Zhang, Peng et al. (2020). "TRIE: End-to-End Text Reading and Information Extraction for Document Understanding". In: *Proceedings of the 28th ACM International Conference on Multimedia*. MM '20: The 28th ACM International Conference on Multimedia. Seattle WA USA: ACM, pp. 1413–1422. ISBN: 978-1-4503-7988-5. DOI: 10.1145/3394171.3413900. URL: https://dl.acm.org/doi/10.1145/3394171.3413900 (visited on 02/16/2021).

Zhao, Xiaohui et al. (2019). "CUTIE: Learning to Understand Documents with Convolutional Universal Text Information Extractor". In: *arXiv:1903.12363 [cs]*. DOI: 10.48550/arXiv.1903.12363. URL: http://arxiv.org/abs/1903.12363 (visited on 02/16/2021).

Zhaokai, Yan and Kevin C. Moffitt (2019). "Contract Analytics in Auditing". In: *Accounting Horizons* 33.3, pp. 111–126. ISSN: 0888-7993, 1558-7975. DOI: 10.2308/acch-52457. URL: http://meridian.allenpress.com/accounting-horizons/article/33/3/111/427543/Contract-Analytics-in-Auditing (visited on 10/06/2021).

Publications

Leveraging Big Data and Analytics for Auditing ... 55
Explaining the (Non-) Adoption of Advanced ... 69
Information Extraction from Invoices: A Graph Neural ... 107
Automated Invoice Processing: Machine Learning-Based ... 125
Benchmarking Machine Learning Models in Auditing: ... 154

Chapter I

Leveraging Big Data and Analytics for Auditing: Towards a Taxonomy

Outline

I.1 Introduction . . . 57
I.2 Related Research . . . 57
I.3 Method . . . 60
I.4 Results . . . 63
I.4.1 Analytics . . . 63
I.4.2 Data Management . . . 63
I.4.3 Auditing . . . 64
I.5 Discussion and Limitations . . . 64
I.6 Conclusion and Outlook . . . 65
I.7 References . . . 66

Bibliographic Information

Krieger, F., Drews, P. (2018). "Leveraging Big Data and Analytics for Auditing: Towards a Taxonomy". In: ICIS 2018 Proceedings. Association for Information Systems. ISBN: 978-0-9966831-7-3. Researchgate: 328902212.

Author's contribution

The author's share of the publication is 65%. Table C.1 in appendix C shows the contributions of all authors of the publication in detail.

Copyright Notice

Abstract

The application of big data and analytics to auditing is sparking a lot of interest in both research and practice. Along with other technological developments, big data and analytics are expected to drive the digitization of auditing and to improving its effectivity and efficiency. While several use cases and first literature reviews on this topic have already been published, the categories for classifying the use cases are still fragmented. By employing a systematic taxonomy development process, we developed a taxonomy that draws upon conceptual work and use cases from the academic literature. This taxonomy provides dimensions and characteristics that help to classify use cases for big data in analytics in auditing in a structured manner.

I.1 Introduction

Self-driving connected cars, smart manufacturing, the Internet of Things – modern technology is transforming established business models, industries, and whole economies. Business leaders are developing and implementing digital transformation strategies to keep up with the pace of the technological advancement in order to grow and protect their business. This also applies to public accounting, an industry facing stagnating revenue growth in their core business external auditing (Rapoport, 2018). Though the employment of computer-assisted auditing tools (CAAT) is established in industry practice, the field is known to lag in the adoption of new technologies (Alles, 2015). This mentality is starting to change: The applicability of technologies like blockchain, robotic automation, cloud computing and big data to the audit of financial statements is being discussed and realized in practice (Justenhoven et al., 2017).

The term big data describes the phenomenon of companies being able to measure, record and digitally capture almost anything. Captured contents range from transaction records over text to images and video content, and are being generated at a frequency that enables the tracking of events in real-time (Manyika et al., 2011). The big data assets are low in value density, if not subjected to further processing and analysis. To derive value from them, analytics is a crucial element of big data (Gandomi and Haider, 2015). Advanced analytical methods can also be applied to smaller data sets for generating new insights relevant to the auditing processes.

The application of big data and analytics in auditing (BDAA) is a rather new research field. The existent research is geared towards specific, auditing-related topics which are loosely connected. It is influenced by neighboring fields like accounting and fraud detection and being discussed from different perspectives. The aim of the presented research in progress is to construct a taxonomy for use cases of big data and analytics in auditing. The research question we address is: Which dimensions and characteristics should a taxonomy for BDAA use cases comprise? The taxonomy can also be used to identify combinations of case characteristics that have not yet been reported about.

The further structure of this paper is as follows: First, we present the relevant related literature and identify the need for a unifying taxonomy. Then we present our approach for the construction of the taxonomy, followed by the intermediate results. We conclude by discussing the results and by giving an outlook on how the work will be continued and on further research.

I.2 Related Research

The research field of BDAA lays at the intersection of the two major research streams Big Data and Auditing. Stewart (2015) describes the goal of Auditing as the issuing of an assurance that a company's financial statements do not exhibit any material misstatements, whether intended or not, with respect to some financial reporting framework, e.g. local or international accounting principles. This assurance is to be of a 'reasonable high' degree, which is generally assumed to be 95% and is a matter of the auditor's subjective judgment (Byrnes et al., 2015). An audit engagement can be expressed as a cycle consisting of

different phases, e.g. as described by Louwers et al. (2008). They identified the phases of audit pre-planning, contracting, understanding of internal controls and identification of risk factors, assessment of control risk, substantive testing, evaluation of evidence and audit reporting.

Big data is generally characterized by the three V's that constitute what qualifies data assets as 'big': volume, velocity, and variety (Chen, Chiang, et al., 2012; Gandomi and Haider, 2015). In addition, further dimensions were introduced that relate to the assets inherent characteristics: veracity, variability, complexity and value (Gandomi and Haider, 2015). Manyika et al. (2011) state that the creation of economic value through big data can be realized (among others) through the automation of human decision making and the innovation of new business models, products and services. They further argue that big data can improve both operational efficiency and effectiveness (ibid.). Transferred to the context of auditing, this is supported by both (Byrnes et al., 2015) and (Stewart, 2015), which see technology as a means of achieving a higher degree of assurance (improvement of audit effectiveness) or to remain at the same level, but at lower cost (improvement of efficiency). Byrnes et al. (2015) observed that the latter is receiving more attention than the former, which by their opinion is due to the economics in the public accounting industry.

Chen, Chiang, et al. (2012) discuss big data in the historical context of business intelligence & analytics and consider it the next evolutional step. They analyze promising applications of big data and present a research framework in which emerging research and foundational technologies are classified into the areas data analytics, text analytics, web analytics, network analytics and mobile analytics. Gandomi and Haider (2015) focus on big data analytics on semi- and unstructured data. The process of analyzing data is broken down into the two stages data management and analytics. Data management are the technologies and processes related to the acquisition, storage and retrieval of data, analytics refers to the techniques used for analyzing the data. They review techniques for the analysis of text, audio, video, and social media data. Yaqoob et al. (2016) review state-of-the-art processing techniques and methods for big data, as well as analysis techniques. They discuss the related opportunities and challenges and present emerging technologies.

Amani and Fadlalla (2017) conducted a literature review on data mining applications in accounting and propose a framework for mapping these applications to the three main accounting topics financial accounting, management accounting and assurance and compliance. They conclude from their analysis that the latter benefits the most from data mining. Appelbaum et al. (2018) reviewed the analytical methods employed in external auditing and map it to an external audit cycle model. They identify that there is no existent research aimed at employing prescriptive methods and a lack of research regarding the methodological support continuous auditing activities. Debreceny and Gray (2011) show the applicability of social network analysis in auditing, Jans et al. (2013) the applicability of process mining.

Analytics can be employed towards the automation of single auditing tasks, as well as the whole auditing cycle (Issa et al., 2016; Kokina and Davenport, 2017). Kokina and Davenport (2017) argue that many audit tasks are highly suitable for automation. Issa et al. (2016) conceptualize the audit cycle as a theoretically fully automatable production line, where the output of one phase becomes the input of the subsequent one. They

provide examples of techniques, data sources, and technologies that come potentially into play. Closely related to automation is the concept of continuous auditing, a novel approach to auditing which accounts for the increasing frequency of data generation and benefits from big data (Bumgarner and Vasarhelyi, 2018). Zhang et al. (2015) discuss the gaps between existing and required data analysis capabilities in continuous auditing for the application of big data. Automation and continuous auditing can be related to the velocity aspect of big data and the respective stream-processing based data management technologies (Gandomi and Haider, 2015; Yaqoob et al., 2016).

The audit of financial statements relies by nature on financial accounting data. Accounting is facing a potential paradigm shift caused by big data, as business transactions can be traced earlier and deeper (Vasarhelyi et al., 2015). Vasarhelyi et al. (ibid.) discuss different big data sources relevant for accounting and auditing and how they can be integrated with financial accounting data. He identifies the need for 'data bridges'. Further examples for relevant data sources are given by Issa et al. (2016). The structural differences in these sources can be expressed by the variety and complexity dimensions of big data (Gandomi and Haider, 2015). Yoon et al. (2015) discuss big data-based audit evidence in the context of the audit evidence criteria framework given by the AICPA SAS nr. 106. They find that such evidence can benefit auditor independence but express concerns regarding data quality and the lack of causal implications. Their concerns can be referred to the veracity dimension of big data (Gandomi and Haider, 2015). In addition to the change in what data is used, there are shifts in the way how data is being used. The common practice of using data samples could be replaced by analyzing full populations. Full-population testing is widely considered to be beneficial to auditing effectiveness and directly relates to the volume aspect of big data (Alles and Gray, 2016; Alles, 2015; Byrnes et al., 2015; Jans et al., 2013; Kokina and Davenport, 2017).

The applicability of Big Data and Analytics to Auditing practice in terms of legally binding professional standards is a complex topic. Essentially, this topic has to be split into two questions: (1) Are data analytics applicable? (2) Are Big Data items a suitable data source? To address these questions, we refer to the International Auditing and Assurance Standards Board (IAASB) which issues the International Standards for Auditing (ISA). Regarding the first question, a working paper by the IAASB's Data Analytics Working Group clearly states that "The ISAs do not prohibit, nor stimulate, the use of data analytics." (Group, 2016). Hence, the position of auditing standards regarding the applicability of data analytics is neutral.

The suitability of Big Data items for Auditing purposes can be evaluated against the audit evidence framework provided in ISA 500, which consists of the criteria sufficiency, reliability and relevance. The evaluation of data items has to be conducted by the auditor as part of his professional judgement (ISA 500, 2009). ISA 500 generally encourages auditors to refer to information from different sources, especially those independent from the audited entity, e.g. social media, which is often "big" in nature. However, the IASSB (2016) also states that the relevance and reliability of external data have to be handled with special care (Group, 2016; ISA 500, 2009). Therefore, auditors are encouraged to consider additional data sources that might be "big", while the standards also stress that special care should be applied for external data.

Our search for related literature showed that a number of literature reviews, conceptional research papers and application-oriented experimental research papers already exist in the young research field of BDAA. A taxonomy for BDAA use cases could advance research for several reasons: First, the related research we present above examines the phenomenon of big data from a technical auditing perspective and a technological perspective. The former is primarily concerned with the exogenous effects and potential benefits of big data and is predominantly conceptional in nature. The latter surveys state-of-the-art methods and technologies to handle big data across different areas of application. We identified intersections between the two that have not yet been explicitly addressed. Second, the different concepts that can be found in the literature overlap technically but are often discussed independently, e.g. auditing process automation and big data-based audit evidence. Third, especially in the domain of analytical methods, we identified a great heterogeneity in the terms and classification schemes used. Taxonomies can be classified as 'theory for analyzing' (Gregor, 2006). This kind of theory describes and analyzes a phenomenon and is especially useful in areas where there is still not much known about the phenomenon under study because it forms the basis for all other kinds of theory in IS research (ibid.).

I.3 Method

For developing the BDAA use case taxonomy, we draw upon the taxonomy development method published by Nickerson et al. (2013). According to them, a taxonomy consists of characteristics that describe objects. The characteristics are grouped into dimensions. Nickerson et al. (ibid.) address the question of how to devise the dimensions and characteristics. The authors conceptualize an iterative process, in a detailed manner and illustrated on an example. The methodology takes alternative approaches for constructing taxonomies into account and is straightforward in its application, which is why it has adopted as methodology for this paper. The objects we intend to describe through the taxonomy are BDAA use cases that reflect how big data and analytics can be employed to benefit auditing.

We started by constructing a first version of the taxonomy by deducting the dimensions and characteristics from the related literature ('conceptual-to-empirical'). Following the 'build/test'-cycle, we applied the taxonomy to a series of selected use cases to see if it would hold, and if new dimensions or characteristics could be devised bottom-up ('empirical-to-conceptual'). The use cases we identified so far are taken from the related literature. A total of twelve use cases were used to devise the taxonomy.

We found that depending on the background and focus of their research, the authors may not include all of the information required to assign a characteristic for every dimension to each object, which is we divert from Nickerson et al. (ibid.) and allow for an object to not be assigned a characteristic. We further allow for an object to be assigned multiple characteristics within one dimension, as e.g. multiple analysis methods or data sources may be used in the same use case. If possible, we inferred not explicitly mentioned characteristics from context.

Table I.1: The Big Data and Analytics in Auditing Taxonomy

Dimension		Characteristics								
Analytics	Analyis Method	Frequent Patterns	Classification	Regression	Cluster Analysis	Outlier Analysis	Visualization	Optimization	Process Mining	SNS
	Analysis Goal	Descriptive			Predictive			Prescriptive		
	Data Scope	Sample				Full-Population				
	Data Bridge	Natural Language Processing								
	Data Reliability	High				Low				
Data Management	Data Origin	Internal				External				
	Data Source	Financial		Communication		Multimedia		Machine Generated Data		
	Data Format	Numerical		Textual		Image		Audio		Geospatial
	Data Structure	Structured			Semi-Structured			Unstructured		
	Processing Technique	Batch				Stream				
Auditing	Audit Cycle Phase	Pre-Planning	Contracting	Understanding of Internal Controls and Identification of Risk Factors		Assessment of Control Risk		Substantive Testing	Evaluation of Evidence	Audit Reporting
	Task Structure	Structured			Semi-Structured			Unstructured		
	Improvement Potential	Effectiveness				Efficiency				

Table I.2: Exemplary Application of the Taxonomy

Use Case	Network Analysis of E-Mail communication for Fraud Detection	Identification of anomalous journal entries using Deep Autoencoder Networks
Source reference	Debreceny and Gray (2011)	Schreyer et al. (2017)
Dimension	Assigned characteristics	
Analysis Method	Social Network Analysis	Outlier Analysis
Analysis Goal	Descriptive	Descriptive
Data Scope	Sampling	Full-Population
Data Bridge	Natural Language Processing	-
Data Reliability	Not defined	High
Data Origin	Internal	Internal
Data Source	Communication	Financial
Data Format	Text	Numerical
Data Structure	Semi-Structured	Structured
Processing Technique	Batch	Batch
Audit Cycle Phase	Assessment of Control Risk	Substantive Testing
Task Structure	Semi-Structured	Structured
Improvement Potential	Effectiveness	Efficiency

I.4 Results

Table I.1 depicts the taxonomy under development after one build/test'-cycle. We allocated the dimensions into three groups; analytics, data management and auditing. The groups analytics and data management reflect the process of extracting value from big data (Gandomi and Haider, 2015), the auditing dimensions cover the aspect of how and to what end BDAA is employed.

I.4.1 Analytics

The dimension analysis method describes the approach to data analytics which is employed in the use case. As already pointed out, we found different terms and classification schemes for analytical methods in the literature. To attain a unifying structure, we abstract from individual models and algorithms. Instead, we focus on the task that is being addressed. Han et al. (2011) defined five typical data mining tasks: frequent pattern mining, classification, regression, cluster analysis, and outlier analysis. We add further analysis methods that we found in the literature but were not able to map to the existing five tasks: visualization (Appelbaum et al., 2018; Chen and Zhang, 2014), social network analysis (SNS) (Chen and Zhang, 2014; Debreceny and Gray, 2011) process mining (Jans et al., 2013) and optimization methods (Amani and Fadlalla, 2017; Chen and Zhang, 2014).

Analysis goal characterizes the three perspectives on data-driven business decision making: descriptive, predictive and prescriptive (Amani and Fadlalla, 2017; Appelbaum et al., 2018). Han et al. (2011) classify the above-outlined data mining tasks into descriptive and predictive. Descriptive tasks seek to describe the properties of the data set, whereas predictive tasks are used to make inductions from data, which is also referred to as predictive analytics (Gandomi and Haider, 2015; Han et al., 2011). If optimization methods are employed complimentary to predictive analytics to minimize or maximize an objective function, it is considered prescriptive analytics (Amani and Fadlalla, 2017; Evans, 2012).

Data scope refers to whether a sampling technique is employed to select a subset of the available data, which is common to Auditing, or the complete data is being analyzed (Full population testing). Data bridges are methods that are used to process semi- or unstructured data in such a way that they are suitable for analysis towards the intended goal. Data reliability picks up the discussion led by Yoon et al. (2015) and characterizes the validity of the data under analysis.

I.4.2 Data Management

As pointed out by Alles and Gray (2016) and Earley (2015), the notion of what constitutes big data in auditing and accounting relies more on the type of analysis that can be conducted with the data, rather than its source and format. We adopt a more technical view of the analyzed data and thus decompose its nature in the dimensions data source, data structure, data format and data type.

The data origin and data structure of the analyzed data reflect the variety aspect of big data on a high level. Internal data refers to the data that is generated and captured by the audit client in his systems and (potentially) provided to the auditors, e.g. financial data,

sensor data, etc. External data is generated from sources outside of the company and holds potentially relevant information, like consumer opinions in social media (Appelbaum et al., 2018; Chen, Chiang, et al., 2012). The data format describes the variety from a data processing standpoint, as different formats pose challenges in terms of the handling their structures and semantics and require different processing approaches to gain insights from it, e.g. textual data (Gandomi and Haider, 2015; Han et al., 2011). The data source indicates the source domain of the data. In addition to financial data, we examined the big data sources that are relevant to auditing as described by Moffitt and Vasarhelyi (2013) and Issa et al. (2016). To keep the taxonomy flexible and concise, we aggregated the sources into higher-level groups. Communication encompasses all forms of internal and external exchange, e.g. emails as well as social media. Multimedia refers to audio-, image- and video-based content, like video material from security systems or Youtube videos. Machine generated data is being generated automatically from technical equipment, e.g. from Internet-of-Things-enabled devices, sensors or web servers.

The last technological dimension we derived is the processing technique, which refers to the two data processing paradigms which are batch and stream processing. Batch processing addresses the problem of having to analyze big volumes of data. A prominent technology for distributed batch-processing is Apache Hadoop. Stream processing addresses the problem of having to process data in (near) real-time that requires a low latency response. Known technologies are Apache Storm and Spark (Gandomi and Haider, 2015).

I.4.3 Auditing

Issa et al. (2016) address how Analytics can be used in an external audit engagement, based on the audit cycle process model introduced by Louwers et al. (2008). They conceptualize the audit cycle as a production line where the output of one step becomes the input of the subsequent step, and thus theoretically allows an end-to-end automation. We adopt this process model of the auditing process, as the provided ideas for the employment of analytics per process step supports the mapping of use cases to audit cycle steps. Issa et al. (2016) point out that if broken down into single tasks, auditing consists mainly of repetitive work and decision making, which has implications for the automation of the audit cycle. Therefore, we included the dimension task structure. Abdolmohammadi (1999) examined the structure of auditing tasks. He considered a task to be structured if the underlying problem is well defined, with a limited number of alternatives and requires little judgment. Unstructured tasks are hard to define, have many alternative solutions and requires substantial judgment. Tasks on the spectrum between these two are considered semi-structured. The auditing improvement potential describes the above-outlined type of value contribution (explicitly or implicitly) intended by the use case, whether it contributes to the effectiveness or efficiency of auditing.

I.5 Discussion and Limitations

In a first build/test-cycle we derived 12 dimensions from the literature and tested them against a sample of use cases. We found on two use cases (Debreceny and Gray, 2011; Yang et al., 2017) that the taxonomy lacks a dimension that reflects intermediate processing steps that were applied to prepare the semi-structured data for analysis. We introduced,

therefore, the dimension "Data Bridge", which relates to the concept given by Vasarhelyi et al. (2015). Table I.2 exhibits the exemplary application of the taxonomy after the first iteration to two uses cases given by Debreceny and Gray (2011) and Schreyer et al. (2017). We found that the taxonomy is apt to describe the use cases but is not yet exhaustive enough to reflect all of the complexities thoroughly. From this, we plan to further diversify the taxonomy in terms of exploring additional dimensions and refining existing dimensions through additional characteristics.

I.6 Conclusion and Outlook

BDAA is a promising research area with regard to its contribution to the audit profession. Auditing is already a complex topic by itself, driven by different regulatory frameworks like GAAPs and professional standards. This complexity is further increased by the broad range of technologies and methods employed in the context of big data. The existent literature on BDAA only covers the phenomenon partially. To close this gap, we present a taxonomy that connects the Auditing and technological perspectives and puts the different concepts discussed in BDAA literature into context. The intended use of the taxonomy is to identify research gaps and to provide a guideline for future research in the field. In Auditing practice, possible applications of the taxonomy are to support the ideation process for new use cases and to navigate through existing use cases. The ideation process can use the taxonomy for identifying application areas for BDAA which have not been explored so far. The use cases listed in this paper can be checked for applicability in the own enterprise. The taxonomy also provides an overview of techniques and prerequisites for deriving requirements for a data analytics platform to be used in Auditing. Furthermore, it could also be employed as an input for use case-oriented knowledge representations in ontologies or knowledge graphs.

As the taxonomy is not yet sufficiently exhaustive, our further research will diversify the taxonomy and improve its robustness. To this end, we will subject it to further 'build/test'-cycles. In addition to the refinement conducted by us, we intend to present the taxonomy to scholars and practitioners active in the BDAA field. This way hope to identify use cases that are relevant in practice but have not been mentioned in the literature.

The taxonomy will be evaluated and refined in further cycles according to the taxonomy development method. The next step in our research will be of empirical nature. We will further investigate this research topic based on expert interviews. The taxonomy will support this study by contributing relevant dimensions and characteristics for structuring the interviews.

I.7 References

Abdolmohammadi, Mohammad J (1999). "A Comprehensive Taxonomy of Audit Task Structure, Professional Rank and Decision Aids for Behavioral Research". In: *Behavioral Research in Accounting* 11. Publisher: American Accounting Association, pp. 51–92.

Alles, Michael and Glen L. Gray (2016). "Incorporating big data in audits: Identifying inhibitors and a research agenda to address those inhibitors". In: *International Journal of Accounting Information Systems* 22, pp. 44–59. ISSN: 1467-0895. DOI: https://doi.org/10.1016/j.accinf.2016.07.004. URL: https://www.sciencedirect.com/science/article/pii/S1467089516300811.

Alles, Michael G. (2015). "Drivers of the Use and Facilitators and Obstacles of the Evolution of Big Data by the Audit Profession". In: *Accounting Horizons* 29.2. _eprint: https://meridian.allenpress.com/accounting-horizons/article-pdf/29/2/439/1592722/acch-51067.pdf, pp. 439–449. ISSN: 0888-7993. DOI: 10.2308/acch-51067. URL: https://doi.org/10.2308/acch-51067.

Amani, Farzaneh A. and Adam M. Fadlalla (2017). "Data mining applications in accounting: A review of the literature and organizing framework". In: *International Journal of Accounting Information Systems* 24, pp. 32–58. ISSN: 14670895. DOI: 10.1016/j.accinf.2016.12.004. URL: http://dx.doi.org/10.1016/j.accinf.2016.12.004.

Appelbaum, Deniz A. et al. (2018). "Analytical procedures in external auditing: A comprehensive literature survey and framework for external audit analytics". In: *Journal of Accounting Literature* 40, pp. 83–101. ISSN: 07374607. DOI: 10.1016/j.acclit.2018.01.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S0737460716300611 (visited on 08/23/2021).

Bumgarner, Nancy and Miklos A. Vasarhelyi (2018). "Continuous Auditing—A New View". In: *Continuous Auditing*. Ed. by David Y. Chan et al. Emerald Publishing Limited, pp. 7–51. ISBN: 978-1-78743-413-4 978-1-78743-414-1. DOI: 10.1108/978-1-78743-413-420181002. URL: https://doi.org/10.1108/978-1-78743-413-420181002 (visited on 09/16/2022).

Byrnes, Paul et al. (2015). "Reimagining Auditing in a Wired World". In: *Audit Analytics and Continuous Audit: Looking Toward the Future*. New York: American Institute of Certified Public Accountants, Inc. ISBN: 978-1-943546-08-4. URL: https://us.aicpa.org/content/dam/aicpa/interestareas/frc/assuranceadvisoryservices/downloadabledocuments/auditanalytics_lookingtowardfuture.pdf (visited on 09/16/2022).

Chen, C. L. Philip and Chun-Yang Zhang (2014). "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data". In: *Information Sciences* 275, pp. 314–347. ISSN: 0020-0255. DOI: https://doi.org/10.1016/j.ins.2014.01.015. URL: https://www.sciencedirect.com/science/article/pii/S0020025514000346.

Chen, Hsinchun, Roger H L Chiang, et al. (2012). "Business Intelligence And Analytics: From Big Data to Big Impact". In: *MIS Quarterly* 36.4, pp. 1165–1188.

Debreceny, Roger S. and Glen L. Gray (2011). "Data Mining of Electronic Mail and Auditing: A Research Agenda". In: *Journal of Information Systems* 25.2. _eprint: https://meridian.allenpress.com/jis/article-pdf/25/2/195/1719069/isys-10167.pdf, pp. 195–226. ISSN: 0888-7985. DOI: 10.2308/isys-10167. URL: https://doi.org/10.2308/isys-10167.

Evans, James R. (2012). *Business Analytics: The Next Frontier for Decision Sciences*. URL: http://faculty.cbpp.uaa.alaska.edu/afef/business_analytics.htm (visited on 09/16/2022).

Gandomi, Amir and Murtaza Haider (2015). "Beyond the hype: Big data concepts, methods, and analytics". In: *International Journal of Information Management* 35.2, pp. 137–144. ISSN: 0268-4012. DOI: https://doi.org/10.1016/j.ijinfomgt.2014.10.007. URL: https://www.sciencedirect.com/science/article/pii/S0268401214001066.

Gregor, Shirley (2006). "The Nature of Theory in Information Systems". In: *MIS Quarterly* 30.3. Publisher: Management Information Systems Research Center, University of Minnesota, pp. 611–642. DOI: https://doi.org/10.2307/25148742. URL: http://www.jstor.org/stable/25148742 (visited on 01/28/2023).

Group, IAASB Data Analytics Working (2016). *Exploring the Growing Use of Technology in the Audit, with a Focus on Data Analytics*. Exposure Drafts and Consultation Papers. New York: International Auditing and Assurance Standards Board. URL: https://www.iaasb.org/publications/exploring-growing-use-technology-audit-focus-data-analytics (visited on 09/16/2022).

Han, Jiawei et al. (2011). *Data Mining: Concepts and Techniques*. 3rd. The Morgan Kaufmann Series in Data Management Systems. Waltham, USA: Elsevier, Morgan Kauffman. ISBN: 978-0-12-381479-1.

ISA 500, IAASB (2009). *International Standard On Auditing 500 Audit Evidence*. URL: https://www.ifac.org/system/files/downloads/a022-2010-iaasb-handbook-isa-500.pdf (visited on 09/16/2022).

Issa, Hussein et al. (2016). "Research Ideas for Artificial Intelligence in Auditing: The Formalization of Audit and Workforce Supplementation". In: *Journal of Emerging Technologies in Accounting* 13.2, pp. 1–20. ISSN: 1558-7940, 1554-1908. DOI: 10.2308/jeta-10511. URL: https://meridian.allenpress.com/jeta/article/13/2/1/115980/Research-Ideas-for-Artificial-Intelligence-in (visited on 10/05/2021).

Jans, Mieke et al. (2013). "The case for process mining in auditing: Sources of value added and areas of application". In: *International Journal of Accounting Information Systems* 14.1, pp. 1–20. ISSN: 14670895. DOI: 10.1016/j.accinf.2012.06.015. URL: http://dx.doi.org/10.1016/j.accinf.2012.06.015.

Justenhoven, Petra et al. (2017). *Digital Audits of Financial Statements: Study on the use of technology in finance and accounting*. Ed. by PricewaterhouseCoopers GmbH WPG. URL: https://www.pwc.de/en/digitale-transformation/studie-digitale-abschlusspruefung-en.pdf.

Kokina, Julia and Thomas H. Davenport (2017). "The Emergence of Artificial Intelligence: How Automation is Changing Auditing". In: *Journal of Emerging Technologies in Accounting* 14.1, pp. 115–122. ISSN: 1558-7940, 1554-1908. DOI: 10.2308/jeta-51730. URL: https://meridian.allenpress.com/jeta/article/14/1/115/116001/The-Emergence-of-Artificial-Intelligence-How (visited on 04/19/2021).

Louwers, Timothy J. et al. (2008). "Auditing & assurance services". In:

Manyika, James et al. (2011). *Big data: The next frontier for innovation, competition, and productivity*. Ed. by McKinsey Global Institute.

Moffitt, Kevin C. and Miklos A. Vasarhelyi (2013). "AIS in an age of big data". In: *Journal of Information Systems* 27.2. Publisher: American Accounting Association, pp. 1–19. ISSN: 0888-7985. DOI: 10.2308/isys-10372.

Nickerson, Robert C. et al. (2013). "A method for taxonomy development and its application in information systems". In: *European Journal of Information Systems* 22.3, pp. 336–359.

Rapoport, Michael (2018). *How Did the Big Four Auditors Get $17 Billion in Revenue Growth? Not From Auditing: Consulting is now a cash cow for accounting firms, raising concerns about conflicts of interest*. Ed. by The Wall Street Journal. URL: https://www.wsj.com/articles/how-did-the-big-four-auditors-get-17-billion-in-revenue-growth-not-from-auditing-1523098800.

Schreyer, Marco et al. (2017). "Detection of Anomalies in Large Scale Accounting Data using Deep Autoencoder Networks". In: *Arxiv*.

Stewart, Trevor R. (2015). "Data analytics for financial statement audits". In: *AICPA Audit Analytics and Continuous Audit* 105, pp. 105–128.

Vasarhelyi, Miklos A. et al. (2015). "Big Data in Accounting: An Overview". In: *Accounting Horizons* 29.2, pp. 381–396. ISSN: 0888-7993. DOI: 10.2308/acch-51071.

Yang, Rong et al. (2017). "Corporate Risk Disclosure and Audit Fee: A Text Mining Approach". In: *European Accounting Review* 24.4, pp. 1–12. ISSN: 0963-8180. DOI: 10.1080/09638180.2017.1329660.

Yaqoob, Ibrar et al. (2016). "Big data: From beginning to future". In: *International Journal of Information Management* 36.6, pp. 1231–1247. ISSN: 02684012. DOI: 10.1016/j.ijinfomgt.2016.07.009.

Yoon, Kyunghee et al. (2015). "Big Data as Complementary Audit Evidence". In: *Accounting Horizons* 29.2, pp. 431–438. ISSN: 0888-7993. DOI: 10.2308/acch-51076.

Zhang, Juan et al. (2015). "Toward Effective Big Data Analysis in Continuous Auditing". In: *Accounting Horizons* 29.2, pp. 469–476. ISSN: 0888-7993. DOI: 10.2308/acch-51070.

Chapter II

Explaining the (Non-) Adoption of Advanced Data Analytics in Auditing: A Process Theory

Outline

II.1 Introduction . . . 71
II.2 Related Literature . . . 72
II.2.1 Technology adoption in auditing . . . 73
II.2.2 Advanced data analytics in auditing . . . 74
II.2.3 Use of IT specialists in financial statement audits . . . 76
II.3 Methods . . . 78
II.3.1 Data . . . 78
II.3.2 Analysis . . . 79
II.4 Results . . . 84
II.4.1 Process . . . 86
II.4.2 Contextual factors . . . 89
II.4.3 Effects of the contextual factors on the adoption process . . . 94
II.5 Discussion . . . 97
II.6 Conclusion and Outlook . . . 99
II.7 References . . . 101

Bibliographic Information

Krieger, F., Drews, P., Velte, P. (2021). "Explaining the (Non-) Adoption of Advanced Data Analytics in Auditing: A Process Theory". International Journal of Accounting Information Systems, 41. Elsevier. DOI 10.1016/j.accinf.2021.100511.

Author's contribution

The author's share of the publication is 70%. Table C.2 in appendix C shows the contributions of all authors of the publication in detail.

Copyright Notice

Abstract

Audit firms are increasingly engaging with advanced data analytics to improve the efficiency and effectiveness of external audits through the automation of audit work and obtaining a better understanding of the client's business risk and thus their own audit risk. This paper examines the process by which audit firms adopt advanced data analytics, which has been left unaddressed by previous research. We derive a process theory from expert interviews which describes the activities within the process and the organizational units involved. It further describes how the adoption process is affected by technological, organizational and environmental contextual factors. Our work contributes to the extent body of research on technology adoption in auditing by using a previously unused theoretical perspective, and contextualizing known factors of technology adoption. The findings presented in this paper emphasize the importance of technological capabilities of audit firms for the adoption of advanced data analytics; technological capabilities within audit teams can be leveraged to support both the ideation of possible use cases for advanced data analytics, as well as the diffusion of solutions into practice.

II.1 Introduction

External audits by professional accounting firms have a long tradition. With the increasing adoption of information systems in the previous three decades, particularly enterprise resource planning (ERP) systems, technology has become increasingly important for obtaining audit evidence (Alles, 2015; Braun and Davis, 2003). The large-scale adoption of ERP systems has made IT-based auditing a necessity, as the ubiquity of these systems has forced auditors to adopt the approach of "auditing through the computer" instead of auditing around it (Alles, 2015). This accelerated the development of computer-assisted audit techniques and tools (CAATTs)—particularly generalized audit software (GAS)— which supports auditors in the extraction and analysis of data, thereby improving audit efficiency and effectiveness (Braun and Davis, 2003). Currently, a cluster of new technologies has been discussed in recent audit literature as a means to further improve both the efficiency and effectiveness of audits: Data analytics, artificial intelligence (AI), robotic process automation (RPA), and big data (Alles and Gray, 2016; Cao et al., 2015; Chan et al., 2018; Kokina and Davenport, 2017; Moffitt et al., 2018). The term big data itself relates to the nature of the data source (Chen et al., 2012). It describes the increasing rate of data generation ("velocity"), the resulting masses of data that are generated ("volume"), and the structural heterogeneity of the data ("variety") (Russom, 2011). Data analytics is used to extract information from data; visualizations, statistics, and data mining techniques can be applied to data to extract information, which in turn can be used in decision-making (Chen et al., 2012). Appelbaum et al. (2018) conducted an extensive literature review on data analytics methods that are applicable for analytical audit procedures. They found a wide spectrum of methods that could be applied through all phases of audit engagement. The authors also indicate that audit practitioners still referred to a rather narrow set of techniques compared to the spectrum of techniques employed in research. Among the techniques that they found applicable is process mining (PM). PM is a data analytics method utilized to generate process models from transactional data stored in ERP systems (Chiu and Jans, 2019; Jans et al., 2013, 2014; Van Der Aalst et al., 2010). This method can be employed for the assessment of control risk and replaces the walkthrough interviews that auditors conduct regularly (ibid.). Data analytics is commonly used for structured data. Deep learning, a form of machine learning, further enables the extraction of structured representations from semi- and unstructured data such as images, text, and sound (Issa et al., 2016; Sun, 2019). Issa et al. (2016) refer to deep learning in their definition of AI, along with expert systems. Expert systems employ rule-based programming for making or informing decisions but these have fallen out of use in audit practice [1] (Gray et al., 2014). Another technology that leverages the use of rules for data processing is RPA. RPA refers to software that utilizes business rules and activity choreographies on regular user interfaces to automate human tasks (Moffitt et al., 2018). Lacity and Willcocks (2016) argue that RPA is best suited for so-called "swivel-chair" processes, where a professional

> "[...] takes in work from many electronic inputs (like emails and spreadsheets), processes it using rules, adds data as necessary by accessing more systems, and then inputs the completed work to yet other systems [...]." (ibid.)

[1] According to Gray et al. (2014), the reasons for the demise of expert systems in auditing are not well explored.

For a task to be automated through RPA, it must be well defined; ambiguous tasks are problematic (Moffitt et al., 2018). In order to increase the ability of RPA to handle more complex tasks, it can be combined with machine learning (Huang and Vasarhelyi, 2019).

This cluster of technologies is referred to as advanced data analytics (ADA), as all the technologies that are part of it are related to the processing and analysis of data that goes beyond traditional audit procedures. Audit firms, usually known to lag in the adoption of new technologies (Alles, 2015), are now recognizing the impact that these emerging technologies can have on their profession, particularly as the audit industry is facing stagnating revenues from their core business (Rapoport, 2018). The efficiency of both financial and non-financial external audits can be improved from the automation of audit work, whereas the effectiveness of such audits can be increased from the analysis of data generated by the client and third parties, thereby enabling a more thorough view of the client's business and the associated risks. If adopted, the potential impact on auditing practice is significant. Issa et al. (2016) indicate the disruptive nature of ADA and envision a highly effective, fully automated audit that is similar to a production line. Several conceptual and empirical studies have explored the individual drivers and inhibitors of ADA adoption in audit firms (Alles and Gray, 2016; Alles, 2015; Dagilienė and Klovienė, 2019; Haddara et al., 2018; Salijeni et al., 2019). Thus far, extant research has not determined how the process of ADA adoption works in audit firms and how the related drivers and inhibitors affect this process. Our research aims to close this gap by addressing the following research question:

> *Which process do audit firms utilize to adopt ADA technologies and how is this process affected by contextual factors?*

To this end, we conducted 15 semi-structured interviews with auditors and other industry experts involved in the development and implementation of ADA solutions in auditing practice. This paper contributes to the growing field of research on ADA in auditing by introducing a mid-range process theory that outlines the process underlying ADA adoption in auditing. Further, the organizational units within audit firms that are involved in the process are described and contextual factors from prior research and extent theoretical frameworks are related to the process. This paper aims to address both the AIS research community and audit practitioners. It presents a theoretical contribution to the body of research on technology adoption in auditing by using a previously unused theoretical perspective, while also providing practical insights on ADA. Further, this research makes a case for strengthening the IT-capabilities of auditors to enable audit firms, small and big, to leverage interdisciplinary skill sets for technology adoption. Therefore, it could also serve as motivation for audit standard-setters and regulators to strengthen the required IT capabilities for certified auditors.

II.2 Related Literature

Audit firms look to information technology (IT) to improve the quality, efficiency, and effectiveness of external audits (Janvrin, Bierstaker, et al., 2008; Lowe et al., 2018). Generally, there has been an increase in IT adoption in the previous decade in the audit profession (ibid.). Moreover, this development is not limited to the Big Four audit firms, who were argued to have an advantage over smaller audit firms due to their "deep pockets"

(Janvrin, Bierstaker, et al., 2008). Mid-sized audit firms have caught up in terms of their use and perceived relevance of IT and even surpassed the Big Four in certain areas, whereas small audit firms are still lagging in their adoption of IT (Lowe et al., 2018). However, the application of ADA to auditing is a rather recent phenomenon that has been perceived as potentially disruptive to the audit profession (Alles, 2015; Eilifsen et al., 2019; Hampton and Stratopoulos, 2016; Moffitt et al., 2018; Salijeni et al., 2019). Hence, ADA represents a "new breed" of technology in auditing, which also requires a new set of skills in auditing firms (Dagilienė and Klovienė, 2019; Haddara et al., 2018; Salijeni et al., 2019).

The remainder of this section is structured in the following manner: First, we introduce empirical studies on technology adoption in auditing, along with the theoretical frameworks and contextual factors used. We then review the recent research on ADA adoption in auditing, which is more explorative in nature and less guided by theory. Finally, we touch on the use of IT specialists in auditing, as they are relevant for the adoption of ADA.

II.2.1 Technology adoption in auditing

The adoption of IT in auditing has been extensively studied. Several empirical studies have been dedicated to determine which factors affect the adoption of audit technology in general (Janvrin, Bierstaker, et al., 2008; Lowe et al., 2018; Vasarhelyi and Romero, 2014), or the adoption of more specific IT applications such as CAATTs (Curtis and Payne, 2014; Janvrin, Bierstaker, et al., 2009; Janvrin, Lowe, et al., 2008; Li et al., 2018; Pedrosa, Costa, and Aparicio, 2020; Pedrosa, Costa, and Laureano, 2015; Rosli et al., 2012; Siew et al., 2020) and GAS (Ahmi and Kent, 2012; Widuri et al., 2016). The studied factors are derived from popular theoretical frameworks, such as the unified theory of acceptance and use of technology (UTAUT) (Venkatesh et al., 2003), the technology organization environment (TOE) framework (DePietro et al., 1990), and the diffusion of innovations (DoI) theory (Rogers, 2003). The utilization of these frameworks depends on the unit of analysis, since technology adoption can be studied at different organizational levels (Molinillo and Japutra, 2017; Salahshour Rad et al., 2018).

The UTAUT is predominantly used to study technology adoption at an individual level (Curtis and Payne, 2014; Janvrin, Bierstaker, et al., 2009; Janvrin, Lowe, et al., 2008; Li et al., 2018; Payne and Curtis, 2006; Pedrosa, Costa, and Laureano, 2015). The factors typically studied in UTAUT-based adoption studies are performance expectancy, effort expectancy, social influence, and facilitating conditions, which are moderated by individual variables such as gender, age, experience, and the voluntariness of use (Venkatesh et al., 2003). However, the theory has been modified and extended to accommodate the specifics of the audit domain to include factors such as budget constraints (Curtis and Payne, 2014; Payne and Curtis, 2006), the adoption preference of superiors and the firm (Payne and Curtis, 2006; Pedrosa, Costa, and Aparicio, 2020; Pedrosa, Costa, and Laureano, 2015), the effect of regulation and standards (Pedrosa, Costa, and Aparicio, 2020; Pedrosa, Costa, and Laureano, 2015), or whether support was provided by IT members of the engagement team (Pedrosa, Costa, and Aparicio, 2020).

DoI and TOE have been used to examine technology adoption at the audit firm-level (O'Donnell, 2010; Rosli et al., 2012; Siew et al., 2020; Widuri et al., 2016). In DoI theory,

technology (or innovation) adoption is the outcome of an individual's or organization's decision process (Rogers, 2003). The organizational process encompasses five stages, which are divided into two sub-processes: The initiation phase and the implementation phase. The sub-processes are divided by the decision to adopt an innovation. During the initiation phase, the organization recognizes an organizational problem that creates a need for an innovation (*agenda setting*); this leads to the creation of an innovation that addresses this problem (*matching*). The initiation phase leads up to the *decision* to (non-) adopt an innovation. The implementation phase is initiated only if the decision to adopt is made. During this phase, the innovation is modified or reinvented to fit the organization (*redefining or restructuring*). Hereafter, the innovation is put into use throughout the organization, such that the use of the innovation becomes clearer to the members of the organization (*clarifying*). The process concludes with the *routinizing* of the innovation in which it has become so ingrained in the regular activities that the innovation loses its identity (ibid.). Most studies on technology adoption limit their reference to DoI to the 'perceived innovation characteristics' (e.g. relative advantage, compatibility, complexity), a set of variables which describe the innovation and affect the individual's innovation decision process (Molinillo and Japutra, 2017; O'Donnell, 2010; Rogers, 2003; Salahshour Rad et al., 2018; Siew et al., 2020). These innovation characteristics are often mapped to the "technology" dimension of the TOE framework; TOE and DoI are frequently used in conjunction to study technology adoption at the organizational level (Oliveira and Martins, 2010). The TOE framework is not concerned with the nature of the innovation decision-making itself, but the contextual factors that affect it. It further expands the contextual factors beyond the technology characteristics and includes factors related to the adopting organization itself (e.g. size, available resources, communication processes, technological readiness, top management support), and to its environment (e.g., competition, regulation, third party sponsorship, customer readiness) (DePietro et al., 1990; Molinillo and Japutra, 2017; Oliveira and Martins, 2010; Yoon and George, 2013). The abovementioned factors used in DoI and TOE have been extended by audit-specific factors, such as the complexity of the audit client's IT (Li et al., 2018; Rosli et al., 2012; Siew et al., 2020), the support of regulators and professional bodies and the encouragement through standards (Li et al., 2018; Siew et al., 2020; Widuri et al., 2016), the commitment of the audit firm's management (Li et al., 2018; Rosli et al., 2012; Siew et al., 2020), the technological competence and resources within the adopting audit firm (Li et al., 2018; Rosli et al., 2012; Siew et al., 2020; Widuri et al., 2016), the existence of IT support staff (Li et al., 2018; Widuri et al., 2016), and the fit between the technology and audit task (Rosli et al., 2012; Widuri et al., 2016).

II.2.2 Advanced data analytics in auditing

While the usage and, hence, the research on the adoption of general IT, CAATTs, and GAS is already at a very advanced stage, the growing body of research dedicated to the application of ADA to auditing remains rather explorative in nature. Prior research has investigated potential areas of application for ADA, the drivers and inhibitors to adoption, as well as the potential impact on the profession if adopted. Apart from conceptual contributions (Alles and Gray, 2016; Alles, 2015; Appelbaum et al., 2018; Cao et al., 2015; Chiu and Jans, 2019; Huang and Vasarhelyi, 2019; IAASB, 2016a; Issa et al., 2016; Jans et al., 2013, 2014; Moffitt et al., 2018; Sun, 2019), there have been several

empirical enquiries that have explored the adoption of ADA in auditing of which most either focused on big data analytics or data analytics (Barr-Pulliam et al., 2020; Dagilienė and Klovienė, 2019; Eilifsen et al., 2019; Haddara et al., 2018; Hampton and Stratopoulos, 2016; Al-Htaybat and Alberti-Alhtaybat, 2017; Manita et al., 2020; Michael and Dixon, 2019; Rose et al., 2017; Salijeni et al., 2019).[2] ADA is considered as a means to increase audit quality as well as efficiency (Dagilienė and Klovienė, 2019; Manita et al., 2020; Salijeni et al., 2019). Increased audit quality is commonly associated with the volumes of data that are analyzed in an audit, moving beyond sampling to the analysis of entire populations or even unstructured data (Manita et al., 2020; Salijeni et al., 2019). Hampton and Stratopoulos (2016) show that the use of ADA increases the confidence of auditors in an audit opinion. Apart from audit quality, the technical ability to access client data remotely in conjunction with standardization efforts provides the possibility of achieving efficiency gains (Salijeni et al., 2019). This enables auditors to focus on more complex and higher valued-added tasks (Manita et al., 2020; Salijeni et al., 2019), thereby ultimately leading to a more relevant audit (Manita et al., 2020). According to the results of Manita et al. (ibid.) and Salijeni et al. (2019), the integration of ADA is likely to enable audit firms to introduce additional, possibly non-audit, service offers.

From the evidence presented by Eilifsen et al. (2019) and Salijeni et al. (2019), it can be concluded that the current adoption of ADA remains limited, despite the potentials mentioned above. However, the adoption of ADA by audit firms is a complex matter that both affects and is affected by several factors that are related to audit firms, their clients, and the institutional environment surrounding audit firms (Dagilienė and Klovienė, 2019; Eilifsen et al., 2019; Haddara et al., 2018; Salijeni et al., 2019).

According to Dagilienė and Klovienė (2019), the characteristics of audit firms —such as size, structure, and the existence of a data-driven strategy— and the availability of professionals with ADA experience affect the adoption of ADA. With regard to the latter, Salijeni et al. (2019) and Haddara et al. (2018) identify a lack thereof in audit firms, which they consider an inhibitor to ADA adoption. ADA skills can be cultivated through training, which has been shown to increase the use of ADA by auditors (Hampton and Stratopoulos, 2016). In addition, Manita et al. (2020) argue that ADA adoption generally pushes auditors toward developing more technological skills and foster a culture of innovation in audit firms. Apart from the skill set of auditors and ADA professionals, Haddara et al. (2018) identify challenges to ADA adoption that are related to the technological preparedness for ADA in audit firms, such as a lack of hardware infrastructure and potential issues with data control and security as well as issues with data integration and storage. Additional challenges in the adoption of ADA by audit firms are the possibility of encountering high numbers of "false positives," which are data points that are falsely identified as requiring the attention of an auditor, as well as the apprehension of auditors regarding the growing influence of data scientists in the auditing practice (Salijeni et al., 2019).

Since auditing is a client-facing business, the characteristics of audit clients and the auditor-client relationship are also relevant factors for the adoption of ADA. According to Dagilienė and Klovienė (2019), the client's size, business model, industry sector, ownership

[2] A tabular summary of the empirical studies cited here on ADA adoption in the audit domain can be found in A.2.

structure, and level of use of technology affect the use of ADA by auditors. Eilifsen et al. (2019) found that, if used, audit firms would use ADA mostly for clients with integrated IT systems and in newly acquired engagements. Further, such clients would have expectations from auditors regarding their use of ADA, which affects its adoption (Hampton and Stratopoulos, 2016). Simultaneously, audit clients are reportedly reluctant to share their data with audit firms due to concerns regarding the auditor's motives and data security (Salijeni et al., 2019). However, the clients also benefit from ADA adoption, as it increases the transparency of the audit for the clients (ibid.) and strengthens the role of an audit as a corporate governance tool (Manita et al., 2020). Further, Salijeni et al. (2019) report technology overspill effects, as audit clients observe the benefits of software used by the audit firm and intend to adopt it.

The market for audit services, regulatory policies, and professional standards form the institutional environment in which audit firms operate. This environment also affects the auditor's adoption of ADA (Dagilienė and Klovienė, 2019; Eilifsen et al., 2019; Salijeni et al., 2019). The analysis by Dagilienė and Klovienė (2019) reveals that the sharp competition within the audit market motivates the use of ADA. Salijeni et al. (2019) indicate that ADA is also referred to as a tool employed to comply with regulatory requirements. From the theory constructed by Eilifsen et al. (2019), the authors argue that the limited use of ADA is likely to persist until it is completely accepted by standard-setters (ibid.). Moreover, the effect of professional standards on the adoption of ADA has been discussed in recent literature, but their effect on ADA is not yet clear. Salijeni et al. (2019) indicate that there is a lack of guidance regarding the use of ADA in professional standards, which has been perceived both as an opportunity for ADA and an inhibitor. The Data Analytics Working Group of the International Auditing and Assurance Standards Board (IAASB) considered the International Standards for Auditing (ISA) technology-agnostic:

> "[. . .] ISAs do not prohibit, nor stimulate, the use of data analytics" (IAASB, 2016a).

In December 2019, the IAASB issued a revised version of the ISA 315 "Identifying and Assessing the Risks of Material Misstatement," which emphasizes the importance of technology (Brown et al., 2018; IAASB, 2019b). The standard refers to "Automated Tools and Techniques," which auditors can refer to when performing audit procedures (IAASB, 2019b). The definition of this term is intentionally broad, as it includes emerging technologies, such as AI and RPA, in addition to data analytics (IAASB, 2019a). This development manifests the relevance of ADA technologies for the profession. Apart from this, the revision of ISA 315 does not touch on the general audit approach; the audit risk model remains the methodological basis for financial statement audits. Further, the usage of these technologies remains optional for auditors and functions as an addition to traditional audit procedures if utilized. This may slow down the adoption of such technologies in audit firms.

II.2.3 Use of IT specialists in financial statement audits

As noted above, previous studies mention the lack of professionals with ADA skills as an inhibiting factor for the adoption of ADA (Haddara et al., 2018; Salijeni et al., 2019). The professional standards require auditors to refer to domain specialists if they do not possess the required expertise outside of accounting and auditing to obtain sufficiently

appropriate audit evidence (IAASB, 2009). In the case of the client's information systems, IT specialists—or IT auditors—fill this skill gap. The required IT expertise and, therefore, the necessity for IT auditors, increases with complexity of the client's IT environment (Curtis, Gregory Jenkins, et al., 2009). They are mainly called upon for the testing of IT-related controls as part of the risk assessment, but they are also increasingly involved across different aspects of financial statement audits (Bauer and Estep, 2014; Bauer, Estep, and Malsch, 2019; Boritz et al., 2017; Otero, 2015). IT auditors are usually required to have a solid understanding of ERP systems, which enables them to better identify risks and select relevant controls in IT processes than financial auditors (Hoitash et al., 2008; Hunton et al., 2004; Tucker, 2001). IT auditors (and specialists in general) are more likely to question client positions (Boritz et al., 2017) and their involvement in financial statement audits is positively associated with the identification of manipulated financial information in IT systems (Otero, 2015). Despite the value they can add to audit engagements, they remain frequently underused in audit engagements (Boritz et al., 2017; Janvrin, Bierstaker, et al., 2009). The main reasons are cost restrictions, communication challenges between specialists and auditors due to differing backgrounds, and their involvement not being perceived as useful by audit teams (Boritz et al., 2017; Otero, 2015). While the involvement of IT specialists in audit engagements can also be considered as an (non-) adoption case itself, prior research has also examined the effect of the support of IT members on the adoption of audit technology by auditors (Vasarhelyi and Romero, 2014). Vasarhelyi and Romero (ibid.) found that IT auditors can support the adoption of audit technology by audit teams if they have a background in accounting, thereby enabling them to identify how technology could help the auditors. While the perceived lack of ADA skills inhibiting ADA adoption has been noted (Haddara et al., 2018; Salijeni et al., 2019), this notion seems at odds with the prevalent practice in audit firms to employ IT specialists (Bauer and Estep, 2014; Bauer, Estep, and Malsch, 2019; Boritz et al., 2017; Otero, 2015). Further investigation is warranted to explore the connection between these skill sets further.[3]

In this section, different research streams are reviewed. We identify several research gaps in the literature that we aim to address in this paper. First, the literature on ADA adoption mainly discusses the potential impacts of ADA on the auditing practice along with the factors that affect ADA adoption in the auditing industry. How this adoption happens—that is, the underlying process that leads to a successful adoption—is widely ignored. Further, the research on this topic remains rather fragmented and lacks coherence, which might be attributed to its explorative nature. Therefore, a theoretical framework can help to guide future research. Second, the literature on general technology adoption in auditing is foremost concerned with variance theories, in which one dependent variable or outcome (technology adoption) is affected by several independent variables or factors, thereby implying an invariant relationship between the factors and the outcome (Markus and Robey, 1988). Variance theories are static in nature and treat the underlying adoption decision process as a black box (Langley, 1999; Markus and Robey, 1988). Uncovering this process can help contextualize the studied factors, such that the mechanism by which they affect the adoption can be further understood. Third, there is a missing link between the inclusion of IT specialists in auditing and the apparent lack of ADA professionals, which inhibits ADA adoption.

[3]This paragraph has been slightly altered with respect to the published version of the paper to improve clarity and readibility.

II.3 Methods

The goal of our study is to assess how ADA is being adopted in audit firms and which factors affect the adoption process. According to the framework of Gregor (2006), this corresponds to a theory of explanation. Common to theories of explanation are process theories, which is what we aim for in our research. As auditing firms are still at the initial stages of engaging with ADA, there is a lack of quantitative data on this subject matter that can be exploited for research purposes (Alles and Gray, 2016). This calls for explorative qualitative approaches such as interviews and case studies (ibid.). Our approach is to derive a theory using the grounded theory method given by Corbin and Strauss (1990). The grounded theory methodology usually leads to process theories that are high in accuracy (in terms of being true to the data) but rather low in terms of generalization (Langley, 1999). We collected data from interviews with industry experts, as we aim to establish a broader understanding of the entire subject and not just one specific case. The data was analyzed as we collected it, which enabled us to shift our research focus for the upcoming interviews to questions that we encountered while analyzing the data, following the idea of theoretical sampling (Corbin and Strauss, 1990). Further, we followed the recommendations for qualitative interviewing given by Myers and Newman (2007) and attempted to avoid the pitfalls and problems that the authors described.

II.3.1 Data

Between June 2018 and June 2019, we conducted a total of 15 interviews. Each interviewee was interviewed separately, except for IP15 and IP16, who were interviewed together, as they are from the same organization. We chose semi-structured interviews as the interview method (ibid.). This enabled us to further pursue interesting cues when they came up in an interview while still providing a structure. We first developed an initial catalogue of questions that we derived from the related literature. As we progressed in our understanding of the phenomenon at hand, we generated new questions and, thus, continually developed the catalogue. Further, we partially adapted the catalogue to the experience or expertise of the interview partner if it was necessary and beneficial for the study. We acquired interview participants through our professional and personal networks as well as through internet research and cold contacting. The interviews were opened by assuring the interview partner of the confidentiality of the collected material and the non-disclosure of personal information and organizational affiliation; in addition, approval was obtained for recording the interview. If the approval was granted (in 11 cases), the interview was recorded and subsequently transcribed. If the approval was not granted or the circumstances of the interview did not allow for audio recording, notes were taken (four cases). Ten interviews were conducted via phone or Skype and the remainder were conducted in person.

Table II.1 presents the interview participants along with their role(s) in their respective organization, their technical background, their professional focus within their roles, and their country of residence. We aimed at acquiring interviewees from different educational or technical backgrounds, hierarchical levels, and organizations of different sizes. This enabled us to triangulate evidence and explore the phenomenon from different perspectives, as suggested by Myers and Newman (2007).

Germany (G) is heavily represented in our sample, with fourteen of our interview partners

working in Germany, one in the Unites States of America (U.S.), and one in Switzerland (CH). Germany is an interesting environment to study the adoption of ADA. Statutory external audits have a long tradition in Germany and have been encoded in commercial law since 1931 (Köhler et al., 2007). In addition to the information function for stakeholders and capital markets, external audits serve an important monitoring and control function in the German two-tier corporate governance system: The supervisory board, not the executive board, appoints the auditor and is the addressee of the long form audit report, which is issued in addition to the regular audit report (ibid.). Apart from this local characteristic, the German audit market and its institutional environment are highly internationalized. As a EU member state, the directives and regulations of the European Commission directly apply to Germany. These are aimed at harmonizing the European regulation to the post-Sarbanes-Oxley Act regulation in the U.S. through strengthening auditor independence and introducing public oversight in addition to the profession's self-administration (EU, 2006, 2014a,b; Humphrey et al., 2011). They also made the adoption of the ISA mandatory for all statutory audits in the EU (Fédération des Experts Comptables Européens, 2015), including the disclosure of key audit matters (IAASB, 2016b). In Germany, the ISA are adapted into local standards that take the local legislation into account (Fédération des Experts Comptables Européens, 2015; Köhler et al., 2007), which is conducted by the professional body "Institut der Wirtschaftspruefer" (IDW). The IDW and the chamber of auditors (Wirtschaftsprueferkammer, WPK), which is the other professional body, are members in the overarching professional network Accountancy Europe, just as the public auditor oversight board is a member of the International Forum of Independent Audit Regulators.

Most of the interviewees occupy leadership positions. The most frequently represented technical backgrounds are business (accounting, auditing, and tax) and information technology (IT). Although auditors are most affected by it, the adoption of ADA in auditing is a highly interdisciplinary phenomenon, which involves individuals with different professional backgrounds. Therefore, we followed the advice given by Alles and Gray (2016) and did not constrain ourselves to external auditors as interview partners but also regarded interview partners from other types of organizations, such as professional bodies, regulatory bodies, and software providers. With regard to audit firms, we interviewed individuals from the Big Four as well as mid-tier audit firms. Similar to previous research on the adoption of ADA, this study began with an explorative intention; incorporating a professionally diverse, multidisciplinary sample of interviewees enabled us to consider the phenomenon from different perspectives. This also enabled us to explore the aforementioned availability of ADA professionals in audit firms and the related required skill sets.

II.3.2 Analysis

In order to construct a coherent theory from the data collected through the interviews, the grounded theory methodology, as described by Corbin and Strauss (1990), was applied. The analysis was accompanied by a phased literature review (Thornberg and Dunne, 2019; Urquhart, 2012). After a first *uncommitted* review, we began with an open coding approach that gave room to consider all possible connections in the data, without a specific kind of theory in mind. During this process, 377 codes were assigned to 1,164 text passages. In the fashion of a *delayed* or *integrative* literature review in grounded theory (Urquhart, 2012), which is informed by the data, we sought theoret-

Table II.1: Interview Participants

Participant #	Role	Background	Organization	Focus	Country
IP1	Data Scientist	STEM	Big Four	ADA	G
IP2	Data Scientist	Business, IT	Software Provider	ADA, Internal Audit	G
IP3	Partner/Director	Business	Big Four	Audit Digital Innovation	G
IP4	Partner/Director	Business	Big Four	Audit Digital Innovation	G
IP5	Partner/Director	Business	Mid-tier Audit Firm	Audit	U.S.
IP6	Head of Department	Business	Mid-tier International Company	Internal Audit	G
IP7	Partner/Director	Business, IT	Mid-tier Audit Firm	IT Audit, Audit Digital Innovation	G
IP8	Partner/Director	Business, IT	Big Four	IT Audit, Audit Digital Innovation	G
IP9	Data Scientist	IT	Mid-tier Audit Firm	IT Audit, ADA	G
IP10	Subject Matter Expert	Business	Professional Body	Audit, Audit Digital Innovation	G
IP11	Manager	IT	Big Four	Robotic Process Automation, Audit Digital Innovation	G
IP12	Subject Matter Expert	Business	Professional Body	ADA, Audit Digital Innovation	G
IP13	Audit Associate	Business	Big Four	Audit	G
IP14	Partner/Director	IT	Big Four	PM, Audit Digital Innovation	CH
IP15	Subject Matter Expert	IT	Auditor Oversight Board	(IT-) Audit	G
IP16	Subject Matter Expert	Business	Auditor Oversight Board	Audit	G

ical frameworks to which we could relate our emergent theory. Relating the emergent theory to prior research and extent formal theories can support the researcher in developing concepts and improves the analytic generalizability of the emergent theory (ibid.).

The results presented in this paper were obtained through multiple data-theory iterations. We first aimed for a causal theory that would explain the differences in technology adoption in audit firms, with the technologies being the unit of analysis. In this theory, the actual innovation process remained a black box, whose output is a binary variable describing a technology adoption or a rejection (non-adoption) of the technology. We found our theory to be relatable to the TOE framework by DePietro et al. (1990), as it addresses the adoption of technologies on an enterprise level while explicitly considering contextual factors within the adopting company as well as its environment. To this end, we categorized our concepts along the axis of the TOE framework. During this step, we concluded that the organizational concepts relate to various organizational units and would affect the technology adoption at different points in time. Therefore, the previous objective was revisited and we decided that the adoption process must be made explicit along with various organizational units. In the TOE framework, the contextual factors affect the so-called "technological innovation decision making" (ibid.). For the further analysis, we considered this "decision-making" to be a process and referred to the innovation decision process in organizations described by Rogers (2003). Drawing upon the DoI theory, we developed concepts that relate to the ADA adoption process—with the audit firm becoming the unit of analysis—as well as concepts relating to contextual factors affecting this process. As the concepts were developed, codes that did not fit the theory were excluded (selective coding). Out of the 377 initial codes, 157 codes were related to the phenomenon of ADA adoption and, thus, selected for further analysis. These codes were merged thematically into 61 codes, which in turn were abstracted into 22 first-order concepts and 8 second-order concepts. The first- and second-order concepts form the building blocks of the theory. Figure II.1 depicts the data model of our results. The arrows indicate the direction of abstraction, thereby indicating which first- and second-order concepts were derived from their respective subclass. Descriptions of the concepts are provided in the results section.

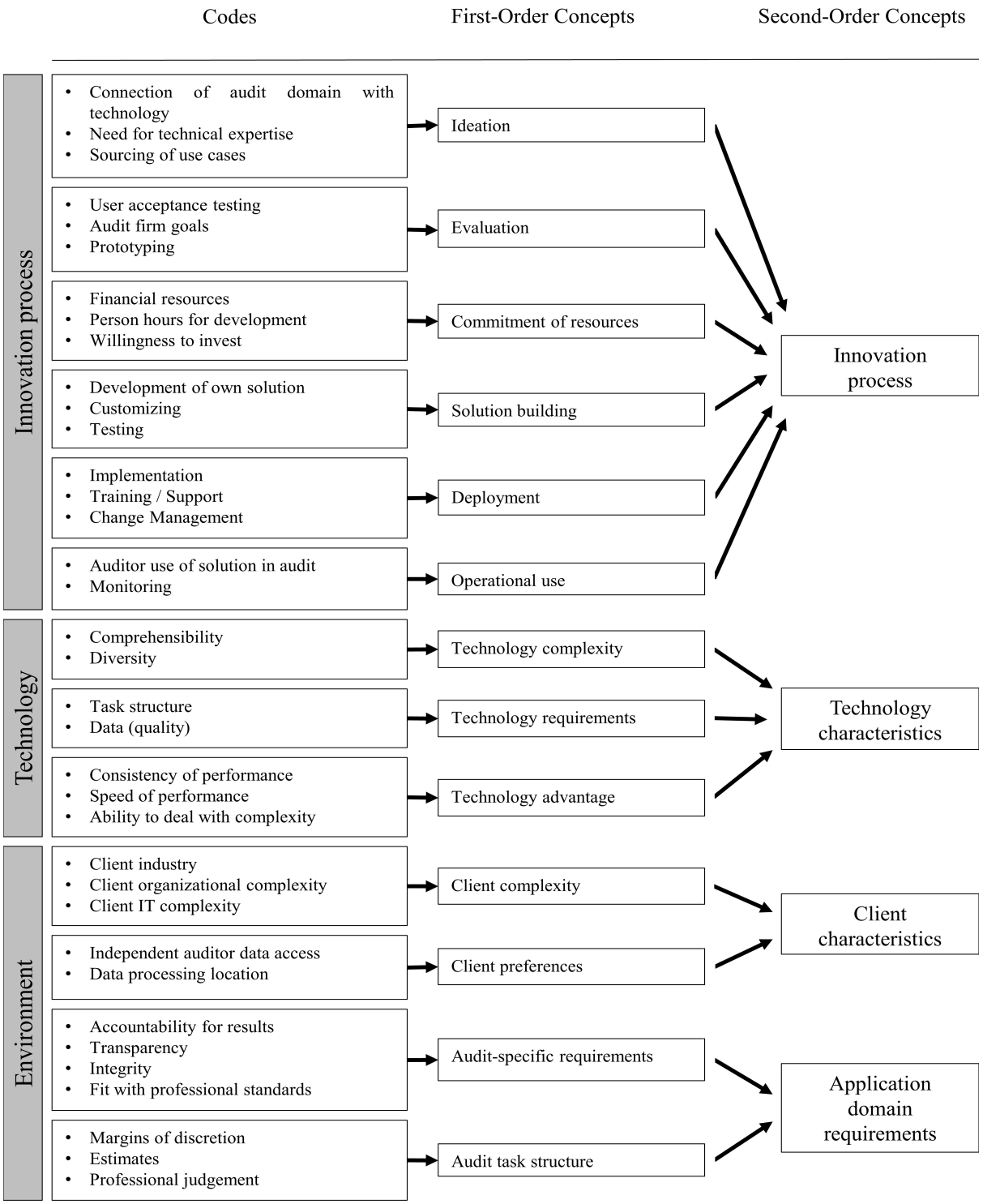

Figure II.1: Data Model

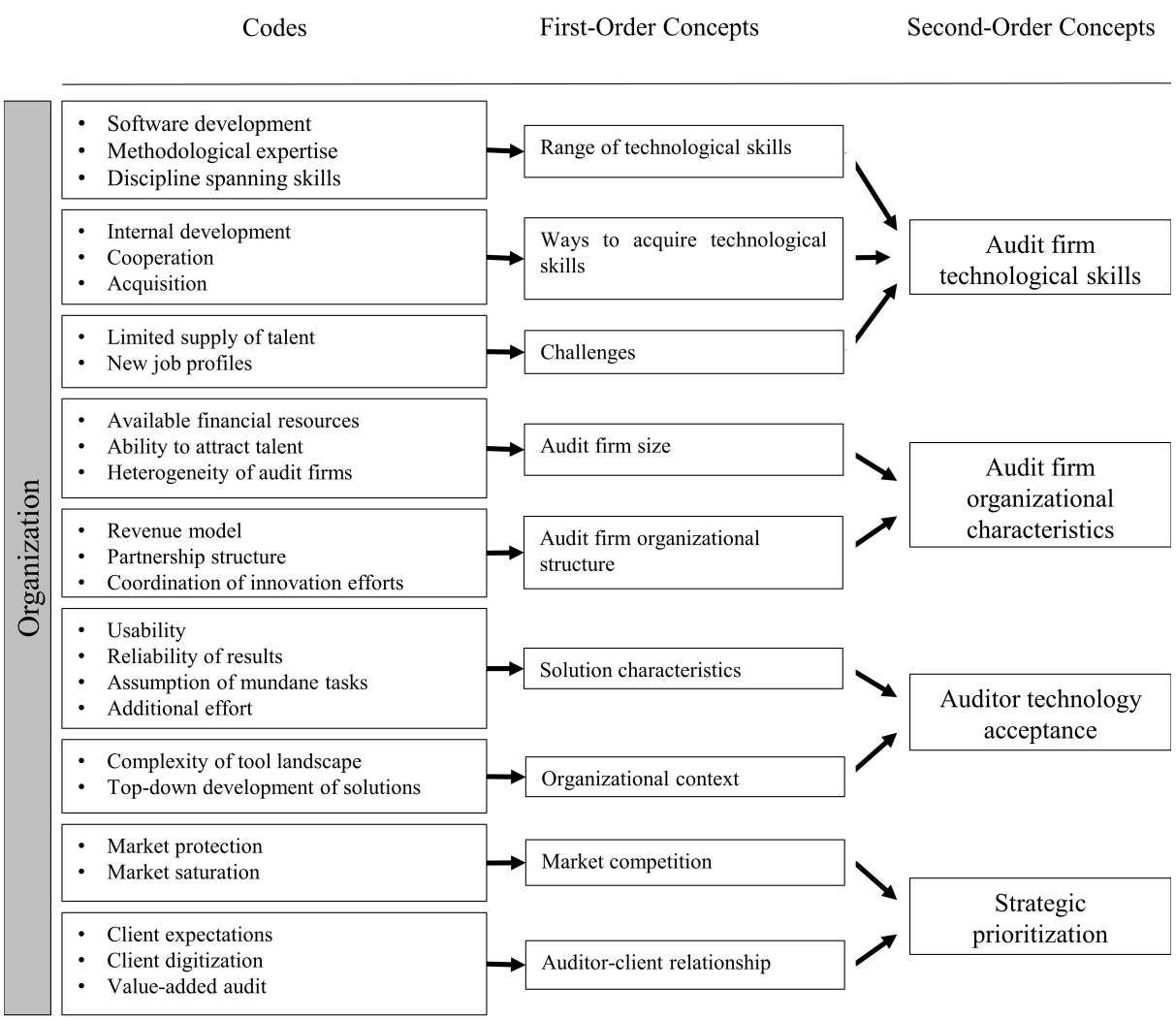

Figure II.1: Data Model (cont.)

II.4 Results

From the interviews, we found that the phenomenon of audit digitization has two dimensions. One dimension is the digitization of audit tasks that relate to the management of audits, such as the organization and delegation of work packages, communication with the client, or the documentation of the audit. From the interviews, we concluded that the digitization of the management of audits is already at an advanced stage. The second dimension is the digitization of the "fieldwork"—that is, the assessment of risks—and the obtaining of audit evidence used to support the audit opinion. The digitization of the latter has a greater impact on the profession and continues to have vast potential. However, our results further indicate that the digitization of the audit fieldwork is progressing rather slowly, and despite the potential of ADA that is being discussed in the literature, its adoption is rather limited (see A.1). From the ADA technologies that are expected to have an influence on auditing practice, we found that mainly RPA and PM are being adopted on a larger scale. This is foremost driven by the Big Four audit firms and their close followers, which invest in the in-house development of solutions. The theory we present in this section explains

1. how audit firms adopt ADA,
2. how this adoption process is affected by contextual factors relating to the audit firms themselves, their environment, and the technology to be adopted,
3. how this affects the outcome of the adoption process.

Figure II.2 depicts a model of the theory. Central to the theory is the adoption process, which comprises several activities. In the model, the process is described within the box and the rounded squares represent the activities. These activities are performed by various organizational units: audit teams, innovation teams, and management. This is represented through swim lanes in the model. Activities breaking through swim lanes indicate that more than one organizational unit is involved in this activity. The **innovation teams** within the audit firm are tasked with the conceptualization and development of ADA-backed solutions and typically consist of data and computer scientists as well as technologically inclined auditors or consultants. The **management** comprises the audit firm's partners and directors involved in the strategic decision-making of the firm. They decide on the allocation of resources for ADA adoption. The **audit teams** are the users of the provided solutions. They employ ADA-backed solutions or resort to using traditional non-ADA audit procedures, depending on whether they see the solutions fit for implementation, given the circumstances of the audit engagements. The activities in the process are performed in a sequential order; however, there can be feedback from one activity to the previous one. In the model, the sequential order is described though arrows—the dashed arrows denote feedback (FB) loops among activities. The activities within the process are affected by contextual factors. These are represented by square boxes outside of the process and are connected via arrows to the activities and organizational units, which represent the relationship (R) through which the factors affect the process.

The findings indicate that there exists a gap between the technological and audit domains in terms of knowledge and professional mentality. In successful ADA adoptions, this gap is closed through cooperation between the audit and innovation teams. This gap is primarily relevant in the ideation activity and the diffusion phase of the process. In the

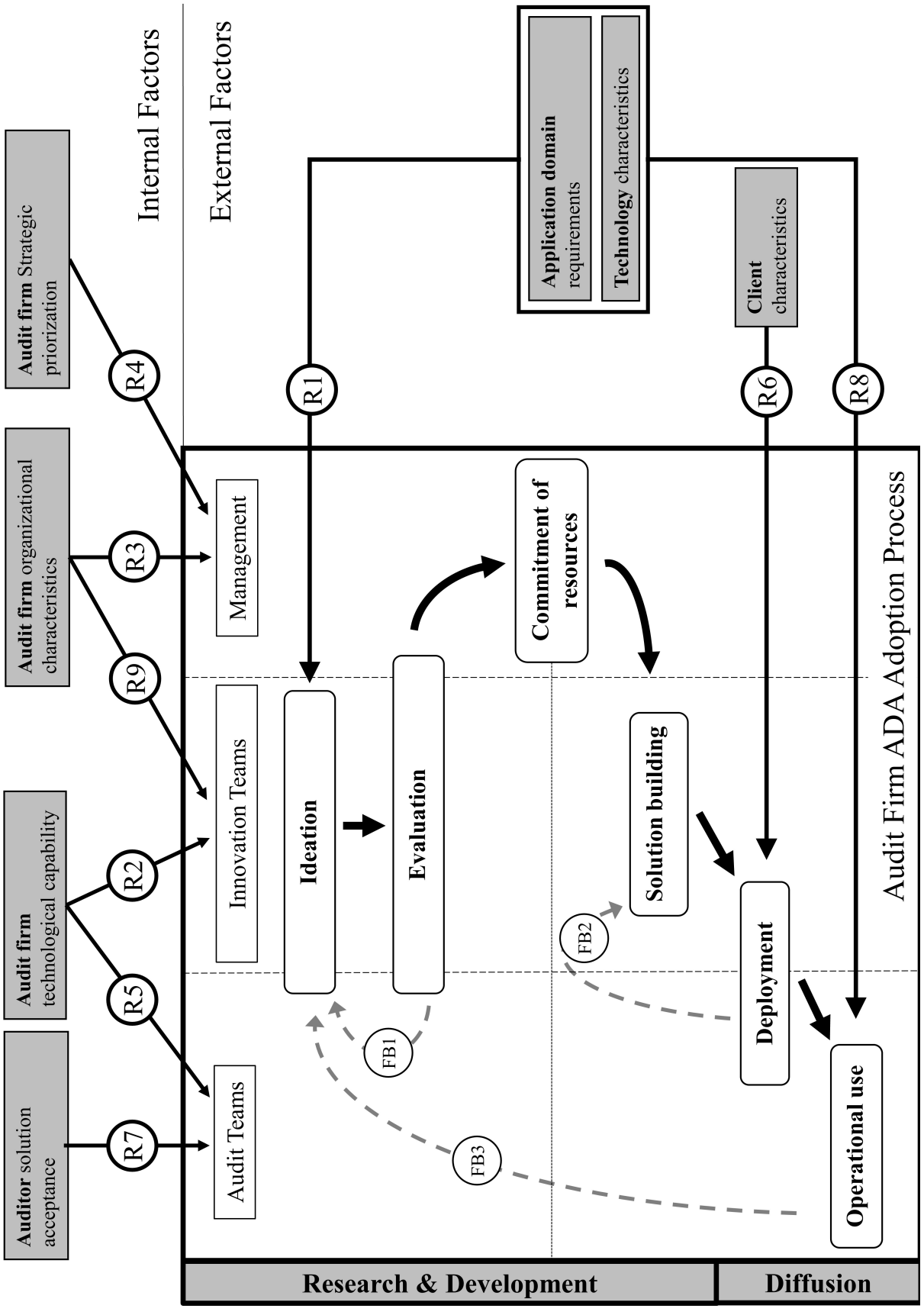

Figure II.2: Process Theory of ADA Adoption in Audit Firms

ideation activity, bridging the knowledge gap between the audit and technological domain is required to identify use-cases for ADA. The identification of use-cases requires knowledge of both the audit domain and the technologies, as well as discipline-spanning skills to connect them. In the diffusion phase, the solution is handed over to the auditors. Here, it is crucial for the solution to be accepted by the auditors so it can diffuse into operational use. The gap in the operational use activity relates to the different mentalities—auditors and data/computer scientists differ in terms of their way of working and approaching challenges. Innovation teams can positively affect the auditor's acceptance of a solution by adapting it's design to the auditor's preferences. On the other hand, auditors take interest in ADA and its application to the audit domain. According to our interviewees, audit firms provide auditors the freedom to develop themselves to take on more technology-oriented roles. These auditors join innovation teams or help in the deployment of solutions as trainers and lead users. In the case of RPA, audit teams also actively approach innovation teams with proposals for use-cases and joint ideation workshops. In this manner, the audit teams support the ideation of solutions through audit knowledge. However, RPA exhibits a lower degree of complexity and, therefore, is more comprehensible for users of the technology. AI would be a counter-example in terms of technology complexity, which is an aspect for why audit firms are struggling to find use-cases. Therefore, from the fact that auditors can be involved in ideation activities for RPA adoption, we hypothesize that there is a relationship between an auditor's technology capability and his/her contribution to the ideation of ADA solutions. This implies that increasing the technological capability across audit teams would not only benefit audit firms through more potential lead users and trainers but also in the ideation of ADA solutions.

II.4.1 Process

The adoption process consists of six activities: Ideation, evaluation, commitment of resources, solution building, deployment, and operational use. Ideation, evaluation, and solution-building constitute the research and development process phase through which a solution is developed from a use-case. A broader adoption of the technology is only achieved if the solution is diffused into operational use. In order to ensure this, the solution must be made available to the users and the users must be enabled to employ it, which is achieved in the deployment activity. Within the ADA adoption process, the activity of the commitment of resources represents a phase gate. The solution-building and subsequent activities only take place if resources are allocated for them.

Ideation

Ideation is the first activity in the technology adoption process. In this step, a possible use-case for a technology is identified. Use-cases are application scenarios for a technology in which the technology is mapped to the application domain (how it is being used and to which end). Ideation is a non-trivial task and is still considered a challenge for audit firms (see A.2). The ideation activity is primarily performed by the innovation teams and can be assisted by the audit teams. It is affected by the technology characteristics (foremost complexity and requirements) and the application domain requirements (R1): The technology advantage motivates the use of a method or technology. A highly complex technology increases the technological capability—particularly the methodological

knowledge—required by the audit firm's innovation team to identify use-cases, as they need to be able to navigate the possibilities provided by ADA. The more complex a technology or method is, the more effort must also be put into making it understandable for different stakeholder groups, such as business users and regulators.

The requirements of the application domain and the technological requirements determine the technical and technological frame for the ideation step (R1). The use-case must comply with professional standards and be relatable to business users and regulatory bodies. However, as mentioned earlier, professional standards can also motivate the use of ADA. Simultaneously, the following prerequisites of a technology or method to be used must be met: The structure of a task can be improved through internal standardization and homogenization by the audit firm, but the data quality and data volumes available are, in most cases, beyond the influence of the audit firm and depend on the use-case. Since it is primarily the innovation team that is concerned with the ideation phase, their technological capability is crucial, both in terms of methodological knowledge and discipline spanning knowledge (R2). The greater the methodological knowledge within the team, the better it is able to deal with complex technologies or methods. Further, discipline-spanning skills help to bridge the knowledge gap between auditing and technology, which—according to our interviewees—is crucial to close the aforementioned gap (see A.3).

Evaluation

Upon identifying the use-case, it is evaluated against the goals of the audit firm; efficiency, effectiveness and relevance. Technologies and methods from the ADA spectrum present promising possibilities to address these goals in general. Efficiency can be realized through automation as well as reducing or omitting substantive procedures based on ADA. Audit assurance can be increased through expanding the scope of analyzed data and through increased transparency, e.g., PM enabling the auditor to quantify the number of process instances being conformant to a certain standard process or the exact number of process instances that violate control tests. Further, additional information for the client can be generated as a by-product from the (better) analysis of client data. They help auditors to deliver a value-added audit, which contributes to the audit's relevance. A prototype of the latter solution may be built, which can be used for evaluation. If the prototype is not sufficiently convincing in the evaluation, another ideation cycle can be triggered (FB1). This evaluation step is mostly performed by the innovation teams, but can also involve auditors (for user acceptance testing) and management in order to ensure alignment with goals. The stronger (in the sense of his contribution to the audit firm's goals) a use-case is, the higher is the likelihood that resources are provided for the development of a solution, which is the next step in the technology-adoption process.

Commitment of resources

The development of a solution from a technology or method requires both financial and human resources. Financial resources are required for hardware, software, and employee training, while human resources are required for development and further actions. If the audit firm management is not willing to invest in these resources, there will be no development or only minimal development. Therefore, this activity represents a phase gate in the adoption process. Committed resources can be existing resources as well as

resources that must be acquired first, depending on the pre-existing technological capabilities. Further, the commitment of resources involves a make-or-buy decision. In an inter- or intra-organizational setting, where solutions are created by one organizational unit to be employed by another (see the solution-building activity), the employing organizational unit can decide to assign resources to the acquisition and/or customization of the solution. This commitment of resources is affected by the size of the audit firm, both in terms of financial resources and the ability to attract talent (see A.4).

Bigger firms have more financial resources available that can be allocated for ADA adoption efforts as compared to smaller firms (R3). Further, larger auditing firms have a greater ability to attract talent to improve their technological capabilities. The organizational structure of an audit firm determines the frame for the process step. Both the revenue model as well as the partnership structure affect the management's willingness to invest. The strategic prioritization of technology adoption in the audit firm affects the management's willingness to invest (R4). The strategic prioritization correlates with the size of the audit firm, as larger audit firms usually deal with larger and more digitized clients that have a different set of expectations as compared to smaller audit firms. The same applies to the competitive situation, as the Big Four and Next Ten, in particular, are currently facing a highly competitive market situation.

Solution-building

During the solution-building process, the use-case or prototype is turned into a solution. In this context, the term solution-building refers to both the in-house development of a solution or the customization of an existing solution, depending on the make-or-buy decision taken in the commitment of resources activity. Developing a solution requires technology capabilities that differ from the capabilities required to identify and evaluate use-cases (R2): It requires capabilities in the areas of software engineering, information systems (e.g., ERP systems), and business processes. Further, the activity is also affected by the organizational structure of the audit firm (R9): Audit firms adopt different approaches to coordinating their innovation efforts. Solution-building takes place in both centralized and decentralized settings. In a centralized setting, one organization (e.g., software company) or organizational unit (e.g., an audit firm's central innovation unit or service center) develops solutions and deploys them for other organizational units (e.g., an audit firms' national subsidiaries). In a decentralized setting, the development takes place in multiple organizational units. In a combination of both, the software can be deployed by one organizational unit and then be customized by another organizational unit to fit its needs.

Deployment

When a version of the solution has been developed, it can be deployed. During deployment, the solution is tested, implemented, and pushed into practice. If t becomes evident during testing that the solution still requires refinement, another development iteration can take place (FB2). Pushing the solution into operational use is challenging, as this may require change management depending on the nature of the solution. For example, in the case of PM, it was described that this new technique changed the manner in which an auditor approaches the analysis of business processes by replacing the walkthrough interview with

data analysis.

The deployment activity benefits from the technological capabilities of audit teams (R5), as technology-savvy auditors can be involved in the deployment process as lead users and/or trainers and support other auditors when employing the solution. The deployment is further affected by client characteristics (R6). When implementing the solution, the complexity of the client in terms of the different information systems and data models must be considered if client data is to be accessed by the solution. Further, the greater the organizational complexity of the client company, the more scalable the solution must be in order to be applied across the client company—for example, if the client has affiliates across different branches of the industry.

Operational use

The operational use marks the last step in the technology adoption process. After the solution has been tested and is approved for release, auditors can go on to use it. The solution is either handed off by the innovation teams to regular IT departments for monitoring, or further developments (e.g., additional features) are initiated in by the innovation team. In the latter case, the adoption process is reiterated. As an audit is seldom a production-like process, the audit teams must be able to tailor their approach to the audit around the client. For this reason, audit firms do not coerce auditors into using the provided solution or do so only to a limited extent. Therefore, the solution must be adopted by the individual auditor; hence, the auditor's solution acceptance affects the adoption and diffusion (R7). The solution acceptance correlates with the characteristics of the underlying technology (complexity and advantage) and the application domain characteristics (R8). The advantage of a technology motivates the operational use of a solution, particularly with regard to the assumption of mundane tasks. The complexity of the technology can discourage auditors from employing a solution if it is not absorbed by the usability of the solution. The application domain requirements may motivate utilization (e.g., ISA 240 and Journal Entry Testing) but can also have a negative effect, as it may require additional documentation to be prepared by the auditor. From the interviews, we learned that in the case of RPA, auditors likely directly propose the tasks for automation. Mimicking human interactions with software, RPA is rather comprehensible for the business user. Therefore, with a complex technology that is absorbable by the auditors given their technological capabilities (R5), it becomes possible to source use-cases for the technology directly from them.

II.4.2 Contextual factors

The process presented above is affected by several contextual factors that are related to the audit firms (internal factors) and their environment (external factors). The internal factors vary among audit firms, but also characterize how audit firms are different from companies in other industries. The strategic prioritization of technology adoption describes how much importance a firm assigns to its digitization efforts. Central to the strategic prioritization is the relationship between the audit firm and its client as well as the competition amongst audit firms. Audit firms pursue digitization to improve the effectiveness and efficiency of financial statement audits and also seek to maintain and improve their relevance. The latter is related to the audit firm's ability to conduct audits and deliver value-added audits

to their clients. In the innovation process, these aspects form the basis for evaluation and influence the commitment of resources. The organizational characteristics represent the effects that the firm's size and organizational structure have on the process. Smaller audit firms usually have fewer resources at their disposal and face more challenges in attracting talent with the skills required to adopt ADA. In terms of talent, the audit firm's technological capability is indicative of the technological skills across the organizational units within the audit firm. They enable the staff to conceptualize, develop or customize, and evaluate ADA-based solutions. The end users of any ADA-based solution are the audit teams. The extent to which the individuals in the audit teams are inclined to use a solution provided to them is characterized by the level of technology acceptance among auditors. Their acceptance is influenced by the characteristics of the solutions as well as the organizational context of the solution deployment. A few of the internal factors—for example, the technological capabilities, strategic prioritization, and organizational structure—can be influenced by the audit firm. The external factors lie beyond the influence of the audit firm and relate to the ADA technology, the specifics of the audit domain, and the characteristics of audit clients. Although ADA technologies form a cluster, they still have different characteristics and requirements, which primarily affects the ideation and operational use-activities that are part of the adoption process (R1, R8). The counterpart of these activities are the requirements of the audit domain; the audit task structure determines how applicable ADA is to a specific audit task in the first place, and the audit-specific requirements define the conditions that are to be met for ADA that must be applied in the audit domain. Audit clients exhibit varying degrees of complexity and have differing preferences for data access and processing by auditors. These characteristics affect the deployment activity that is part of the adoption process (R6). Details on the contextual factors are provided in Table II.2.

Table II.2: Contextual factors affecting the ADA adoption process

Contextual Factor	Relationship(s)	Affected Process Activity	Factor Dimensions	Description
Technology Characteristics	R1, R8	Ideation, Operational Use	Advantage	The advantage of technology over human cognitive abilities and/or traditional audit procedures—such as speed and consistency of performance, ability to deal with complexity, or ability to process large amounts of data.
			Complexity	ADA technologies exhibit varying degrees of complexity in terms of comprehensibility and versatility. Versability is the spectrum of possible uses and comprehensibility of the amount of technological knowledge required to understand and employ the technology.
			Requirements	The conditions that must be met to employ the technology: Foremost data quality, availability of (labeled) data, and task structure.
Client Characteristics	R6	Deployment	Complexity	The complexity of the audit client's organization, accounting system, information systems and their underlying data models, which differ across audit clients (see A.5 and A.6).
			Preferences	Audit clients have differing preferences regarding independent data access of auditors and the location of data processing, mostly due to concerns of data ownership and security.
Application Domain Requirements	R1, R8	Ideation, Operational Use	Audit-specific requirements	The professional standards prescribe the general audit appraoch, which was considered as possibly limiting the possibilities of ADA use (see A.7). This includes the audit risk model, which is encoded in the ISA. As auditors are legally accountable for the issued audit opinion, a high reliability of ADA solutions used to collect audit evidence must be ensured (see A.8). Further, the solutions must be transparent to regulators (see A.9).
			Audit task structure	From our interviewees, we learned that unstructured tasks that involve the auditor's professional judgment are considered more difficult to automate than more structured tasks (see A.10). RPA is usually used to automate highly structured tasks, whereas ML can deal with more complex tasks.
Audit Firm Technological Capability	R2, R5	Ideation, Evaluation, Solution building, Deployment, Operational Use	Range of Technological Skills	A range of technological skills is required to develop ADA solutions related to methods, software engineering, knowledge of business processes, and information systems. Further, a few interviewees emphasized the importance of discipline spanning skills to bridge the knowledge gap between auditing and computer or data science (see A.11).

Table II.2: Contextual factors affecting the ADA adoption process (cont.)

Contextual Factor	Relationship(s)	Affected Process Activity	Factor Dimensions	Description
			Ways to Acquire Technological Capabilities	Audit firms resort to different ways of acquiring technologcal capabilities—internal development, hiring, and cooperation with research facilities and specialized companies.
			Challenges	The acquisition of talent is associated with several challenges, such as the increasing dependence of audit firms on technological experts, the "war for talents" (IP15) along with a limited ability of smaller audit firms to attract such talent, as well as the new hiring practices and new career paths that are required to attract this kind of talent.
Audit Firm Organizational Characteristics	R3, R9	Ideation, Evaluation, Commitment of Resources, Deployment	Size	Audit firms are rather heterogeneous in size. The size of the audit firm affects the firm's available resources (both financial and human) as well as its ability to attract talent. Smaller audit firms usually have fewer resources available to allocate for ADA adoption and they struggle with attracting talented indivduals with a technological background.
			Organizational Structure	In most cases, audit firms follow the the partnership model through which shareholders and management are unified; the firm's cashflow affects the partner's earnings, which, in turn, affects the amount of resources devoted to ADA adoption. The revenue model of most audit firms, person hours × fee, does not inlcude development costs for ADA solutions and entails opportunity costs for employees who do not generate billable hours. While these aspects are common among most audit firms, each of them differs in terms of how they coordinate their innovation efforts. Moreover, the innovation teams are embedded differently in the organization—for example, audit-specific or cross-service and national or international.
Strategic Prioritization of Technology Adoption	R4	Commitment of Resources	Market for Audit Services	Interviewees described the audit market as fiercely competitive and simultaneously protected against outsiders. The protection is established through professional titles and the knowledge capital accumulated in audit firms. This made innovation less of a necessity in the past, but competition has lead audit firms to seek efficiency gains and provide better services.

Table II.2: Contextual factors affecting the ADA adoption process (cont.)

Contextual Factor	Relationship(s)	Affected Process Activity	Factor Dimensions	Description
			Auditor/Client Relationship	In addition to effectiveness and efficiency, auditors expect ADA to uphold their standing as experts in all financial matters, as clients increasingly digitize their financial operations. Further, clients have varying expectations from the auditor's use of innovative solutions, depending on how much they are looking to leverage the audit for themselves. They leverage the audit by obtaining additional information that is yielded as a by-product of the audit (see A.12). The ability to provide such a value-added audit can lead to advantages in the marketplace (see A.13).
Auditor Solution Acceptance	R7	Operational Use	Solution Characteristics	Characteristics of ADA solutions relevant for use by auditors are their usability, reliability of their results, the assumption of mundane tasks, and potential additional effort caused by their use from training and/or documentation. Creating usability has been associated by interviewees with bridging the gap between the mindsets of the innovation teams and the auditor (see A.14).
			Organizational Context	A few interviewees stated or implied a resistance from audit teams with regard to solutions that they felt were developed in a "top-down" manner by innovation teams, thereby indicating a disconnect between innovation teams and auditors. Another aspect here is the complexity of the tool landscape available to auditors—audit teams lose track of which tool serves which purpose and feel overwhelmed.

II.4.3 Effects of the contextual factors on the adoption process

In order to illustrate the effects of the contextual factors on the adoption process, the process is instantiated in this section through various configurations. The configurations comprise different sets of manifestations of contextual factors and the resulting manifestations of the process activities. They were derived from the different manifestations of the respective contextual factors along with the different process outcomes we encountered in the data. Therefore, while the configurations do not represent case studies, they are rooted in the data. We illustrate the respective effects of the technology and organizational characteristics on the process, as they represent the most significant adoption gaps encountered in the interviews. One adoption gap is between ADA technologies; RPA and PM are witnessing wider adoption, whereas ML and DL are not. The other adoption gap is among audit firms: From our data, we conclude that the Big Four and their close peers are leading in the adoption of ADA, whereas smaller firms are lagging behind.

Table II.3: Effects of technology characteristics on ADA adoption

	RPA	PM	ML	DL
Technology characteristics				
Advantage	Speed, Consistency	Ability to process large amounts of data, quantification of previously qualitative information	Ability to deal with complexity, automation of less structured tasks possible	Ability to deal with complexity, tasks that require human cognitive skills feasible—for example, processing of unstructured data
Complexity	Low	Mid	Mid-high	High
Requirements	Task structure	Data, Data quality	Data, Data quality	Large data volumes
Effects on adoption process				
Ideation	Use-cases can be sourced from audit teams	Use-cases have to be designed top-down and be a good fit between audit requirements and technology	Use-cases have to be designed top-down, yet can be problematic as the knowledge gap between auditing and technology must be bridged	Use-cases have to be designed top-down and are yet problematic as the knowledge gap between auditing and technology needs to be bridged
Commitment of resources	Efficiency aspect clear	Efficiency and effectiveness aspect clear	Efficiency and effectiveness to be evaluated	Efficiency and effectiveness to be evaluated
Deployment	Can be easily deployed; no dependence on client	Dependency on client data; solution must be able to deal with different ERP systems and data structures; change management required	Dependency on client data depends on use-case, Change management and training required	Dependence on client data depends on use-case, availability of large data amounts depending on use-case, change management and training required
Operational use	Minimal to no training required	Guidance for use required	Guidance for use required	Guidance for use required
Successful adoption	**Yes**	**Yes**	**Not on a broader scale**	**Not on a broader scale**

The effect of technology characteristics on ADA adoption

Table II.3 presents configurations that indicate the effect of technology characteristics on the adoption process. The different characteristics of RPA, PM, ML, and DL are examined. The technologies exhibit varying degrees of complexity and have different advantages and different requirements. RPA exhibits a low complexity compared to the

other technologies presented in the table, which makes the technology easy to comprehend. This, in turn, enables the sourcing of use-cases from audit teams, even if they posses a lower technological capability. The foremost use of RPA is to automate existing audit procedures, thereby making it independent from the clients' preferences on data access and processing. The technology is also focused on efficiency aspects, through which the evaluation and, hence, the decision to commit resources becomes rather straightforward. However, RPA depends heavily on the structure of a task or procedure to be automated. This requires a high level of standardization and homogenization of tasks and processes as well as makes it important for auditors to carefully prepare the inputs for RPA solutions. More complex tasks are out of reach for RPA if not combined with machine learning and, thus, becoming more intelligent.

Relative to RPA, PM exhibits greater complexity. This increase in complexity shifts the responsibility for ideation more towards the innovation team. A greater complexity also leads to more efforts during the deployment event, as solutions based on more complex ADA technologies require guidance in their operational use. Although RPA requires only minimal training, the use of PM and more complex technologies requires training and possibly the employment of new workflows and processes. More technologically inclined audit team members can function as lead users, trainers, and multipliers that support the deployment of the solution. On the other hand, innovation teams need to design the solutions in such a manner that maximizes acceptance by audit teams. PM relies heavily on the access of the client's transactional data. This implies that the solution must be able to deal with different IT and accounting systems as well as accommodate different database models. ML and DL are currently witnessing less broad applications in practice. Audit firms are reportedly struggling with the identification of use-cases for ML and DL in the first place (A.2), thereby hindering adoption right at the beginning of the process. There appears to be a mismatch between the audit domain and technology, which is yet to be addressed. This could be attributed to lacking technology capabilities in the audit firms, given the complexity of the technologies (also in terms of disciplinary-spanning skills) or difficulties in identifying use-cases that comply with the requirements of both the technology and audit domains.

Table II.4: Effects of organizational characteristics on ADA adoption

Audit firm characteristics	Small firm	Big Four firm
Size	Small, mid-sized	Large
Coordination of innovation	Centralized	International, cross service line
Technological capability	Low	High
Effects on adoption process		
Ideation	None possible	Identification of use-cases from both innovation and audit teams
Make or buy decision (Commitment of resources)	Buy	Make
Solution building	Customizing	Own development
Deployment	Deployed to local teams, external training required	Deployed to international teams, internal training possible
Adopted ADA technologies	**None**	**RPA, PM**

The effect of the organizational characteristics of the audit firm

This section explores how the organizational characteristics of the audit firm in conjunction with its technological capability affects the adoption process. The related configurations are presented in Table II.4; as above, they are based on the current expression of the respective contextual factors. The table juxtaposes instances of ADA adoption of any given ADA technology in a small audit firm and a Big Four firm. In terms of organizational characteristics, the main distinction here is the size of the audit firm, which in turn affects the available resources within the firm and its ability to attract talent with technological capabilities. Due to its limited technological capabilities, the small audit firm will not be able to identify use-cases for ADA on its own and must resort to solutions provided by third-party developers. The adoption process here begins with the commitment of resources. Given that the audit firm identifies a strategic need for a third party solution, they may choose to license or buy. If the level of technological capability is sufficient, the audit firm may customize the solution to the needs of their audit teams. Further, depending on the complexity of the solution's underlying ADA technology, its use may require training. In the case of a small audit firm, training must be sourced externally, thereby adding to its dependency on third parties for ADA adoption. A Big Four firm has more resources at its disposal and better channels to attract the necessary talent. The resulting technological capability prevalent in the firm enables it to develop own-solutions as well as establish internal training. The bigger audit firm also has the possibility of coordinating their research and development efforts across country units and/or service lines, thereby enabling international deployment of solutions. Further, the in-house development of solutions in the Big Four firm enables a close feedback loop between users (audit teams) and developers of a solution (FB3) and, hence, the sourcing of use-cases from audit teams.

The configurations presented in the above sections reveal how the contextual factors that are internal and external to audit firms affect the adoption process. The course of individual process instances and even its outcome depends upon the manifestation of contextual factors. While all process activities are affected by internal contextual factors, the external factors affect only the ideation, deployment, and operational use activities. The technology characteristics and application domain requirements affect both ideation and operational use. In both activities, the audit firm is faced with the task of finding a fit between the two aspects; first in terms of identifying a use-case and subsequently in the diffusion of the solution. The latter can be supported during the deployment activity and through cooperation between audit and innovation teams throughout the process.

II.5 Discussion

The aim of this research was to reveal the process by which audit firms adopt ADA technologies and the contextual factors relevant to the audit domain that affect the process.

Our research adopts a different perspective on ADA adoption in auditing than that adopted by previous research. Instead of discussing the adoption of ADA from a standpoint where ADA is considered a single technology, our results focus on individual solutions based on ADA. These solutions are the product of a process that can be divided into a research and development process phase and a diffusion process phase. Both phases are affected by contextual factors. The only exception are the client characteristics, which only affect the diffusion phase. In the presented theory, we acknowledge the effect of the contextual factors on the adoption process but refrain from assuming these concepts to be drivers for or inhibitors to technology adoption. The relationship between these contextual factors and the adoption process is not yet fully understood, and our research approach is oriented toward the generation of hypotheses rather than quantitative reductionism.

This paper offers a theoretical contribution to the extant body of research by employing a theoretical perspective that has not been used in prior research on technology adoption in auditing, which is primarily concerned with variance theories. The process theory introduces a sequential logic and allows for a contextualization of the related factors. The theory we present in this paper extends, in terms of certain aspects, the theoretical framework that was used for its development—the organizational innovation process as featured in Rogers (2003). Our theory depicts an organizational innovation process but focuses on the development of single ADA solutions across different organizational units. Therefore, our theory makes room for more complexity in the process and highlights the specifics of the audit domain. This results in a lower overall generalizability of the theory, which is consistent with the classification of Langley (1999) regarding the generalizability of the grounded theory. Similar to our process theory, the process in Rogers (2003) is divided into two phases: the initiation and implementation phases. The decision to (non-) adopt an innovation partitions both phases. In our theory, this adoption decision is made in the commitment of resource activity. However, it does not define the phases of the process. Rather, the research and development and diffusion phases relate to different objectives within the adoption process. The objective of the research and development phase is to develop a deployable solution from a use-case, while the objective of the diffusion phase is to ensure the operational use of the solution. The process theory in Rogers (ibid.) is also strictly linear, where each activity must be completed in a sequential order. Our theory allows for feedback loops between single activities. These feedback loops originate from reiterations of single-process activities in the evaluation and deployment stages, if the use-case (or prototype) or solution require revision. For example, the decision to not adopt a technology is not static, but it can be refuted if a better use case is presented. The process as a whole can be further reiterated after successful adoption for continued development (e.g., new features).

The process perspective emphasizes the active role of audit firms; the adoption of ADA requires the firm to make an effort. This contrasts the passive perspective in variance studies, which model technology adoption as a dependent variable. For example, previous

studies employed the task-technology fit (Rosli et al., 2012; Widuri et al., 2016) and the related performance expectancy (Curtis and Payne, 2014; Pedrosa, Costa, and Laureano, 2015; Rosli et al., 2012) as exogenous variables to explain technology adoption in variance studies. In our theory, this fit of task to technology is achieved through the ideation activity in which the *technology characteristics* have to be matched to the *requirements of the audit domain*. The *complexity* of the underlying technology can be absorbed throughout the development of the solution, such that it maximizes the usability for auditors, which is reflected in the *solution characteristics* in the *auditor technology acceptance*. The *auditor technology acceptance* can be related to the research on technology adoption in auditing at the individual level (Curtis and Payne, 2014; Janvrin, Bierstaker, et al., 2009; Janvrin, Lowe, et al., 2008; Li et al., 2018; Payne and Curtis, 2006; Pedrosa, Costa, and Laureano, 2015). One interesting finding in our study, which has not been addressed in previous research, is that the *organizational context* through which ADA solutions are made available to audit teams is of relevance for the adoption.

Similar to previous research (Li et al., 2018; Rosli et al., 2012; Siew et al., 2020), our theory relates the available resources to the size of the audit firm. Our results indicate that, currently, the Big Four take the lead in the adoption of ADA. This contradicts the results of Lowe et al. (2018), which state that mid-sized audit firms have caught up to the Big Four firms in terms of their use of technology. The available resources in audit firms enable them to build own technological capabilities and engage in the development of own-solutions based on ADA. According to our results, the size of the firm also correlates with its ability to attract talent, which is crucial to the adoption process. Further, several studies have referred to the technological competence and resources within the adopting audit firm as factors for technology adoption (Haddara et al., 2018; Hampton and Stratopoulos, 2016; Li et al., 2018; Rosli et al., 2012; Salijeni et al., 2019; Siew et al., 2020; Widuri et al., 2016). These are also reflected in our theory in the *audit firm technological capability*. However, we differ between the capabilities of different groups (audit teams, innovation teams) and between the skills required for different activities that are part of the process (ideation, solution building, operational use). Here, our results also touch upon the inclusion of IT specialists and the apparent lack of ADA professionals in audit firms that inhibit ADA adoption: The innovation teams mentioned in our results represent a different kind of IT specialists—ones who are not directly involved in audit engagements themselves. Instead, they are tasked with developing solutions, which can then be employed in audits. They also possess a different set of skills, which leans more toward software development and methodological knowledge. The latter is particularly important during the ideation phase, where they are required to identify use cases. A few studies refer to the inclusion of IT support staff in the audit teams as a factor for technology adoption (Li et al., 2018; Widuri et al., 2016). In our theory, this is reflected in the existence of sufficiently high technological capability in audit teams, such that the technology-savvy staff (if available) may be leveraged to support other auditors in the deployment activity. Our results further support the findings of Vasarhelyi and Romero (2014) that interdisciplinary skills of IT specialists can have a positive impact on technology adoption. However, according to our findings, this is a two-way street—increased IT capabilities in audit teams can also help the adoption of ADA throughout the process.

Our results also have a few practical implications. The theory relates the contextual factors to the individual activities within the adoption process. The factors are separated

into internal factors, which can be influenced by the audit firms, and external factors, which lie beyond its reach. Therefore, the theory provides a framework to guide both auditors and regulators when referring to the adoption of ADA. Specifically, our findings emphasize the importance of ADA capabilities in audit teams and the importance of being able to identify use-cases for ADA technologies in the first place.

The presented findings and the developed theory must be viewed in light of certain limitations, which are mostly grounded in the sampling of our interview partners and affect the generalizability of the theory. The first limitation is that many of our interview partners are in upper management positions, with most of them holding occupations related to audit digitization. This could imply that they exhibit a bias toward the potential benefits of technology and the importance this aspect holds for practice. We addressed this issue by including interviewees who are employed in audit field work. Another limitation is the country of employment of our interview partners. Most of them work in Germany; therefore, the results can be biased toward the development of the phenomenon there. However, we anticipate these effects to be weak, as a large number of our interview partners work in global companies and the German economy is the fourth-largest in the world by nominal gross domestic product and has global economic ties. As we indicate in section 3.1., the institutional environment of the German audit market is highly internationalized. A further limitation is manifested through the role of audits in the German two-tier corporate governance setting, which is referred to in the methodology section. The motivation of audit firms to use ADA to deliver client insights beyond the normal regular audit report might be stronger in Germany than in other countries. Another limitation is associated with a design choice for the theory. For the sake of simplicity, we chose to not explicitly model existing relationships among contextual factors.

II.6 Conclusion and Outlook

While prior research has addressed the application of ADA to auditing to a large extent, the process underlying ADA adoption in audit firms has barely received attention. This paper contributes to the research field of ADA adoption in auditing by introducing a process theory that reflects how audit firms adopt ADA. The theory further focuses on single solutions derived from ADA technologies instead of discussing the adoption of single ADA technologies into auditing practice as a whole. The process encompasses six activities through which a solution from an ADA technology is derived and diffused into practice: ideation, evaluation, solution-building, commitment of resources, deployment, and operational use. The first four activities are concerned with the development of a solution, whereas the latter two are concerned with the diffusion of solutions into operational use. The commitment of resources is a phase gate within the process, as any adoption of ADA requires financial or human resources. The derived theory further highlights how this adoption process is affected by contextual factors related to the adopting audit firm, characteristics of ADA technologies, characteristics of its clients, and requirements of the audit domain. The contextual factors related to the audit firm are its strategic prioritization of ADA, its organizational characteristics, its technological capability, and the acceptance of the solution on the part of the firm's auditors. All the activities involved in the adoption process are affected by the contextual factors related

to the audit firm, whereas the external contextual factors mainly affect the ideation, deployment, and operational use activities.

Our results indicate that the technological capability of the audit firm, both amongst auditors and among innovation teams, affects successful technology adoption across different phases of the adoption process, most notably during the ideation and deployment activities. One of the greatest challenges in ADA adoption is the identification of use-cases in the ideation phase. Therefore, it is important to bridge the gap between the audit domain and technology, which can be achieved by equipiing auditors with technological knowledge and by encouraging cross-disciplinary thinking in innovation teams. This effect is reinforced if a feedback loop between operational use and the ideation phase is established. Auditors with technological affinity can support the ideation phase, which in turn can improve the acceptance of solutions by other auditors, as they are involved in the development process. Involving auditors in the ideation phase can help to align the solution's design with the auditors' mindset in order to ensure usability, which affects the diffusion of the solution into operational use. Further, auditors with technological capabilities can support the deployment of a solution, where they act as trainers and lead users. However, this requires the commitment of the management of the audit firm to invest corresponding resources, even in the face of potential opportunity costs from non-billable person hours. Based on the above argument, we encourage regulators and standard-setters to consider the importance of ADA capabilities for the skill set of future auditors. Equipped with sufficient background on ADA technologies, they can identify use cases for technologies and drive the digitization of the profession from a conceptual perspective, independent from the size or pockets of their respective firms, such that, once again, the adoption gap between the Big Four and smaller firms can be closed.

This research aims to inspire further inquiries in this field. The results can be followed up by further qualitative and quantitative empirical analyses to extend the depth of the explanations of the results, further examine the nature of the relationships between the process and contextual factors, and test the hypothesized relationships between the associated factors and process outcome. Of special interest is the relationship between the audit domain requirements and the adoption process. In the presented theory, the audit risk model encoded in the professional standards affects the adoption of ADA, with some interviewees stating that they perceived the professional standards as a limiting factor for ADA adoption. However, the adoption of ADA on a broader scale could also transform the general approach of audit firms when conducting audits, thereby possibly leading to further development of the ISA. In this context, it would be interesting to examine the adoption of ADA for external audits of non-financial and diversity disclosures. Assurance services for non-financial reporting are a rather young phenomenon, where large volumes of data (e.g. environmental data) might be more prevalent than in financial reporting. Further, future research could compare the adoption of ADA across different countries to examine possible differences. Another topic that could be addressed in future research is the interaction between auditing and consulting service lines in the development of ADA solutions in Big Four firms and how this is affected by top-level management. In addition, our results indicate an interconnection among service lines, but we did not undertake a deeper examination of this aspect in this paper.

II.7 References

Ahmi, Aidi and Simon Kent (2012). "The utilisation of generalized audit software (GAS) by external auditors". In: *Managerial Auditing Journal* 28.2, pp. 88–113. ISSN: 02686902. DOI: 10.1108/02686901311284522.

Alles, Michael and Glen L. Gray (2016). "Incorporating big data in audits: Identifying inhibitors and a research agenda to address those inhibitors". In: *International Journal of Accounting Information Systems* 22, pp. 44–59. ISSN: 14670895. DOI: 10.1016/j.accinf.2016.07.004. URL: http://dx.doi.org/10.1016/j.accinf.2016.07.004.

Alles, Michael G. (2015). "Drivers of the Use and Facilitators and Obstacles of the Evolution of Big Data by the Audit Profession". In: *Accounting Horizons* 29.2. _eprint: https://meridian.allenpress.com/accounting-horizons/article-pdf/29/2/439/1592722/acch-51067.pdf, pp. 439–449. ISSN: 0888-7993. DOI: 10.2308/acch-51067. URL: https://doi.org/10.2308/acch-51067.

Appelbaum, Deniz A. et al. (2018). "Analytical procedures in external auditing: A comprehensive literature survey and framework for external audit analytics". In: *Journal of Accounting Literature* 40.September 2016, pp. 83–101. ISSN: 07374607. DOI: 10.1016/j.acclit.2018.01.001. URL: https://doi.org/10.1016/j.acclit.2018.01.001.

Barr-Pulliam, Dereck D. et al. (2020). "Data Analytics and Skeptical Actions: The Countervailing Effects of False Positives and Consistent Rewards for Skepticism". In: *SSRN Electronic Journal*. ISSN: 1556-5068. DOI: 10.2139/ssrn.3537180. URL: https://www.ssrn.com/abstract=3537180.

Bauer, Tim and Cassandra Estep (2014). "The IT Auditor Function on Financial Statement and Integrated Audits: Description of Practice and Avenues for Future Research". In: *SSRN Electronic Journal*. ISSN: 1556-5068. DOI: 10.2139/ssrn.2579193.

Bauer, Tim D., Cassandra Estep, and Bertrand Malsch (2019). "One Team or Two? Investigating Relationship Quality between Auditors and IT Specialists: Implications for Audit Team Identity and the Audit Process". In: *Contemporary Accounting Research* 36.4, pp. 2142–2177. ISSN: 19113846. DOI: 10.1111/1911-3846.12490.

Boritz, J. Efrim et al. (2017). "Auditors and Specialists Views About the Use of Specialists During an Audit". In: *SSRN Electronic Journal* 53.9, pp. 1689–1699. ISSN: 1556-5068. DOI: 10.2139/ssrn.2534506. URL: http://www.ssrn.com/abstract=2534506.

Braun, Robert L. and Harold E. Davis (2003). "Computer-assisted audit tools and techniques: Analysis and perspectives". In: *Managerial Auditing Journal* 18.9, pp. 725–731. ISSN: 02686902. DOI: 10.1108/02686900310500488.

Brown, Veena Looknanan et al. (2018). "Comments of the Auditing Standards Committee of the Auditing Section of the American Accounting Association on International Auditing and Assurance Standards Board Exposure Draft, Proposed International Standard on Auditing 315 (Revised): Identifying and Assessing the Risks of Material Misstatement and Proposed Consequential and Conforming Amendments to Other ISAs". In: *Current Issues in Auditing* 13.1, pp. C1–C9. ISSN: 1936-1270. DOI: 10.2308/ciia-52338. URL: https://doi.org/10.2308/ciia-52338 (visited on 12/26/2023).

Cao, Min et al. (2015). "Big data analytics in financial statement audits". In: *Accounting Horizons* 29.2, pp. 423–429. ISSN: 15587975. DOI: 10.2308/acch-51068.

Chan, David Y. et al. (2018). "New Perspective: Data Analytics as a Precursor to Audit Automation". In: *Continuous Auditing*, pp. 315–322. DOI: 10.1108/978-1-78743-413-420181016.

Chen, Hsinchun et al. (2012). "Business Intelligence And Analytics: From Big Data to Big Impact". In: *MIS Quarterly* 36.4, pp. 1165–1188.

Chiu, Tiffany and Mieke Jans (2019). "Process mining of event logs: A case study evaluating internal control effectiveness". In: *Accounting Horizons* 33.3, pp. 141–156. ISSN: 15587975. DOI: 10.2308/acch-52458.

Corbin, Juliet M. and Anselm Strauss (1990). "Grounded theory research: Procedures, canons, and evaluative criteria". In: *Qualitative Sociology* 13.1, pp. 3–21. ISSN: 01620436. DOI: 10.1007/BF00988593.

Curtis, Mary B., J. Gregory Jenkins, et al. (2009). "Auditors' training and proficiency in information systems: A research synthesis". In: *Journal of Information Systems* 23.1, pp. 79–96. ISSN: 08887985. DOI: 10.2308/jis.2009.23.1.79.

Curtis, Mary B. and Elizabeth A. Payne (2014). "Modeling voluntary CAAT utilization decisions in auditing". In: *Managerial Auditing Journal* 29.4, pp. 304–326. ISSN: 02686902. DOI: 10.1108/MAJ-07-2013-0903.

Dagilienė, Lina and Lina Klovienė (2019). "Motivation to use big data and big data analytics in external auditing". In: *Managerial Auditing Journal* 34.7, pp. 750–782. ISSN: 02686902. DOI: 10.1108/MAJ-01-2018-1773.

DePietro, Rocco et al. (1990). "The context for change: Organization, technology and environment". In: *The processes of technological innovation*. Ed. by Louis G. Tornatzky and Mitchell Fleischer. Lexington, MA: Lexington Books, pp. 151–175.

Eilifsen, Aasmund et al. (2019). "An Exploratory Study into the Use of Audit Data Analytics on Audit Engagements". In: *SSRN Electronic Journal* 1, pp. 1–18. ISSN: 1556-5068. DOI: 10.2139/ssrn.3458485. URL: https://www.ssrn.com/abstract=3458485.

EU (2006). "Directive 2006/43/EC of the European Parliament and of the Council of 17 May 2006 on statutory audits of annual accounts and consolidated accounts, amending Council Directives 78/660/EEC and 83/349/EEC and repealing Council Directive 84/253/EEC". In: *Official Journal of the European Union: Legislation L157* 49, pp. 87–107.

— (2014a). "Directive 2014/56/EU of the European Parliament and of the Council of 16 April 2014 amending Directive 2006/43/EC on statutory audits of annual accounts and consolidated accounts". In: *Official Journal of the European Union: Legislation L158* 57, pp. 196–226.

— (2014b). "Regulation (EU) No 537/2014 of the European Parliament and of the Council of 16 April 2014 on Specific Requirements Regarding Statutory Audit of Public-Interest Entities and Repealing Commission Decision 2005/909/EC". In: *Official Journal of the European Union: Legislation L158* 57, pp. 77–112.

Fédération des Experts Comptables Européens (2015). *Overview of ISA Adoption in the European Union*. URL: https://www.accountancyeurope.eu/wp-content/uploads/MA%7B%5C_%7DISA%7B%5C_%7Din%7B%5C_%7DEurope%7B%5C_%7Doverview%7B%5C_%7D150908%7B%5C_%7Dupdate.pdf.

Gray, Glen L. et al. (2014). "The expert systems life cycle in AIS research: What does it mean for future AIS research?" In: *International Journal of Accounting Information Systems* 15.4, pp. 423–451. ISSN: 14670895. DOI: 10.1016/j.accinf.2014.06.001.

Gregor, Shirley (2006). "The Nature of Theory in Information Systems". In: *MIS Quarterly* 30.3, pp. 611–642.

Haddara, Moutaz et al. (2018). "Applications of Big Data Analytics in Financial Auditing-A Study on The Big Four". In: *Twenty-fourth Americas Conference on Information Systems, New Orleans* 2018, pp. 1–10. DOI: 10.1201/b18737-189.

Hampton, Clark and Theophanis C. Stratopoulos (2016). "Audit Data Analytics Use: An Exploratory Analysis". In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.2877358.

Hoitash, Rani et al. (2008). "Internal control quality and audit pricing under the Sarbanes-Oxley Act". In: *Auditing* 27.1, pp. 105–126. ISSN: 02780380. DOI: 10.2308/aud.2008.27.1.105.

Al-Htaybat, Khaldoon and Larissa von Alberti-Alhtaybat (2017). "Big Data and corporate reporting: impacts and paradoxes". In: *Accounting, Auditing and Accountability Journal* 30.4, pp. 850–873. ISSN: 09513574. DOI: 10.1108/AAAJ-07-2015-2139.

Huang, Feiqi and Miklos A. Vasarhelyi (2019). "Applying robotic process automation (RPA) in auditing: A framework". In: *International Journal of Accounting Information Systems* 35.xxxx, p. 100433. ISSN: 14670895. DOI: 10.1016/j.accinf.2019.100433. URL: https://doi.org/10.1016/j.accinf.2019.100433.

Humphrey, Christopher et al. (2011). "Regulating Audit beyond the Crisis: A Critical Discussion of the EU Green Paper". In: *European Accounting Review* 20.3, pp. 431–457. ISSN: 09638180. DOI: 10.1080/09638180.2011.597201.

Hunton, James E et al. (2004). "Are financial auditors overconfident in enterprise resource planning systems?" In: *Journal of Information Systems* 18.2, pp. 7–28. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract%7B%5C_%7Did=691683.

IAASB (2009). *International Standard On Auditing 620 Using the Work of an Auditor's Expert*. URL: https://www.ifac.org/system/files/downloads/a035-2010-iaasb-handbook-isa-620.pdf.

— (2016a). *Exploring the Growing Use of Technology in the Audit, with a focus on data analytics*. URL: https://www.ifac.org/system/files/publications/files/IAASB-Data-Analytics-WG-Publication-Aug-25-2016-for-comms-9.1.16.pdf.

— (2016b). *International Standard On Auditing 701 Communicating Key Audit Matters in the Independenten Auditor's Report*. URL: https://www.ifac.org/system/files/publications/files/ISA-701%7B%5C_%7D2.pdf.

— (2019a). *Basis for Conclusions: International Standard on Auditing 315 (Revised 2019): Identifying and Assessing the Risks of Material Misstatement, Including Conforming and Consequential Amendments to other International Standards*.

— (2019b). *International Standard on Auditing 315 (Revised 2019): Identifying and Assessing the Risks of Material Misstatement*. URL: https://www.ifac.org/system/files/publications/files/ISA-315-Full-Standard-and-Conforming-Amendments-2019-.pdf.

Issa, Hussein et al. (2016). "Research ideas for artificial intelligence in auditing: The formalization of audit and workforce supplementation". In: *Journal of Emerging Technologies in Accounting* 13.2, pp. 1–20. ISSN: 15587940. DOI: 10.2308/jeta-10511.

Jans, Mieke et al. (2013). "The case for process mining in auditing: Sources of value added and areas of application". In: *International Journal of Accounting Information Systems* 14.1, pp. 1–20. ISSN: 14670895. DOI: 10.1016/j.accinf.2012.06.015. URL: http://dx.doi.org/10.1016/j.accinf.2012.06.015.

Jans, Mieke et al. (2014). "A field study on the use of process mining of event logs as an analytical procedure in auditing". In: *Accounting Review* 89.5, pp. 1751–1773. ISSN: 00014826. DOI: 10.2308/accr-50807.

Janvrin, Diane, James Bierstaker, et al. (2008). "An examination of audit information technology use and perceived importance". In: *Accounting Horizons* 22.1, pp. 1–21. ISSN: 08887993. DOI: 10.2308/acch.2008.22.1.1.

— (2009). "An Investigation of Factors Influencing the Use of Computer-Related Audit Procedures". In: *Journal of Information Systems* 23.1, pp. 97–118. ISSN: 0888-7985. DOI: 10.2308/jis.2009.23.1.97. URL: https://meridian.allenpress.com/jis/article/23/1/97/75367/An-Investigation-of-Factors-Influencing-the-Use-of.

Janvrin, Diane, D. Jordan Lowe, et al. (2008). "Auditor Acceptance of Computer-Assisted Audit Techniques". In: *American Accounting Association Auditing Mid Year Meeting AAA* April, pp. 1–26.

Köhler, Annette G et al. (2007). "Audit regulation in Germany: Improvements driven by internationalization". In: *Auditing, Trust and Governance: Regulation in Europe*. Ed. by Reiner Quick et al. Routledge, pp. 129–161.

Kokina, Julia and Thomas H. Davenport (2017). "The emergence of artificial intelligence: How automation is changing auditing". In: *Journal of Emerging Technologies in Accounting* 14.1, pp. 115–122. ISSN: 15587940. DOI: 10.2308/jeta-51730.

Lacity, Mary C. and Leslie P. Willcocks (2016). "Robotic process automation at telefónica O2". In: *MIS Quarterly Executive* 15.1, pp. 21–35. ISSN: 15401979.

Langley, Ann (1999). "Strategies for theorizing from process data". In: *Academy of Management Review* 24.4, pp. 691–710. ISSN: 03637425. DOI: 10.5465/AMR.1999.2553248. URL: http://www.jstor.org/stable/259349.

Li, He et al. (2018). "Understanding usage and value of audit analytics for internal auditors: An organizational approach". In: *International Journal of Accounting Information Systems* 28.November 2017, pp. 59–76. ISSN: 14670895. DOI: 10.1016/j.accinf.2017.12.005.

Lowe, D. Jordan et al. (2018). "Information technology in an audit context: Have the big 4 lost their advantage?" In: *Journal of Information Systems* 32.1, pp. 87–107. ISSN: 15587959. DOI: 10.2308/isys-51794.

Manita, Riadh et al. (2020). "The digital transformation of external audit and its impact on corporate governance". In: *Technological Forecasting and Social Change* 150.September 2019, p. 119751. ISSN: 00401625. DOI: 10.1016/j.techfore.2019.119751. URL: https://doi.org/10.1016/j.techfore.2019.119751.

Markus, M Lynne and Daniel Robey (1988). "Information Technology and Organizational Change: Causal Structure in Theory and Research". In: *Management Science* 34.5, pp. 583–598. ISSN: 0025-1909. DOI: 10.1287/mnsc.34.5.583. URL: http://pubsonline.informs.org/doi/abs/10.1287/mnsc.34.5.583.

Michael, Amir and Rob Dixon (2019). "Audit data analytics of unregulated voluntary disclosures and auditing expectations gap". In: *International Journal of Disclosure and Governance* 16.4, pp. 188–205. ISSN: 1746-6539. DOI: 10.1057/s41310-019-00065-x. URL: https://doi.org/10.1057/s41310-019-00065-x.

Moffitt, Kevin C. et al. (2018). "Robotic process automation for auditing". In: *Journal of Emerging Technologies in Accounting* 15.1, pp. 1–10. ISSN: 15587940. DOI: 10.2308/jeta-10589.

Molinillo, Sebastian and Arnold Japutra (2017). "Organizational adoption of digital information and technology: a theoretical review". In: *Bottom Line* 30.1, pp. 33–46. ISSN: 0888045X. DOI: 10.1108/BL-01-2017-0002.

Myers, Michael D. and Michael Newman (2007). "The qualitative interview in IS research: Examining the craft". In: *Information and Organization* 17.1, pp. 2–26. ISSN: 14717727. DOI: 10.1016/j.infoandorg.2006.11.001.

O'Donnell, Jb (2010). "Innovations in Audit Technology: A Model of Continuous Audit Adoption". In: *Journal of Applied Business and Economics* 10.5, pp. 11–20. URL: http://m.www.na-businesspress.com/odonnellweb.pdf.

Oliveira, Tiago and Maria Fraga Martins (2010). "Information technology adoption models at Firm Level: Review of literature". In: *4th European Conference on Information Management and Evaluation, ECIME 2010* May, pp. 312–322.

Otero, Angel (2015). "Impact of IT Auditors' Involvement in Financial Audits". In: *International Journal of Research in Business and Technology* 6.3. DOI: 10.17722/ijrbt.v6i3.404.

Payne, Elizabeth A. and Shirley Curtis (2006). "An Examination of Contextual Factors and Individual Characteristics Affecting Technology Implementation Decisions in Auditing". In: *Journal of Chemical Information and Modeling* 53.9, pp. 1689–1699. ISSN: 1098-6596. DOI: 10.1017/CBO9781107415324.004. arXiv: arXiv:1011.1669v3.

Pedrosa, Isabel, Carlos J. Costa, and Manuela Aparicio (2020). "Determinants adoption of computer-assisted auditing tools (CAATs)". In: *Cognition, Technology and Work* 22.3, pp. 565–583. ISSN: 14355566. DOI: 10.1007/s10111-019-00581-4. URL: https://doi.org/10.1007/s10111-019-00581-4.

Pedrosa, Isabel, Carlos J. Costa, and Raul M.S. Laureano (2015). "Motivations and limitations on the use of information technology on statutory auditors' work: An exploratory study". In: *2015 10th Iberian Conference on Information Systems and Technologies, CISTI 2015* June. DOI: 10.1109/CISTI.2015.7170623.

Rapoport, Michael (2018). *How Did the Big Four Auditors Get $17 Billion in Revenue Growth? Not From Auditing: Consulting is now a cash cow for accounting firms, raising concerns about conflicts of interest*. Ed. by The Wall Street Journal. URL: https://www.wsj.com/articles/how-did-the-big-four-auditors-get-17-billion-in-revenue-growth-not-from-auditing-1523098800.

Rogers, Everett M (2003). *Diffusion of innovations*. 5th ed. New York: Free Press. ISBN: 9780743222099.

Rose, Anna M et al. (2017). "When Should Audit Firms Introduce Analyses of Big Data Into the Audit Process?" In: *Journal of Information Systems* 31.3, pp. 81–99. ISSN: 0888-7985. DOI: 10.2308/isys-51837. URL: https://doi.org/10.2308/isys-51837.

Rosli, Khairina et al. (2012). "Factors Influencing Audit Technology Acceptance by Audit Firms: A New I-TOE Adoption Framework". In: *Journal of Accounting and Auditing: Research & Practice* 2012, pp. 1–11. ISSN: 2165-9532. DOI: 10.5171/2012.876814.

Russom, Philip (2011). *Introduction to Big Data Analytics*. URL: https://vivomente.com/wp-content/uploads/2016/04/big-data-analytics-white-paper.pdf.

Salahshour Rad, Maryam et al. (2018). "Information technology adoption: a review of the literature and classification". In: *Universal Access in the Information Society* 17.2, pp. 361–390. ISSN: 16155297. DOI: 10.1007/s10209-017-0534-z.

Salijeni, George et al. (2019). "Big Data and changes in audit technology: contemplating a research agenda". In: *Accounting and Business Research* 49.1, pp. 95–119. ISSN: 21594260. DOI: 10.1080/00014788.2018.1459458.

Siew, Eu Gene et al. (2020). "Organizational and environmental influences in the adoption of computer-assisted audit tools and techniques (CAATTs) by audit firms in Malaysia".

In: *International Journal of Accounting Information Systems* 36, p. 100445. ISSN: 14670895. DOI: 10.1016/j.accinf.2019.100445. URL: https://doi.org/10.1016/j.accinf.2019.100445.

Sun, Ting Sophia (2019). "Applying deep learning to audit procedures: An illustrative framework". In: *Accounting Horizons* 33.3, pp. 89–109. ISSN: 15587975. DOI: 10.2308/acch-52455.

Thornberg, Robert and Ciarán Dunne (2019). "The literature review in grounded theory". In: *The Sage handbook of current developments in grounded theory*, pp. 206–221.

Tucker, George H (2001). "IT and the audit". In: *Journal of Accountancy* 192.3, p. 41.

Urquhart, Cathy (2012). *Grounded Theory for Qualitative Research: A Practical Guide*. Sage. ISBN: 978-1847870544. DOI: 10.4135/9781526402196.

Van Der Aalst, Wil M.P. et al. (2010). "Auditing 2.0: Using process mining to support tomorrow's auditor". In: *Computer* 43.3, pp. 90–93. ISSN: 00189162. DOI: 10.1109/MC.2010.61.

Vasarhelyi, Miklos A. and Silvia Romero (2014). "Technology in audit engagements: A case study". In: *Managerial Auditing Journal* 29.4, pp. 350–365. ISSN: 02686902. DOI: 10.1108/MAJ-06-2013-0881.

Venkatesh et al. (2003). "User Acceptance of Information Technology: Toward a Unified View". In: *MIS Quarterly* 27.3, p. 425. ISSN: 02767783. DOI: 10.2307/30036540. URL: https://www.jstor.org/stable/10.2307/30036540.

Widuri, Rindang et al. (2016). "Adopting generalized audit software: an Indonesian perspective". In: *Managerial Auditing Journal* 31.8-9, pp. 821–847. ISSN: 02686902. DOI: 10.1108/MAJ-10-2015-1247.

Yoon, Tom E. and Joey F. George (2013). "Why aren't organizations adopting virtual worlds?" In: *Computers in Human Behavior* 29.3, pp. 772–790. ISSN: 07475632. DOI: 10.1016/j.chb.2012.12.003.

Chapter III

Information Extraction from Invoices: A Graph Neural Network Approach for Datasets with High Layout Variety

Outline

III.1 Introduction 109
III.2 Related Work 110
III.3 Methodology 112
III.4 Experimental setup 116
III.5 Results 117
III.6 Discussion 120
III.7 Conclusion and outlook 122
III.8 References 123

Bibliographic Information

Krieger, F., Drews, P., Funk, B., Wobbe, T. (2021). "Information Extraction from Invoices: A Graph Neural Network Approach for Datasets with High Layout Variety". In: Ahlemann, F., Schütte, R., Stieglitz, S. (eds.) Innovation Through Information Systems. WI 2021. Lecture Notes in Information Systems and Organisation, 47. Springer. DOI 10.1007/978-3-030-86797-3_1 Researchgate: 348200686.

Author's contribution

The author's share of the publication is 80%. Table C.3 in appendix C shows the contributions of all authors of the publication in detail.

Abstract

Extracting information from invoices is a highly structured, recurrent task in auditing. Automating this task would yield efficiency improvements, while simultaneously improving audit quality. The challenge for this endeavor is to account for the text layout on invoices and the high variety of layouts across different issuers. Recent research has proposed graphs to structurally represent the layout on invoices and to apply graph convolutional networks to extract the information pieces of interest. However, the effectiveness of graph-based approaches has so far been shown only on datasets with a low variety of invoice layouts. In this paper, we introduce a graph-based approach to information extraction from invoices and apply it to a dataset of invoices from multiple vendors. We show that our proposed model extracts the specified key items from a highly diverse set of invoices with a macro F_1 score of 0.8753.

III.1 Introduction

According to the study conducted by Frey and Osborne (2017), auditing is among the professions which are most likely to be impacted by computerization, as they involve a high number of repetitive, structured tasks. One crucial task that fits this description is the extraction of information from invoices (EII), which is performed during tests of details. Tests of details are substantive audit procedures to obtain evidence that the balances and disclosures related to the audited company's financial statement and the corresponding transactions have been recorded and reported correctly (IAASB, 2009). Invoices are used frequently here, as they are the most elemental source of data used in accounting. They hold the details of any commercial exchange of goods or services between companies and/or consumers. Details of interest to the auditor are e.g. *invoice numbers*, invoice and due dates, total and tax amounts, VAT numbers, and line items. When performing tests of details, auditors draw samples of invoices, which can range from dozens to hundreds of documents in size, depending on the beforehand conducted risk assessment. Reviewing the sampled invoices by hand for tests of details requires many person-hours per audit engagement. Automating EII can thus increase the efficiency of audits, while simultaneously increasing audit quality by allowing auditors to focus on higher value-added tasks, and through the ability to test more invoices by increasing the processing speed. Initially proposed solutions for automating EII employ rules-based processing and template-matching (Dengel and Bertin, 2002; Esser et al., 2012; Schuster et al., 2013), which require human input to construct business rules and templates. In an audit context, the scalability of such solutions quickly reaches its limits: Especially bigger audit firms audit a wide range of clients from multiple industries, which receive invoices from a multitude of different business partners. The layouts of invoices can vary highly between issuing companies (hereafter referred to as 'vendors'). The efficiency gains from rules- or template-based automation solutions would soon be canceled out by the effort required for their adaption to individual vendor layouts. For a solution to be employed in audits, it should therefore be able to capture the general patterns prevalent on invoices and generalize to unseen invoice layouts. The applicability of such a solution would also extend beyond auditing and could support administrative processes, especially accounts payable, in public and private organizations. To address the complexity of dealing with a multitude of invoice layouts, recent research in the area of EII has proposed machine learning (ML)-based approaches (Denk and Reisswig, 2019; Katti et al., 2018; Liu et al., 2019; Lohani et al., 2019; Majumder et al., 2020; Palm et al., 2017). The challenge for the application of ML to EII is that the text follows a 2-dimensional layout, as opposed to the sequential, unformatted text usually assumed by natural language processing (NLP) methods. Previous studies proposed to employ graphs for representing the text on an invoice document such that the layout is preserved (Liu et al., 2019; Lohani et al., 2019). The key items are extracted from this document graph by using graph convolutional neural networks (GCN) (Liu et al., 2019; Lohani et al., 2019). GCN leverage a context diffusion mechanism, which is functionally similar to the local receptive fields in (grid) convolutional networks used in computer vision (CV) (Bacciu et al., 2020). Graph representations of invoices are albeit less granular than their pixel-based CV counterparts (Denk and Reisswig, 2019; Katti et al., 2018), making them more computationally efficient. However, the ability to extract key items from invoices of GCNs has so far only been demonstrated on invoice datasets with minor variations in layouts (Liu et al., 2019). In line with the above-mentioned requirements for the audit domain, our research therefore addresses the

following research question:

> "How can graph-based neural networks be applied to extract key items from invoices with a high variety of layouts?"

The contribution of this paper lies in the introduction of a graph attention-based model to extract information from invoices, and its application on a dataset of invoices sourced from a multitude (277) of vendors.

III.2 Related Work

Early works concerned with the automation of EII have studied rule- and template-based approaches to automating this task (Dengel and Bertin, 2002; Esser et al., 2012; Schuster et al., 2013). These approaches are able to extract the desired information, albeit only from known invoice templates, and require human input to create new rules or templates. One of the first proposed systems to apply ML was CloudScan (Palm et al., 2017), which uses an LSTM-based neural network to extract key items via sequence labeling, a common approach to information extraction in NLP. However, such approaches assume the text to be sequential and unformatted and do not account for the 2-dimensional layout of invoices.

Recent studies have proposed different approaches to represent the text on invoices such that the layout is preserved. The approaches can be broadly classified into grid-based (Denk and Reisswig, 2019; Katti et al., 2018) and graph-based (Liu et al., 2019; Lohani et al., 2019). In the former, the text is mapped to a grid, as in Chargrid (Katti et al., 2018) and BERTgrid (Denk and Reisswig, 2019). The latter approaches model documents as graphs, in which either words (Lohani et al., 2019) or whole text segments (Liu et al., 2019) are represented as nodes, and their spatial relationships are represented as edges. In (ibid.), the edges are furthermore weighted with the distances between text segments. As is shown in Table III.1, the document representation and the methods used are intertwined. Table III.1 summarizes the methodology of previous studies and provides details on the respective data sets and the reported performance metrics: Grid-based approaches lean methodologically more towards CV and define the task of extracting key items as semantic segmentation and/or bounding box regression (Denk and Reisswig, 2019; Katti et al., 2018). The graph-based approaches lean more towards NLP, using word/node classification and sequence labeling to identify key items. (Liu et al., 2019; Lohani et al., 2019). Majumder et al. (2020) use a different approach. Their work is based on representation learning and leverages prior knowledge about key items which is used to generate prototypical embeddings for each key item. For each field, candidate words are selected based on their inferred data type. To determine whether a candidate is a key item, it is embedded together with its contextual information, and the cosine distance between the candidate's embedding and the respective key item embedding is calculated (ibid.).

While there are different approaches to document representation, a notion common to all mentioned works is the importance of context for the detection of key items. To this end, grid (Denk and Reisswig, 2019; Katti et al., 2018) or graph (Liu et al., 2019; Lohani et al., 2019) convolutions are employed in the recent literature, as well as the attention

mechanism (Majumder et al., 2020), or a combination thereof (Liu et al., 2019). Further similarities can be found in the nature of the features used. Usually, some combination of syntactical, positional, and/or semantic features are employed. Syntactical features capture (dis-) similarities in the syntax of words and are obtained via character (Denk and Reisswig, 2019; Katti et al., 2018) or byte pair (Liu et al., 2019) encoding, as well as through the inference of data types (string, date, alphanumeric, etc., Lohani et al., 2019; Majumder et al., 2020). Positional features are usually bounding box coordinates either used as explicit features (Majumder et al., 2020), implicitly encoded in the document representation (Denk and Reisswig, 2019; Katti et al., 2018; Liu et al., 2019; Lohani et al., 2019), or Euclidean distances between text boxes (Liu et al., 2019; Lohani et al., 2019). Semantic features are obtained through word embedding layers (Majumder et al., 2020) or from language models such as word2vec (Liu et al., 2019) or BERT (Denk and Reisswig, 2019). In addition to semantic, positional, and semantic features, Lohani et al. (2019) use external databases to discover known entities (cities, zip codes, etc.). While the general type of features used is similar across approaches, the specific utilization and the respective implementation of the model vary. The works reviewed in this section are difficult to compare in terms of performance, as each work relies on different proprietary invoice datasets. As is shown in Table III.1, the datasets vary in size and variety; Majumder et al. (Majumder et al., 2020) use the most exhaustive dataset of the works presented in this section, with each invoice coming from a different vendor. The dataset used by Katti et al. (2018) and Denk and Reisswig (2019) is comparable in size and variety. Liu et al. (2019) use a set of Chinese invoices, which all follow the same government-regulated layout. In terms of size, it is comparable to the dataset used by Lohani et al. (2019). The authors however do not provide any further details, such as the number of vendors or templates or the exact distribution of languages.

Apart from the datasets, another difficulty in comparing approaches is given through the different evaluation methodologies and metrics. Katti et al. (2018) and Denk and Reisswig (2019) evaluate the performance of their models on the character level. To this end, they use a metric similar to the word error rate. As they are based on the same dataset and use the same metric, they are the most comparable works reviewed in this section. Denk and Reisswig (ibid.) show that BERTgrid is able to outperform Chargrid with 65.49% average accuracy over 61.48% by extracting BERT features for every word on the invoice from a BERT model trained on invoices. The other works evaluate their models on the word level. Lohani et al. (2019) present very detailed results for the most exhaustive list of extracted key items of all works reviewed in this section. They report F1 scores, precision, and recall for 27 extracted key items, with micro-averages of 0.93, 0.93 and 0.929 respectively. The other graph-based approach presented by Liu et al. (Liu et al., 2019) achieves an averaged F1 score of 0.881 on an invoice dataset. No details on the averaging method are provided. They furthermore report F1 scores for 6 out of 16 extracted key items on the invoice dataset. Majumder et al. (2020) present F1 scores for 7 extracted key items along with a macro-averaged F1 score of 0.878. Naively observed, it may seem that the approach introduced by Lohani et al. (2019) performs better than the approaches presented by Liu et al. (2019) and especially Majumder et al. (2020). Due to the specifics of the datasets, a direct comparison of the presented results is not meaningful. The limited variety of the invoice dataset used by Liu et al. (2019) and the incomplete information regarding the variety of the dataset used by Lohani et al. (2019) leave doubt as to whether graph-based approaches would perform as well in a setting with a higher

Table III.1: Methodology, dataset details and reported performance measures of related studies

	Document representation structure and granularity	Model type	Information extraction task	Dataset size	Dataset languages	Number of vendors / layouts in dataset	Reported averaged performance over all key items
Katti et al. (2018)	Grid; Characters	(Grid) Convolutional Neural Network	Semantic segmentation, bounding box regression	12,000	Several, mainly English	Most vendors appear once or twice (Katti et al., 2018)	61.48% Accuracy measure (as reported in Denk and Reisswig (2019))
Denk and Reisswig (2019)							65.48% Accuracy measure
Liu et al. (2019)	Graph; Text segments	Graph Attention Network, BiLSTM-CRF	Sequence labelling	3,000	Chinese	Single layout	0.881 F_1 score
Lohani et al. (2019)	Graph; Words	Graph Convolutional Network	Node classification	3,100	English, French	No reference	0.93 F_1 score (Micro) 0.93 Precision (Micro) 0.929 Recall (Micro)
Majumder et al. (2020)	Candidate representations; Words	Attention-based Neural Network	Measuring candidate embedding similarity to field embedding	14,327	English	14,327	0.878 F_1 score (Macro)

variety of layouts. Our work aims to close this research gap by exploring the performance of a graph-based neural network for EII on a dataset with invoices from a multitude of vendors.

III.3 Methodology

To evaluate the performance of graph-based models in EII, we introduce a graph network model that draws inspiration from the above presented recent research. Figure III.1 depicts the document representation and model architecture, and how they are intertwined. The model takes document graphs as input, in which each node represents a word in the document. Syntactic, positional, and semantic features are attached to each node, which are derived from the word the node represents. The edges in the document graphs represent the relative positional relationship between the words. Key items are then extracted via node classification. In this section, we introduce the representation of documents through node features and document graphs and the architecture of the proposed model.

III.3.1 Document representation

As mentioned before, we use graphs to represent documents, which are constructed from optical character recognition (OCR) outputs of invoices. We use word-level OCR outputs,

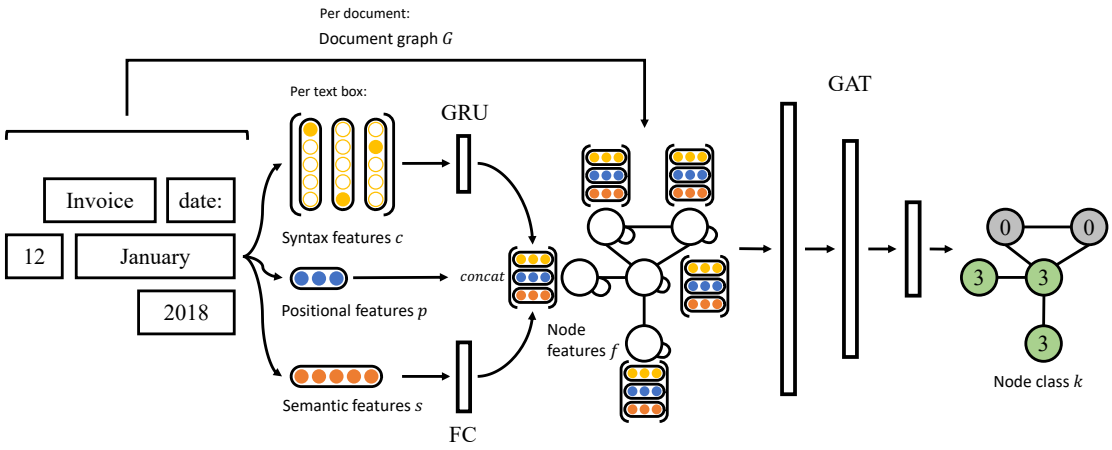

Figure III.1: The node features are embedded using fully connected (FC) and recurrent (GRU) layers and are attached to the document graph, which is passed into graph attention layers (GAT) for node classification

such that OCR yields a set of text boxes $\mathcal{D}$, which contains all n recognized words w on the document. Each text box corresponds to a word. We ignore empty and whitespace text boxes. The bounds of the text boxes are described by the cartesian coordinates x_1, y_1, x_2, y_2 of the box' corners, measured in pixels. The width and height of a document are denoted by W, H such that $0 \leqslant x_{1,2} \leqslant W$ and $0 \leqslant y_{1,2} \leqslant H$. An OCRed document can hence be formally described as a set of n text boxes on a 2-dimensional plane $\mathcal{D} = \{(w^{(j)}, x_1^{(j)}, y_1^{(j)}, x_2^{(j)}, y_2^{(j)}) | j \in \{1, ..., n\}\}$, where the superscript refers to a text box. Each text box in $\mathcal{D}$ is represented as node in the document graph $G = (V, E)$. $V = \{v^{(i)} | i \in \{1, \ldots, n\}\}$ is a set of nodes, and $E = \{e_{ij} | i, j \in \{1, \ldots, n\}\}$ a set of edges between nodes $v^{(i)}$ and $v^{(j)}$. Figure III.2 depicts an example of the graph representation used in our approach. E is then constructed from the text box coordinates. We use the following algorithm to construct E: Using the bounding box coordinates, each node $v^{(i)}$ is connected through e_{ji} with its neighbors $v^{(j)}$ to the top, bottom, left and right. The neighborhood $N(i)$ of $v^{(i)}$ are all nodes which are connected to it via an edge: $N(i) = \{v^{(j)} \in V | e_{ji} \in E\}$. $N(i)$ can contain more than four elements, as the edges are in rare instances not symmetrical.

For $v^{(j)} | j \neq i$ to become a candidate for a horizontal neighbor, it must fulfill either $x_2^{(j)} x_1^{(i)}$ (left neighbor) or $x_2^{(j)} \geqslant x_1^{(i)}$ (right neighbor), while simultaneously fulfilling $y_1^{(i)} \leqslant y_1^{(j)} \leqslant y_2^{(i)}$, $y_1^{(i)} \leqslant y_2^{(j)} \leqslant y_2^{(i)}$ or $y_1^{(j)} \leqslant y_1^{(i)}$, $y_2^{(i)} \leqslant y_2^{(j)}$. From these candidates, the candidate with the smallest Euclidean distance between the respective outer coordinates is then selected as neighbor. An example for this heuristic is given in Figure III.3.

Vertical neighbors are determined analogically. In addition to the neighbors, each node includes a self-loop e_{ii}. Through the self-loops, the model proposed in section III.3.2

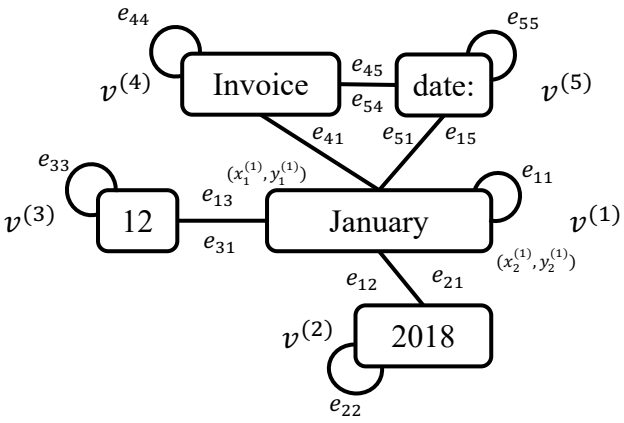

Figure III.2: Example for the constructed document graph, showing the neighborhood $N(1)$ for the node $v^{(1)}$ representing $w^{(1)}$ "January"

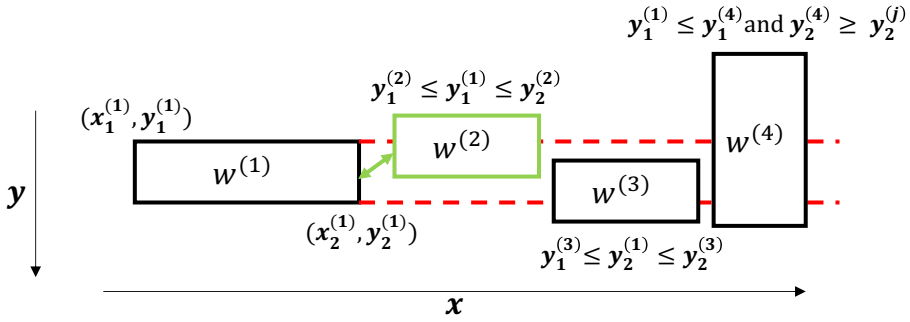

Figure III.3: Valid right neighbor candidates for $v^{(1)}$, with $v^{(2)}$ being selected as neighbor

can access the node's own features. The edges in E are unweighted and undirected, the in-degree of $v^{(i)}$ is equal to its out-degree.

For each node in G, word-level features are extracted. We use three types thereof: Syntactic features, positional features, and semantic features. The syntactic features are used to capture the fine-grained syntactical (dis-) similarities between tokens. Character level one-hot representations of $w^{(j)}$ are extracted, using a fixed dictionary of 110 capital and lowercase Western European letters, numbers and special characters. The length of each token is padded to ten characters. The one-hot encoding process yields a tensor $C = [c_1, \ldots, c_n]$ with dimensions $(n \times 110 \times 10)$ The coordinates of the text boxes are used to extract positional features. The $x_{1,2}^{(j)}$ and $y_{1,2}^{(j)}$ coordinates are scaled to W and H respectively. The positional features include the width, height, and area of the text box, and the Euclidean distances to the nearest neighbors. Missing values for distances are imputed using the maximum possible distance 1.0. Missing values appear if a node has no neighbor in one of the directions. In sum, 13 positional features are extracted for each text box, yielding a matrix $P = [p_1, \ldots, p_n]$ of shape $(n \times 13)$. The semantic features are supposed to capture the meaning behind a token and its relationship to other tokens. To this end, we extract word embedding vectors for $w^{(j)}$ using a pretrained multilingual

BERT (Devlin et al., 2019) base model. To ensure the scalability of the system, we refrain from building dictionaries or embedding tables of a fixed set of words. Another reason for this is the susceptibility of OCR to noise, depending on the quality of the underlying document. For each $w^{(j)}$, BERT outputs a feature vector of length 768. Passing each token into BERT, a feature matrix $S = [s_1, ..., s_n]$ is obtained. Four model inputs are generated in summary: The document graph G, a feature tensor containing the one-hot encoded tokens C, the positional feature matrix P, and the semantic feature matrix S.

III.3.2 Model architecture

We use the document representation described above as input to the model to perform node classification on the document graph G; each node is assigned a corresponding class label. The model proposed in this paper is composed of recurrent, linear, and graph attention layers. Figure III.1 shows the model architecture, along with the correspondent in- and outputs. The first layers are designated for feature embedding. The one-hot encoded character sequences C are passed into a gated recurrent unit (GRU) (Cho et al., 2014) layer, which extracts syntax-sensitive word embeddings C'. Recurrent layers (such as GRU) have been shown to be useful for character-level language modeling (Karpathy et al., 2015). The semantic features S are embedded into a lower dimensional vector space using a fully connected layer (FC) with $ReLu$ nonlinearity, yielding S'. The embedded syntactic and semantic features are then concatenated with the positional features to form the node features $F = (C'||P||S')$, where $||$ denotes the concatenation operator. G and F are passed into the graph attention (GAT) layers (Veličković et al., 2018). GAT layers perform weighted neighborhood feature aggregation over G, using attention scores as weights. The attention mechanism allows the model to focus on specific neighboring nodes. For a node $v^{(i)}$ and its neighborhood $N(i)$ of adjacent nodes, which includes its self-loop e_{ii}, the forward propagation rule of a GAT layer l can be written as

$$h_i^{(i+1)} = \sigma\left(\sum_{j \in N(i)} \alpha_{ij}^{(l)} z_j^{(l)}\right) \tag{III.1}$$

where $\alpha_{ij}^{(l)} = softmax(w_{ij}^{(l)})$ and $w_{ij}^{(l)}$ are the raw attention scores $w_{ij}^{(l)} = LeakyRelu\left(a^{(l)^T}(z_i^{(l)} z_j^{(l)})\right)$. $a^{(l)^T}$ is a vector learned by the model and $z_i^{(l)}$, $z_j^{(l)}$ are the linear activations of layer l. σ denotes the nonlinearity, for which we use $ReLu$ on the GAT layers, similar to Lohani et al. (2019). We use multiple GAT layers to extend the context used to classify a node beyond its direct neighborhood $N(i)$ to include the neighborhoods $N(j)$ of the nodes in $N(i)$ (Bacciu et al., 2020). In the first two GAT layers, we additionally employ multi-headed attention. Multi-headed attention has been applied to increase the stability of the learning process in GAT layers (Veličković et al., 2018). The attention heads compute equation III.1 independently, their outputs are then concatenated and passed into the next layer. The last layer is a single-headed GAT layer with a softmax activation function, which performs the node classification. It returns a probability distribution over the classes for each node in the document graph.

III.4 Experimental setup

We test the above-proposed document representation and model using a set of invoices to determine its performance in a realistic setting. For now, we focus on a limited set of key items to be extracted: The *invoice number*, the invoice date, and the *total amount*. We define a fourth class, "unlabeled", for all other text on the invoice. Details on the dataset and the model implementation and training are given below.

III.4.1 Dataset

As of now, there are no publicly available sets of labeled invoices, which are sufficient in size and variety to train sophisticated ML models. For this research, we were provided a set of invoices by an audit firm, in which they are the recipient. The dataset is composed of 1,129 English one-page invoices from 277 different vendors. We annotated the invoices ourselves by hand for the key items. Individual key items can be composed of multiple words and appear more than once on one invoice. The classes (i.e. key items) in our dataset are sharply imbalanced: Out of the 243,704 textboxes retrieved in our dataset, only 1,427 ($\tilde{0}.58\%$) contain *invoice numbers*, 2,221 ($\tilde{0}.91\%$) contain *total amounts*, and 2,600 ($\tilde{1}.06\%$) contain *invoice dates*. Table III.2 details the split of the dataset into training, validation, and test sets. We chose validation and test set sizes of 10% each, to save as many examples for training as possible. The data splits were stratified across vendors. This way we ensure that both the validation and test splits contain invoices from vendors that remained unseen during training.

Table III.2: Details on data splits

	Number of invoices	Number of unique vendors	Number of vendors unique to split
Training set	903	239	178
Validation set	113	58	18
Test set	113	62	18

III.4.2 Implementation and training

The model described above is implemented using Pytorch 1.7.0 with CUDA 10.1 and the Deep Graph Library (Wang, 2019) 0.5.2. We use Tesseract 4.0 as OCR engine. Table III.3 outlines the chosen model specification in terms of layer sizes (number of hidden nodes) and the number of attention heads in the GAT layers. The size of the GAT 1 layer equals the sum of sizes of the FC and GRU layers and the number of positional features p.

The model is trained using the multi-class cross-entropy loss between the predicted and target labels for the nodes, with class weighting to address the class imbalance described above. The 'unlabeled' class is weighted with 0.1112 and the key item classes with 1.0. The weighting increases the misclassification cost for key items compared to the 'unlabeled' class. Training batches are constructed from batches of documents, using 8 invoices per batch. The invoices in the training set are shuffled. ADAM (Kingma and

Ba, 2015) is used as optimizer with the standard configuration $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 1e-8$. We employ gradient clipping to control for the exploding gradients problem in recurrent layers; all gradients with L^2 norm over 0.5 are clipped. A fixed stepwise learning rate (α) schedule is applied: The model training starts with $\alpha = 5.4452 * 10^{(-4)}$, and α is decreased by factor 10 every 50 epochs. We furthermore use an early stopping criterion, which aims to maximize the macro F_1 score on the validation set with a patience of 50 epochs.

Table III.3: Selected model specification

Layer	Size	# Attention heads
FC	256	-
GRU	128	-
GAT 1	397	12
GAT 2	192	8
GAT 3	512	1

The above described model specification (layer sizes, number of attention heads per layer) and training hyperparameters (α, batch size, weighting of the "unlabeled" class) were selected using a hyperparameter search with Hyperband (Li et al., 2017). 50 hyperparameter configurations were tried with the objective to maximize the macro F_1 score on the validation set. For each configuration, the model was trained for a minimum of 10 and a maximum of 60 epochs. The best configuration achieved a macro F_1 score (incl. the "unlabeled" class) of 0.8956 after 60 epochs.

III.5 Results

We report F_1, precision, and recall scores for the extracted key items. We furthermore include their macro averages, both including and excluding the unlabeled class. The scores are calculated by comparing the model outputs with the annotated instances in the test set. The model outputs were generated using the model state after the completion of epoch 64, as the early stopping criterion ended the model training on epoch 115. We furthermore analyze the attention weights inferred on the document graph edges by the model.

Table III.4: Classification results on test set

	Unlabeled	Invoice number	Total amount	Invoice date	Macro avg. incl. unlabeled	Macro avg. excl. unlabeled
Precision	0.9959	0.9391	0.8333	0.9196	0.9220	0.8974
Recall	0.9971	0.8571	0.8072	0.8996	0.8902	0.8546
F_1 Score	0.9965	0.8963	0.8200	0.9095	0.9055	0.8753

III.5.1 Classification results

Table III.4 shows the classification results on the test set. The model is able to detect all three key items, though performing better on *invoice numbers* and *invoice dates* than *total amounts*. For *total amounts* and *invoice dates*, precision and recall are well balanced, leading to reasonable high F_1 scores. For *invoice numbers*, the spread between precision and recall is higher.

This can also be seen in Figure III.4, which shows a plot of the precision and recall scores for different probability thresholds (precision-recall curve), along with isometric lines for several levels of F_1 scores. The curve for *invoice numbers* indicates high trade-off costs between precision and recall if a recall over 0.9 was to be achieved. Generally, the figure shows that higher F_1 scores are attainable through threshold moving; F_1 scores over 0.9 for *invoice numbers* and close to 0.85 for *total amounts* could be achieved.

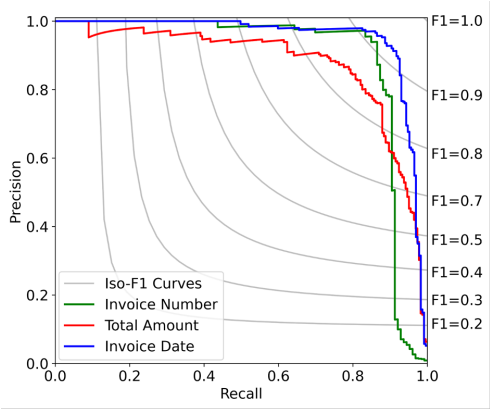

Figure III.4: Precision-recall curves for different probability thresholds for the key items invoice number, total amount, and invoice date

III.5.2 Attention analysis

To classify a node $v^{(i)}$, the model has not only access to the node's own features but to all features of $N(i)$. The distinguishing ability of GAT is to perform feature aggregation weighted by the attention weights allocated to the connecting edges e_{ij}. Analyzing the attention weights inferred by the model therefore allows to gain an understanding if and how contextual relationships affect the node classification. To this end, we analyze the attention weights allocated by the model on the edges of the document graphs in the test set.

Figure III.5 depicts the distributions of attention weights inferred by the last GAT layer on the edges which connect all nodes classified as key items with their surrounding nodes. For *invoice numbers*, the model infers sharp attention weights, i.e. the weights tend to be closer to either 0.0 or 1.0, resembling a bimodal distribution. For *invoice dates* and *total amounts*, this distribution is flatter; weights in between the two extremes are allocated more frequently. As the model allocates weights close to 0.0 very often, we furthermore narrow the analysis down to the attention allocated towards the self-loops. This way, we can analyze whether the model disregards the neighborhood features in favor of the node features or vice versa.

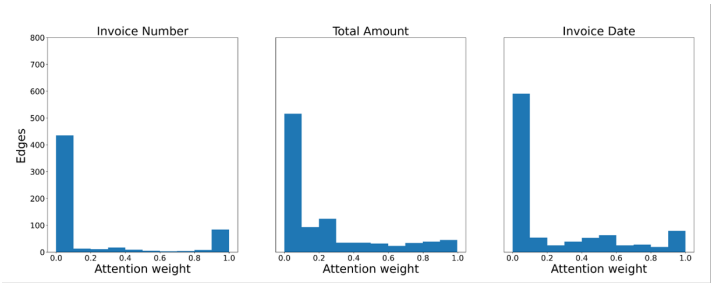

Figure III.5: Distribution of attention weights over all edges on predicted key item nodes

Figure III.6 shows the distribution of attention weights on the node's self-loop edges e_{ii}. The model infers primarily very small attention weights for *invoice numbers*, which indicates that the classified node's own features informed the classification not as much as the features of its neighboring nodes $v^{(j)}$. In the case of *total amounts* and *invoice dates*, the self-loop edges received much higher attention weights. In summary, the attention weights on the self-loops imply that *invoice numbers* are rather identified via their context, whereas *total amounts* and *invoice dates* are identified via their own features.

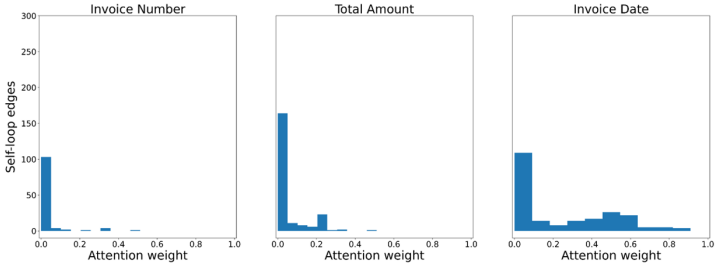

Figure III.6: Distribution of attention weights over self-loop edges on predicted key item nodes

This is also reflected in Figure III.7: The figure depicts graphs that summarize the attention weights inferred on the edges of the 2-hop neighborhood of predicted key items. The attention weights are summed by similar words. The thickness of the edges connecting the tokens reflects the attention placed by the model on the respective relationships, summed across all document graphs in the test set. For each key item, we exemplarily choose the 10 terms which have received the highest attention weights. In the figure, <Key item> denotes a neighboring token which has also been classified as a key item, <Self> denotes the attention on the self-loop.

In the case of *invoice numbers*, the model assigns most attention to combinations of words that form some variation of "invoice number." For the invoice date, the model allocates most attention to the self-loop and to neighboring nodes which have also been classified as *invoice dates*. Similarly, the classification of *total amounts* is mainly based on the node's own features, and the context receives only small attention weights. For these two key items, their context is less important for their classification than their own features.

The analysis of attention weights shows that the model classifies nodes both based on the node's own features, as well as based on the features of neighboring nodes. The

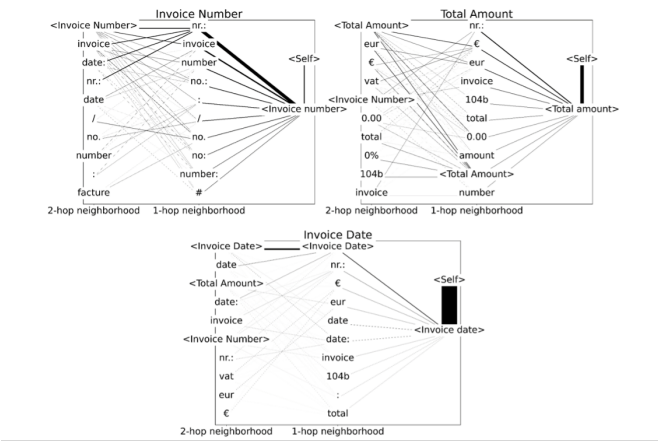

Figure III.7: Attention weights allocated towards the edges of extracted key items, summed across unique words

weighting hereby varies by key item class. Furthermore, it shows that the model is capable of identifying contextual relationships over multiple hops in the graph. This is highly relevant as discriminative terms such as “invoice number”, “total amount” and “invoice date” are usually composed of multiple words.

III.6 Discussion

Prior research has presented promising results for graph-based approaches to EII. However, these were tested on invoice datasets with a low variety of invoice layouts. The purpose of this research is therefore to evaluate the effectiveness of a graph-based approach on a dataset containing invoices from a multitude of vendors, based on a novel model architecture. Our results show that the model is able to capture the patterns prevalent on invoices to extract the defined key items. They further show that the model emphasizes either the context of a word or the word itself to classify it, depending on the key item class. One interesting finding is that the model identifies relevant context over multiple hops in the graph, thereby combining multiple words.

The intended area of application for the model proposed in this paper is the test of details in audits. The overall goal of the test of details is to see whether the details (key items) from the invoices have been recorded correctly. To this end, the key items extracted by the model are reconciled against other, mostly structured, sources of data. Hereby, false positives and false negatives incur different costs with respect to the overall goal of automating EII: False positives (falsely detected key items) require the active attention of an auditor. False negatives (falsely not detected key items), can be offset by further testing of details until sufficient audit evidence is collected. In that regard, the current performance of the model is adequate, yet offers room for further improvement. The model achieves macro averaged precision and recall of 0.8974 and 0.8546 on the key items. While the current performance of the model would not be enough to fully automate this task, it can still lead to efficiency gains in an audit engagement, as long as the effort to review possible false positives is smaller than the effort to extract all key items from the invoices by hand. The raw outputs of the model could also be further enhanced by heuristics and business rules to reduce false positives, which we did not explore in this

paper. For example, rules could be applied which retrieve only the key items with the highest probability per document. Another possibility is to perform logical checks on the data type of a retrieved key item, e.g. evaluating whether retrieved *invoice dates* can be parsed as dates. For full automation of the task, the model should be tunable such that precisions close to 1.0 can be achieved by threshold moving, without sacrificing too much recall. In that case, it could substitute human labor by only requiring limited amounts of additional testing.

In the usage of graphs, our work relates to the approaches by Lohani et al. (2019) and Liu et al. (2019). As anticipated, we do not match their respective results. We achieve F1 scores on the invoice number, total amount, and invoice date of 0.8963, 0.8200, and 0.9095, while Liu et al. (ibid.) report 0.961, 0.910, and 0.963 for the respective key items, and Lohani et al. (2019) report 0.90, 0.99, and 0.95. For Liu et al. (2019), this gap in performance can be attributed to the much higher layout variety in our dataset. As Lohani et al. (2019) do not report in detail on the variety in their dataset, we can only safely attribute this performance gap to the difference in size: Our dataset is smaller than the invoice datasets used in related works. The differences in variety and size render the results hardly comparable. This also applies to a comparison with the results presented by Majumder et al. (2020), which use the biggest and most diverse dataset in all of the related research. However, similar to our results and the results of Liu et al. (2019), they report worse performance on *total amounts* than the other two key items. They achieve F1 scores of 0.949 and 0.940 on *invoice numbers* and *invoice dates*, and 0.858 on *total amounts*. A possible explanation for this is that *total amounts* can appear multiple times on one invoice and are difficult to identify by their context.

In general, we see advantages to the graph approach: The model can access an arbitrary number of direct or indirect neighbors of each node, instead of restricting it to a fixed-sized number of neighbors, like in the approach used by Majumder et al. (2020). This advantage is however contrasted by the computational cost of constructing the document graph in the first place. To extract node embeddings from the document graph, we use GAT layers, similar to Liu et al. (2019). This way, edges between nodes can be individually weighted, unlike the graph convolutions used in Lohani et al. (2019). GAT networks are also better suited for transductive graph learning tasks (Veličković et al., 2018). Hence, document graphs can be processed individually instead of requiring one large graph composed of multiple documents for both learning and inference. Our approach deviates from Liu et al. (2019) in the granularity of the graph: Similar to Lohani et al. (2019), we construct the graph from single words instead of using paragraph-like text blocks.

The presented research must be seen in light of some limitations, which are mainly grounded in the dataset. First, multiple recurring vendors are present in the dataset, hence recurring layouts. Though we try to control for this effect by applying stratified sampling in the training, validation, and test splits, overfitting might still be an issue. Furthermore, all invoices in our dataset are addressed to the same recipient. These however are realistic circumstances for the audit domain, where a client might have recurrent business with suppliers, and all invoices are addressed to the client. Second, we only used English invoices. We have yet to assess how the model responds to invoices from several languages. A further limitation is grounded in the size of the dataset; compared to other works in the area, our dataset is quite small. The classification results of our

models are therefore not comparable to other research.

III.7 Conclusion and outlook

EII is a highly structured, repetitive task in auditing, which can highly benefit from automation. The high variety of invoice layouts faced by auditors calls for approaches that are able to capture the general patterns on invoices. Recent research in this area has proposed graph-based approaches to EII, showing promising results. However, these approaches have been so far applied to datasets with a low variety of invoice layouts. In this paper, we introduce a novel graph-based model architecture, perform an experiment using a dataset from 277 different vendors. The dataset resembles a realistic setting faced by auditors. We show that the model extracts specified key items with high F_1 scores, by leveraging contextual relationships between words on the invoices. While our results do not match the scores achieved by previous works due to the higher variety of invoice layouts in our dataset, they indicate that graph-based models are capable of learning the general patterns prevalent on invoices and extrapolate them.

As of now, we do not see our research on this topic as concluded. As the dataset used in this research is small compared to other related works, further research with bigger datasets and more invoice layouts needs to be conducted to strengthen our results. Complementary to that, we aim to explore more architectural options for the model, such as replacing the GRU with a convolutional layer to extract character-level word embeddings. In our usage of graphs, we further aim to explore edge features. Interesting features could be both continuous, such as the semantic and spatial distance between words, as well as categorical, such as whether words are linked entities or the direction of the edge. Further research should also include more key items to be extracted; especially line items represent an interesting area of investigation. We also plan to extend this model to extract key items from further document types such as receipts, purchase orders, etc.

As pointed out in section 2, the whole research field of EII lacks comparability. Unfortunately, as of now, there are no publicly available labeled sets of invoices that are sufficient in size and variety to train sophisticated ML models. It could therefore vastly benefit from a study that benchmarks different approaches on the same dataset, or from an openly accessible annotated dataset.

III.8 References

Bacciu, Davide et al. (2020). "A gentle introduction to deep learning for graphs". In: *Neural Networks* 129, pp. 203–221. ISSN: 08936080. DOI: 10.1016/j.neunet.2020.06.006. arXiv: 1912.12693. URL: http://arxiv.org/abs/1912.12693%20https://linkinghub.elsevier.com/retrieve/pii/S0893608020302197.

Cho, Kyunghyun et al. (2014). "Learning phrase representations using RNN encoder-decoder for statistical machine translation". In: *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 1724–1734. DOI: 10.3115/v1/d14-1179. arXiv: 1406.1078.

Dengel, Andreas and Klein Bertin (2002). "smartFIX: A Requirements-Driven System for Document Analysis and Understanding". In: *Lecture Notes in Computer Science* 2423.August 2002, pp. 272–282. DOI: 10.1007/3-540-45869-7. URL: http://www.springerlink.com/index/C4WW94M0BTJQJ4L6.pdf%7B%5C%%7D5Cnhttp://www.springerlink.com/index/10.1007/3-540-45869-7.

Denk, Timo I. and Christian Reisswig (2019). "BERTgrid: Contextualized Embedding for 2D Document Representation and Understanding". In: *arXiv:1909.04948 [cs]*. arXiv: 1909.04948. URL: http://arxiv.org/abs/1909.04948 (visited on 02/16/2021).

Devlin, Jacob et al. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://www.aclweb.org/anthology/N19-1423.

Esser, Daniel et al. (2012). "Automatic indexing of scanned documents: a layout-based approach". In: *Document Recognition and Retrieval XIX* 8297.May 2014, 82970H. ISSN: 0277786X. DOI: 10.1117/12.908542.

Frey, Carl Benedikt and Michael A. Osborne (2017). "The future of employment: How susceptible are jobs to computerisation?" In: *Technological Forecasting and Social Change* 114, pp. 254–280. ISSN: 00401625. DOI: 10.1016/j.techfore.2016.08.019. URL: https://linkinghub.elsevier.com/retrieve/pii/S0040162516302244 (visited on 05/04/2021).

IAASB (2009). *International Standard On Auditing 330 The Auditor's Responses To Assessed Risks*. URL: https://www.ifac.org/system/files/downloads/a019-2010-iaasb-handbook-isa-330.pdf.

Karpathy, Andrej et al. (2015). "Visualizing and Understanding Recurrent Networks". In: *ArXiv Preprint* 1506.02078, pp. 1–12. arXiv: 1506.02078. URL: http://arxiv.org/abs/1506.02078.

Katti, Anoop Raveendra et al. (2018). "Chargrid: Towards Understanding 2D Documents". In: *arXiv:1809.08799 [cs]*. arXiv: 1809.08799. URL: http://arxiv.org/abs/1809.08799 (visited on 02/16/2021).

Kingma, Diederik P. and Jimmy Lei Ba (2015). "Adam: A method for stochastic optimization". In: *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pp. 1–15. arXiv: 1412.6980.

Li, Lisha et al. (2017). "Hyperband: A novel bandit-based approach to hyperparameter optimization". In: *The Journal of Machine Learning Research* 18.1. Publisher: JMLR.org, pp. 6765–6816.

Liu, Xiaojing et al. (2019). "Graph Convolution for Multimodal Information Extraction from Visually Rich Documents". In: *arXiv:1903.11279 [cs]*. arXiv: 1903.11279. URL: http://arxiv.org/abs/1903.11279 (visited on 04/19/2021).

Lohani, D. et al. (2019). "An Invoice Reading System Using a Graph Convolutional Network". In: *Computer Vision – ACCV 2018 Workshops*. Lecture Notes in Computer Science 11367. Ed. by Gustavo Carneiro and Shaodi You, pp. 144–158. DOI: 10.1007/978-3-030-21074-8_12. URL: http://link.springer.com/10.1007/978-3-030-21074-8_12 (visited on 04/19/2021).

Majumder, Bodhisattwa Prasad et al. (2020). "Representation Learning for Information Extraction from Form-like Documents". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, pp. 6495–6504. DOI: 10.18653/v1/2020.acl-main.580. URL: https://www.aclweb.org/anthology/2020.acl-main.580 (visited on 04/19/2021).

Palm, Rasmus Berg et al. (2017). "CloudScan - A configuration-free invoice analysis system using recurrent neural networks". In: *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. Vol. 01, pp. 406–413. DOI: 10.1109/ICDAR.2017.74. arXiv: 1708.07403. URL: http://arxiv.org/abs/1708.07403 (visited on 05/08/2021).

Schuster, Daniel et al. (2013). "Intellix - End-user trained information extraction for document archiving". In: *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, pp. 101–105. ISSN: 15205363. DOI: 10.1109/ICDAR.2013.28.

Veličković, Petar et al. (2018). "Graph attention networks". In: *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, pp. 1–12. arXiv: 1710.10903.

Wang, Minjie Yu (2019). "Deep Graph Library: Towards efficient and scalable deep learning on graphs". In: *ICLR Workshop on Representation Learning on Graphs and Manifolds*. URL: https://par.nsf.gov/biblio/10311680.

Chapter IV

Automated Invoice Processing: Machine Learning-Based Information Extraction for Long Tail Suppliers

Outline

IV.1 Introduction 127
IV.2 Related Literature 128
IV.3 Study design 134
IV.4 Implementation 139
IV.5 Results 142
IV.6 Discussion 146
IV.7 Conclusion 148
IV.8 References 149

Bibliographic Information

Krieger, F., Drews, P., Funk, B. (2023). "Automated Invoice Processing: Machine Learning-Based Information Extraction for Long Tail Suppliers". Manuscript submitted for publication.

Author's contribution

The author's share of the publication is 80%. Table C.4 in appendix C shows the contributions of all authors of the publication in detail.

Copyright Notice

Abstract

Automation of the processing of incoming invoices promises to yield vast efficiency improvements in accounting. Until a universal adoption of fully electronic invoice exchange formats has been achieved, machine learning can help bridge the adoption gaps in electronic invoicing by extracting structured information from unstructured invoice formats. Machine learning especially helps the processing of invoices of suppliers who only send invoices infrequently, as the models are able to capture the semantic and visual cues of invoices and generalize them to previously unknown invoice layouts. Since the population of invoices in many companies is skewed toward a few frequent suppliers and their layouts, this research examines the effects of training data taken from such populations on the predictive quality of different machine-learning approaches for the extraction of information from invoices. Comparing the different approaches, we find that they are affected to varying degrees by skewed layout populations: The accuracy gap between in-sample and out-of-sample layouts is much higher in the Chargrid and random forest models than in the LayoutLM transformer model, which also exhibits the best overall predictive quality. To arrive at this finding, we designed and implemented a research pipeline that pays special attention to the distribution of layouts in the splitting of data and the evaluation of the models.

IV.1 Introduction

Invoices are essential documents for different business processes such as procurement and accounts payable. They hold the details of transactions between clients and suppliers, due to which they also bear legal value (Cristani et al., 2018) and are frequently used by external auditors as evidence for the existence of transactions and their correct recording (ISA 330, 2009; Krieger et al., 2021).

Invoicing is becoming increasingly digitalized, which increases the productivity and efficiency of the associated processes (Baviskar, Ahirrao, Potdar, et al., 2021; Cristani et al., 2018; Tanner and Richter, 2018). Electronic invoicing encompasses different degrees of digitization, from printed invoices, which are scanned over natively electronically readable invoices in PDF format that are exchanged by email, to invoicing as an integrated business-to-business process via electronic data interchange (EDI) (Tanner and Richter, 2018).

PDFs contain information in an unstructured format and require an information extraction (IE) (Sarawagi, 2008) step to gain a semantic representation. Invoices are, however, a particularly interesting type of business document, which makes the IE step simultaneously challenging and significant. While there are several pieces of information that can be expected to appear on an invoice, their positioning and schematical organization, the layout, is usually unregulated and to be freely decided on by the issuing company (Cristani et al., 2018).

In EDI-based invoicing, on the other hand, the information is directly delivered in a structured format, usually XML based, which facilitates any downstream processing (Tanner and Richter, 2018). Despite this advantage, the adoption of EDI invoicing is still progressing slowly, which can be attributed in part to the fact that companies benefit to varying degrees from its adoption (ibid.).

Whereas companies receiving large quantities of invoices benefit heavily from EDI invoicing, the adoption barriers might be too great for companies for which this is not the case. Thus, companies looking to adopt EDI invoicing might need to actively onboard their suppliers; they can achieve this onboarding most easily with suppliers from which they receive a large quantity of invoices. Infrequent suppliers, which make up the biggest share of suppliers of most larger companies (Klein et al., 2004), are less inclined to adopt EDI invoicing, as they do not benefit to the same extent from its adoption (Tanner and Richter, 2018). IE from unstructured invoices therefore represents an attractive intermediate step toward automated invoice processing until EDI-based invoicing has universally superseded PDFs and paper-based invoices.

As mentioned before, the challenge for IE lies in the diversity of these long tail suppliers and the resulting variety of layouts (Cristani et al., 2018). Template-based IE solutions require a substantial amount of human involvement to create and maintain the different supplier templates, which can number hundreds (ibid.) or even thousands (Klein et al., 2004) and therefore offset the efficiency gains from automation. This shortcoming of template-based IE can be addressed through machine learning (ML).

A growing body of research proposes different ML approaches for IE from invoices, which would allow the training of models that are able to generalize to previously unseen invoice layouts. These approaches employ precomputed or learned structural representations that embed the visual and semantic properties of the invoices while preserving the layout. They can be broadly classified into graph-based (Krieger et al., 2021; Liu, Gao, et al., 2019; Lohani et al., 2019; Yu et al., 2020) and grid-based (Denk and Reisswig, 2019; Katti et al., 2018; Zhang, Xu, et al., 2020; Zhao et al., 2019) approaches, as well as approaches relying on positional embeddings (Garncarek et al., 2021; Majumder et al., 2020; Xu, Xu, et al., 2022; Xu, Li, et al., 2020). However, ML models are sensitive to the biases in the data being used for training (Shah et al., 2020). One crucial type thereof, selection bias, can originate from samples being used for training that are not representative of the population to which the model is then applied (ibid.). Given the structure of suppliers present in many larger companies—few high-frequency suppliers, many low-frequency long tail suppliers— we are interested in understanding how the layout-aware ML approaches respond to training data taken from such a population. Therefore, the question that has motivated us to conduct this research is the following:

> *How do ML-based approaches to IE from invoices respond to skewed vendor distributions?*

The contribution of this paper is twofold. First, we show that ML models can suffer from layout bias and that different models are affected to varying degrees by this bias. We train and evaluate different ML-based approaches to IE from a set of invoices that is skewed toward a few suppliers. The evaluation results are then disaggregated into in-sample and out-of-sample layouts, showing that all models are more accurate[1] on in-sample layouts. We also contribute to the literature by conducting a benchmark of different layout-aware ML models over a common set of invoices using a common evaluation metric. The results show that the pretrained transformer model LayoutLM (Xu, Li, et al., 2020) outperforms all other models in our benchmark by a comfortable margin, especially for out-of-sample layouts. LayoutLM is also the most robust against layout bias. The results further indicate that the other models in our study have different strengths with respect to detecting certain information entities.

The remainder of this paper is structured as follows: In section IV.2, we review relevant previous research and delineate the research gap. In section IV.3, we describe the employed methods and materials, after which the details of the model implementation are provided (section IV.4). Thereafter, we present (section IV.5) and discuss (section IV.6) the results, and we conclude our research in section IV.7.

IV.2 Related Literature

Due to the large possible varieties of layouts in invoices, they are considered a particularly challenging and interesting case for automated business document processing (Cristani et al., 2018). In essence, any form of automation requires the information contained on invoices to be made available for downstream applications in a structured fashion.

[1] In this paper, we refer to the "accuracy" of an ML model as its predictive quality, not as the evaluation metric.

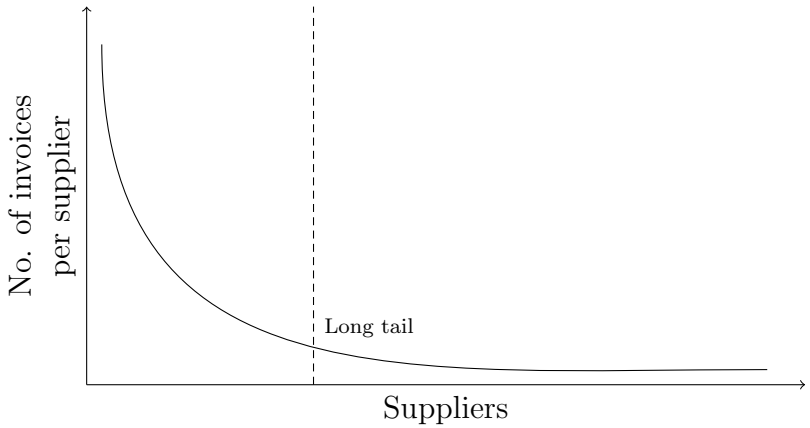

Figure IV.1: Distribution of invoices per supplier as described by Koch (2019). Figure based on Koch (2019) and Tanner and Richter (2018).

IV.2.1 Challenges Associated with the Automated Processing of Invoices

The information on invoices can be broadly classified into two types of entities: header fields and line items. Header fields encompass, inter alia, the invoice number, the issue date, the total amount due, and the value-added-tax identification number (VAT ID) of the supplier (Cristani et al., 2018). Line items provide the details of the exchanged goods or services, such as a description of the provided good or service, the unit price, the number of units provided, and the total amount due per line item (Katti et al., 2018). EDI-based electronic invoicing facilitates the capture of invoice data by directly providing the information in a structured, usually XML-based, format (Tanner and Richter, 2018). However, the adoption of EDI invoicing remains limited (Koch, 2017; Tanner and Richter, 2018). Its main beneficiaries are large companies that receive large quantities of invoices. For many of their smaller suppliers, the incentive is weaker in the face of the organizational (Tanner and Richter, 2018) and technical (Cristani et al., 2018; Tanner and Richter, 2018) obstacles associated with the adoption of EDI-based invoicing. According to Koch (2017), even several years after adopting an EDI invoicing system, only 25%–30% of the invoices received by most large companies are fully electronic, and those they do receive are usually from business partners with whom they share a large number of transactions. This, however, represents only a small proportion of an organization's suppliers, as typically most suppliers only send invoices infrequently (Koch, 2019). The distribution of invoices received per supplier is therefore skewed toward a few high-frequency suppliers; for this reason, Tanner and Richter (2018) refer to lower-frequency suppliers, usually small and medium-sized enterprises, as the long tail. Figure IV.1 provides a visual example of this distribution. This adoption gap can be bridged by technologies that allow the extraction of information from unstructured formats such as document images or PDFs.

IV.2.2 Machine Learning for IE from Invoices

Early systems proposed for IE from business documents relied on manually preconfigured layout templates and rules (Baviskar, Ahirrao, Potdar, et al., 2021; Cristani et al., 2018; Klein et al., 2004). A pioneering study in this field was presented by Palm et al. (2017). The authors introduced "CloudScan," a system that serializes the text of invoices into

sequences of tokens and performs sequence tagging using a recurrent neural network to extract key-value pairs. CloudScan, however, dismisses the layout of the invoice documents, forcing the authors to decide on the ordering of the words (Palm et al., 2017). Subsequent approaches have addressed this shortcoming by developing document representations that retain the layout information. The approaches can broadly be classified into graph-based approaches (Liu, Gao, et al., 2019; Lohani et al., 2019; Yu et al., 2020), grid-based approaches (Denk and Reisswig, 2019; Katti et al., 2018; Zhao et al., 2019), and positional embeddings (Garncarek et al., 2021; Majumder et al., 2020; Xu, Xu, et al., 2022; Xu, Li, et al., 2020).

Graph-based approaches model the document as a graph, in which text elements (words or paragraphs) are represented as nodes and the edges represent the spatial relationship between the text elements. The task of extracting key-value pairs is then modeled as node classification (Lohani et al., 2019; Yu et al., 2020; Zhang, Xu, et al., 2020) or sequence tagging (Liu, Gao, et al., 2019), and graph convolution is employed to capture the contextual information between the text elements.

Grid-based approaches map the document text to a grid such that the relative spatial positions of the text elements are preserved. The grids can be of varying granularity: Chargrid (Katti et al., 2018) and BERTgrid (Denk and Reisswig, 2019) employ pixel-level grids, whereas CUTIE (Zhao et al., 2019) uses a coarser grid, in which each word is mapped to a single cell. The extraction of key-value pairs is then modeled as a semantic segmentation or object detection task (Denk and Reisswig, 2019; Katti et al., 2018; Zhao et al., 2019). Chargrid and BERTgrid also combine semantic segmentation with object detection to detect line item bounding boxes (Denk and Reisswig, 2019; Katti et al., 2018). The respective models capture contextual patterns through (grid) convolution mechanisms common in computer vision.

Another stream of study proposes models based on transformer architecture (Garncarek et al., 2021; Xu, Xu, et al., 2022; Xu, Li, et al., 2020) and the self-attention-based mechanism (Majumder et al., 2020). The layout of the documents is accounted for by using learned (ibid.) or precomputed positional embeddings (Garncarek et al., 2021; Xu, Xu, et al., 2022; Xu, Li, et al., 2020), which use the two-dimensional coordinates of the respective text's bounding box, analogical to the positional encodings proposed in Vaswani et al. (2017). The transformer-based models further employ pretraining strategies like BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) over a multitude of different layout-rich document types, aiming for a general understanding of documents. The pretrained models can then be fine-tuned for different downstream tasks over different types of layout-rich documents such as token classification to extract information from invoices.

IV.2.3 Limits to Comparability across Different Approaches to IE from Invoices

All of the abovementioned studies address the problem of extracting information from invoices by using a different approach, in terms of either document representation, model architecture, or input features. Although they all present promising results, a direct comparison between approaches from the literature is hardly possible. Table IV.1 summarizes

Table IV.1: Achieved results reported in the related literature along with the used data sets, evaluation metrics and baselines. Proprietary data sets are indicated with the abbreviation (pr.).

Model (Reference)	Data set(s)	Evaluation metrics	Reported aggregated results	Baselines
BERTgrid (Denk and Reisswig, 2019)	Invoices (pr.)	Character-level accuracy (Word-error rate) for three header entities and one line item entity, as well as aggregated over all entities and line item entities	65.48% ±0.58	Chargrid, Wordgrid (similar to CUTIE), variants of the proposed model
LAMBERT (Garncarek et al., 2021)	Mix of open and proprietary data sets for training, open data sets for evaluation (SROIE, CORD)	Entity-level F_1 scores averaged across entities for SROIE and CORD	94.41 F_1 on CORD, 98.17 F_1 on SROIE	RoBERTa (Liu, Ott, et al., 2019), LayoutLM, LayoutLMv2
Chargrid (Katti et al., 2018)	Invoices (pr.)	Character-level accuracy (Word-error rate) for five header entities and three line item entities	61,99% Macro average computed from the results reported for the individual entity types	Sequential (Bidirectional GRUs), variants of the proposed model
GAT+BiLSTM-CRF (Liu, Gao, et al., 2019)	Invoices (pr.), receipts (pr.)	F_1 score for six header entities for the invoice data set, averaged F_1 scores across all entities for invoices and receipts	0.873 F_1 on invoices, 0.836 F_1 on receipts	Sequential (BiLSTM+CRF)
GCN (Lohani et al., 2019)	Invoices (pr.)	Word-level F_1 score, precision, recall for 36 entities, including line item entities	0.93 F_1 (micro) on invoices	None
Self-attention (Majumder et al., 2020)	Invoices (pr.), receipts (SROIE)	F_1 score, ROC-AUC for seven resp. two header entities for invoices and receipts	0.878 F_1 (macro) on invoices	Variants of the proposed model
LayoutLM (Xu, Li, et al., 2020)	Open data sets for training and evaluation (incl. SROIE)	Entity-level precision, recall, F_1 score for SROIE	0.9524 F_1 on SROIE	Previous best on the SROIE leaderboard
LayoutLMv2 (Xu, Xu, et al., 2022)	Open data sets for training and evaluation (incl. SROIE)	Entity-level precision, recall, F_1 score for SROIE	0.9601 F_1 on CORD, 0.9781 F_1 on SROIE	BERT (Devlin et al., 2019), UniLMv2 (Bao et al., 2020), LayoutLM (Xu, Li, et al., 2020); SROIE leaderboard, including PICK (Yu et al., 2020) and TRIE (Zhang, Xu, et al., 2020)
PICK (Yu et al., 2020)	Medical invoices (pr.), train tickets (pr.), receipts (SROIE)	Mean entity precision (mEP), recall (mER), and F score (mEF) for six header entities on the medical invoice data set, as well as aggregated mEF on the train ticket and SROIE data sets	98.6 mEF on train tickets, 96.1 mEF on SROIE, 87.0 mEF on medical invoices	Sequential (BiLSTM+CRF), LayoutLM (Xu, Li, et al., 2020) (only on SROIE)
CUTIE (Zhao et al., 2019)	Receipts (pr.; SROIE)	Average precision (AP) and soft AP (tolerance for false positives) across keys for different receipt types, including SROIE	94.0, 81.5, 74.6 AP on different receipt type subsets	Sequential (Palm et al., 2017), BERT (Devlin et al., 2019), variants of the proposed model
TRIE (Zhang, Xu, et al., 2020)	Taxi invoices (pr.), receipts (SROIE), resumes (pr.)	The quantity of goods or services provided. In case of services, this might relate to temporal units.	93.26 F_1 on taxi invoices, 96.18 F_1 on SROIE, 76.3 F_1 on resumes	Chargrid, Sequential (Ma and Hovy, 2016), GAT+BiLSTM-CRF (Liu, Gao, et al., 2019)

the used data sets, evaluation metrics, and evaluation baselines of the studies. It shows the basis on which the limited comparability between the proposed approaches is grounded. First, most of the studies use some sort of proprietary set of invoices; this is because invoices often contain highly sensitive data and are therefore not made available to the general public (Baviskar, Ahirrao, and Kotecha, 2021; Baviskar, Ahirrao, Potdar, et al., 2021). The next reason is the granularity of the reported results: If results are reported for individual entity types, there is no established common set of entity types to be extracted from the invoices. This primarily affects line items: Only a few studies (Denk and Reisswig, 2019; Katti et al., 2018; Lohani et al., 2019) report the model's performance on line item entities.

A further limitation is given through the evaluation methodology. As can be seen in Table IV.1, the performance of the models is evaluated on different levels of granularity. The evaluations range from character-level (Denk and Reisswig, 2019; Katti et al., 2018) through word-level (Lohani et al., 2019) to entity-level (Garncarek et al., 2021; Xu, Xu, et al., 2022; Xu, Li, et al., 2020; Yu et al., 2020). Most studies use established measures in information retrieval such as precision, recall, F_1 score, and average precision to evaluate the performance of their approach. With respect to the different levels of evaluation, however, it is not always clearly stated how the results are aggregated from lower-level predictions to higher-level measures. For example, the transformer-based models LAMBERT, LayoutLM and LayoutLMv2 employ byte-pair-encoding (BPE) (Sennrich et al., 2016) and WordPiece (Wu et al., 2016) subword tokenization, which subsequently yield subword-level predictions. Here, only Garncarek et al. (2021) state the use of the geometric mean as an aggregation method to arrive at entity-level measures from the predictions. In the case of Chargrid and BERTgrid, the evaluation is only conducted at the character level and not further aggregated at all; the authors use an accuracy measure similar to the word error rate (Denk and Reisswig, 2019; Katti et al., 2018). Due to the use of this metric, Chargrid and BERTgrid may not be compared to the other models that have not been used as evaluation baselines in the respective studies.

The use of baselines is another factor affecting the comparability between approaches. Especially in experiments where proprietary data sets have been used (Denk and Reisswig, 2019; Katti et al., 2018; Liu, Gao, et al., 2019; Lohani et al., 2019; Majumder et al., 2020; Yu et al., 2020; Zhang, Xu, et al., 2020; Zhao et al., 2019), the authors resort to using a sequential model after the example of CloudScan and variants of their own models. Only a few studies use non-sequential approaches as baselines; in addition, Denk and Reisswig (2019) benchmark their proposed model against a variant of CUTIE, and Zhang, Xu, et al. (2020) use Chargrid and the graph attention-based model proposed in Liu, Gao, et al. (2019) as baselines.

In conclusion, the utilization of proprietary data sets, different evaluation methodologies and metrics, and the inconsistent use of baselines lead to poor comparability of different approaches in this field of research.

IV.2.4 Open Data Sets for IE from Business Documents and Previous Benchmarking Studies on IE from Invoices

A partial remedy to the points outlined above is offered by the scanned receipts OCR and information extraction (SROIE) challenge (Huang et al., 2019) and the consolidated receipt data set for post-OCR parsing (CORD) (Park et al., 2019), as well as the Kleister data sets (Stanisławek et al., 2021), and the multi-layout invoice document data set (MIDD) (Baviskar, Ahirrao, and Kotecha, 2021).

SROIE and CORD are composed of receipts. Both data sets are similar in size; SROIE contains 973 receipts, of which 347 are dedicated to testing. CORD contains 1,000 receipts, with 800 examples dedicated to training and 100 each for validation and testing. The SROIE data set is annotated for four header entity types: company, address, receipt date, and total amount. The CORD data set, on the other hand, encompasses only line item–like entities; most header fields have been left out due to privacy concerns (Park, 2021).

Stanisławek et al. (2021) introduce two data sets, Kleister-NDA and Kleister-Charity, which exhibit a greater complexity stemming from the multipage nature of the underlying document types. The data sets feature two types of layout-rich documents, nondisclosure agreements (Kleister-NDA) and financial statements (Kleister-Charity), which are annotated for several information entities. The MIDD data set is the only open data set containing invoices. It encompasses the OCR outputs for 630 invoices from four different layouts, annotated for seven invoice header fields. However, MIDD provides only the text and the corresponding annotations without the associated bounding box coordinates, due to which the layout information is lost.

SROIE, CORD, and the Kleister data sets are split into predefined train, validation, and test sets. For the SROIE, CORD, and Kleister-NDA data sets, there is no information provided on the quantity of the different layouts and their distribution in the respective splits, and for Kleister-Charity, the information is only rudimentary (Huang et al., 2019; Park et al., 2019; Stanisławek et al., 2021). Majumder et al. (2020) found that the receipts in the SROIE training and validation sets stem from 234 layouts, with 46 out of 626 receipts pertaining to one layout. As SROIE offers a public ladder in which researchers and practitioners may participate, the test set is kept private.

Introducing the Kleister data sets and Stanisławek et al. (2021) the benchmark sequential (Flair, (Akbik et al., 2019); BERT, (Devlin et al., 2019); RoBERTa, (Liu, Ott, et al., 2019)) and layout-aware models (LAMBERT, (Garncarek et al., 2021); LayoutLM, (Xu, Li, et al., 2020)) on the Kleister data sets shows that the layout-aware LAMBERT model outperforms all other models.Similar benchmarks have also been conducted on invoice data sets: Liu, Zhang, et al. (2016) constructed a comprehensive set of features and compare naive Bayes, logistic regression and support vector machine classifiers on a rather small data set composed of 97 invoices. In their study, the logistic regression and support vector machine classifiers outperformed the naive Bayes by a significant margin. Baviskar, Ahirrao, and Kotecha (2021) compared different word embedding techniques (Word2Vec, GloVe, FastText, embedding layer) for a sequential BiLSTM model on a data set of 1,646 invoices from eight suppliers. They found that embedding words through a dedicated

embedding layer achieved the best results.

To conclude, the studies proposing approaches to IE use mostly proprietary sets of invoices. The distribution of different layouts in the respective data sets is widely left unaddressed. The same applies to the open data sets that are being used to benchmark IE on visually rich documents. This leaves doubt as to whether the distribution of invoice layouts has an effect on the models. In addition, obtaining a comparison of the predictive quality of different layout-aware models is difficult due to the utilization of different evaluation metrics and the reliance on sequential models as baselines. Previous benchmarking studies on IE from invoices employ only either very small data sets (Liu, Zhang, et al., 2016) or data sets with small numbers of invoice layouts Baviskar, Ahirrao, and Kotecha (2021). Furthermore, neither study employs any of the layout-aware approaches introduced in section IV.2.2. We therefore see the need for a study that benchmarks different approaches to IE from invoices and explicitly addresses the distribution of layouts.

IV.3 Study design

The goal of this study is to close the gaps mentioned in the previous section by conducting an independent benchmark study of different ML-based approaches to information extraction from invoices. We devised a pipeline similar to the pipelines employed by Baviskar, Ahirrao, and Kotecha (ibid.), Liu, Zhang, et al. (2016), and Stanisławek et al. (2021), which is pictured in Figure IV.2. The main differences between our pipeline and the one presented in previous research (Baviskar, Ahirrao, and Kotecha, 2021; Liu, Zhang, et al., 2016; Stanisławek et al., 2021) are that we specifically accounted for suppliers—and therefore layouts—in the data splitting. Furthermore, we tuned the neural networks for a determined set of hyperparameters to achieve the best possible accuracy for each model. This section provides details on the data set and the methodology utilized for annotating the data set, tuning and training the models, and evaluating their accuracy.
As there are only few implementations available of the abovementioned models, we had to limit our study to a few selected examples. Our goal is to represent a broad range of methodologies; hence, we have chosen examples for each of the three model types described in section IV.2.2: graph-based, grid-based and transformer-based. We chose Chargrid and BERTgrid as grid-based approaches, and the GCN- and GAT-based models introduced in Lohani et al. (2019) and Liu, Gao, et al. (2019). All of them have been specifically designed for invoice documents. We also included the LayoutLM transformer model, whose authors aimed for a more general understanding of business documents to see how it compares to invoice-specific approaches. In addition to the rather complex neural network models, we included a random forest model in our evaluation as a baseline.

IV.3.1 Data Set

Our data set was composed of 1,059 invoices provided by a large German firm that sources products and services from a multitude of countries across the world. The predominant language of the documents was English (955 invoices), followed by German (76 inv.). The less frequent languages were French (8 inv.), Dutch (8 inv.), Spanish (7 inv.) and Italian (5 inv.). The invoices in our data set documented the procurement of physical goods as well as services. Out of the 1,059 invoices, 63 were multipage documents, resulting in a total of

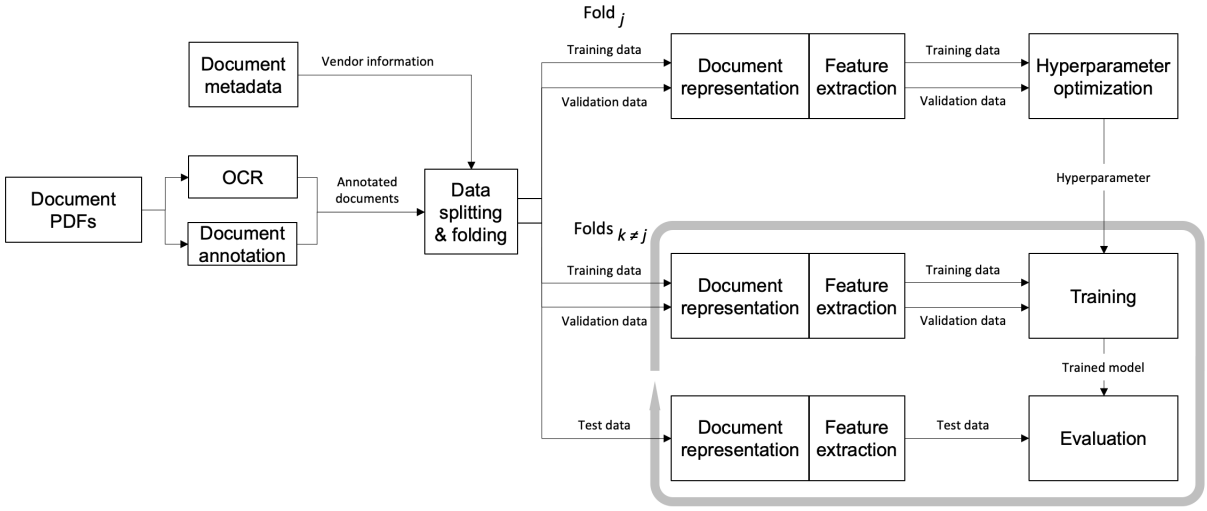

Figure IV.2: Study pipeline. A set of invoices is passed through an optical character recognition (OCR) engine and annotated for several information entities. The annotated OCR outputs are used to train and evaluate different ML models over several iterations (folds). One fold is set aside for hyperparameter optimization.

1,126 invoice pages. The invoices stemmed from 259 different suppliers, reflecting a wide variety of invoice layouts. The distribution of suppliers was, however, skewed. Figure IV.3 plots the distribution of suppliers in the data set. The figure shows that the distribution of suppliers follows a long-tailed distribution; a large subset of invoices stemmed from a handful of suppliers, including one supplier who was disproportionately represented with 348 invoices. While the associated invoices followed the same layout structure, they still exhibited variance between them; the documents could contain different counts of line items, which in turn may be of varying length.

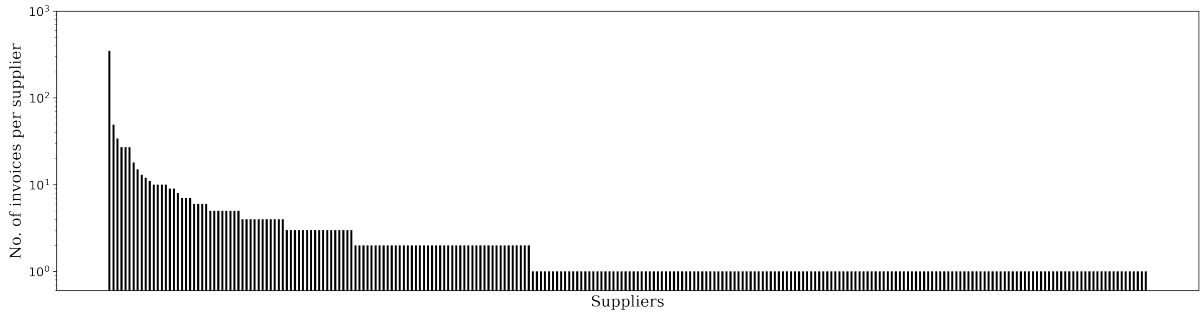

Figure IV.3: Barplot of the supplier distribution in the data set; each bar represents one supplier. The distribution follows the pattern described by Koch (2019).

IV.3.2 Data Set Annotation

We extracted the text from the abovementioned invoices using an optical character recognition (OCR) engine. The OCR step yielded 249,032 word-level text boxes from

Table IV.2: Models selected for benchmarking

Model	Layout representation	Input features	Information extraction	Prediction granularity
BERTgrid	Grid	BERT embeddings	Semantic segmentation, object detection	Pixel-level
Chargrid	Grid	One-hot-encoded characters	Semantic segmentation, object detection	Pixel-level
GCN	Word-level graph	BPE, manually constructed features	Node classification	Word-level
GAT+ BiLSTM-CRF	Paragraph-level graph	Word2Vec (Mikolov et al., 2013) embeddings, manually constructed distance-related edge features	Sequence tagging, node classification	Word-level
LayoutLM	Positional embedding	Tokens	Word classification	Token-level
Random Forest	-	Manually constructed text box and string features	Word classification	Word-level

the 1,126 invoice pages. The documents were annotated by two domain experts for 14 different information entities. The annotation was done by drawing bounding boxes over the document images using Microsoft Azure Machine Learning's labeling tool for object detection. A corresponding class label was assigned to a text box if its area was at least 80% covered by an annotation bounding box. We further assigned a background class to all text boxes that were not sufficiently covered by an annotation box. In total, 54,023 text boxes were assigned an entity class, with recipient address being the most frequently assigned class (13,683 annotated text boxes) and total tax amount the least frequently assigned class (383 annotated text boxes). The data set therefore exhibits a sharp class imbalance, both between the background class and the information entity classes and between the entity classes themselves. Note that some entity types did not appear on every invoice, whereas others could appear multiple times on the same invoice, as was the case with invoice numbers. Table B.1 in the appendix provides the details on the assigned entity classes, along with the data types for each entity class. The data types were also used later as features for the random forest and GCN models. This section and the previous section outline two biases in the data set: a skewed distribution of suppliers and a steep class imbalance. To gain a valuable assessment of the models' accuracy, they must be accounted for during training and evaluation.

IV.3.3 Evaluation

Given the characteristics of the data set, an appropriate evaluation methodology was required to accurately address the research question. The skewed supplier distribution called for an according stratification when the data were split into the respective training, validation, and test sets. On the other hand, the evaluation metric should have been unaffected by the class imbalance.

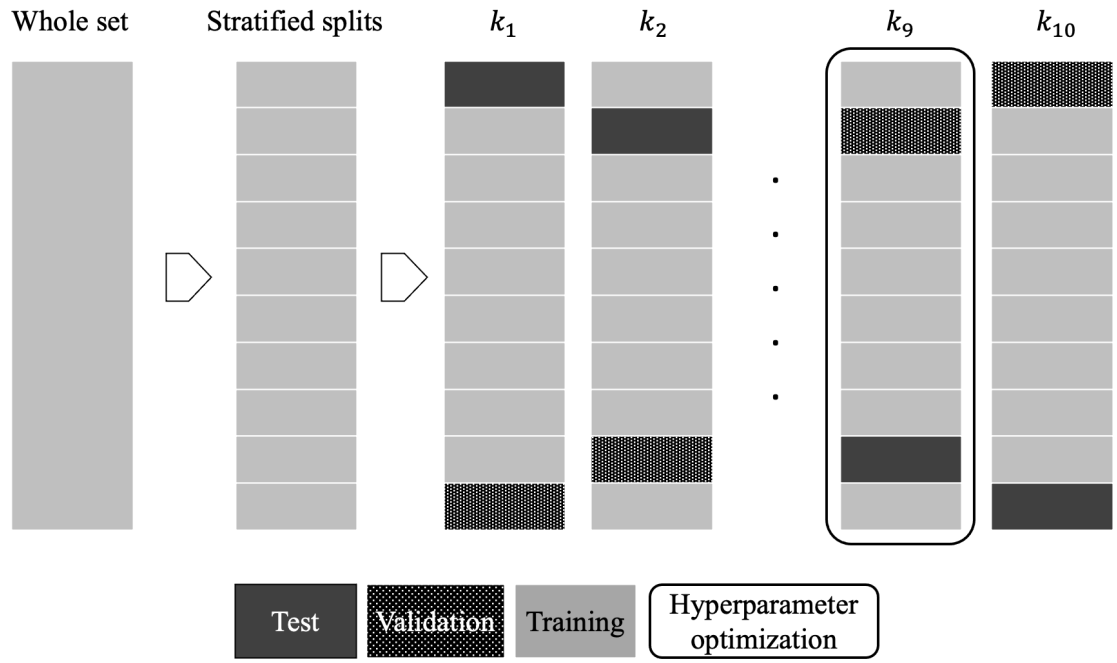

Figure IV.4: The invoices are split into 10 subsets. Each box represents one split; the colors indicate its utilization as training, validation, or testing data in the different folds. The tuning fold is set aside for hyperparameter optimization.

Data Splitting and Folding Strategy

As we wanted to assess the accuracy of the models on both invoices whose layouts have been part of the training set and invoices whose layouts the models have not been exposed to, we ensured that each split contained a proportion of invoices from suppliers that were exclusive to it. In this way, we maximized the variance of invoice layouts in the training set while simultaneously enabling a decomposition of the evaluation results on the respective invoices with in-sample and out-of-sample layouts. To address the stochastic influences governing the data splitting, a 10-fold cross-validation (Han et al., 2011) was employed. This also allowed us to judge the robustness of the models toward varying training and testing data. Figure IV.4 provides a graphical representation of the folding strategy. The documents in the data set were split into 10 (approximately) equally sized, mutually exclusive subsets of invoices. We then generated 10 different folds from the subsets so that each split was used as the test set once. Figure IV.5 shows the distribution of suppliers per split and highlights the suppliers contributing the documents with out-of-sample layouts. To prevent the models from overfitting to the training data, one split in each fold was assigned as the validation set, on which a stopping criterion was applied during the training (see section IV.3.4). One split was set aside as the validation set for the hyperparameter optimization (see section IV.3.5) and subsequently discarded from the evaluation; the evaluation metrics were therefore computed over 9 folds.

Evaluation Metric

To assess the model performance during training and for the final evaluation, we employed the F_1 score, similar to Stanisławek et al. (2021). The score is commonly used in information retrieval and is well suited to judging the predictive quality of a model in the presence of unbalanced class distributions (Sarawagi, 2008). The models were evaluated

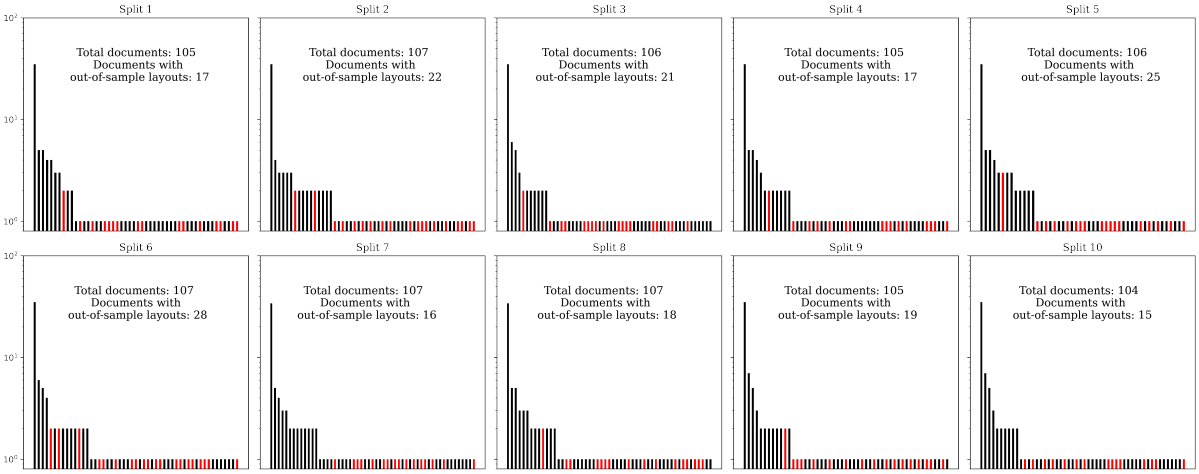

Figure IV.5: Bar charts of the supplier distribution in the individual splits analogous to Figure IV.3. The red bars indicate suppliers that do not appear in the training set (out-of-sample suppliers) if the respective split is used as the test set.

at the text-box level against the ground truth described in section IV.3.2. For the models returning a probability distribution P over the entity types Y, we selected the entity type with the highest probability as the prediction. If the model yielded more granular predictions j associated with one text box i, such as n subword tokens or pixels, the predictions for each text box were aggregated using the geometric mean, as has been used by Garncarek et al. (2021):

$$\hat{y}_i = arg\,max_y \left(\prod_{j=1}^{n} P(Y = y)_{i,j} \right)^{\frac{1}{n}}$$

This way, we ensured the comparability of results across different models that yield predictions of differing granularity (see Table IV.2).

IV.3.4 Training

During training, an early stopping criterion kept the models from overfitting to the training data (Bengio, 2012). We monitored the prediction-level macroaveraged F_1 score, the unweighted mean of the F_1 scores per entity, on the validation set with a patience of 10 epochs. The macroaveraged F_1 score included the background class, as the models needed to be able to distinguish between relevant entities and background text. Where applicable, we followed the class-weighting scheme proposed by Paszke et al. (2016) to address the class imbalance described above, similar to Katti et al. (2018):

$$w_{class} = \frac{1}{ln(c + p_{class})}$$

Using the class frequency p_{class}, the scheme yielded static class weights w_{class}, which affected how much the examples of each class affected the training loss; the greater the weight, the more a falsely predicted instance of the associated class contributed to the loss. The class weights were governed by a single hyperparameter c, which could be easily tuned for during hyperparameter optimization.

IV.3.5 Hyperparameter Optimization

The performance of neural networks is significantly tied to the chosen hyperparameters (Smith, 2018). We therefore optimized their hyperparameters using Bayesian optimization (Snoek et al., 2012), aiming to maximize the prediction-level F_1 score on the validation set. The optimization was performed with one dedicated split as the validation set (see section IV.3.3), to which the fold using the same split as the test set is excluded from the evaluation. Following the recommendations given by Bengio (2012), when working with limited resources, we focused on the learning rate α and the minibatch size B. We further tuned for a third hyperparameter: the class weighting factor c, which governs the distribution of class weights (see section IV.3.4). As the models and data representations consumed varying amounts of GPU memory, we determined the sample interval for B for each model by starting with $B_1 = 2$ and exponentially increasing B until the GPU memory limit was reached. The sample interval for α was determined using a α-range test (Smith, 2017), exponentially increasing α from 10^{-8} to 1.0 using the maximum previously determined B. For c, we chose a sample interval between 1.0 and 2.0. A factor of 1.0 completely leveled out any imbalances between the classes, whereas 2.0 assigned each class identical weights, leaving the class distribution untouched.

IV.4 Implementation

We developed a research environment in Python using different ML packages and frameworks, mainly around the PyTorch ecosystem. We used PyTorch 1.7.0 with CUDA 10.1 to implement, train, and test all models and referred to PyTorch-geometric (Fey and Lenssen, 2019) for any graph layers in the models. The BERTgrid, Chargrid, GAT+BiLSTM-CRF, and GCN models were implemented directly by us; the implementations of LayoutLM and random forest were obtained from the huggingfaces transformers library (Wolf et al., 2020) and scikit-learn (Pedregosa et al., 2011). Tesseract 4.0 was employed as the OCR engine to extract the text boxes from the documents. It should be noted that Tesseract is able to extract paragraph-level text boxes, which, however, can be of arbitrary accuracy on invoices. We utilized this functionality of Tesseract for the GAT+BiLSTM-CRF model (see section IV.4.3). To streamline and execute the experiments of the PyTorch-based models, we employed PyTorch Lightning. All experiments were run on a single computing instance with 6 processor cores, 118 GB RAM, and an Nvidia K80 GPU with 8 GB video RAM. In this section, we provide the implementation details for the individual models. Table IV.3 gives an overview of the chosen hyperparameters that resulted from the optimization step on the neural network models. The following subsections provide the implementation details for the individual models.

Table IV.3: Tuned hyperparameters per model

	α		B		c	
Model	Search interval	Selected	Search interval	Selected	Search interval	Selected
BERTgrid	$\{2.7542\times10^{-4}, 2.7542\times10^{-2}\}$	1.4550×10^{-2}	{2, 4}	2	{1.0, 2.0}	1.2292
Chargrid	$\{2.2909\times10^{-4}, 2.2909\times10^{-2}\}$	2.1209×10^{-2}	{2, 4}	4	{1.0, 2.0}	1.1319
GCN	$\{1.9055\times10^{-6}, 1.9055\times10^{-4}\}$	1.9055×10^{-4}	{2, 64}	16	{1.0, 2.0}	1.2149
GAT+BiLSTM-CRF	$\{3.0200\times10^{-5}, 3.0200\times10^{-3}\}$	9.8624×10^{-4}	{2, 16}	16	-	-
LayoutLM	$\{1.6596\times10^{-6}, 1.6596\times10^{-4}\}$	3.2771×10^{-4}	{2, 8}	4	{1.0, 2.0}	1.2112

IV.4.1 Chargrid

Our implementation of Chargrid deviated in minor aspects from the description given in Katti et al. (2018). We did not oversample invoices with multiple line items during training—the data set was already quite small, with repeating invoice layouts, and we wanted to limit further risk of overfitting. We maintained the target resolution for the Chargrid representations used in the original paper. Furthermore, Katti et al. (ibid.) mention employing anchor boxes (Ren et al., 2015) to support the object detection task; their parametrization is, however, not further specified. Following the recommendations given by Zhang, Lipton, et al. (2021), we therefore decided to generate a total of 13 anchor boxes per pixel, using ten linearly increasing aspect ratios $\mathbf{r}$ (relation of anchor box width to height) from 5 to 50, $\mathbf{r} = \left[5, \ldots, 50\right]$ and four linearly increasing scales $\mathbf{s}$ from 0.1 to 0.4, $\mathbf{s} = \left[0.1, \ldots, 0.4\right]$. The chosen $\mathbf{s}$ and $\mathbf{r}$ form rectangular anchor boxes of flat and wide shape and allowed us to accommodate line items of varying sizes. To evaluate the model, we upsampled the predicted segmentation mask back to the document's original resolution by inverting the interpolation from the Chargrid construction step (Katti et al., 2018). The upsampled segmentation mask was then compared against the ground-truth's labeled text boxes.

IV.4.2 BERTgrid

BERTgrid (Denk and Reisswig, 2019) employs the same model architecture as Chargrid, the main difference between the two being the document representation fed into the model. Whereas Chargrid represents a document by one-hot-encoding characters on the pixel level, BERTgrid embeds tokens using a pretrained BERT model. Our implementation deviated from the approach described in Denk and Reisswig (ibid.) in the usage of said BERT model: We used a standard multilingual pretrained BERT model obtained from the huggingface transformers library (Wolf et al., 2020), which was not further pretrained on invoices. As the BERTgrid document representation yields large tensors and WordPiece tokens are generally coarser than characters, we downsampled the document representation to half the target resolution of our Chargrid representation. In all other aspects, our implementation of BERTgrid equalled our implementation of Chargrid.

IV.4.3 GAT+BiLSTM-CRF

The model described in Liu, Gao, et al. (2019) combines different mechanisms to model contextual relationships; graph-attention layers are combined with BiLSTM layers and a CRF sequence tagging head. For our implementation of the model, we opted for two graph-attention layers in light of the results of the ablation study presented in Liu, Gao, et al. (ibid.). A two-layer multiperceptron (MLP) with ReLu nonlinearity embeds the node-edge-node features in each graph layer. As the authors did not specify an optimizer for training in the original paper, we employed ADAM (Kingma and Ba, 2017), similar to Lohani et al. (2019) and Xu, Li, et al. (2020). The sizes for the model layers—that is, the number of hidden nodes per layer—were not given in the paper. They govern the complexity of the model and therefore the ability of the model to generalize (Bengio, 2009). Hence, we performed a bandit-based random parameter search with successive halving (Karnin et al., 2013) aimed at maximizing the macroaveraged F_1 score on the validation set. The same cross-validation iteration as for the hyperparameter optimization described above was used. The search ran for 50 iterations, each time training the model for 30

epochs. Table IV.4 gives the results of the parameter search. Note that the MLPs that embed the edge-to-edge features have the same number of nodes as the respective GAT layers. To build the underlying document graph as described by Liu, Gao, et al. (2019), we

Table IV.4: Results of the parameter search for the GAT+BiLSTM-CRF model

Layer	Search interval	Selected
BiLSTM_1	[64, 2,048]	512
GAT_1	[64, 2,048]	256
GAT_2	[64, 2,048]	256
BiLSTM_2	[64, 2,048]	512

extracted paragraph-level OCR outputs, which can span text passages ranging from single words to multiple lines. For each word in the paragraph, we extracted word embeddings using pretrained Word2Vec (Mikolov et al., 2013) vectors via gensim (Řehůřek and Sojka, 2010). To enable batch-wise computation, we padded the length of all paragraphs to the length of the longest paragraph in our data set, which was 78 words. We assigned a distinct vector to each padding element and constructed a padding mask for each document indicating non-padding elements for the CRF layer. Out-of-vocabulary words were similarly assigned a designated embedding vector. When evaluating the model, the most likely sequence of tags for each paragraph was obtained using the Viterbi algorithm (Forney, 1973). The sequence was then converted from the IOB tagging scheme back to the original set of labels and compared against the ground truth. Note that each word in a paragraph corresponded to one text box in the ground truth.

Table IV.5: Results of the parameter search for the GCN model

Layer	Search space	Selected
GCN_1	[64, 2,048]	512
GCN_2	[64, 2,048]	256
GCN_3	[64, 2,048]	256
GCN_4	[64, 2,048]	128

IV.4.4 GCN

The model presented by Lohani et al. (2019) employs graph convolution over a precomputed document graph similar to Liu, Gao, et al. (2019); however, it relies solely on node classification to extract any relevant information pieces. Our implementation only deviated in a minor detail from the description provided in Lohani et al. (2019): We did not implement any features that rely on external knowledge bases or databases. The BPE embeddings were obtained from BPEmb (Heinzerling and Strube, 2018). As the description of the model architecture given by the authors did not include layer size specifications, we ran a parameter search in the same fashion as for the GAT+BiLSTM-CRF model. Table IV.5 contains the results.

IV.4.5 LayoutLM

We obtained an implementation of the LayoutLM model with pretrained weights via the huggingfaces transformers library (Wolf et al., 2020), which corresponds to the LayoutLM_{Base} configuration with text and layout modality, complemented with a token classification head composed of a two-layer MLP with dropout. As proposed by Xu, Li, et al. (2020), we fine-tuned the parameters of the whole (pretrained) model on the token classification task using an ADAM optimizer (Kingma and Ba, 2017). The learning rate was warmed up linearly for one epoch. When extracting the tokens for each document, we scaled the x and y of the textboxes to the range $\{0, 1{,}000\}$. The extracted token sequences were then padded to the maximum length possible (512). If their length already exceeded the maximum length, we applied a sliding window approach.

IV.4.6 Random Forest

As described above, we also trained and evaluated a random forest model (Breiman, 2001). Random forests are known to be robust, with only very limited performance gains to be achieved via hyperparameter tuning (Probst et al., 2019). We therefore abstained from tuning the random forest and used the standard parameters. The model was trained on the OCR outputs, which had been enriched with several manually constructed features. One row in the data set corresponded to one single text box. The manual features captured the properties of the text associated with each text box, such as its length and the number of (special) characters and digits, and data types (see appendix B.1). We furthermore calculated dictionary-based features that searched for the presence of discriminative terms such as "Invoice no." or "Total" in the vicinity of the text box. These features mimicked the context diffusion mechanisms in the neural networks. Handling text boxes independently from each other further allowed us to employ subsampling to address the class imbalance. We subsampled the background class so that the numbers of labeled and unlabeled instances in the training set were equal.

IV.5 Results

Using the implementation described above, we tuned, trained, and evaluated the models as described in section IV.3.3. We report the F_1 scores with standard deviation per model and entity type averaged across all folds and the corresponding macro averages in Table IV.6. The results are further disaggregated down into results obtained from invoices with in-sample and out-of-sample layouts. In the tables, **bold** and underlined results indicate the best and second-best results per entity class. Figure IV.6 summarizes the tables and provides box plots of the attained F_1 scores per model. The figure shows a significant gap in predictive quality between in-sample and out-of-sample layouts, both in terms of macro average (green dashed line) and scattering. This gap—which we will refer to from here on as the accuracy gap—varies between the different models. While the accuracy gap of the LayoutLM model—the overall best model in the study—is 0.1742 points, it is much larger for the other models. The most affected are the random forest and Chargrid models, with gaps of 0.4299 and 0.4206 F_1, respectively. The GCN and BERTgrid modes are less affected: Their accuracy drops by 0.3287 and 0.3187, respectively. LayoutLM achieves the best overall results, both for in- and out-of-sample layouts in terms of macroaveraged F_1 scores: 0.8761, and 0.7019, respectively. Looking at the overall results and the results over

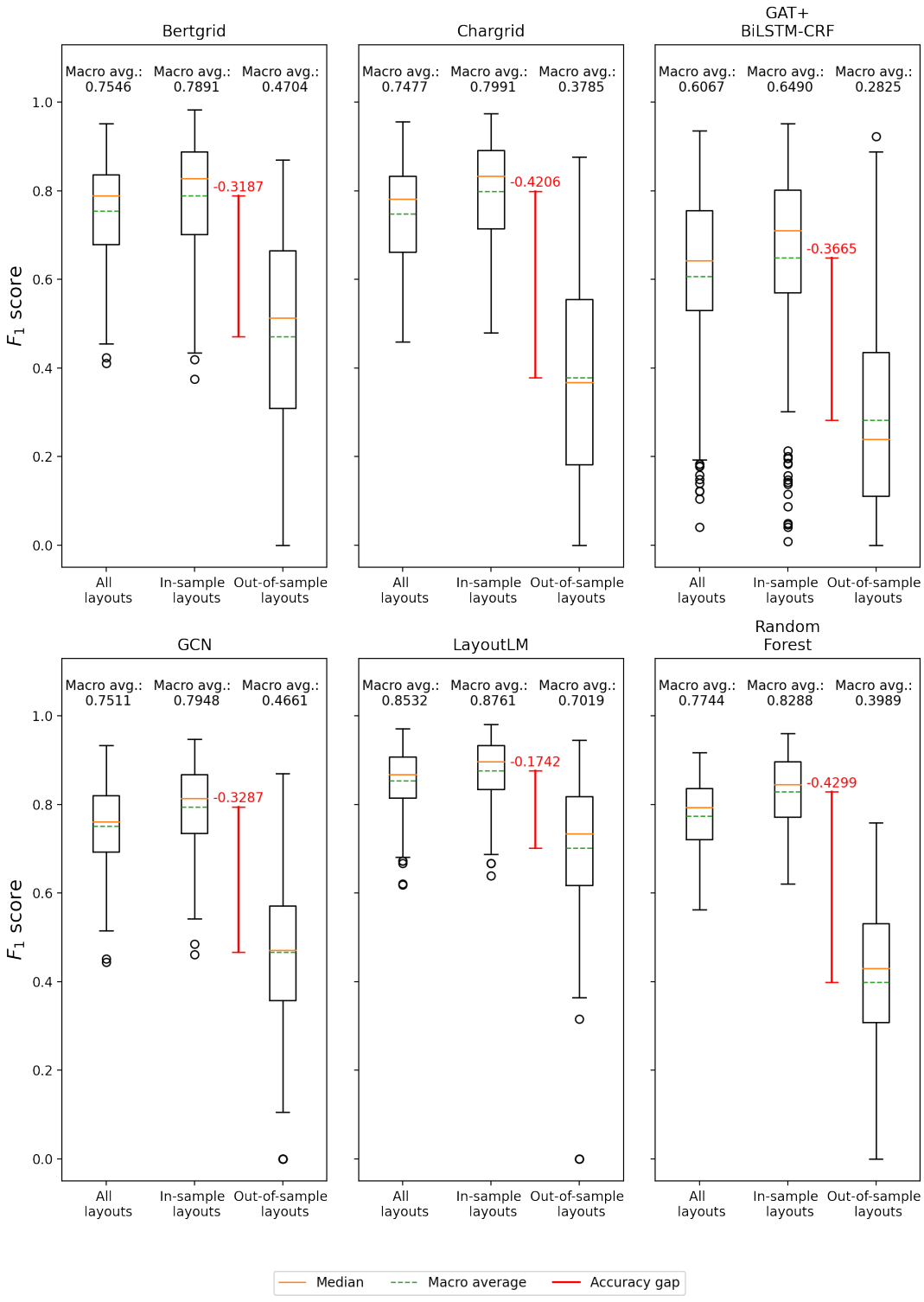

Figure IV.6: Box plots of the F_1 scores across all folds for both in-sample and out-of-sample layouts by model. The red line emphasizes the gap in predictive quality between in-sample and out-of-sample layouts. The macro average is the unweighted average of class-wise F_1 scores across all folds.

in-sample layouts, its closest competitor is the random forest model, with macro in-sample and out-of-sample F_1 scores of 0.7744 and 0.8288, respectively. It outperforms all neural networks other than LayoutLM in terms of macroaveraged F_1, and it even manages to outperform LayoutLM in the detection of line item quantities over in-sample layouts. This picture changes as we look at the out-of-sample results. Random forest's accuracy drops to 0.3989, being subsequently outperformed by both the BERTgrid and the GCN models. While BERTgrid is slightly better than the GCN model over out-of-sample layouts, GCN achieves a lower standard deviation, indicating more consistent results. This observation also holds true for in-sample layouts and overall results. The overall worst-performing model is the GAT+BiLSTM-CRF model in terms of both macro average F_1 and standard deviation. However, its accuracy gap between in- and out-of-sample layouts is smaller than that of random forest and Chargrid. Comparing the overall results for the individual entity classes, all models extract the recipient address with the highest accuracy and the total tax amount with the lowest. However, when reduced to the out-of-sample layouts, the models show worse performance on the line item tax amounts than on the total tax amounts; even LayoutLM achieves an F_1 of merely 0.2826. This is surprising, considering that LayoutLM achieves a very good F_1 score of 0.8811 on the class over in-sample layouts. All of the models fail in at least one fold to extract the line item tax amounts; the GAT+BiLSTM-CRF does not succeed at all. Apart from LayoutLM, which almost always outperforms all other models, the models exhibit individual strengths in extracting certain entities. For example, the GAT+BiLSTM-CRF achieves the second-highest F_1 score for the vendor VAT ID on out-of-sample layouts, whereas Chargrid almost matches the accuracy of LayoutLM for line item tax amounts.

Table IV.6: F_1 scores with standard deviation by model and entity class across **in-sample** and **out-of-sample** layouts in the test sets

	All layouts						**In-sample layouts**						**Out-of-sample layouts**					
Entity type	BERTgrid	Chargrid	GAT +BiLSTM-CRF	GCN	LayoutLM	Random Forest	BERTgrid	Chargrid	GAT +BiLSTM-CRF	GCN	LayoutLM	Random Forest	BERTgrid	Chargrid	GAT +BiLSTM-CRF	GCN	LayoutLM	Random Forest
Invoice Number	0.8122	0.7624	0.5797	0.7761	**0.9115**	0.8505	0.8989	0.9103	0.7600	0.8578	**0.9324**	0.8938	0.3830	0.2104	0.2278	0.4203	**0.7476**	0.4762
	(0.0292)	(0.0380)	(0.1620)	(0.0336)	(0.0252)	(0.0284)	(0.0364)	(0.0520)	(0.2542)	(0.0647)	(0.0315)	(0.0474)	(0.1005)	(0.0754)	(0.1073)	(0.0914)	(0.0985)	(0.1127)
Issue Date	0.8702	0.8157	0.6953	0.8291	**0.9241**	0.8177	0.8827	0.8826	0.6465	0.8469	**0.9465**	0.9150	0.6028	0.3237	0.3972	0.5914	**0.8376**	0.4884
	(0.0414)	(0.0708)	(0.2140)	(0.0463)	(0.0244)	(0.0447)	(0.0327)	(0.0344)	(0.2057)	(0.0371)	(0.0258)	(0.0266)	(0.1257)	(0.1067)	(0.2354)	(0.1212)	(0.0673)	(0.0933)
Total Amount	0.8130	0.7857	0.5539	0.7309	**0.8288**	0.7394	0.9137	0.8820	0.7304	0.8757	**0.9407**	0.8843	0.6262	0.4966	0.2410	0.4861	**0.7302**	0.4693
	(0.0477)	(0.0398)	(0.1755)	(0.0575)	(0.0555)	(0.0582)	(0.0435)	(0.0693)	(0.2611)	(0.0479)	(0.0257)	(0.0460)	(0.0988)	(0.0566)	(0.1428)	(0.1255)	(0.0886)	(0.0958)
Vendor VAT ID	0.5933	0.6155	0.5611	0.7317	**0.8563**	0.7363	0.8154	0.8137	0.6943	0.6823	**0.8609**	0.7343	0.2850	0.0797	0.3639	0.2641	**0.7068**	0.2059
	(0.0535)	(0.0591)	(0.1819)	(0.0479)	(0.0359)	(0.0505)	(0.0391)	(0.0364)	(0.2090)	(0.0644)	(0.0313)	(0.0394)	(0.1448)	(0.0690)	(0.1819)	(0.0833)	(0.1082)	(0.1247)
Due Date	0.8536	0.8444	0.7156	0.7822	**0.9097**	0.8275	0.8323	0.8299	0.6095	0.8210	**0.8636**	0.8418	0.4140	0.1374	0.1945	0.2609	**0.7578**	0.1928
	(0.0313)	(0.0638)	(0.2169)	(0.0697)	(0.0257)	(0.0496)	(0.0547)	(0.0524)	(0.1708)	(0.0415)	(0.0373)	(0.0471)	(0.2306)	(0.1645)	(0.2173)	(0.1918)	(0.1092)	(0.1555)
Provision Date	0.7045	0.5971	0.4468	0.6008	**0.7822**	0.6407	0.7817	0.8610	0.6362	0.8563	**0.9097**	0.8879	0.6077	0.1514	0.2110	0.4522	**0.6916**	0.3873
	(0.0337)	(0.0428)	(0.1284)	(0.0753)	(0.0554)	(0.0601)	(0.0991)	(0.0680)	(0.1785)	(0.0358)	(0.0221)	(0.0334)	(0.1073)	(0.1487)	(0.1463)	(0.1881)	(0.1024)	(0.2043)
Recipient Address	0.9361	0.9344	0.8396	0.9136	**0.9602**	0.8630	0.8718	0.8778	0.6784	0.8277	**0.8956**	0.8822	0.8004	0.7850	0.7761	0.8048	**0.9066**	0.5795
	(0.0135)	(0.0199)	(0.2321)	(0.0126)	(0.0094)	(0.0174)	(0.0306)	(0.0343)	(0.2055)	(0.0327)	(0.0420)	(0.0297)	(0.0662)	(0.0825)	(0.0861)	(0.0457)	(0.0282)	(0.0936)
Vendor Address	0.5619	0.5133	0.7523	0.7801	**0.8897**	0.7838	0.7394	0.7618	0.5765	0.7452	**0.7775**	0.7686	0.6137	0.3752	0.5163	0.5142	**0.7999**	0.4072
	(0.0471)	(0.0339)	(0.2155)	(0.0448)	(0.0326)	(0.0247)	(0.0454)	(0.0417)	(0.2285)	(0.0368)	(0.0574)	(0.0317)	(0.0546)	(0.1018)	(0.1096)	(0.0827)	(0.0589)	(0.0849)
Total Tax Amount	0.5464	0.6206	0.3626	0.6036	**0.6917**	0.6912	0.7279	0.6804	0.4966	0.6338	**0.8005**	0.6904	0.2931	0.3846	0.1019	0.4133	**0.5442**	0.3106
	(0.0997)	(0.0673)	(0.1741)	(0.0837)	(0.0544)	(0.0681)	(0.0411)	(0.0542)	(0.1523)	(0.0772)	(0.0675)	(0.0497)	(0.1994)	(0.2143)	(0.1398)	(0.1196)	(0.0441)	(0.2080)
Line Item Description(s)	0.7797	0.7708	0.6400	0.6413	**0.8317**	0.6851	0.9579	0.9588	0.8487	0.9321	**0.9692**	0.9155	0.6232	0.5784	0.3967	0.4704	**0.6967**	0.5173
	(0.0321)	(0.0322)	(0.1764)	(0.0543)	(0.0270)	(0.0216)	(0.0138)	(0.0150)	(0.2650)	(0.0104)	(0.0089)	(0.0194)	(0.1022)	(0.1000)	(0.1120)	(0.0677)	(0.0868)	(0.0427)
Line Item Price(s)	0.7857	0.7903	0.5727	0.7908	**0.8441**	0.7925	0.8460	0.8427	0.6064	0.7754	**0.8513**	0.7886	0.3275	0.4020	0.0951	0.5452	**0.6840**	0.4276
	(0.0462)	(0.0507)	(0.1596)	(0.0378)	(0.0318)	(0.0509)	(0.0560)	(0.0511)	(0.2198)	(0.0633)	(0.0621)	(0.0632)	(0.1271)	(0.1770)	(0.1234)	(0.1294)	(0.1156)	(0.0880)
Line Item Quantity(s)	0.7491	0.8381	0.5999	0.8228	**0.8938**	0.8456	0.5974	0.6703	0.4089	0.6458	0.7284	**0.7594**	0.2738	0.5154	0.2182	0.5410	**0.7262**	0.4447
	(0.0943)	(0.0692)	(0.1651)	(0.0436)	(0.0197)	(0.0292)	(0.1320)	(0.0427)	(0.1989)	(0.0917)	(0.0609)	(0.0683)	(0.1993)	(0.1948)	(0.1299)	(0.1715)	(0.1006)	(0.1621)
Line Item Subtotal(s)	0.8360	0.8352	0.6184	0.7881	**0.8665**	0.8237	0.5493	0.5376	0.7983	0.8310	**0.9080**	0.8553	0.6237	0.5899	0.2149	0.5942	**0.7154**	0.5257
	(0.0324)	(0.0295)	(0.1884)	(0.0278)	(0.0425)	(0.0314)	(0.0520)	(0.0345)	(0.2423)	(0.0490)	(0.0347)	(0.0246)	(0.1235)	(0.0722)	(0.1383)	(0.1112)	(0.1037)	(0.0985)
Line Item Tax Amount(s)	0.7226	0.7445	0.5564	0.7237	**0.7540**	0.7449	0.6324	0.6785	0.5946	0.7962	**0.8811**	0.7861	0.1111	0.2696	0.0000	0.1675	**0.2826**	0.1515
	(0.0424)	(0.0403)	(0.2284)	(0.0389)	(0.0584)	(0.0362)	(0.0593)	(0.0644)	(0.2442)	(0.0507)	(0.0326)	(0.0483)	(0.2357)	(0.2000)	(0.0000)	(0.1643)	(0.2224)	(0.1743)
Macro Avg.	0.7546	0.7477	0.6067	0.7511	**0.8532**	0.7744	0.7891	0.7991	0.6490	0.7948	**0.8761**	0.8288	0.4704	0.3785	0.2825	0.4661	**0.7019**	0.3989
	(0.1241)	(0.1231)	(0.2141)	(0.0982)	(0.0791)	(0.0786)	(0.1334)	(0.1228)	(0.2363)	(0.1000)	(0.0776)	(0.0816)	(0.2360)	(0.2336)	(0.2346)	(0.1979)	(0.1727)	(0.1800)

IV.6 Discussion

The goal of our research was to determine how ML-based approaches to IE from invoices respond to a skewed distribution of suppliers that is characterized by few highly frequent suppliers and a long tail of infrequent suppliers. Specifically, we were interested in the accuracy of the models on the invoices of long tail suppliers.

Our research approach pays special attention to the distribution of layouts in the training, validation, and test data sets, which has been largely neglected in previous research on IE from invoices. The data in our study are split into mutually exclusive subsets such that each subset contains a proportion of out-of-sample layouts issued by long tail suppliers. We found that all models in our study were significantly less accurate on out-of-sample layouts than on in-sample layouts. This is reflected in both lower macro F_1 scores and higher standard deviations between folds.

These findings suggest that there exists a layout bias of which designers of IE systems for invoices need to be aware. If we had not disaggregated the classification results into in-sample and out-of-sample layouts, the accuracy gap between them could have gone undetected. Considering our aim of automating the IE from invoices from long tail suppliers, letting this bias go undetected could have had negative effects in practice as the models would yield considerable amounts of false positives and false negatives. We expected to find the existence of this bias, but we were surprised to find that the models were affected to varying degrees by this it.

Therefore, with respect to previous research, we see the main contribution of our research as discovering the layout bias and describing and implementing an evaluation methodology to detect it. The distribution of layouts has seldom been reported or methodologically addressed in the related literature. This unfortunately also holds true for the open benchmark data sets SROIE (Huang et al., 2019), CORD (Park et al., 2019) and Kleister (Stanisławek et al., 2021), due to which we suspect that the results obtained from these benchmark data sets might also be affected, albeit to a possibly different degree. A further contribution of this paper is the development and implementation of a benchmark on a dedicated invoice data set, using a common set of accuracy metrics. Previous studies of IE from layout-rich documents either relied on proprietary data sets, used different metrics and aggregation levels for evaluation, or did not report results for individual entities. Benchmarking studies dedicated to IE from invoices (Baviskar, Ahirrao, and Kotecha, 2021; Liu, Zhang, et al., 2016) did not include layout-aware approaches to IE.

The results further hold practical implications. They show that designers of systems that seek to automate IE from invoices should carefully evaluate the distribution of layouts in the population to which the system is applied. In the case presented in this paper, the layout bias can represent a form of selection bias (Shah et al., 2020): The invoices used for model training were taken from a ground population that contained a skewed distribution of invoice layouts. If the models are only intended to be applied to long tail suppliers, this approach will most certainly yield suboptimal outcomes, as shown in our results. If, however, the models are to be applied to the whole range of incoming invoices, failing to appropriately capture highly recurrent invoice layouts would lead to inefficiencies. In this case, a significant change in the supplier structure of the company would require a retrain-

ing of the models, and therefore also the annotation of a new set of invoices. Furthermore, the skewed distribution of layouts could be addressed using multicriterion sample weights, which take into account the distribution of layouts in addition to the distribution of classes.

Of all models in our experiment, LayoutLM shows both the best accuracy over both in-sample and out-of-sample layouts. It continues the victory march of transformer models in NLP, which benefit from extensive pretraining over large unlabeled document corpora. In this context, one surprising finding in our study was the accuracy of the random forest model. The model's overall accuracy was on a par with that of the nontransformer neural networks, and it was able to outperform the Chargrid and GAT+BiLSTM-CRF models over out-of-sample layouts. Random forests could be employed very effectively to handle cases in which all possible layouts can be represented in the training data. Another unexpected finding was how the predictive quality of the BERTgrid and Chargrid models compared to the GCN model. While Chargrid and BERTgrid are very close in overall accuracy, BERTgrid is almost 0.1 points better on the out-of-sample layouts. This hints at the fact that the semantic BERT embeddings help the model draw generalizations. The drawback of both models is the memory consumption of the respective document grids: BERTgrid especially yields large tensors with the dimensions $\{width_{target}, height_{target}, \text{features}\}$. The GCN model relies on more efficient graph representations of the shape $\{nodes, features\}$; however, it achieves results very close to those of BERTgrid. As BERTgrid and Chargrid employ object detection to detect line items, we especially expected them to perform better over line item–related entities, which was not the case. Finally, the poor performance of the GAT+BiLSTM-CRF model, both overall and on unseen layouts, might be attributed to the quality of the OCR outputs; the paragraph recognition on layout-rich documents of Tesseract is prone to errors, and the resulting paragraphs fluctuate heavily in length.

With respect to the individual entities, we observed that the models had problems extracting the total tax amount. The total tax amount appeared infrequently in our data set; only 383 text boxes have been annotated as such. The recipient address was the easiest information entity in our data set: the recipient is always the same company, and only small variations in the address appear, such as street names, house numbers, and cities for the different offices. Considering its importance to both accountants and auditors, we were disappointed to see the low performance of all models on the total amount as compared to the invoice number and issue date.

These results and the subsequent discussion have to be seen in the light of some limitations. First, the data set was rather small. The models might respond better to larger data sets and show fewer signs of bias, even if the distribution of layouts were similarly skewed. Apart from its size, the data set was also composed of mainly English invoices. Therefore, the effect of multiple languages on the predictive quality of the models was not adequately considered. Furthermore, the evaluation methodology was quite strict and did not account for further rule-based post-processing to reduce false positives and false negatives, such as checking the internal consistency of the extracted amounts. Also, most models in the experiment were our own implementations, and they might not have been true in all aspects to the original implementations. This was a necessity, as the original implementations were not openly accessible. However, we have specified all relevant details in section IV.4.

IV.7 Conclusion

IE from invoices offers a way to automate invoice processing that causes less friction with suppliers that are not suitable for the implementation of an EDI-based invoicing tool. The results from our research show that designers of IE systems for invoices should carefully consider how to collect invoices for training and testing data, as the distribution of invoices per supplier in a company is usually biased toward a few suppliers. According to our results, this skewed distribution of suppliers, and therefore layouts, causes the models to perform worse on invoices with out-of-sample layouts, such as invoices whose layouts have not been part of the training set. However, this effect varied across the models used in our study. LayoutLM is the least affected by layout bias, while it simultaneously exhibits the best macroaveraged F_1 scores on both in-sample and out-of-sample layouts; this result can most likely be attributed to the extensive pretraining of the model. This kind of bias can go undetected if not appropriately accounted for. As it has received little attention in previous research, we also suspect that the results obtained from popular open data sets for IE from layout-rich documents might be affected. We therefore strongly encourage researchers to investigate the distribution of layouts in the respective training, validation, and test sets for these data sets.

IV.8 References

Akbik, Alan et al. (2019). "FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 54–59. DOI: 10.18653/v1/N19-4010. URL: https://aclanthology.org/N19-4010.

Bao, Hangbo et al. (2020). "UniLMv2: Pseudo-Masked Language Models for Unified Language Model Pre-Training". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 642–652. URL: http://proceedings.mlr.press/v119/bao20a.html.

Baviskar, Dipali, Swati Ahirrao, and Ketan Kotecha (2021). "Multi-Layout Unstructured Invoice Documents Dataset: A Dataset for Template-Free Invoice Processing and Its Evaluation Using AI Approaches". In: *IEEE Access* 9, pp. 101494–101512. DOI: 10.1109/ACCESS.2021.3096739.

Baviskar, Dipali, Swati Ahirrao, Vidyasagar Potdar, et al. (2021). "Efficient Automated Processing of the Unstructured Documents Using Artificial Intelligence: A Systematic Literature Review and Future Directions". In: *IEEE Access* 9, pp. 72894–72936. DOI: 10.1109/ACCESS.2021.3072900.

Bengio, Yoshua (2009). "Learning Deep Architectures for AI". In: *Foundations and Trends in Machine Learning* 2.1. Place: Hanover, MA, USA Publisher: Now Publishers Inc., pp. 1–127. ISSN: 1935-8237. DOI: 10.1561/2200000006. URL: https://doi.org/10.1561/2200000006.

— (2012). "Practical Recommendations for Gradient-Based Training of Deep Architectures". In: *Neural Networks: Tricks of the Trade*. Ed. by Grégoire Montavon et al. 2nd. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 437–478. ISBN: 978-3-642-35289-8. URL: https://doi.org/10.1007/978-3-642-35289-8_26.

Breiman, Leo (2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324. URL: https://doi.org/10.1023/A:1010933404324.

Cristani, Matteo et al. (2018). "Future paradigms of automated processing of business documents". In: *International Journal of Information Management* 40, pp. 67–75. ISSN: 0268-4012. DOI: https://doi.org/10.1016/j.ijinfomgt.2018.01.010. URL: https://www.sciencedirect.com/science/article/pii/S0268401217309994.

Denk, Timo I. and Christian Reisswig (2019). "BERTgrid: Contextualized Embedding for 2D Document Representation and Understanding". In: *arXiv:1909.04948 [cs]*. arXiv: 1909.04948. URL: http://arxiv.org/abs/1909.04948 (visited on 02/16/2021).

Devlin, Jacob et al. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://www.aclweb.org/anthology/N19-1423.

Fey, Matthias and Jan Eric Lenssen (2019). "Fast graph representation learning with PyTorch Geometric". In: *arXiv:1903.02428 [cs.LG]*. URL: https://arxiv.org/abs/1903.02428 (visited on 12/09/2022).

Forney, G David (1973). "The viterbi algorithm". In: *Proceedings of the IEEE* 61.3. Publisher: Ieee, pp. 268–278. ISSN: 1558-2256. DOI: 10.1109/PROC.1973.9030.

Garncarek, Lukasz et al. (2021). "LAMBERT: Layout-Aware Language Modeling for Information Extraction". In: *Document Analysis and Recognition – ICDAR 2021*. Ed. by Josep Lladós et al. Cham: Springer International Publishing, pp. 532–547. ISBN: 978-3-030-86549-8.

Han, Jiawei et al. (2011). *Data Mining: Concepts and Techniques*. 3rd. The Morgan Kaufmann Series in Data Management Systems. Waltham, USA: Elsevier, Morgan Kauffman. ISBN: 978-0-12-381479-1.

Heinzerling, Benjamin and Michael Strube (2018). "BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Ed. by Nicoletta Calzolari (Conference chair) et al. Miyazaki, Japan: European Language Resources Association (ELRA). ISBN: 979-10-95546-00-9.

Huang, Zheng et al. (2019). "ICDAR2019 Competition on Scanned Receipt OCR and Information Extraction". In: *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1516–1520. DOI: 10.1109/ICDAR.2019.00244.

ISA 330, IAASB (2009). *International Standard On Auditing 330 The Auditor's Responses To Assessed Risks*. URL: https://www.ifac.org/system/files/downloads/a019-2010-iaasb-handbook-isa-330.pdf.

Karnin, Zohar et al. (2013). "Almost Optimal Exploration in Multi-Armed Bandits". In: *Proceedings of the 30th International Conference on Machine Learning*. Ed. by Sanjoy Dasgupta and David McAllester. Vol. 28. Proceedings of Machine Learning Research. Issue: 3. Atlanta, Georgia, USA: PMLR, pp. 1238–1246. URL: https://proceedings.mlr.press/v28/karnin13.html.

Katti, Anoop Raveendra et al. (2018). "Chargrid: Towards Understanding 2D Documents". In: *arXiv:1809.08799 [cs]*. arXiv: 1809.08799. URL: http://arxiv.org/abs/1809.08799 (visited on 02/16/2021).

Kingma, Diederik P. and Jimmy Ba (2017). "Adam: A Method for Stochastic Optimization". In: *arXiv:1412.6980 [cs.LG]*. DOI: 10.48550/arXiv.1412.6980. URL: https://arxiv.org/abs/1412.6980 (visited on 12/09/2022).

Klein, Bertin et al. (2004). "Results of a Study on Invoice-Reading Systems in Germany". In: *Document Analysis Systems VI*. Ed. by Simone Marinai and Andreas R. Dengel. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 451–462. ISBN: 978-3-540-28640-0.

Koch, Bruno (2017). *Business Case E-Invoicing / E-Billing*. URL: https://www.billentis.com/e-invoicing-businesscase.pdf (visited on 03/25/2022).

— (2019). *The E-Invoicing Journey 2019-2025*. URL: https://www.billentis.com/The_einvoicing_journey_2019-2025.pdf (visited on 03/25/2022).

Krieger, Felix et al. (2021). "Information Extraction from Invoices: A Graph Neural Network Approach for Datasets with High Layout Variety". In: *Wirtschaftsinformatik 2021 Proceedings*. International Conference on Wirtschaftsinformatik.

Liu, Wenshun, Y Zhang, et al. (2016). "Unstructured document recognition on business invoice". In: *CS229: Machine Learning, Stanford iTunes University, Stanford, CA, USA, Tech. Rep.* URL: https://cs229.stanford.edu/proj2016/report/LiuWanZhang-UnstructuredDocumentRecognitionOnBusinessInvoice-report.pdf (visited on 12/09/2022).

Liu, Xiaojing, Feiyu Gao, et al. (2019). "Graph Convolution for Multimodal Information Extraction from Visually Rich Documents". In: *arXiv:1903.11279 [cs]*. arXiv: 1903.11279. URL: http://arxiv.org/abs/1903.11279 (visited on 04/19/2021).

Liu, Yinhan, Myle Ott, et al. (2019). "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *arXiv:1907.11692 [cs.CL]*. DOI: 10.48550/arXiv.1907.11692. URL: https://arxiv.org/abs/1907.11692 (visited on 12/09/2022).

Lohani, D. et al. (2019). "An Invoice Reading System Using a Graph Convolutional Network". In: *Computer Vision – ACCV 2018 Workshops*. Lecture Notes in Computer Science 11367. Ed. by Gustavo Carneiro and Shaodi You, pp. 144–158. DOI: 10.1007/978-3-030-21074-8_12. URL: http://link.springer.com/10.1007/978-3-030-21074-8_12 (visited on 04/19/2021).

Ma, Xuezhe and Eduard Hovy (2016). "End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 1064–1074. DOI: 10.18653/v1/P16-1101. URL: https://www.aclweb.org/anthology/P16-1101.

Majumder, Bodhisattwa Prasad et al. (2020). "Representation Learning for Information Extraction from Form-like Documents". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, pp. 6495–6504. DOI: 10.18653/v1/2020.acl-main.580. URL: https://www.aclweb.org/anthology/2020.acl-main.580 (visited on 04/19/2021).

Mikolov, Tomas et al. (2013). "Efficient Estimation of Word Representations in Vector Space". In: *arXiv:1301.3781 [cs.CL]*. DOI: 10.48550/arXiv.1301.3781. URL: https://arxiv.org/abs/1301.3781 (visited on 12/09/2022).

Palm, Rasmus Berg et al. (2017). "CloudScan - A configuration-free invoice analysis system using recurrent neural networks". In: *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. Vol. 01, pp. 406–413. DOI: 10.1109/ICDAR.2017.74. arXiv: 1708.07403. URL: http://arxiv.org/abs/1708.07403 (visited on 05/08/2021).

Park, Seunghyun (2021). *CORD: A Consolidated Receipt Dataset for Post-OCR Parsing*. CORD: A Consolidated Receipt Dataset for Post-OCR Parsing. URL: https://github.com/clovaai/cord (visited on 07/12/2021).

Park, Seunghyun et al. (2019). "CORD: A Consolidated Receipt Dataset for Post-OCR Parsing". In: *Workshop on Document Intelligence at NeurIPS 2019*. URL: https://openreview.net/forum?id=SJl3z659UH (visited on 12/09/2022).

Paszke, Adam et al. (2016). "Enet: A deep neural network architecture for real-time semantic segmentation". In: *arXiv:1606.02147 [cs.CV]*. DOI: 10.48550/arXiv.1606.02147. URL: https://arxiv.org/abs/1606.02147 (visited on 12/09/2022).

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Probst, Philipp et al. (2019). "Hyperparameters and tuning strategies for random forest". In: *WIREs Data Mining and Knowledge Discovery* 9.3. ISSN: 1942-4787, 1942-4795. DOI: 10.1002/widm.1301. URL: https://onlinelibrary.wiley.com/doi/10.1002/widm.1301 (visited on 11/06/2021).

Řehůřek, Radim and Petr Sojka (2010). "Software Framework for Topic Modelling with Large Corpora". In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, pp. 45–50.

Ren, Shaoqing et al. (2015). "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes et al. Vol. 28. Curran Associates, Inc., pp. 91–99. URL: https://proceedings.neurips.cc/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf (visited on 12/09/2022).

Sarawagi, Sunita (2008). "Information Extraction". In: *Foundations and Trends in Databases* 1.3. Place: Hanover, MA, USA Publisher: Now Publishers Inc., pp. 261–377. ISSN: 1931-7883. DOI: 10.1561/1900000003. URL: https://doi.org/10.1561/1900000003.

Sennrich, Rico et al. (2016). "Neural Machine Translation of Rare Words with Subword Units". In: *arXiv:1508.07909 [cs.CL]*. DOI: 10.48550/arXiv.1508.07909. URL: https://arxiv.org/abs/1508.07909 (visited on 12/09/2022).

Shah, Deven Santosh et al. (2020). "Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Publisher: Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.468. URL: http://dx.doi.org/10.18653/v1/2020.acl-main.468.

Smith, Leslie N (2017). "Cyclical Learning Rates for Training Neural Networks". In: *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE Winter Conference on Applications of Computer Vision (WACV). Santa Rosa, California, United States of America: IEEE, pp. 464–472. DOI: 10.1109/WACV.2017.58.

— (2018). "A disciplined approach to neural network hyper-parameters: Part 1 - Learning rate, batch size, momentum, and weight decay". In: *arXiv:1803.09820 [cs.LG]*. DOI: 10.48550/arXiv.1803.09820. URL: https://arxiv.org/abs/1803.09820 (visited on 12/09/2022).

Snoek, Jasper et al. (2012). "Practical Bayesian Optimization of Machine Learning Algorithms". In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira et al. Vol. 25. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2012/file/05311655a15b75fab86956663e1819cd-Paper.pdf.

Stanisławek, Tomasz et al. (2021). "Kleister: Key Information Extraction Datasets Involving Long Documents with Complex Layouts". In: *Document Analysis and Recognition – ICDAR 2021*. Ed. by Josep Lladós et al. Cham: Springer International Publishing, pp. 564–579. ISBN: 978-3-030-86549-8.

Tanner, Christian and Sarah-Louise Richter (2018). "Digitalizing B2B Business Processes—The Learnings from E-Invoicing". In: *Business Information Systems and Technology 4.0: New Trends in the Age of Digital Change*. Ed. by Rolf Dornberger. Cham: Springer International Publishing, pp. 103–116. ISBN: 978-3-319-74322-6. DOI: 10.1007/978-3-319-74322-6_7. URL: https://doi.org/10.1007/978-3-319-74322-6_7.

Vaswani, Ashish et al. (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wolf, Thomas et al. (2020). "Transformers: State-of-the-art natural language processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45.

Wu, Yonghui et al. (2016). "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation". In: *arXiv:1609.08144 [cs.CL]*. DOI: 10.48550/arXiv.1609.08144. URL: https://arxiv.org/abs/1609.08144 (visited on 12/09/2022).

Xu, Yang, Yiheng Xu, et al. (2022). "LayoutLMv2: Multi-Model Pre-Training For Visually-Rich Document Understanding". In: *arXiv:2012.14740 [cs.CL]*, p. 17. DOI: 10.48550/arXiv.2012.14740. URL: https://arxiv.org/abs/2012.14740 (visited on 12/09/2022).

Xu, Yiheng, Minghao Li, et al. (2020). "LayoutLM: Pre-training of Text and Layout for Document Image Understanding". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1192–1200. DOI: 10.1145/3394486.3403172. arXiv: 1912.13318. URL: http://arxiv.org/abs/1912.13318 (visited on 05/09/2021).

Yu, Wenwen et al. (2020). "PICK: Processing Key Information Extraction from Documents using Improved Graph Learning-Convolutional Networks". In: *arXiv:2004.07464 [cs]*. DOI: 10.48550/arXiv.2004.07464. arXiv: 2004.07464. URL: http://arxiv.org/abs/2004.07464 (visited on 02/16/2021).

Zhang, Aston, Zachary C. Lipton, et al. (2021). "Dive into Deep Learning". In: *arXiv:2106.11342 [cs.LG]*. DOI: 10.48550/arXiv.2106.11342. URL: https://arxiv.org/abs/2106.11342 (visited on 12/09/2022).

Zhang, Peng, Yunlu Xu, et al. (2020). "TRIE: End-to-End Text Reading and Information Extraction for Document Understanding". In: *Proceedings of the 28th ACM International Conference on Multimedia*. MM '20: The 28th ACM International Conference on Multimedia. Seattle WA USA: ACM, pp. 1413–1422. ISBN: 978-1-4503-7988-5. DOI: 10.1145/3394171.3413900. URL: https://dl.acm.org/doi/10.1145/3394171.3413900 (visited on 02/16/2021).

Zhao, Xiaohui et al. (2019). "CUTIE: Learning to Understand Documents with Convolutional Universal Text Information Extractor". In: *arXiv:1903.12363 [cs]*. DOI: 10.48550/arXiv.1903.12363. URL: http://arxiv.org/abs/1903.12363 (visited on 02/16/2021).

Chapter V

Benchmarking Machine Learning Models in Auditing: Toward an Information Extraction Pipeline for the Test of Details

Outline

V.1 Introduction 156
V.2 Related Literature 157
V.3 Methodology 159
V.4 Results 164
V.5 Discussion 169
V.6 Conclusion 170
V.7 References 171

Bibliographic Information

Krieger, F., Drews, P., Funk, B. (2023). "Benchmarking Machine Learning Models in Auditing: Toward an Information Extraction Pipeline for the Test of Details". Manuscript submitted for publication.

Author's contribution

The author's share of the publication is 80%. Table C.5 in appendix C shows the contributions of all authors of the publication in detail.

Copyright Notice

Abstract

Artificial intelligence (AI) has frequently been predicted to change the auditing profession. Yet, the observed adoption of AI in this domain remains rather limited. We aim to actively advance the adoption of AI by designing and implementing an information extraction pipeline for an application supporting the test of details, a frequently performed audit procedure. To this end, we employ the action design research methodology, through which we reflect how the audit domain shapes the emerging artifact. The multitude of audit clients and the resulting variety of languages and legal regulation requires continuous benchmarking of the machine learning models employed. This is amplified by the continuous stream of new and updated models arising from the research on information extraction from layout-rich documents. We developed a benchmarking pipeline to support the test of details and captured the results of reflecting the design process in design principles.

V.1 Introduction

Artificial intelligence (AI) is expected to have a vast, even possibly disruptive impact on the audit profession (Issa et al., 2016; Kokina and Davenport, 2017); however, its actual adoption remains limited (Krieger et al., 2021). Krieger and colleagues put forward the suggestion that audit firms adopt advanced data analytics technologies by embedding them into software applications, which are tailored toward one or more specific use cases (ibid.). The applicability of AI to auditing ranges across various steps in the audit process (Appelbaum et al., 2018) and different types of audit procedures (Sun, 2019). One example for such a use case is the "test of details" (TOD) ISA 500. An essential procedure in audit fieldwork, the TOD is used to collect evidence that individual transactions or balances have been reported correctly. To this end, audit teams are often required to sift through large samples of business documents to extract pieces of information that are then reconciliated against other, usually structured sources of data stemming from enterprise resource planning (ERP) systems. This information extraction (IE) task costs audit teams valuable time given the sharp budget constraints under which they operate. Hence, its automation through the means of AI—machine learning (ML) and natural language processing (NLP)—can alleviate their workload and allows them to focus on higher value-added tasks. It would allow also for the processing of larger quantities of documents, which is commonly associated with a higher audit quality (Manita et al., 2020; Salijeni et al., 2019). Business documents relevant for the TOD, such as invoices, purchase orders, or delivery notes, are characterized by sparsity and a two-dimensional layout, as opposed to the sequential text usually assumed in NLP techniques. Subsequently, they have drawn the attention of a fast-growing body of research, in which many highly specialized layout-preserving document representations and neural network architectures are proposed. These representations rely on graphs, grids, or positional embeddings of varying granularity to capture the layout; the models then leverage semantic and visual features of the document to extract the desired information. Introducing these techniques to automate the TOD presents a particularly interesting technology adoption case in the audit domain: the TOD is a highly repetitive task in auditing, which is performed in every audit engagement. While the task of extracting information from an invoice and reconciling it against another data source is of low complexity for human cognitive capabilities, it requires highly specialized ML approaches. These approaches, in turn, permit the processing of large quantities of invoices in a fraction of the time a human operator would require.

Our work is motivated by two goals: to advance the adoption of AI in auditing, and to contribute to the ongoing research on the adoption of AI to auditing. Embedded within an audit firm, we design and implement the IE component for an application seeking to automate the TOD for invoices. Utilizing a ML model in production requires a workflow with multiple steps around data preparation as well as model training, evaluation, and deployment. In the context of ML, this workflow is also referred to as a "pipeline" (Google, 2022; Hapke and Nelson, 2020). Our research aims to design such a pipeline in a way that it meets the requirements of the audit domain, posing the following research question (RQ):

> *How can an information extraction pipeline for the TOD be designed and which design principles could guide the development of ML benchmarking pipelines?*

We address the RQ by utilizing the methodology of action design research (ADR) (Sein et al., 2011). Following the ADR process, the artifact under construction – the IE component – has gone through several building, intervention, and evaluation (BIE) cycles. The contribution of this paper lies in the created artifact and the design principles (DPs) developed for ML benchmarking pipelines in auditing.

V.2 Related Literature

External audits encompass several tasks that are both highly repetitive and highly structured (Abdolmohammadi, 1999) (Abdolmohammadi, 1999), making them ideal for automation through either robotic process automation (RPA) (Moffitt et al., 2018) or AI (Issa et al., 2016; Kokina and Davenport, 2017; Krieger et al., 2021; Sun, 2019). While RPA requires very well-defined tasks, ML and its subfield deep learning are able to handle more ambiguous and complex tasks than RPA, such as the analysis of text (Sun, 2019). Zhaokai and Moffitt (2019) developed a comprehensive framework that illustrates how text analytics can be employed in auditing, oriented toward the analysis of contracts. The framework is composed of multiple so-called "functional areas", which can be interpreted as process steps: the documents are first imported and categorized ("document management"). Afterwards, the relevant information is extracted ("content identification") which can then be used for downstream tasks belonging to the TOD, such as cutoff testing, term verification, or record confirmation (ibid.). In the content identification area, the authors propose ML as a means to extract the information. Contracts are usually rich, sequential text, through which the IE task can be addressed by employing sequential models such as transformers, (Bi-)LSTMs and conditional random fields (Chalkidis and Androutsopoulos, 2017; Elwany et al., 2019; Hu and Su, 2021). As pointed out in the previous section, this structure does not characterize all business documents; many exhibit a two-dimensional layout which renders purely sequential approaches inefficient. Recent research has addressed this shortcoming by proposing several model architectures and document representations that account for the layout. In this section, we give an overview of the techniques proposed for IE from layout-rich documents. We also deepen the concept of ML pipelines and their implementation and briefly review relevant research on AI adoption in auditing, before delineating the gaps in the literature to be addressed in this paper.

V.2.1 Information Extraction from Layout-Rich Documents

The problem of extracting information from unstructured text is usually formulated as a classification problem, such that a ML model distinguishes between irrelevant text and one or more classes of relevant text. The model either learns to classify single text units of varying granularity, e.g., words (Lohani et al., 2019), tokens (Denk and Reisswig, 2019; Garncarek et al., 2021; Huang et al., 2022; Xu, Xu, et al., 2022; Xu, Li, et al., 2020; Xu, Lv, et al., 2021), or characters (Katti et al., 2018), or to assign a sequence of class labels to a sequence of text units (Liu et al., 2019).

ML models can leverage several types of inputs to classify the text on a document: for layout-rich documents, the position of the individual text units on the document is an important signal, along with their semantics and the structure of the string(s). Another important signal is the context of the text unit, e.g., are the words "invoice no." next to a

string, a strong indicator for belonging to the class "invoice number." Recent research has proposed different document representations and model architectures to appropriately capture those contextual relationships on layout-rich documents. BERTgrid (Denk and Reisswig, 2019) and Chargrid (Katti et al., 2018) use pixel-level token- respectively character grids in conjunction with a model architecture based on convolutional layers, whereas (Zhao et al., 2019) employ less granular word-level grids. A different stream of work proposes graphs to structurally represent the documents and employs models with graph convolution layers to capture contextual relationships (Liu et al., 2019; Lohani et al., 2019; Yu et al., 2020). The document graphs in these approaches may be explicitly precomputed (Liu et al., 2019; Lohani et al., 2019) or learned from the data (Yu et al., 2020). They are thus different from the implicit document graphs that are used in transformer-based models such as LayoutLM(v2, v3) (Huang et al., 2022; Xu, Xu, et al., 2022; Xu, Li, et al., 2020), LayoutXLM (Xu, Lv, et al., 2021), LAMBERT (Garncarek et al., 2021), Donut (Kim et al., 2021) and LiLT (Wang et al., 2022). The transformer model architecture relies on the self-attention mechanism to capture contextual relationships; within this approach, all inputs are related to one another (Vaswani et al., 2017), thus forming a dense graph. The major contribution of the works around transformer models is the introduction of layout-aware pretraining strategies. In conjunction with the parallelization capabilities of the transformer architecture, this enables them to train the models on massive corpora of unlabeled documents, aiming for a general understanding of layout-rich documents (Garncarek et al., 2021; Huang et al., 2022; Kim et al., 2021; Wang et al., 2022; Xu, Xu, et al., 2022; Xu, Li, et al., 2020; Xu, Lv, et al., 2021). The pretrained models can then be finetuned for specific downstream tasks, such as IE.

V.2.2 Machine Learning Pipelines

It is well-established that the application of ML methods requires a workflow composed of different steps to gather and prepare data, train and evaluate models, and deploy a model into production. This workflow has been codified into standard frameworks for data science such as "Knowledge Discovery in Databases" (Fayyad et al., 1996) or the "Cross-Industry Standard Process for Data Mining" (Wirth and Hipp, 2000). More recent publications refer to this workflow in the context of ML as a pipeline (Google, 2022; Hapke and Nelson, 2020; Xin et al., 2021). More formally described, a pipeline is a directed, acyclic graph, in which the nodes represent processing steps and the edges represent in- and output relationships between processing steps (Hapke and Nelson, 2020; Xin et al., 2021). The steps need to be executed in the correct order, such that all inputs for a step are available before it is executed. This is referred to as the orchestration of a pipeline. ML pipelines can be categorized into different levels of maturity (Google, 2022). The most basic level of maturity are manual pipelines, in which all steps within the pipeline are orchestrated manually. The next, more advanced levels are integrated pipelines, which are automatically orchestrated and can be executed via a trigger (ibid.). Such automated pipelines allow for more rapid experimentation and for an architectural symmetry between experimentation and operation pipelines (ibid.). This enables researchers to run a multitude of experiments across different settings, or for practitioners to quickly retrain models using fresh data and deploy them into operation.

V.2.3 Adoption of Artificial Intelligence in Audit Firms

Leveraging AI to automate tasks such as TOD is a new phenomenon in the audit domain, and empirical evidence suggests a slow adoption of AI in audit firms (Krieger et al., 2021). Krieger et al. (ibid.) explored the adoption of emerging technologies in this area, most importantly RPA, (Big) Data Analytics, and AI, and propose a process theory that illustrates how audit firms adopt these technologies. The authors argue that audit firms develop technology applications – or source them from third party vendors - based on use cases. This process is affected by several contextual factors that affect the outcome of the process, relating to such characteristics as the underlying technology, the characteristics of the firm's audit clients, and the specific demands of the audit domain, e.g., the mandatory consideration of professional standards. While the proposed theory seeks to explain the logic of technology adoption in audit firms, scientific accounts of such applications are rare. One example is the "Document Intelligence for Contract Review" (DICR) platform, which has been proposed by the AI Lab of the Big Four audit company EY (Tecuci et al., 2020). While DICR is presented as an application for document review, the authors do not elaborate on the rationale behind its design. Furthermore, it is designed for the analysis of contracts, which—as pointed out above—are structurally different from layout-rich business documents such as invoices.

In this section, we have identified several gaps in the related literature, which we aim to address in our study. We identified the need for the evaluation of models according to specific utilization settings, and the potential for automated ML pipelines to address this need. More specific to the audit domain, we have also identified the lack of scientific accounts of AI applications that reflect their design with respect to the demands of auditing.

V.3 Methodology

The goal of our research is to create an IE pipeline for a TOD application to be employed on external audits by an auditing firm. To address this goal appropriately, we employ the ADR methodology (Sein et al., 2011). Similar to the design research methodology proposed by Hevner et al. (2004), ADR adopts a construction-oriented approach to research, intended to generate design knowledge that is not merely explanatory, but which is relevant to practice. Design Research has been successfully applied to design platforms or applications that employ ML, e.g., for hate speech detection (Bunde, 2021), health communication (Neuhauser et al., 2013), and to enhance bus ticket vending machines in low-bandwidth areas (Butgereit et al., 2018). ADR explicitly incorporates the organizational context into the emergent artifact by placing the researcher within a participating organization (Sein et al., 2011).

Figure V.1 depicts our research approach. The approach follows several stages: after an initial formulation of the problem (section 3.1), the artifact is devised through three BIE cycles (section 3.2). The BIE cycles are reflected in the reflection and learning phase (section 3.2), which enables us to formulate DPs for IE pipelines in audit applications. The artifact and the DPs are presented in the results section (section 4).

Environment. For this research, the participating organization is an international audit

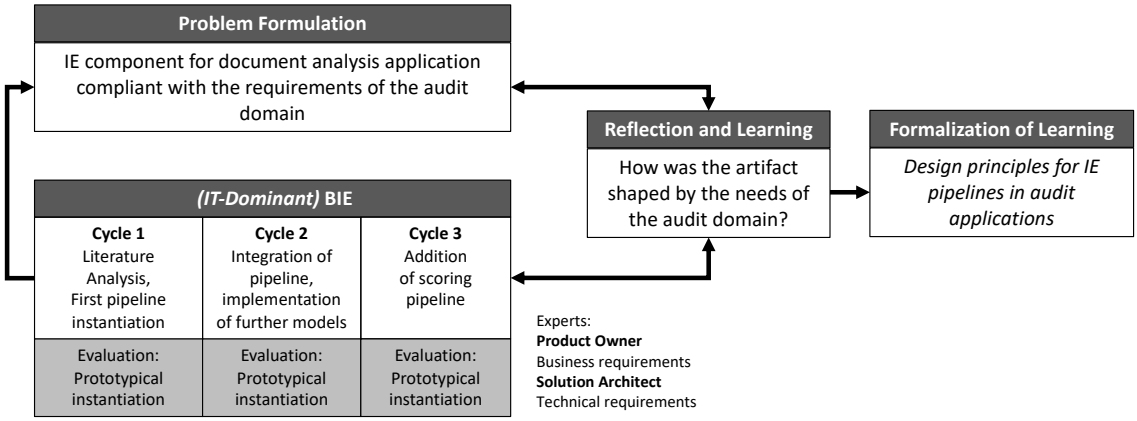

Figure V.1: Research approach based on the ADR methodology.

firm. One member of the research team was placed as a data scientist within a team of software developers and data scientists in the firm developing the TOD application. The ongoing cooperation for this project started at the end of 2019, and the application was still under active development at the time this text was written. The team was led by the application's product owner, who was assisted by the solution architect. Both provided feedback multiple times during the BIE cycles. The product owner could provide requirements from the business side and assess their fulfillment; the solution architect would do the same for technical requirements. For development purposes, we were given an experimental data set, composed of 1059 invoices, which the firm received from 259 different suppliers. The invoices were annotated manually by two domain experts for 14 information entities (e.g., invoice number, total amount).

Scope. The core function of the TOD application is to reconcile two different accounting data sources: a tabular excerpt of the audit client's ERP system, and a set of documents. Figure V.2 visualizes the application's processing flow. The documents are usually provided as PDF files, whereas the ERP excerpt is provided as CSV file. To this end, the application needs to convert the document into electronically readable text (1), extract any relevant information from the document (2), and search for this information in the ERP excerpt (3). If the extracted information is found, it is counted as reconciled. The text extraction step can be performed via OCR, which is well researched and available through several open-source and commercial applications, such as Tesseract, Abbyy, Amazon Textract, Azure Cognitive Services OCR, or Google's Cloud Vision API. The reconciliation step can be realized through well-studied name matching methods such as string distances (Cohen et al., 2003). The IE step is therefore crucial to this application; if it yields poor results, the reconciliation will fail – or even worse – lead to wrongfully reconciled information. As described above, the application of ML to IE from layout-rich documents receives increasing attention from the research community, yielding a continuous stream of new approaches to solve this problem. The utilization of ML within an application is a non-trivial endeavor, as it is associated with both technical and non-technical challenges (Baier et al., 2019). Considering these aspects, we focus in this research on the development of the application's pipeline intended to perform the IE task, as it fundamentally contributes to the overall success of the application. By designing the corresponding artifact, we address a more abstract class of problems: designing IE pipelines for audit applications.

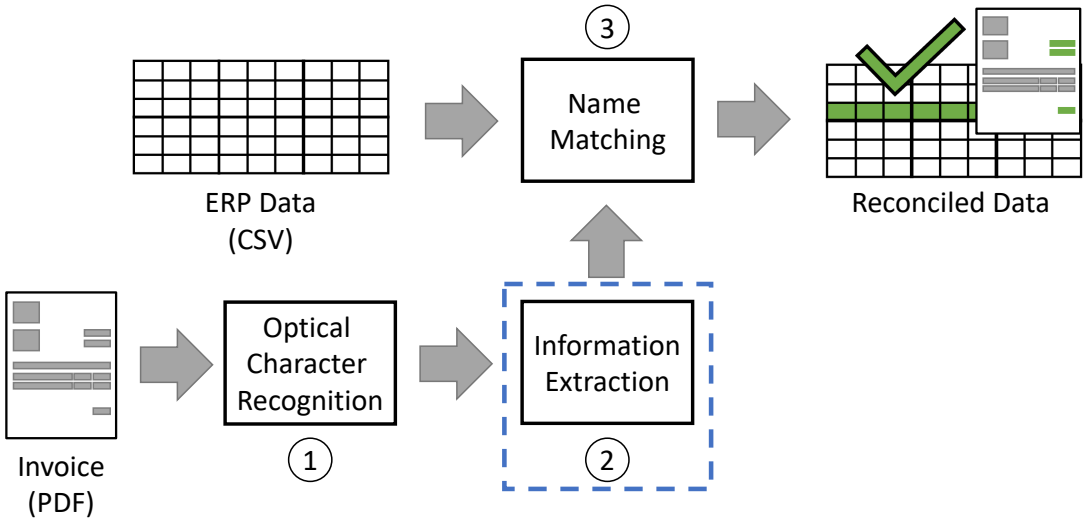

Figure V.2: Processing flow of the TOD application. The blue dashed line indicates the scope of our work within the application.

V.3.1 Problem Formulation

The problem addressed in this research is mainly inspired from practice, which corresponds to the "practice-inspired research" principle in ADR (Sein et al., 2011). The decision to employ IE was made to avoid falsely reconciled information, e.g., falsely reconciling the payment due date from the invoice with an erroneously entered invoice date from the ERP data. Thus, the base requirement for the application's IE component was to perform the extraction task with the highest achievable accuracy. Adding to this base requirement, the product owner and solution architect set a list of business- and technology related requirements:

- **Adaptability to locales**: The IE component should be able to process different languages and sets of extracted entities. The audit firm faces a multitude of clients, operating across a variety of countries and each receiving invoices from multiple countries. This implies that a multitude of languages will be encountered in the invoices, and also that the diverse legal requirements will determine which information must be included on an invoice, which also impacts their respective formats (e.g., different date formats or bank account number formats).
- **Flexibility with respect to labeled data**: This requirement amends the previous one. As labelled data is expensive, an abundance thereof cannot be assumed for all languages or sets of extracted entities.
- **Capability to process large amounts of invoices**: To play out its strength, the IE component should be able to process large amounts of invoices. Less so for the training of IE models, but rather for scoring (i.e., retrieving predictions from) the invoices to be audited.

- **Integration**: The application is intended to enable users to train IE models for a set of labelled data by pressing a button. This requires that the IE component encompasses and automates the whole training procedure in addition to the scoring.

V.3.2 Design Cycles

Following the ADR research methodology, the design of the IE pipeline emerged through several BIE cycles (Sein et al., 2011), which we describe in this section.

Cycle 1: Laying the foundation. The first cycle was conducted from late 2019 until the middle of 2020. The goal of the first cycle was to devise a research pipeline to allow the implementation and testing of models for IE from invoices and similar documents. We started the cycle by conducting a literature study. At the time, few articles had been published on layout-aware IE from invoices (Denk and Reisswig, 2019; Katti et al., 2018; Liu et al., 2019; Lohani et al., 2019; Zhao et al., 2019). This step gave us an idea of how to construct the pipeline, as most works that propose and evaluate new ML models use a well-established experimental process consisting of data preparation (covering acquisition, cleaning, annotation, and splitting), feature extraction, model training, and model evaluation.

We implemented the individual steps of the pipeline for a self-designed model by using an initial experimental data set and selected state-of-the-art technologies like PyTorch. We designed a model based on graph attention networks, taking a word-level graph as input in the same way as Lohani et al. (2019). Our model initially achieved a satisfactory macro averaged F_1 score of 0.8753 on the experimental data set. While the stakeholders were generally satisfied with this technical implementation, they criticized the lack of proper integration, as the individual steps of the pipeline were still orchestrated manually.

Cycle 2: Integration and expansion. This second cycle started in the second half of 2020. In response to feedback from the evaluation phase of the previous cycle, we aimed for an automation of the pipeline. Additionally, we endeavored to implement more models from recent literature and to evaluate them, to select the best model available to be deployed to the application.

We devised individual python scripts for each of the pipeline steps for our initial model. Using these scripts as blueprints, we implemented further models: Chargrid (Katti et al., 2018), BERTgrid (Denk and Reisswig, 2019), the models proposed by Liu et al. (2019) and Lohani et al. (2019), as well the newly introduced transformer-based model LayoutLM (Xu, Li, et al., 2020). The implementation of these models helped us to identify potential for abstraction and modularization of the code base.

The evaluation of this cycle was conducted by demonstrating evaluation results to show the validity of the model implementations. The results are given in Table V.1. The results show that most models are significantly less accurate on out-of-sample layouts. This bias stems from a skewed distribution of vendors, and therefore invoice layouts, in our experimental data set. Leaving this discrepancy undetected could lead to suboptimal model choices: for example, the random forest model shows a higher F_1 score overall than the model proposed by Lohani et al. (2019) or BERTgrid, but it is significantly worse

Table V.1: Evaluation results (macro averaged F_1 scores) of the training pipeline by model on an experimental data set.

Model	Complete test set	In-sample layouts	Out-of-sample layouts
BERTgrid	0.7546	0.7891	0.4704
Chargrid	0.7477	0.7991	0.3785
Liu et al. (2019)	0.6067	0.6490	0.2825
Lohani et al. (2019)	0.7511	0.7948	0.4661
LayoutLM	0.8532	0.8761	0.7019
Random Forest	0.7744	0.8288	0.3989

when only considering vendors that were not represented in the training set.

The status of the pipeline was further examined by the application's product owner and lead architect. The feedback was that the training pipeline be generalized, so that it could accept arbitrary labels (information entities) and invoice data sets. Furthermore, the models should be deployed in such a way that they could score large quantities of incoming invoices at once.

Cycle 3: Adding the scoring pipeline. The third cycle started toward the end of 2021. In response to the feedback from the last cycle, the goal for this cycle was to operationalize the pipelines such that the models could be trained with an arbitrary set of invoices and then be deployed to production to process large numbers of invoices at once.

Due to the modular and abstract structure of our codebase, it was convenient to devise a scoring pipeline for the previously implemented models. The methods and classes could be reused, effectively mirroring the training pipeline. To prevent the pipeline from breaking when dealing with new data, we added a pipeline step that checked that incoming data for consistency, e.g., whether all columns that our modules expect were in the OCR outputs, and whether the columns had the right data type, and the values were in the right range.

To evaluate the outcomes of the third cycle, we presented the improvements to the training pipeline and the scoring pipeline to the product owner and the solution architect. The outcomes were in general well received; however, it was proposed to export the trained models to the Open Neural Network Exchange (ONNX) format for scoring, to reduce dependencies from the various ML frameworks used in the training pipeline. This way, the scoring could be instantiated using a leaner Python environment. This feedback was implemented after the third cycle.

Reflection and Learning. In the ADR methodology, the reflection and learning step serves the purpose of moving conceptually from the built artifact to a higher level of abstraction, thus addressing a broader class of problems. In our case, this would be the design of IE pipelines for audit applications beyond the TOD. Our approach to generalization was to reflect on our learning process against the theory provided by Krieger et al. (2021). This allowed us to judge whether a design decision made for the constructed artifact would address the needs of the audit domain, and therefore have a wider applicability, justifying its elevation into a DP.

V.4 Results

In this section, we describe the resulting artifact in detail. We further relate our findings back to the process theory of ADA adoption in auditing (Krieger et al., 2021), allowing us to propose DPs that address the design of IE pipelines for audit applications as a broader class of problems.

V.4.1 The Information Extraction Pipeline(s)

The artifact that emerged through the design cycles is the IE backend for an application that automates the TOD. It encompasses two types of pipelines, which are depicted in Figure 3: one pipeline to train, optimize, and evaluate the models (1), and one pipeline for scoring incoming data using a previously trained model (2). This section provides details on the technology used to implement the pipeline, the steps within the pipelines along with the methodology employed, and their implementation.

Core technology. As most of the models we sought to benchmark were novel neural network architectures and relied on customized document representations (e.g., graphs or grids), we needed to implement the models and the associated data representations from scratch. We therefore decided to use Python, which had become the lingua franca of data science and offered various ML libraries and frameworks, such as Tensorflow and PyTorch. To implement the neural networks, we gave preference to PyTorch, as it was gaining traction rapidly in the research community: by the end of 2019, PyTorch implementations of research papers in ML had already superseded TensorFlow (Papers with Code, 2022). To implement any graph-related components of the neural networks, we employed PyTorch Geometric (Fey and Lenssen, 2019). For loading pretrained models such as BERT, LayoutLM, or Word2Vec we used Huggingface's transformers and Gensim. To utilize the Microsoft Azure cloud resources provided by the firm, we employed the Azure Machine Learning (AML) python software development kit. Through AML, we were able to train the models and score data across different GPU-enabled computing nodes in a cluster. To reduce the amount of boilerplate code needed to run the training of ML models on graphics processing units (GPUs), we utilized PyTorch Lightning (PTL). PTL implements several classes and functions that facilitate the configuration of the training process of PyTorch models and the control of ML experiments, such as the logging of training and evaluation metrics for AML. AML utilizes the MLFlow library to track and manage models as well as experiments, and Docker to govern the environment in which the python scripts are executed, such as the version of the above-mentioned libraries.

Model and document representations. Most models rely on a model-specific document representation. Therefore, we decided to implement a single *data set* class for each model. The primary purpose of these classes was to provide the data in a way that they could be consumed by the model through the PyTorch data loaders. The classes implement methods that prepare the document representations and features and fuse them. The class also implements a method that provides the distribution of labels in the data set, which can then be used for class weighting. The implementation of the models follows a similar logic – in PyTorch, models are implemented as classes with a *forward()* method that describes their forward propagation. Through PTL, the model class is extended by several methods, which govern the training and validation steps as well as the optimizer

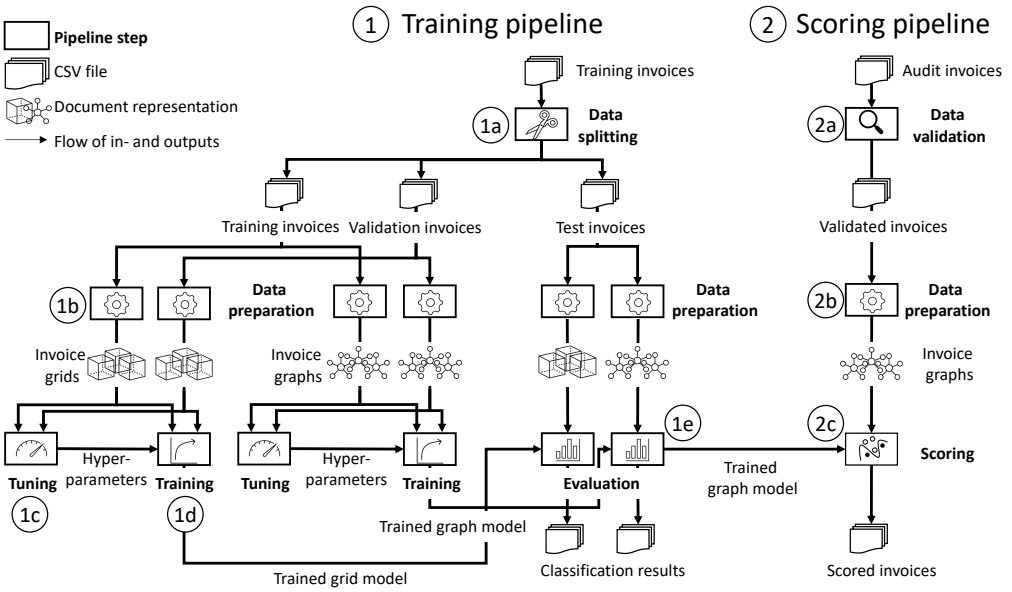

Figure V.3: Exemplary pipeline in- and output scheme for the training and evaluation of two models and the subsequent deployment of one model into the scoring pipeline.

settings. Further, the model classes take a python dictionary as one parameter, which contains the hyperparameters that govern both the model hyperparameter (e.g., number of hidden nodes per layer) and training hyperparameters (e.g., initial learning rate; batch size).

Model training pipeline (1). As described above, our training and evaluation pipeline follows a well-established process in data science and machine learning. It begins by splitting the invoices into training, validation, and test sets (1a). These splits are stratified by vendor by default, such that the most frequent vendors are represented in all three splits, whereas less frequent vendors appear only in one of them. This allows the evaluation results to be disaggregated, at a later stage, into in-sample and out-of-sample layouts. Optionally, the splits can also be randomized. The script further validates the data, i.e., it ensures the right data types and value ranges for the columns in the csv files. Files that do not conform to the expected format are dismissed with a warning. It also checks whether the labels in the files are already numerically encoded and, if not, it creates a label encoding scheme. This label encoding scheme is later attached to the models.

Using the above-described data set and model classes, for each model we devised separate *data preparation (1b)*, *model tuning (1c)*, *model training(1d)*, and *model evaluation(1e)* scripts. The *data preparation (1b)* step generates the document representation and attaches the extracted features. These were implemented per model and were run separately for each of the data splits to avoid accidental data leakage between the splits.

The *model training (1d)* step makes use of PTL's trainer class, which calls the PTL-based *model* classes as parameters and initializes the training through its *fit()* method. The *fit()* method takes several arguments that facilitate the implementation of, for example, early

stopping, or of using GPUs for training. Early stopping is implemented to save computational resources and prevent the models from overfitting to the training data. The early stopping is configured to maximize the F1 score on the validation set. We also implemented a scheme for class weighting, to prevent the models from overfitting to more frequent classes. Class weighting affects the relative contribution of misclassified instances to the training loss per class. We implemented a scheme introduced by Paszke et al. (2016) and also used by Katti et al. (2018). Depending on the chosen parameter, the scheme yields class weights that lie on the span between leaving the class distribution as-is and completely inverting it.

The *model tuning (1c)* step acts as a wrapper around the *model training* step and instantiates the training with different hyperparameter combinations, aiming to maximize the F1 score on the validation set. We chose a set of hyperparameters optimized for all models, using Bayesian hyperparameter optimization (Snoek et al., 2012): learning rate, batch size, and the parameter for class weighting. Running the hyperparameter tuning step is optional. If it is run, the best combination of hyperparameters are then passed to the model training step.

The training step outputs a binary file containing the weights for the trained model. The *model evaluation (1e)* loads the model weights and runs the predictions for the documents in the test split. If needed, the model predictions are aggregated back to the text box level and compared against the ground truth. The evaluation step outputs a detailed classification report containing entity-wise precision, recall, F1 score, true and false positives as well as true and false negatives, and the global confusion matrix. The results were disaggregated into in-sample layouts and out-of-sample results, depending on whether or not the vendor of an invoice used in the test set was present in the training set. The goal of this was to judge whether the models generalize well to new, unseen invoice layouts, while maximizing the variety of vendor layouts for the training process. The evaluation step exports the PTL model into the ONNX format, together with the chosen hyperparameters.

Scoring pipeline (2). When the model training pipeline is run, one model is automatically selected to be deployed for scoring incoming, previously unseen invoices. The decision of which model is deployed is based on the macro-averaged F1 score from the **out-of-sample** evaluation results, unless specified otherwise. The first step in the pipeline is to validate the incoming data (2a), similar to the *data splitting* step in the training pipeline. The validated invoices are therefore prepared for scoring in the *data preparation (2b)* step. Here, the same script as in the *scoring (2c)* step then loads the chosen model, which classifies the text boxes in the invoice. As in the evaluation step, at this stage the model outputs are aggregated back to the text box level. For each text box, a probability distribution is provided for the entities the model was trained to extract. These entities (or labels) are retrieved from the model repository together with the model weights.

Implementation. We structured our codebase into two core modules: *source* and *pipelinesteps*. Figure V.4 illustrates the structure of the modules and their submodules, as well as the dependencies between them. The *source* module contains reusable classes and functions for the models and datasets, including feature engineering and postprocessing. If classes and/or functions had to be implemented model-wise, we made sure to align their signatures to the highest achievable degree. The *pipelinesteps* module then pieces these classes and methods together to form integrated pipelines. The design of

the *pipelinesteps* module was influenced by the pipeline functionality of AML: AML pipelines allow users to execute individual Python scripts with arguments ("PythonScriptStep") and to transfer inputs and outputs ("PipelineData") between them. The execution of such a pipeline can then be triggered through a RESTful API endpoint. *PipelineData* usually references a file storage residing inside Azure storage for binary files. The data is then transferred between the pipeline steps using these references as arguments

By combining PTL's MLFlow logger with AML's run context, metrics from the model training step (e.g., training and validation loss) are directly logged to the AML browser interface to monitor the training. In the evaluation step, the models are saved to AMLs model registry, which versionizes the models and associates the model weights with relevant metadata, such as the evaluation results, hyperparameters, encoded labels, and a timestamp of the training, and connects them with the data set that was used to train it.

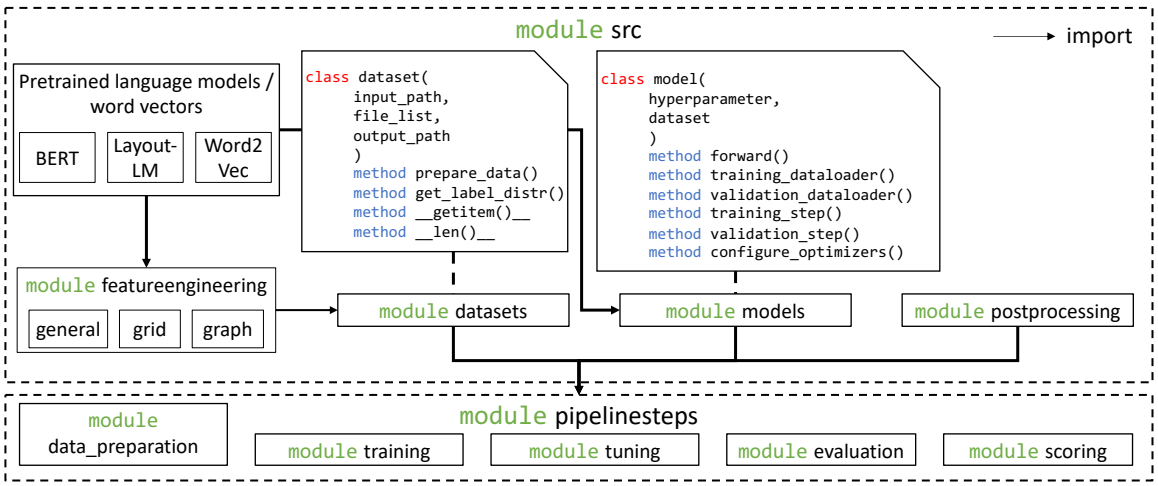

Figure V.4: Relations between individual code modules of the pipeline.

V.4.2 Design Principles for Information Extraction Pipelines in Auditing

Reflecting what we had learned from the design cycles through which the artifact emerged, we can delineate several DPs that help address a greater class of problems: designing IE pipelines for audit applications.

DP1: Separation of source code and pipeline code. Separating source code from pipeline code increases the reusability of code as, in most cases, the code for the individual pipeline steps is an instantiation of the source code. For example, they may be data set classes used to prepare the training and validation data for the training step, as well as the test data for the evaluation step. It also ensures that any changes in the model and data set classes are reflected in any pipeline steps, thus forming a symmetry between the training and scoring pipelines. The decision to devise a pipeline for batch scoring was based on the requirement to process large amounts of invoices. This in turn is mainly driven by the audit methodology and can hence be mapped to the "audit requirements"

factor in the theory by (Krieger et al., 2021).

DP2: Streamlining class and method signatures. By aligning the signatures of classes and methods across different model and dataset classes, the reusability of code can be improved. In our pipeline, the model-wise pipeline steps for data preparation, model training, evaluation and scoring all follow the same pattern. This allows for a fast and easy implementation of these steps for new models generated by the stream of research to evaluate their applicability for the TOD. The variety of models being proposed can be mapped to the complexity aspect of the technology characteristics factor in the theory proposed by (ibid.).

DP3: Standardization of input and output formats. By standardizing the formats of inputs and outputs between pipeline steps, we further improve the reusability not only of the steps within a pipeline, but of the whole pipeline itself. If the incoming CSV files follow the same structure, they can represent invoices from arbitrary languages or even other layout-rich documents. This design decision helps to address the multitude of different application settings that can be attributed to the "requirements" (labeled training data) of the technology in the theory, as well as client characteristics (diversity of client locales). Therefore, the pipeline gains broader applicability. This standardization also encompasses the inputs and outputs of the pipelines as a whole, thus evading friction with other components of the application.

DP4: Aggregation of model outputs. Depending on the type of document representation, the models output predictions on different levels of granularity, e.g., Chargrid returns pixel-level predictions, whereas LayoutLM returns token-level predictions. Aggregating these outputs back to the text box level (as produced by the OCR), we obtain a more meaningful model comparison and ensure the reusability of downstream processing steps. Similar to DP2, this DP addresses the complexity of the technology.

DP5: Flexibility of labels. Depending on the scope of the TOD (e.g., cutoff testing, record confirmation) and the legal ramifications of the country in which the TOD is carried out, the information to be extracted can vary. Allowing for flexibility in the labels (i.e., extracted information entities) allows the client characteristics and the requirements of the audit domain to be taken into account. In our artifact, we achieved this by flexibly encoding the labels and attaching the information to the model in the model repository.

DP6: Evaluation of models according to the biases in the data. While conducting experiments to test the training pipeline implementation, we identified a bias in the experimental data set: some vendors—and therefore invoice layouts—appeared more frequently than others. As we showed in the results in Table 1, this can obfuscate the model choice. Any biases in the training data—in our example the distribution of vendors—that could affect the model's accuracy should be identified and accounted for in the evaluation of models to avoid selecting sub-optimal models. We consider this interrelation between biases in the data and model accuracy to be another instance of the technology complexity.

V.5 Discussion

The goal of our research was to devise an information extraction pipeline for an audit application designed to automate the TOD. To this end, we adopted an ADR approach, embedded within an audit firm. By reflecting our learnings, we can propose DPs for a more abstract class of problems: designing IE pipelines for audit applications.

There have been few scientific accounts of AI applications for audit practice. In this paper, we introduce an IE pipeline for an audit application and provide the details for its implementation. Going beyond the brief technical description of DICR provided by Tecuci et al. (2020), our research covers how the design of the artifact can be shaped to comply with the requirements of the audit domain. The resulting artifact allows ML models to be trained and evaluated for IE from invoices for varying application contexts, such as document languages and information entities to be extracted, or the volume of labeled training data available. The pipeline is fully automatically orchestrated: models can be trained and evaluated by executing a single trigger once (new) labeled data is provided. This also applies to the scoring pipeline. By generating design knowledge in the form of DPs in addition to the artifact, we are also contributing a new methodological perspective to research on the adoption of AI in the audit domain. Previous literature has either been conceptual in nature (Appelbaum et al., 2018; Issa et al., 2016; Kokina and Davenport, 2017; Sun, 2019; Zhaokai and Moffitt, 2019) or explanatory (Krieger et al., 2021), rather than practical.

A further contribution of this paper lies in the artifact's design to facilitate the addition of new models and to support their benchmarking. The research on IE from invoices and other layout-rich documents produces a continuous stream of proposed models and approaches, as researchers pursue the highest evaluation metrics on benchmark data sets like SROIE. Yet, these standardized data sets do not necessarily reflect the reality of their models' application. This question of transferability is a challenge companies face in other application areas for ML, such as computer vision. Especially if the context of application changes, the no-free-lunch theorem suggests that the hypothesis of which model performs best should be validated. Herein also lie the practical implications of our research: we emphasize the need to re-evaluate models depending on different application contexts and propose a conceptual blueprint for practitioners to conduct this re-evaluation according to their needs.

The results presented here are not free from certain limitations, which are rooted in the employed technology and the scope of this paper. The design of the artifact is optimized toward working on AML; it may therefore not be directly transferrable to other cloud computing environments such as Amazon Web Services or Google Cloud. Also, our paper is only concerned with the IE component of the TOD application. It does not consider dependencies with upstream components, e.g., the impact of OCR on the accuracy of the models, nor does it contemplate the effect on the name matching component. These results are also limited with respect to technology adoption in auditing as other highly adoption-relevant components of the TOD application, such as the design of its user interface, are not covered.

V.6 Conclusion

The adoption of AI is still progressing slowly in the audit industry, and scientific accounts of AI applications that go beyond high-level conceptualization are rare. In this paper, we present two IE pipelines—one for training IE models and one for batch processing invoices—designed for an audit application addressing the TOD, a highly recurrent audit procedure. The design of the artifact explicitly addresses the needs of the audit domain, most importantly, the need to benchmark IE models according to the variety of their utilization settings, such as different languages and legal requirements. The structure of the pipelines' codebase further enables an efficient implementation of new models, an important step to evaluate the models proposed by the growing body of research on IE from layout-rich documents. Beyond the specific artifact, we generate DPs for IE pipelines in audit applications.

We also seek to inspire further research on the operationalization of IE within auditing or other areas of application, such as accounting. As stated above, our research is only concerned with the IE pipeline. Further research could additionally examine the interdependencies in the end-to-end process from PDF to standardized information, e.g., how to arrive at a payment date "30-05-2022" from text boxes containing the text "30", "May" and "2022", and how this is affected by the use of different OCR engines. In the context of aggregating and standardizing information from classified text boxes, future inquiries could also examine whether the task of IE could be better formulated as a generative process, e.g., as a question answering task. It could also explore how human-assisted learning paradigms such as active learning could address the need for labeled data.

V.7 References

Abdolmohammadi, Mohammad J (1999). "A Comprehensive Taxonomy of Audit Task Structure, Professional Rank and Decision Aids for Behavioral Research". In: *Behavioral Research in Accounting* 11. Publisher: American Accounting Association, pp. 51–92.

Appelbaum, Deniz A. et al. (2018). "Analytical procedures in external auditing: A comprehensive literature survey and framework for external audit analytics". In: *Journal of Accounting Literature* 40, pp. 83–101. ISSN: 07374607. DOI: 10.1016/j.acclit.2018.01.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S0737460716300611 (visited on 08/23/2021).

Baier, Lucas et al. (2019). "Challenges in the Deployment and Operation of Machine Learning in Practice". In: *Proceedings of the 27th European Conference on Information Systems (ECIS)*. 27th European Conference on Information Systems. Stockholm & Uppsala, Sweden: AIS eLibrary (AISeL), Paper: 163. ISBN: 978-1-73363-250-8.

Bunde, Enrico (2021). "AI-Assisted and Explainable Hate Speech Detection for Social Media Moderators – A Design Science Approach". In: *Proceedings of the 54th Hawaii International Conference on System Sciences*. Hawaii International Conference on System Sciences. Hawaii, United States of America. DOI: 10.24251/HICSS.2021.154. URL: http://hdl.handle.net/10125/70766 (visited on 10/05/2021).

Butgereit, Laurie et al. (2018). "A Design Science Model for the Application of Data Mining and Machine Learning Models on Constrained Devices in Low Bandwidth Areas". In: *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)*. 2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD). Durban, South Africa: IEEE, pp. 1–7. ISBN: 978-1-5386-3060-0. DOI: 10.1109/ICABCD.2018.8465407. URL: https://ieeexplore.ieee.org/document/8465407/ (visited on 10/05/2021).

Chalkidis, Ilias and Ion Androutsopoulos (2017). "A Deep Learning Approach to Contract Element Extraction". In: *Legal Knowledge and Information Systems*, p. 10.

Cohen, William W. et al. (2003). "A Comparison of String Distance Metrics for Name-Matching Tasks". In: *Proceedings of the 2003 International Conference on Information Integration on the Web*. IIWEB'03. event-place: Acapulco, Mexico. AAAI Press, pp. 73–78.

Denk, Timo I. and Christian Reisswig (2019). "BERTgrid: Contextualized Embedding for 2D Document Representation and Understanding". In: *arXiv:1909.04948 [cs]*. arXiv: 1909.04948. URL: http://arxiv.org/abs/1909.04948 (visited on 02/16/2021).

Elwany, Emad et al. (2019). *BERT Goes to Law School: Quantifying the Competitive Advantage of Access to Large Legal Corpora in Contract Understanding*. _eprint: 1911.00473.

Fayyad, Usama et al. (1996). "The KDD process for extracting useful knowledge from volumes of data". In: *Communications of the ACM* 39.11. Publisher: ACM New York, NY, USA, pp. 27–34.

Fey, Matthias and Jan Eric Lenssen (2019). "Fast graph representation learning with PyTorch Geometric". In: *arXiv:1903.02428 [cs.LG]*. URL: https://arxiv.org/abs/1903.02428 (visited on 12/09/2022).

Garncarek, Lukasz et al. (2021). "LAMBERT: Layout-Aware Language Modeling for Information Extraction". In: *Document Analysis and Recognition – ICDAR 2021*. Ed. by Josep Lladós et al. Cham: Springer International Publishing, pp. 532–547. ISBN: 978-3-030-86549-8.

Google (2022). *MLOps: Continuous delivery and automation pipelines in machine learning*. URL: https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning?hl=de (visited on 11/08/2022).

Hapke, Hannes Max and Catherine Nelson (2020). *Building machine learning pipelines: automating model life cycles with TensorFlow*. First edition. OCLC: on1138611607. Sebastopol, California: O'Reilly Media, Inc. 337 pp. ISBN: 978-1-4920-5319-4.

Hevner, Alan R. et al. (2004). "Design Science in Information Systems Research". In: *MIS Q.* 28.1. Place: USA Publisher: Society for Information Management and The Management Information Systems Research Center, pp. 75–105. ISSN: 0276-7783.

Hu, Xiang and Wenwei Su (2021). "Information Extraction from Contract Based on BERT-BiLSTM-CRF". In: *Advancements in Mechatronics and Intelligent Robotics*. Ed. by Zhengtao Yu et al. Singapore: Springer Singapore, pp. 109–115. ISBN: 978-981-16-1843-7.

Huang, Yupan et al. (2022). *LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking*. DOI: 10.48550/ARXIV.2204.08387. URL: https://arxiv.org/abs/2204.08387.

ISA 500, IAASB (2009). *International Standard On Auditing 500 Audit Evidence*. URL: https://www.ifac.org/system/files/downloads/a022-2010-iaasb-handbook-isa-500.pdf (visited on 09/16/2022).

Issa, Hussein et al. (2016). "Research Ideas for Artificial Intelligence in Auditing: The Formalization of Audit and Workforce Supplementation". In: *Journal of Emerging Technologies in Accounting* 13.2, pp. 1–20. ISSN: 1558-7940, 1554-1908. DOI: 10.2308/jeta-10511. URL: https://meridian.allenpress.com/jeta/article/13/2/1/115980/Research-Ideas-for-Artificial-Intelligence-in (visited on 10/05/2021).

Katti, Anoop Raveendra et al. (2018). "Chargrid: Towards Understanding 2D Documents". In: *arXiv:1809.08799 [cs]*. arXiv: 1809.08799. URL: http://arxiv.org/abs/1809.08799 (visited on 02/16/2021).

Kim, Geewook et al. (2021). *OCR-free Document Understanding Transformer*. DOI: 10.48550/ARXIV.2111.15664. URL: https://arxiv.org/abs/2111.15664.

Kokina, Julia and Thomas H. Davenport (2017). "The Emergence of Artificial Intelligence: How Automation is Changing Auditing". In: *Journal of Emerging Technologies in Accounting* 14.1, pp. 115–122. ISSN: 1558-7940, 1554-1908. DOI: 10.2308/jeta-51730. URL: https://meridian.allenpress.com/jeta/article/14/1/115/116001/The-Emergence-of-Artificial-Intelligence-How (visited on 04/19/2021).

Krieger, Felix et al. (2021). "Explaining the (non-) adoption of advanced data analytics in auditing: A process theory". In: *International Journal of Accounting Information Systems* 41, p. 100511. ISSN: 14670895. DOI: 10.1016/j.accinf.2021.100511. URL: https://linkinghub.elsevier.com/retrieve/pii/S1467089521000130 (visited on 08/23/2021).

Liu, Xiaojing et al. (2019). "Graph Convolution for Multimodal Information Extraction from Visually Rich Documents". In: *arXiv:1903.11279 [cs]*. arXiv: 1903.11279. URL: http://arxiv.org/abs/1903.11279 (visited on 04/19/2021).

Lohani, D. et al. (2019). "An Invoice Reading System Using a Graph Convolutional Network". In: *Computer Vision – ACCV 2018 Workshops*. Lecture Notes in Computer Science 11367. Ed. by Gustavo Carneiro and Shaodi You, pp. 144–158. DOI: 10.1007/978-3-030-21074-8_12. URL: http://link.springer.com/10.1007/978-3-030-21074-8_12 (visited on 04/19/2021).

Manita, Riadh et al. (2020). "The digital transformation of external audit and its impact on corporate governance". In: *Technological Forecasting and Social Change* 150 (September 2019). Publisher: Elsevier, p. 119751. ISSN: 00401625. DOI: 10.1016/j.techfore.2019.119751. URL: https://doi.org/10.1016/j.techfore.2019.119751.

Moffitt, Kevin C. et al. (2018). "Robotic Process Automation for Auditing". In: *Journal of Emerging Technologies in Accounting* 15.1, pp. 1–10. ISSN: 1558-7940, 1554-1908. DOI: 10.2308/jeta-10589. URL: https://meridian.allenpress.com/jeta/article/15/1/1/9413/Robotic-Process-Automation-for-Auditing (visited on 10/06/2021).

Neuhauser, Linda et al. (2013). "Using design science and artificial intelligence to improve health communication: ChronologyMD case example". In: *Patient Education and Counseling* 92.2, pp. 211–217. ISSN: 07383991. DOI: 10.1016/j.pec.2013.04.006. URL: https://linkinghub.elsevier.com/retrieve/pii/S073839911300164X (visited on 10/05/2021).

Papers with Code (2022). *Trends - Frameworks*. URL: https://paperswithcode.com/trends (visited on 07/29/2022).

Paszke, Adam et al. (2016). "Enet: A deep neural network architecture for real-time semantic segmentation". In: *arXiv:1606.02147 [cs.CV]*. DOI: 10.48550/arXiv.1606.02147. URL: https://arxiv.org/abs/1606.02147 (visited on 12/09/2022).

Salijeni, George et al. (2019). "Big Data and changes in audit technology: contemplating a research agenda". In: *Accounting and Business Research* 49.1, pp. 95–119. ISSN: 21594260. DOI: 10.1080/00014788.2018.1459458.

Sein et al. (2011). "Action Design Research". In: *MIS Quarterly* 35.1, p. 37. ISSN: 02767783. DOI: 10.2307/23043488. URL: https://www.jstor.org/stable/10.2307/23043488 (visited on 10/05/2021).

Snoek, Jasper et al. (2012). "Practical Bayesian Optimization of Machine Learning Algorithms". In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira et al. Vol. 25. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2012/file/05311655a15b75fab86956663e1819cd-Paper.pdf.

Sun, Ting (Sophia) (2019). "Applying Deep Learning to Audit Procedures: An Illustrative Framework". In: *Accounting Horizons* 33.3, pp. 89–109. ISSN: 0888-7993, 1558-7975. DOI: 10.2308/acch-52455. URL: http://meridian.allenpress.com/accounting-horizons/article/33/3/89/427545/Applying-Deep-Learning-to-Audit-Procedures-An (visited on 08/23/2021).

Tecuci, Dan G. et al. (2020). "DICR: AI Assisted, Adaptive Platform for Contract Review". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.9, pp. 13638–13639. ISSN: 2374-3468, 2159-5399. DOI: 10.1609/aaai.v34i09.7106. URL: https://aaai.org/ojs/index.php/AAAI/article/view/7106 (visited on 10/05/2021).

Vaswani, Ashish et al. (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wang, Jiapeng et al. (2022). *LiLT: A Simple yet Effective Language-Independent Layout Transformer for Structured Document Understanding*. DOI: 10.48550/ARXIV.2202.13669. URL: https://arxiv.org/abs/2202.13669.

Wirth, Rüdiger and Jochen Hipp (2000). "CRISP-DM: Towards a standard process model for data mining". In: *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. Vol. 1. Manchester, pp. 29–39.

Xin, Doris et al. (2021). "Production Machine Learning Pipelines: Empirical Analysis and Optimization Opportunities". In: *Proceedings of the 2021 International Conference on Management of Data*. SIGMOD '21. New York, NY, USA: Association for Computing Machinery, pp. 2639–2652. ISBN: 978-1-4503-8343-1. DOI: 10.1145/3448016.3457566. URL: https://doi.org/10.1145/3448016.3457566 (visited on 11/08/2022).

Xu, Yang, Yiheng Xu, et al. (2022). "LayoutLMv2: Multi-Model Pre-Training For Visually-Rich Document Understanding". In: *arXiv:2012.14740 [cs.CL]*, p. 17. DOI: 10.48550/arXiv.2012.14740. URL: https://arxiv.org/abs/2012.14740 (visited on 12/09/2022).

Xu, Yiheng, Minghao Li, et al. (2020). "LayoutLM: Pre-training of Text and Layout for Document Image Understanding". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1192–1200. DOI: 10.1145/3394486.3403172. arXiv: 1912.13318. URL: http://arxiv.org/abs/1912.13318 (visited on 05/09/2021).

Xu, Yiheng, Tengchao Lv, et al. (2021). *LayoutXLM: Multimodal Pre-training for Multilingual Visually-rich Document Understanding*. DOI: 10.48550/ARXIV.2104.08836. URL: https://arxiv.org/abs/2104.08836.

Yu, Wenwen et al. (2020). "PICK: Processing Key Information Extraction from Documents using Improved Graph Learning-Convolutional Networks". In: *arXiv:2004.07464 [cs]*. DOI: 10.48550/arXiv.2004.07464. arXiv: 2004.07464. URL: http://arxiv.org/abs/2004.07464 (visited on 02/16/2021).

Zhao, Xiaohui et al. (2019). "CUTIE: Learning to Understand Documents with Convolutional Universal Text Information Extractor". In: *arXiv:1903.12363 [cs]*. DOI: 10.48550/arXiv.1903.12363. URL: http://arxiv.org/abs/1903.12363 (visited on 02/16/2021).

Zhaokai, Yan and Kevin C. Moffitt (2019). "Contract Analytics in Auditing". In: *Accounting Horizons* 33.3, pp. 111–126. ISSN: 0888-7993, 1558-7975. DOI: 10.2308/acch-52457. URL: http://meridian.allenpress.com/accounting-horizons/article/33/3/111/427543/Contract-Analytics-in-Auditing (visited on 10/06/2021).

Appendix

Appendix to chapter II **176**
Appendix to chapter IV **182**
Author contributions **183**
Complete list of publications **187**
Curriculum Vitæ **188**

Appendix A

Appendix to chapter II

A.1 Interview quotes

All quotes have been translated from German.

"And if you look at the newer technologies, for example, data mining, big data analytics, AI, or machine learning, yes, we're seeing that audit firms use them, that there are efforts. But currently almost only in their consulting practices and less in their audit practices. It may very well happen at some point, that they will engage more with these technologies. [...] But we have not yet been able to see this in recent audits." (IP15) (A.1)

"Where we admittedly still have a bit of a problem [...] is to find the right use-cases for artificial intelligence." (IP12) (A.2)

"[...] no one can guide these data science people. [...] None of their managers can tell them: We have this and that project going on now, let's have a look at it, maybe we can analyze that better. Because companies, to do data science, in my opinion, do not lack the competence, but the use-cases." (IP2) (A.3)

"[...] because it requires investments, and the big problem is, of course, which all sectors actually have, they also need specialist [technological] knowledge. Small and medium-sized practices, in particular, have difficulties in obtaining suitable specialist staff. And that is why it is simply difficult to introduce new technologies in smaller firms." (IP10) (A.4)

"We have found that clients are quite individual, even if they use the same ERP system, for example." (IP8) (A.5)

"And every system has a different data model. So, it contains the same information, but has a different structure." (IP9) (A.6)

"I must nevertheless take into account everything that the auditing standards prescribe, even if I may be able to audit in a completely different, perhaps more intelligent way, through the use of data analytics." (IP8) (A.7)

"What we as auditors need to make sure is that the data does resemble the reality." (IP8) (A.8)

"[...] I need to prove the reliability [of the technology]: The system has been fed this way [with data], works that way, and then comes for those reasons to this result. This also applies to the interpretation of results; everything must be made comprehensible." (IP15) (A.9)

"Everywhere where professional judgment is required, where I have to consider: Is this okay or not? Do I need to modify the audit opinion? A lot of people think that can't be replaced". (IP12) (A.10)

"You may quickly overlook that [the need for disciplinary spanning skills] and think: We just need a handful of computer scientists and a handful of auditors, and then something good will happen. But that's simply not the case." (IP2) (A.11)

"[...] from an audit point of view, we don't see any irregularities, everything seems fine, but did you actually know that if you pay your invoices three days earlier, you can raise significantly more cash discount [...]" (IP8) (A.12)

"[...] in most pitches, especially when it comes to larger mandates, where it's not just pure price competition, then of course it's all about what additional insights we can generate for the client." (IP8) (A.13)

"We achieved a certain fit there, such that the auditor does not have to change [his mindset] towards the reports, but the we change the reports in such a way, that the auditor can use them properly." (IP8) (A.14)

"Before I invest my time and run analyses which I can throw away afterwards, and possibly even make insinuations towards my client that are plain wrong, I'd rather stick with my old audit procedures." (IP10) (A.15)

A.2 Overview of empirical studies on ADA in auditing

Citation	Type	Country	ADA Technology	Focus	Results
Al-Htaybat and Alberti-Alhtaybat, 2017	Interviews	Saudi Arabia	Big Data (Analytics) (BDA)	Big data (analytics) in corporate report-ing,including auditing	According to the authors, the application of big data to corporate reporting is paradox: While big data is supposed to simplify reporting, the interviewees perceive the technology as very complex and identify a lack of necessary skills (statistics, programming) in their organizations.
Barr-Pulliam et al., 2020	Experiment	U.S.	Data Analytics (DA)	Professional skepticism applied by auditors toward the outputs of analytic tool in the presence of false positives (data points falsely flagged as anomalous)	Applied skepticism is low, even when positively rewarded. The authors recommend carefully tuning data analytics tools towards reducing false positives.
Chiu and Jans, 2019	Case study	U.S.	Process Mining (PM)	Application of process mining to the evaluation of internal controls on the example of a bank	The authors show the applicability of PM to the evaluation of the effectiveness of internal controls. PM is used to analyze process variants, to check the segregation of duties, to investigate personnel data and to perform analyses on timestamp data. The authors stress the importance of data integrity for these kind of analyses and further argue that they need to become automated to increase efficiency.
Dagilienė and Klovienė, 2019	Interviews	Lithuania	Big Data (Analytics) (BDA)	Motivation for audit firms to use big data (analytics)	Three types of motivating factors are identified: • Company-related factors referring to characteristics of both audit firm and client company, for example, size, industry sector, use of technology, relationship between audit firm and client. • Technology-related factors: Digitization of business processes, accounting software used by the client company, and availability of professionals with big data analytics experience. • Institutional factors: Competition in the audit market, regulation, and education.

Citation	Type	Country	ADA Technology	Focus	Results
Jans et al., 2014	Case study	U.S.	Process Mining (PM)	Exploration of the applicability of process mining to auditing on the example of a bank.	The case study demonstrates the applicability of process mining to auditing. Process mining focuses on paths of transactions using meta-data generated automatically by the IT system. It further uses the full population of transactions instead of samples. While this is an important extension of the focus of external audits, it can also create a problem of an "alarm flood" of false positives.
Eilifsen et al., 2019	Interviews, Survey	Norway	Data Analytics (DA)	Use of data analytics in auditing practice	The actual use of DA remains limited and the use of ADA even more so. The authors find that DA is mostly used for clients with integrated IT-systems, and for newly acquired engagements. They further find that the results of DA are mostly used as supplementary audit evidence.
Haddara et al., 2018	Interviews, Survey	Taiwan	(Big) Data Analytics (BDA)	The facilitators and inhibitors of big data analytics adoption	Several adoption barriers are identified in the paper: • Reluctance of clients to provide confidential data to auditors • Cost-Benefit uncertainty • Lack of big data-related skills in audit firms • Lack of hardware infrastructure • Data control and storage • Data integration and storage • Lack of guidelines for data usage and regulation
Hampton and Stratopoulos, 2016	Survey	Canada	Data Analytics (DA)	The study examines the motivation for DA adoption, the effect of supporting environment on DA usage, and the tradeoff underlying training strategies (DA expertise vs. diversity of DA tools)	The study finds that client expectations of DA use positively affect auditor use of DA. The use of DA by auditors in turn positively affects the confidence in the audit opinion, more in big audit firms than small- and medium-sized audit firms. The expectations of clients positively affect the availability of training opportunities within the audit firm, which in turn increases the use of DA. Both the DA proficiency of auditors and the diversity of DA tools are positively associated with confidence in the audit opinion. However, the results indicate that expertise in one tools is more important than basic knowledge of a diversity of tools. The authors conclude that this will pose a challenge to audit firms, professional organizations, and the education system, as auditors are required to be proficient in numerous interrelated emerging technologies.

Citation	Type	Country	ADA Technology	Focus	Results
Manita et al., 2020	Interviews	France	Digitization in general; Data Analytics (DA), Artificial Intelligence (AI)	Impact of digitization on auditing through the lens of auditing as a corporate governance tool	Through digitization, audit firms will be able to deliver a more relevant, value-added audit by shifting the focus on high value-added tasks and transitioning from sample-based auditing to an audit of all data. DA and AI are likely to enable audit firms to introduce new service offers and improve audit quality. Digitization will probably shift the skill profile of auditors toward technological skills and foster a culture of innovation in audit firms.
Michael and Dixon, 2019	Survey	U.K.	Big Data (Analytics) (BDA)	The effect of BDA use by auditors on the expectation gap in audits of voluntary disclosures	The results reveal that the use of BDA decreases the expectation gap in audits of voluntary disclosures. This kind of audit requires BDA as well, as the source data is usually "big" in terms of volume and variety, as opposed to statutory financial audits, which are mainly concerned with structured financial data. The assurance of voluntary disclosures is a promising additional service enabled by ADA that audit firms can offer.
Rose et al., 2017	Experiment	U.S.	Big Data (Analytics) (BDA)	The effect of the timing of BDA visualization in the audit process on auditor judgement	The experiment indicates that auditors do not identify crucial patterns in big data visualizations before forming initial expectations by applying traditional audit procedures. They are also more likely to express concerns regarding potential misstatements when the visualizations do not match their initially formed expectations. These findings contradict the oft-stated belief that BDA visualizations would support auditors in directing the audit toward critical areas before applying traditional procedures.
Salijeni et al., 2019	Interviews	U.K., Belgium	Big Data (Analytics) (BDA)	The impact of BDA adoption on the audit profession	The authors identify three key aspects in this regard: The impact on the relationship between auditor and client (greater transparency for the client, technology spillover effects, provision of additional—possibly non-audit—services). The consequences of technology use for the conduct of audit engagements (improvement of audit quality and efficiency, focus on more complex issues, shift from traditional data interrogation to analysis of full populations, and even unstructured data). Common challenges associated with embedding big data analytics (lack of relevant skills in audit firms, possible high frequency of false positives, reluctance toward growing influence of data scientists)

Appendix B

Appendix to chapter IV

Table B.1: Key items annotated in the data set

Key Item	Data Type	Description	Annotated text boxes
Invoice number	Alphanumeric	The unique identifier of each invoice. May appear multiple times on the same invoice.	1,147
Issue date	Date	The date on which the invoice has been issued. May appear multiple times on the same invoice.	1,377
Total amount	Decimal	The gross total amount due per invoice. Includes any tax, discounts, or other deductions.	1,096
Recipient address	Text, integer, alphanumeric	Sequence of text indicating the name, street, house number, ZIP code, and country of the invoice's recipient.	13,683
Supplier address	Text, integer, alphanumeric	Sequence of text indicating the name, street, house number, ZIP code, and country of the invoice's issuer.	12,857
Supplier VAT ID / tax ID	Alphanumeric	The (value-added) tax ID of the invoice's issuer follows a fixed pattern in EU invoices, varying across other countries, mostly not present in US invoices.	1,345
Total tax amount	Decimal or percentage	The share of the total amount due to value added or sales tax. Usually expressed as a sum or percentage of the net amount.	383
Due date	Date	The date upon which the payment is due.	822
Service date	Date	The date on which the related services have been performed or the delivery of the related products has occurred. In the case of services, this can also be a range of dates and be part of a line item.	1,292
Line item description	Text	A textual description of the goods or services provided.	13,289
Line item quantity	Integer	The quantity of goods or services provided. In the case of services, this might relate to temporal units.	1,747
Line item unit price	Decimal	The price (gross or net) per unit.	1,632
Line item tax amount	Decimal, percentage	The (value-added, sales) tax due for the goods or services provided, expressed either in an amount or as a percentage of the net price.	1,359
Line item subtotal	Decimal	The (gross or net) subtotal of each line item, equivalent to quantity x price (+ tax).	1,994

Appendix C

Author contributions

Table C.1: Publication I: Authors qualitative & quantitative contribution

Krieger	Drews
Conceptual research design	
×	×
Planning of research activities	
×	×
Data collection	
×	
Data analysis and interpretation	
×	×
Manuscript writing	
×	×
Publication equivalence value	
65%	35%

Table C.2: Publication II: Authors qualitative & quantitative contribution

Krieger	Drews	Velte
Conceptual research design		
×	×	×
Planning of research activities		
×	×	
Data collection		
×		
Data analysis and interpretation		
×	×	
Manuscript writing		
×		
Publication equivalence value		
70%	20%	10%

Table C.3: Publication III: Authors qualitative & quantitative contribution

Krieger	Drews	Funk	Wobbe
Conceptual research design			
×	×	×	
Planning of research activities			
×	×		
Data collection			
×			×
Data analysis and interpretation			
×			
Manuscript writing			
×			
Publication equivalence value			
80%	10%	5%	5%

Table C.4: Publication IV: Authors qualitative & quantitative contribution

Krieger	Drews	Funk
Conceptual research design		
×	×	×
Planning of research activities		
×		
Data collection		
×		
Data analysis and interpretation		
×		
Manuscript writing		
×		
Publication equivalence value		
80%	10%	10%

Table C.5: Publication V: Authors qualitative & quantitative contribution

Krieger	Drews	Funk
	Conceptual research design	
×	×	×
	Planning of research activities	
×		
	Data collection	
×		
	Data analysis and interpretation	
×		
	Manuscript writing	
×		
	Publication equivalence value	
80%	10%	10%

Complete List of Publications

Krieger, F., Drews, P. (2018). "Leveraging Big Data and Analytics for Auditing: Towards a Taxonomy". In: ICIS 2018 Proceedings. Association for Information Systems.

Krieger, F., Drews, P., Velte, P. (2021). "Explaining the (Non-) Adoption of Advanced Data Analytics in Auditing: A Process Theory". International Journal of Accounting Information Systems, 41. Elsevier.

Krieger, F., Drews, P., Funk, B., Wobbe, T. (2021). "Information Extraction from Invoices: A Graph Neural Network Approach for Datasets with High Layout Variety". In: Ahlemann, F., Schütte, R., Stieglitz, S. (eds.) Innovation Through Information Systems. WI 2021. Lecture Notes in Information Systems and Organisation, 47. Springer.

Krieger, F., Drews, P., Funk, B. (2023). "Automated Invoice Processing: Machine Learning-Based Information Extraction for Long Tail Suppliers". Manuscript submitted for publication.

Krieger, F., Drews, P., Funk, B. (2023). "Benchmarking Machine Learning Models in Auditing: Toward an Information Extraction Pipeline for the Test of Details". Manuscript submitted for publication.

Curriculum Vitæ

Felix Friedrich Anton Krieger

Personal

Occupation	PhD candidate at Leuphana University Lüneburg | Data scientist at Ernst & Young GmbH
Nationality	German
Residence	Hamburg, Germany
Contact	felix.krieger@leuphana.de | felix.krieger@de.ey.com | ffa.krieger@gmail.com

Education

2018 – 2023	Doctoral studies at Leuphana University Lüneburg
2014 – 2017	Master of Arts in Management & Controlling/Information Systems at Leuphana University Lüneburg
2010 – 2014	Bachelor of Science in Management & Economics at University of Osnabrück

Practical experience

since 2021	Data scientist at Ernst & Young GmbH, Hamburg Assurance Digital
2018 – 2020	Stipendiary research fellow at Leuphana Universität, Lüneburg Institute of Information Systems
2016 – 2017	Working student and master's thesis candidate at KPMG AG, Hamburg Digital Finance Advisory
2014 – 2015	Working student at Kreditech Holding SSL GmbH, Hamburg Financial Business Intelligence
2014 – 2014	Intern at Unipoint Electric Ltd., Taipei (Taiwan) Controlling | Automotive Aftermarket
2013 – 2013	Intern at Robert Bosch GmbH, Karlsruhe Controlling | Automotive Aftermarket

Publications

2023	"Design Principles for the Documentation of AI" - Submitted for publication
2023	"Automated Invoice Processing: Machine Learning-Based Information Extraction for Long Tail Suppliers" - Submitted for publication
2023	"Benchmarking Machine Learning Models in Auditing: Toward an Information Extraction Pipeline for the Test of Details" - Submitted for publication
2021	"Information Extraction from Invoices: A Graph Neural Network Approach for Datasets with High Layout Variety" - International Conference on Wirtschaftsinformatik
2021	"Explaining the (Non-) Adoption of Advanced Data Analytics in Auditing. A Process Theory." - International Journal for Accounting Information Systems
2018	"Leveraging Big Data and Analytics for Auditing: Towards a Taxonomy" - International Conference on Information Systems