LEUPHANA
UNIVERSITÄT LÜNEBURG

# Machine Learning Canvases for Project Support and Application of Graph Neural Networks in Invoice Recognition

By the School of Management and Technology
of Leuphana University Lüneburg for the award of the degree of

Doctor of Economics and Social Sciences
- Dr. rer. pol. -

approved dissertation by
M.Sc. MBA Lukas-Walter Thiée

born on 20 January 1989 in Worms

Submitted on: 30 September 2025

Oral defence (disputation) on: 19 December 2025

| | | |
|---|---|---|
| First supervisor: | Prof. Dr. Burkhardt Funk | Leuphana University Lüneburg |
| First reviewer: | Prof. Dr. Burkhardt Funk | Leuphana University Lüneburg |
| Second reviewer: | Prof. Dr. Paul Drews | Leuphana University Lüneburg |
| Third reviewer: | Prof. Dr. Ciprian Daniel Neagu | University of Bradford |

The individual contributions to the cumulative dissertation project are or will be published as follows, including the framework paper if applicable:

Lukas-Walter Thiée (2021). A systematic Literature Review of Machine Learning Canvases, INFORMATIK 2021 - Die 51. Jahrestagung der Gesellschaft fur Informatik in: Computer Science and Sustainability. Gesellschaft für Informatik e.V. (GI) (Hrsg.). Bonn: Gesellschaft für Informatik e.V., S. 1221-1235 15 S. (Lecture Notes in Informatics (LNI), Proceedings - Series of the Gesellschaft fur Informatik (GI); Band P-314). https://doi.org/10.18420/informatik2021-101

Lukas-Walter Thiée (2022). Developing an Ontology for Data Science Projects to facilitate the Design Process of a Canvas, 17th International Conference on Wirtschaftsinformatik, February 2022, Nürnberg, Germany.
https://aisel.aisnet.org/wi2022/ai/ai/13/

Lukas-Walter Thiée, Felix Krieger and Burkhardt Funk (2023). Extraction of Information from Invoices – Challenges in the Extraction Pipeline, INFORMATIK 2023 - Designing Futures: Zukünfte gestalten. Joint Workshop IntDig 2023 MOC 2023; Intelligente Digitalisierung, (KI-basiertes) Management und Optimierung komplexer Systeme. Berlin. 26.-29. September 2023.
https://doi.org/10.18420/inf2023_180

Lukas-Walter Thiée (2024). Ablation Study of a multi-modal GAT Network on perfect synthetic and real-world Data to investigate the Influence of Language Models in Invoice Recognition, 18th International Conference on Document Analysis and Recognition, ICDAR 2024, held in Athens, Greece, during August 30–31, 2024.
https://doi.org/10.1007/978-3-031-70642-4_13

Lukas-Walter Thiée and Burkhardt Funk (2025). Enhancing invoice recognition with LLM Embeddings in GAT Networks, AMCIS 2025 Proceedings, Generative AI in Information and Service Systems, Montréal.
https://aisel.aisnet.org/amcis2025/sig_svc/sig_svc/6/

Year of publication: 2026

# Abstract

**Motivation:** Modern machine learning (ML) projects are complex in many respects. From an organizational perspective, this concerns, for example, the number and type of stakeholders or the collaboration of development teams. From a technical perspective, high requirements arise from the development of algorithms or the size, availability, and processing of data. These various types of complexity make it difficult for small organizations to implement ML projects purposefully. For this reason, conceptual and technical solutions must be sought to support users, managers, and developers in these contexts. From a conceptual point of view, it is advisable to create clear process descriptions and overviews in order to gain clarity about the project at hand. From a technical perspective, concrete ML use cases should be implemented to overcome technical obstacles. In the specific use case of this thesis, this refers to the improvement of Graph Attention Networks (GAT) for invoice recognition (IR). The research question - which methods and concrete models support Small and medium-sized enterprises in implementing ML - as well as the application context are derived from the collaboration with a partner company.

**Research Method:** In this dissertation, a holistic approach is applied to advance the Data Science domain. Design Science Research and Case Studies are used, to make conceptual as well as technical progress for machine learning projects. A comprehensive literature review builds the foundation for the conceptual section, upon which artifacts are built. Model development in program code as well as evaluation form the central part of the technical section. From the big picture to the detailed view, general collaboration and prioritization aspects in ML projects are first explored and then the ML use case of invoice recognition with Graph Neural Networks is examined in detail. This mixed methods approach allows for a comprehensive view of the field of application.

**Contribution:** Five research papers result in two major contributions for the practice and research community - a catalog and a model. The catalog lists ML-canvases and their corresponding fields and questions. It is derived from a comprehensive review of the related literature. This catalog supports the initiation and consistent implementation of machine learning projects. In particular, specific questions help to understand the problem that needs to be solved, the value that is supposed to be generated through the ML project, and the steps to reach a feasible solution. The second contribution is a specific model for invoice recognition. In this use case, special challenges and information types of invoice documents - visually rich documents - are analyzed and categorized. Subsequently, a pre-trained multi-modal Graph Attention Network is enhanced and tested on different datasets. Integrating an LLM in a Graph Attention Network to provide semantic embeddings is a new research approach in this thesis. Applying the model to English and German invoice documents, a significant improvement in token classification can be achieved. These contributions support machine learning projects on both a conceptual and a technical level.



# Keywords

Machine Learning Canvas, Invoice Recognition, Graph Attention Networks

# Contents

# III Appendix

# List of Figures

# List of Tables

# Acronyms

**AGI** Artificial General Intelligence. 74

**AI** Artificial Intelligence. 1, 24, 41, 42, 61, 73

**ANI** Artificial Narrow Intelligence. 74

**B2B** Business to Business. 73, 93

**B2C** Business to Consumer. 73, 93

**BERT** Bidirectional Encoder Representations from Transformers. 13, 18, 19, 95, 105, 113

**CNN** Convolutional Neural Network. 12, 113

**CORD** Consolidated Receipt Dataset for Post-OCR Parsing. 116

**CRISP-DM** Cross Industry Standard Process for Data Mining. 3, 24, 61, 63, 73

**CWA** Corona-Warn-App. 131, 133

**DASC-PM** Data Science Process Model. 20, 24, 73, 83

**DL** Deep Learning. 75

**DocILE** Document Information Localization and Extraction. 25, 93, 95, 99, 122

**DS** Data Science. 1, 41, 61

**DSR** Design Science Research. 7, 74

**ERP** Enterprise Resouce Planning. 12, 114

**FUNSD** Form Understanding in Noisy Scanned Documents. 93, 95, 116

**GAT** Graph Attention Network. 13, 17–19, 23, 24, 26, 27, 93, 113, 114

**GCN** Graph Convolutional Network. 13

**GNN** Graph Neural Network. 13, 17, 113, 114

**GPT** Generative Pretrained Transformer. 5, 27, 115

**GPU** Graphics Processing Unit. 1

**Inv3D** Invoice 3D. 17, 94, 100, 105, 117, 118, 121

**IR** Invoice Recognition. 18

**IS** Information Systems. 42, 61

**LLaMA** Large Language Model Meta AI. 19, 113

**LLM** Large Language Model. 5, 17–19, 23, 24, 26, 27, 113

**ML** Machine Learning. 1, 18, 41, 42, 61, 73, 93, 113

**MLLM** Multimodal Large Language Model. 25, 26, 122

**NER** Named Entity Recognition. 13

**NLP** Natural Language Processing. 12, 13, 19, 113

**OCR** Optical Character Recognition. 12, 19, 77, 98, 113

**PD** Private Dataset. 96, 101, 105, 117, 118, 121

**PLS** Partial Least Squares Method. 141

**RGB** Red-Green-Blue. 77

**RNN** Recurrent Neural Network. 12

**RoI** Regions of Interest. 95, 103

**RVL-CDIP** Ryerson Vision Lab Complex Document Information Processing. 95, 116

**SEM** Structural Equation Modeling. 21, 139

**SME** Small and Medium-sized Enterprise. 1, 4, 18, 27, 41

**SROIE** Scanned Receipts OCR and Information Extraction. 93, 99, 116

**YOLO** You Only Look Once. 20

# Part I

# Preamble

# Chapter 1

# Introduction

## 1.1 Research Motivation and Context

At a time when countless new digital data are created and stored every day [1], it is important to understand that the data themselves do not provide any value. They are merely an unprocessed raw material. Even if data are referred to as digital *"gold"* [2], only analytics and algorithmic processing turn data into information from which decisions can be made and value can be created. In other words, data are the fuel and algorithms are the engines. The *"digging for gold"* [3] has accelerated considerably in recent decades. With the ability to capture more and more data, to store and quickly retrieve data, and to process the data with state-of-the-art algorithms, unprecedented value can be created in various areas. Key factors here are the amount and availability of data, the provision of computing power, e.g., GPU computing, and the implementation of intelligent algorithms, i.e., machine learning (ML) *"as a branch of AI (artificial intelligence)"* [4, p. 24]. These drivers can be used in research, in medicine, in the military, in the public sector, and above all, in corporate business. The application in various areas has given rise to a whole new domain called data science (DS) [5, p. 3]. It is therefore not surprising that a large number of data science projects have been set up which, due to their inherent research character, are highly complex and place high demands on the experts carrying them out.

Whereas large players in the industry have the financial and human resources to actively research and implement ML and AI, it is difficult for smaller organizations and companies to keep up with the speed of implementation. Nevertheless, Small and medium-sized enterprises (SMEs) also seek to leverage big data and ML/AI to gain competitive advantages and improve decision-making. However, the implementation of such projects in SMEs remains a major challenge, particularly when the business does not originate from the software domain. The challenges of SMEs are driven by a combination of organizational, technical, and strategic factors.

One of the most significant obstacles is the limited availability and quality of data. Unlike large organizations, SMEs often lack the volume, variety, and veracity of data necessary to develop robust ML models [6]. Data silos, inconsistent data collection, and insufficient data infrastructure frequently result in datasets that are either too small or too noisy for effective model training. Consequently, ML solutions in SMEs often suffer from poor generalizability and limited reliability. Although non-customized off-the-shelf models and generative AI can make many tasks easier in SMEs, there is often no real automation

and use of - internal and customer - data. Not least because ethical and regulatory concerns present a major obstacle. Compliance with data protection regulations such as the General Data Protection Regulation (GDPR) demands careful data governance and risk assessment. SMEs, which often lack legal and compliance departments, may find it difficult to ensure that ML applications adhere to standards and current laws.

In addition to data limitations, SMEs face considerable resource constraints. Implementing ML solutions requires substantial investment in skilled personnel, computational infrastructure, and time, all of which are often in short supply in SME environments [7]. The recruitment and retention of data scientists or ML engineers is particularly challenging given the global talent shortage and the high cost of expertise. Furthermore, many SMEs lack the technical capacity to maintain the infrastructure needed for training and deploying ML models. Moving from a successful prototype to a production-level solution requires robust pipelines, monitoring systems, and compatibility with existing business processes, components that are frequently underdeveloped or absent in SMEs [8]. As a result, many DS projects stall after the proof-of-concept stage, never delivering operational benefits. Even when ML models are successfully deployed, SMEs often lack the resources and practices to maintain and monitor them over time. Issues such as data drift, changing business environments, and evolving customer behavior necessitate regular updates and recalibration of models, an aspect frequently overlooked in SMEs due to limited technical capacity [9].

Cultural and organizational barriers also play a non-trivial role in impeding DS and ML adoption. Data-driven decision-making requires a shift in organizational mindset, improved data literacy, and openness to algorithmic insights. SMEs may encounter internal resistance to change or lack a clear understanding of the limitations and benefits of ML technologies [10]. Strategic misalignment between DS initiatives and business goals further hampers the success of ML projects. SMEs may embark on ML projects without a clearly defined problem statement or a measurable outcome, resulting in technically sound models that offer limited business value [6]. This lack of strategic focus often leads to disillusionment with data-driven initiatives. While larger enterprises face similar problems, they can generally allocate more resources to these initiatives. Nevertheless, SMEs also have advantages compared to bigger corporations, such as faster decision-making and less hierarchical structures. However, the data and use cases that SMEs can provide make them an intriguing research object for ML projects. This is particularly because they find themselves asking whether ML/AI should be viewed more from an application or training perspective. Application means using a model exclusively for inference or generation. Training means that models would have to be (further) trained and proprietary data provided. Both options have advantages and disadvantages in terms of data protection and the controllability of the models, as well as hardware and personnel resources.

In summary, SMEs face multifaceted challenges in the adoption and implementation of ML. Addressing these challenges often requires partnerships with academic institutions, technology vendors, or external consultants who can provide the necessary expertise and support frameworks. Without such collaborative strategies, the transformative potential of DS and ML may remain largely untapped in the SME sector. One currently existing field of research in machine learning, in various kinds of enterprises, is the automated processing of documents along business processes, e.g., the recognition of invoice documents. In terms of machine learning, the question arises as to which data and models SMEs can use to create or utilize invoice recognition systems.

## 1.2 Research Gap and Questions

Due to the aforementioned challenges, SMEs lack methodological guidance for the initialization and concrete implementation of ML. The complexity of ML projects, or rather its reduction, can be approached from different directions. For example, one possibility is to divide an ML project into specific phases and, based on the tasks and questions within each project phase, determine which conceptual and technical requirements need to be addressed.

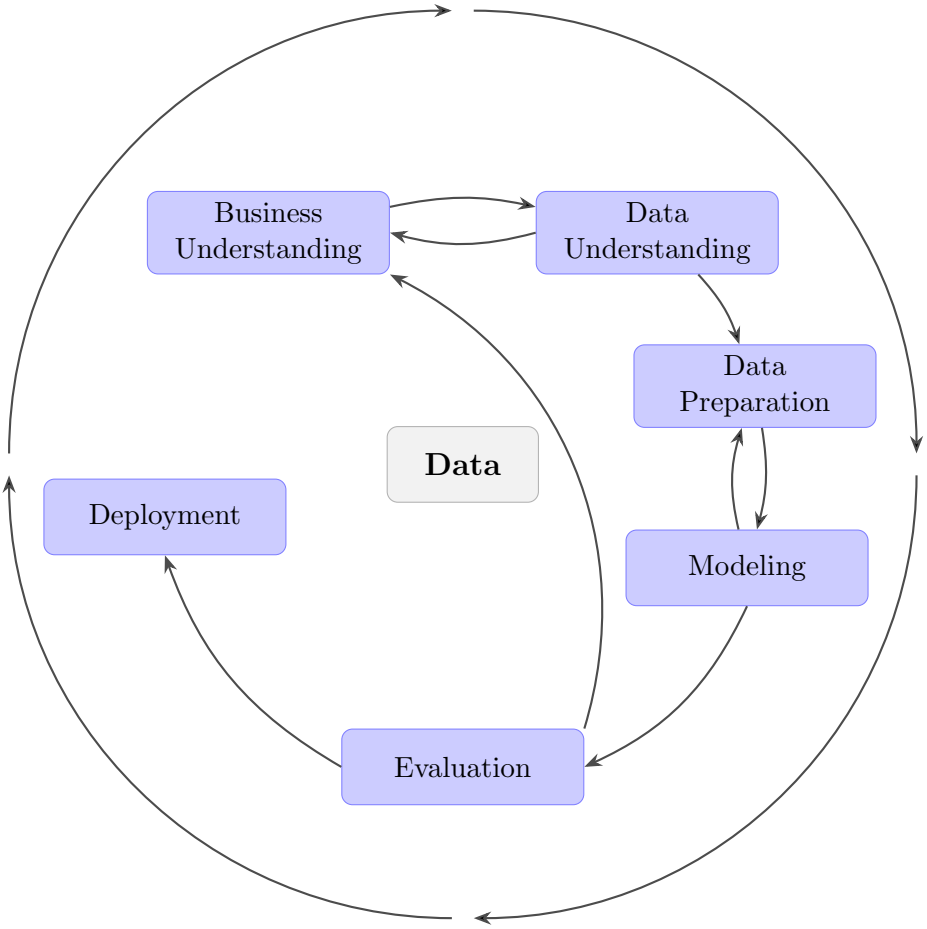


Figure 1: CRISP-DM Process Model from [11]

In the standard process for data projects, the so-called CRISP-DM [11], for example, six successive and recursive phases *"Business Understanding", "Data Understanding", "Data Preparation", "Modeling", "Evaluation"* and *"Deployment"* are mentioned (cf. Fig 1). The individual project phases require different methodological support. Due to its practical relevance, this dissertation takes a closer look at the understanding phases and the modeling phase, resulting in two research paths - a conceptual and a technical path. The research paths are linked in practice, but can be investigated independently of each other. Therefore, they address different research gaps, questions, and methods outlined below.

## Research Path 1 (conceptual)

For the initial phases *"Business Understanding"* and *"Data Understanding"* it is essential to create a common understanding of the project and the specific content, especially in cooperation projects across organizational boundaries, e.g., between universities and SMEs. Various stakeholders, e.g., developers, researchers, designers, architects, managers, or board members, must create a common understanding of the objective at its essence. From a conceptual perspective, the research problem is that although several process descriptions and methodological tools exist, it is not clear which tool should be used in a specific scenario and what the advantages and disadvantages of these approaches are. Even though process models and descriptions offer a very detailed insight into the general process flow, they appear to be too expansive for initial ML projects. As there are usually few personnel and time resources available, simple tools are needed that can be applied efficiently.

### Research Gap Path 1

> *There is no sufficient support for the selection and use of tools for content clarification in initial machine learning projects. A comprehensive - yet lightweight - overview of the tasks and questions that need to be answered in the course of a machine learning project is not available.*

One way of creating an overview of the tasks and issues of the upcoming ML project and discussing the content and values with several stakeholders is to use canvases. Some examples of these are listed in the literature and in practice [12, 13]. A canvas is a conceptual design and documentation template (cf. Paper 1). However, especially for smaller organizations, such as SMEs, the selection of the right canvas poses a challenge in already complex projects. Clear and easy-to-use process descriptions and/or method manuals for ML projects that provide an overview of the key tasks and milestones of the ML project, especially when the projects are initiated, could be very beneficial for the involved project stakeholders. The following research questions are derived from this research gap.

### Research Questions Path 1

1. *Which canvas models, that address ML or AI implementation, are available, and which contents do they cover?*

2. *Where are gaps and what are potential extensions of these canvases in order to address specific challenges and needs in initial ML projects?*

3. *What does a joint working tool, i.e., canvas, need to look like that supports teams during initial and subsequent tasks of ML projects in line with standard data science processes?*

While research path 1 focuses on conceptual support during the initiation of ML projects, path 2 tackles the implementation of a concrete ML task, namely invoice recognition, which includes data preparation, modeling, and evaluation.

## Research Path 2 (technical)

In the *"Modeling"* phase of an ML project (see Figure 1), concrete algorithms are developed or improved, and specific experiments are conducted to find one or more satisfactory models. One of the enduring challenges in the field of document analysis is the development of systems capable of accurately extracting and structuring information from documents with heterogeneous layouts and non-sequential textual content. The ability to logically represent and interpret such content would enable the transformation of unstructured data into structured, machine-readable formats, significantly enhancing automation in information processing workflows. With the advent of advanced AI, particularly the rise of Large Language Models (LLMs) and their generative capacities, substantial progress has been made in modeling and understanding sequential text. LLMs demonstrate an unprecedented ability of computers to interpret and generate text and images, even in zero-shot environments. However, these models still lack the human-like abstraction ability required to quickly discern complex relationships and integrate information from unfamiliar or visually complex sources. Not least because humans can integrate past and latent knowledge that LLMs usually do not have access to. This limitation underscores why a universally applicable solution for document recognition remains elusive and why ongoing research continues to explore a variety of methodological approaches for robust content understanding in documents. Relying solely on online-provided solutions, such as recent GPT models, cumbersome training can be avoided, but control over and interpretability of the models are reduced. In addition, meaningful use often requires confidential data to be sent to external or foreign servers. This often does not comply with data governance concepts. Therefore, despite the prevailing trend toward solving recognition and comprehension tasks exclusively with LLM-based architectures, it remains a strong rationale for pursuing multi-modal and multi-model approaches and actively training rather than just applying models. These strategies leverage the complementary strengths of multiple input modalities - such as textual content, visual layout, images, and structural cues - to enhance model generalization. Techniques such as pre-training and transfer learning play a pivotal role in this context, enabling model adaptation to domain-specific tasks through fine-tuning. This not only mitigates the demand for large-scale annotated datasets and computational resources but also allows for the design of more compact models with reduced memory requirements. Consequently, such models are more suitable for deployment in resource-constrained environments, including mobile devices, thus expanding the applicability of document understanding systems in real-world scenarios.

The *"Modeling"* phase of an ML project requires appropriate software libraries and hardware capacities so that various models can be programmed and tested. Here, the research gap arises from the desire for computers to develop the same or even better skills than humans in recognizing, linking, and abstracting content on documents. In the specific use case of this dissertation, this refers to the contents of invoice documents. Due to their characteristics (see Paper 3), invoice documents represent a very useful yet challenging case for document analysis, as they are visually rich documents with diverse layouts and non-sequential text [14].

### Research Gap Path 2

> *To date, there are no generally valid models to recognize invoice contents. The research gap is that a final architecture still needs to be found, e.g., Large Language*

*Models and Graph Models have not yet been harmonized with each other.*

Invoices contain information from key items, such as addresses, dates, and invoice numbers, billing information, such as account and payment details, and line items, such as product quantities and descriptions. Although similar basic information can always be found, the actual layout is highly variable as there are no universal rules for creating invoices. Moreover, the information is embedded in different types of signals, e.g., syntactic, semantic, or spatial signals, from which the actual core information must be derived (cf. Paper 3). To process these different signals in neural networks, models are needed that can combine multi-modal inputs. The overarching research context arises from the question of which methods contribute to the improvement of invoice recognition in the context of multi-modal inputs. In order to understand the specific problem with invoice documents and explore a corresponding ML model, the following research questions are posed.

**Research Questions Path 2**

1. *What challenges exist in ML/AI pipelines in the context of invoice recognition?*
2. *What kind of information types are present on invoice documents that can influence the extraction pipeline?*
3. *What performance impact does the match between a model's language module and the dataset's language have?*
4. *How does the performance of the model change when switching from perfect (synthetic) to imperfect (real-world) data?*
5. *Can the model performance be increased by integrating an LLM in a Graph Neural Network setup?*

## 1.3 Research Methods

To answer the research questions, this dissertation takes two parallel paths to advance the data science domain and ML projects in particular. On the one hand, approaches to collaboration in ML projects are analyzed, and on the other hand, a concrete ML use case is implemented. The motivation to actively study both paths originates from the fact that practical experience has shown that both paths are mutually dependent in terms of actual implementation. Without careful project planning and an understanding of the data and the value of the solution, i.e., the conceptual side, the appropriate models cannot be explored efficiently. Without the programming, testing, and implementation of actual ML models and the curation of the corresponding datasets (technical side), any planning efforts are pointless. Only by returning the ML system to a business problem does the solution become truly valuable. Consequently, the paths are interconnected in practice (indicated by the dashed line in Fig. 2), although they could be investigated independently of each other.

Path 1 describes the examination of machine learning projects from a conceptual perspective, i.e., for better collaboration and prioritization. Path 2 is the implementation

of a specific use case from the field of document analysis using Graph Neural Networks, thus a technology-oriented perspective of machine learning. The technical path follows general trends in AI research by utilizing various design paradigms, namely, 1. fine-tuning of pre-trained models [15], 2. designing multi-modal inputs [16, 17], 3. integrating (large) language models [18, 19], and 4. utilizing synthetic data [20].

The conceptual perspective helps structure projects and prioritize tasks. The technical perspective, i.e., researching algorithms and models, and working with (big) data, is the essence of data science projects. Eventually, both research paths support each other in conducting ML projects that deliver valuable end-to-end products. In order to contribute to the scientific body of knowledge, both paths must be researched and followed. Since both research paths are highly related to practice, the basic research approaches used in this dissertation are design science research and case study methodologies to create valuable artifacts. The relationship between research paths and research approaches is illustrated in Figure 2.

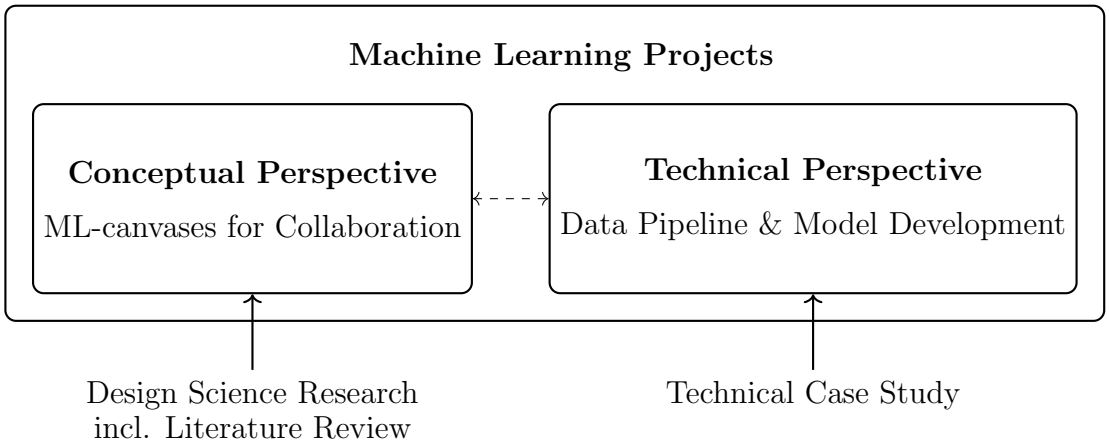


Figure 2: Research approach - Two research perspectives to facilitate ML projects

**Design science research** (DSR) is a methodology primarily used in information systems and computing disciplines to create and evaluate artifacts designed to solve real-world problems. Artifacts may include models, methods, constructs, or instantiations [21]. The key objective of DSR is not only to build innovative solutions but also to contribute to knowledge by rigorously justifying and evaluating these artifacts. The three core cycles of DSR include a relevance cycle, which connects the research to the environment and practical problems, a rigor cycle, which draws from the existing knowledge base to inform the design, and a design cycle, which involves iterative building and evaluation of the artifact [22]. Gregor and Hevner (2013) [23] later expanded on the theory-building aspect of DSR, emphasizing the importance of purposeful artifact design that contributes to both practice and theory. DSR is particularly well-suited for this dissertation, where one goal is to develop and evaluate new ML project guidelines. Its structured framework helps ensure that the proposed solutions are grounded in a real-world problem (relevance). Existing theoretical and technical knowledge informs the artifact (rigor). And the artifact is evaluated through systematic methods (design). At the beginning of this study's DSR, a comprehensive review of the related literature is included.

A **technical case study** is a research approach that investigates a specific technical problem or system in detail, focusing on computational processes, algorithms, and per-

formance in a realistic setting. Unlike general case studies that emphasize organizational or social context, technical case studies are rooted in a systematic, technical exploration of a well-defined engineering or computational task [24, 25]. This methodology is often used in fields like software engineering, data science, and artificial intelligence to explore the behavior of algorithms on real-world data, the challenges of implementation, performance bottlenecks, and trade-offs in design [26, 27, 28]. It typically involves detailed experimentation, architectural decisions, evaluation metrics, error analysis, and iterative refinements - all anchored in a real use case [29]. Therefore, a technical case study is particularly well-suited to this dissertation because it enables deep technical insight into how and why an ML model performs in a given setting. It also provides context-aware evaluation, linking performance with domain-specific challenges, and supports documenting iterative development, which is central to ML research [30].

## 1.4 Project Context

In order to design valuable and sustainable tools for ML projects they have to be derived from and tested in practice. For the application of concrete ML models and the evaluation of design artifacts, an industry partner from the *'Digital Entrepreneurship'* project at Leuphana University was available during the course of this dissertation ($2020-2023$). The *'Digital Entrepreneurship'* project was funded by the European Union for the regional development of digitization competence, including research on the handling and implementation of ML/AI. The project involved various cooperating SMEs and researchers in different fields of research, e.g., digital marketing, digital strategy, and digital transformation, as well as machine learning.

One of the SME cooperation partners in the project was planning a process automation system for tax consulting and invoice processing. A major part of this automation included the task of invoice recognition. Specifically, the question was how exactly machine learning could be used to obtain quantitatively and qualitatively better output data in the process of key value and line item extraction from invoices, and how to transition from a rule-based to an ML-based recognition system. This question lent itself very well to the scope of this thesis. The cooperation partner not only provided real-world datasets and GPU computing power but also offered the opportunity to test conceptual artifacts such as the ML-canvas in practice. In the project initialization phase, the ML-canvas from Dorard [12] (cf. Appendix B) was filled out in several meetings and feedback loops. This created a good picture of the problem statement, the data, and the expected output. The task was to find and implement a machine learning algorithm, that would enhance and eventually replace an existing rule-based algorithm for information recognition for German invoices and receipts. The key classes in this classification task included, general classes, such as *'Rechnungsnummer' (invoice number), 'Rechnungsdatum' (date)*, contact classes, such as *'Adressen' (addresses), 'Kontaktdaten' (contacts), 'Referenz' (reference)*, numeric classes, such as *'Beträge' (totals)*, or special classes like *'IBAN' (international bank account number)*, and *'Handelsregisternummer' (register number)* (cf. Paper 4 Table 4.2). Although there are different solutions for mobile banking apps to scan and recognize invoices already on the market, so-called *scan2bank* applications, these solutions are usually limited to general classes and classes with a fixed structure. The goal in the project was not to build a consumer app, but to find a solution that processes thousands

of invoice documents and eventually provides structured data.

The application of a machine learning canvas helped to understand the nature and complexity of the classification task. It was used to clarify the data, process, collaboration, and potential models. This reflects the indicated connection between the ML canvas concepts and the machine learning use case (see Figure 2). At the beginning of the project, simple model pipelines were implemented, e.g., decision trees and random forests were used to create an initial benchmark against the existing rule-based system. As the class distribution was highly unbalanced, class weights were introduced to account for minority classes, e.g., the total sum of the invoice is usually just one number compared to all other words/tokens in the document. As the project progressed and the ML-canvas was regularly updated, it was decided to use more complex models and feature extraction methods. Different approaches from academia and practice were tested, such as *AttendCopyParse* [31], *InvoiceNet*[1], and *Chargrid* [32]. Finally, due to the nature of the data, Graph Neural Networks could be considered, which fit into the technical scope of this thesis.

[1] https://github.com/naiveHobo/InvoiceNet

# Chapter 2

# Related Work

To better understand and contextualize the research papers, their underlying questions, and contributions, the following sections briefly introduce the three main subject areas relevant to this thesis - ML-canvases, invoice recognition, and Graph Attention Networks.

## 2.1 Machine Learning Canvas

The development and deployment of machine learning systems require structured methodologies to ensure efficiency, reproducibility, and alignment with business objectives. Machine learning canvases have emerged as strategic tools that provide a structured blueprint for organizing ML projects. These canvases offer a systematic way to capture essential aspects of an ML system, such as problem definition, data sources, model selection, evaluation metrics, and deployment strategies. They foster communication and collaboration between technical teams and business stakeholders so that the success of the ML project adds value to the organization. ML-canvases can best be understood by recalling the previously developed *Business Model Canvases*, which are intended to clearly and quickly identify and demonstrate the benefits of a business idea [33].

A typical ML-canvas consists of multiple interconnected components, each addressing a critical aspect of the project lifecycle (cf. Appendix B). The background section outlines the problem domain, detailing the significance of the problem and the motivation behind the ML solution. This is followed by the objective definition, which specifies the success criteria and expected outcomes of the model. The data component plays a crucial role in determining the quality of the ML system, encompassing data collection, preprocessing, augmentation, and labeling strategies. The modeling section explores different algorithmic approaches, feature engineering techniques, and hyperparameter tuning strategies necessary for optimizing model performance. The evaluation criteria include performance metrics such as accuracy, precision, recall, F1-score, and AUC-ROC, ensuring that the model meets predefined standards. Finally, the deployment strategy covers the technical and infrastructural requirements for integrating the ML model into a production environment, including aspects such as model versioning, monitoring, and real-time inference [12].

The concept of structured ML-canvases has been discussed in recent literature. Different titles of these artifacts describe similar constructs. For example, Agrawal et al. [34] introduced the *AI Canvas*, which emphasizes economic trade-offs in AI applications, particularly focusing on prediction capabilities and cost-benefit analyses. Similarly, Shteingart

et al. [35] proposed a *Machine Learning Prescriptive Canvas*, which focuses on ensuring that ML models produce actionable decisions rather than mere predictions. Furthermore, Kerzel et al.[13] presented an *Enterprise AI Canvas*, which integrates ML methodologies within enterprise workflows to facilitate large-scale AI adoption. These frameworks highlight the importance of structured planning in ML projects, reducing risks and improving overall system efficiency.

The advantages of employing ML-canvases are multifaceted. First, they enhance collaboration between interdisciplinary teams by providing a shared conceptual framework. Second, they promote strategic alignment by ensuring that ML initiatives contribute to broader business objectives. Third, they enable risk mitigation by identifying potential challenges at an early stage, allowing for proactive solutions. Lastly, they facilitate resource optimization by clearly defining data requirements, modeling strategies, and computational needs. Given these benefits, the adoption of ML-canvases is becoming increasingly widespread across academia and industry, serving as a fundamental tool in the planning and execution of ML-projects.

## 2.2 Invoice Recognition

Automated invoice recognition is a critical task in intelligent document processing systems, significantly enhancing efficiency in financial operations, accounting, and enterprise resource planning (ERP) systems [36]. *"Invoices contain always rather similar information, fostered by legal requirements for information items on invoices. However, the information items are distributed according to all different layout styles"* [37]. This variance in layouts and languages is a challenge to standardizing invoice recognition.

The meaning of the raw invoice data, such as words, numbers, or abbreviations, often only emerges from the interaction of different signals. These signals may be of segmental, syntactic, semantic, spatial, external, graphical, or logical nature [38], e.g., letterhead conventions, sums, tables, percentages, or font styles.

Traditional invoice processing relied heavily on manual data entry, which was not only time-consuming, but also prone to errors and resulting in average *"processing costs of about 9 Euro"* [37]. The advent of machine learning and deep learning methodologies has revolutionized invoice recognition, enabling high accuracy in extracting structured information from diverse invoice formats. Invoices - visually rich documents - are transformed into machine-readable formats, to gain structured information [39, 14, 31].

Earlier approaches to invoice recognition were predominantly template- or rule-based, leveraging Optical Character Recognition (OCR) combined with predefined layouts to extract key information such as invoice numbers, dates, amounts, and supplier details [37, 40, 41, 42]. These methods, however, suffered from several limitations. Since invoices vary widely in format, template-based systems require constant updates whenever a new layout is encountered, making them inflexible and difficult to scale. Additionally, OCR engines struggled with noisy, low-quality, or skewed scanned documents, leading to significant accuracy degradation [43]. Applying simple NLP techniques on the OCR results is also sub-optimal, as the information is usually not in a clear sequential order [44].

With the introduction of deep learning, modern invoice recognition systems have adopted end-to-end trainable models that combine Convolutional Neural Networks (CNNs) for image-based feature extraction [45] with Recurrent Neural Networks (RNNs) [46] or

transformers for sequence modeling [47, 48]. Different grid-based approaches have also been applied to invoice recognition [32, 49, 50]. Furthermore, Named Entity Recognition (NER) models have been widely employed to classify and extract key-value pairs from textual content [51]. State-of-the-art transformer architectures, such as BERT (Bidirectional Encoder Representations from Transformers) and its domain-specific variant LayoutLM, have demonstrated superior performance in extracting structured information from unstructured invoices [47, 52].

Another critical advancement in invoice recognition is graph-based learning, which treats invoice documents as structured graphs rather than plain text sequences [53, 14, 39]. In this approach, nodes represent text blocks, and edges encode spatial relationships between them. Graph Neural Networks (GNNs) have shown promising results in extracting hierarchical invoice structures. These models enable robust generalization across diverse invoice layouts, significantly improving recognition accuracy.

The adoption of AI/ML-driven invoice recognition systems has led to substantial benefits, including improved processing speed, reduced manual intervention, and enhanced accuracy in financial workflows. Future research in this domain is expected to focus on improving cross-lingual adaptation, robustness to document noise, and the integration of self-supervised learning techniques for data-efficient model training.

## 2.3 Graph Attention Networks

Graph-based machine learning has gained significant traction in recent years due to its ability to model relationships between entities in non-sequential data structures. Graph Neural Networks (GNNs) extend traditional deep learning techniques to operate on graph-structured data, making them highly effective for tasks such as node classification, link prediction, and graph clustering. Among various GNN architectures, Graph Attention Networks (GATs) have emerged as a powerful model for learning graph representations by dynamically weighting node influences based on attention mechanisms [54].

Traditional GNNs, such as Graph Convolutional Networks (GCNs), aggregate information from neighboring nodes using fixed weighting schemes, which often fail to capture the varying importance of different nodes. In contrast, GATs employ a self-attention mechanism, allowing the model to learn attention coefficients that determine the relative importance of neighboring nodes in a given graph. This mechanism enhances the model's ability to capture complex dependencies, particularly in heterogeneous graphs where node relationships vary significantly.

The architecture of GATs consists of multiple attention heads, which independently compute attention weights for each node's neighbors. The final node representation is obtained by aggregating information from these attention heads, allowing for better feature extraction and robustness to noise. This multi-head attention mechanism improves generalization, making GATs particularly effective for tasks such as social network analysis, fraud detection, and recommendation systems [55]. In practical applications, GATs have been widely utilized in various domains. In natural language processing (NLP), they enhance text classification and knowledge graph completion tasks by modeling word dependencies in a structured format.

The main advantages of GATs include their ability to dynamically assign importance to neighboring nodes, improved scalability compared to traditional GNNs, and enhanced

expressiveness in capturing hierarchical graph structures. However, challenges remain in optimizing attention computation efficiency, reducing overfitting in highly dense graphs, and extending GATs to large-scale real-world datasets. Ongoing research in this area is focused on improving scalability through sparse attention mechanisms and integrating GATs with other deep learning architectures for hybrid learning paradigms (cf. Paper 5).

# Chapter 3

# Publications and Contributions

The following section presents the publications relevant to this dissertation and their contributions (see Table 1). In addition, works are presented that were published during the course of the dissertation project but are not included in the direct evaluation of this work. The five publications included in the main body of this dissertation are peer-reviewed papers that were presented and discussed at national and international IS conferences. The corresponding papers were published in the respective conference proceedings. The proceedings of AMCIS, WI, and INFORMATIK are listed in the VHB journal ranking for Information Systems[1]. ICDAR is a leading international conference in the document analysis domain.

## 3.1 Publications

### Publications in the Conceptual Research Path (1)

#### Paper 1: A systematic Literature Review of Machine Learning Canvases

Paper 1 provides a literature review of machine learning canvas approaches. 18 canvas artifacts (Table 1.2) are compared and categorized into 4 categories (Table 1.1). The comparison leads to a set of guiding questions, summarized in Table 1.3, that represent the main contribution of this paper. The question catalog allows project stakeholders in an ML-project to identify the ideal canvas and the right questions for their specific use case. This lightweight approach fosters collaboration and understanding between stakeholders to implement solutions faster.

#### Paper 2: Developing an Ontology for Data Science Projects to facilitate the Design Process of a Canvas

In paper 2 an ontology (see Figure 2.2) of data science projects is proposed. The paper presents the derivation of the design requirements and the development of the underlying ontology. The contribution of this research project is to combine both an integrative view of the overall project and an appropriate level of detail in the individual sections, thus addressing the dichotomy between holistic and compact representations. The mapped ontology helps to see the nuanced sections of a data science project in an easy overview.

[1] https://www.vhbonline.org/fileadmin/vhb/Services/vhb-rating/WI/

Table 1: List of publications

| No. | Publication | Ref. |
|---|---|---|
| 1 | **A systematic Literature Review of Machine Learning Canvases**<br>Thiée, Lukas-Walter (2021). INFORMATIK 2021<br>05.10.2021 https://doi.org/10.18420/informatik2021-101 | [56]<br>Paper 1 |
| 2 | **Developing an Ontology for Data Science Projects to facilitate the Design Process of a Canvas**<br>Thiée, Lukas-Walter (2022). Wirtschaftsinformatik 2022<br>17.01.2022 https://aisel.aisnet.org/wi2022/ai/ai/13/ | [57]<br>Paper 2 |
| 3[1] | **Extraction of Information from Invoices – Challenges in the Extraction Pipeline**<br>Thiée, Lukas-Walter; Krieger, Felix; Funk, Burkhardt (2023). INFORMATIK 2023<br>29.11.2023 https://doi.org/10.18420/inf2023 _180 | [38]<br>Paper 3 |
| 4 | **Ablation study of a multi-modal GAT network on perfect synthetic and real-world data to investigate the influence of Language Models in invoice recognition**<br>Thiée, Lukas-Walter (2024). ICDAR 2024 Workshops<br>11.09.2024 https://doi.org/10.1007/978-3-031-70642-4 _13 | [58]<br>Paper 4 |
| 5[1] | **Enhancing Invoice Recognition with LLM Embeddings in GAT Networks**<br>Thiée, Lukas-Walter and Funk, Burkhardt (2025). AMCIS 2025<br>14.08.2025 https://aisel.aisnet.org/amcis2025/sig_svc/sig_svc/6/ | [59]<br>Paper 5 |
| 6[1,2] | **DASC-PM v1.1 "Ein Vorgehensmodell für Data-Science-Projekte"**<br>Schulz et al. (2022). Universitäts- und Landesbibliothek Sachsen-Anhalt<br>30.03.2022 http://dx.doi.org/10.25673/85296 | [60] |
| 7[1,2] | **Parking Space Management Through Deep Learning – An Approach for Automated, Low-Cost and Scalable Real-Time Detection of Parking Space Occupancy**<br>Schulte et al. (2021). Innovation Through Information Systems (WI 2021 Proceedings)<br>16.10.2021 https://doi.org/10.1007/978-3-030-86797-3 _42 | [61] |
| 8[1,2] | **'Spread the app, not the virus' – An extensive SEM-approach to understand pandemic tracing app usage in Germany**<br>Thiée et al. (2021). ECIS 2021 Proceedings<br>11.05.2021 https://aisel.aisnet.org/ecis2021_rp/123/ | [62] |

[1] Co-authorship.

[2] Additional contribution, not in direct scope of this thesis.

## Publications in the Technical Research Path (2)

### Paper 3: Extraction of Information from Invoices – Challenges in the Extraction Pipeline

Paper 3 shows the characteristics of invoice documents and discusses specific challenges in invoice recognition pipelines. The contribution lies in the systematic detection and cataloging of information types (Table 3.2) and pipeline challenges (Table 3.3). The proposed framework is the result of an initial design cycle. It is a systematic catalog of challenges and information types encountered in invoice recognition, based on a comprehensive review of literature and practice approaches. With the help of the framework, model selection and data preparation can be better structured, and it improves the comparability between different invoice recognition pipelines. The framework is used to initialize a comparative study between two GNN approaches for invoice recognition.

### Paper 4: Ablation Study of a multi-modal GAT Network on perfect synthetic and real-world Data to investigate the influence of Language Models in invoice recognition

Paper 4 presents an ablation study on an existing state-of-the-art pre-trained multi-modal Graph Attention Network (GAT). Based on the knowledge gained from the papers 1, 2 and 3 a technical use case is implemented. Therein, two kinds of modifications are performed to understand the sensitivity of a classification task (see Table 4.6) by exchanging the language module and applying both the original and modified network on a perfect synthetic (Inv3D, cf. [63]) and an imperfect real-world dataset. The results of the study show the importance of language modules for semantic embeddings in multi-modal invoice recognition and illustrate the impact of data annotation quality. The contribution is twofold. Firstly, the performance of a multi-modal state-of-the-art approach is analyzed and confirmed. The impact of matching the language of the semantic module to the dataset language is demonstrated. Secondly, the approach is extended with a German language model.

### Paper 5: Enhancing Invoice Recognition with LLM Embeddings in GAT Networks

Paper 5 extends the approach of Paper 4 by integrating LLM embeddings into GATs. The multiple attention mechanism plays a critical role in capturing both semantic and relational information, leading to robust performance across diverse extraction classes. Experiments show that the model increases classification results on the selected datasets, with respect to their benchmarks. The approach contributes to the improvement of invoice recognition through the harmonization of language and layout embeddings and offers a model that can be transferred to general supervised document analysis. The findings highlight the potential of advanced embedding techniques to overcome challenges inherent to document analysis, particularly in contexts where traditional approaches may struggle to generalize.

## 3.2 Contributions

The individual research papers, but also the thesis as a whole, make various contributions. In both the conceptual and the technical research streams methodological contributions play a central role in enabling the formulation and resolution of the research questions. Figure 3 provides an overview of the contributions of this dissertation along both the conceptual and technical dimensions, illustrating how each contribution - and the corresponding publications - advance the development and execution of machine learning projects. Although they can be investigated independently of each other, the two research streams are linked to the extent that the conceptual perspective in this thesis sharpens the understanding of the machine learning problem in the given project and thus allows for a differentiated model selection in the technical path (indicated by the dashed line in Figure 3). The main contributions of this thesis are a catalog of questions (cf. Table 1.3) for ML-canvas applications and a GAT+LLM model for invoice recognition tasks (cf. Appendix D.1).

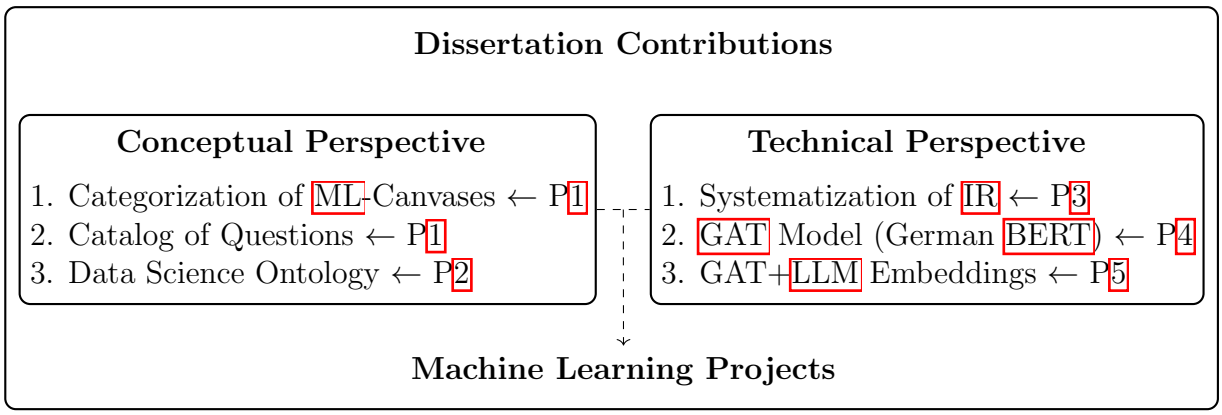


Figure 3: Contributions in the two research paths and corresponding papers

To address the identified research gap concerning the lack of structured support - for SMEs - in the early phases of data science projects, this study presents a systematic categorization of tools designed to enhance decision-making and collaboration. The table provides a comprehensive listing and categorization of existing ML-canvases relevant to data science workflows (cf. Table 1.1). By offering a curated catalog of guiding questions, it facilitates the selection of an appropriate canvas based on the project context and stakeholder needs. These tools support project teams in establishing a shared understanding of objectives, assumptions, and roles, thereby reducing ambiguities and improving communication.

Complementing this, the derivation of a data science ontology contributes a formalized representation of project phases and their interdependencies. The ontology (cf. Appendix C.1) not only structures the lifecycle of data science initiatives but also clarifies the logical sequence and prerequisites of each phase. This dual contribution - pragmatic support through canvases and theoretical structuring via ontology - closes a key methodological gap by linking practical guidance with conceptual clarity, thereby fostering more efficient and transparent project planning in ML projects.

To address the second research gap concerning the limitations of current document analysis techniques in invoice recognition, this study investigates the task-specific chal-

lenges and information types inherent in such use cases. A detailed analysis of invoice documents reveals complexities such as diverse text layouts, varied semantic content, and the integration of numerical and categorical information, all of which pose challenges to traditional OCR- and NLP-based methods. There has not yet been such a detailed examination of invoice documents and their specifics in relation to information extraction using machine learning. The results help users and researchers to incrementally improve the entire ML extraction pipeline as well as individual sections of it.

This thesis further contributes empirically by applying and evaluating different transformer-based language models, including BERT and LLaMA2, on both real-world and synthetically generated datasets for the task of token classification in invoice recognition. Results show consistent improvements over existing benchmarks [39, 38], with gains of 3% and 4% in classification metrics, i.e., F1-score, highlighting the effectiveness of LLMs in extracting semantically meaningful representations under diverse data conditions.

A key innovation of this study lies in the application of LLM-based semantic embeddings within a multi-modal, pre-trained Graph Attention Network (GAT). This novel integration enables the model to jointly capture semantic content and document structure through multiple attention layers, leading to enhanced performance across heterogeneous data classes. By leveraging graph-based learning in conjunction with LLM-derived embeddings, the approach bridges the gap between unstructured textual content and visual layout representations, an essential step forward for tasks requiring multi-modal comprehension.

Beyond invoice recognition, the proposed architecture demonstrates strong transferability to other document analysis tasks that involve the recognition of both textual and numerical elements within a spatial context. This suggests broader applicability in domains such as customer data processing, compliance auditing, and financial document parsing. While the results affirm the promise of LLM+GAT integration for improved document understanding, the study also underscores the need for continued research to refine the model architecture, ensure scalability, and address domain-specific limitations.

In order to answer the aforementioned research questions, several research papers have been written and published over the past years. Collectively and individually, these papers contribute to the body of knowledge in the realm of machine learning and especially invoice recognition. Paper 1 and 2 deal with the question of how ML projects can be supported by corresponding canvas approaches to facilitate the entry into a project and the result-oriented collaboration between project stakeholders, e.g., researchers and practitioners, particularly in small organizations. Papers 3, 4, and 5 deal with the application of machine learning in the context of invoice recognition. Challenges and specific technical solutions are presented. GAT models are applied to English and German datasets and different semantic embeddings are compared. Table 1 lists the publications of the main body of this thesis as well as further contributions to the data science domain published during the time of this dissertation project.

## 3.3 Additional Contributions

In addition to the articles directly relevant to this thesis, further research papers were written and published in collaboration with other authors over the past year. As these articles complement this dissertation project, they are briefly listed and described below.

## DASC-PM v1.1 "Ein Vorgehensmodell für Data-Science-Projekte"

The DASC-PM [60], which translates to "A Process Model for Data-Science-Projects", is a scientifically grounded framework designed to support the structured execution of data science projects. Developed through a collaboration between academia and industry, this process model provides a comprehensive and phase-oriented approach that reflects the interdisciplinary nature of data science. It consists of four main phases: *Project Order, Data Provisioning, Analysis, and Deployment and Application*. The process begins with the *Project Order phase*, where the objectives, scope, and constraints of the project are clearly defined. This foundational stage ensures alignment among stakeholders and sets the direction for all subsequent steps. In the *Data Provisioning phase*, relevant data is identified, collected, cleaned, transformed, and integrated, preparing it for meaningful analysis. The third phase, *Analysis*, focuses on selecting appropriate statistical and machine learning methods to generate insights or predictions from the data, ensuring scientific rigor and relevance to the project's goals. Finally, in the *Deployment and Application phase*, the results are transferred into operational use. This includes implementation, monitoring, and performance evaluation to ensure that the developed solutions provide tangible value. A notable feature of DASC-PM v1.1 is its explicit consideration of scientific methodology, domain-specific requirements, IT infrastructure, and impact assessment throughout all phases. It also includes detailed role and competence profiles, helping organizations to allocate responsibilities effectively and foster interdisciplinary collaboration. The process model is designed to be both robust and flexible, making it suitable for academic research as well as practical industry applications.

The author's contribution to DASC-PM v1.1 is limited to participation in a survey on data science processes, group interview sessions, and selective content review.

## Parking Space Management Through deep learning – An Approach for Automated, Low-Cost and Scalable Real-Time Detection of Parking Space Occupancy

In this work, Schulte et al. [61] present a novel solution to urban parking challenges by leveraging deep learning techniques. The authors develop a camera-based system designed for real-time detection of parking space occupancy, emphasizing affordability and scalability. Utilizing a deep neural network implemented with *TensorFlow*, specifically employing YOLOv4 and *DeepSORT* algorithms, the system effectively tracks vehicles in parking lots. An integral component of their approach is a web interface that visualizes parking capacity and provides additional insights, such as average parking durations. This research contributes to the field of smart parking management by offering a cost-effective, scalable, and real-time detection system suitable for public, open-air parking facilities.

The author's contribution to Schulte et al. [61] is limited to designing the structure, as well as writing and revising individual sections of the manuscript.

## ‘Spread the app, not the virus’ – An extensive SEM-approach to understand pandemic tracing app usage in Germany

The complete paper is appended to this work in Appendix A. In this study, Thiée et al. [62] conduct a comprehensive analysis to understand the factors influencing the adoption and usage of pandemic tracing apps in Germany. Recognizing the pivotal role of digital contact tracing in mitigating the spread of infectious diseases, the authors employ Structural Equation Modeling (SEM) to dissect the interplay between user attitudes, perceived risks, trust in technology, and demographic variables. Their findings reveal that while *technological efficacy* and *data privacy concerns* significantly impact user acceptance, *social influence* and *perceived benefits* also play crucial roles. The study underscores the necessity for transparent communication strategies and user-centric design to enhance public trust and encourage widespread adoption of tracing apps. By providing empirical insights into user behavior, Thiée et al. [62] contribute valuable knowledge to the development of effective digital interventions during health crises.

The author is the main author of this publication and provides the majority of the content in collaboration with co-author Hannes M. Petrowsky, particularly in the areas of research design, data collection, statistical analysis, manuscript, and presentation of the results.

# Chapter 4

# Outlook

The papers and contributions of this dissertation thesis, listed in the previous chapter, contribute to the continuation of the discourse in the DS domain and to the expansion of the body of knowledge in ML projects. Certainly, they are to be viewed in the context of their field of application, that is, essentially within the associated research project *'Digital Entrepreneurship'* and the two subject areas *ML-canvas* and *invoice recognition*. It is particularly because of this close integration into practice that the limitations of this thesis need to be highlighted at the beginning of this outlook chapter. The boundaries of the results in relation to the datasets applied must be taken into account. Furthermore, based on contributions and limitations, directions for future research are outlined in this chapter, and the research field of document recognition with LLMs is critically reflected upon. A summary and a conclusion complete this preamble.

## 4.1 Limitations

Despite the promising results and novel contributions of this work, particularly the ML-canvas categorization and question catalog as well as the implementation of a GAT architecture enhanced with LLM embeddings, several limitations must be acknowledged. These limitations present opportunities for refinement and further exploration. While this work supports machine learning projects on both a conceptual and technical level, its artifacts are only used in their specific application context, so an evaluation beyond this context is still pending.

### Methodological Constraints

One of the most significant limitations lies in the evaluation methodology. While the ML-canvas catalog and ontology were developed with conceptual rigor, a comprehensive empirical evaluation, particularly with a larger group of domain experts across varied industries, was beyond the scope of this thesis. The lack of a formal evaluation restricts the ability to generalize the utility and usability of the question catalog.

Similarly, the ontology that underpins parts of the system architecture was not systematically assessed for completeness. Without such validation, the robustness of the knowledge representation remains a potential concern. Nevertheless, both artifacts represent valuable tools in the context of initial machine learning projects in general and for the

*'Digital Entrepreneurship'* project in particular, leaving the impact of the contributions undiminished.

### Dataset and Case Study Constraints

The case studies used for demonstrating the GAT approach and the integration of LLM embeddings involved a dataset of limited size and domain scope. Although it served its illustrative and programming purpose, this raises questions regarding the scalability and adaptability of the proposed model across larger and more heterogeneous datasets. Furthermore, the model's performance may vary when applied to domains with different graph structures or data distributions.

Another core limitation involves the generalizability of both the ML-canvas question catalog and the GAT+LLM model. The solutions were tailored to their specific contexts and datasets, which means their applicability and performance in other use cases or organizational environments might be smaller - an inherent problem with case study approaches. Additionally, due to potential dependencies on specific LLM versions and frameworks, reproducibility may pose a challenge, especially as the AI landscape continues to evolve rapidly.

## 4.2 Future Research

Both the contributions and the limitations of this study lay the foundation for further research endeavors. Future research can move in both directions, the conceptual area of machine learning projects, as well as the actual application and programming of multi-modal document analysis systems.

### Conceptual, ML-canvas related Future Research

An opportunity to further develop the conceptual findings of this study is to integrate the question catalog into standard process descriptions, such as CRISP-DM or DASC-PM. Such integration could significantly enhance transparency, reproducibility, and operational efficiency across various stages of ML project execution. In this regard, the question catalog may function as a guiding tool for practitioners, aligning stakeholder expectations and supporting structured decision-making.

To validate and refine the organizational framework proposed in this thesis, future research should empirically assess the ML-canvas through structured interviews, participatory workshops, or user studies involving data scientists, ML engineers, and project stakeholders. Such mixed-method evaluations could yield both qualitative and quantitative insights into the practical utility, comprehensiveness, and adaptability of the question catalog. In parallel, ontology refinement and extension should be conducted. This may involve formal evaluation techniques and LLM-driven methods for automated or semi-automated knowledge extraction, thereby enabling scalable and context-aware ontology development. In particular, the technical availability of models, e.g., on mobile devices, LLM pipelines, and their usability in relation to data protection guidelines could be topics for expansion.

From an applied perspective, embedding both the ML-canvas and the GAT model within end-to-end data science workflows, such as AutoML platforms or MLOps pipelines,

represents an important next step. Such integration would provide critical insights into their real-world robustness, adaptability, and efficiency in further production settings.

### Technical, invoice recognition related Future Research

As introduced in the methods section, the research paradigms - fine-tuning of pre-trained models, designing multi-modal inputs, integrating (large) language models, and utilizing synthetic data augmentation - will continue to prevail in the technical path in the future.

In the field of invoice recognition, future work may focus on the development and inclusion of more extensive datasets, with an emphasis on expanding the range of labeled entity classes. This would not only improve model generalizability but also extend its applicability to a broader spectrum of document types encountered in practical financial workflows. Integrating more classes in the recognition pipeline can lead to a comprehensive recognition of key values and line items, eventually guiding to class-free recognition of arbitrary classes.

Applying the model to well-established benchmark datasets, such as DocILE [64], would provide a robust evaluation framework and facilitate meaningful comparisons with existing methods. The scalability of the proposed GAT architecture must also be rigorously tested through experimentation with larger, multi-domain datasets. Understanding how the model's performance scales with data complexity will be essential to gauging its practical viability. Research into generalizability frameworks, particularly meta-learning techniques and domain adaptation strategies, may offer solutions for enhancing cross-domain performance without necessitating exhaustive retraining procedures.

Moreover, from a model-centric perspective, the proposed Graph Attention Model offers several potential research trajectories. One of these involves increasing the dimensionality of the LLM embeddings, for instance, scaling up to 4096 dimensions, to enhance the semantic granularity and contextual fidelity of the representations. This augmentation could yield measurable improvements in classification and information extraction performance.

Another technically promising avenue involves the exploration of graph node fusion techniques. These methods aim to integrate multiple feature types - textual, visual, and structural - at the node level, potentially improving the model's capacity to capture intricate interdependencies and boosting overall classification performance.

Furthermore, the growing field of multi-modal machine learning presents an essential opportunity for comparative research. A structured evaluation contrasting the current approach with state-of-the-art multi-modal transformer architectures, such as *LayoutLMv2* [52] and *DocFormer* [65], may provide valuable insights. These models incorporate textual content, visual features, and spatial layout information into a unified representation, substantially improving performance on document understanding tasks. This trend toward modality integration underscores the importance of benchmarking against the latest advancements, including emerging architectures such as *DocLLM* [66] and *LayoutLLM* [67], which complement the forefront of layout-aware multi-modal document processing. A systematic comparison with MLLMs utilizing pre-trained encoders could further elucidate the strengths and weaknesses of graph-based approaches relative to transformer-based multi-modal systems. Finally, leveraging ensemble techniques might as well be a promising future directive, i.e., applying multi-modal-multi-model approaches. These approaches depend heavily on the available computing capacity and the amount of usable

training data.

In summary, by addressing these interconnected research challenges, future work can significantly advance the theoretical underpinnings and practical impact of GAT+LLM-based methods within the broader context of intelligent document processing. The convergence of conceptual best practices and cutting-edge model architectures holds considerable promise for shaping the next generation of robust, scalable, and multi-modal document understanding systems.

## 4.3 Reflection

An emerging frontier in AI research is the rise of *multi-modal Large Language Models* (MLLMs), which combine text and vision understanding within a single unified architecture. These models, capable of interpreting images, documents, charts, and even handwriting alongside natural language, are opening new possibilities in areas such as document analysis and information extraction from scanned or semi-structured sources. For example, MLLMs can be leveraged to automatically extract key fields from invoices, detect layout structures in forms, or interpret diagrams and annotate them with semantic information. These capabilities are particularly valuable in business domains where large volumes of documents are processed manually, leading to potential automation and efficiency gains. Looking ahead, both ends of the AI pipeline, namely, data inputs and model outputs, are expected to undergo substantial advancements. On the input side, the availability and scale of datasets will continue to increase, enabling the capture of greater real-world variability and thereby facilitating the training of more robust and generalizable models. Concurrently, model performance is anticipated to improve even more, driven by advances in computational power as well as increases in model complexity and architectural sophistication. Emerging approaches, particularly those employing multi-modal and multi-model strategies, referred to as ensemble learning, will play a pivotal role. These methods integrate the strengths of diverse deep learning architectures, thereby enhancing performance, robustness, and generalizability across a wider range of tasks and domains. Within the context of this thesis, such technological developments hold particular promise for the comprehensive extraction of information from invoices. Regardless of their language, layout, or modality, these enhanced AI systems will increasingly be capable of identifying and processing the full spectrum of relevant information embedded within visually rich documents.

Despite recent advances in (multi-modal) LLMs and generative approaches, it remains valuable to pursue invoice recognition using modular, non-generative methods. This strategy offers greater control over discrete pipeline components - such as OCR, tokenization, feature extraction, representation of input signals (e.g., images, geometric structures, and tables), and their fusion - enabling more targeted optimization. Furthermore, this modularity enhances model explainability and facilitates fine-tuning, particularly in scenarios with limited data availability. A persistent challenge in this domain is that self-supervised approaches, such as masked language modeling, remain difficult to apply effectively due to insufficient dataset sizes for pre-training or domain adaptation.

The convergence of advanced machine learning models, graph-based architectures, and human-centered frameworks is reshaping the future of data science. As tools like the ML-canvas help project teams navigate conceptual complexity, and state-of-the-art mod-

els, e.g. transformer architectures, offer deeper knowledge integration, a central question emerges: *what role should increasingly powerful AI models play in the research process itself?* Large Language Models such as the modern GPT architectures are no longer simply tools for automating text generation. They are becoming co-creators, agents, and even research aides that support literature reviews, code generation, hypothesis formation, and drafting of academic texts. This raises an intriguing question: *LLM – is this science?* On the one hand, LLMs extend human capability. They can process vast amounts of information, suggest non-obvious connections, and accelerate iteration in both research and application. Used responsibly, they can democratize access to scientific knowledge and reduce barriers to interdisciplinary collaboration. On the other hand, their use challenges traditional definitions of scientific authorship and originality. If a model generates part of the analysis, how is intellectual credit assigned? If the model's inner workings are opaque or based on untraceable training data, how do we ensure scientific transparency and reproducibility? Moreover, LLMs are trained on existing knowledge. They do not innovate in the philosophical sense but remix and interpolate. These tensions suggest that the integration of LLMs into research practice must be accompanied by clear ethical, methodological, and epistemological frameworks. Academic communities need to engage with these models not just as tools, but as actors that reshape the very process of knowledge creation.

In the context of this thesis, the use of LLM embeddings within the GAT architecture serves as a productive case study. It shows how human-curated domain knowledge and LLM-derived semantic representations can be fused without relinquishing the critical thinking, validation, and interpretation that define scientific inquiry. The future of data science and research more broadly may not lie in choosing between human and machine, but in designing partnerships where each complements the other's strengths, and where the scientific method evolves to account for this collaboration.

## 4.4 Summary and Conclusion

This thesis introduced two key innovations aimed at addressing complexity in machine learning projects: the ML-canvas question catalog, a structured framework for guiding project collaboration and understanding, and a Graph Attention Network architecture enhanced with LLM embeddings, designed to integrate rich semantic knowledge into graph-based machine learning for invoice recognition.

Beginning with a discussion of the increasing complexity and interdisciplinarity in modern data science, the work outlined the motivation for tools that support both the conceptual and technical aspects of such projects. The ML-canvas catalog was proposed as a lightweight, modular guide to foster structured communication, planning, and alignment in data-driven initiatives, especially in SMEs. Complementing this, the GAT model enriched with LLM embeddings demonstrated a novel approach to knowledge representation and reasoning on graph-structured data in the context of invoice recognition. The model architecture is one of the state-of-the-art approaches and increases performance on selected datasets.

While these contributions represent meaningful steps forward, the thesis also highlighted several limitations, including the need for broader evaluation, the constraints imposed by the dataset size, and challenges related to scalability. These limitations offer

fertile ground for future research, which can further mature the concepts presented here and bring them closer to deployment.

# Bibliography


[1] Statista. *Data growth worldwide 2010-2028 — Statista*. 2025. URL: https://www.statista.com/statistics/871513/worldwide-data-created/ (visited on 03/26/2025).

[2] Sandro Shubladze. "How To Make Use Of The New Gold: Data". In: *Forbes* (2023). URL: https://www.forbes.com/councils/forbestechcouncil/2023/03/27/how-to-make-use-of-the-new-gold-data/ (visited on 03/27/2025).

[3] R. Shortland and R. Scarfe. "Digging for gold". In: *IEE Review* 41.5 (1995), pp. 213–217. ISSN: 0953-5683. DOI: 10.1049/ir:19950504. URL: https://digital-library.theiet.org/doi/10.1049/ir:19950504.

[4] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge: Cambrige University Press, 2014. ISBN: 9781107057135. DOI: 10.1017/CBO9781107298019.

[5] Wil van der Aalst. "Data Science in Action". In: *Process mining*. Ed. by Wil van der Aalst. Berlin et al.: Springer, 2016, pp. 3–23. ISBN: 978-3-662-49851-4. DOI: 10.1007/978-3-662-49851-4_1. URL: https://link.springer.com/chapter/10.1007/978-3-662-49851-4_1.

[6] Daan A Kolkman and Ruud Sneep. *Challenges to Data Science Projects with SMEs: An Analysis and Decision Support Tool*. 2019. DOI: 10.13140/RG.2.2.25212.80006.

[7] Johannes Otterbach and Thomas Wollmann. *Chameleon: A Semi-AutoML framework targeting quick and scalable development and deployment of production-ready ML systems for SMEs*. May 8, 2021. URL: http://arxiv.org/pdf/2105.03669.

[8] Andrei Paleyes, Raoul-Gabriel Urma, and Neil D. Lawrence. "Challenges in Deploying Machine Learning: A Survey of Case Studies". In: *ACM Computing Surveys* 55.6 (2023), pp. 1–29. ISSN: 0360-0300. DOI: 10.1145/3533378.

[9] Karthik Shivashankar and Antonio Martini. "Maintainability Challenges in ML: A Systematic Literature Review". In: *48th Euromicro Conference on Software Engineering and Advanced Applications*. Ed. by Gustavo M. Callico. Piscataway, NJ: IEEE, 2022, pp. 60–67. ISBN: 978-1-6654-6152-8. DOI: 10.1109/SEAA56994.2022.00018.

[10] Nadia Nahar et al. *Collaboration Challenges in Building ML-Enabled Systems: Communication, Documentation, Engineering, and Process*. Oct. 19, 2021. URL: http://arxiv.org/pdf/2110.10234.

[11] Rüdiger Wirth and Jochen Hipp. "CRISP-DM: Towards a standard process model for data mining". In: *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining* (2000). URL: https://cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf.

[12] Louis Dorard. *The Machine Learning Canvas: A handbook for innovators and visionary managers striving to design tomorrow's Machine Learning systems: DRAFT version 0.1 released on 12 February 2019*. 2019. URL: https://www.ownml.co/machine-learning-canvas.

[13] Ulrich Kerzel. "Enterprise AI Canvas Integrating Artificial Intelligence into Business". In: *Applied Artificial Intelligence* 35.1 (2021), pp. 1–12. ISSN: 0883-9514. DOI: 10.1080/08839514.2020.1826146.

[14] Xiaojing Liu et al. *Graph Convolution for Multimodal Information Extraction from Visually Rich Documents*. Mar. 27, 2019. URL: https://arxiv.org/pdf/1903.11279.

[15] Paul Ohm. "Focusing On Fine-Tuning". In: *Science and Technology Law Review* 25.2 (2024). DOI: 10.52214/stlr.v25i2.12762.

[16] Sushant Gautam. "Bridging Multimedia Modalities: Enhanced Multimodal AI Understanding and Intelligent Agents". In: *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION*. Ed. by Elisabeth André et al. New York, NY, USA: ACM, 2023, pp. 695–699. ISBN: 9798400700552. DOI: 10.1145/3577190.3614225.

[17] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. "Challenges and applications in multimodal machine learning". In: *The Handbook of Multimodal-Multisensor Interfaces: Foundations, User Modeling, and Common Modality Combinations - Volume 2*. Ed. by Sharon Oviatt et al. Association for Computing Machinery, 2018, pp. 17–48. ISBN: 9781970001716. DOI: 10.1145/3107990.3107993.

[18] Xinyi Hou, Yanjie Zhao, and Haoyu Wang. *The Next Frontier of LLM Applications: Open Ecosystems and Hardware Synergy*. Mar. 6, 2025. URL: http://arxiv.org/pdf/2503.04596.

[19] James Boyko et al. *An Interdisciplinary Outlook on Large Language Models for Scientific Research*. Nov. 3, 2023. URL: http://arxiv.org/pdf/2311.04929.

[20] Shuang Hao et al. *Synthetic Data in AI: Challenges, Applications, and Ethical Implications*. Jan. 3, 2024. URL: http://arxiv.org/pdf/2401.01629.

[21] Alan Hevner et al. "Design Science in Information Systems Research". In: *MIS Quarterly* 28.1 (2004), p. 75. ISSN: 02767783. DOI: 10.2307/25148625. URL: https://arizona.pure.elsevier.com/en/publications/design-science-in-information-systems-research.

[22] Alan Hevner. "A Three Cycle View of Design Science Research". In: *Scandinavian Journal of Information Systems* 19 (2007). URL: https://www.researchgate.net/publication/254804390_A_Three_Cycle_View_of_Design_Science_Research.

[23] Shirley Gregor and Alan R. Hevner. "Positioning and Presenting Design Science Research for Maximum Impact". In: *MIS Quarterly* 37.2 (2013), pp. 337–355. ISSN: 02767783. DOI: 10.25300/MISQ/2013/37.2.01.

[24] Robert K. Yin. *Case study research and applications: Design and methods*. Sixth edition. Los Angeles et al.: SAGE, 2018. ISBN: 9781506336169.

[25] Forrest Shull, Janice Singer, and Dag I. K. Sjøberg. *Guide to advanced empirical software engineering*. London: Springer, 2008. ISBN: 978-1-84800-043-8. DOI: 10.1007/978-1-84800-044-5.

[26] B. A. Kitchenham et al. “Preliminary guidelines for empirical research in software engineering”. In: *IEEE Transactions on Software Engineering* 28.8 (2002), pp. 721–734. ISSN: 0098-5589. DOI: 10.1109/TSE.2002.1027796.

[27] Per Runeson and Martin Höst. “Guidelines for conducting and reporting case study research in software engineering”. In: *Empirical Software Engineering* 14.2 (2009), pp. 131–164. ISSN: 1382-3256. DOI: 10.1007/s10664-008-9102-8.

[28] M. V. Zelkowitz and D. R. Wallace. “Experimental models for validating technology”. In: *Computer* 31.5 (1998), pp. 23–31. ISSN: 00189162. DOI: 10.1109/2.675630.

[29] D. Sculley et al. “Hidden Technical Debt in Machine Learning Systems”. In: *Advances in Neural Information Processing Systems* 28 (2015). URL: https://proceedings.neurips.cc/paper_files/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf.

[30] Saleema Amershi et al. “Software Engineering for Machine Learning: A Case Study”. In: *2019 IEEE/ACM 41st International Conference on Software Engineering: Software engineering in practice*. Piscataway, NJ: IEEE, 2019, pp. 291–300. ISBN: 978-1-7281-1760-7. DOI: 10.1109/ICSE-SEIP.2019.00042.

[31] Rasmus Berg Palm, Florian Laws, and Ole Winther. “Attend, Copy, Parse - End-to-end information extraction from documents”. In: *ICDAR* (2019). URL: https://arxiv.org/pdf/1812.07248.

[32] Anoop Raveendra Katti et al. “Chargrid: Towards Understanding 2D Documents”. In: *Proceedings of EMNLP*. 2018. URL: https://arxiv.org/pdf/1809.08799.

[33] Alexander Osterwalder and Yves Pigneur. *Business model generation: A handbook for visionaries, game changers, and challengers*. New York: Wiley&Sons, 2010. ISBN: 9780470876411.

[34] Ajay Agrawal, Avi Goldfarb, and Joshua Gans. *A Simple Tool to Start Making Decisions with the Help of AI*. Ed. by Harvard Business Review Cases. 2018. URL: https://hbr.org/2018/04/a-simple-tool-to-start-making-decisions-with-the-help-of-ai (visited on 04/16/2021).

[35] Hanan Shteingart et al. *Machine Learning Prescriptive Canvas for Optimizing Business Outcomes*. June 21, 2022. URL: http://arxiv.org/pdf/2206.10333.

[36] Graham A. Cutting and Anne-Francoise Cutting-Decelle. *Intelligent Document Processing -Methods and Tools in the real world*. Dec. 28, 2021. URL: http://arxiv.org/pdf/2112.14070.

[37] Bertin Klein, Stevan Agne, and Andreas Dengel. “Results of a Study on Invoice-Reading Systems in Germany”. In: *Document analysis systems VI*. Ed. by David Hutchison. Vol. 3163. Lecture Notes in Computer Science. Berlin: Springer, 2004, pp. 451–462. ISBN: 978-3-540-23060-1. DOI: 10.1007/978-3-540-28640-0_43.

[38] Lukas-Walter Thiée, Felix Krieger, and Burkhardt Funk. “Extraction of Information from Invoices – Challenges in the Extraction Pipeline”. In: *Informatik 2023*. Ed. by Maike Klein et al. Lecture notes in Informatics (LNI) Proceedings. Bonn: Gesellschaft für Informatik, 2023, pp. 1777–1792. ISBN: 9783885797319. DOI: 10.18420/INF2023_180. (Visited on 01/10/2024).

[39] Felix Krieger et al. “Information Extraction from Invoices: A Graph Neural Network Approach for Datasets with High Layout Variety”. In: *Wirtschaftsinformatik 2021 Proceedings* (2021). URL: https://aisel.aisnet.org/wi2021/RDataScience/Track09/4.

[40] Serif Adali, A. Coskun Sonmez, and Mehmet Gokturk. “An Integrated Architecture for Processing Business Documents in Turkish”. In: *Computational linguistics and intelligent text processing*. Ed. by Alexander Gelbukh. Vol. 5449. Lecture notes in computer science Theoretical Computer Science and General Issues. Berlin and Heidelberg: Springer, 2009, pp. 394–405. ISBN: 978-3-642-00381-3. DOI: 10.1007/978-3-642-00382-0_32.

[41] Yolande Belaid and Abdel Belaid. “Morphological tagging approach in document analysis of invoices”. In: *Proceedings of the 17th International Conference on Pattern Recognition, 2004.* 2004, 469–472 Vol.1. ISBN: 0-7695-2128-2. DOI: 10.1109/ICPR.2004.1334166.

[42] Bill Janssen et al. “Receipts2Go”. In: *Proceedings of the 2012 ACM symposium on Document engineering*. Ed. by Cyril Concolato. ACM Conferences. New York, NY: ACM, 2012. ISBN: 9781450311168. DOI: 10.1145/2361354.2361381.

[43] Henrik Nell. “Quantifying the noise tolerance of the OCR engine Tesseract using a simulated environment”. Master Thesis. Karlskrona, Sweden: Blekinge Tekniska Högskola, 2015. URL: http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A831347&dswid=6545.

[44] Felix Krieger and Paul Drews. “Leveraging Big Data and Analytics for Auditing: Towards a Taxonomy”. In: *ICIS 2018 Proceedings* (2018). URL: https://aisel.aisnet.org/icis2018/datascience/Presentations/16.

[45] Brian Davis et al. “Deep Visual Template-Free Form Parsing”. In: *15th International Conference on Document Analysis and Recognition* (2019). URL: https://arxiv.org/pdf/1909.02576.

[46] R. B. Palm, O. Winther, and F. Laws. “CloudScan - A Configuration-Free Invoice Analysis System Using Recurrent Neural Networks”. In: *12th International Conference on Document Analysis and Recognition*. 2013, pp. 406–413. ISBN: 978-0-7695-4999-6. DOI: 10.1109/ICDAR.2017.74.

[47] Yiheng Xu et al. *LayoutLM: Pre-training of Text and Layout for Document Image Understanding*. 2020. DOI: 10.1145/3394486.3403172. URL: https://arxiv.org/pdf/1912.13318.

[48] Łukasz Garncarek et al. “LAMBERT: Layout-Aware Language Modeling for Information Extraction”. In: *Document Analysis and Recognition - ICDAR 2021* Vol. 12821 (2021), pp. 532–547. DOI: 10.1007/978-3-030-86549-8_34. URL: https://arxiv.org/pdf/2002.08087.

[49] Timo I. Denk and Christian Reisswig. “BERTgrid: Contextualized Embedding for 2D Document Representation and Understanding”. In: *33rd Conference on Neural Information Processing Systems, Vancouver, Canada.* (2019). URL: https://arxiv.org/pdf/1909.04948.

[50] Arsen Yeghiazaryan et al. “Tokengrid: Toward More Efficient Data Extraction From Unstructured Documents”. In: *IEEE Access* 10 (2022), pp. 39261–39268. DOI: 10.1109/ACCESS.2022.3164674.

[51] Bodhisattwa Prasad Majumder et al. “Representation Learning for Information Extraction from Form-like Documents”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020), pp. 6495–6504. DOI: 10.18653/v1/2020.acl-main.580. URL: https://aclanthology.org/2020.acl-main.580/.

[52] Yang Xu et al. *LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding*. 2020. URL: https://arxiv.org/pdf/2012.14740.

[53] D. Lohani, A. Belaïd, and Y. Belaïd. “An Invoice Reading System Using a Graph Convolutional Network”. In: *ACCV Workshops*. 2018, pp. 144–158. DOI: 10.1007/978-3-030-21074-8_12. URL: https://link.springer.com/chapter/10.1007/978-3-030-21074-8_12.

[54] Petar Veličković et al. *Graph Attention Networks*. Oct. 30, 2017. URL: http://arxiv.org/pdf/1710.10903.pdf.

[55] Jie Zhou et al. “Graph neural networks: A review of methods and applications”. In: *AI Open* 1 (2020), pp. 57–81. ISSN: 26666510. DOI: 10.1016/j.aiopen.2021.01.001.

[56] Lukas-Walter Thiée. “A systematic literature review of machine learning canvases”. In: *Informatik 2021*. GI-edition Proceedings. Bonn: Gesellschaft für Informatik e.V. (GI), 2021, pp. 1221–1235. ISBN: 9783885797081. DOI: 10.18420/informatik2021-101. URL: https://doi.org/10.18420/informatik2021-101.

[57] Lukas-Walter Thiée. “Developing an ontology for data science projects to facilitate the design process of a canvas”. In: *Wirtschaftsinformatik 2022 Proceedings*. Association for Information Systems (AIS), 2022. URL: https://aisel.aisnet.org/wi2022/ai/ai/13.

[58] Lukas-Walter Thiée. “Ablation Study of a Multimodal Gat Network on Perfect Synthetic and Real-world Data to Investigate the Influence of Language Models in Invoice Recognition”. In: *Document analysis and recognition - ICDAR 2024 workshops*. Ed. by Harold Mouchère and Anna Zhu. Lecture Notes in Computer Science. Cham: Springer, 2024, pp. 199–212. ISBN: 978-3-031-70642-4. DOI: 10.1007/978-3-031-70642-4_13. URL: https://link.springer.com/chapter/10.1007/978-3-031-70642-4_13.

[59] Lukas-Walter Thiée and Burkhardt Funk. “Enhancing Invoice Recognition with LLM Embeddings in GAT Networks”. In: *AMCIS 2025 Proceedings* (2025). URL: https://aisel.aisnet.org/amcis2025/sig_svc/sig_svc/6.

[60] Michael Schulz et al. *DASC-PM v1.1 - Ein Vorgehensmodell für Data-Science-Projekte*. Ed. by Universitäts- und Landesbibliothek Sachsen-Anhalt and Martin-Luther Universität. 2022. DOI: 10.25673/85296. URL: http://dx.doi.org/10.25673/85296.

[61] Michael René Schulte et al. “Parking Space Management Through Deep Learning – An Approach for Automated, Low-Cost and Scalable Real-Time Detection of Parking Space Occupancy”. In: *Innovation through information systems*. Ed. by Frederik Ahlemann, Reinhard Schütte, and Stefan Stieglitz. Springer eBook Collection. Cham: Springer International Publishing and Springer, 2021, pp. 642–655. ISBN: 978-3-030-86797-3. DOI: 10.1007/978-3-030-86797-3_42. URL: https://link.springer.com/chapter/10.1007/978-3-030-86797-3_42.

[62] Lukas-Walter Thiée et al. "'Spread the app, not the virus' – An extensive SEM-approach to understand pandemic tracing app usage in Germany". In: *ECIS 2021 Research Papers*. Association for Information Systems (AIS), 2021. URL: https://aisel.aisnet.org/ecis2021_rp/123/.

[63] Felix Hertlein, Alexander Naumann, and Patrick Philipp. *Inv3D: a high-resolution 3D invoice dataset for template-guided single-image document unwarping - Meta data*. Karlsruhe Institute of Technology, 2023. DOI: 10.35097/1730. URL: https://publikationen.bibliothek.kit.edu/1000161884.

[64] Stěpán Šimsa et al. *DocILE Benchmark for Document Information Localization and Extraction*. Feb. 11, 2023. URL: https://arxiv.org/pdf/2302.05658.

[65] Srikar Appalaraju et al. *DocFormer: End-to-End Transformer for Document Understanding*. 2021. DOI: 10.48550/ARXIV.2106.11539.

[66] Dongsheng Wang et al. *DocLLM: A layout-aware generative language model for multimodal document understanding*. 2024. DOI: 10.48550/ARXIV.2401.00908.

[67] Masato Fujitake. *LayoutLLM: Large Language Model Instruction Tuning for Visually Rich Document Understanding*. 2024. DOI: 10.48550/ARXIV.2403.14252.

# Part II

# Publications

# Paper 1

# A systematic Literature Review of Machine Learning Canvases

## Outline



## Bibliographic Information


Lukas-Walter Thiée.
Leuphana Universität Lüneburg, Institute for Information Systems, Lüneburg, Germany.
05.10.2021, https://doi.org/10.18420/informatik2021-101
INFORMATIK 2021 - Die 51. Jahrestagung der Gesellschaft fur Informatik in: Computer Science and Sustainability. Gesellschaft für Informatik e.V. (GI) (Hrsg.). Bonn: Gesellschaft für Informatik e.V., S. 1221-1235 15 S. (Lecture Notes in Informatics (LNI), Proceedings - Series of the Gesellschaft fur Informatik (GI); Band P-314).


## Copyright Notice

# Abstract

The use of machine learning technology is still significantly lower in small and medium sized enterprises than in large enterprises. It seems that there are specific challenges in the implementation of data-driven methods, that hinder SMEs in their adoption. One approach to support the initialization and execution of such methods is the use of boundary objects, e.g., canvases, serving as a visual communication document. As it is not clear which approaches are being pursued in detail and how they are interrelated, in this paper, a systematic literature review is being presented, that identifies and analyzes 18 canvas artifacts. These canvases represent the status quo and they can be grouped into four distinct categories of different foci. The aggregation of the fields and questions provides an essence of canvas contents, to point out gaps and ultimately to expand the canvas approach as well as ML adoption.



**Keywords**

Machine Learning, Canvas, Literature Review, SME

# 1 Introduction and Approach

The extent to which companies apply machine learning (ML) has increased in recent years and with over a quarter of German companies using multiple ML technologies today [1], it is far from a niche technology. While larger companies have been taking advantage of data-driven analytics for some time, mainly with a focus on process optimization [2], [3] and product development [4], for small and medium sized companies (SME) in Germany it is still difficult to start and use ML applications in their businesses [5]. As of 2020 only 10% of the small companies utilize multiple ML technologies and over 22% admit that ML applications are still not a topic in the company at all [1]. Furthermore, only 29% of SMEs assess ML as a driver for innovation and product development [1], which could indicate an imbalance in the assessment of opportunities and challenges of data science (DS), artificial intelligence (AI) in particular. Nevertheless, two out of five SMEs are beginning to plan digitalization projects and another 29% are considering to do so [6]. Even though studies have shown benefits of data usage for SMEs [7], SMEs have been focusing descriptive approaches and conventional business intelligence [8], rather than leveraging predictive or prescriptive data analytics to its full potential [9]. It seems that to the same extent as SMEs were challenged with the introduction of information technology in its beginnings [10], they are now challenged to implement advanced data analytics and data strategies [11]. However, in practice there seem to be specific challenges for SMEs regarding the initialization and successful execution of ML projects [12]. One of the main reasons for the slow adoption of ML technology in SMEs is probably the lack of resources [13]. Compared to larger companies, SMEs have less financial power and fewer or no skilled personnel for specific tasks in ML projects. Consequently, this gives rise to a variety of challenges. Lack of management focus and experience [14], lack of internal and external experts [15], or complexity of (sequential) data [16] are exemplary challenges that can impede both project initiation and success. One of these challenges is the detailed description and expedient guidance in the use case definition and execution phase of an ML project. Naturally, these projects include multiple stakeholders at varying levels of data-literacy [17], which unfortunately promotes the development of communication silos [1]. This communication barrier, which exists both within the company and externally, must be overcome. Therefore, in addition to general methods that foster ML adoption, such as design thinking workshops, specific tools have been proposed, such as the DUCAR process model for smart picking of ML use cases [18], or procedures to identify and prioritize use cases [19], [20]. Also the use of Guided Analytics [21] has been proposed, so that non-experts can access ML tools [22]. Nevertheless, this approach currently still mainly focuses model tuning. Another promising approach to facilitate ML communication and execution is the use of canvases. *"A canvas is just a visual chart to describe a complex object, in a better way than a simple text document. Each key aspect of that object has its own "block", and blocks are arranged on the chart in a way that makes visual sense"* [23]. Canvases are a kind of boundary object, i.e., a document that serves as a communication platform between multiple stakeholders, motivates cross-disciplinary collaboration [24], and provides common identity [25]. Originating from the Business Model Canvas by Osterwalder and Pigneur [26], this approach has been further developed and analyzed in many ways, for example in data-driven business models [27]. However, it is not clear which of these canvas approaches is most promising for initial ML projects.

**Research Approach and Contribution**. For this reason, it is essential to review

relevant resources systematically. Comparing different data-related canvases not only streamlines research in this field, but also provides useful guidance in business practice. A compilation of existing tools and a possible extension that can easily be applied in practice, can help bridging between the aforementioned corporate silos, i.e., decision-makers, developers, domain experts, and external partners, as well as promote the execution of ML tasks. Furthermore, the result of this review will be the foundation of conducting empirical research in longitudinal case studies to investigate the applicability and usefulness of such boundary objects. The detailed research questions for this paper are: (1) *Which canvas models, that address ML or AI implementation, are available, and which contents do they cover?* and (2) *Where are gaps and what are potential extensions of these canvases in order to address specific challenges and needs in initial ML projects?* Providing answers to these questions might support initialization and successful execution of ML projects in small organizations. Therefore, the goal for this review is to find relevant canvas tools, compare these tools systematically, and potentially enhance them.

# 2 Methodology: Literature search process documentation

In order to find existing evidence, identify gaps, and build an appropriate background on the topic a systematic literature review is being conducted [28]. The review shall identify relevant sources within a defined scope, synthesize the findings [29], and most of all provide guidance on research and practice [30]. As rigor and reproducibility are key to qualitative IS reviews [31], the search process is documented in the following. The selection of databases (see Table 1.1) comprises ten international renowned databases, such as the Web of Science Core Collection, the ACM digital library and IEEE Xplore. The latter belong to the most impactful databases in computer science [32]. The selection of databases was inspired by the journal rating for IS literature from VHB JOURQUAL3 [33], which also comprises IS conference proceedings. The AIS Electronic Library features important IS outlets, such as MISQ or BISE. Furthermore, two online archives, Google Scholar and ArXiv.org; as well as three German online libraries were included, because they basically supplement the corpus of resources. To further integrate practitioner views, the Harvard Data Science Review, which is a non-peer-reviewed open access journal of the MIT Press, was also considered.

The formulation of the search terms is very important in this matter, because there are often no clear dividing lines between the terms, e.g., the phrases "data mining", "data-driven", "digitalization", or "big data analytics" all have a large intersection, which is mainly due to the fact that in the field of IS there is a plethora of terms and abbreviations that describe the topic area or sub-areas. Not even the distinction between AI, ML and deep learning is unambiguous. For example, an initial search in the EBSCOhost Business Source library, yields over 30,000 results. Therefore, we limit the search to the search string: {*"Machine Learning Canvas OR "Artificial Intelligence Canvas" OR "ML Canvas" OR "AI Canvas"*}. This initially excluded phrases such as digital canvas or data science canvas. The query was individually adapted to the requirements of the respective database. No keyword was integrated that was specifically suitable for SMEs, since it was assumed that so far rather general approaches are available. The query was executed on all fields, i.e., title, abstract, keywords and full text, and the time span for results was limited

Table 1.1: Databases and literature search process results

| Database | Results | Dupl. | TAK | |
|---|---|---|---|---|
| AISeL | 4 | 4 | 4 | |
| EBSCOhost BSC | 2 | 2 | 2 | Backward: |
| ACM Digital Library | 2 | 1 | 0 | +15 |
| Emerald Insight | 0 | 0 | 0 | |
| ScienceDirect | 3 | 3 | 1 | |
| SCOPUS | 8 | 4 | 1 | Forward: |
| IEEE Xplore | 1 | 1 | 1 | +1 |
| Web of Science | 15 | 14 | 2 | |
| Google Scholar | 3 | 0 | 0 | |
| ArXiv | 1 | 0 | 0 | Language: |
| Taylor&Francis | 3 | 1 | 0 | -2 |
| Springer Link | 3 | 3 | 1 | |
| HMD | 0 | 0 | 0 | |
| Duncker & Humblot | 0 | 0 | 0 | Artifact Identity: |
| RonPub | 0 | 0 | 0 | -1 |
| Harvard DS Review | 0 | 0 | 0 | |
| Total | 45 | 33 | 12 | 25 |

to the years 2000 to 2021. The search was conducted end of March 2021. As the search term is rather strict, there were 45 initial hits within the included databases (see Table 1.1). Eliminating duplicates left 33 results and screening titles, abstracts, and keywords (TAK) reduced the results to 12 articles. The screening excluded hits that didn't relate to ML or AI canvases or projects, such as the "Business Ethics Canvas" [34]. A backward search through the listed references as proposed by Webster and Watson [29] complements the search and adds 15 articles to the results. The backward search was performed in such a way that a TAK screening was performed for references that showed a promising title related to canvases, such as "Data-Driven Business Models" [35]. Two articles were excluded from the results, because they were not written in English and another article was excluded, because it was a foundational research article to a different article in the results, describing the same artifact, "ML-Process Canvas" [36]. The results all originated from the time window of the years 2016-2021. This period is very recent, therefore, conducting a forward search in the mentioned databases didn't reveal any further relevant results. However, a sample forward search on Google yielded citations on the professional online network LinkedIn, which produced another result [37]. In total there were 25 relevant articles left, of which 18 contained a canvas for in-depth analysis, which is the answer to part one of research question one (Table 1.2).

# 3 Analysis and Results

The analysis of the canvases is performed in two parts, structural analysis and content categorization. The structural analysis considers the layout and the process of filling. For this purpose, both the artifacts themselves and a full-text analysis are performed. Also,

Table 1.2: Literature search results – 18 canvas artifacts

| Year | Source | Canvas Artifact | #Fields | Structure |
|---|---|---|---|---|
| 2016 | [38] | Data Canvas: Data-Need Fit | 10 | F, D, E |
| 2017 | [39] | Data Value Map | 14 | F, S, D |
| 2017 | [40] | Digitalization Canvas | 9 | F, S, Q |
| 2018 | [41] | AI Canvas | 7 | F, Q |
| 2018 | [42] | AI Canvas | 8 | F, Q |
| 2018 | [37] | The ML Canvas (Big Data MBA Version) | +2 | F, D, Q |
| 2018 | [37] | Hypothesis Development Canvas v1.1 | 10 | F, Q |
| 2019 | [35] | Data Insight Generator | 6 | M, D, Q |
| 2019 | [23] | Machine Learning Canvas v0.4 | 10 | F, D, Q |
| 2019 | [43] | Data Innovation Board | 14 | F, S, Q |
| 2019 | [44] | Data Collection Map | 12 | F, E |
| 2019 | [45] | AI Project Canvas | 9 | F, S, Q |
| 2020 | [46] | AI performance canvas (prototype) | 10 | F, S, Q |
| 2020 | [47] | Data Product Canvas | 7 | F, Q, E |
| 2020 | [48] | Key Activity Canvas | 10 | M, Q, E |
| 2020 | [49] | Canvas for the use of AI (author) | 7 | F, Q |
| 2020 | [50] | ML Lifecycle Canvas | 6 | M, D, Q |
| 2021 | [17] | Enterprise AI Canvas | 12 | F, S, Q |

Note: F=Fields, M=Matrix, S=Sections, E=Examples, D=Descriptions, Q=Questions

it is being determined whether and which roles or persons are being addressed to fill the respective canvas and if the filling is self-sufficient and detailed guidelines published. Further, if applicable, the scientific derivation and evaluation of the artifacts are investigated.
**Structural analysis**. Almost all artifacts are structured as a canvas with fields that have a descriptive title accompanied by either questions or examples in the fields (see Table 1.2, Structure column). Only a few use a 2-dimensional matrix arrangement with rows and columns. Some artifacts provide sections or arrows in the layout to guide the user in filling in the fields, e.g., the "Key Activity Canvas" provides dotted arrows to visualize interactions between the customer, the company, and the partners [48], or Zawadzki draws arrows that indicate how to go from section to section in the "AI Project Canvas" [45]. 4 articles provide written guidelines how to proceed in filling in the canvas fields, such as "Explore, Ideate, Evaluate" [43], "Think, Validate, Know" [35], design loops [50], or agile development [40]. Whereas 3 others point to the iterative nature of the tool [23], [37], [48], the remaining 11 do not specify the process (once or iterative), so that a single pass of the filling has to be assumed. In terms of size, or number of fields respectively, the 18 artifacts range from 7 fields in the "AI Canvas" [41] to 14 fields in the "Data Innovation Board" [43]. The highest number of cells, logically, follows from one of the matrix approaches, namely 21 cells in the "Key Activity Canvas" [48]. As an overarching finding, it can be noted that all articles see data-driven projects, and the respective canvas in particular, as an interdisciplinary task. Nevertheless, 5 of the 18 articles don't specify the person or department which should fill in the canvas. 5 others only mention general terms, such as "business stakeholders" [37], "heterogeneous stakeholder groups" [35], "pioneers" [42], or "different departments and diverse expertise" [38]. Specifically mentioned are Data

and AI project managers, IT departments, domain experts, service design teams, data science teams, "senior executives, middle management, frontline staff, business stakeholders, technology stakeholders and customers" [39], or in general "managers, who provide the glue between everyone" [23].

Regarding the scientific derivation, 6 articles explicitly name Design Science Research, Action Design Research, Research through Design, or Design Thinking as methodological procedures in their articles, e.g., applying questionnaires, triangulation, or design principles as methods. 5 of these also describe the evaluation procedure in their research, which is predominantly a focus group workshop. Also, the ontology, which the artifact is based upon, is mentioned in 2 articles, namely for the "Data Collection Map" [44], and for the "Data Innovation Board" [43]. 2 other articles describe the interviews and workshops conducted as part of a case study [40], [38]. [42] and [23] ground their artifacts on the Business Model Canvas by [26]. However, nearly half of the results (8/18) don't specify the scientific method, which the artifact is based upon, and these contributions also don't mention the evaluation technique applied.

**Categorization**. In the second part of the analysis, the individual canvases and their core contents and objectives are being investigated. 6 of the 18 canvases explicitly label the canvas with the term AI, e.g., the "AI Canvas" by Agrawal et al. (2018) [41], and 3 assign ML to their artifact, e.g., the "ML Lifecycle Canvas" by Zhou et al. (2020) [50]. The naming already indicates that different foci are being set. As the artifacts all belong to the same realm of data science, categorizing them is ambitious due to the proximity of their contents. Nevertheless, four categories can be proposed, as summarized in Figure 1.1. This methodological procedure is in line with Webster and Watson's (2002) call for a concept-centric approach in IS literature reviews [29].

| **ML / DS Focus** | **(AI) Project Focus** | **Data Value Focus** | **Data Source Focus** |
|---|---|---|---|
| • [23], [37] [48], [50] | • [17],[41], [42], [45], [49], [46], [40] | • [47], [39], [38], [43] | • [35], [44] |

Figure 1.1: Categories of canvases with different thematic foci

**Machine Learning / Data Science Focus**. The first category is formed by canvases with a technical focus on machine learning and data science. It includes the "Machine Learning Canvas v0.4" [23] (see Appendix B) and its extension the "Machine Learning Canvas (Big Data MBA Version)" [37], as well as the "Hypothesis Development Canvas" [37], the "ML Lifecycle Canvas" [50], and the "Key Activity Canvas" [48]. These artifacts belong to this category, because they describe concrete machine learning or data science process steps, e.g., the definition of inputs and outputs or engineering of corresponding features. *"The Machine Learning Canvas [is] the first step towards making sure you connect what ML can do to your organization's objectives, and towards assessing feasibility. It should be filled in before starting any implementation work, and even before Exploratory Data Analysis"* [23]. This artifact is the most technical, as it contains fields like "ML task" or "Features". Dorard also integrates metrics – offline and online – and value proposition into his canvas. However, it is supposed to be an initial document, and *"the canvas results [have to be] translated to a technical specification document"* [51] later on. Schmarzo builds upon Dorard's canvas and proposes two additional fields, namely

Prescription and Automation in the "Big Data MBA Version", to adopt the canvas to "data science requirements" [37]. This extension shall form the path from a small ML use case to an integrated, scaled application. Nevertheless, it's not obvious how to really do that. Schmarzo also proposes the "Hypothesis Development Canvas v1.1" in order to facilitate *"collaboration between the business stakeholder and the Data Science team to identify the hypothesis requirements that underpin Data Science engagement success"* [37].

Another rather technical approach is the "ML Lifecycle Canvas". It's a *"conceptual design tool featuring the holistic visualization of cooperation among ML, users, and scenarios during the ML lifecycle"* [50]. This canvas is unique in the regard, that it provides a detailed question list and "persona cards" to fill in the canvas. Questions like *"Is there any ML model feasible for completing the required tasks?"* illustrate the level of detail regarding the final ML design process. The authors align the questionnaire with questions from "existing guidebooks on human-AI interaction" [50], like Google PAIR [52] or [53]. The "Key Activity Canvas" [48] is a matrix arrangement and integrates three views, Customer, Company and Partners, for the *"methodological assistance (key activities) guiding the actual conceptualization of necessary activities in analytics-based services"*.

**(AI) Project Focus**. The second category includes canvases with a holistic view on AI projects, "AI Canvas" [41], "AI Canvas" [42], "AI Project Canvas" [45], "Canvas for the use of AI" [49], "Enterprise AI Canvas" [17], "AI performance canvas" [46], and "Digitalization Canvas" [40]. In principle, these could synonymously be applied to ML projects. However, the distinction from the first category arises from the fact that a focus is placed on the overall project rather than on the technical details, e.g., the cost and revenue structure of an AI project. The "AI Canvas" [41] is supposed to support corporate decision-making through prediction models. *"How can you decide whether employing a prediction machine will improve matters? The AI Canvas is a simple tool that helps you organize what you need to know into seven categories [Prediction, Judgment, Action, Outcome, Input, Training, Feedback] in order to systematically make that assessment"* [41]. It's rather designed "for a non-technical audience" [46]. Technical details and business integration are not being covered [17]. This is also true for the "AI Canvas" by Dewalt and Rands. They try to connect a business opportunity via a strategy with a solution using AI models. The canvas is the summary to lay the path to "Become an AI Company in 90 Days" [42]. Similarly, based on the original Business Model Canvas the "AI Project Canvas" by Zawadzki is helpful for "project managers" [17] and is intended to "structure and convey the holistic idea of your AI project to others" [45]. Another way to "determine the relevance of artificial intelligence for your company" [49] is the "Canvas for the use of AI", which is embedded in a corporate transformation process towards AI. The approach by Kerzel "Enterprise AI Canvas" [17] features two parts, one with a technical focus and one with a business focus. It's supposed to *"bring business and Data Science experts together and systematically evaluate potentially new business opportunities"* [17]. Although part 2 "model and data view" shows similarities to Dorard's approach, technical details on modeling are omitted. The AI Project Focus category also includes the "AI performance canvas", whose main objective is the *"collaborative construction of performance goals for data & AI products in organizations"* [46]. In this approach, there is a strong focus on feasibility of the product and legal/compliance issues in data governance. Nevertheless, only a prototype canvas is being presented and the "trigger questions" of the fields are not yet published. The "Digitalization Canvas" [40] is the only canvas specifically designed

for SMEs and corresponds to a holistic digitalization strategy. The approach promotes easy-to-implement and strategic data projects. While the canvas itself isn't self-sufficient, there are detailed questionnaires within the case study. *"The Digitalization Canvas together with a project portfolio defines a concrete roadmap to digitalization and summarizes the arguments to apply for the necessary budget"* [40].

**Data Value Focus**. The third category are canvases with a data value or data product focus, namely the "Data Product Canvas" [47], the "Data Value Map" [39], the "Data Canvas – Data-Need Fit" [38], and the "Data Innovation Board" [43]. These canvases seek to identify and implement the value of data through tangible customer or user benefits and information gain through data. The "Data Product Canvas" [47] can be used when an organization in an initiation phase aims to develop ideas for a data product. Similarly, the *"Data Canvas and Data-Need Fit are intended to spark a discussion on available data in organizations among diverse stakeholders. The Data Canvas provides trigger questions and a visual representation that help to develop a common understanding of available data"* [38]. *"A Data-Need Fit is found when data sources contribute gain creators and pain relievers that users find valuable"* [38]. *"To facilitate a shared understanding for data initiatives"* [39] is also the main objective of the "Data Value Map". This canvas focuses the process from data creator to data user and emphasizes data governance topics, such as data principles and access, and business-related topics such as cost reduction or revenue generation. Finally, the "Data Innovation Board" [43] features three design thinking steps, namely exploration, ideation, and evaluation, and promotes the description of performance goals in "a visual collaboration tool that anyone can work with". The artifact is clearly supposed to facilitate initial progress in a data-driven project, and the authors note that *"it is a beginner's tool* [, which needs] *to be accompanied by other visual tools with more specific views on technology and algorithms"* [43].

**Data Source Focus**. The fourth category includes canvases that focus on explaining the data source and the data processing, in order to gain a better understanding of data as an asset: the "Data Insight Generator" [35] and the "Data Collection Map" [44]. The "Data Insight Generator" [35] is a workshop canvas that connects key data resources with a value proposition for data-driven business models. It contains columns for Pipes, Analytics and Insight. The process of filling in is guided by the rows Think, Validate, and Know, which are to be processed one after the other. Finally, the *"Data Collection Map was designed as an entry point in the ideation process of data-driven use cases. Hence, the purpose of the tool is to get people to think about data (e.g. clicks and engagement metrics) instead of IT systems (e.g. Google Analytics) and to raise the necessary data awareness about the available data resources within the organization"* [44]. It's basically an add-on to the "Data Innovation Board" by Kronsbein and Mueller (2019) [43].

**Categorization of the fields and questions**. As it is not sufficient to only categorize the artifacts on a title level, the fields are being analyzed. In order to conceptualize the core content of the canvases and thus answer the second part of research question one, all fields (or headers in matrix patterns) are being captured and groups as well as top categories are being proposed. This should clarify where the canvases overlap and where the focus has been placed so far. In total there are 163 fields in the results, e.g., the "Machine Learning Canvas v0.4" [23] contains 10 fields that contain one or more questions to guide the filling (see Table 1.2). Logically, this count includes multiple entries. Therefore, the fields that cover a similar area in terms of content are grouped together. In the next step, multiple entries are eliminated and fields are combined that either de-

scribe exactly or almost the same thing. For example, fields like "Data Sources", "AI data base", "Metadata", and all the various data types from the "Data Collection Map" [44], belong to the same group of "Data Sources" (see Table 1.3). As most fields contain more than one guiding question or example and not every field headline describes the same content, in order to refine the assignment, all guiding questions or examples are also individually examined. Ultimately, 39 groups and 11 top categories can be created. The content intersection of the final clusters compiles the name of their top category, e.g., the groups "Data Quality", "Data Policies", and "Data Lifecycle" build the top category "Data Governance". The result of this assignment is presented in Table 1.3.

Table 1.3: Categories and groups of all canvas fields and questions with examples

| Category | Group | # | Example Questions |
|---|---|---|---|
| Business & Value | Strategy | 13 | "What trends, market facts are relevant for the topic [...]?" [43] |
| | Risks | 10 | "What risks are associated with the use of AI for our industry?" [49] |
| | Operat. value | 7 | "How does the use-case generate value?" [17] |
| | Revenue | 6 | "How will the project generate revenue?" [45] |
| | Cost | 9 | "Will the project reduce internal costs [...]?" [45] |
| Product & Customer | Product/Service | 14 | "Which potentials in the production area can be leveraged?" [49] |
| | Delivery | 7 | "In which form do we provide the data service to our users [...]?" [47] |
| | Customer | 16 | "Who is our customer?" [47] |
| | Gains/Value | 9 | "What is the customer value of the hypothesis?" [37] |
| | Pains/Needs | 9 | "What customer pain is the AI project solving?" [45] |
| Organization | Implementation | 8 | "How might we implement the idea?" [43] |
| | Internal skills | 7 | "Is the required know-how for the implementation available inhouse?" [40] |
| | Stakeholders | 3 | "Sponsor: Which senior manager is responsible?" [17] |
| | Domain | 2 | "Which domain expertise is needed?" [17] |
| | Partners | 8 | "Which external services and products are required?" [40] |
| Technology | Systems | 7 | "Which systems are required and already available to handle data?" [17] |
| | Infrastructure | 7 | "How are the models served? Edge, on-premise or Cloud?" [17] |
| | Integration | 4 | "Which networks along the value chain are necessary?" [49] |
| Data Characteristics | Data types | 25 | "What kind of data do we need for training?" [50] |
| | Data sources | 9 | "Which raw data sources can we use (internal and external)?" [23] |
| | Data availability | 14 | "What data is currently collected in the organization?" [43] |
| | Data collection | 8 | "How might we collect the needed data?" [43] |
| | Data pipeline | 3 | "Which interfaces can I use to combine this data?" [35] |
| Data Governance | Data quality | 6 | "How is the validity of the data, [...] consistency, and completeness?" [40] |
| | Data policies | 9 | "Are there any compliance requirements [...]?" [40] |
| | Data lifecycle | 5 | "Determining the definition, [...] and retirement of data." [39] |
| Preprocessing | Data preparation | 4 | "What do we have to do to prepare the data [...]?" [48] |
| | Features | 2 | "Which features are likely important?" [17] |
| | Inputs | 4 | "What are the model inputs?" [42] |
| Modelling | Learning | 6 | "Is there any ML model suitable for the available dataset?" [50] |
| | Analytics | 5 | "With which data analytics methods do we generate insights [...]?" [47] |
| | Interpretation | 4 | "How can we interpret the mined patterns?" [48] |
| | Prediction | 5 | "What should be predicted?" [17] |
| | Decision | 9 | "Prescription: Once we have a prediction, what do we do?" [37] |
| Evaluation | KPI Model | 3 | "Which key metric are you optimizing for?" [45] |
| | KPI Business | 6 | "Outcome: What are your metrics for task success?" [41] |
| | Improvements | 3 | "How can you use the outcomes to improve the algorithm?" [41] |
| | Automation | 6 | "When do we create/update models with new training data?" [23] |
| | Live / Ex-post | 6 | "Methods and metrics to evaluate the system after deployment [...]." [23] |

# 4 Discussion and Conclusion

Using one of the categorized canvases is a solid starting point for initial ML projects. Practitioners can use the four categories as guidance and pick one of the mentioned canvases, e.g., if they want to explore their data, the canvases from the category "Data sources" will help. Although *"an over emphasis on technology"* [39] has been mentioned as a potential barrier, we feel that diving deeper into the technical details, i.e., data processing and modeling, is key to foster ML adoption. Therefore, with regard to research question two, the canvases with a ML/DS Focus can be recommended. SMEs can use the canvases and/or the catalog of questions to promote cooperation internally or externally, i.e., with research or consultancy. The canvases are all standalone artifacts for individual, valid use cases. Nevertheless, there is still room for improvement of the canvas approach for three reasons. (1) The first reason is, that there are still gaps regarding content. For example no detailed questions for algorithms, hyper parameter tuning, visualization of results, or concept drift [54] could be identified. Also scalability and feasibility checks were not really mentioned. (2) The second reason regards the applicability of the canvases. Clear guidelines on who, when, and how to use the canvas are needed, describing the explicit benefit. Otherwise, their usefulness is mitigated and scientific artifacts will not be favored against grey literature. A multipage description might not be too handy. However, a balance between detailed description, e.g., as in Google PAIR [52], and self-sufficiency, will provide the greatest benefit. Eventually, the evaluation benchmark of the canvases must be: "What's the artifact from the artifact?" That means, did using the canvas result in or facilitate building an ML application that provides value. (3) The third reason is concerning the fact that *"canvases help us ask the right questions, but they don't provide the answers"* [42]. In order to lift the canvas approach from pure ideation to application, providing answers to the questions is necessary to foster initial ML adoption, especially in SMEs. In particular, evaluating the concrete data potential, as a result of information gain through algorithmic processing, has to be elaborated. Therefore, guiding the software tool chain or the model selection as in the *scikit learn cheat sheet*[1] might be fruitful. Other potential extensions of the canvas approach could be the integration of cloud service platforms, e.g., Amazon Web Services, or other MLaaS providers [55], or questions regarding "ground truth" [56], or the inclusion of data classification schemes [44] like the Dublin Core Elements [57]. Taking these thoughts further, future research may include three key points. First, the question list has to be compared to a) the challenges of ML adoption in SMEs and b) to existing process descriptions, such as CRISP-DM [58]. Second, a comprehensive canvas for initial ML projects can be conceptualized from the findings of this review. And third, this concept can then be used and evaluated in empirical research, especially in workshops and case studies. The conceptualization of a new artifact and its evaluation would also address the inherent limitations of this review paper, as personal bias and experience could not totally be omitted, especially in the categorization parts.

[1] https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

# References


[1] Jürgen Hill et al. *Studie Machine Learning 2020*. Ed. by IDG Business Media GmbH. 2020. URL: https://www.lufthansa-industry-solutions.com/de-de/studien/idg-studie-machine-learning-2020 (visited on 02/04/2021).

[2] Jörg Krüger et al. *WGP-Standpunkt KI in der Produktion: Künstliche Intelligenz erschliessen für Unternehmen*. Ed. by WGP Wissenschaftliche Gesellschaft für Produktionstechnik e.V. 2019. URL: https://wgp.de/wp-content/uploads/WGP-Standpunkt_KI-final_20190906-2.pdf (visited on 05/07/2021).

[3] VDMA. *Quick Guide Machine Learning im Maschinen- und Anlagenbau*. Ed. by VDMA Software und Digitalisierung. Frankfurt, 2018. URL: https://www.vdma.org/documents/34570/1052572/Quick-Guide+Machine+Learning+-KI.pdf/8021ab42-79a6-5c72-c4f6-20b6c49ad54a?t=1615363921245 (visited on 04/14/2021).

[4] Dietmar Wolff and Richard Göbel. *Digitalisierung: Wie Die Digitalisierung Unsere Lebens- und Arbeitswelt Verändert: Segen oder Fluch*. Berlin, Heidelberg: Springer, 2018. ISBN: 3662548410.

[5] Matthias Parlings, Marie Lindemann, and Arno Kühn. *Künstliche Intelligenz im Mittelstand - Potenziale und Anwendungsbeispiele*. Ed. by Digital in NRW - Kompetenz für den Mittelstand. 2020. URL: https://www.mittelstand-digital.de/MD/Redaktion/DE/Publikationen/zentrum-dortmund-ki-im-mittelstand.pdf?__blob=publicationFile&v=2.

[6] Volker Zimmermann. *Unternehmensbefragung 2020: Anteil der Digitalisierungsplaner stagniert auf hohem Niveau*. Ed. by KfW Bankengruppe. 2020. URL: https://www.kfw.de/PDF/Download-Center/Konzernthemen/Research/PDF-Dokumente-Unternehmensbefragung/Unternehmensbefragung-2020-%E2%80%93-Digitalisierung.pdf (visited on 05/07/2021).

[7] Siti Aishah Mohd Selamat et al. "Big data analytics-A review of data-mining models for small and medium enterprises in the transportation sector". In: *WIREs (Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery)* 3 (2018). DOI: 10.1002/widm.1238. URL: https://onlinelibrary.wiley.com/doi/full/10.1002/widm.1238.

[8] Wolfgang Becker and Christoph Feichtinger. "Data Analytics in jungen Unternehmen: Ergebnisse einer Online-Befragung". In: *Der Betriebswirt* 59.2 (2018), pp. 26–31. ISSN: 0172-6196. DOI: 10.3790/dbw.59.2.26.

[9] Shirley Coleman et al. "How Can SMEs Benefit from Big Data? Challenges and a Path Forward". In: *Quality and Reliability Engineering International* 32.6 (2016), pp. 2151–2164. ISSN: 07488017. DOI: 10.1002/qre.2008.

[10] M. Ghobakhloo et al. "Information Technology Adoption in Small and Medium-sized Enterprises: An Appraisal of Two Decades Literature". In: *Interdisciplinary Journal of Research Business* 1.7 (2011), pp. 53–80. URL: https://www.semanticscholar.org/paper/Information-Technology-Adoption-in-Small-and-An-of-Ghobakhloo-Sabouri/168756609f5c36d8476f98d8347bd13d962b1a91.

[11] G. Vossen, J. Lechtenbörger, and D. Fekete. "Big Data in kleinen und mittleren Unternehmen: Eine empirische Bestandsaufnahme (Nr. 135)". In: *Arbeitsberichte des Instituts für Wirtschaftsinformatik, Westfälische Wilhelms-Universität Münster* (2015). URL: https://www.wi.uni-muenster.de/sites/wi/files/public/research/arbeitsberichte/ab135.pdf.

[12] Markus Bauer, Clemens van Dinther, and Daniel Kiefer. "Machine Learning in SME: An Empirical Study on Enablers and Success Factors". In: *AMCIS 2020 Proceedings* (2020). URL: https://aisel.aisnet.org/amcis2020/adv_info_systems_research/adv_info_systems_research/3.

[13] Volker Zimmermann. *Unternehmensbefragung 2017: Digitalisierung der Wirtschaft: breite Basis, vielfältige Hemmnisse*. Ed. by KfW Bankengruppe. 2017. URL: https://www.kfw.de/PDF/Download-Center/Konzernthemen/Research/PDF-Dokumente-Unternehmensbefragung/Unternehmensbefragung-2017-%E2%80%93-Digitalisierung.pdf (visited on 05/07/2021).

[14] A. Moeuf et al. "Industry 4.0 and the SME: a technology-focused review of the empirical literature". In: *Proceedings of 7th International Conference on Industrial Engineering and Systems Management IESM* (2017). URL: https://hal.science/hal-01836173v1/document.

[15] Silvia Russegger et al. "Big Data und Data-driven Business für KMU". In: *Digital networked Data* (2015). URL: https://www.salzburgresearch.at/publikation/big-data-und-data-driven-business-fuer-kmu/.

[16] Burkhardt Funk, Matthias Rettenmeier, and Tobias Lang. "Deep Learning auf sequenziellen Daten als Grundlage unternehmerischer Entscheidungen". In: *Wirtschaftsinformatik & Management* 9.5 (2017), pp. 16–25. ISSN: 1867-5913. DOI: 10.1007/s35764-017-0104-4. URL: https://link.springer.com/article/10.1007%2Fs35764-017-0104-4.

[17] Ulrich Kerzel. "Enterprise AI Canvas Integrating Artificial Intelligence into Business". In: *Applied Artificial Intelligence* 35.1 (2021), pp. 1–12. ISSN: 0883-9514. DOI: 10.1080/08839514.2020.1826146.

[18] Franziska Schäfer et al. "Smart Use Case Picking with DUCAR: A Hands-On Approach for a Successful Integration of Machine Learning in Production Processes". In: *Procedia Manufacturing* 51 (2020), pp. 1311–1318. ISSN: 2351-9789. DOI: 10.1016/j.promfg.2020.10.183. URL: https://www.sciencedirect.com/science/article/pii/S2351978920320400.

[19] Felix Kuschicke et al. "A Data-based Method for Industrial Big Data Project Prioritization". In: *Proceedings of the International Conference on Big Data and Internet of Thing* (2017), pp. 6–10. DOI: 10.1145/3175684.3175687.

[20] Bill Schmarzo. *Big data MBA: Driving business strategies with data science*. Indianapolis, Indiana: Wiley, 2016. ISBN: 9781119181118.

[21] Adrian Bourcevet et al. "Guided Machine Learning for Business Users". In: *BLED 2019 Proceedings* (2019). URL: https://aisel.aisnet.org/bled2019/47.

[22] Qian Yang et al. “Grounding Interactive Machine Learning Tool Design in How Non-Experts Actually Build Models”. In: *Proceedings of the 2018 Designing Interactive Systems Conference* (2018). URL: https://www.semanticscholar.org/paper/Grounding-Interactive-Machine-Learning-Tool-Design-Yang-Suh/8d4b89081ba250f2e35585af1c5eb8d764b11ab9.

[23] Louis Dorard. *The Machine Learning Canvas: A handbook for innovators and visionary managers striving to design tomorrow's Machine Learning systems: DRAFT version 0.1 released on 12 February 2019*. 2019. URL: https://www.ownml.co/machine-learning-canvas.

[24] Christoph Kollwitz, Maximilian Perez Mengual, and Barbara Dinter. “Cross-Disciplinary Collaboration for Designing Data-Driven Products and Services”. In: *Pre-ICIS SIGDSA Symposium on Decision Analytics Connecting People, Data & Things, San Francisco 2018* (2018). URL: https://aisel.aisnet.org/sigdsa2018/11/.

[25] S. L. Star and J. Griesemer. “Institutional Ecology, ‘Translations’ and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39”. In: *Social Studies of Science* 19.3 (1989), pp. 387–420. URL: https://www.semanticscholar.org/paper/Institutional-Ecology%2C-%60Translations%27-and-Boundary-Star-Griesemer/d5216326373656d42b9d08df8bda2a9b8cb3c95a.

[26] Alexander Osterwalder and Yves Pigneur. *Business model generation: A handbook for visionaries, game changers, and challengers*. New York: Wiley&Sons, 2010. ISBN: 9780470876411.

[27] P. Hartmann et al. “Big Data for Big Business? A Taxonomy of Data-driven Business Models used by Start-up Firms”. In: *Cambrigde Service Alliance* (2014). URL: https://www.semanticscholar.org/paper/Big-Data-for-Big-Business-A-Taxonomy-of-Data-driven-Hartmann-Zaki/2e6fc4039125c3151d885845bb8da7869a6f4322.

[28] Barbara Kitchenham. *Procedures for Performing Systematic Reviews: Keele University Technical Report TR/SE-0401*. 2004. URL: https://www.bibsonomy.org/bibtex/75c82aef0bd6a41e833647512d5e78d6.

[29] J. Webster and R. T. Watson. “Analyzing the past to prepare for the future: Writing a literature review”. In: *MIS Quarterly* 26.2 (2002), pp. 13–23. ISSN: 02767783. URL: https://www.jstor.org/stable/4132319?seq=1.

[30] Guido Schryen. “Writing Qualitative IS Literature Reviews—Guidelines for Synthesis, Interpretation, and Guidance of Research”. In: *Communications of the Association for Information Systems* 37.1 (2015). ISSN: 1529-3181. DOI: 10.17705/1CAIS.03712. URL: https://aisel.aisnet.org/cais/vol37/iss1/12.

[31] J. vom Brocke et al. “Reconstructing the giant: on the importance of rigour in documenting the literature search process”. In: *Proceedings of the 17th European Conference on Information Systems*. 2009. URL: https://aisel.aisnet.org/ecis2009/161/.

[32] Stefan Kehrer, Dirk Jugel, and Alfred Zimmermann. “A systematic literature review of big data literature for EA evolution”. In: *Digital Enterprise Computing 2016, Lecture Notes in Informatics (LNI)* (2016). URL: https://dl.gi.de/handle/20.500.12116/955.

[33] Thorsten Prof. Dr. Hennig-Thurau and Henrik Prof. Dr. Sattler. *VHB-JOURQUAL3*. 2015. URL: https://www.vhbonline.org/fileadmin/vhb/Services/vhb-rating/JOURQUAL/JOURQUAL3_Gesamtliste.pdf.

[34] Richard Vidgen, Giles Hindle, and Ian Randolph. "Exploring the ethical implications of business analytics with a business ethics canvas". In: *European Journal of Operational Research* 281.3 (2020), pp. 491–501. ISSN: 03772217. DOI: 10.1016/j.ejor.2019.04.036.

[35] Babett Kühne and Tilo Böhmann. "Data-Driven Business Models - Building the Bridge between Data and Value". In: *Twenty-Seventh European Conference on Information Systems (ECIS2019), Stockholm-Uppsala, Sweden.* (2019). URL: https://www.semanticscholar.org/paper/Data-Driven-Business-Models-Building-the-Bridge-and-K%C3%BChne-B%C3%B6hmann/d68a5d6d7873510002dc26f251088ac6da29bc50.

[36] Zhibin Zhou et al. "ML-Process Canvas: A Design Tool to Support the UX Design of Machine Learning-Empowered Products". In: *CHI'19 Extended Abstracts, 2019, Glasgow, Scotland UK* (2019). URL: https://www.semanticscholar.org/paper/ML-Process-Canvas%3A-A-Design-Tool-to-Support-the-UX-Zhou-Gong/ac196807c94cb66d76e69f75d38a6fc4cb282956.

[37] Bill Schmarzo. *Data Science "Paint by the Numbers" with the Hypothesis Development Canvas*. 2018. URL: https://www.linkedin.com/pulse/data-science-paint-numbers-hypothesis-development-canvas-schmarzo/ (visited on 04/27/2021).

[38] Katrin Mathis and Felix Köbler. "Data-Need Fit – Towards Data-Driven Business Model Innovation". In: *ServDes. 2016, Fifth Service Design and Innovation conference* (2016). URL: https://www.semanticscholar.org/paper/Data-Need-Fit-%E2%80%93-Towards-Data-Driven-Business-Model-Mathis-K%C3%B6bler/e284913e68fbd7d5955c9b17cff80fe6177e1cbc.

[39] David Sammon and Tadhg Nagle. "The Data Value Map: A framework for developing shared understanding on data initiatives". In: *ECIS 2017: 25th European Conference on Information Systems* (2017), pp. 1439–1452. URL: https://cora.ucc.ie/handle/10468/5167.

[40] Andreas Heberle et al. "Digitalization Canvas - Towards Identifying Digitalization Use Cases and Projects". In: *Journal of Universal Computer Science* 23.11 (2017), pp. 1070–1097. URL: http://www.jucs.org/jucs_23_11/digitalization_canvas_towards_identifying/jucs_23_11_1070_1097_heberle.pdf.

[41] Ajay Agrawal, Avi Goldfarb, and Joshua Gans. *A Simple Tool to Start Making Decisions with the Help of AI*. Ed. by Harvard Business Review Cases. 2018. URL: https://hbr.org/2018/04/a-simple-tool-to-start-making-decisions-with-the-help-of-ai (visited on 04/16/2021).

[42] K. Dewalt and R. Rands. *Become an AI Company in 90 Days: The No-Bullshit Guide for Understanding AI, Identifying Opportunities, and Launching Your First Product*. Prolego, 2018. ISBN: 9780692192337. URL: https://books.google.de/books?id=esb5uwEACAAJ.

[43] Tizian Kronsbein and Roland Mueller. "Data Thinking: A Canvas for Data-Driven Ideation Workshops". In: *Hawaii International Conference on System Sciences 2019 (HICSS-52)* (2019). URL: https://aisel.aisnet.org/hicss-52/cl/visual_tools/4.

[44] Liza Kayser, Roland Mueller, and Tizian Kronsbein. "Data Collection Map: A Canvas for Shared Data Awareness in Data-Driven Innovation Projects". In: *Proceedings of the 2019 Pre-ICIS SIGDSA Symposium* (2019). URL: https://aisel.aisnet.org/sigdsa2019/18.

[45] Jan Zawadzki. *Introducing the AI Project Canvas - Towards Data Science*. 2019. URL: https://medium.com/data-science/introducing-the-ai-project-canvas-e88e29eb7024 (visited on 04/15/2021).

[46] Michael Engel and Fabian Lang. "A Pilot Study on Designing a Data & AI Performance Canvas". In: *AMCIS 2020 Proceedings* (2020). URL: https://aisel.aisnet.org/amcis2020/data_science_analytics_for_decision_support/data_science_analytics_for_decision_support/9.

[47] Michael Fruhwirth, Gert Breitfuss, and Viktoria Pammer-Schindler. "The Data Product Canvas - A Visual Collaborative Tool for Designing Data-Driven Business Models". In: *BLED 2020 Proceedings* (2020). URL: https://aisel.aisnet.org/bled2020/8.

[48] Fabian Hunke, Stefan Seebacher, and Hauke Thomsen. "Please Tell Me What to Do – Towards a Guided Orchestration of Key Activities in Data-Rich Service Systems". In: *Designing for Digital Transformation. Co-Creating Services with Citizens and Industry*. Ed. by Sara Hofmann, Oliver Müller, and Matti Rossi. Information Systems and Applications, incl. Internet/Web, and HCI. Cham: Springer International Publishing and Imprint: Springer, 2020, pp. 426–437. ISBN: 978-3-030-64823-7. URL: https://link.springer.com/book/10.1007/978-3-030-64823-7.

[49] Ralf T. Kreutzer and Marie Sirrenberg. "AI Challenge - How Artificial Intelligence Can Be Anchored in a Company". In: *Understanding Artificial Intelligence*. Ed. by Ralf T. Kreutzer and Marie Sirrenberg. Management for Professionals. Cham: Springer International Publishing and Imprint: Springer, 2020, pp. 235–273. ISBN: 978-3-030-25271-7. DOI: 10.1007/978-3-030-25271-7_10.

[50] Zhibin Zhou et al. "ML Lifecycle Canvas: Designing Machine Learning-Empowered UX with Material Lifecycle Thinking". In: *Human–Computer Interaction* 35.5-6 (2020), pp. 362–386. ISSN: 0737-0024. DOI: 10.1080/07370024.2020.1736075.

[51] Ivan Marin. "Data Science and Development Team Remote Communication: the use of the Machine Learning Canvas". In: *2019 ACM/IEEE 14th International Conference on Global Software Engineering (ICGSE 2019)*. Piscataway, NJ: IEEE, 2019, pp. 18–21. ISBN: 978-1-5386-9196-0. DOI: 10.1109/ICGSE.2019.00018.

[52] Google. *People + AI Guidebook*. 2019. URL: https://pair.withgoogle.com/guidebook/ (visited on 04/16/2021).

[53] Saleema Amershi et al. "Guidelines for Human-AI Interaction". In: *CHI 2019*. Ed. by Stephen Brewster et al. New York, NY: ACM, 2019, pp. 1–13. ISBN: 9781450359702. DOI: 10.1145/3290605.3300233.

[54] Christian Weber et al. “A New Process Model for the Comprehensive Management of Machine Learning Models”. In: *Proceedings of the 21st International Conference on Enterprise Information Systems ICEIS 2019* (2019), pp. 415–422. DOI: 10.5220/0007725304150422.

[55] Mauro Ribeiro, Katarina Grolinger, and Miriam A.M. Capretz. “MLaaS: Machine Learning as a Service”. In: *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2015, pp. 896–902. ISBN: 978-1-5090-0287-0. DOI: 10.1109/ICMLA.2015.152.

[56] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge: Cambrige University Press, 2014. ISBN: 9781107057135. DOI: 10.1017/CBO9781107298019.

[57] DCMI. *DCMI Metadata Terms*. 2020. URL: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/ (visited on 05/13/2021).

[58] Rüdiger Wirth and Jochen Hipp. “CRISP-DM: Towards a standard process model for data mining”. In: *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining* (2000). URL: https://cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf.

# Paper 2

# Developing an Ontology for Data Science Projects to facilitate the Design Process of a Canvas

## Outline



## Bibliographic Information


Lukas-Walter Thiée.
Leuphana Universität Lüneburg, Institute for Information Systems, Lüneburg, Germany.
21.02.2022, https://aisel.aisnet.org/wi2022/ai/ai/13/
17th International Conference on Wirtschaftsinformatik, February 2022, Nürnberg, Germany.
Best Short Paper Award: 1st Runner-Up. https://wi22.de/en/awards-2/


## Copyright Notice

# Abstract

Data science projects can become very complex, due to the complexity of their content, but also due to the nature and composition of their stakeholders. There are several approaches to remedy this, e.g., canvases, which support ideation and common understanding. However, previous approaches are limited to single details or abstract too much, so that it is difficult to carry out entire projects successfully based on them. This paper describes one part of the design process, namely the derivation of the underlying ontology, of a new canvas that integrates both the overall project and detail steps. The ontology is mainly derived from CRISP-DM, literature review and project work.



**Keywords**

Data Science Project, Machine Learning, Canvas, Ontology

# 1 Introduction

With the advent of powerful local and cloud hardware, open source software, and even online collaboration tools for big data projects, e.g., Google Colab, data science (DS) should be within reach for all the interested. However, machine learning (ML) and artificial intelligence (AI) are still elusive topics for many companies and individuals. These Information Systems (IS) topics, nevertheless, are not only important drivers for the optimization of existing products and business processes, but also for the (digital) transformation and foundation of companies. So far-reaching, in fact, that data has been called a new commodity [1]. In order to take part in data science, organizations start their own data-driven projects and due to the complexity of these projects they *"need clear and structured guidance at the beginning of the innovation process to formulate and communicate business ideas with data"* [2]. Therefore, standard processes have been suggested [3], maturity models have been developed [4], and competency profiles have also been proposed [5]. Yet, there is no generally accepted definition of DS and the discipline's practice. A major obstacle in the projects is the high spread of data literacy among the stakeholders, which hinders the process of value generation [6]. One approach to make the entry and execution of complex projects easier are canvases, such as the Business Model Canvas [7]. Canvases are boundary objects to facilitate teamwork and generate common understanding of complex topics. For the overarching field of DS, several canvas approaches have been proposed in recent years, for example the 'Machine Learning Canvas' [8], the 'Key Activity Canvas' [9], or the 'Data Value Map' [10]. These tools offer support in various phases of a DS project. Most approaches focus on identifying potential before project start. However, it remains unclear how to get from this ideation to concrete tasks in the course of the project. The challenge is that most approaches are structured in such a way that they either address a specific part, e.g., ideation, thus ignoring the overall project, or they look at the overall project very generically, e.g., project justification and budgeting, so that the necessary level of detail is not achieved. Accordingly, an approach that integrates the standard process, i.e., CRISP-DM [3], is missing. Therefore, the question underlying this research project is *"What does a joint working tool, i.e., canvas, need to look like that supports teams during initial and subsequent tasks of ML projects in line with standard data science processes?"* The goal is to design a canvas for DS projects, especially suitable for small organizations. The contribution of this research project is in combining both an integrative view of the overall project and an appropriate level of detail in the individual sections, thus addressing the dichotomy between holistic and compact. This paper presents the derivation of the design requirements and the development of the underlying ontology to facilitate the design process of a canvas.

# 2 Related Work and Methodology

The approach of using a canvas to facilitate the development of ML or AI solutions has been prominent in IS research in recent years. At its core, most contributions try to support (parts of) the process from data exploration and ideation to a concrete business value. Four categories of such canvases with different thematic foci, namely ML/DS, (AI) Project, Data Value, and Data Source, have been identified in prior work [11], as shown in Table 2.1.

Table 2.1: Canvas artifacts with different foci

| Focus | Year | Source | Canvas Artifact |
|---|---|---|---|
| ML/DS | 2018 | [12] | The ML Canvas (Big Data MBA Version) |
| ML/DS | 2018 | [12] | Hypothesis Development Canvas v1.1 |
| ML/DS | 2019 | [8] | Machine Learning Canvas v0.4 |
| ML/DS | 2020 | [9] | Key Activity Canvas |
| ML/DS | 2020 | [13] | ML Lifecycle Canvas |
| (AI) Project | 2017 | [14] | Digitalization Canvas |
| (AI) Project | 2018 | [15] | AI Canvas |
| (AI) Project | 2018 | [16] | AI Canvas |
| (AI) Project | 2019 | [17] | AI Project Canvas |
| (AI) Project | 2020 | [18] | AI performance canvas (prototype) |
| (AI) Project | 2020 | [19] | Canvas for the use of AI |
| (AI) Project | 2021 | [20] | Enterprise AI Canvas |
| Data Value | 2016 | [21] | Data Canvas: Data-Need Fit |
| Data Value | 2017 | [10] | Data Value Map |
| Data Value | 2019 | [6] | Data Innovation Board |
| Data Value | 2020 | [2] | Data Product Canvas |
| Data Source | 2019 | [22] | Data Collection Map |
| Data Source | 2019 | [23] | Data Insight Generator |

All of these practice and scientific artifacts are intended as initial tools for generating ideas and/or as a communication platform at the beginning of DS projects. Although many of the approaches include parts of standard processes, such as the 'Data Preparation' phase, there is no explicit alignment between canvas and project progress, i.e., subsequent tasks after project initiation. The literature review preceding this work posed the research question *"Which canvas models, that address ML or AI implementation, are available, and which contents do they cover?"* The answering of the latter resulted in a catalog of 163 fields with 287 (non-exclusive) questions categorized in a total of 11 categories and 39 subgroups. On the one hand this catalog with its categories is a good starting point for DS projects, on the other hand it also shows that there are still areas that are underrepresented, such as the connection between data preparation and modeling. Thus, the review is not only suitable as a basis for the following design part of an own canvas, but also shows where a new content focus must be set. The methodological approach of this research follows the design science paradigm, which at its core seeks to create useful (IS) artifacts through creative problem-solving techniques, thereby enhancing the scientific corpus and practical utility [24]. Since design science is meant to solve an observed (organizational) problem [25], we formulate the problem statement in two parts: (1) Existing canvases and process models are not aligned, which makes it hard for organizations to use them coherently and (2) existing canvases either only focus on parts of the whole project or lack a level of detail, when they take an abstract view on the whole project, which makes it difficult to get to successful solutions. Wirth and Hipp have already addressed this dilemma between detailed (exhaustiveness) and generic (parsimony) process descriptions [3]. We therefore propose the design requirements for the artifact: The canvas should be exhaustive, in order to address the whole project, and

provide the right level of detail in order to be useful, while at the same time the canvas should be kept as simple as possible to provide ease-of-use and common understanding. Optionally, as the canvas might be too complex for a paper based version and workshop, a digital tool could be the preferred solution [26], which would demand the inclusion of user-centered design. Since various stakeholders, e.g., data scientists, managers, domain experts, and IT specialists, are usually involved in DS projects and the project itself deals with complex topics, two main issues, namely collaboration/communication and structuring of tasks, have to be managed. Avdiji et al. (2018) [26] propose IS *"design principles for tools that both support collaboration and are tailored for specific ill-structured problems"*. These principles include *"(1) framing the ill-structured problem by developing an ontology, (2) representing the ontology into a shared visualization, and (3) instantiating the visualization in a way that supports shared prototyping of the solution"* [22]. We build upon these guidelines in the design process. In the following the derivation of the ontology is being described.

# 3 Developing the Ontology

The first step in the design of the future artifact, is the development of an ontology. In computer science an ontology is a representation of entities and the relationship between these entities in a specific subject area. It can be understood as reference model [27] and helps reducing *"conceptual and lexical confusion by providing a unifying framework within an organization"* [26], thereby sharpening problem understanding. The field of DS encompasses a wide range of disciplines and skills, e.g. computer science, programming, statistics, or data management. In order to make this complex term more tangible, attempts have already been made to develop an ontology for DS, e.g., the data science ontology [28]. This ontology indexes various concepts from the DS discipline as well as annotations of commonly used software libraries. However, the relationship between the elements is not evident from the index. The context and usefulness of the integrated software libraries can only be understood with appropriate prior knowledge. *"Its long-term objective is to improve the efficiency and transparency of collaborative, data-driven science."*[1] However, this publicly available ontology does not lay claim to completeness, rather it is a living and editable online document. We must therefore use other references.

For this reason, we integrate three main sources to initialize the entities in our ontology (Figure 2.1). We use selected entities from these sources and then build the relationships between these entities. First of all, we utilize the CRISP-DM process stages (I) [3], as they are fundamental to DS. These stages include the entities 'Business Understanding', 'Data Understanding', 'Data Preparation', 'Modeling', 'Evaluation', and 'Deployment' Then we integrate two sources from prior work, namely the results of the aforementioned literature review regarding ML canvases (II) [11], and an item list of topics and questions, which is a part of project work regarding the development of a DS process model (III) [29]. Our approach ensures, that on the one hand the ontology is built on a tested scientific artifact, as CRISP-DM can be seen as a fundamental basis for DS projects, and on the other hand, both a content focus on DS and ML as well as actuality are taken into account. Additionally, we consider enhancements of the standard process, such as CRISP-ML(Q), which integrates quality assurance in ML projects [30]. Exemplarily, we

[1] https://www.datascienceontology.org/about

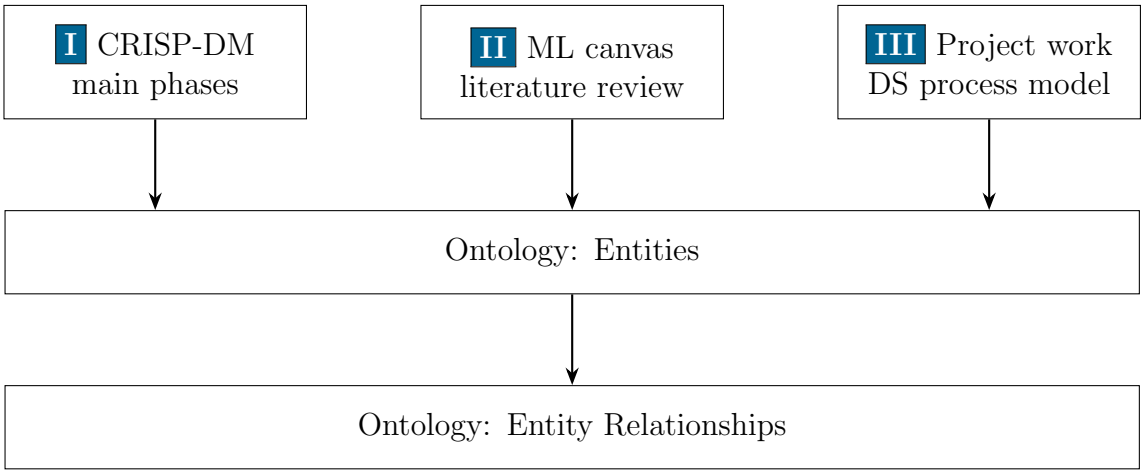


Figure 2.1: Process and references for the development of the ontology

describe the main path of the ontology (Figure 2.2), which was taken from CRISP-DM (blue): 'Business Understanding' enables 'Data Understanding', 'Data Understanding' in turn supports 'Business Understanding' and is simultaneously the basis for 'Data Preparation'. 'Modeling' requires 'Data Preparation' and is assessed by the 'Evaluation', which in turn influences the 'Deployment', and recursively updates the 'Business Understanding'. The ontology is structured in such a way that the main entities are composed of sub-elements, e.g. (shaded), 'Business Understanding' is composed of 'Business Key Performance Indicators' (KPI), which in turn are derived from 'Customer', 'Financial', 'Product', 'Organizational', and 'Technological Understanding'. The interlinking of the sub-elements results in a web structure, which reflects the iterative nature of DS projects, and CRISP-DM respectively. Another central aspect of DS projects is captured in the ontology, namely the paths between 'Data Quality' and 'Model Training' (red). Extent and kind of the entire data preparation is substantially dependent on the selection and programming of an appropriate ML/AI algorithm, vice versa. The estimator selection in turn affects the model training and tuning. Therefore, depending on project maturity, different entities have to be incorporated. The ontology in Figure 2.2 represents an interim result and contribution of our research. It provides a holistic overview of a DS project.

# 4 Conclusion and Future Research

In this article we call for the design of a new canvas for DS projects that is both holistic and compact, as previous approaches either address only partial aspects or are too generic. The artifact should support (small) organizations not only in generating ideas , but also in supporting the overall project. The design of the canvas is to be done in three steps, defining the ontology, designing a shared visualization and initializing the canvas. In this paper, the derivation of the ontology is presented as an intermediate result and contribution of our research. Future research consequently involves finalizing the design process, integrating principles of user-centered design [31], and evaluating the artifact. The evaluation is to be done essentially through qualitative methods, e.g., workshops and case studies, as is common in design science projects, and refers mainly to plausibility, usability, and perceived usefulness of the artifact.

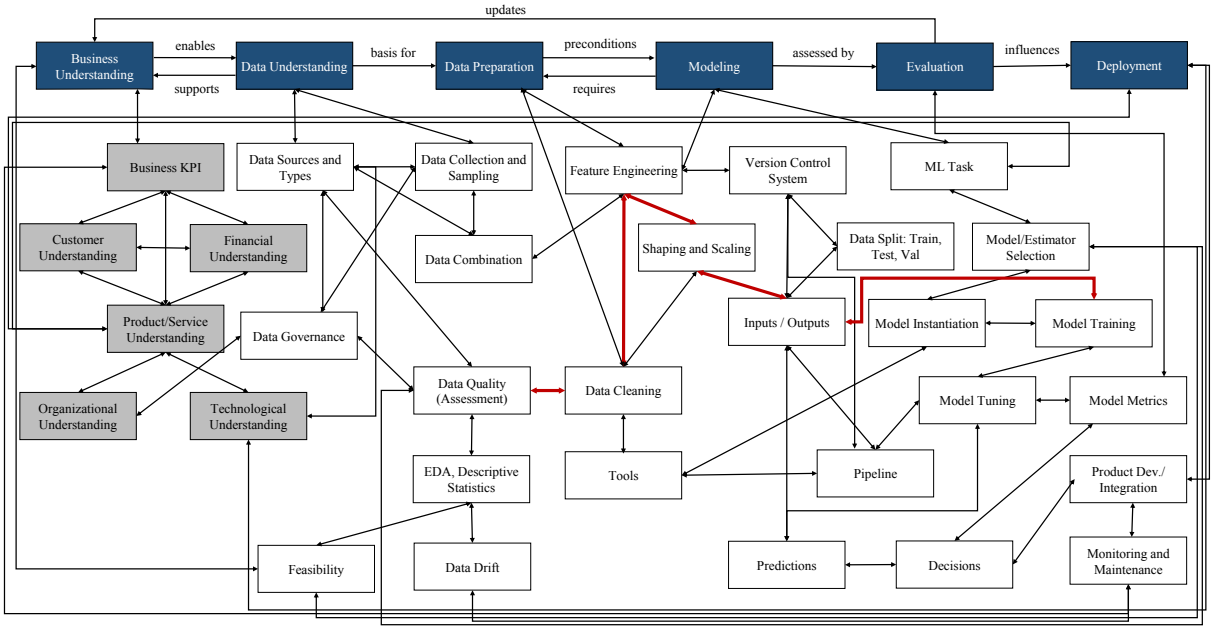


Figure 2.2: Simplified view of the ontology of a data science project (see Appendix C.1)

# References


[1] The Economist. *The world's most valuable resource is no longer oil, but data*. 2021. URL: https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data (visited on 08/30/2021).

[2] Michael Fruhwirth, Gert Breitfuss, and Viktoria Pammer-Schindler. "The Data Product Canvas - A Visual Collaborative Tool for Designing Data-Driven Business Models". In: *BLED 2020 Proceedings* (2020). URL: https://aisel.aisnet.org/bled2020/8.

[3] Rüdiger Wirth and Jochen Hipp. "CRISP-DM: Towards a standard process model for data mining". In: *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining* (2000). URL: https://cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf.

[4] Sulaiman Alsheiabni, Yen Cheung, and Chris Messom. "Towards An Artificial Intelligence Maturity Model: From Science Fiction To Business Facts". In: *PACIS 2019 Proceedings* (2019). URL: https://aisel.aisnet.org/pacis2019/46.

[5] Foster Provost and Tom Fawcett. "Data Science and its Relationship to Big Data and Data-Driven Decision Making". In: *Big data* 1.1 (2013), pp. 51–59. ISSN: 2167-6461. DOI: 10.1089/big.2013.1508.

[6] Tizian Kronsbein and Roland Mueller. "Data Thinking: A Canvas for Data-Driven Ideation Workshops". In: *Hawaii International Conference on System Sciences 2019 (HICSS-52)* (2019). URL: https://aisel.aisnet.org/hicss-52/cl/visual_tools/4.

[7] Alexander Osterwalder and Yves Pigneur. *Business model generation: A handbook for visionaries, game changers, and challengers*. New York: Wiley&Sons, 2010. ISBN: 9780470876411.

[8] Louis Dorard. *The Machine Learning Canvas: A handbook for innovators and visionary managers striving to design tomorrow's Machine Learning systems: DRAFT version 0.1 released on 12 February 2019*. 2019. URL: https://www.ownml.co/machine-learning-canvas.

[9] Fabian Hunke, Stefan Seebacher, and Hauke Thomsen. "Please Tell Me What to Do – Towards a Guided Orchestration of Key Activities in Data-Rich Service Systems". In: *Designing for Digital Transformation. Co-Creating Services with Citizens and Industry*. Ed. by Sara Hofmann, Oliver Müller, and Matti Rossi. Information Systems and Applications, incl. Internet/Web, and HCI. Cham: Springer International Publishing and Imprint: Springer, 2020, pp. 426–437. ISBN: 978-3-030-64823-7. URL: https://link.springer.com/book/10.1007/978-3-030-64823-7.

[10] David Sammon and Tadhg Nagle. "The Data Value Map: A framework for developing shared understanding on data initiatives". In: *ECIS 2017: 25th European Conference on Information Systems* (2017), pp. 1439–1452. URL: https://cora.ucc.ie/handle/10468/5167.

[11] Lukas-Walter Thiée. "A systematic literature review of machine learning canvases". In: *Informatik 2021*. GI-edition Proceedings. Bonn: Gesellschaft für Informatik e.V. (GI), 2021, pp. 1221–1235. ISBN: 9783885797081. DOI: 10.18420/informatik2021-101. URL: https://doi.org/10.18420/informatik2021-101.

[12] Bill Schmarzo. *Data Science "Paint by the Numbers" with the Hypothesis Development Canvas*. 2018. URL: https://www.linkedin.com/pulse/data-science-paint-numbers-hypothesis-development-canvas-schmarzo/ (visited on 04/27/2021).

[13] Zhibin Zhou et al. "ML Lifecycle Canvas: Designing Machine Learning-Empowered UX with Material Lifecycle Thinking". In: *Human–Computer Interaction* 35.5-6 (2020), pp. 362–386. ISSN: 0737-0024. DOI: 10.1080/07370024.2020.1736075.

[14] Andreas Heberle et al. "Digitalization Canvas - Towards Identifying Digitalization Use Cases and Projects". In: *Journal of Universal Computer Science* 23.11 (2017), pp. 1070–1097. URL: http://www.jucs.org/jucs_23_11/digitalization_canvas_towards_identifying/jucs_23_11_1070_1097_heberle.pdf.

[15] Ajay Agrawal, Avi Goldfarb, and Joshua Gans. *A Simple Tool to Start Making Decisions with the Help of AI*. Ed. by Harvard Business Review Cases. 2018. URL: https://hbr.org/2018/04/a-simple-tool-to-start-making-decisions-with-the-help-of-ai (visited on 04/16/2021).

[16] K. Dewalt and R. Rands. *Become an AI Company in 90 Days: The No-Bullshit Guide for Understanding AI, Identifying Opportunities, and Launching Your First Product*. Prolego, 2018. ISBN: 9780692192337. URL: https://books.google.de/books?id=esb5uwEACAAJ.

[17] Jan Zawadzki. *Introducing the AI Project Canvas - Towards Data Science*. 2019. URL: https://medium.com/data-science/introducing-the-ai-project-canvas-e88e29eb7024 (visited on 04/15/2021).

[18] Michael Engel and Fabian Lang. "A Pilot Study on Designing a Data & AI Performance Canvas". In: *AMCIS 2020 Proceedings* (2020). URL: https://aisel.aisnet.org/amcis2020/data_science_analytics_for_decision_support/data_science_analytics_for_decision_support/9.

[19] Ralf T. Kreutzer and Marie Sirrenberg. "AI Challenge - How Artificial Intelligence Can Be Anchored in a Company". In: *Understanding Artificial Intelligence*. Ed. by Ralf T. Kreutzer and Marie Sirrenberg. Management for Professionals. Cham: Springer International Publishing and Imprint: Springer, 2020, pp. 235–273. ISBN: 978-3-030-25271-7. DOI: 10.1007/978-3-030-25271-7_10.

[20] Ulrich Kerzel. "Enterprise AI Canvas Integrating Artificial Intelligence into Business". In: *Applied Artificial Intelligence* 35.1 (2021), pp. 1–12. ISSN: 0883-9514. DOI: 10.1080/08839514.2020.1826146.

[21] Katrin Mathis and Felix Köbler. "Data-Need Fit – Towards Data-Driven Business Model Innovation". In: *ServDes. 2016, Fifth Service Design and Innovation conference* (2016). URL: https://www.semanticscholar.org/paper/Data-Need-Fit-%E2%80%93-Towards-Data-Driven-Business-Model-Mathis-K%C3%B6bler/e284913e68fbd7d5955c9b17cff80fe6177e1cbc.

[22] Liza Kayser, Roland Mueller, and Tizian Kronsbein. "Data Collection Map: A Canvas for Shared Data Awareness in Data-Driven Innovation Projects". In: *Proceedings of the 2019 Pre-ICIS SIGDSA Symposium* (2019). URL: https://aisel.aisnet.org/sigdsa2019/18.

[23] Babett Kühne and Tilo Böhmann. "Data-Driven Business Models - Building the Bridge between Data and Value". In: *Twenty-Seventh European Conference on Information Systems (ECIS2019), Stockholm-Uppsala, Sweden.* (2019). URL: https://www.semanticscholar.org/paper/Data-Driven-Business-Models-Building-the-Bridge-and-K%C3%BChne-B%C3%B6hmann/d68a5d6d7873510002dc26f251088ac6da29bc50.

[24] Alan Hevner et al. "Design Science in Information Systems Research". In: *MIS Quarterly* 28.1 (2004), p. 75. ISSN: 02767783. DOI: 10.2307/25148625. URL: https://arizona.pure.elsevier.com/en/publications/design-science-in-information-systems-research.

[25] Ken Peffers et al. "A Design Science Research Methodology for Information Systems Research". In: *Journal of Management Information Systems* 24.3 (2007), pp. 45–77. ISSN: 0742-1222. DOI: 10.2753/MIS0742-1222240302.

[26] Hazbi Avdiji et al. *Designing Tools for Collectively Solving Ill-Structured Problems*. 2018. ISBN: 978-0-9981331-1-9. DOI: 10.24251/HICSS.2018.053. URL: http://hdl.handle.net/10125/49940.

[27] A. Osterwalder. "The business model ontology a proposition in a design science approach". In: *undefined* (2004). URL: https://www.semanticscholar.org/paper/The-business-model-ontology-a-proposition-in-a-Osterwalder/87bbedf0efbf010515ed54086bdf31c7cb33e4a3.

[28] Svetlana Chuprina, Vassil Alexandrov, and Nia Alexandrov. "Using Ontology Engineering Methods to Improve Computer Science and Data Science Skills". In: *Procedia Computer Science* 80 (2016), pp. 1780–1790. ISSN: 18770509. DOI: 10.1016/j.procs.2016.05.447.

[29] Michael Schulz et al. *DASC-PM v1.0: Ein Vorgehensmodell für Data-Science-Projekte*. Hamburg, Elmshorn, and Halle (Saale): valantic Business Analytics GmbH, Nordakademie gAG Hochschule der Wirtschaft, and Universitäts- und Landesbibliothek Sachsen-Anhalt, 2021. ISBN: 978-3-00-064898-4. DOI: 10.25673/32872.2. URL: https://opendata.uni-halle.de/handle/1981185920/33065.2.

[30] Stefan Studer et al. *Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology*. Mar. 11, 2020. URL: https://arxiv.org/pdf/2003.05155.

[31] Zahid Hussain, Wolfgang Slany, and Andreas Holzinger. "Investigating Agile User-Centered Design in Practice: A Grounded Theory Perspective". In: *HCI and Usability for e-Inclusion*. Ed. by David Hutchison et al. Vol. 5889. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 279–289. ISBN: 978-3-642-10307-0. DOI: 10.1007/978-3-642-10308-7_19.

# Paper 3

# Extraction of Information from Invoices – Challenges in the Extraction Pipeline

## Outline



## Bibliographic Information


Lukas-Walter Thiée, Felix Krieger and Burkhardt Funk.
Leuphana Universität Lüneburg, Institute for Information Systems, Lüneburg, Germany.
29.11.2023, https://doi.org/10.18420/inf2023_180
INFORMATIK 2023 - Designing Futures: Zukünfte gestalten. Joint Workshop IntDig 2023 MOC 2023; Intelligente Digitalisierung, (KI-basiertes) Management und Optimierung komplexer Systeme. Berlin. 26.-29. September 2023.


## Copyright Notice

# Abstract

Data from invoices are key information for business processes. In order to use the data and create business value, the information must be captured in a digital and structured form. Leveraging digital tools and AI/ML is state-of-the-art in the extraction of information from invoices. However, the existing approaches are trained on specific languages and layouts, and while focusing on the performance of individual metrics, they neglect the demonstration of the pipeline from raw data to processable information. In this paper, we investigate the types of information on invoices and address the challenges in the extraction pipeline. We contribute by providing a morphological framework for the problematization and design of a pipeline as part of a design science study.

**Keywords**

Invoice recognition, Information extraction, Data pipeline.

# 1 Introduction, Motivation and Method

Extraction of information from business documents is an evolving area of research and practice, as structured, digital information support numerous business processes. While we focus on invoices in this paper, the research approaches can be applied to other document types, such as receipts or checks. Digitizing incoming invoices, i.e., capturing structured information, can not only save a considerable amount of time, but can also add value. E.g., supply chain management can utilize these data by automatically integrating delivery dates and quantities into ERP systems [1]. In addition, structured invoice data enable business analytics, e.g., for purchasing patterns [2, 3]. Furthermore, auditing firms can leverage the data to simplify and enhance financial audits from sample testing to substantive test of details [4]. Despite the possibilities of the electronic creation, transfer and standardized integration of documents [1], it is still common practice today to send invoices on a paper or pdf basis, so that the information must be extracted from the document or file. This refers to both B2B and B2C invoicing processes. In contrast to digital invoice recognition human invoice reading (as well as annotation) is error-prone and costly with average *"processing costs of about 9 Euro"* [1]. Nevertheless, humans are good at the cognitive task of information extraction, i.e., infer abbreviations, link tabular data, and form composite information.

As in many fields of digitalization and research, methods from artificial intelligence (AI) and machine learning (ML) are increasingly being examined and used. The ultimate goal of this field of research is to digitally capture all (relevant) information from arbitrary business documents and make it available in a structured, processable format. In particular, the applications shall translate the raw data on a document into a machine-readable data type, so that the correct learning of the relationships in the data can be used to infer the original information. Existing industry solutions leverage these methods (see Paper 3 Chapter 2). However, the solutions are still far from comprehensive recognition of all relevant information, because invoices are designed in a plethora of layouts and languages. In addition, data protection concerns arise in the external processing of business documents, as sensitive information must not end up on external/foreign storages. Although the use of external services can provide access to (pretrained) models, cloud- and development environments, it creates a dependency that reduces both the ability to influence and the understanding of the model output. For these reasons – extraction quality, data privacy, and model dependency – there is a demand in research and practice to create and deploy proprietary models. However, two main difficulties become apparent when creating own models. On the one hand, the availability of a correspondingly large and qualitative/labeled dataset, which is necessary for training, validation, and testing, and on the other hand, the model selection and the effort for the corresponding data preparation and pipeline implementation. An extraction process usually consists of several steps which depend on each other. We call this process a pipeline. It describes the data flow of an ML/AI model from raw data over algorithmically processable to complete and relatable information, also referred to as 'Information Extraction' [5]. Since errors propagate within the pipeline, it is important to improve and optimally connect individual steps to ensure the best possible extraction of information, which strongly motivates this research. However, in order to create such an ML/AI pipeline, a wide variety of decision points and development stages must be passed through. General process descriptions such as CRISP-DM [6] or DASC-PM [7] are not sufficient in themselves for this

case. While they provide a good structure of general project stages and stakeholders, they don't offer practical guidelines that support the development and implementation of invoice recognition pipelines. Moreover, existing pipelines are not directly comparable, w.r.t. the deployed model, extracted classes and reported metrics, e.g., the F1-score in [8] refers to other classes than the F1-score in [9]. Furthermore, the capability of the models to generalize to unknown layouts and languages is limited, far from human capabilities, so that we speak of weak, or Artificial Narrow Intelligence (ANI) [10]. In order to support standardization in the field of document analysis, and invoice recognition in particular, and to promote practical solutions towards Artificial General Intelligence (AGI), the specific challenges must first be captured systematically. Related literature has neglected the systematization of the plethora of information and the specific challenges of the extraction, in particular in the context of different languages and layout conventions. We identify this as an open research gap and postulate that a comprehensive description of challenges and systematic categorization of information types can close this gap. The proposed framework can ultimately support extraction performance through better understanding of the data pipeline and the model in various countries/languages. Therefore, we address the following research questions: *What challenges exist in ML/AI pipelines in the context of invoice recognition? What kind of information types are present on invoice documents that can influence the extraction pipeline?*

**Research Goal and Method**. Our overall research goal is to improve invoice recognition pipelines, which covers the general phases of such projects, including data preparation, model selection, training, evaluation, and implementation. To achieve this goal we apply a Design Science Research (DSR) method, that helps us to generate practice-oriented artifacts [11, 12]. Hevner defines three inherent research cycles for a DSR project – the relevance, design and rigor cycle [13]. In this paper, we address the relevance cycle. At this early stage of our design study, we capture the details of the problem in order to design (software) artifacts in subsequent cycles. In our case, the derivation of the specific challenges and information types (Chapter 3), represents an intermediate result that can be used both in our design study and beyond, e.g., in the recognition of other business documents containing text, numbers, forms, or tables. The framework (Chapter 4) we present in this paper is the result of an initial design requirements cycle. It is a systematic catalog of challenges and information types encountered in invoice recognition, based on comprehensive review of literature and practice approaches (Chapter 2).

## 2 Review of Literature and Practice Approaches

Invoices are a characteristic type of business document, that represent proof of a business transaction, e.g., the purchase of goods or the provision of a service. *"Invoices contain always rather similar information, fostered by legal requirements for information items on invoices. However, the information items are distributed according to all different layout styles"* [1]. Besides different languages, country-specific conventions are often reflected in their appearance and content. In contrast to pure, sequential text, invoices represent a form of visually rich documents [14], on which various sources, especially semantics and 2-dimensional layout, contain the information. The plethora of information on invoices can range from dates and invoice numbers, over account and payment details, to product quantities and descriptions (line items), most of which is not arranged sequentially.

Since the information on invoices do not follow a predefined sequence, conventional natural language processing (NLP) techniques are difficult to apply [15]. Rather, meaning of words, their positions, and linguistic nuances, such as abbreviations and certain template conventions, form the superordinate signal basis of invoice information, e.g., a field for recipient address on the top left. To capture the information and to cope with the abundance of unstandardized data and layouts, a wide variety of approaches from the field of ML/AI have been presented in recent years. Therefore, we perform a literature review in relevant online (IS) libraries, such as AIS (19 results), IEEE (108 results), Web of Science (47 results), Scopus (194 results), SpringerLink (4 results) for literature from 2010 to 2022. We use combinations of the search terms 'invoice', 'recognition', 'information', and 'extraction', and enhance the results through forward and backward screening. Since we focus invoice document types, we do not include 'receipts' or similar search terms explicitly. We sort out duplicates and filter the results according to whether concrete ML/AI models and pipelines are described, and obtain 26 results for in-depth analysis. We classify the literature results according to the described model types and find, that in digital invoice recognition, a distinction can basically be made between three successive approaches: rule-based, conventional machine learning, and deep learning approaches [8]. Rule-based approaches to invoice recognition [1, 16, 17, 18] rely on consistent layouts or templates, defined patterns, and human input, which involves a lot of manual effort [8] that is difficult to scale, and therefore does not achieve the desired level of automation and cost reduction. ML-based approaches [19, 20, 21] integrate training based inference and NLP- based models include language models [22, 23].

Deep learning (DL) represents a subcategory of ML, characterized by the use of neural networks. For example, [22] apply a Long-Short-Term-Memory network to match eight different key-value-pairs from labeled sequences. Other approaches include end-to-end deep learning models [24], [25]. Convolutional Neural Networks (CNN) are applied, for example, in [26], but also in grid-based approaches, such as Chargrid [27], who represent invoice documents as a grid of characters and predict segmentation masks and bounding boxes for several key classes using a fully-convolutional encoder-decoder architecture. Other grid approaches include Wordgrid [28], BERTgrid [29], ViBERTgrid [30], or Tokengrid [31]. Another approach is presented by [9], who use representation learning to identify specific key-value-pairs on invoices. For each value element, candidates are selected based on predefined data types and a priori knowledge ("business rules"), which are then compared to the corresponding key via a multi-stage embedding.

In graph-based approaches with Graph Convolutional Networks (GCN), entities (usually words or tokens) are transformed into a graph structure in which the nodes have relationships to each other via edges [32, 14, 15]. Both nodes and edges can obtain features. [8] and [33] present transformer architectures. LayoutLM, for example, an extension of the BERT model [34], integrates 2D position (layout information) and image embeddings (visual information) in the document understanding task. Table 3.1 summarizes state-of-the-art DL approaches. The table illustrates, that while the approaches all address a similar extraction problem, they are difficult to compare against each other, w.r.t. the model, dataset, class labels and metrics used. For example, [15] report a macro average F1-score of 0.875 for the classes 'invoice date', 'invoice number' and 'total amount', excluding, and 0.905 including the 'no label' class. [9] report a macro-average F1-score of 0.878 for seven key items in their invoice dataset. [35] report F1-scores for six key items for two different datasets (Chinese, English), whereas [8] report an F1-score

of 0.952 for the SROIE[1] dataset. Evidently, the approaches deploy different pipelines. While this leads to limited comparability, they share the assumption that context plays an important role in the recognition of key information on invoices [15].

Table 3.1: Selection of deep learning approaches in invoice recognition

| Reference | Approach | Segmentation | Number of Classes: Keys |
|---|---|---|---|
| [15] | GCN/GAT | Entity-level node classification | 3: Invoice number, Invoice date, Total amount |
| [14] | GCN/GAT BiLSTM-CRF | Entity-level node classification | 6: Invoice number, Date, Price, Tax, Buyer, Seller |
| [32] | GCN | Word-level node classification | 27: Invoice number, Invoice date, Total without tax, Total tax amount, ... |
| [36] | LayoutLM | Token-level classification | FUNSD: 4 (question, answer, header, other) SROIE: 4 (company, date, address, total) |
| [8] | LayoutLMv2 | Token-level classification | CORD: 4: menu, void menu, subtotal, total Kleister NDA: 4: date, jurisdiction, party, term |
| [9] | Self-attention Network | Token-wise candidate similarity | 7: Amount due, Due date, Invoice date, Invoice ID, Purchase order, Total amount, Total tax amount |
| [27] | Chargrid | Character-wise segmentation | 8: Invoice number, Amount, Date, Vendor name, Address, Line item description, Quantity, Amount |
| [29] | BERTgrid | Character-wise segmentation | 6: Amount, Number, Date, Vendor name, LI mean, LI quantity |
| [31] | Tokengrid | 1D Anchor & Heatmap | 12: Vendor/Importer address & name, Freight, Insurance, Invoice date & number, Product, Prices |
| [24] | MFCN (fully conv. network) | Pixel-wise segmentation | 6 format segments (figure, table, section heading, caption, list, paragraph) |

**Practice Review**. The literature review shows state-of-the-art approaches from science, some of which are already being used in practice. Klein et al. (2004) [1] study various suppliers for automated invoice processing. However, we need to account for technical progress since 2004. Therefore, we perform a Google search with the aforementioned search terms. We find a large number of providers of document processing software, API services, and receipt scanning apps (e.g., 'scan2bank' services from banks), which we cannot present here all. *Klippa, Nanonets, Kofax, Super.AI, Rossum, MS Azure Form Recognizer or Veryfi* represent only a selection that offer intelligent invoice document processing. In the latter various key-value-pairs can be extracted through a Graph Neural Network. Google[2] offers a whole suite within their Document AI, e.g., *Docsumo*,

[1] Scanned Receipts OCR and Information Extraction Dataset: https://rrc.cvc.uab.es/?ch=13
[2] https://cloud.google.com/document-ai

or Human-in-the-loop training. Among other things, they provide services such as table recognition, optical character recognition (OCR) [37], and (non-finite) key-to-value matching. The aforementioned services are not open-source and not free of charge. We sample test documents, for providers that offer to test one file. From this (non-standardized) test no conclusions can be drawn on transferable performance. Generally, there is hardly any (scientific) record of the deployed pipelines and their performance available. That is, the details and performance of these approaches are difficult to validate.

# 3 Design Challenges in our Application Context

## 3.1 Information Types

Based on our literature review, we propose to distinguish seven different basic types of information on invoices, namely **segmental, syntactic, semantic, spatial, external, graphical**, and **logical** information, which are not equivalent to the type and number of classification labels, such as the class label 'total amount'. They rather represent the underlying signal of the data. The categorization was performed through systematic grouping of themes within the literature results. Segmentation refers to the level of abstraction at which data from invoices are processed. These levels can also be thought of as a hierarchy. On the lowest level, images are represented by pixels using the RGB encoding [38]. Depending on the language, individual symbols/characters/letters, words, and sentences constitute further levels of abstraction. Tokens differ from words in the sense that they may represent only a part of a word, e.g., the word stem, or even only a single character. Furthermore, paragraphs, rows, columns, and fields, if necessary, pages can be segmented. Syntactic information form an intermediate stage between segmentation and semantics. For example, they can be useful to label individual entities by their data type as string, digit, or alphanumeric. Upper and lower case or the number of letters or special characters can also be included as information. We refer to this category as syntax, because this information reflects rules for different entity types. For instance, a regular e-mail address requires three parts, a local name, an @ sign, and a server name with its top-level domain. These three parts therefore represent the syntax. The syntactic category also includes so-called tagging, e.g., POS (Part-of-Speech) [39], IOB- (Inside-Outside-Beginning) [40], or NER- (Named Entity Recognition) [41] tagging. While these tags can of course represent semantic information (logical semantics) [42], e.g., whether a word is tagged as a noun or a verb, they mostly provide structure (grammar respectively). In terms of semiotics, syntax stands for the sign (and the rules for their concatenation) and semantics stands for the norm, i.e., the conceptual interpretation of a sign or a series of signs [43]. Thus, we distinguish syntax from semantics by integrating so-called language or word models that transfer tokens, words, and sentences into a vector space (embedding), spanning a cognitive map [44]. This part of semantics (lexical semantics) [42] is concerned with the meaning of words and their relations. Semantically similar entities are located closer together in the vector space, indicated by their cosine similarity. Popular word models represent, for example, Word2Vec [45] or GloVe [46]. Another word model that has been increasingly used in recent years is BERT[3], which has been developed and open-sourced by Google. It is available in a wide variety of variants, e.g., RoBERTa [35] or DistilBERT

[3] https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

[47]. Its strength lies primarily in analyzing words in the context of sentences and thus extracting the semantic. An example of different meaning, that only arises in the context are homonyms, e.g., the word drop in the sentences *"I hope I don't drop my cup of coffee"* vs. *"I enjoyed every last drop of my coffee"*. In the invoicing context, subtle linguistic nuances and abbreviations emerge that are not necessarily used in everyday speech or prose, e.g., 'Inv.#.' could mean 'invoice number'. These semantic characteristics have to be accounted for. Whereas syntax and semantics can be found in the NLP context as well, 2D spatial information add a new information. The position can be integrated explicitly (Cartesian coordinates) or implicitly through pixel position. Absolute and relative position of elements like the coordinates of corners or vertices of bounding boxes represent a significant signal. In addition, center points, areas and shapes can be used. Distance measures between elements, i.e., the Euclidean distance, can be applied to establish positional neighborhood relations [32], e.g., the left, right, top and bottom neighbor of a word. Content neighborhoods are also conceivable through 'semantic linking' [8]. Graphical information are images such as logos, icons or product pictures, or visual elements such as table frames, boxes, arrows, dashed or dotted lines. Especially the latter can strongly contribute to the structure of the document. In addition, fonts, font styles and font sizes are also graphical information, e.g., an underlined, bold or highlighted element can have a special meaning in the document [37]. Also, white-space can convey information. External information represent knowledge which helps to interpret the content of invoices in the form of data bases [32] or knowledge graphs [9]. External information can be distinguished in content and format-oriented elements. Format-oriented information are conventions for certain entities like tax or account numbers (IBAN), which are often country specific, e.g., five-digit ZIP codes. Content oriented external information can be drawn from data bases or self-generated dictionaries on certain topics [32]. For example, order numbers from a company's ERP system. The final category here is logical information [24, 48]. They are a result of the context and/or the interaction of several information. To illustrate that, we investigate the token '4.07'. Syntactically it could be a decimal number or a date. Semantically it is ambiguous, it can express both a quantity or a currency amount. Only in combination of the currency sign ('$4.07' or '4,07€') the content is specified in this trivial example. Logic can also refer to calculations, e.g., the calculation of a gross amount from net amount and percentage tax amount. Another example of logic would be missing and multiple values, e.g., if an invoice states "billing address same as shipping address", we could logically infer to fill both keys with the same value.

## 3.2 Pipeline Challenges

The information types form the substantive basis for the generation of input features from raw data for ML/AI pipelines. Several challenges arise in the development of such a pipeline. For example, the correct classification of single elements does not equate to the retrieval of the full information. This means, that the pre- and post-processing of labels and entities belongs to the extraction pipeline, as label data and raw data are usually not stored in the same location (Figure 3.1). To elucidate this: the ground truth could be stored in a json file like this 'invoice date': '7/04/2022'.

The corresponding OCR could yield four independent tokens: 'July', '4', 'th', '2022'. Now, assuming correct positional data, we could assign the label 'invoice date' to all four

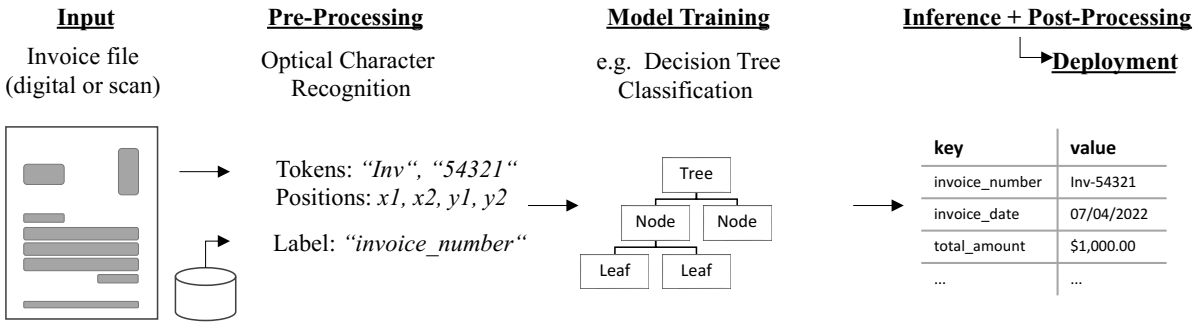


Figure 3.1: Exemplary pipeline of invoice information extraction, author's figure

tokens. If our model correctly classifies all four tokens as 'invoice date', we achieve great performance, however, we would still need to perform post-processing steps to assemble the final value, here for example, connecting tokens of the same line. Since many scientific articles focus dataset performance and do not specify pre- and post-processing procedures [15, 27, 8], it is difficult to validate the deployed pipeline, in terms of full information retrieval.

In order to specify quality standards and comparison criteria in the future design and comparison of extraction pipelines, we distinguish here four categories of challenges: **Document-, Data-, Model- and Assembly Challenges** for ML pipelines in invoice recognition. While Document and Data Challenges refer to quality criteria, Model and Assembly Challenges are usually consequences of design decisions. Due to the fact that there are a variety of factors influencing the overall pipeline in these four categories, it is difficult to determine or to rank the influence of any single parameter. However, different quality criteria and design decisions influence and potentially reinforce each other. This interdependency may propagate errors through the pipeline, e.g., poor scan quality can lead to incorrectly recognized characters, which limits the tokenization quality. Although digitally created and sent documents have a great advantage over paper-based documents as they eliminate the factor of scan quality, and therefore generally provide better OCR results (or make them obsolete), they still face downstream challenges in the pipeline, such as adequate label quality. **Document Challenges** include the image/scan/file quality as well as the syntactic and semantic quality. The former is due to image noise, fading colors, or skewness [49]. Syntactic and semantic quality refer to errors that arise when the invoice is created, e.g., spelling mistakes or incorrect abbreviations. Even if these errors are out of the scope of the actual invoice recognition task, they can still influence training and inference in ML pipelines. In the **Data Challenges** category we distinguish between text quality, visual information, tokenization, and label quality. Since we cannot assume that underlying meta data (i.e., words, labels, positions) of the documents are available and we cannot mine the content stream of the pdf file, an electronic conversion of the text must be performed by OCR engines, e.g., *Tesseract, Paddle, Abbyy, MS Azure Read, Google Cloud Vision* [50]. If the pixels are not translated into the correct characters, words, sentences, or paragraphs, with corresponding bounding boxes, difficulties arise in the subsequent process. For pure computer vision approaches, the OCR step is omitted since only bounding boxes and class labels are needed. For graphical information, the main challenge is to translate them into features, i.e., how to deal with table margins or white spaces, which give the document an essential structure, but are difficult to

standardize in terms of data processing. Token quality refers primarily to pipelines that rely on language models. If the tokens are too small or big, information may be lost. Label quality refers to a major problem in data storage. Usually, image data and label data are stored separately, i.e., an assignment of labels to corresponding elements has to be programmed first. In addition, it may happen that several values are assigned to a single key or several keys, e.g., 'invoice date' and 'shipping date' receive the same value. That means that the assignment could be ambiguous. The labels themselves can be inaccurate, as they usually stem from manual annotation. An essential quality criterion of the data is therefore the existence of an accurate ground truth, which also influences the number and kind of classification labels. Another challenge is the inherently strong class imbalance, i.e., the valuable information are usually contained in minority classes, which increases the chance of misclassification. **Model Challenges**. In ML features are used to represent the learning task. Feature engineering refers to the transformation steps and enrichment of the raw data, e.g., normalization or the procedure that calculates neighborhood relations [32]. In general, converting raw data into model inputs can lead to information loss. The selected model of course influences the feature engineering, since the inputs must be prepared exactly as the model expects them to be, e.g., for pixel models with segments, segmentation masks must be created, or graph representations for graph models. The model itself then specifies how the corresponding outputs look like, e.g., probabilities per class. Furthermore, the integration of pre-trained models and the determination of hyperparameters also influence the pipeline. In addition, the characteristics of the dataset must be considered, i.e., the general sample size, the distribution of the classes, and the variance of layouts and tokens. Here, model and data challenges overlap. Lastly, attention must be drawn to challenges in pre- and post-processing [51]. Depending on the outputs of the model, only fragments of the actual information are predicted. **Assembly** quality means how well abstract predictions at the entity level are converted into complete information. e.g., by defining pre- and post-processing heuristics. For example, it could be useful to assemble a numeric date from a string-based date.

# 4 Results and Conclusion

The result of our initial design cycle are the catalogs of information types and pipeline challenges, summarized in Table 3.2 and Table 3.3, which we apply in collaboration with our design process stakeholders (SME industry partner) in subsequent design cycles. The catalogs generally help to better understand the underlying data problem in each extraction case and to optimize the overall pipeline. As they address the relevance cycle of our design study, they can be used in two beneficial ways. First, they help to compare different approaches, as they offer structuring elements, e.g., whether and which OCR engine is used (Table. 3.4). Second, they support the selection of models and data preparation strategies based on available datasets, and vice versa. The following questions are suitable for the application of the framework:

1. What type of document and what language are at hand?
2. Which classes and annotations are available and which outputs are required?
3. Which information types are involved?

4. Which models and which hardware/implementation structure can be used?

5. Which challenges are derived from the constellation of questions 2, 3 and 4?

Regarding the information types, we recommend using Table 3.2 to identify all the types that are relevant to the specific case. Based on model selection, the details and challenges in the implementation of the pipeline can then be identified (Table 3.3). We briefly showcase how we utilize the catalogs to design a pipeline, by means of the examples shown in Table 3.2 and 3.3.

Table 3.2: Information Types with Examples

| Information Type | Specification Examples |
|---|---|
| Segmental | Pixel, Symbol/Letter/Character, Token/Word, Sentence, Paragraphs/Block/Columns, Pages/Slides |
| Syntactic | Data type (string, digit, decimal, date, alphanumeric), Case (upper, lower, mixed), Special characters (e.g., @, %), Count, Tagging (e.g., PoS, IOB, NER), character or byte pair encoding |
| Semantic | Numerical representation (e.g., vectorization) of an entity, based on context provided by a (word) model, such as Word2vec, GloVe, or BERT |
| Spatial/ Positional | Explicit: Bounding box (vertices), coordinates: left, top, width, height, right, bottom (absolute/relative/scaled); Implicit: image pixel structure (Euclidean) Distance, Size, Area, Center points |
| Neighborhood | Spatial neighborhood, semantic connection, neighbor features, self-loop |
| External | Dictionaries, e.g., ZIP codes, City names Conventions/ Abbreviations, e.g. Tax ID, IBAN, Spell check |
| Graphical | Font (type, size, style); Tables/Frames, Lines, dots, arrows, connectors, boxes, contours, Machine-readables (QR, barcode), Images |
| Logical | Calculations: sums, percentages, quantities, units, rebates, discounts; References |

Table 3.3: ML pipeline challenges in four categories

| C1: Document | C3: Model | C4: Assembly |
|---|---|---|
| Image/Scan Quality<br>*e.g. noise, skewness*<br>Syntactic/Semantic Quality<br>*e.g. false/missing information* | Feature Engineering<br>*e.g. word model, normalization*<br>Model Design<br>*e.g. outputs, hyperparameters* | Pre/Post-Processing Heuristics<br>*e.g. class aggregation, ambiguity* |
| C2: Data | | |
| Sample Quality<br>*e.g. size, distribution*<br>Text Quality<br>*e.g. OCR conversion* | Tokenization Quality<br>*e.g. stemming*<br>Graphics Quality<br>*e.g. whitespace* | Label Quality<br>*e.g.incorrect annotation* |

In our present use case, the goal is to develop a model that recognizes information on German invoices. We have a real-world dataset available, consisting of 977 pdf files with rule-based label annotations and OCR extracts for over 60 classes, which provides segmentation on word level. The labels to be classified in the model include among others the invoice date, invoice number, total amount, payment information, and also specifics such as the IBAN (international bank account number) and commercial register number, i.e., digit and character syntax. As for this design cycle we apply a neighborhood algorithm, however, we don't integrate external, graphical and logical information. We also exclude product line items and their details (quantities, amounts) in this cycle. We identify the general size of the dataset, class distributions (minority classes), and the label quality as potential challenges. Regarding the choice of model, we want to leverage the ability to integrate a word model, since we deal with German documents, instead of English documents (as predominant in science).

Therefore, the semantic aspect and neighborhood should be explicitly modifiable in the pipeline (Table 3.2). Based on these requirements, we prototypically decide to investigate graph-based models. With regards to our literature review, we consider the graph-based approaches from [35], [32], and [15]. We select [15], because it is the most recent one and integrates semantic, syntactic and positional information types. This model was originally trained on a dataset of 1,129 manually labeled English invoices, which is comparable in terms of size. Differences emerge in the language and dataset variance, as only a single recipient was considered there, while we look at various recipients in our dataset. Table 3.4 showcases how the categorization helps to draw a comparison between approaches. Both GCN models include graph attention layers implemented in Pytorch and Deep Graph Library (DGL), which are accessible from open-source. For a detailed description and exact training parameters we refer to [15]. We enhance their approach by integrating a German-BERT[4] model for semantic features into the pipeline (Figure 3.2).

Table 3.4: Exemplary comparison of approaches

| | Quality category | Krieger et al. (2021) [15] Pytorch DGL | Adopted Prototype Pytorch DGL |
|---|---|---|---|
| Doc. | Scan/born digital | n.a. | scanned |
| Doc. | Date range of invoice | n.a. | 2010-2019 |
| Data | # of inv. (vendors/recip.) | 1129 (277/1) | 977 (494/531) |
| Model | Training method and split | supervised 80/20 | supervised 80/10/10 |
| Process. | Labeling method | manually | rule-based |
| Data | File Formats | n.a. | .pdf, .json |
| Process. | OCR engine | Tesseract | Abbyy |
| Model | Model type | GCN | GCN |
| Model | Labels excl. *'no_label'* | *date,number,total* | *date,number,total* |
| Metric | Test F1-score | 0.905 | 0.823 |

[4] https://www.deepset.ai/german-bert

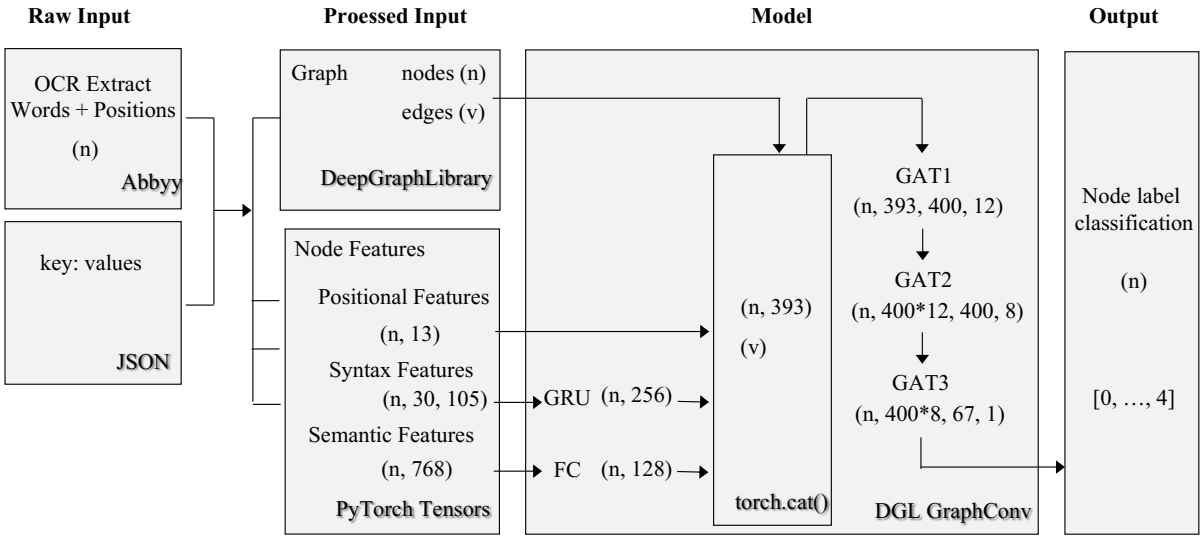


Figure 3.2: Prototype pipeline section, author's figure

Based on initial trainings on our dataset our prototype in this stage of the design study does not achieve the same performance (F1-score of 0.823) as the benchmark model (0.905). The difference could be due to the fact that we use a different word model and a different algorithm for neighborhood generation. Nevertheless, the application of the pipeline and the model results represent a proof of concept for our framework.

**Contribution and Future Research**. As the prior analysis and detailed (de-) composition of the data problem is an essential step in ML/AI projects, the catalogs are very beneficial tools. They form the morphological basis for the problem analysis and the derivation of design requirements in our specific as well as in general use cases beyond specific datasets and key items. With the help of the framework, model selection and data preparation can be better structured, and it improves the comparability between different approaches. Furthermore, as the classification problem is better understood, it helps to augment and synthesize further training data, which also highlights a path for future research. The novelty of our work is that the specific challenges and information types have not been cataloged before. The application of the framework supports both practice pipeline implementations as well as descriptive process research, into which these kinds of catalogs can be adopted, e.g., in DASC-PM. We validate our results here by demonstrating the selection, comparison, and enhancement of a model pipeline within our design study. We plan to use the framework in our future design path. The framework enables targeted work in the IS community and related disciplines towards more comprehensive document analysis and recognition, as it provides guidance to identify and implement improved pipelines. Especially, in terms of line item recognition many training opportunities remain [52], provided that appropriate (labeled) data can be accessed. Too little training data have already been discussed in other papers as a main source of suboptimal results [35]. Future research could leverage our work to rank the categories' influence and develop a quantitative scoring model for extraction pipelines, which can be applied alongside standard performance metrics, such as F1-score. This would further enhance the comparability of different approaches from both research and practice.

# References


[1] Bertin Klein, Stevan Agne, and Andreas Dengel. "Results of a Study on Invoice-Reading Systems in Germany". In: *Document analysis systems VI*. Ed. by David Hutchison. Vol. 3163. Lecture Notes in Computer Science. Berlin: Springer, 2004, pp. 451–462. ISBN: 978-3-540-23060-1. DOI: 10.1007/978-3-540-28640-0_43.

[2] Alea Fairchild. "Using Electronic Invoicing to Manage Cash Forecasting and Working Capital in the Financial Supply Chain". In: *ECIS 2004 Proceedings* (2004). URL: https://aisel.aisnet.org/ecis2004/29.

[3] Mussadiq Abdul Rahim et al. "RFM-based repurchase behavior for customer classification and segmentation". In: *Journal of Retailing and Consumer Services* 61 (2021), p. 102566. ISSN: 0969-6989. DOI: 10.1016/j.jretconser.2021.102566. URL: https://www.sciencedirect.com/science/article/pii/S0969698921001326.

[4] Lisa Koonce, Urton Anderson, and Garry Marchant. "Justification of Decisions in Auditing". In: *Journal of Accounting Research* 33.2 (1995), p. 369. ISSN: 00218456. DOI: 10.2307/2491493. URL: http://www.jstor.org/stable/2491493.

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

[6] Rüdiger Wirth and Jochen Hipp. "CRISP-DM: Towards a standard process model for data mining". In: *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining* (2000). URL: https://cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf.

[7] Michael Schulz et al. *DASC-PM v1.1 - Ein Vorgehensmodell für Data-Science-Projekte*. Ed. by Universitäts- und Landesbibliothek Sachsen-Anhalt and Martin-Luther Universität. 2022. DOI: 10.25673/85296. URL: http://dx.doi.org/10.25673/85296.

[8] Yang Xu et al. *LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding*. 2020. URL: https://arxiv.org/pdf/2012.14740.

[9] Bodhisattwa Prasad Majumder et al. "Representation Learning for Information Extraction from Form-like Documents". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020), pp. 6495–6504. DOI: 10.18653/v1/2020.acl-main.580. URL: https://aclanthology.org/2020.acl-main.580/.

[10] Rosario Girasa. *Artificial Intelligence as a Disruptive Technology: Economic Transformation and Government Regulation*. 1st ed. 2020. Cham: Springer International Publishing and Imprint Palgrave Macmillan, 2020. ISBN: 9783030359751. URL: https://link.springer.com/book/10.1007/978-3-030-35975-1.

[11] Alan Hevner et al. "Design Science in Information Systems Research". In: *MIS Quarterly* 28.1 (2004), p. 75. ISSN: 02767783. DOI: 10.2307/25148625. URL: https://arizona.pure.elsevier.com/en/publications/design-science-in-information-systems-research.

[12] Ken Peffers et al. “A Design Science Research Methodology for Information Systems Research”. In: *Journal of Management Information Systems* 24.3 (2007), pp. 45–77. ISSN: 0742-1222. DOI: 10.2753/MIS0742-1222240302.

[13] Alan Hevner. “A Three Cycle View of Design Science Research”. In: *Scandinavian Journal of Information Systems* 19 (2007). URL: https://www.researchgate.net/publication/254804390_A_Three_Cycle_View_of_Design_Science_Research.

[14] Xiaojing Liu et al. *Graph Convolution for Multimodal Information Extraction from Visually Rich Documents*. Mar. 27, 2019. URL: https://arxiv.org/pdf/1903.11279.

[15] Felix Krieger et al. “Information Extraction from Invoices: A Graph Neural Network Approach for Datasets with High Layout Variety”. In: *Wirtschaftsinformatik 2021 Proceedings* (2021). URL: https://aisel.aisnet.org/wi2021/RDataScience/Track09/4.

[16] Serif Adali, A. Coskun Sonmez, and Mehmet Gokturk. “An Integrated Architecture for Processing Business Documents in Turkish”. In: *Computational linguistics and intelligent text processing*. Ed. by Alexander Gelbukh. Vol. 5449. Lecture notes in computer science Theoretical Computer Science and General Issues. Berlin and Heidelberg: Springer, 2009, pp. 394–405. ISBN: 978-3-642-00381-3. DOI: 10.1007/978-3-642-00382-0_32.

[17] Bill Janssen et al. “Receipts2Go”. In: *Proceedings of the 2012 ACM symposium on Document engineering*. Ed. by Cyril Concolato. ACM Conferences. New York, NY: ACM, 2012. ISBN: 9781450311168. DOI: 10.1145/2361354.2361381.

[18] Yolande Belaid and Abdel Belaid. “Morphological tagging approach in document analysis of invoices”. In: *Proceedings of the 17th International Conference on Pattern Recognition, 2004*. 2004, 469–472 Vol.1. ISBN: 0-7695-2128-2. DOI: 10.1109/ICPR.2004.1334166.

[19] Daniel Esser et al. “Automatic indexing of scanned documents: a layout-based approach”. In: SPIE, 2012, pp. 118–125. DOI: 10.1117/12.908542. URL: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/8297/1/Automatic-indexing-of-scanned-documents-a-layout-based-approach/10.1117/12.908542.short?SSO=1.

[20] Andreas R. Dengel and Bertin Klein. “A Requirements-Driven System for Document Analysis and Understanding”. In: *DAS 2002* (2002), pp. 433–444. DOI: 10.1007/3-540-45869-7_47. URL: https://link.springer.com/chapter/10.1007/3-540-45869-7_47.

[21] D. Schuster et al. “Intellix - End-User Trained Information Extraction for Document Archiving”. In: *12th International Conference on Document Analysis and Recognition*. 2013, pp. 101–105. ISBN: 978-0-7695-4999-6. DOI: 10.1109/ICDAR.2013.28.

[22] R. B. Palm, O. Winther, and F. Laws. “CloudScan - A Configuration-Free Invoice Analysis System Using Recurrent Neural Networks”. In: *12th International Conference on Document Analysis and Recognition*. 2013, pp. 406–413. ISBN: 978-0-7695-4999-6. DOI: 10.1109/ICDAR.2017.74.

[23] Sonit Singh. *Natural Language Processing for Information Extraction*. Australia, July 6, 2018. URL: https://arxiv.org/pdf/1807.02383.

[24] Xiao Yang et al. *Learning to Extract Semantic Structure from Documents Using Multimodal Fully Convolutional Neural Network*. June 7, 2017. URL: https://arxiv.org/pdf/1706.02338.

[25] Rasmus Berg Palm, Florian Laws, and Ole Winther. "Attend, Copy, Parse - End-to-end information extraction from documents". In: *ICDAR* (2019). URL: https://arxiv.org/pdf/1812.07248.

[26] Brian Davis et al. "Deep Visual Template-Free Form Parsing". In: *15th International Conference on Document Analysis and Recognition* (2019). URL: https://arxiv.org/pdf/1909.02576.

[27] Anoop Raveendra Katti et al. "Chargrid: Towards Understanding 2D Documents". In: *Proceedings of EMNLP*. 2018. URL: https://arxiv.org/pdf/1809.08799.

[28] Timo I. Denk. "Wordgrid: Extending Chargrid with Word-level Information". PhD thesis. Unpublished, 2019. DOI: 10.13140/RG.2.2.19846.11844.

[29] Timo I. Denk and Christian Reisswig. "BERTgrid: Contextualized Embedding for 2D Document Representation and Understanding". In: *33rd Conference on Neural Information Processing Systems, Vancouver, Canada.* (2019). URL: https://arxiv.org/pdf/1909.04948.

[30] Weihong Lin et al. "ViBERTgrid: A Jointly Trained Multi-Modal 2D Document Representation for Key Information Extraction from Documents". In: *Proceedings of ICDAR 2021*. 2021. URL: https://arxiv.org/pdf/2105.11672.

[31] Arsen Yeghiazaryan et al. "Tokengrid: Toward More Efficient Data Extraction From Unstructured Documents". In: *IEEE Access* 10 (2022), pp. 39261–39268. DOI: 10.1109/ACCESS.2022.3164674.

[32] D. Lohani, A. Belaïd, and Y. Belaïd. "An Invoice Reading System Using a Graph Convolutional Network". In: *ACCV Workshops*. 2018, pp. 144–158. DOI: 10.1007/978-3-030-21074-8_12. URL: https://link.springer.com/chapter/10.1007/978-3-030-21074-8_12.

[33] Łukasz Garncarek et al. "LAMBERT: Layout-Aware Language Modeling for Information Extraction". In: *Document Analysis and Recognition - ICDAR 2021* Vol. 12821 (2021), pp. 532–547. DOI: 10.1007/978-3-030-86549-8_34. URL: https://arxiv.org/pdf/2002.08087.

[34] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Oct. 11, 2018. URL: https://arxiv.org/pdf/1810.04805.

[35] Yinhan Liu et al. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. July 26, 2019. URL: https://arxiv.org/pdf/1907.11692.

[36] Yiheng Xu et al. *LayoutLM: Pre-training of Text and Layout for Document Image Understanding*. 2020. DOI: 10.1145/3394486.3403172. URL: https://arxiv.org/pdf/1912.13318.

[37] Henrik Nell. "Quantifying the noise tolerance of the OCR engine Tesseract using a simulated environment". Master Thesis. Karlskrona, Sweden: Blekinge Tekniska Högskola, 2015. URL: http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A831347&dswid=6545.

[38] Saurabh Gupta et al. "Learning Rich Features from RGB-D Images for Object Detection and Segmentation". In: *European Conference on Computer Vision (ECCV)* (2014). DOI: appear. URL: https://arxiv.org/pdf/1407.5736.

[39] Hajar Farahmand, Ali Harounabadi, and S. Javad Mirabedini. “Document features selection using background knowledge and word clustering technique”. In: *Management Science Letters* (2014), pp. 241–250. ISSN: 19239335. DOI: 10.5267/j.msl.2013.12.033.

[40] Erik Tjong Kim Sang and Jorn Veenstra. “Representing Text Chunks”. In: *Ninth Conference of the European Chapter of the Association for Computational Linguistics* (1999), pp. 173–179. URL: https://aclanthology.org/E99-1023/.

[41] David Nadeau and Satoshi Sekine. “Named Entities: Recognition, classification and use”. In: *Lingvisticæ Investigationes* 30.1 (2007), pp. 3–26. ISSN: 1569-9927. DOI: 10.1075/li.30.1.03nad. URL: https://www.jbe-platform.com/content/journals/10.1075/li.30.1.03nad.

[42] Barbara H. Partee. “Semantics”. In: *The MIT Encyclopedia of the Cognitive Sciences, Cambridge, MA* (1999), pp. 739–742.

[43] Ronald Stamper et al. “Understanding the roles of signs and norms in organizations - a semiotic approach to information systems design”. In: *Behaviour & Information Technology* 19.1 (2000), pp. 15–27. DOI: 10.1080/014492900118768.

[44] Ronald Stamper. *A semiotic theory of information and information systems*. Enschede, 1993. URL: https://research.utwente.nl/files/5383733/101.pdf.

[45] Tomas Mikolov et al. *Efficient Estimation of Word Representations in Vector Space*. Jan. 16, 2013. URL: https://arxiv.org/pdf/1301.3781.

[46] Jeffrey Pennington, Richard Socher, and Christopher Manning. “Glove: Global Vectors for Word Representation”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2014), pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://aclanthology.org/D14-1162/.

[47] Victor Sanh et al. *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. Oct. 2, 2019. URL: https://arxiv.org/pdf/1910.01108.

[48] Song Mao, Azriel Rosenfeld, and Tapas Kanungo. “Document structure analysis algorithms: a literature survey”. In: SPIE, 2003, pp. 197–207. DOI: 10.1117/12.476326. URL: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/5010/1/Document-structure-analysis-algorithms-a-literature-survey/10.1117/12.476326.short?SSO=1.

[49] K. M. Yindumathi, Shilpa Shashikant Chaudhari, and R. Aparna. “Analysis of Image Classification for Text Extraction from Bills and Invoices”. In: *11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. Piscataway, New Jersey: IEEE, 2020. ISBN: 978-1-7281-6851-7. DOI: 10.1109/ICCCNT49239.2020.9225564.

[50] Ahmad P. Tafti et al. “OCR as a Service: An Experimental Evaluation of Google Docs OCR, Tesseract, ABBYY FineReader, and Transym”. In: *Advances in visual computing*. Ed. by George Bebis. Vol. 10072. Lecture Notes in Computer Science. Cham: Springer, 2016, pp. 735–746. ISBN: 978-3-319-50834-4. DOI: 10.1007/978-3-319-50835-1_66.

[51] Tobias Grüning et al. “A two-stage method for text line detection in historical documents”. In: *International Journal on Document Analysis and Recognition (IJDAR)* 22.3 (2019), pp. 285–302. ISSN: 1433-2833. DOI: 10.1007/s10032-019-00332-1.

[52] Stěpán Šimsa et al. *DocILE Benchmark for Document Information Localization and Extraction*. Feb. 11, 2023. URL: https://arxiv.org/pdf/2302.05658.

# Paper 4

# Ablation Study of a multi-modal GAT Network on perfect synthetic and real-world Data to investigate the Influence of Language Models in Invoice Recognition

## Outline



## Bibliographic Information


Lukas-Walter Thiée.
Leuphana Universität Lüneburg, Institute for Information Systems, Lüneburg, Germany.
11.09.2024, https://doi.org/10.1007/978-3-031-70642-4_13
18th International Conference on Document Analysis and Recognition, ICDAR 2024, held in Athens, Greece, during August 30–31, 2024.


## Copyright Notice

# Abstract

Document analysis and invoice recognition have been significantly advanced in recent years by grid-based, graph-based and transformer architectures. However, it is not only the model architecture that influences an approach's results, but also the quality of training and test data. In this paper, we perform an ablation study on an existing state-of-the-art pre-trained multi-modal GAT network. Therein we investigate two kinds of modifications to understand the sensitivity of the results by (1) exchanging the language module and (2) applying both the original and modified network on a perfect synthetic and an imperfect real-world dataset. The results of the study show the importance of language modules for semantic embeddings in multi-modal invoice recognition and illustrate the impact of data annotation quality. We further contribute an adapted GAT model for German invoices.



**Keywords**

Invoice recognition, GAT, Synthetic data, Inv3D, *GraphDoc*

# 1 Introduction

In the digital age, businesses are increasingly reliant on data-driven processes to streamline operations and enhance efficiency. Among these, automated invoice processing stands out as a critical component, playing a pivotal role in the financial and administrative functions of organizations. The ability to swiftly and accurately handle invoices not only accelerates payment cycles but also reduces errors associated with manual data entry, thereby optimizing resource allocation and improving overall financial management. Using structured invoice data, various stakeholders can obtain information and optimize processes, e.g., for financial auditing [1]. Despite the potential for electronic transmission and standardized integration of invoice data [2], the prevailing practice, in both B2B and B2C relationships, remains sending invoices in paper or pdf format, necessitating the extraction of information from the document or file once again. Leveraging structured, meta, or blockchain data is not common practice.

**Characteristics of Invoice Recognition**. Automated invoice processing involves the extraction and interpretation of pertinent information from a myriad of invoice documents. This task, while crucial, is often challenging due to the inherent variability in invoice formats, diverse language structures, and the complexity of interrelations between different data elements. In contrast to pure, sequential text, invoices represent a form of visually rich documents [3], on which various signals, especially semantics and 2-dimensional layout, contain the information. The plethora of information on invoices can range from key items like addresses, dates, and invoice numbers, over account and payment details, to product quantities and descriptions (line items). Prior research has investigated the information types and pipeline challenges in information extraction from invoices [4]. Due to the privacy of sensitive personal and financial data, another hurdle in invoice recognition research is to get access to datasets that are sufficient in annotation quality and sample quantity to train machine learning (ML) algorithms. Existing publicly available datasets, such as SROIE[1] or FUNSD[2], are very beneficial for the research community and used for various benchmark comparisons. However, these are usually tailored to a specific domain, contain too few samples, only consider a certain set of class labels, or generally have poor annotation quality. Therefore, it is promising to utilize a combination of synthetic and real data to improve both annotation quality and sample size, such as the *Document Information Localization and Extraction* (DocILE[3]) dataset. Conventional methods, such as rule-based systems and traditional machine learning models, have demonstrated limitations in handling the intricate nature of invoice data, prompting the exploration of advanced techniques capable of capturing nuanced relationships within these documents. In order to effectively capture information and address the challenges posed by the abundance of unstructured data and diverse layouts, a multitude of new approaches within the field of machine learning have been introduced in recent years. These approaches encompass various techniques, including NLP-based models [5, 6], computer vision methodologies [7], graph-based methods [3, 8, 9], deep learning networks [10, 11], and transformer architectures [12, 13, 14]. One such promising avenue is the application of Graph Attention Networks (GAT), a subset of neural networks specifically designed to model complex relationships within graph-structured data. Graphs, in this context,

[1] SROIE: https://rrc.cvc.uab.es/?ch=13
[2] FUNSD: https://guillaumejaume.github.io/FUNSD
[3] DocILE: https://docile.rossum.ai/

represent the intricate network of connections between various entities present in an invoice, such as vendor details, line items, and transaction amounts. The inherent ability of GATs to weigh and prioritize different elements of the graph based on contextual relevance makes them a compelling choice for automated invoice processing [15].

**Research questions and goal**. Current research points to the benefits of four application streams, 1. fine-tuning of pre-trained models, 2. designing multi-modal inputs, 3. integrating (large) language models, and 4. utilizing synthetic data. This research paper wants to leverage all of these streams and delves into the application of Graph Attention Networks to two distinct datasets, each presenting unique challenges and characteristics in the realm of automated invoice processing. By harnessing the power of GATs, we aim to enhance the accuracy and efficiency of information extraction from invoices, ultimately contributing to the broader goal of seamless and error-free automated financial workflows. In particular, we perform a modified ablation study on an existing GAT approach. Whereas a conventional ablation study would omit parts of the model to understand the contribution of the component, we exchange an English word model with a German word model. We then apply both models on two different datasets to understand the impact of input data quality. To this end, we leverage two state-of-the-art approaches for our study both related to the 17th International Conference on Document Analysis and Recognition (ICDAR). Inv3D by Hertlein et al. [16] was presented at ICDAR 2023[4] and *GraphDoc* by Zhang et al. [17] run the leaderboard in the ICDAR 2023 Competition on DocILE[5]. We strongly believe it is imperative that not only the models of state-of-the-art approaches, but also the entire ML pipeline, including the original and additional datasets, are validated to ultimately minimize ML-related debt, such as reproducibility debt and data quality debt, as introduced by Sculley et al. [18]. The approaches selected for this study have been chosen to represent state-of-the-art in invoice recognition and dataset synthesis. They will be briefly described in our related work section. The comparative analysis of GATs on the datasets will provide insights into the adaptability and generalizability of the proposed approach across different domains. As we embark on this research journey, the following key questions will guide our investigation:

1. *What performance impact does the match between a model's language module and the dataset's language have?*

2. *How does the performance of the model change when switching from perfect (synthetic) to imperfect (real-world) data?*

**Relevance and Contribution**. Addressing these questions will not only contribute to the advancement of automated invoice processing but will also shed light on the broader applicability and limitations of Graph Attention Networks in handling complex and dynamic relational data structures. Through this research, we aim to provide valuable insights that will inform the development of more robust and adaptive solutions for automated invoice processing in diverse business environments. We contribute a performance test of a multi-modal state-of-the-art approach, to analyze the impact of language models. Furthermore, we extend *GraphDoc* with an existing German language transformer model.

[4] https://icdar2023.org/program/accepted-papers/
[5] https://icdar2023.org/program/competitions/

# 2 Related Work: Models and Datasets

## 2.1 GraphDoc

Zhang et al. present *"a multimodal graph attention-based model for various document understanding tasks"* [17], called *GraphDoc*[6]. *GraphDoc* leverages three multi-modal inputs to the network, namely positional, textual and image information, and unites those in a gate fusion layer. The textual features are integrated via the pre-trained English Sentence-BERT (EB) model as a language module, which provides sentence embeddings. The weights of the language module are not being pre-trained in *GraphDoc*. Zhang et al. build upon the idea that individual elements of a document can depend on their direct neighborhood, and therefore *"inject the graph structure into the attention mechanism to form a graph attention layer so that each input node can only attend to its neighborhoods"* (see Figure 4.1). *GraphDoc* is pre-trained on a sample of 320k images from the RVL-CDIP dataset to learn document representations through masked sentence modelling, and it is evaluated on the FUNSD dataset. *GraphDoc* is based on regions of interest (RoI), which represent contiguous areas of the layout of a document. *EasyOCR* is used to obtain texts, bounding boxes, and regions. The structure of the document is induced by the graph attention mechanism so that k=36 of the nearest neighbors of a region are taken into account. Zhang et al. not only report very good results of the approach on various benchmark datasets in their paper, but also lead the ranking in the DocILE challenge by Šimsa et al. [19]. Table 4.1 lists selected results.

Table 4.1: Collection of F1-scores of *GraphDoc* on different datasets [17]

| Model | Dataset | FUNSD (Form) F1 | SROIE (Receipt) F1 | CORD (Receipt) F1 | DocILE (Invoice) F1 |
|---|---|---|---|---|---|
| *GraphDoc* (EB) | | 0.8777 | 0.9845 | 0.9693 | 0.7425[7] |

## 2.2 (German) BERT

With the rise of vectorized word models in natural language processing, like Word2Vec [20] or GloVe [21], also transformer architectures were introduced to represent language. One of those architectures is BERT (*Bidirectional Encoder Representations from Transformers* [22]), which has been developed and open-sourced by Google[8]. It *"is basically a multi-layer bidirectional transformer"* [12], to generate semantic embeddings of words within the context of its sentence. It uses the self-attention mechanism to weigh the significance of different words in a sentence. By training on large amounts of textual data, BERT learns to generate contextually rich word embeddings, allowing it to grasp the relationships between words and their contextual meanings.

English Sentence BERT is an extension of the BERT model, specifically tailored for sentence-level embeddings. It focuses on learning sentence representations. This makes

[6] https://github.com/ZZR8066/GraphDoc

[7] Robust Reading Competition: DocILE 2023, Key Information Localization and Extraction: https://rrc.cvc.uab.es/?ch=26&com=evaluation&task=1

[8] https://research.google/blog/open-sourcing-bert

it suitable for various NLP tasks such as sentence similarity, classification, and semantic textual similarity. On the other hand, the German BERT model[9] by deepset.ai [23] is a BERT-based model fine-tuned for the German language. Fine-tuning involves training a pre-trained model on a specific task or dataset to adapt it for domain-specific or language-specific applications. The German BERT model captures the intricacies of the German language and can be employed in a range of NLP tasks for German text.

## 2.3 Inv3d, perfect data

Hertlein et al. [16] establish a new benchmark dataset called Inv3D[10] for automated invoice processing, mainly targeted at dewarping invoice images and invoice recognition tasks. Their large-scale high-resolution dataset consists of 25,000 samples including the flatbed image and pdf file and two ground-truth annotation json-files, i.e., complete list of words and relevant areas (key and line items), split into 0.7 train, 0.15 validation, and 0.15 test sets. *"The dataset creation pipeline consists of four stages: resource preparation, invoice rendering, invoice warping, and finally the auxiliary map generation"* [16]. For the purpose of our study the first two stages are relevant. Inv3D is based on 100 real template layouts. They leverage the *Python Faker* package to randomly generate coherent invoice content, e.g., sales orders and personas, and apply random modifications to the design of the invoice, e.g., font sizes and colors, and exchange company logos. The content is mapped to a machine-readable tag structure, such as buyer.shipment.address. The train dataset contains 162 classes. They also provide the code of their pipeline to generate further samples with new content, layout, and other types of documents. We utilize Inv3D in this study, because the ground-truth of this dataset is perfect, meaning the annotations and boxes are unambiguous, which is crucial for understanding class based inference on token level.

## 2.4 Private Dataset (PD), imperfect data

We have a non-public dataset available that consists of German invoices and receipts, hereinafter referred to as PD. The data originates from the accounting systems of different German agricultural businesses. These invoices encompass a diverse range of transactions and commodities, extending beyond exclusively agricultural products. Instead, they comprise a substantial number of service-related invoices and invoices associated with the acquisition of diverse industrial and consumer goods. The pdf files correspond to scanned invoices sourced from different companies and various scanning devices, leading to variations in scan quality. In total, our dataset comprises 977 invoices originating from 494 distinct vendors and involving 531 distinct recipients. Abbyy Finereader OCR results in 196,548 (43,621 distinct) tokens, with a mean average of 201 tokens per invoice. The results are subject to various challenges, such as the umlauts in the German language or unstandardized abbreviations. The invoices have been annotated for 67 different classes, such as *adresse, beleg_datum, and rechnung_nr, but also steuer_nr, iban*, and *ust_id_number*, which are more specific to the German-speaking area (see Table 4.2).

The annotation is based on a sophisticated rule-based algorithm, which yields decent, however, not perfect results. Annotations could be inaccurate and incomplete. For exam-

[9] https://www.deepset.ai/german-bert
[10] https://felixhertlein.github.io/inv3d

Table 4.2: Selection of class labels in the datasets

| **Type** | **Inv3D classes** | | **PD classes** |
|---|---|---|---|
| Key-items: contact | buyer.bill.address | seller.address (1) | adresse (1) |
| | buyer.bill.address.city | seller.address.city | telefon_nr (4) |
| | buyer.bill.address.city.postcode | seller.address.city.postcode | telefax_nr |
| | buyer.bill.address.street | seller.address.street | email (3) |
| | buyer.bill.contact | seller.contact | kunden_nr (2) |
| | buyer.bill.email | seller.email (3) | bestell_nr |
| | buyer.bill.customer_id (2) | seller.phone_number (4) | referenz_nr |
| | buyer.bill.phone_number | seller.company.name | hndlreg_nr |
| | buyer.company.name | seller.company.slogan | steuer_nr |
| | buyer.shipment.address | seller.fax_number | ust_id_number |
| | buyer.shipment.contact | seller.salesperson | url |
| | buyer.shipment.phone_number | | bic |
| | beneficiary.iban.account_code | | blz |
| | beneficiary.name | | konto_nr |
| | beneficiary.bic | | iban |
| | beneficiary.iban | | karten_nr |
| | beneficiary.bank.address.city | | |
| | beneficiary.bank.address.street | | |
| | beneficiary.bank.name | | |
| | beneficiary.bank.address | | |
| Key-items: general | summary.balance (7) | invoice_date (5) | beleg_datum (5) |
| | summary.discount | invoice_due_date | rechnung_nr (6) |
| | summary.subtotal | invoice_number (6) | betrag (7) |
| | summary.tax_rate | payment_terms | zahlungsart |
| | summary.tax_total | | referenz_nr |
| | summary.discount | | bezahlt_flag |
| | summary.shipping.method | | zu_zahlen_flag |
| | summary.shipping.price | | globale_lokations_nr |
| Line-items | products.0.description | | |
| | products.0.id | | |
| | products.tax_rate | | |
| | products.0.quantity | | |
| | products.0.total | | |
| | products.0.unit_price | | |

(n) indicates counterparts in the datasets and integration in our train/test set

ple, they do not distinguish between vendor and recipient address, nor do they consider line items. For various tokens, the assignment between ground-truth data and OCR results is not unambiguous. Some of the documents may contain handwritten notes and they vary considerably in type and layout.

# 3 Experiments

Our study design consists of four experiments – two models and two datasets – each training and testing the corresponding model on one of the aforementioned datasets (see Table 4.3). The first model is the pre-trained model from [17], presented in section 2.1 with English Sentence-BERT (EB). We use the pre-trained main network and *GraphDocForTokenClassification* adapted to our experiments. This network tail is used to obtain token level classifications. As a second model we use *GraphDoc* again, but replace the language module with an implementation of German BERT (GB). We then fine-tune the pre-trained models on our datasets. Rather than omitting a component of the model, we exchange the component. We do this because we consider it necessary to keep the semantic module, as the model is designed for multi-modal inputs. Our ablation study is therefore not designed as a hyperparameter optimization, but rather as a general model verification. We are committed to using existing models, as well as testing and expanding their performance as part of scientific rigor in business informatics. Experiment E1 and E4 are our focus experiments, as the language of the dataset and the language module within the model match, i.e., English for E1 and German for E4. The proposed experiment setup provides us with six expedient comparison pairs, two that test the influence of the language module (E1 vs. E2 and E3 vs. E4) and four that test the combined influence of the dataset quality and language fit (E1 vs. E3, E2 vs. E4, E2 vs. E3 and E1 vs. E4).

**Expected results** (a priori qualitative hypotheses). With regard to our research questions and the possibilities for comparison between the experiments, we put forward the following qualitative hypotheses for the model performance P, measured as macro average F1-score. The comparisons are illustrated with arrows in Table 4.3. The direction of the arrow indicates an expected relative increase in performance.

Table 4.3: Experimental setup and comparison options

| Model | Dataset | Inv3D (English) | PD (German) |
|---|---|---|---|
| | Samples | Train: 17.500<br>Test: 3.750<br>Classes: 7/162 | Train: 782<br>Test: 195<br>Classes: 7/67 |
| *GraphDoc* (English Sentence BERT) | | E1 | E3 |
| *GraphDoc* (German BERT) | | E2 | E4 |

- **H1**: Isolated impact of the language module: The performance of the experiments with a matching language of the language module and dataset language is expected

to be higher than the performance of experiments with a non-matching language, when keeping the dataset. This means that in the case of Inv3D we expect a higher performance with *GraphDoc* EB, $P_{E1} > P_{E2}$. And in the case of PD, better performance with *GraphDoc* GB, $P_{E3} < P_{E4}$.

- **H2**: Combined effects: Since we are not modifying the datasets themselves, we can only investigate the combined impact of data quality and language fit. Generally, we expect the performance to decrease when switching from perfect synthetic data to imperfect real-world data, $P_{E1} > P_{E3}$, $P_{E2} > P_{E3}$, and $P_{E1} > P_{E4}$. However, for E2 vs. E4 we anticipate the effect of language fit to outweigh disadvantages of data quality, hence $P_{E2} < P_{E4}$.

### Preprocessing and Training

In order to train and test the models we have to prepare the datasets according to the model input structure. Inv3D contains 17.500 train and 3.750 test samples, and PD contains 782 train and 195 test samples. For both datasets we use the provided bounding boxes of the ground truth of all words and set these to the class *none*. We then match all words and bounding boxes with the class label ground truth to obtain word-level inputs.

To ensure that we have the same semantic classes in both datasets, we focus on 7 selected classes, such as *invoice_date* = *beleg_datum*, *invoice_number* = *rechnung_nr*, and *summary.balance* = *betrag*. Table 4.2 gives an impression of the possible and used classes. Whereas PD does not distinguish seller and buyer addresses, Inv3D uses a more detailed structure of the classes (buyer, seller, and beneficiary) and also explicitly enumerates product details. With this structure Inv3D can potentially be used for line item recognition training. We apply the model hyperparameters suggested by Zhang et al. [17], batch size 4, learning rate $5 * 10^{-5}$, and fine-tune the models for 50 epochs each. Fig. 4.1 illustrates the general process of the pipeline, i.e., word-level tokens, their different embeddings, and graph structure.

# 4 Results and Discussion

The overall and class-level performance metrics of the experiments are summarized in Table 4.4 and Table 4.5. We use the macro average F1-score for the evaluation of the model performance P. F1-score is the harmonic mean between precision and recall, which is evaluated per class. A macro average in multi-class classification is the arithmetic mean of all per-class scores regardless of their support, whereas a weighted average considers the class distribution.

Since we use new combinations of models and datasets, there is no direct benchmark comparison. Nevertheless, we can put the macro F1-score of the experiments in relation to results of the original model on other datasets (see Section 2.1 in Related Work). The F1-scores of the focus experiments, i.e., the ones with a language fit (E1 = 0.9324 and E4 = 0.8593), represent a significant improvement compared to the application of *GraphDoc* on the DocILE dataset (0.7425). However, in terms of dataset and class complexity, the DocILE benchmark is more complex. The performance of *GraphDoc* in the SROIE challenge is not achieved (0.9845). These comparisons provide a basic ranking but are limited in their informative value due to different dataset properties, such as document

Table 4.4: Classification report on Inv3D test dataset

| Model/Class | | Precision | Recall | F1-score |
|---|---|---|---|---|
| ***GraphDoc* (English BERT): Experiment 1** | | | | |
| (0) none | | 0.9948 | 0.9923 | 0.9935 |
| (1) buyer.bill.address | | 0.8476 | 0.9434 | 0.8929 |
| (2) buyer.bill.customer_id | | 0.8338 | 0.8576 | 0.8455 |
| (3) seller.email | | 0.9742 | 0.9029 | 0.9372 |
| (4) seller.phone_number | | 0.9903 | 0.9507 | 0.9701 |
| (5) invoice_date | | 0.9940 | 0.9327 | 0.9624 |
| (6) invoice_number | | 0.9538 | 0.9526 | 0.9532 |
| (7) summary.balance | | 0.9054 | 0.9030 | 0.9042 |
| accuracy | 0.9865 | | | |
| macro avg. | | 0.9367 | 0.9294 | **0.9324** |
| weighted avg. | | 0.9871 | 0.9865 | 0.9867 |
| ***GraphDoc* (German BERT): Experiment 2** | | | | |
| (0) none | | 0.9878 | 0.9376 | 0.9621 |
| (1) buyer.bill.address | | 0.3510 | 0.7992 | 0.4877 |
| (2) buyer.bill.customer_id | | 0.3452 | 0.7898 | 0.4804 |
| (3) seller.email | | 0.9350 | 0.8881 | 0.9109 |
| (4) seller.phone_number | | 0.9578 | 0.9207 | 0.9389 |
| (5) invoice_date | | 0.9826 | 0.8806 | 0.9288 |
| (6) invoice_number | | 0.9395 | 0.9421 | 0.9408 |
| (7) summary.balance | | 0.8500 | 0.8522 | 0.8511 |
| accuracy | 0.9296 | | | |
| macro avg. | | 0.7936 | 0.8763 | **0.8126** |
| weighted avg. | | 0.9603 | 0.9296 | 0.9410 |

Table 4.5: Classification report on PD test dataset

| Model/Class | | Precision | Recall | F1-score |
|---|---|---|---|---|
| ***GraphDoc* (English BERT): Experiment 3** | | | | |
| (0) none | | 0.9509 | 0.9115 | 0.9308 |
| (1) adresse | | 0.5837 | 0.7291 | 0.6484 |
| (2) kunden_nr | | 0.5661 | 0.8045 | 0.6646 |
| (3) email | | 0.8516 | 0.8201 | 0.8356 |
| (4) telefon_nr | | 0.9133 | 0.8103 | 0.8587 |
| (5) beleg_datum | | 0.7781 | 0.8704 | 0.8217 |
| (6) rechnung_nr | | 0.7863 | 0.7244 | 0.7541 |
| (7) betrag | | 0.7388 | 0.8049 | 0.7704 |
| accuracy | 0.8838 | | | |
| macro avg. | | 0.7711 | 0.8094 | **0.7855** |
| weighted avg. | | 0.8967 | 0.8838 | 0.8887 |
| ***GraphDoc* (German BERT): Experiment 4** | | | | |
| (0) none | | 0.9829 | 0.9740 | 0.9784 |
| (1) adresse | | 0.8690 | 0.9169 | 0.8923 |
| (2) kunden_nr | | 0.7914 | 0.8271 | 0.8088 |
| (3) email | | 0.9364 | 0.8571 | 0.8950 |
| (4) telefon_nr | | 0.9213 | 0.8410 | 0.8794 |
| (5) beleg_datum | | 0.8438 | 0.9000 | 0.8710 |
| (6) rechnung_nr | | 0.8099 | 0.7717 | 0.7903 |
| (7) betrag | | 0.7063 | 0.8211 | 0.7594 |
| accuracy | 0.9618 | | | |
| macro avg. | | 0.8576 | 0.8636 | **0.8593** |
| weighted avg. | | 0.9629 | 0.9618 | 0.9622 |

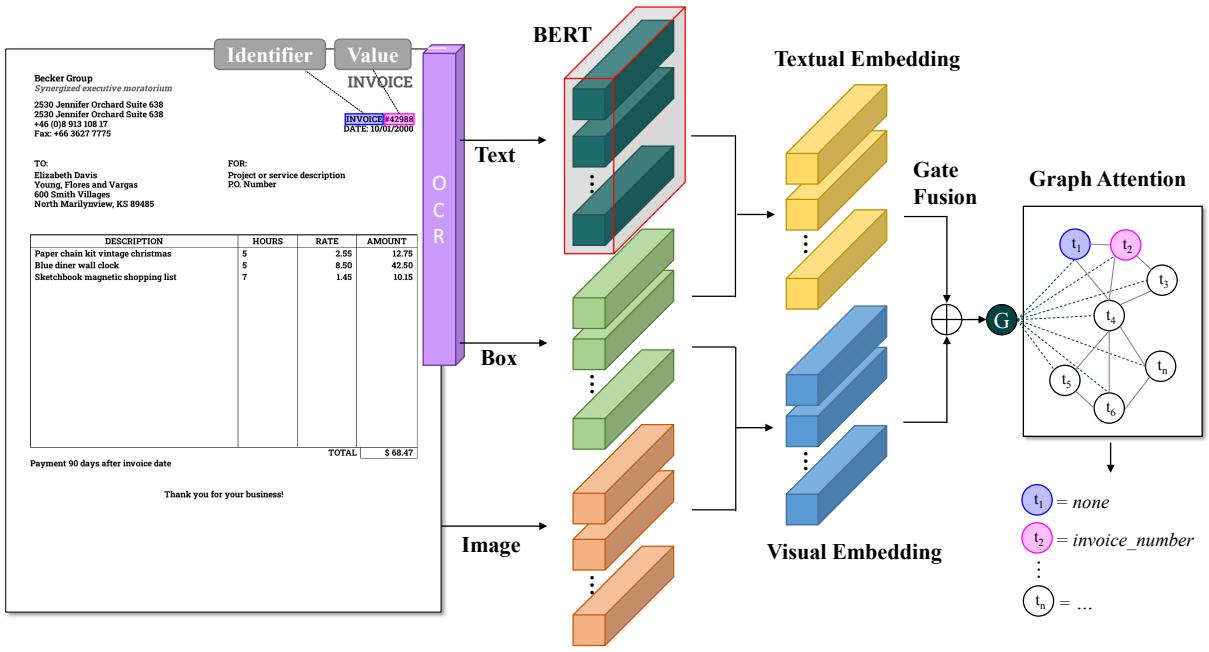


Figure 4.1: Exemplary document from Inv3D [16] testset and model structure adapted from [17]

type, sample size, type and distribution of class labels, as well as variance of layouts. For example, more classes are considered in the DocILE challenge and fewer classes in the SROIE challenge.

All experiments show high accuracy scores. However, the classes are strongly imbalanced and the majority class (*none*) consumes the result. Nonetheless, precision, recall and F1 also achieve good scores. The difference between macro average precision and recall is balanced for all experiments, which shows that the models learn to discriminate the classes despite the class imbalance. Table 4.6 summarizes the macro average F1-scores of the experiments and shows the difference between the experiments in percentage points (pp). The comparison of the F1-scores serves to test our qualitative hypotheses.

Table 4.6: Macro average F1-scores and deltas between the experiments

| Model | Dataset | Inv3D (English) | PD (German) |
|---|---|---|---|
| *GraphDoc* (EB) | | **E1 0.9324** | **E3 0.7855** |
| *GraphDoc* (GB) | | **E2 0.8126** | **E4 0.8593** |



- **H1**: $P_{E1} > P_{E2}$ and $P_{E3} < P_{E4}$. The hypothesis regarding the influence of the network's language module on performance is confirmed in both comparisons. In both cases, there is a significant delta between the experiment with a language fit and the experiment without language fit. In the case of Inv3D almost 12pp.

This means that a language module that matches the dataset has a strong positive influence on the classification results, if the dataset is maintained. It is interesting to note that the effect of the language module is stronger for synthetic data than for real-world data. Hypothesis 1 reflects the real-world use case, as the dataset is not usually exchanged, but the model is adapted.

- **H2**: $P_{E1} > P_{E3}$, $P_{E2} > P_{E3}$, and $P_{E1} > P_{E4}$. Our hypothesis about the combined impact of data quality and language fit is also confirmed. In general, better results are achieved with synthetic data. The comparison between E1 and E3 shows the largest effect, of approximately 15pp. E2 vs. E3 and E1 vs. E4 also confirm this. In the case of E2 vs. E4, where we expected the effects of language and data quality to counteract, we see that a matching language can outweigh issues with data quality, in fact $P_{E2} < P_{E4}$. However, the total delta between E2 and E4 of 4.67pp is only a third of E1 vs. E3.

As expected, the case of language fit and synthetic data achieves the best results (E1). Unfavorable influences from inferior data quality, e.g., from annotation or OCR errors, can be compensated by adjusting the network's language module. Purely synthetic data provide best training results, as these data inherit synthesized patterns. In real-world scenarios, augmenting the real-word data with synthetic data can benefit the training. As E2 and E3 also achieve good results, it shows that features from the visual embedding also have predictive value in the model, which strengthens the approach of multi-modal inputs. The confusion matrices of our focus experiments E1 (Table 4.7) and E4 (Table 4.8) show that, in general, most misclassifications occur between the majority class *'none'* and each foreground class. For example, the address tokens can be recognized well as an overall construct. This is probably due to the exposed position and size of the tokens in that class. Classes with a syntactic structure such as email, phone numbers, and dates can also be assigned well. Identifier classes, such as invoice numbers or customer ID, also show good classification results in our experiments. This is due to the attention mechanism in the network. The weights in the GAT layers for these tokens show a strong influence of their neighborhood context [8]. This means that the actual invoice number is influenced by the identifier, also represented in Fig.4.1. The total amount class (**summary.balance** and **betrag**) shows confusion with the *'none'* class. This could be due to the fact that several amounts can be found on invoices, but these were labeled as *'none'*.

**Discussion and Future Work**. Our ablation study shows that language modules within a network, that featurizes semantic embeddings, have a strong impact on performance. Overall, the multi-modal approach, which combines semantic embeddings, image elements, and layout structure, shows promising results on both datasets. The approach of using the pre-trained models also proves to be expedient, as good results are consistently achieved. Nevertheless, our training is a fine-tuning, so that a full training including the optimization of hyperparameters of the models can possibly lead to even better results in the future. Although the original *GraphDoc* model is designed for RoI, i.e., paragraphs, text blocks, tables, etc., it also has good performance in the case where *"both the region-level and word-level boxes are the same"* [17]. This means that modeling at token level is perfectly permissible. However, it is to be expected that preprocessing RoI instead of word-level tokens can achieve even better results. This would also support the actual approach of the BERT models, as they should embed the sentence context and not only single words. Our approach of measuring the influence of data quality via the

Table 4.7: Confusion matrix of E1

| | | Prediction | | | | | | | | Support |
|---|---|---|---|---|---|---|---|---|---|---|
| **Actual** | | ⓪ | ① | ② | ③ | ④ | ⑤ | ⑥ | ⑦ | |
| none | ⓪ | 374502 | 2401 | 108 | 12 | 25 | 11 | 10 | 345 | 377414 |
| seller.address | ① | 585 | 13899 | 22 | 8 | 2 | 4 | 2 | 211 | 14733 |
| b.b.customer_id | ② | 71 | 34 | 873 | 2 | 9 | 5 | 3 | 21 | 1018 |
| seller.email | ③ | 99 | 4 | 1 | 1097 | 0 | 0 | 4 | 10 | 1215 |
| s.phone_number | ④ | 244 | 3 | 5 | 2 | 5923 | 0 | 45 | 8 | 6230 |
| invoice_date | ⑤ | 306 | 10 | 8 | 3 | 3 | 5969 | 0 | 101 | 6400 |
| invoice_number | ⑥ | 97 | 10 | 18 | 2 | 17 | 6 | 3094 | 4 | 3248 |
| summary.balance | ⑦ | 572 | 38 | 12 | 0 | 2 | 10 | 86 | 6700 | 7420 |

Table 4.8: Confusion matrix of E4

| | | Prediction | | | | | | | | Support |
|---|---|---|---|---|---|---|---|---|---|---|
| **Actual** | | ⓪ | ① | ② | ③ | ④ | ⑤ | ⑥ | ⑦ | |
| none | ⓪ | 31574 | 711 | 7 | 4 | 12 | 41 | 7 | 62 | 32418 |
| adresse | ① | 405 | 4765 | 6 | 6 | 1 | 2 | 3 | 9 | 5197 |
| kunden_nr | ② | 12 | 2 | 110 | 0 | 2 | 0 | 5 | 2 | 133 |
| email | ③ | 20 | 5 | 0 | 162 | 0 | 0 | 0 | 2 | 189 |
| telefon_nr | ④ | 49 | 0 | 8 | 0 | 328 | 0 | 4 | 1 | 390 |
| beleg_datum | ⑤ | 19 | 0 | 4 | 1 | 0 | 243 | 0 | 3 | 270 |
| rechnung_nr | ⑥ | 16 | 0 | 4 | 0 | 4 | 0 | 98 | 5 | 127 |
| betrag | ⑦ | 29 | 0 | 0 | 0 | 9 | 2 | 4 | 202 | 246 |

exchange of the dataset is only valid to a limited extent. With a different approach, e.g., by manipulating the labels in the original dataset, this effect could be investigated in more isolated form. Nevertheless, the observation that in one of our experimental comparisons the effect of language fit outweighs inferior data quality (E2 vs. E4) is quite interesting. This effect indicates that the capabilities for generating automatic embeddings can partly compensate issues with data quality. Nevertheless, annotation and preprocessing quality must still be assigned a high level of importance. There are multiple possibilities to continue this study. One option is to integrate further class labels of the datasets, e.g., for line-item recognition and to generate further synthetic documents via the Inv3D pipeline, so that the variance of the dataset increases. An investigation of other language modules, such as GBERT and GELECTRA [24], could provide further insights into the influence of semantic embeddings in invoice recognition and also promise even better performance, as they achieve state-of-the-art performance across document classification and named entity recognition tasks. Also, manipulating the datasets (annotation, text, boxes), and investigating different entity levels (character, word, region) with different OCR engines could yield further insights. While the confirmation of the hypotheses appears to be straight-forward, it is nevertheless important to deal with the conclusions drawn from them. It means, for example, that competitions with a generally higher data quality show better results. However, since models should be valid beyond the benchmark dataset, a

validation on unseen data from different datasets is essential — considering consistent requirements for comparison, e.g., number and type of prediction classes. The key takeaway from our experiments is that the right choice of language model can partially compensate for problems with data quality. The multi-modal combination of specialized models, e.g., (large) language models for semantic tasks, graphs for structure related tasks, or convolutional networks for image tasks, is a promising research path. We therefore argue that the use of multi-modal and multi-model approaches, in the sense of ensemble learning, can achieve highest generalizability on unseen datasets.

**Conclusion**. In this paper, we conduct an ablation study in which we exchange the language module (BERT) of a state-of-the-art document understanding model, namely multi-modal pre-trained *GraphDoc*. The original and the modified model are fine-tuned on two different datasets, a perfect synthetic (Inv3D) and a real-world dataset (PD). The classification metrics achieve good results comparable to other benchmark approaches. Our hypotheses that both a matching language model and the annotation quality have a significant influence on macro F1-score performance are confirmed. Our contribution is twofold. Firstly, we analyze and confirm the performance of a multi-modal state-of-the-art approach. We show the impact of matching the language of the semantic module to the dataset language. Secondly, we extend the approach with a German language model.

# References


[1] Felix Krieger and Paul Drews. "Leveraging Big Data and Analytics for Auditing: Towards a Taxonomy". In: *ICIS 2018 Proceedings* (2018). URL: https://aisel.aisnet.org/icis2018/datascience/Presentations/16.

[2] Bertin Klein, Stevan Agne, and Andreas Dengel. "Results of a Study on Invoice-Reading Systems in Germany". In: *Document analysis systems VI*. Ed. by David Hutchison. Vol. 3163. Lecture Notes in Computer Science. Berlin: Springer, 2004, pp. 451–462. ISBN: 978-3-540-23060-1. DOI: 10.1007/978-3-540-28640-0_43.

[3] Xiaojing Liu et al. *Graph Convolution for Multimodal Information Extraction from Visually Rich Documents*. Mar. 27, 2019. URL: https://arxiv.org/pdf/1903.11279.

[4] Lukas-Walter Thiée, Felix Krieger, and Burkhardt Funk. "Extraction of Information from Invoices – Challenges in the Extraction Pipeline". In: *Informatik 2023*. Ed. by Maike Klein et al. Lecture notes in Informatics (LNI) Proceedings. Bonn: Gesellschaft für Informatik, 2023, pp. 1777–1792. ISBN: 9783885797319. DOI: 10.18420/INF2023_180. (Visited on 01/10/2024).

[5] R. B. Palm, O. Winther, and F. Laws. "CloudScan - A Configuration-Free Invoice Analysis System Using Recurrent Neural Networks". In: *12th International Conference on Document Analysis and Recognition*. 2013, pp. 406–413. ISBN: 978-0-7695-4999-6. DOI: 10.1109/ICDAR.2017.74.

[6] Sonit Singh. *Natural Language Processing for Information Extraction*. Australia, July 6, 2018. URL: https://arxiv.org/pdf/1807.02383.

[7] Brian Davis et al. "Deep Visual Template-Free Form Parsing". In: *15th International Conference on Document Analysis and Recognition* (2019). URL: https://arxiv.org/pdf/1909.02576.

[8] Felix Krieger et al. “Information Extraction from Invoices: A Graph Neural Network Approach for Datasets with High Layout Variety”. In: *Wirtschaftsinformatik 2021 Proceedings* (2021). URL: https://aisel.aisnet.org/wi2021/RDataScience/Track09/4.

[9] D. Lohani, A. Belaïd, and Y. Belaïd. “An Invoice Reading System Using a Graph Convolutional Network”. In: *ACCV Workshops*. 2018, pp. 144–158. DOI: 10.1007/978-3-030-21074-8_12. URL: https://link.springer.com/chapter/10.1007/978-3-030-21074-8_12.

[10] Xiao Yang et al. *Learning to Extract Semantic Structure from Documents Using Multimodal Fully Convolutional Neural Network*. June 7, 2017. URL: https://arxiv.org/pdf/1706.02338.

[11] Rasmus Berg Palm, Florian Laws, and Ole Winther. “Attend, Copy, Parse - End-to-end information extraction from documents”. In: *ICDAR* (2019). URL: https://arxiv.org/pdf/1812.07248.

[12] Yiheng Xu et al. *LayoutLM: Pre-training of Text and Layout for Document Image Understanding*. 2020. DOI: 10.1145/3394486.3403172. URL: https://arxiv.org/pdf/1912.13318.

[13] Łukasz Garncarek et al. “LAMBERT: Layout-Aware Language Modeling for Information Extraction”. In: *Document Analysis and Recognition - ICDAR 2021* Vol. 12821 (2021), pp. 532–547. DOI: 10.1007/978-3-030-86549-8_34. URL: https://arxiv.org/pdf/2002.08087.

[14] Yang Xu et al. *LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding*. 2020. URL: https://arxiv.org/pdf/2012.14740.

[15] Petar Veličković et al. *Graph Attention Networks*. Oct. 30, 2017. URL: http://arxiv.org/pdf/1710.10903.pdf.

[16] Felix Hertlein, Alexander Naumann, and Patrick Philipp. *Inv3D: a high-resolution 3D invoice dataset for template-guided single-image document unwarping - Metadata*. Karlsruhe Institute of Technology, 2023. DOI: 10.35097/1730. URL: https://publikationen.bibliothek.kit.edu/1000161884.

[17] Zhenrong Zhang et al. *Multimodal Pre-training Based on Graph Attention Network for Document Understanding*. Mar. 25, 2022. URL: http://arxiv.org/pdf/2203.13530.pdf.

[18] D. Sculley et al. “Hidden Technical Debt in Machine Learning Systems”. In: *Advances in Neural Information Processing Systems* 28 (2015). URL: https://proceedings.neurips.cc/paper_files/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf.

[19] Stěpán Šimsa et al. *DocILE Benchmark for Document Information Localization and Extraction*. Feb. 11, 2023. URL: https://arxiv.org/pdf/2302.05658.

[20] Tomas Mikolov et al. *Efficient Estimation of Word Representations in Vector Space*. Jan. 16, 2013. URL: https://arxiv.org/pdf/1301.3781.

[21] Jeffrey Pennington, Richard Socher, and Christopher Manning. “Glove: Global Vectors for Word Representation”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2014), pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://aclanthology.org/D14-1162/.

[22] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Oct. 11, 2018. URL: https://arxiv.org/pdf/1810.04805.

[23] deepset.ai. *German BERT*. 2020. URL: https://www.deepset.ai/german-bert (visited on 01/07/2024).

[24] Branden Chan, Stefan Schweter, and Timo Möller. *German's Next Language Model*. Oct. 21, 2020. URL: http://arxiv.org/pdf/2010.10906.pdf.

# Paper 5

# Enhancing invoice recognition with LLM Embeddings in GAT Networks

## Outline



## Bibliographic Information


Lukas-Walter Thiée and Burkhardt Funk.
Leuphana Universität Lüneburg, Institute for Information Systems, Lüneburg, Germany.
15.08.2025, https://aisel.aisnet.org/amcis2025/sig_svc/sig_svc/6/
AMCIS 2025 Proceedings, Generative AI in Information and Service Systems, Montréal.


## Copyright Notice

# Abstract

We propose a novel approach for invoice recognition by integrating Large Language Model embeddings as semantic features into the nodes of a Graph Attention Neural Network. Both the language model and the graph structure provide rich contextual information for our model to enhance the classification of OCR tokens from invoice documents. The experimental results demonstrate improvements in the classification performance on our datasets by over 3%, highlighting the effectiveness of our multiple attention mechanism. The approach is transferable to all kinds of service systems that process visually rich documents.

**Keywords**

Invoice recognition, GAT, LLM.

# 1 Introduction

State-of-the-art invoice recognition leverages various methods from the field of machine learning (ML), such as natural language processing (NLP), Convolutional Neural Networks CNN, Graph Neural Networks (GNN), or Large Language Models (LLM) [1, 2]. Applying these methods, the goal is to gain structured information, e.g., names, amounts, or dates, and also line items in a key:value format from unstructured or semi-structured data. This type of information extraction and utilization has different fields of application, ranging from banking apps in private use to (fully) automated financial operations in the business sector. Whereas forms and sequential text have defined structures or grammar, invoices represent visually rich documents [3] whose content results from the interaction of various information signals. These signals can include text, numbers, layout, tables, graphics, and colors, as well as latent information, such as conventions for the letterhead. The recognition pipeline from unstructured to structured data poses various steps and challenges, such as optical character recognition (OCR), classification, metric definition, or pre- and post-processing [4]. Current research streams are divided into OCR-free and OCR+model approaches. OCR-free approaches, such as Donut [5], integrate the visual abstraction step of character recognition into the model. The same applies to LLM-based document reading systems or foundational models, which can analyze the content of documents through multimodal inputs and corresponding prompts [6, 7]. Approaches with OCR rely on pre-processed data, i.e., the OCR output, to represent inputs. Subsequently, these inputs are fed into a learning algorithm to predict outputs. Various machine learning techniques can be used for the training. One promising approach for this is the use of GNNs, specifically Graph Attention Networks (GAT) [3, 8].

We see a research gap in the fact that generative LLMs have not yet been integrated as a semantic component in multimodal GAT approaches to invoice recognition. In this paper, we present an enhancement of an OCR+model approach. We leverage a pre-trained multimodal model that combines positional, visual, and semantic embeddings from OCR and image inputs in a GAT. The original model (*GraphDoc*) was trained and published by Z. Zhang et al. [9]. Instead of using a BERT model (Bidirectional Encoder Representations from Transformers) [10], we now apply LLaMA 2 (Large Language Model Meta AI) [11] as the semantic part of the model. Both the language model and the graph structure provide rich contextual information for our classification pipeline. In both the LLM and the GAT, attention mechanisms are used to map the intricate relationships between the input items. We expect this multiple attention mechanism to improve the classification of OCR tokens from invoice documents. Our research is motivated by the goal of improving the classification results in invoice recognition pipelines and we pose the research question *if an LLM, such as LLaMA2, can increase performance with respect to the classification results of a BERT language model in a graph setup.* We strongly believe that rich feature representations in GAT nodes are beneficial for the classification task. In our experiments, we show that our model increases classification results on our selected datasets. With our approach we contribute to the improvement of invoice recognition through the harmonization of language and layout embeddings and offer a model that can be transferred to general supervised document analysis.

# 2 Related Work

## 2.1 Invoice Recognition

Invoice recognition, as a subdomain of document analysis, is the systematic process of automating the identification, extraction, and semantic interpretation of semi-structured and unstructured data from invoice documents. Utilizing advanced computational techniques such as optical character recognition, natural language processing, and machine learning models, this process facilitates the transformation of visually rich documents into machine-readable formats [3, 12, 13]. invoice recognition plays a pivotal role in streamlining financial operations, particularly in the context of accounts payable and enterprise resource planning (ERP) systems, by minimizing manual data entry, reducing error rates, and optimizing processing efficiency [14]. This technological advancement is integral to the broader domain of intelligent document processing, contributing to enhanced automation and precision in financial and administrative workflows [15]. Furthermore, not only database systems, also user-centric apps, such as mobile banking apps, form a field of application [16]. The difficulty in recognizing invoice information lies in the fact that the meaning of various elements often only emerges from the interaction of different signals. These signals may be of a segmental, syntactic, semantic, spatial, external, graphical, or logical nature [4]. The interaction of these types of information can be shown well in tabular representations. The table as an independent segment is generally based on spatial (columns, rows, cells), graphical (borders, connectors) and logical (headers, sums) information types. The message of the table only becomes clear through the interaction of the different types of information, e.g., a bold value or a cell in a specific column. To map the relationships, models are needed that take into account the relationship between different elements. Furthermore, rich feature embeddings are required that integrate as many of the input signals as possible. Both are available in Graph Neural Networks.

## 2.2 Graph Neural Networks

Graph Neural Networks (GNNs) have emerged as a powerful instrument for modeling data with graph structures, where relationships between entities are represented as nodes and edges. GNNs leverage message-passing mechanisms to propagate and aggregate information across graph nodes, enabling the learning of node, edge, or graph-level representations. This architecture has been instrumental in domains such as social network analysis, molecular graph modeling, and recommendation systems [17]. A critical advancement in GNNs is the introduction of Graph Attention Networks (GATs), which incorporate attention mechanisms to dynamically assign importance weights to neighboring nodes during message passing. Unlike traditional GNNs, GATs enhance flexibility by focusing on the most relevant neighbors, improving performance in heterogeneous or complex graphs [18]. The attention mechanism in GATs ensures scalability and interpretability, addressing key challenges in graph-based learning. Together, GNNs and GATs have redefined how relational data is processed, with applications spanning natural language processing, protein interaction networks, and fraud detection [19]. Examples of the use of GNNs in the context of invoice recognition can be found in [20] and [12], where Graph Convolutional Neural Networks are applied or in [3] and [9], who leverage Graph Attention Networks. A prerequisite for GNNs to perform well is that they are fed with sufficiently good input

features, such as semantic embeddings from LLMs.

### 2.3 Large Language Models

Recent advancements in large language models, like the introduction of GPT-4, have significantly influenced research in and application of AI. LLMs leverage deep learning architectures - primarily transformer-based models - to understand, generate, and manipulate human language at scale. Built upon vast corpora of text and trained using self-supervised learning techniques, LLMs such as GPT [21] or T5 [22] have demonstrated remarkable performance across a broad range of NLP tasks, including text summarization, translation, question answering, and document understanding [23]. One of the core strengths of LLMs is their ability to generalize across tasks with minimal task-specific tuning - a property known as few-shot or zero-shot learning [21]. Furthermore, due to their deep contextual understanding, LLMs are particularly effective in handling noisy or unstructured data, making them well-suited for complex information extraction tasks, such as invoice parsing and document classification [24]. Some of the models are being open-sourced, such as LLaMA2, introduced by Meta AI in 2023, which has promoted significant innovation and evaluation in the research community. By providing access to a highly capable and scalable LLM, researchers were encouraged to explore new applications and integrations, including potential synergies with Graph Neural Networks. LLaMA2 represents a major milestone in the development of scalable and efficient LLMs, building upon its predecessor, LLaMA, by incorporating enhanced training datasets, improved fine-tuning techniques, and architectural refinements. It demonstrates state-of-the-art performance across various NLP tasks, while also being computationally efficient and scalable to a wide range of deployment scenarios. LLaMA2 is *"a collection of pre-trained and fine-tuned large language models (LLMs) ranging in scale from 7 billion to 70 billion parameters* [11]. LLaMA2's contributions are notable in the context of understanding and processing large, complex datasets - a challenge that overlaps with the focus of GATs. Specifically, LLaMA2 employs transformer architectures capable of modeling intricate relationships within textual data and captures dependencies and interactions within sequences of text. The model's improvements in token-level attention mechanisms and its emphasis on resource-efficient training align conceptually with the attention paradigms used in GATs, where selective attention to graph nodes enables the capture of meaningful relationships without excessive computational overhead. While LLaMA is set up as a decoder-only network for next token generation, it is still valid to use it to get semantic embeddings because its internal hidden states represent deep, contextual understanding of language. Even though it is designed primarily for left to right text generation, internally it must understand the input sequence very deeply to make good predictions. Each token, as it passes through the layers, gets transformed into a dense vector that captures meaning, syntax (grammar), and semantics (relationships, implications).

## 3 Methodology

We apply a pre-trained multimodal Graph Attention Network (*GraphDoc*) for invoice recognition, which was originally designed, trained, and published by [9]. The model presented in *"Multimodal Pre-training Based on Graph Attention Network for Document*

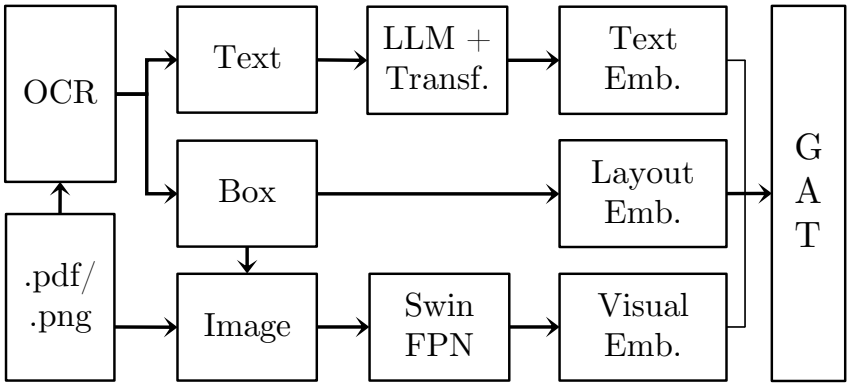


Figure 5.1: Pipeline schema

*Understanding"* was pre-trained on 320k unlabeled documents from PubLayNet [25] using a masked sentence modeling technique. The model was fine-tuned and evaluated on different datasets, including RVL-CDIP [26], FUNSD [27], SROIE [28], and CORD [29], showing excellent results. Further research using this model confirmed a strong relationship of classification results with the applied language model [30]. The model consists essentially of three input branches, namely textual, layout, and visual inputs. Figure 5.1 shows the structure of our adapted pipeline. The first branch are the bounding boxes of the OCR tokens (Figure 5.1, middle). In our case, we consider tokens at the word-level, that is, whole words or fragments, numbers, or alphanumeric segments that have been captured by OCR. The second branch are the visual inputs (bottom). Here, the corresponding sections of the document are fed into a visual encoder using the bounding boxes. A Swin Transformer [31] with a Feature Pyramid Network (FPN) [32] builds the visual encoder that produces visual embeddings for regions of interest. The third branch is the semantic input (top). In the original model, a BERT model is applied to generate the semantic embeddings from paragraphs of text. Since we want to use the entire document context when generating the semantic representations, we apply LLaMA as the language model in this paper, in particular LLaMA2 7b Chat from Hugging Face, and use concatenated OCR tokens as input sequence. LLaMA2 is publicly available and the version of 7b parameters has the lowest computational overhead for our embedding inference. Although the sequence does not necessarily reflect a grammatically correct sentence, we assume that the LLM accounts for missing structure and surface-level errors [33], and generates contextualized word embeddings that already embed essential information in one of the three input branches of the model. While LLaMA2 works best for English language and prompts, it still performs well across other languages [34]. We apply it to two different datasets, one with English and one with German invoice documents.

While *GraphDoc* was pre-trained with a hidden feature size of 768 and LLaMA2 outputs word vectors with a context length of 4096, we need to reduce the embedding size to fit the network. Therefore, we implement three additional transformer encoder layers, with output size 2048, 1024 and 768, after the LLM to gradually reach the pre-trained dimension of 768. The semantic, visual, and layout encodings are merged in a gate fusion layer and attached to the individual nodes of the graph as a node feature. The construction of the graph is based on the original model with 12 attention heads. The adjacency of elements is calculated by *"the top-k nodes nearest to node i (including itself) according to the Euclidean distance"* [9]. We maintain a neighborhood size k of 36 and

we use *GraphDocForTokenClassification* as network tail for output classification. Figure 5.2 depicts the structure of the model layers.

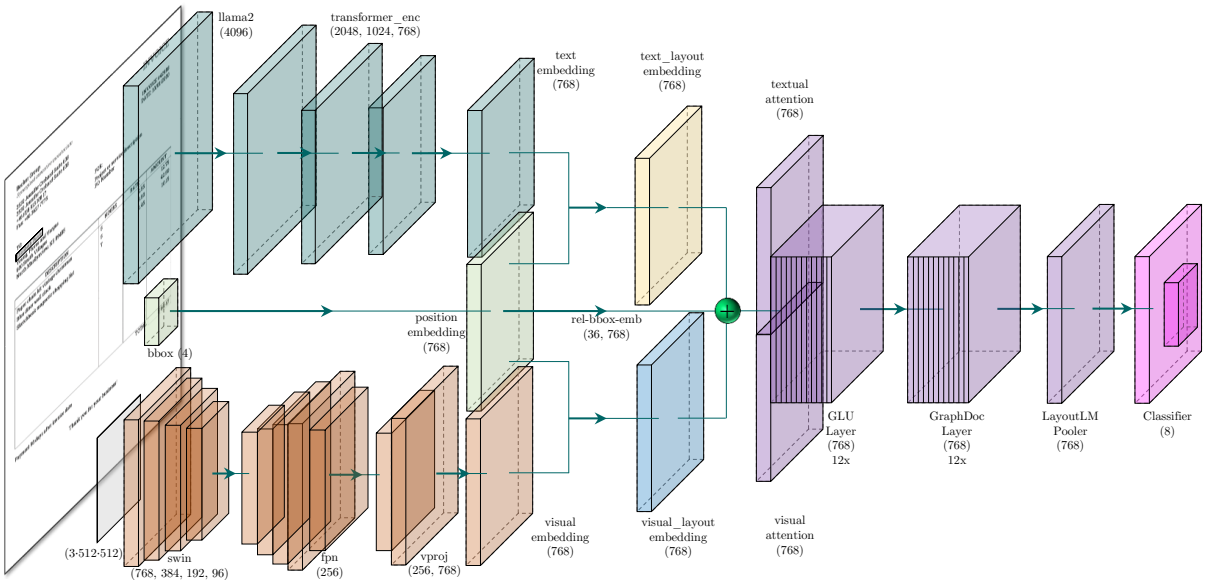


Figure 5.2: Simplified model layers (see Appendix D)

# 4 Experiments and Results

We train and test on two datasets, Inv3D [35] and a private dataset (PD), both of which were also applied in [30]. Inv3D is a synthetic dataset that is based on 100 real template invoice layouts. The original dataset provided consists of several thousand documents in English that contain more than 160 classes. Since the data are synthesized, there is no need for additional OCR steps. In order to have comparability among our experiments, we limit the dataset to 1000 train and 300 test samples and 7 specific classes, namely *Seller Address, Buyer Bill Customer ID, Seller Email, Seller Phone Number, Invoice Date, Invoice Number, and Summary Balance*, cf. Table 5.1. From our second dataset, a private dataset (PD) of German invoices, we use the 7 classes that have a corresponding meaning, namely *Adresse, Kundennummer, Email, Telefonnummer, Beleg Datum, Rechnungsnummer, and Betrag*. The OCR tokens from this dataset were generated with *AbbyyFineReader*. We use 671 documents for training and 169 for testing. This makes the two datasets roughly the same size in terms of the number of tokens to be classified during testing ($\sim 33k$). For both datasets, a batch size of 4, a learning rate of $5 * 10^{-5}$, and 50 training epochs are applied [9]. During training, files are randomly shuffled. The macro F1-score is used to evaluate the performance of the model on the datasets and compared to baseline performances of *GraphDoc* (GD) with English BERT (EB) $GD_{EB,Inv3D} = 0.9324$ for Inv3D and German BERT (GB) $GD_{GB,PD} = 0.8593$ for PD from [30]. The results presented in the table showcase classification performance across several metrics. Table 5.1 lists the results for our experiment with the Inv3D dataset. Table 5.2 lists the results for our experiment with the PD.

Starting with Table 5.1, our GAT model achieves an overall macro F1-score of 0.9699, surpassing the baseline of $GD_{EB,Inv3D} = 0.9324$ by over 3%. This improvement is mirrored

Table 5.1: Classification results Inv3D

| Class | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| none | 0.9951 | 0.9877 | 0.9914 | 29,558 |
| buyer.bill.address | 0.8859 | 0.9420 | 0.9131 | 2,276 |
| buyer.bill.cust_id | 0.9878 | 0.9310 | 0.9585 | 87 |
| seller.email | 1.0 | 1.0 | 1.0 | 88 |
| seller.phone_num | 0.9139 | 0.9922 | 0.9514 | 514 |
| invoice_date | 0.9399 | 0.9918 | 0.9651 | 489 |
| invoice_number | 0.9695 | 0.9960 | 0.9826 | 256 |
| summary.balance | 0.9949 | 1.0 | 0.9974 | 594 |
| accuracy | | | 0.9849 | 33,862 |
| macro avg | 0.9609 | 0.9801 | **0.9699** | 33,862 |
| weighted avg | 0.9855 | 0.9849 | 0.9851 | 33,862 |

Table 5.2: Classification results PD

| Class | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| none | 0.9683 | 0.9796 | 0.9739 | 26,711 |
| adresse | 0.9163 | 0.8700 | 0.8925 | 4,116 |
| kunden_nr | 0.9420 | 0.9348 | 0.9384 | 261 |
| email | 0.9047 | 1.0 | 0.9500 | 133 |
| telefon_nr | 0.9251 | 0.9293 | 0.9272 | 439 |
| beleg_datum | 0.8897 | 0.8861 | 0.8879 | 246 |
| rechnung_nr | 0.856 | 0.7753 | 0.8136 | 276 |
| betrag | 0.8899 | 0.8133 | 0.8499 | 1,034 |
| accuracy | | | 0.9575 | 33,216 |
| macro avg | 0.9115 | 0.8985 | **0.9042** | 33,216 |
| weighted avg | 0.9569 | 0.9575 | 0.9570 | 33,216 |

in accuracy, recall and precision points. These scores suggest that the model is both highly accurate overall and well-balanced in its treatment of different classes. For the majority class (*none*), the F1-score is 0.9914. The foreground classes also perform well, particularly *seller.email, invoice_number*, and *summary.balance*, with F1-scores of 1.0, 0.9826, and 0.9974, respectively.

As e-mail addresses have a clear structure, it seems reasonable, that these tokens are identified perfectly. However, the model does show some limitations when dealing with minority classes such as *buyer.bill.customer_id*, which has an F1-score of 0.9585. This lower score is primarily due to a recall of only 0.9310, even though it has better precision at 0.9878, indicating difficulty in consistently identifying instances of this class despite high specificity. In Table 5.2, the classification model achieves an overall macro F1-score of 0.9042, which exceeds its baseline $GD_{GB,PD} = 0.8593$ by over 4%. As expected, due to lower dataset quality, the metrics are weaker compared to Table 5.1. Among the minority classes, some, like *email* and *kunden_nr*, perform relatively well, with F1-scores of 0.9500 and 0.9384, respectively. However, certain classes such as *rechnung_nr* and *betrag* exhibit notable weaknesses, with F1-scores of 0.8136 and 0.8499, driven by lower recall. While both models perform well for their respective majority classes, Table 5.1 excels in handling several key classes, such as *invoice_number, rechnung_nr*, and *summary.balance*. Table 5.2 shows more variability in its class-wise metrics and struggles, particularly with minority classes.

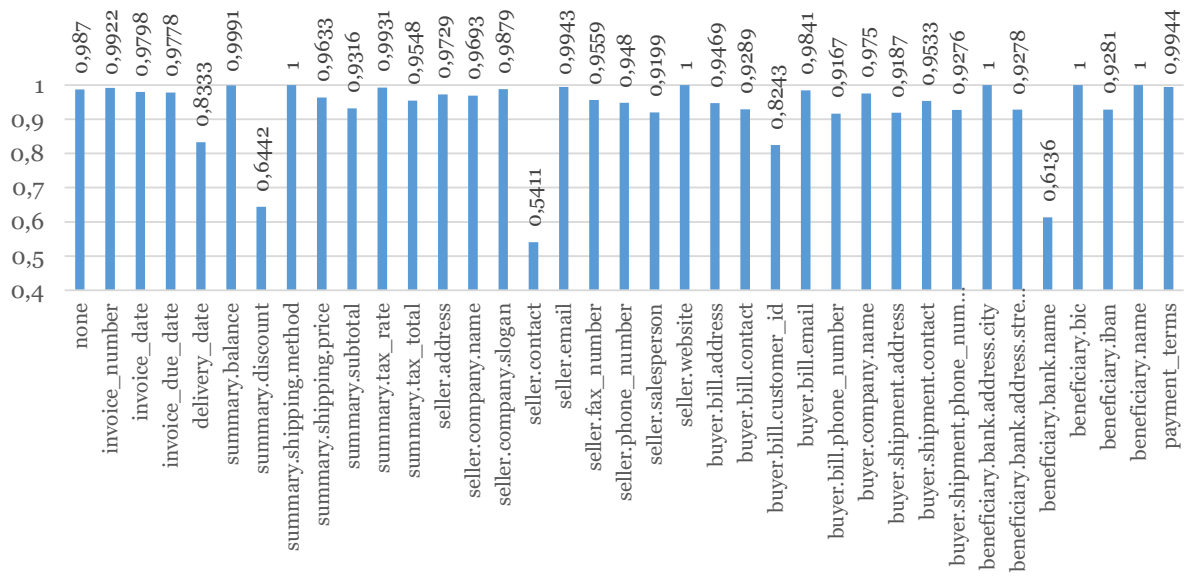


Figure 5.3: F1-scores for Inv3D testset (key classes)

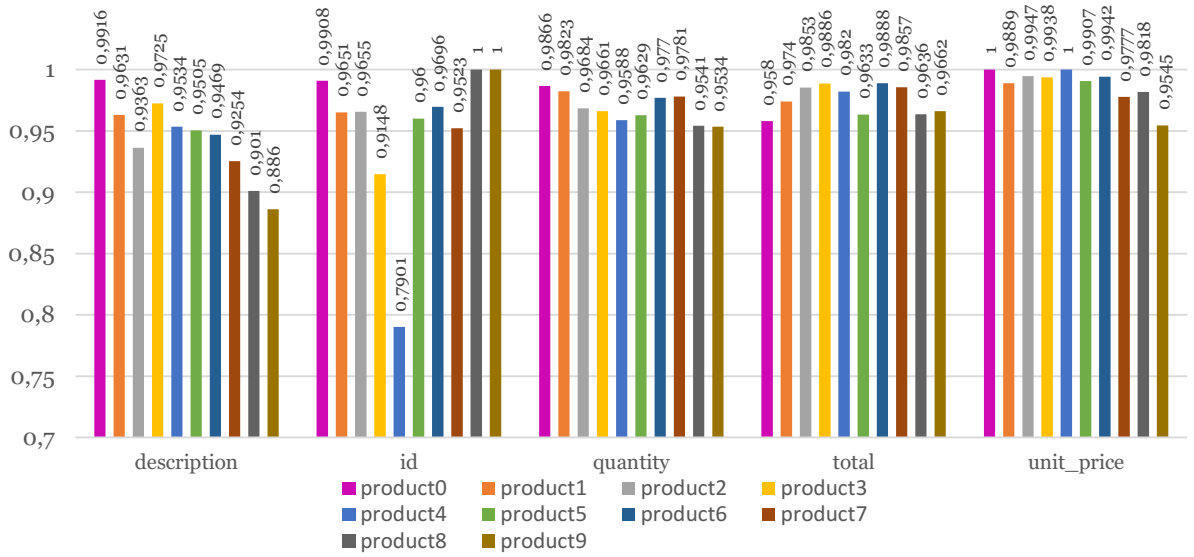


Figure 5.4: F1-scores for Inv3D testset (line items)

As the Inv3D dataset contains labeled line items, e.g., *product id, description, unit price*, and *total*, we retrain the model for an adjusted set of classes with up to 10 line items. We use the same training parameters as mentioned before. Classification results are displayed in Figure 5.3. Compared to the first training, the classes that have been retained even show slightly improved scores, e.g., *summary.balance*. In Figure 5.4 we display the F1-scores for the 10 line items. The line item *total, quantity* and *unit price* perform well on average. *Description* and *ID* show more variability.

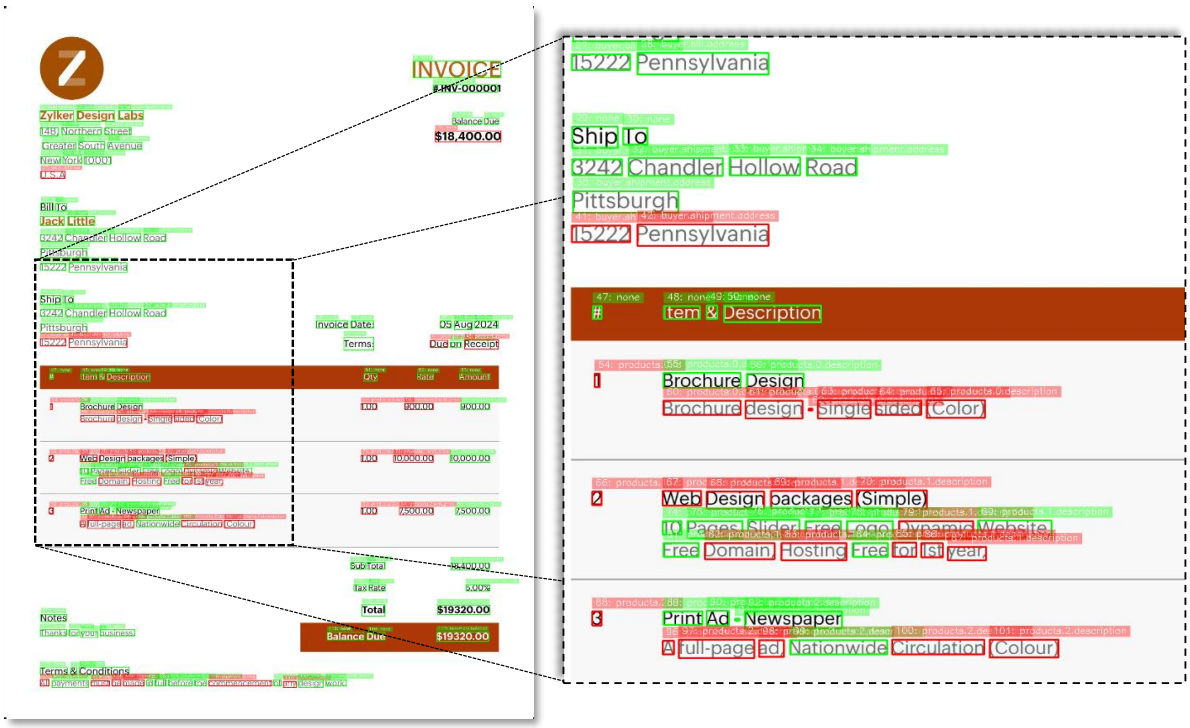


Figure 5.5: Out-of-sample prediction

To analyze the generalizability of the method, we apply the Inv3D model, trained on key classes and line items, on an out-of-sample invoice document. The document is randomly selected from an online search of invoice documents. We manually label the contained classes and apply a minimum measure of preprocessing steps, such as document resizing and filtering of OCR tokens with above-average width or height. We feed this document into the model and predict the class labels. Figure 5.3 shows the prediction results. A green border illustrates the prediction of the right class and red border a false prediction. The full size image shows a lot of green bounding boxes 5.5. From a total of 135 tokens 93 are predicted correctly. This includes key classes like *invoice number, invoice date* and *summary balance*. Most of the address tokens are also classified correctly. We also see a good bit of line items to be predicted as the correct line item. Although 68% accuracy does not exactly look like zero-shot capability, most classes are still recognized. However, the model struggles to identify the longer descriptions of the products coherently. We see that it recognizes the descriptions as line item descriptions, but assigns the wrong line item *ID* to them.

# 5 Discussion and Conclusion

The results presented in this study provide several key insights into the integration of LLM embeddings with Graph Attention Networks for invoice recognition tasks. One of the most notable findings is the marked improvement in classification performance when incorporating LLM embeddings, as compared to benchmark experiments. The integration of LLaMA2 and the transformer layers improves the classification results, which answers our research question. Once again, we see that attention in neural networks has a major influence on performance [36]. In this study, we focus on LLaMA2 because it supports multi-language functionality and can be run locally as an open-source model. We therefore can access the hidden states of the model and retrieve dynamic contextualized vectors that can be used within our OCR+Model pipeline. This enhancement demonstrates the potential of LLM embeddings to contribute meaningful semantic information, even in domains where numerical data play a significant role, such as invoice documents. However, it is important to note that the handling of numerical values by LLMs is not inherently their strongest capability, which is reflected in slightly lower performance for classes involving numerical amounts. Nevertheless, the overall improvement indicates that the multiple attention mechanism effectively leverages both the contextual richness of LLM embeddings and the relational structure captured by GATs. In this study, we focused on 7 key items and 10 line items, i.e., a considerably reduced scope of the actual possible information content of invoices. It is surprising that the *summary.balance* class is recognized very well in one dataset, whereas the corresponding *betrag* class is recognized significantly worse in the other dataset. The difference in performance between the datasets might be due to differences in dataset quality. Compared to the synthesized data of Inv3D, PD suffers real-world drawbacks in OCR and annotation quality [30]. Despite the advancements, there are limitations to the proposed method. The diffusion effect within the LLM attention mechanism presents challenges, particularly when handling non-sequential text that is concatenated to a sequence. This may lead to a dilution of focus on critical tokens, affecting the model's capacity to accurately capture important relationships in complex document structures. Additionally, the dimensionality constraints of the current embed-

ding space (786 dimensions) may limit the full potential of the approach, suggesting room for optimization. Further improvements can be achieved by increasing the dataset size for training and also by increasing the training variance.

In this study we apply LLaMA2 for semantic embeddings in a multimodal pre-trained GAT for invoice recognition. We improve classification performance for two datasets with respect to their benchmark results (3% and 4%). The multiple attention mechanism plays a critical role in capturing both semantic and relational information, leading to robust performance across diverse classes. These findings highlight the potential of advanced embedding techniques to overcome challenges inherent to document analysis, particularly in contexts where traditional approaches may struggle to generalize. Our approach is transferable to all types of document analysis that need to recognize textual and numerical elements in a visual layout. The results of this study have implications for the field of document analysis in general and in particular for OCR+model based approaches in invoice recognition. By integrating LLM techniques with graph-based learning, this approach bridges the gap between unstructured text and layout-based document representations. Our method can be applied to other document analysis tasks that rely on multimodal input signals. It opens avenues for more nuanced and accurate classification in real-world applications, such as customer data analysis, and beyond. However, the study also underscores the need for further research to refine the method and address its limitations. Looking ahead, several directions for future research could address these limitations and further enhance the method. Increasing the embedding dimensionality to a larger space, such as 4096 dimensions, may allow for richer representations and improved performance. Applying this approach to widely recognized benchmark datasets, such as DocILE [37] would provide a more robust evaluation and facilitate comparisons with existing methods. A direct comparison with Multimodal Large Language Models (MLLMs) [38] with pre-trained encoders could also provide valuable insights. Recent models like LayoutLMv2 [39] and DocFormer [40] have extended traditional LLMs by incorporating both textual and spatial (layout) information, significantly improving performance in document-based applications. These multimodal LLMs highlight the growing trend toward integrating multiple modalities - text, vision, and layout - into unified models capable of more holistic document understanding. Our future research therefore naturally includes a comparison of our approach with the very latest research findings, e.g., DocLLM [41] and LayoutLLM [42]. Furthermore, implementing pre- and post-processing steps as part of a unified end-to-end pipeline could streamline the workflow and enhance efficiency. Another promising avenue involves exploring graph node fusion techniques, which may enable better integration of diverse document features and improve classification accuracy. These steps will be essential in pushing the boundaries of what can be achieved with LLM and GAT-based approaches in the evolving landscape of intelligent document processing.

# References


[1] Faiza Loukil et al. *LLM-centric pipeline for information extraction from invoices*. 2024. URL: https://hal.science/hal-04772570/ (visited on 11/08/2024).

[2] Oshi Varma, Samarth Srivastava, and M. Gayathri. "Technical Invoice Data Extraction System: State of the Art, Research Challenges and Countermeasures". In:

*Ambient Communications and Computer Systems* 356 (2022), pp. 201–210. ISSN: 2367-3389. DOI: 10.1007/978-981-16-7952-0_19. URL: https://link.springer.com/chapter/10.1007/978-981-16-7952-0_19.

[3] Felix Krieger et al. "Information Extraction from Invoices: A Graph Neural Network Approach for Datasets with High Layout Variety". In: *Wirtschaftsinformatik 2021 Proceedings* (2021). URL: https://aisel.aisnet.org/wi2021/RDataScience/Track09/4.

[4] Lukas-Walter Thiée, Felix Krieger, and Burkhardt Funk. "Extraction of Information from Invoices – Challenges in the Extraction Pipeline". In: *Informatik 2023*. Ed. by Maike Klein et al. Lecture notes in Informatics (LNI) Proceedings. Bonn: Gesellschaft für Informatik, 2023, pp. 1777–1792. ISBN: 9783885797319. DOI: 10.18420/INF2023_180. (Visited on 01/10/2024).

[5] Geewook Kim et al. "OCR-Free Document Understanding Transformer". In: *Computer Vision – ECCV 2022*. Ed. by Shai Avidan et al. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland and Imprint Springer, 2022, pp. 498–517. ISBN: 978-3-031-19815-1. DOI: 10.1007/978-3-031-19815-1_29. URL: https://link.springer.com/chapter/10.1007/978-3-031-19815-1_29.

[6] Muhammad Awais et al. *Foundational Models Defining a New Era in Vision: A Survey and Outlook*. July 25, 2023. URL: http://arxiv.org/pdf/2307.13721.

[7] Anni Zou et al. *DOCBENCH: A Benchmark for Evaluating LLM-based Document Reading Systems*. July 15, 2024. URL: http://arxiv.org/pdf/2407.10701.

[8] Chao Zhang et al. "Extract Data Points from Invoices with Multi-layer Graph Attention Network and Named Entity Recognition". In: *2022 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*. Piscataway, NJ: IEEE, 2022, pp. 1–6. ISBN: 978-1-6654-9991-0. DOI: 10.1109/ICAICA54878.2022.9844508.

[9] Zhenrong Zhang et al. *Multimodal Pre-training Based on Graph Attention Network for Document Understanding*. Mar. 25, 2022. URL: http://arxiv.org/pdf/2203.13530.pdf.

[10] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Oct. 11, 2018. URL: https://arxiv.org/pdf/1810.04805.

[11] Hugo Touvron et al. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. July 18, 2023. URL: http://arxiv.org/pdf/2307.09288.

[12] Xiaojing Liu et al. *Graph Convolution for Multimodal Information Extraction from Visually Rich Documents*. Mar. 27, 2019. URL: https://arxiv.org/pdf/1903.11279.

[13] Rasmus Berg Palm, Florian Laws, and Ole Winther. "Attend, Copy, Parse - End-to-end information extraction from documents". In: *ICDAR* (2019). URL: https://arxiv.org/pdf/1812.07248.

[14] Bertin Klein, Stevan Agne, and Andreas Dengel. "Results of a Study on Invoice-Reading Systems in Germany". In: *Document analysis systems VI*. Ed. by David Hutchison. Vol. 3163. Lecture Notes in Computer Science. Berlin: Springer, 2004, pp. 451–462. ISBN: 978-3-540-23060-1. DOI: 10.1007/978-3-540-28640-0_43.

[15] Graham A. Cutting and Anne-Francoise Cutting-Decelle. *Intelligent Document Processing -Methods and Tools in the real world*. Dec. 28, 2021. URL: http://arxiv.org/pdf/2112.14070.

[16] Gianni Fenu and Pier Luigi Pau. "An Analysis of Features and Tendencies in Mobile Banking Apps". In: *Procedia Computer Science* 56 (2015), pp. 26–33. ISSN: 18770509. DOI: 10.1016/j.procs.2015.07.177. URL: https://www.sciencedirect.com/science/article/pii/S1877050915016580.

[17] Zonghan Wu et al. "A Comprehensive Survey on Graph Neural Networks". In: *IEEE transactions on neural networks and learning systems* 32.1 (2021), pp. 4–24. DOI: 10.1109/TNNLS.2020.2978386.

[18] Petar Veličković et al. *Graph Attention Networks*. Oct. 30, 2017. URL: http://arxiv.org/pdf/1710.10903.pdf.

[19] Jie Zhou et al. "Graph neural networks: A review of methods and applications". In: *AI Open* 1 (2020), pp. 57–81. ISSN: 26666510. DOI: 10.1016/j.aiopen.2021.01.001.

[20] D. Lohani, A. Belaïd, and Y. Belaïd. "An Invoice Reading System Using a Graph Convolutional Network". In: *ACCV Workshops*. 2018, pp. 144–158. DOI: 10.1007/978-3-030-21074-8_12. URL: https://link.springer.com/chapter/10.1007/978-3-030-21074-8_12.

[21] Tom B. Brown et al. *Language Models are Few-Shot Learners*. 2020. DOI: 10.48550/ARXIV.2005.14165.

[22] Colin Raffel et al. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *Journal of Machine Learning Research* 21.140 (2020), pp. 1–67. URL: http://jmlr.org/papers/v21/20-074.html.

[23] Yupeng Chang et al. "A Survey on Evaluation of Large Language Models". In: *ACM Transactions on Intelligent Systems and Technology* 15.3 (2024), pp. 1–45. ISSN: 2157-6904. DOI: 10.1145/3641289.

[24] Yiheng Xu et al. *LayoutLM: Pre-training of Text and Layout for Document Image Understanding*. 2020. DOI: 10.1145/3394486.3403172. URL: https://arxiv.org/pdf/1912.13318.

[25] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. *PubLayNet: largest dataset ever for document layout analysis*. Aug. 16, 2019. URL: http://arxiv.org/pdf/1908.07836.

[26] Adam W. Harley, Alex Ufkes, and Konstantinos G. Derpanis. *Evaluation of Deep Convolutional Nets for Document Image Classification and Retrieval*. Feb. 25, 2015. URL: http://arxiv.org/pdf/1502.07058.

[27] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. *FUNSD: A Dataset for Form Understanding in Noisy Scanned Documents*. May 27, 2019. URL: http://arxiv.org/pdf/1905.13538.

[28] Zheng Huang et al. *ICDAR2019 Competition on Scanned Receipt OCR and Information Extraction*. 2021. DOI: 10.1109/ICDAR.2019.00244. URL: http://arxiv.org/pdf/2103.10213.

[29] Seunghyun Park et al. “CORD: A Consolidated Receipt Dataset for Post-OCR Parsing”. In: *Workshop on Document Intelligence at NeurIPS 2019* (2019). URL: https://openreview.net/forum?id=SJl3z659UH.

[30] Lukas-Walter Thiée. “Ablation Study of a Multimodal Gat Network on Perfect Synthetic and Real-world Data to Investigate the Influence of Language Models in Invoice Recognition”. In: *Document analysis and recognition - ICDAR 2024 workshops*. Ed. by Harold Mouchère and Anna Zhu. Lecture Notes in Computer Science. Cham: Springer, 2024, pp. 199–212. ISBN: 978-3-031-70642-4. DOI: 10.1007/978-3-031-70642-4_13. URL: https://link.springer.com/chapter/10.1007/978-3-031-70642-4_13.

[31] Ze Liu et al. *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*. Mar. 25, 2021. URL: http://arxiv.org/pdf/2103.14030.

[32] Tsung-Yi Lin et al. “Feature Pyramid Networks for Object Detection”. In: *CVPR*. 2017, pp. 2117–2125. URL: https://openaccess.thecvf.com/content_cvpr_2017/html/Lin_Feature_Pyramid_Networks_CVPR_2017_paper.html.

[33] Jeehaan Algaraady and Mohammad Mahyoob. “ChatGPT’s Capabilities in Spotting and Analyzing Writing Errors Experienced by EFL Learners”. In: *Arab World English Journal* 9 (2023), pp. 3–17. DOI: 10.24093/awej/call9.1.

[34] Chris Wendler et al. *Do Llamas Work in English? On the Latent Language of Multilingual Transformers*. Feb. 16, 2024. URL: http://arxiv.org/pdf/2402.10588.

[35] Felix Hertlein, Alexander Naumann, and Patrick Philipp. *Inv3D: a high-resolution 3D invoice dataset for template-guided single-image document unwarping - Metadata*. Karlsruhe Institute of Technology, 2023. DOI: 10.35097/1730. URL: https://publikationen.bibliothek.kit.edu/1000161884.

[36] Ashish Vaswani et al. *Attention Is All You Need*. June 12, 2017. URL: http://arxiv.org/pdf/1706.03762.

[37] Stěpán Šimsa et al. *DocILE Benchmark for Document Information Localization and Extraction*. Feb. 11, 2023. URL: https://arxiv.org/pdf/2302.05658.

[38] Shukang Yin et al. *A survey on multimodal large language models*. 2024. DOI: 10.1093/nsr/nwae403. URL: http://arxiv.org/pdf/2306.13549.

[39] Yang Xu et al. *LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding*. 2020. URL: https://arxiv.org/pdf/2012.14740.

[40] Srikar Appalaraju et al. *DocFormer: End-to-End Transformer for Document Understanding*. 2021. DOI: 10.48550/ARXIV.2106.11539.

[41] Dongsheng Wang et al. *DocLLM: A layout-aware generative language model for multimodal document understanding*. 2024. DOI: 10.48550/ARXIV.2401.00908.

[42] Masato Fujitake. *LayoutLLM: Large Language Model Instruction Tuning for Visually Rich Document Understanding*. 2024. DOI: 10.48550/ARXIV.2403.14252.

# Part III

# Appendix

# Appendix A

# ‘Spread the App not the Virus’ - An extensive SEM-approach to understand Pandemic Tracing App Usage in Germany

## Outline



## Bibliographic Information


Lukas-Walter Thiée, Hannes M. Petrowsky, Marie-Lena Frech, David D. Loschelder and Burkhardt Funk. Leuphana Universität Lüneburg, Institute for Information Systems and Institute for Management and Organization, Lüneburg, Germany.
14.06.2021, https://aisel.aisnet.org/ecis2021_rp/123
29. European Conference on Information Systems (ECIS 2021), Marrakesh, Morocco.


## Copyright Notice

# Abstract

The release of the Corona-Warn-App (CWA), a governmental pandemic tracing app to track infection chains related to COVID-19 in Germany, marks an unprecedented situation that offers a unique opportunity for investigating population-wide adoption of novel technology. We develop a conceptual model to investigate the effects and path relationships of multiple constructs related to technology adoption, data security, morality, social influence, trust, and COVID-19 to predict behavioral intentions and actual usage behavior. We use structural equation modeling with the partial least squares method and identify effort expectancy, social influence, prevailing opinions on COVID-19 and the CWA, as well as moral and ethical considerations as the most influential predictors. We are able to explain moderate to high amounts of variance with our model. Our results offer valuable insights for the technology acceptance literature and enable practical recommendations for improving the public communication and elevating user numbers of pandemic tracing apps in Germany.



### Keywords

COVID-19, Coronavirus, Tracing App, Technology Acceptance

# 1 Introduction

The rapid global spread of COVID-19 (coronavirus disease) presents modern society with an extraordinary challenge which is reflected in the dilemma of saving lives by minimizing social contact while maintaining social and economic life at the same time. At a closer look, this challenge is not only a virological or medical issue, but also an ethical and social one, as it marks a great balancing act between personal restrictions and societal benefits [1]. Since COVID-19 seems to occupy a sweet spot between danger and ability to spread, and since no adequate mass treatments or vaccinations are yet available, the only way to contain its spread is to reduce it [2, 3]. In this context, various measures are used to contain the virus, e.g., the use of face masks in public. Although a majority of German citizens agree with these restrictive measures, compliance varies in different societies and different strata, even more so if people are personally affected by losing jobs or neglecting social interaction [1].

One promising measure to reduce the spread of the virus is the tracking and breaking of infection chains by contact tracing. Since conventional manual contact tracing by health authorities is rather slow and requires high deployment of personnel, digital solutions are preferred. Here, the possibility for information technology to contribute with so-called tracing apps opens up [4]. The technical implementation of identification and notification depends on the respective tracing app, e.g., centralized or decentralized storage of IDs [5]. In any case, the goal is to warn users who have had risk contacts and to suggest possible measures, e.g., self-isolation.

In Germany, the official proximity tracing app, called Corona-Warn-App (CWA), was released by governmental authorities on 16th of June 2020 with the goal to decelerate the spread of COVID-19 [6]. The development of the app was commissioned by the German federal government and was conducted through an open-programming project (https://github.com/corona-warn-app/) to build a basis for trust in the app. In cooperation between Robert Koch-Institute, an official health authority, and the two technology companies SAP and Telekom, the app is offered in all relevant mobile app stores. After download, there are no registration or login requirements for users. However, users need to enable localization on their devices and activate tracking in the app. The app is based on the native iOS and Android exposure notification API. It uses low-energy Bluetooth to detect interpersonal encounters and stores temporary IDs onto users' devices to anonymously assess local risk [7]. Positively tested individuals can voluntarily report their infection status in the app. IDs who have had contact with an infected person will then be notified. Although the app has been downloaded more than 21 million times as of November 2020 [8], the willingness to use the app varies substantially within the general public. An online survey from June 2020 shows that about 52% of Germans consider it very or rather unlikely to download the app, with the most common arguments against using the app being concerns regarding the usefulness, data security, potential governmental surveillance, and battery usage [9]. Findings from other countries show similar patterns [10, 11, 12]. Nevertheless, governments worldwide have high hopes for digital tracing apps as a part of their confinement strategies against COVID-19 [13]. Proximity tracing apps of this kind, however, can only really exploit their potential if population-wide acceptance is sufficiently high and notification times are as short as possible [14]. Current studies estimate that more than half of the overall population must use the app in order to achieve effective traceability, with respect to different infection rates, doubling

times, and smartphone use [15]. Consequently, it is crucial to investigate how individuals adopt to the app and identify factors that might increase its usage.

## 1.1 Research Gap & Goals

Current research examines potential app adoption with regard to app design specifications and app store description [16], highlighting the societal context of tracing apps and different propensities within the population. Furthermore, some studies provide a theoretical conceptualization of IT governance for rapid population-wide adoption of tracing apps and describe mass adoption as a collective action problem on a societal level [13]. This collective action problem is one of the main reasons for the exceptionality of the situation, because the effectiveness of the app only occurs when a large mass of a society installs the app. This creates the difficulty of convincing an individual of the usefulness, which, however, is not given for an individual alone, but only arises through mass adoption. A so-called "free rider" mentality, i.e., an individual relies on everyone else to participate, but refuses to participate himself, could undermine the whole concept.

Whereas prior studies on tracing apps have already examined the acceptance of different hypothetical app characteristics [17], potential privacy and surveillance concerns [18], and the willingness to adopt future pandemic tracing apps [19], we conduct a survey regarding a truly existing app, the official German Corona-Warn-App, available for public download on all relevant mobile app stores. Related work in the field of technology acceptance often only measures behavioral intentions to use a technological product or service, works with hypothetical, non-existing apps or technologies, or combines both these characteristics. Since the CWA is already released in Germany and already features a considerable number of users, we measure actual behavior, namely actual download of the CWA. We see this as a clear benefit of our research approach, since we are able to capture actual behavioral data for a real and publicly known application, rather than limiting our insights on the behavioral intention for the usage of hypothetical services. German authorities act as one of the international pioneers with regard to governmental pandemic tracing apps, which offers the opportunity to generate empirical data for an unprecedented pandemic situation.

We contribute to the research community by collecting and publishing needed empirical data [20]. Building upon prior findings, and based on existing theories on technology acceptance [21] and privacy concerns [22], we conduct an online survey and seek to establish a conceptual model regarding the usage intention and usage behavior for governmental pandemic tracing apps in Germany (i.e., CWA). The analyses offer insights into the relative importance of different constructs for predicting individuals' proclivity (vs. reservations) to use the CWA. So far, research does not provide a comprehensive model for app adoption in a pandemic. Therefore, we take an explorative approach to test a multitude of relevant influencing factors in this context. The primary goal of our research is to identify the most influential factors that predict individuals' usage intention and actual CWA usage and to draw practical implications for respective policy makers in order to increase app uptake within the population to ultimately decelerate the spread of COVID-19.

Our research benefits the appropriate communication of the app and might help to improve functionality itself. In addition to shedding theoretical light on the adoption of digital tracing apps, this approach also promises to support respective campaigns to

increase app uptake and overcome inertia in app adoption. Our model contributes to the academic and political discourse on measures to contain COVID-19. Finally, our findings might not only support app adoption, but also yield understanding to societal relations in a pandemic and the acceptance of various measures in an overall confinement strategy.

## 2 Theoretical Background

Technology acceptance models (see [23, 21]), which are anchored in the theory of reasoned action [24] and the theory of planned behavior [25], are widely used in IS research and have often been adapted for multiple purposes [26]. However, these universally applicable models also feature some limitations, such as a general lack of specificity for particular use cases, an insufficient inspection of group processes and collective intentions, and the neglection of underlying goals that trigger technology adoption [27]. Since the COVID-19 pandemic with the accompanying mass adoption of pandemic tracing apps marks a unique situation that is heavily influenced by group processes and a multitude of motivations for app usage, we believe that this use case transcends the pre-existing models of technology acceptance.

We therefore propose that the current models for general technology adoption need to be complemented by other theoretical accounts and constructs to represent the complex societal reality and decision-making processes in this case, and to offer a sufficient descriptive and predictive validity. To accomplish this goal, we base our research model on several different models, predictors, and interrelations that are derived from existing theories [28] and briefly introduced in the following section. In Figure A.1, we provide a higher-level explanatory overview of our research design. For a more detailed description, we refer to the literature listed.

**General Technology Acceptance (UTAUT) and Technical Affinity**. Since the adoption of the CWA is in its core a matter of technology acceptance, we base our model on the Unified Theory of Acceptance and Use of Technology (UTAUT) model [21], one of the most renowned models in this field. The UTAUT was developed by testing and aggregating several constructs of pre-existing technology acceptance models. The model assumes that there is a link between personal attitudes, intention to do something, and the actual behavior of an individual, which we echo (H1, see figure A.1).

The three main predictors of the intention to accept a technology are performance expectancy, effort expectancy, and social influence. Performance and effort expectancy map the potential personal benefit, respectively the degree of expected difficulties accompanying the use of a novel technology. Social influence accounts for the amount of peer pressure an individual experiences when deciding in favor or against the use of a new technology. Further, 'facilitating conditions' influence the intention and the actual behavior. The UTAUT has shown a great predictive validity across many technology-related examinations, for instance, user adoption of mobile commerce [29], mobile banking [30, 31] and social media [32], as well as technology adoption of health information systems [33]. We therefore propose that its core constructs will also play a major role for the acceptance of pandemic tracing apps, especially social influence (H2).

We base our model on the original UTAUT instead of UTAUT2 [34] since the three novel aspects of UTAUT2, namely price value, habit, and hedonic motivation, are in large parts not applicable to the case of a freely available app that is based on passive activation

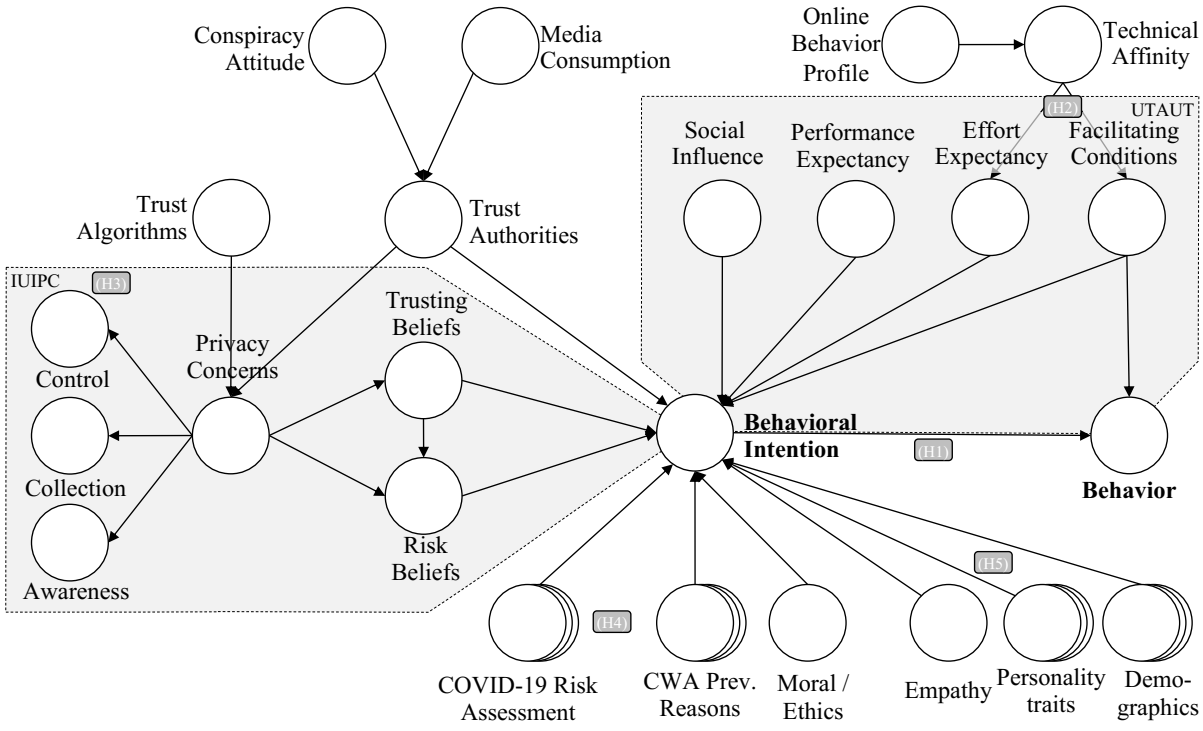


Figure A.1: Simplified overview of the research design

with limited active user interaction. Since the CWA is offered for free, we also cannot account for transition or sunk costs, that have been proposed as influencing factors in other technology acceptance literature [35]. Not even the case of "zero costs", i.e., free use of a service in exchange for personal data [36], applies here, since no personal data is being stored and there is no commercial interest on the part of the providers. Therefore, instead of looking at these inapplicable hedonic or cost-related facets of technology acceptance, we extend the UTAUT by incorporating measures for technical affinity in our research design. Technical affinity has been shown to be an important antecedent of technology adoption in other use cases (see [37, 38]), and we therefore propose that it also serves as an antecedent of effort expectancy and facilitating conditions in our model. We further complement technical affinity by measuring an 'online behavior profile' that maps the use of online services, such as social networks, messengers, cloud services, or AI assistants. We propose that a strong use of online services reflects individual technical ability, which may act as a facilitating condition in the UTAUT and ultimately decrease individual effort expectancy.

**Privacy Concerns (IUIPC) and Trust Motives**. As the UTAUT was proposed as a general technology acceptance model, it does not cover data security concerns. Therefore, regarding CWA-related privacy concerns, we utilize another well-known and previously confirmed model, namely the Internet Users' Information Privacy Concerns (IUIPC) model to account for our specific application case [22]. This allows us to address overall technology acceptance on the one hand while also accounting for the influence of data security and privacy concerns on the decision to use the CWA. The IUIPC model by Malhotra, Kim and Agarwal (2004) [22] was established to explain the influence of privacy concerns on the behavioral intention to use digital services. It features the three

constructs 'collection', 'control', and 'awareness', which are all influenced by the latent construct 'privacy concerns'. 'Collection' measures the amount and perceived importance of the data collection, whereas 'control' and 'awareness' account for the perceived personal influence and transparency in the data collection process. The privacy concerns influence the magnitude of trust and risk beliefs of an individual, which in turn predict the behavioral intention to use a digital service.

Similar to the UTAUT, the IUIPC model has been used to explain behavioral intention in various digital case-studies [39, 40]. Due to the predictive power of this model, and due to the fact that data security and privacy protection are generally of high importance in Germany [41] and have been named as a reason against using the app by many individuals in representative media polls [9], we include these constructs in our model and propose that they influence usage intention (H3). Since 'trust beliefs' in the IUIPC essentially only cover trust related to a digital application, we argue that additional trust motives need to be incorporated in the context of CWA.

Therefore, we extend the privacy concerns of the IUIPC by two other trust-related constructs, namely trust in algorithms, which captures an individual's attitude towards artificial intelligence and computational data processing, and trust in authorities, which has been shown to be an influencing factor for the usage intention of mobile health services as well as the intention to use pandemic tracing apps [42, 43]. We argue that if an individual feels great trust in computer algorithms and the authorities and organizations responsible for the development of an app (German federal government, Robert Koch-Institute for disease control, SAP and Telekom), the privacy concerns are reduced and the intention to install an officially provided app is being supported.

**Prevailing Opinions, Morals, and Individual Cognitions**. Since the combination and extension of the aforementioned models, UTAUT and IUIPC, may still neglect some important influencing factors in the specific case of widespread pandemic tracing apps, we enrich our model with further concepts reflecting prevailing opinions towards the CWA and attitudes towards the coronavirus disease, which are both likely to have a considerable influence on usage intention (H4). We distinguish these two because an individual may have fundamentally different opinions regarding the disease and the app. These specific characteristics of our model have not been explored in previous literature to our knowledge. Therefore, we define multiple constructs to obtain the broadest possible picture of factors influencing intention to use the CWA.

Since the use or non-use of the CWA is not only a personal affair, but has also implications for the macrosocial health and well-being of other people, we define constructs that capture the assessment of COVID-19-related economic and health risks for oneself and for others, as well as moral concerns in the context of the coronavirus pandemic, which have been shown to increase individual health protection motivation [42, 44]. Regarding opinions about the app, we include selected prevailing reasons in favor or against app usage that were discussed in media polls prior to our study [9] to see whether these biases really are drivers of CWA adoption intention and to capture an individual's general attitude towards the CWA. In doing so, we attempt to account for the dual-factor concepts proposed by Tsai et al. (2019)[35], i.e., factors in favor (enablers) and against (inhibitors) the use of a technology.

**Personality Traits and Demographics**. While the mentioned constructs show an individual's specific attitude towards a specific topic, i.e., disease or app, we further want to examine general attitudes anchored in an individual's personality. The 'Big Five'

main human personality traits, namely extraversion, agreeableness, openness, conscientiousness, and neuroticism, are fundamental to personality research and psychology as a whole (e.g., [45, 46]). We include them in our model since they have been shown to affect individuals' technology acceptance and usage attitudes in prior studies [47]. Exemplarily, agreeableness is positively correlated with perceived usefulness, whereas extraversion is positively linked to behavioral intention to adopt new technologies [48, 49]. Therefore, we assume that personality traits may also play a role in CWA adoption (H5). Additionally, we include measures for empathy and perspective-taking [23], since these capabilities may increase the proclivity to use the app for societal benefits.

Furthermore, we record several demographic variables, such as age, gender, or income level of our participants, that may also influence the intention to use the CWA. Age and gender were important factors in explaining technology use in a plethora of previous studies on technology acceptance (e.g., [50, 51, 52]). Income may not account for the formation of usage intention, but it may indirectly affect actual CWA usage since the app is only usable on latest smartphones which have the technical requirements for low-energy Bluetooth technology [7].

# 3 Method

## 3.1 Research Model & Questionnaire Design

For our research model, we include constructs regarding UTAUT and IUIPC, trust motives, attitudes towards COVID-19 and CWA, as well as personality and various demographic variables. We expand current research by testing and quantifying the (incremental) effects of these different predictors. To our knowledge, this is the first paper to simultaneously combine a plethora of factors and prior models into one comprehensive model for pandemic tracing. We use unaltered items out of validated scales whenever possible (e.g., BFI-10, [53]) or slightly modified items of existing scales to fit our specific experimental setting. Exemplarily, we adapt the items measuring 'performance expectancy' in the UTAUT model, since the CWA offers no personal 'performance boost' but rather a group benefit for multiple users ('collective action problem', e.g., [54]).

Therefore, and since the CWA is usually not used actively and rather acts passively without the user noticing, we drop items such as "Using the system boosts my productivity" and replace them with "The CWA contributes effectively to the containment of the coronavirus" (Item: UT_Use2, see Figure A.2) to suit this specific case of COVID-19 tracing. We measure social influence with an item adjusted to our specific case, namely "People I care about think that I should use the CWA" (UT_SI, see Table A.1). In a similar fashion, we adapt some general items from the original IUIPC scale and align them with our specific research focus. For example, instead of assessing trusting beliefs towards online corporations in general, we capture the specific trusting beliefs concerning the CWA ("I trust that my data will be handled responsibly and safely when using the Corona-Warn-App", IU03_Trust).

All our changes are therefore directed at increasing the specificity and understandability of the questionnaire. The final questionnaire consists of 128 items and over 20 constructs (see Table A.1 and A.2 for a brief overview). Prior to the data collection, we pre-registered this research project on the Open Science Framework to ensure pro-

cess transparency. The full questionnaire, collected data, and SmartPLS model files are publicly available (https://osf.io/kepfr/).

Table A.1: Excerpt of items, questions and scales of technology acceptance contructs

| Construct | Item | Question (translated) | Scale |
|---|---|---|---|
| Actual Behavior | BH01_01 | Did you download the CWA? | binary |
| Beh. Intention | BI01_01 | I will be using the CWA in the next few days. | 7-point Likert |
| Eff. Expectancy | UT_EE | It is effortful for me to install and use the CWA | 7-point Likert |
| Perf. Expectancy | UT_Use1 | By using the Corona-Warn-App, I get a personal benefit. | 7-point Likert |
| Social Influence | UT_SI | People I care about think that I should use the CWA. | 7-point Likert |
| Facil. Conditions | UT_FC | I have the knowledge and tech. understanding to use the CWA. | 7-point Likert |

**Sample Characteristics**. Our online survey was conducted from July 1st to July 7th 2020 via the software SoSciSurvey. Participants were German residents that were primarily recruited through the online platform Clickworker. We opted for this platform since it provides benefits and quality measures comparable to other crowdsourcing platforms while featuring a larger German population [56, 57]. Crowdsourcing is especially adequate when trying to assess diverse cognitions of a wide range of individuals [58]. We supplemented this sample by sharing the questionnaire in our internal university research participation system and through social messenger providers (e.g., Whatsapp) to further increase sample diversity. In total, we recruited 458 participants. 33 participants were excluded because they did not pass the integrated attention check [59] or did not complete the questionnaire. Our final sample included 425 participants (Mage = 33.7, SD = 13.7; 18-71 years), with 286 Clickworker, 106 student pool and 33 social network participants. At 53.4% (227 of 425), slightly more males participated than females. Furthermore, our recruiting procedure provided a sample with nearly equal shares of CWA users (55.1%) and non-users (44.9%). Although this is not exactly representative for the German population, it is highly suitable to gain insights on essential drivers for CWA use as well as non-use in our research. Regarding COVID-19, 7.5% of the participants had already been tested and a total of 0.7% of our sample had already been diagnosed with a COVID-19 infection prior to answering our questionnaire.

## 3.2 Analysis Method: Structural Equation Model (SEM)

To empirically test our pre-registered, a-priori theoretical assumptions, we apply a structural equation model (SEM), which belongs to the family of causal analyses. SEMs are helpful to find interdependencies of effects of latent variables [28]. Since we place particular emphasis on the correct modeling of formative and reflective variables [60], we use

Table A.2: Research Model Components with Descriptions and References

| **Component** | **Description** | **References** |
|---|---|---|
| Actual Behavior | Download and activation of the CWA | [21] |
| Beh. Intention | An individual's willingness to use the CWA | [21] |
| Tech. Acceptance | Performance and effort expectancy, social influence, and other facilitating conditions with respect to CWA adoption | [21] |
| Privacy Concerns | Privacy concerns with respect to collection, awareness and control of data, trust and risk beliefs | [22] |
| COVID-19 Risk Assessment | COVID-19-related personal, societal, and economic risks | self-defined |
| Trust in Authorities | Trust in different governmental authorities and enterprises | [55] |
| Trust in Algorithms | Trust in computational algorithms and artificial intelligence | self-defined, Nickel and Pinto (1986) |
| Conspiracy Attitude | COVID-19 related conspiracy attitudes | [55] |
| Media Consumption | Consumption of different official and alternative media sources | [55] |
| Online Behavior Profile | Use of different online services, i.e., social networks, mailing, cloud services, and digital assistants | self-defined |
| Technical Affinity | An individual's expertise on the internet | self-defined |
| CWA prevailing opinons | Prevailing opinions in favor or against digital tracing apps | self-defined, [9] |
| Moral and Ethics | Moral and ethical attitudes with respect to CWA usage | self-defined |
| Empathy | Individual empathy and perspective-taking | [23] |
| Personality Big 5 | Main personality traits, comprising extraversion, agreeableness, conscientiousness, neuroticism and openness | [53] |
| Demographics | Various socio-economic variables | self-defined |

a variance-analytical approach, i.e., the partial least squares method (PLS) for our analysis. In contrast to the covariance approach (linear structural relationships, LISREL), PLS allows a simultaneous estimation of reflective and formative measurement models [61]. Furthermore, according to Chin and Newsted (1999) [62], the PLS approach is most suitable for comparably new phenomena without profound construct theories, complex models with a high number of measured variables, and predictive research with a moderate sample size. Since these characteristics are all given in our research, and since our goal is to explain as much variance as possible by combining and integrating multiple models and constructs, we apply the PLS approach for data analysis using SmartPLS, closely following the research procedure proposed by Weiber and Mühlhaus [63, pp. 323-330].

# 4 Data Analysis & Results

The analysis of the data includes the confirmation of the measurement model (outer model) and the test of the structural model itself (inner model). Since our model features both formative and reflective constructs, we apply respective methods in the evaluation of the structural and in the measurement model. We test reflective constructs with three or more items for uni-dimensionality through exploratory factor analysis. The sample adequacy for conducting factor analysis is confirmed via the Kaiser-Meyer-Olkin criterion (KMO) above 0.6 and mean sampling adequacy (MSA) greater than 0.5. All constructs are confirmed to be uni-dimensional, i.e., the constructs are the main cause for the items. Subsequently we examine the reliability, i.e., the accuracy of a measuring instrument, of our constructs through scale-wise inspection of Cronbach's alpha with a cutoff value of $\alpha \geq 0.7$. Scales with insufficient values are being pruned by removing incongruous items until satisfactory reliability is accomplished, since the items should capture the same underlying construct. Furthermore, we test the convergence and discriminant validity of the reflective constructs with the Fornell-Larcker-Criterion with a threshold of Average-Variance-Extracted ($AVE$) $> 0.5$ [64]. The AVE of the reflective constructs lies between 0.688 and 0.953 and should be greater than the squared correlations, which lie between 0.001 and 0.416. Hence, adequate validity is present for all constructs.

The formative constructs are examined through testing for collinearity. The variance inflation factor (VIF) is $< 3$ for all respective items, signaling a low (multi-) collinearity and adequate distinctness. The reliability of the formative constructs is investigated through bootstrapping with 1,000 samples. Items with non-significant path weights, with $t < 1.96$ at a 95% confidence level, are eliminated from the measurement model [63, p. 327 & 336].

The structural model, i.e., the paths between constructs, is tested by applying a bootstrapping approach with 1,000 samples. Corresponding to existing literature on SEM testing, we don't eliminate non-significant paths in a first step, since they were created with deductive, theoretical considerations in mind [65, 66]. The direction of the effects, i.e., path coefficients, are in line with our theoretical assumptions, which generally confirms the structural model. The included predictors explain a significant amount of variance for both behavioral intentions ($R^2_{adj} = 0.643$) and factual behavior ($R^2_{adj} = 0.594$). The relationship between behavioral intention and behavior was also confirmed in an additional logistic regression analysis with behavioral intention as the sole predictor (Nagelkerke's $R^2 = .684, B = 1.00, SE_B = .084, p < .001, Odds-Ratio = 2.718, CI_{95\%} = [2.305, 3.206]$).

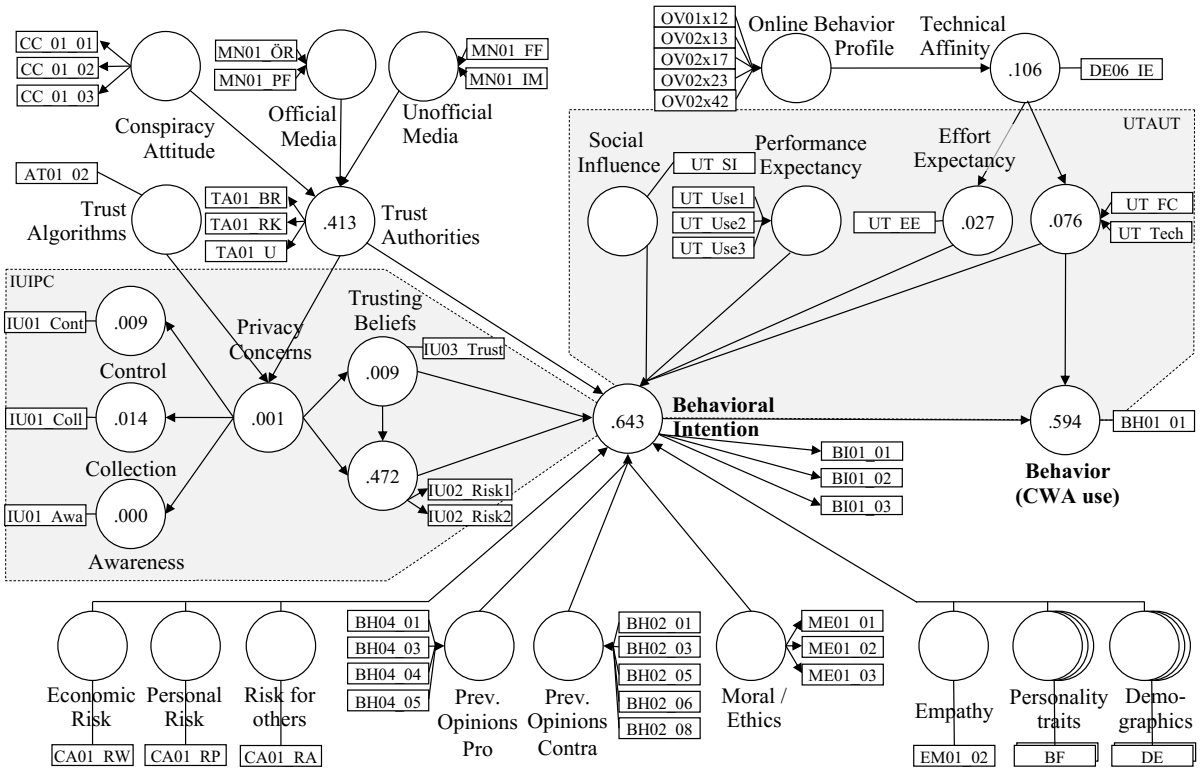


Figure A.2: Full model with explained variance in the respective construct circles ($adj.R^2$, only computable for endogenous constructs)

Furthermore, we test cross-validated predictive relevance of the PLS path model ($Q^2$) with the Stone-Geisser-criterion under a blindfolding procedure [67, 68]. Resulting $Q^2$ values should be larger than 0 to indicate that the exogenous constructs have predictive relevance for the endogenous constructs under consideration. Our model shows a predictive relevance for behavioral intention of $Q^2 = 0.611$ and for behavior $Q^2 = 0.589$, indicating a strong predictive relevance ($> 0.35$) [65, 69]. The full model is illustrated in Figure A.2.

The significance analysis indicates that some of the paths in the full model are non-significant and do not explain a substantial amount of variance in our dependent variables (i.e., intended and factual CWA use). To streamline our model and to increase overall clarity, we successively eliminate all non-significant paths (pruning). In order to keep 'facilitating conditions', we connected it directly to 'effort expectancy'. The remaining predictors still explain a moderate to large amount of variance for both behavioral intentions ($R^2_{adj} = 0.639$) and behavior ($R^2_{adj} = 0.590$), with a negligible prediction difference compared to the full model ($\Delta R^2_{adj} = 0.004$ for both intended and actual CWA use). The predictive validity remains strong at $Q^2 = 0.607$ for intentions and at $Q^2 = 0.587$ for behavior. The resulting (pruned) model is presented in Figure A.3 with all path coefficients and effect sizes $f^2$ displayed in Table A.3. The total predictive effect of the remaining constructs on behavioral intention and the effect between intention and behavior in our model are considered to be strong according to common conventions ($\geq 0.35$; [70]).

Table A.3: Structural model evaluation of pruned model ordered by relationship and $f^2$

| **Significant Relationship ↓** | **Path coefficient** | **t-stat.** | **Confidence 2,5%** | **97,5%** | $f^2$ ↓ |
|---|---|---|---|---|---|
| Intention → Behavior | $0.769^{***}$ | 29.356 | 0.715 | 0.819 | 1.444 |
| Effort Expectancy → Intention | $-0.201^{***}$ | 5.201 | -0.275 | -0.123 | 0.097 |
| Social Influence → Intention | $0.216^{***}$ | 5.046 | 0.130 | 0.297 | 0.072 |
| Prevailing Reasons Contra → Intention | $-0.190^{***}$ | 5.070 | -0.273 | -0.123 | 0.061 |
| Moral/Ethics → Intention | $0.239^{***}$ | 4.253 | 0.123 | 0.342 | 0.059 |
| Prevailing Reasons Pro → Intention | $0.194^{***}$ | 4.549 | 0.114 | 0.281 | 0.056 |
| Risk for Others → Intention | $-0.117^{**}$ | 3.298 | -0.187 | -0.048 | 0.027 |
| Income → Intention | $0.084^{**}$ | 3.101 | 0.028 | 0.132 | 0.019 |
| Performance Expectancy → Intention | $0.119^{*}$ | 1.999 | 0.013 | 0.249 | 0.014 |
| Economic Risk → Intention | $0.062^{*}$ | 2.218 | 0.008 | 0.119 | 0.011 |
| Facil. Conditions → Effort Expectancy | $-0.474^{***}$ | 9.879 | -0.567 | -0.381 | 0.289 |
| Online Profile → Technical Affinity | $0.345^{***}$ | 9.788 | 0.282 | 0.423 | 0.135 |
| Technical Affinity → Facil. Conditions | $0.275^{***}$ | 4.905 | 0.165 | 0.388 | 0.082 |

Notes: Evaluation of effect size according to [70]:
$f^2 \geq 0.02$ small impact, $f^2 \geq 0.15$ moderate impact, $f^2 \geq 0.35$ large impact.
Significance levels: $^{*} = p < .05$, $^{**} = p < .01$, $^{***} = p < .001$

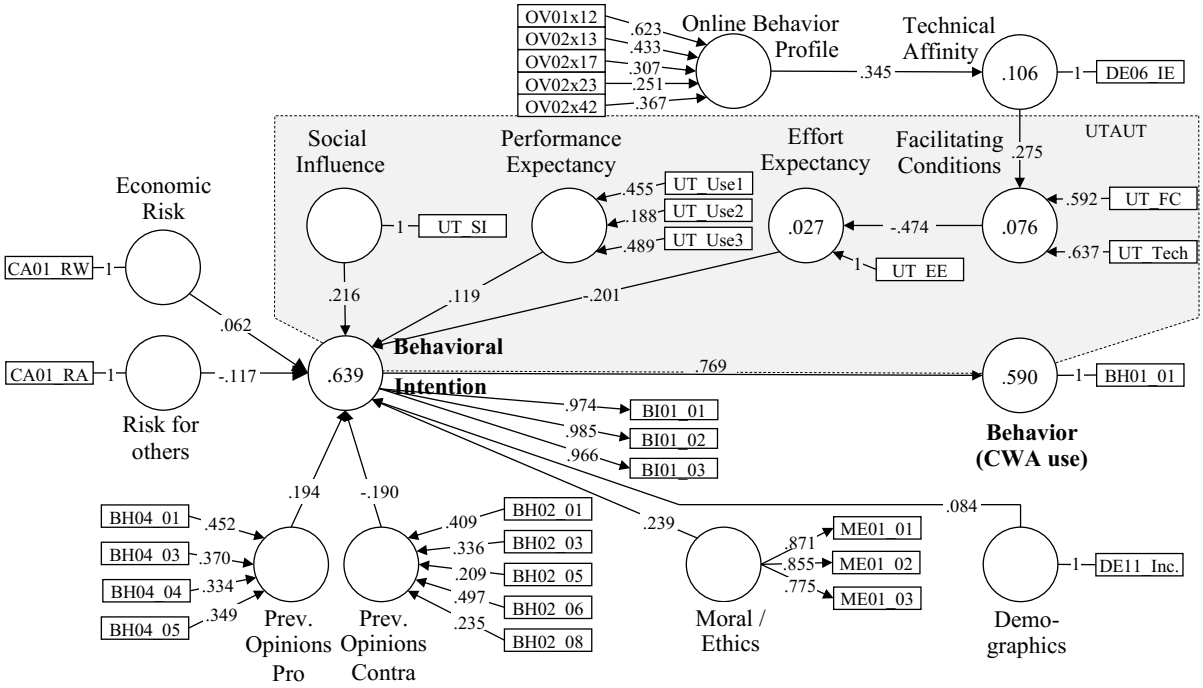


Figure A.3: Pruned model (adj. $R^2$ for endogenous constructs, path coeff., weights and loadings)

# 5 Discussion

Our goal was to create and empirically test an extensive model of usage intention and actual use of pandemic tracing apps. Specifically, we investigated the German Corona-Warn-App and tried to identify the most influential factors that predict individuals' usage intention and actual CWA usage. The included constructs were able to explain a substantial amount of variance for behavioral intention and behavior (i.e., CWA use) in the full model as well as in a pruned model. We find that usage intention strongly influences actual use. This is not surprising, however, it represents a core statement of our specific model for the adoption of tracing apps. In other words, the factors that influence intention also influence actual behavior in this pandemic context (H1). Therefore, it can be assumed that the conclusions we draw for behavioral intention will also have an impact on the actual download and usage numbers. In the following section, we will now discuss the most important predictors of usage intention and actual use.

**General Technology Acceptance (UTAUT) and Technical Affinity**. In terms of effect size, effort expectancy proves to have the strongest influence on behavioral intention. As expected, we found a negative path coefficient, indicating that the higher the expected effort, the lower the intention to install the Corona-Warn-App. Furthermore, we investigated some antecedents of effort expectancy, as we see strong path coefficients between online behavior profile and technical affinity, technical affinity and facilitating conditions, as well as between facilitating conditions and effort expectancy. This branch of our model helps to understand how the expectation of effort arises in this context. If individuals have a basic technical affinity, which is measured via the use of online services (online profile and experience items), then this, together with the required technical resources and the corresponding practical knowledge, forms the facilitating conditions. The facilitating conditions have in turn a strong influence on the effort expectancy (negative path coefficient), i.e., the lower the supporting factors, the higher the expected effort and the lower the intention. Furthermore, performance expectancy, i.e., perceived effectiveness of the CWA, positively influences usage intention, another relation that was expected based on the UTAUT. Considering path coefficient and effect size, social influence exerted the second strongest positive influence on CWA usage intention (H2). If an individual experiences its social environment participating or advocating the tracing app, it's likely that this individual will also adopt. This coincides with research on conditional cooperation [71]. However, we cannot make a statement here as to whether a single social contact or a social group is necessary and whether this contact must take place on a real or digital level for the effect to occur. Nevertheless, we can still assume that more favorable attitudes of peers towards the CWA and an encouraging communication within the direct social environment positively affect intention and usage. This relation provides a solid basis for the practical implications later on. Furthermore, our findings concerning the UTAUT constructs once again underline the universality and predictive validity of the model, which proves to be applicable in this specific pandemic tracing context.

**Privacy Concerns (IUIPC) and Trust Motives**. Our assumption that data security concerns would have a major influence on intention was not confirmed, with neither the IUIPC model nor our additional trust measures exerting a significant influence on our dependent variables (H3). Despite a broad sample distribution of privacy concern perceptions in the measurement model, the IUIPC did not possess any predictive relevance. This was somewhat surprising, but could be explained by two reasons. First, given the

unique and unprecedented situation of a pandemic, the magnitude of individual data security concerns could be of relatively minor importance compared to the possible benefit of saving lives and tracing infection chains. Second, as the German federal government and media emphasized the high CWA data security in the communication and followed an open app development approach, the data security concerns might have been alleviated beforehand. When asked for trust in the Robert Koch-Institute, which is the official app distributor, our sample shows relatively high trust ($M_{TrustRKI} = 5.61, SD = 1.63$), which might also explain the reduced data privacy concerns. Nevertheless, these findings need to be validated in further studies. In general, we are still confident that the IUIPC model is a reasonable approach to include privacy concerns in a SEM for digital tracing apps.

**Prevailing Opinions, Morals, and Individual Cognitions**. The construct moral and ethics exerted a significant influence on intention. We measured this construct through three items regarding moral and ethical considerations on the individual and collective level. The respective path coefficient was positive, meaning that individuals perceiving a strong moral obligation to use the CWA and a high willingness to impose personal consequences in case of a risk notification or COVID-19 infection respectively, e.g., self-isolation, also showed a higher intention to use the app. Furthermore, we can say that prevailing reasons in favor or against the usage of the app show a significant impact (H4). At first glance, this may appear somewhat trivial, however, a more thorough investigation of singular reasons leads to interesting insights. Within the construct, the most influential reason in favor of using the app is personal protection. If an individual perceives a personal health benefit through using the CWA, the likelihood to use the app increases. In contrary, the most influential reason against using the app is a lack of personal interest, resulting in a reduced usage intention. This contradictory positioning, i.e., the assessment of a personal health advantage, which should be related to a high level of self-interest, however, is possibly diminished or even completely reduced by individual disinterest. Possible reasons for this may be that some groups within the population might still underestimate the risks concerning COVID-19, or that the value of the app, the urgency of a large participation, or the low effort to participate may not be clear to everybody, indicating room for improvement in CWA communication. The bivariate correlation between estimated risk for others and usage intention indicate a positive relationship ($r = .313, p < 0.001$). Surprisingly, this effect is reversed ($\beta = -.117$) in our SEM calculation, so that a higher risk assessment coincides with a lower intention to use the app. From a methodical standpoint, this can most likely be explained due to multicollinearity of some included constructs, but logically, we cannot give a sound interpretation of this surprising effect and it may be wise to refrain from further interpretation.

**Personality Traits and Demographics**. The Big Five personality traits constructs do not influence the behavioral intention to use the CWA in our dataset (H5). Regarding demographical influence, only income proves to be a significant predictor. Higher income predicts a higher intention to use the CWA. Participants' age does not have a significant influence on intention in the SEM. However, additionally to our SEM analysis, we find a negative bivariate correlation between age and CWA use ($r = -.128, p < 0.01$). The older the participants in our sample, the less likely is their actual usage of the app. This might be explained due to technical difficulties, reflected in a negative correlation between age and technical affinity ($r = -.117, p < 0.05$), which may represent an existing digital divide between age groups [72].

## 5.1 Theoretical and Practical Implications

Our theoretical contribution lies within the creation and testing of a comprehensive exploratory model that integrates different existing theories and expands them with additional psychological factors and concepts. We enriched theory by testing the validity of important models in information technology, namely UTAUT and IUIPC, in the unique societal and technological context of pandemic tracing. Moreover, our exploratory findings yield interesting insights for further theory development. We were able to identify substantive antecedents of UTAUT's facilitating conditions and detected the construct's significant influence on effort expectancy. Since further research on these antecedents and consequences was highlighted as an important task for future research in a recent meta-analysis on the UTAUT [73], we illuminated some of these interrelations and offer a starting point for further examination.

Since we captured actual behavioral data for a real and publicly known app, our practical implications are based on a real application case and can be specifically implemented by CWA policy makers. Nevertheless, these implications might not hold for different countries or societies, e.g., mandatory measures in collectivistic societies versus voluntary measures in individualistic societies. The importance of effort expectancy, moral and ethical considerations, as well as social influence offer practical implications for the communication and marketing of the Corona-Warn-App in Germany. Future campaigns should potentially focus on addressing these constructs to facilitate diffusion of the app. In the following, we suggest actionable implications in three general categories, namely marketing channels, communication content, and public restrictions.

Regarding marketing channels, it is necessary to gain broad visibility for the app in public. This could be achieved by traditional marketing in public areas, e.g., sports events and TV advertising. As we identified direct social influence as a major predictor for usage, we suggest social media marketing through 'influencers'. This could encourage individuals, especially in the younger age groups, to use the app. As collective action is needed, it might also help to engage already existing users by sending them push messages, thanking them for participating, and giving them the possibility to easily share a download link with their peers, using slogans such as "spread the app, not the virus". In terms of communication content, the message of advertisements should highlight the individual and the collective societal benefit of a pandemic tracing app, which is also valid for app design specifications [16]. Therefore, the usefulness of the app in a general confinement strategy should be emphasized, e.g., by describing real situations in which a chain of infection has already been broken by a CWA notification. Further, explanatory videos for the download and the technology of proximity tracing apps and the related data security could be published. Regarding public restrictions, health authorities could partner with event management or owners of museums, restaurants, or other places of interest to start a campaign where individuals and groups are granted access to events or places, e.g., soccer matches, only if they can prove their installation of the CWA. Since the installation is voluntary, this will connect a concrete benefit to having the app installed and at the same time implies some kind of obligation. Nevertheless, using the app should not promote digital divide, because as we mentioned earlier, success is a joint effort in this case of societal action. This could be even more important because people's morale might be a great factor for their stamina and overall strategy to contain the virus. The app should also support traditional contact tracing. For example, it could offer the possibility

to leave selected contact data through push messages (digital ID card), in a restaurant for instance, when customers have to leave their contact data anyway, which could further enhance the app's perceived practical benefits.

## 5.2 Limitations and Future Research

In our study we investigated the specific case of a governmental pandemic tracing app in Germany. Thus, our results may not be valid for other countries and cultural groups and should be transferred and generalized with caution. Furthermore, although our sample featured participants of different age and income groups, as well as nearly equal shares of males/females and CWA users vs. non-users, it is not entirely representative for the German population with regard to its demographic characteristics, e.g., age. Nevertheless, we argue that our sample, which is rather young with a mean age of 33.7 years, is suitable for our cause, since young people are key factors in both technology adoption [50] and community transmission of COVID-19 [74]. The fact that some constructs for which we assumed a theoretical connection to CWA usage, such as the IUIPC model or the Big Five personality traits, did not have a significant influence on behavior might not hold in future studies and might be due to sample characteristics, e.g., younger people having less privacy concerns as they grew up with smartphone apps and social media. We furthermore found that behavioral intention did not entirely predict actual behavior, which echoes the critique by Bagozzi, who noted that the "intention-behavior linkage is probably the most uncritically accepted assumption in [...] IS research" [27, p. 245], and is consistent with scientific literature on the attitude-behavior gap (see [75, 76]).

Although structural equation modeling with the PLS approach is used very frequently in information systems research, there is currently no global, universally valid criterion for evaluating the overall model fit [63, p. 330]. Recent research streams are concerned with developing indices for this purpose, but since they are still in their early research stage and not yet fully understood or established, these criteria should not be used to assess model quality just yet [77]. Nonetheless, a model is usually assessed as reliable when both the measurement and structural model meet the relevant quality criteria [63, p. 330]. As we showed in the previous sections, this is the case for our model. Even though our SEM met all relevant quality criteria and explained a significant amount of variance with regard to usage intention and usage behavior, there are still some methodological limitations to our research. Despite the many advantages of the PLS approach, PLS assumes an error-free measurement of the formative measurement model, which can potentially confound the structural model [65],[63, p. 77-78]. Further experiments and validations of our results should therefore choose a different approach to data analysis, e.g., the LISREL approach that accounts for measurement errors [63, p. 51]. Furthermore, future studies within this field of research should possibly strive for examining the effects in specific demographic groups. Given that our explorative research model is rather complex with a multitude of relevant constructs, we were not able to measure all of these constructs with multiple items regarding all of their facets and dimensions due to practical reasons and limited questionnaire space. Our study should therefore be seen as an explorative basis for future research with a narrower focus on particular constructs and path relationships, e.g., a country comparison of tracing apps and interoperability [78].

# 6 Conclusion

The purpose of this study was to develop an extensive model of use and intended use of pandemic tracing apps, namely the official German Corona-Warn-App, by using an integrative approach and combing and enhancing existing theories. We designed a comprehensive, exploratory structural equation model, analyzed the model with a PLS approach, and pruned the model to exclude non-significant relationships. The remaining constructs are able to explain a substantial amount of variance ($R^2$) for behavioral intention and behavior (CWA use). We also examined the differential predictive importance of the respective constructs. Both our models show strong predictive relevance ($Q^2$). We identify effort expectancy, social influence, prevailing opinions, as well as moral and ethical considerations as the most important predictors. Jointly, the predictors account for a substantial amount of variance within both our full and the pruned model. Our results offer valuable insights for the technology acceptance literature, such as the integration of moral and ethical consideration in a technology acceptance model, and enable practical recommendations for improving the public communication and user numbers of pandemic tracing apps in Germany. Our results imply that a shift of communication focus from data security towards group benefits and moral obligation could further increase the CWA user numbers. Nevertheless, our extensive model analysis should only be viewed as a starting point for further in-depth analyses of selected path relationships, e.g., on the effect of ethical and moral considerations as a driver for technology usage intention. We encourage other researchers to review our pre-registration and original questionnaire, and to use the collected data, which are all publicly available (https://osf.io/s3zq2/).

# References


[1] G. G. Wagner, S. Kühne, and N. A. Siegel. *Akzeptanz der einschränkenden Corona-Maßnahmen bleibt trotz Lockerungen hoch*. 2020.

[2] D. Baud et al. "Real estimates of mortality following COVID-19 infection". In: *The Lancet Infectious Diseases* (2020). DOI: 10.1016/S1473-3099(20)30195-X.

[3] M. A. Shereen et al. "COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses". In: *Journal of Advanced Research* (2020). DOI: 10.1016/j.jare.2020.03.005.

[4] K. Sun and C. Viboud. "Impact of contact tracing on SARS-CoV-2 transmission". In: *The Lancet Infectious Diseases* (2020), pp. 876–877. DOI: 10.1016/S1473-3099(20)30357-1.

[5] J. Li and X. Guo. "COVID-19 Contact-tracing Apps: A Survey on the Global Deployment and Challenges". In: *arXiv preprint arXiv:2005.03599* (2020).

[6] German Federal Press Office. *Veröffentlichung der Corona-Warn-App*. 2020.

[7] J. Mueller. *COVID-19: Die technische Grundlage der Corona-Warn-App in Deutschland*. 2020.

[8] Robert Koch-Institute. *Kennzahlen zur Corona-Warn-App*. 2020. URL: https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/WarnApp/Archiv_Kennzahlen/Kennzahlen_06112020.pdf?__blob=publicationFile (visited on 11/13/2020).

[9] A.-K. Sonnenberg. *Für die Hälfte der Deutschen ist Nutzung der Corona-Warn-App unwahrscheinlich*. 2020.

[10] L. Simko et al. "COVID-19 Contact Tracing and Privacy: Studying Opinion and Preferences". In: *arXiv preprint arXiv:2005.06056* (2020).

[11] M. Birnbaum and C. Spolar. *Coronavirus tracking apps meet resistance in privacy-conscious Europe*. 2020.

[12] J. Taylor. *Covidsafe app: How to download Australia's coronavirus contact tracing app and how it works*. 2020.

[13] K. Riemer et al. "Digital contact-tracing adoption in the COVID-19 pandemic: IT governance for collective action at the societal level". In: *European Journal of Information Systems* (2020), pp. 1–15.

[14] L. Ferretti et al. "Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing". In: *Science* 368.6491 (2020).

[15] R. Hinch et al. *Effective configurations of a digital contact tracing app: A report to NHSX*. 2020.

[16] S. Trang et al. "One app to trace them all? Examining app specifications for mass acceptance of contact-tracing apps". In: *European Journal of Information Systems* 29.4 (2020), pp. 415–428.

[17] S. Becker et al. *Akzeptanz von Corona-Apps in Deutschland vor der Einführung der Corona-Warn-App*. 2020.

[18] B. Zhang, S. Kreps, and N. McMurry. *Americans' perceptions of privacy and surveillance in the COVID-19 Pandemic*. 2020.

[19] G. Kaptchuk, E. Hargittai, and E. M. Redmiles. "How good is good enough for COVID19 apps? The influence of benefits, accuracy, and privacy on willingness to adopt". In: *arXiv preprint arXiv:2005.04343* (2020).

[20] T. Jahnel et al. "Contact-Tracing-Apps als unterstützende Maßnahme bei der Kontaktpersonennachverfolgung von COVID-19". In: *Gesundheitswesen (Bundesverband Der Arzte Des Offentlichen Gesundheitsdienstes (Germany))* 82.08-09 (2020), p. 664.

[21] V. Venkatesh et al. "User acceptance of information technology: Toward a unified view". In: *MIS Quarterly* (2003), pp. 425–478. ISSN: 02767783.

[22] N. K. Malhotra, S. S. Kim, and J. Agarwal. "Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model". In: *Information Systems Research* 15.4 (2004), pp. 336–355.

[23] F. D. Davis, R. P. Bagozzi, and P. R. Warshaw. "User acceptance of computer technology: a comparison of two theoretical models". In: *Management Science* 35 (1989), pp. 982–1002.

[24] M. Fishbein. "A theory of reasoned action: some applications and implications". In: *Nebraska Symposium on Motivation*. Vol. 27. 1979, pp. 65–116.

[25] Icek Ajzen. "From intentions to actions: A theory of planned behavior". In: *Action Control*. Ed. by J. Kuhl and J. Beckmann. SSSP Springer Series in Social Psychology. Springer, 1985, pp. 11–39. DOI: 10.1007/978-3-642-69746-3_2.

[26] E. M. van Raaij and J. J. L. Schepers. "The acceptance and use of a virtual learning environment in China". In: *Computers & Education* 50 (2008), pp. 838–852. DOI: 10.1016/j.compedu.2006.09.001.

[27] R. P. Bagozzi. "The Legacy of the Technology Acceptance Model and a Proposal for a Paradigm Shift". In: *Journal of the Association for Information Systems* 8.4 (2007), pp. 244–254.

[28] J. Hair et al. *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*. 2nd ed. Sage Publications, 2017.

[29] Q. Min, S. Ji, and G. Qu. "Mobile commerce user acceptance study in China: a revised UTAUT model". In: *Tsinghua Science and Technology* 13.3 (2008), pp. 257–264.

[30] C. S. Yu. "Factors affecting individuals to adopt mobile banking: Empirical evidence from the UTAUT model". In: *Journal of Electronic Commerce Research* 13.2 (2012), p. 104.

[31] T. Zhou, Y. Lu, and B. Wang. "Integrating TTF and UTAUT to explain mobile banking user adoption". In: *Computers in Human Behavior* 26.4 (2010), pp. 760–767.

[32] A. Gruzd, K. Staves, and A. Wilk. "Connected scholars: Examining the role of social media in research practices of faculty using the UTAUT model". In: *Computers in Human Behavior* 28.6 (2012), pp. 2340–2350.

[33] B. Kijsanayotin, S. Pannarunothai, and S. M. Speedie. "Factors influencing health information technology adoption in Thailand's community health centers: Applying the UTAUT model". In: *International Journal of Medical Informatics* 78.6 (2009), pp. 404–416.

[34] V. Venkatesh, J. Y. Thong, and X. Xu. "Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology". In: *MIS Quarterly* (2012), pp. 157–178. ISSN: 02767783.

[35] J. M. Tsai et al. "Acceptance and resistance of telehealth: The perspective of dual-factor concepts in technology adoption". In: *International Journal of Information Management* 49 (2019), pp. 34–44. ISSN: 0268-4012. DOI: 10.1016/j.ijinfomgt.2019.03.003.

[36] C. J. Hoofnagle and J. Whittington. "Free: accounting for the costs of the internet's most popular price". In: *UCLA Law Review* 61 (2013), p. 606.

[37] A. B. Ozturk et al. "Understanding the mobile payment technology acceptance based on valence theory". In: *International Journal of Contemporary Hospitality Management* 29.8 (2017), pp. 2027–2049.

[38] E. Bigné, C. Ruiz, and S. Sanz. "Key drivers of mobile commerce adoption. An exploratory study of Spanish mobile users". In: *Journal of Theoretical and Applied Electronic Commerce Research* 2.2 (2007), pp. 48–60.

[39] S. E. Chang, W. C. Shen, and A. Y. Liu. "Why mobile users trust smartphone social networking services? A PLS-SEM approach". In: *Journal of Business Research* 69.11 (2016), pp. 4890–4895. ISSN: 0148-2963.

[40] R. Balebako et al. “The impact of timing on the salience of smartphone app privacy notices”. In: *Proceedings of the 5th Annual ACM CCS Workshop on Security and Privacy in Smartphones and Mobile Devices*. 2015, pp. 63–74.

[41] G. Hornung and C. Schnabel. “Data protection in Germany I: The population census decision and the right to informational self-determination”. In: *Computer Law & Security Review* 25.1 (2009), pp. 84–88.

[42] M. Al-Rasheed. “Protective Behavior against COVID-19 among the Public in Kuwait: An Examination of the Protection Motivation Theory, Trust in Government, and Sociodemographic Factors”. In: *Social Work in Public Health* 35.7 (2020), pp. 546–556.

[43] R. Schnall et al. “Trust, perceived risk, perceived ease of use and perceived usefulness as factors related to mHealth technology use”. In: *Studies in Health Technology and Informatics* 216 (2015), p. 467.

[44] S. Milne, P. Sheeran, and S. Orbell. “Prediction and intervention in health-related behavior: A meta-analytic review of protection motivation theory”. In: *Journal of Applied Social Psychology* 30.1 (2000), pp. 106–143.

[45] S. D. Gosling, P. J. Rentfrow, and W. B. Swann Jr. “A very brief measure of the Big-Five personality domains”. In: *Journal of Research in Personality* 37.6 (2003), pp. 504–528.

[46] B. De Raad. *The Big Five Personality Factors: The psycholexical approach to personality*. Hogrefe & Huber Publishers, 2000.

[47] E. Karahanna et al. “Individual differences and relative advantage: the case of GSS”. In: *Decision Support Systems* 32.4 (2002), pp. 327–341.

[48] G. B. Svendsen et al. “Personality and technology acceptance: the influence of personality factors on the core constructs of the Technology Acceptance Model”. In: *Behaviour & Information Technology* 32.4 (2013), pp. 323–334.

[49] S. Devaraj, R. F. Easley, and J. M. Crant. “Research note—how does personality matter? Relating the five-factor model to technology acceptance and use”. In: *Information Systems Research* 19.1 (2008), pp. 93–105.

[50] S. A. Brown, A. R. Dennis, and V. Venkatesh. “Predicting collaboration technology use: Integrating technology adoption and collaboration research”. In: *Journal of Management Information Systems* 27.2 (2010), pp. 9–54. ISSN: 0742-1222.

[51] M. G. Morris, V. Venkatesh, and P. L. Ackerman. “Gender and age differences in employee decisions about new technology: An extension to the theory of planned behavior”. In: *IEEE Transactions on Engineering Management* 52.1 (2005), pp. 69–84.

[52] V. Venkatesh and M. G. Morris. “Why don’t men ever stop to ask for directions? Gender, social influence, and their role in technology acceptance and usage behavior”. In: *MIS Quarterly* (2000), pp. 115–139. ISSN: 02767783.

[53] B. Rammstedt and O. P. John. “Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German”. In: *Journal of Research in Personality* 41.1 (2007), pp. 203–212.

[54] R. Willer. "Groups reward individual sacrifice: The status solution to the collective action problem". In: *American Sociological Review* 74.1 (2009), pp. 23–43.

[55] Lotte Pummerer and Kai Sassenberg. *Conspiracy Theories in Times of Crisis and their Societal Effects: Case "Corona"*. 2020. URL: https://files.de-1.osf.io/v1/resources/y5grn/providers/osfstorage/5e960f0d4301660436a04b68?action=download&direct&version=2.

[56] Clickworker GmbH. *Our Clickworker community*. 2020. URL: https://www.clickworker.com/clickworker-crowd/.

[57] Andy T. Woods et al. "Conducting perception research over the internet: a tutorial review". In: *PeerJ* 3 (2015), e1058. ISSN: 2167-8359. DOI: 10.7717/peerj.1058. URL: https://peerj.com/articles/1058/?sa=x&ved=0ccuq9qewb2ovchmi6za8ilx3xgivugfbch3auw7l.

[58] R. Jia, Z. R. Steelman, and B. H. Reich. "Using mechanical turk data in IS research: risks, rewards, and recommendations". In: *Communications of the Association for Information Systems* 41.1 (2017), p. 14.

[59] D. M. Oppenheimer, T. Meyvis, and N. Davidenko. "Instructional manipulation checks: Detecting satisficing to increase statistical power". In: *Journal of Experimental Social Psychology* 45.4 (2009), pp. 867–872.

[60] S. Albers and L. Hildebrandt. "Methodische Probleme bei der Erfolgsfaktorenforschung — Messfehler, formative versus reflektive Indikatoren und die Wahl des Strukturgleichungs-Modells". In: *Schmalenbachs Zeitschrift für betriebswirtschaftliche Forschung* 58.1 (2006), pp. 2–33.

[61] C. M. Ringle. *Messung von Kausalmodellen. Ein Methodenvergleich*. 2004.

[62] W. W. Chin and P. R. Newsted. "Structural equation modeling analysis with small samples using partial least squares". In: *Statistical strategies for small sample research*. Vol. 1. 1999, pp. 307–341.

[63] R. Weiber and D. Mühlhaus. *Strukturgleichungsmodellierung: Eine anwendungsorientierte Einführung in die Kausalanalyse mit Hilfe von AMOS, SmartPLS und SPSS*. Springer-Verlag, 2014.

[64] C. Fornell and D. F. Larcker. "Evaluating structural equation models with unobservable variables and measurement error". In: *Journal of Marketing Research* 18.1 (1981), pp. 39–50.

[65] M. Secka. *Einfluss von Kommunikationsmaßnahmen mit CSR-Bezug auf die Einstellung zur Marke: Entwicklung und Überprüfung eines konzeptionellen Modells*. Peter Lang International Academic Publishers, 2015.

[66] A. Herrmann, F. Huber, and F. Kressmann. "Varianz-und kovarianzbasierte Strukturgleichungsmodelle—ein Leitfaden zu deren Spezifikation, Schätzung und Beurteilung". In: *Schmalenbachs Zeitschrift für betriebswirtschaftliche Forschung* 58.1 (2006), pp. 34–66.

[67] S. Geisser. "A predictive approach to the random effect model". In: *Biometrika* 61.1 (1974), pp. 101–107.

[68] M. Stone. "Cross-validatory choice and assessment of statistical predictions". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 36.2 (1974), pp. 111–133.

[69] W. W. Chin. "How to write up and report PLS analyses". In: *Handbook of Partial Least Squares*. Berlin, Heidelberg: Springer, 2010, pp. 655–690.

[70] W. W. Chin. "The partial least squares approach to structural equation modeling". In: *Modern Methods for Business Research*. Vol. 295. 1998, pp. 295–336.

[71] C. Thöni and S. Volk. "Conditional cooperation: Review and refinement". In: *Economics Letters* 171 (2018), pp. 37–40. DOI: 10.1016/j.econlet.2018.06.022.

[72] B. Niehaves and R. Plattfaut. "Internet adoption by the elderly: employing IS technology acceptance theories for understanding the age-related digital divide". In: *European Journal of Information Systems* 23.6 (2014), pp. 708–726. DOI: 10.1057/ejis.2013.19.

[73] H. Khechine, S. Lakhal, and P. Ndjambou. "A meta-analysis of the UTAUT model: Eleven years later". In: *Canadian Journal of Administrative Sciences* 33.2 (2016), pp. 138–152.

[74] T. K. Boehmer et al. "Changing age distribution of the COVID-19 pandemic—United States, May–August 2020". In: *Morbidity and Mortality Weekly Report* 69.39 (2020), p. 1404.

[75] C. J. Armitage and M. Conner. "Efficacy of the theory of planned behaviour: A meta-analytic review". In: *British Journal of Social Psychology* 40.4 (2001), pp. 471–499.

[76] E. Boulstridge and M. Carrigan. "Do consumers really care about corporate responsibility? Highlighting the attitude-behaviour gap". In: *Journal of Communication Management* 4.4 (2000), pp. 355–368.

[77] J. Hair et al. "An updated and expanded assessment of PLS-SEM in information systems research". In: *Industrial Management & Data Systems* 117.3 (2017), pp. 442–458.

[78] N. Lomas. *How will Europe's Coronavirus contact-tracing apps work across borders?* 2020.

# Appendix B

# Appendix to Paper 1

Figure B.1: Machine Learning Canvas from L. Dorard (2019) [12]

# Appendix C

# Appendix to Paper 2

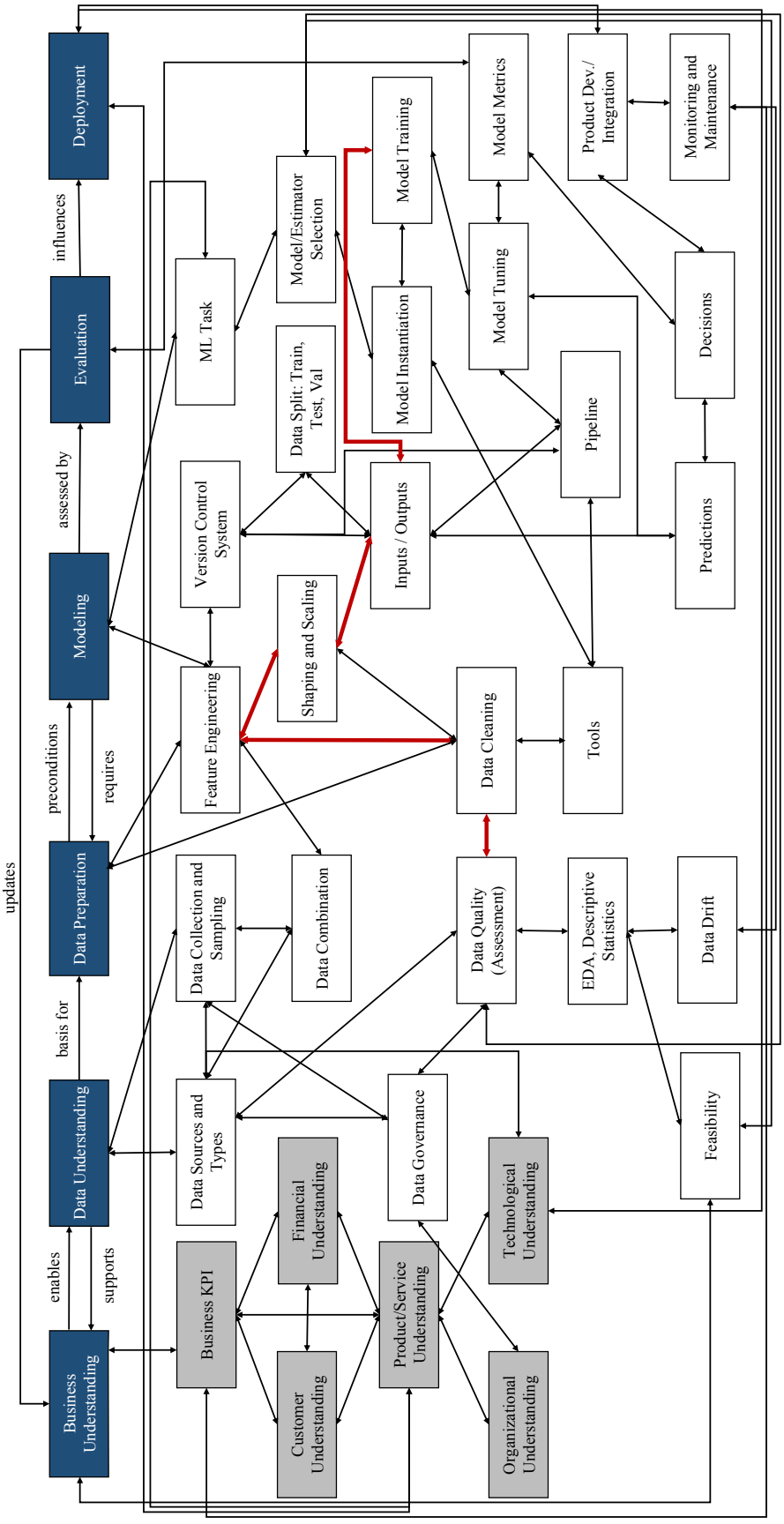


Figure C.1: Simplified view of the ontology of a data science project full page (see Paper 2, Figure 2.2)

# Appendix D

# Appendix to Paper 5

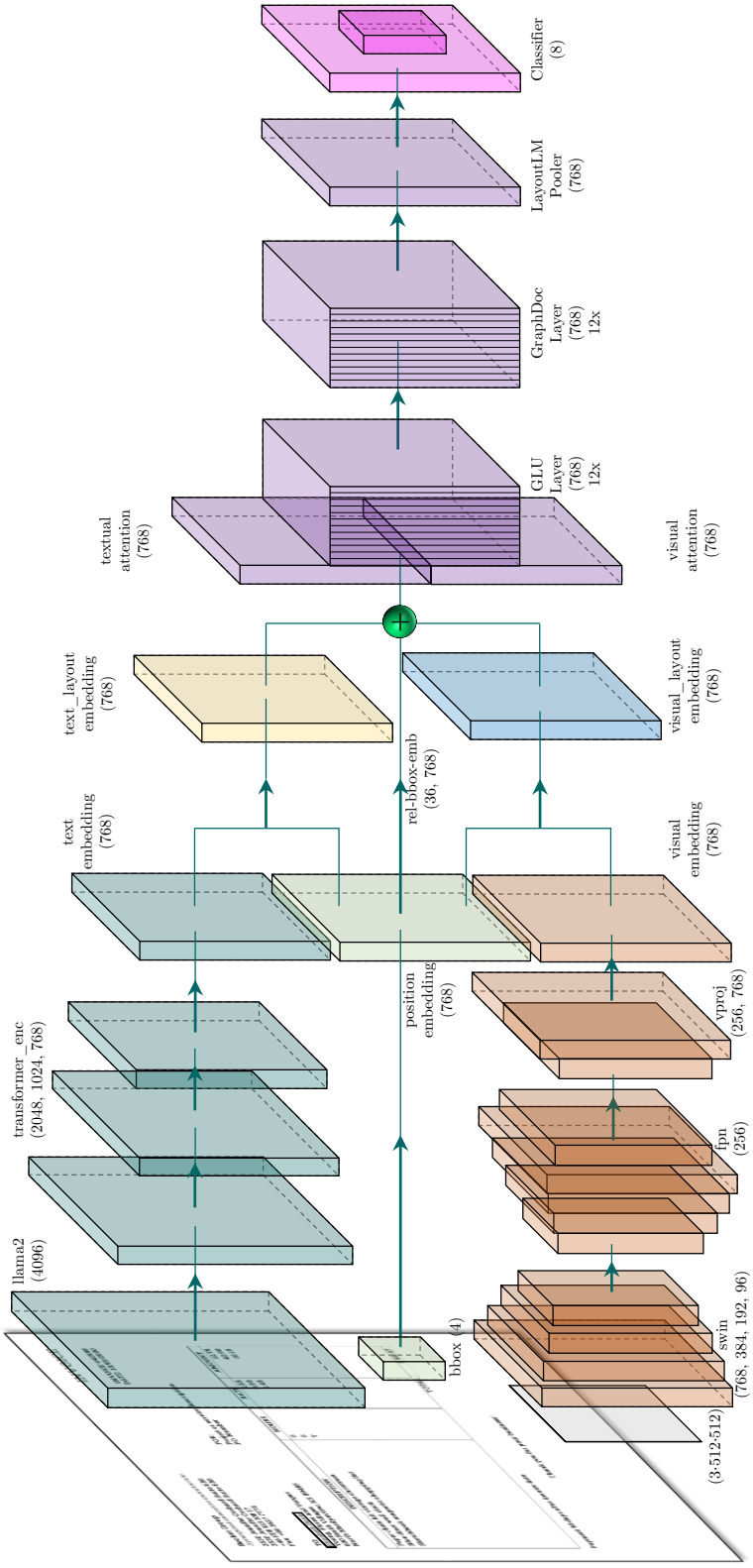


Figure D.1: Simplified model layers: GAT Model full page (see Paper 5, Figure 5.2)