Communications of the Association for Information Systems

Volume 54 Paper 47

9-27-2024

Image-Mining-Based Decision Support Systems: Design Knowledge and its Evaluation in Tool Wear Analysis

Jannis Walk
Karlsruhe Institute of Technology, jannis.walk@googlemail.com

Max Schemmer
Karlsruhe Institute of Technology, max.schemmer@kit.edu

Niklas Kühl
University of Bayreuth, kuehl@uni-bayreuth.de

Gerhard Satzger
Karlsruhe Institute of Technology, gerhard.satzger@kit.edu

Recommended Citation

Walk, J., Schemmer, M., Kühl, N., & Satzger, G. (2024). Image-Mining-Based Decision Support Systems: Design Knowledge and its Evaluation in Tool Wear Analysis. Communications of the Association for Information Systems, 54, 1124-1152. https://doi.org/10.17705/1CAIS.05447

Image-Mining-Based Decision Support Systems: Design Knowledge and its Evaluation in Tool Wear Analysis

Cover Page Footnote

[Note: *Jannis Walk and Max Schemmer contributed equally in a shared first authorship.] This manuscript underwent peer review. It was received 12/21/2022 and was with the authors for 24 months for two revisions. Andreas Drechsler served as Associate Editor.

Communications of the Association for Information Systems

Research Article DOI: 10.17705/1CAIS.05447 ISSN: 1529-3181

Image-Mining-Based Decision Support Systems: Design Knowledge and its Evaluation in Tool Wear Analysis

Jannis Walk*
Karlsruhe Institute of Technology
jannis.walk@googlemail.com
0000-0002-8410-1065

Max Schemmer*
Karlsruhe Institute of Technology
max.schemmer@kit.edu
0000-0001-6341-2051

Niklas Kühl
University of Bayreuth
kuehl@uni-bayreuth.de
0000-0001-6750-0876

Gerhard Satzger
Karlsruhe Institute of Technology
gerhard.satzger@kit.edu
0000-0001-8731-654X

Abstract:

Many decision processes are based on image analysis, for instance, medical diagnoses or visual monitoring of industrial processes. At the same time, advances in deep learning have significantly improved information extraction from images. While recent research strongly focuses on extracting information from single images, the potential of mining entire image collections for decision processes has been neglected so far. In this work, we develop design knowledge to use image collections for improved decision-making. We derive design requirements for image-mining-based decision support systems from literature and expert interviews. Drawing on research in image mining and decision support systems, we conceptualize design principles to address the design requirements. Subsequently, we instantiate and evaluate them in the machining industry with the help of an artifact to support tool wear analysis. The results prove the validity of our design knowledge. Our study contributes to research and practice by deriving nascent design knowledge for image-mining-based decision support systems.

Keywords: Image Mining, Decision Support Systems, Design Science, Deep Learning, Tool Wear Analysis.

[Note: *Jannis Walk and Max Schemmer contributed equally in a shared first authorship.]

This manuscript underwent peer review. It was received 12/21/2022 and was with the authors for 24 months for two revisions. Andreas Drechsler served as Associate Editor.

1 Introduction

Over the past years, rapid and major advances have been made in image processing techniques that target image data and turn it into useful information. Deep learning in particular is often considered a breakthrough (Lecun et al., 2015) in image processing—with novel algorithms well able to outperform humans in image classification tasks (He et al., 2015).

Most existing research and practical applications commonly focus on analyzing individual images to identify or categorize objects. However, there is still a significant untapped opportunity to analyze entire collections of images for new insights. In fields like manufacturing, where processes have inherent variations, relying solely on single images for analysis is unreliable. To put it simply, a single image is like a single entry in a database—it has only limited potential for analysis. To extract knowledge from these collections of data, techniques from data mining can be used (Han et al., 2011; Spangler et al., 1999). However, dealing with images presents challenges that go beyond standard data mining approaches. The concept of “image mining” addresses this challenge and aims to uncover patterns within image data and create value from image collections (Bhatt & Kankanhalli, 2011).

Image data has a high potential to be used in decision support systems (Kohli & Devaraj, 2004) because of its information richness (Lambin et al., 2017). Due to this information richness and recent advances in deep learning, we see great potential in combining research on image mining and decision support systems. “Image-mining-based decision support systems” depict a novel class of systems and can create value in any domain where images need to be analyzed for decision-making. There is a wide variety of potential application areas, such as manufacturing (Trinks & Felden, 2019), medicine (Sollini et al., 2019), sports (W. Sun et al., 2008), geo-information (Coenen & Dittakan, 2016), or services (Villarroel Ordenes & Zhang, 2019). Most of these domains have already conducted initial research on image mining which confirms the potential for value creation. However, there is no general design knowledge that supports the design of these image-mining-based decision support systems. Existing work focuses either on single images rendering decision support (e.g., Shihavuddin et al., 2019; Jourdan et al., 2021) or on creating AI systems for purposes other than decision support (e.g., Ofori et al., 2022; Shin et al., 2020). Therefore, uncovering novel design knowledge is essential to facilitate value creation from image collections, a field with great potential due to the information richness of image data and technological advances in the field of automatic image processing (Dey et al., 2015). The information systems body of knowledge on designing decision support systems (DSSs) should be expanded to enable the incorporation of this important type of data. While other special data types (e.g., text), are traditionally embodied in DSS research (Turban et al., 2011) and design knowledge for text mining is already formulated (Abbasi & Chen, 2008), the potential of image mining has not yet been unlocked—even though the data contains potentially still more information. In the past, the complexity of transforming entire image databases into information may have prevented the systematic development of design knowledge for corresponding systems. Recent advances in the field of deep learning, though, have reduced this complexity and opened the door for systematic development of decision support systems in this area (Lecun et al., 2015). We, therefore, formulate the following research question:

> **Research Question: What design knowledge should guide the development of image-mining-based decision support systems (IM-DSSs)?**

In his outlook for IS research, Lee (2010) stressed that the predominant form of theory should become theory for design and action. We followed common guidelines from design science research (Hevner et al., 2004; March & Smith, 1995; Peffers et al., 2007; Winter, 2008) to derive design knowledge and evaluate this design knowledge in practice. Design knowledge always refers to a class of problems (Gregor & Jones, 2007; Hevner et al., 2004). With our work, we want to shed first light on a problem class that we define as image-mining-based decision support systems. Based on this new problem class, we follow the dual mission of design science research of generating theoretical knowledge and developing usable artifacts (Gregor & Jones, 2007; March & Smith, 1995).

To ensure practical grounding, a key issue in former DSS research (Arnott & Pervan, 2005; Miah et al., 2019), we conducted a design science research project in a manufacturing company that produces machining tools. Machining is one of the most important manufacturing techniques (Arrazola et al., 2013). It is applied in various industries, such as aerospace (Nabhani, 2001), automotive (Dasch et al., 2005), and medicine (Kreiss et al., 1996). Our case company is well-suited for developing and evaluating an image mining artifact because many of the workers have use cases where they need to interpret image collections as part of their daily job. These use cases are concerned with analyzing wear on machining tools, either as customer service or as part of developing new machining tools. This analysis is currently done manually by using magnifiers and microscopes. Based on optical inspection, the domain experts make decisions such as selecting machining parameters (Lukić et al., 1991), tools (Alberti et al., 2011), or coatings (Athanasopoulos et al., 2009). These tasks obtain high economic importance; research suggests that tool failures are responsible for about 20% of production downtime in machining processes (Kurada & Bradley, 1997). Furthermore, cutting tools and their replacement account for 3–12% of total production cost (Castejón et al., 2007).

This research contributes to theory as well as practice. Our work provides design knowledge for image-mining-based decision support systems as an important, but neglected class of information systems. Therefore, our first theoretical contribution is depicting and discussing a new problem class. Based on that, we conducted an exploratory study in the machining industry to derive initial design requirements. To address these design requirements, we conceptualized design principles based on previous work in image mining, deep learning, and decision support systems. These design principles could work as a "blueprint" for upcoming image-mining-based decision support systems (Gregor & Jones, 2007). Based on these design principles, we derived design features as specific implementations for the machining industry. These specific design features were used to develop an artifact that allowed us to rigorously evaluate the design knowledge in practice. We use this instantiation to solve a real-world problem at our case company. The concrete artifact supports developers and researchers at the case company by removing manual work and supporting the knowledge generation process.

The remainder of this work is structured as follows: In Section 2, we introduce our design science research methodology. On that basis, we derive tentative design requirements based on an exploratory interview study and literature in Section 3. We then present related work and the key concepts of our research (Section 4). With the necessary foundations at hand, in Section 5, we derive and evaluate design knowledge for image processing. Based on this, in Section 6, we derive further initial design principles, following the image mining process and evaluate them. We then refine our derived design knowledge and the artifact based on the prior feedback and evaluate it as part of Section 7. In Section 8 we discuss our results and present our developed design theory. Lastly, in Section 9, we summarize this research, explain the limitations of our study, and provide an outlook on future work.

2 Research Design

Our research follows the design science research approach (Hevner et al., 2004; March & Smith, 1995). In particular, we follow the three-cycle design science research guidelines from Hevner (2007) – the relevance, rigor, and design cycle (DC). The relevance cycle provides the research with environmental requirements and ensures field testing (input from the practical knowledge base). The rigor cycle provides the research with grounding theories and methodology from the knowledge base (input from the theoretical knowledge base). A design cycle incorporates the design, development, and evaluation of artifacts. For our research, we conducted three DCs each refining the proposed design knowledge and the artifact. Each DC follows the steps of awareness of problem, suggestion, development, and evaluation, based on Kuechler and Vaishnavi (2008). In the awareness of problem step, we draw from the practical knowledge base and gather requirements. In the suggestion phase, we derive design principles (DPs) based on the theoretical knowledge base. In the development phase, we map the abstract DPs in specific design features (DFs) and implement them in an artifact that is used to evaluate the design.

A key aspect of design science research is the rigorous evaluation (Brendel et al., 2021; Peffers et al., 2012). To structure the evaluation, we follow the guidelines of Venable, Pries-Heje, and Baskerville (2016). In their framework, the evaluation is structured in distinct evaluation episodes (EEs), and each DC can be evaluated with multiple distinct EEs.

In general, Venable et al. (2016) differentiate two major purposes of evaluation – formative and summative. While formative evaluation addresses artifact refinement, summative evaluation is used to depict the results of the completed development. We conduct multiple formative and summative EEs, addressing different goals and following a mixed-method approach by conducting quantitative and qualitative EEs to provide a comprehensive analysis.

In the first DC, we derive design knowledge for image processing. In the evaluation phase, we show the feasibility of deep learning (DL) as a prepossessing step for image mining by conducting a technical experiment, as Peffers et al. (2012) proposed. After targeting the technical viability of the image processing in DC1, in DC2 we derive further design knowledge for IM-DSSs and show the general desirability. We, therefore, develop an artifact, the automatic tool wear analyzer, and evaluate it in a formative way by using exploratory focus groups (Tremblay et al., 2010a). In the final DC3, we refine our design knowledge and the prototype from the previous DC and perform a summative evaluation conducting four distinct EEs. First, we assess the effectiveness with an additional technical experiment that compares humanly derived features with features that our artifact automatically extracts. Second, we calculate the efficiency of our system by measuring potential savings in human working time. Additionally, we use confirmatory focus groups, as Tremblay et al. (2010a) propose, to gather qualitative feedback about the artifact's usefulness. Lastly, to validate perceived usefulness, we conducted a survey based on questions from the technology acceptance model (TAM) (Venkatesh & Bala, 2008). Figure 1 visualizes the interplay of the three DCs and our EEs, including the objectives, the methods, and our results.

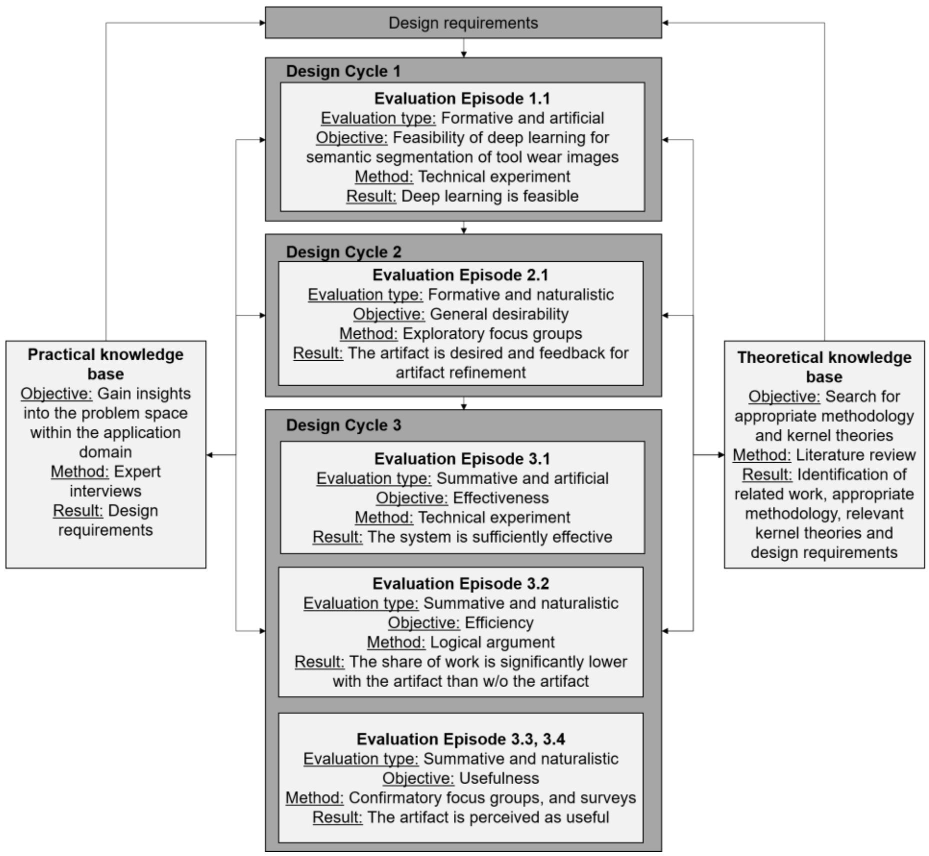

Figure 1. Design Science Research Methodology (Based on: Hevner (2007), Kuechler and Vaishnavi (2008), and Venable et al. (2016))

This work is part of a larger research program on DSSs based on image data. It was developed in parallel with Landwehr, Kühl, Walk, and Gnädig (2022) and elaborated on with overlapping author teams. While both manuscripts focus on DSSs based on image data there is a crucial difference: Landwehr et al. (2022) develop design knowledge for DSSs based on *single* images and the work at hand develops design knowledge for DSSs based on *collections of* images. The differences and similarities of the design knowledge developed by us and Landwehr et al. (2022) are further discussed in detail in Section 8.2.

3 Relevance Cycle: Design Requirements for Image-Mining-Based DSSs

To ensure a grounding in practice, we derive the design requirements (DRs) based on domain requirements collected through an exploratory study in a case company. We purposefully selected the industry and the case company based on the importance of the image data analysis. We selected a manufacturing company that produces cutting tools for the machining industry. Because the analysis of these tools is mainly done visually, the employees have a lot of experience in image analysis. An additional advantage of the case industry is that there is potential to get high-quality standardized images that facilitate image processing. The industry in general, as well as the particular case company, are therefore highly suitable for IM-DSS research.

In general, machining is a key manufacturing process (Arrazola et al., 2013) that describes the mechanism of removing unwanted material from a workpiece using a cutting tool (Black, 1989). There are several types of cutting tools, for example, drills and indexable inserts. We focus on the latter, shown in Figure 2.

The exploratory study had two goals: First, collecting potential use cases of IM-DSSs in the machining industry, and second, gathering domain requirements for the chosen use case. Based on the domain requirements, we then derive the first preliminary DRs for IM-DSSs. In the following subsection, we describe the methodology and results.

Figure 2. Exemplary Machining Process with an Indexable Insert

3.1 Methodology: Exploratory Study

To conduct insightful interviews, we used purposive sampling (Coyne, 1997) and snowballing as a sampling technique (Palinkas et al., 2015). We started the study by interviewing managers who are well-connected in the company and have a good overview of applications of tool wear analysis. Afterward, we used the snowballing technique and asked each interviewee for other relevant interviewees. The suggestions were clustered into stakeholder groups based on their job descriptions. We found two major stakeholder groups that can directly profit from IM-DSSs. First, there are application engineers, who are the technical interface to the customers. Analyzing images is a crucial building block in their task of developing recommendations for customers to improve their machining processes. DSSs can support them in tool or machining parameter selection. The second group of stakeholders is developers who improve and develop products using image analysis to get insights regarding the tool wear mechanisms. These insights can be used to improve specific tool development decisions, like the coating choice.

We chose semi-structured interviews because they allow for flexibility (Whiting, 2008). We interviewed 19 experts (9 application engineers, 3 managers, and 7 developers). Some were interviewed in group sessions because their application domain was similar. Each interview lasted between 30 and 70 minutes. After conducting the interviews, we transcribed and coded them in MaxQDA (Kuckartz, 2012). Potential application areas were summarized, whereafter we chose the most promising use case. We then analyzed nine sessions in-depth to extract domain requirements. Because we conducted exploratory interviews, we decided to use inductive coding for the analysis (Thomas, 2006). We chose this route instead of the deductive approach because we were in an exploratory phase, aiming to find the stakeholders' important business issues. With the deductive approach, key topics could be ignored (Thomas, 2006). Based on the identified codes, we derived eight initial tentative DRs.

3.2 Results of the Exploratory Study

We found 12 image mining use cases in the case company[1]. For our research, we chose a service for customers of our case company where experts (application engineers and developers) analyze machining

[1](1) Tool condition monitoring; (2) Single-image recommender system; (3) Quality assurance; (4) Benchmark tests; (5) Consumer complaint analysis; (6) Coating analysis; (7) Chip breaking analysis; (8) Predictive maintenance; (9) Recommender systems; (10) Customer services; (11) Market-driven development; (12) Market trend analysis

processes based on visual inspection of cutting tools and give recommendations on how to improve the process — optimization-as-a-service. During the process experts inspect, with the help of a microscope, the surface of many cutting tools from a given machining process to identify patterns. Subsequently, they make recommendations regarding machining process improvements based on these patterns. We identified this use case as highly suitable for an IM-DSS because of two main reasons. First, many different groups of employees are involved in the creation of this service – hence, we can involve many domain experts with diverse backgrounds for the development and evaluation of the design knowledge. Second, image mining needs huge image collections to be useful (Hsu et al., 2002). This use case is suitable in this regard since it enables an efficient collection of relevant images.

In the following part, we present and discuss our preliminary DRs, derived from the coded interviews. Because we develop generalized design knowledge for IM-DSSs, we formulate them on the relevant abstraction level. Table 1 shows exemplary quotes from the interviews done in the case company to illustrate the derivation. We also support the DRs with evidence from the literature.

Table 1. Tentative Design Requirements, Exemplary Quotes, and Support from Literature

Design requirement	Exemplary quote	Support from literature
DR1 (Context)	"[...] without the metadata, it is half the truth, or it can falsify the truth. There is a danger of something being misinterpreted."	(Lambin et al., 2017)
DR2 (Image quality)	"Unfortunately, we have had the experience that misinterpretations are made. That's why I am so fussy about the quality of the original picture. Because you can only make a clear statement by the quality of the original picture."	(Afshar et al., 2019)
DR3 (Comparison)	"The approach is mostly that you have some kind of benchmark, let's say when turning. Then you find out that your competitors are reaching this tool life and then you make your test variants."	(Liu et al., 2018)
DR4 (Scalability)	"Now we're going to do a long-term experiment of 300 or 400 records, it depends. They'll be tested."	(Kumar et al., 2012)
DR5 (Cost-effectiveness)	"I think that's the biggest challenge because we can't afford to have five million images labeled by any user like Google."	(Wang et al., 2016)
DR6 (Reproducibility)	"If you ask three experts now, you'll get five opinions."	(Patel & Sethi, 2007)
DR7 (Dispersion)	"Reality simply has a certain spread, on the machine side, cooling, material being machined. Our products also have scatters. Then we haven't talked about the tool holder. If it's brand new, the plate tends to be more stable and work better when it fits, like a tool that's already worn out and you just get vibration. Sometimes it's just the screw. Doesn't tighten properly anymore. If it's screw tension, the plate can be as good as it wants if it doesn't fit properly."	(Grove et al., 2015)
DR8 (Exploration)	"Yes, above all, maybe we can draw conclusions, maybe our phase or our geometry is not stable enough at that point, because you always get wear at the same place."	(Gillies et al., 2016)

According to the experts, context is needed to generate significant value from image data. In terms of tool wear analysis, this means information about the cutting process parameters like the cutting speed, the workpiece, and the tool. We therefore formulate the following DR:

DR1 (Context): The system should ensure the availability of context.

Another important aspect is the quality of the input images. The necessary quality of the recording depends on the goal of the IM-DSS. For example, if the goal is generating very general recommendations for cutting parameters, a low resolution might be sufficient. Nevertheless, the recording quality must match the goal of the IM-DSS.

DR2 (Image quality): The system should ensure appropriate image quality.

A recurring theme of the experts was the comparison of image collections in benchmark tests. An image collection in the machining industry usually represents the wear on cutting tools that were used in a defined production process. After changing process parameters such as the type of cutting tool, the outcomes need to be compared with each other. The image collections are currently being compared with small sample sizes and are subject to individual assessment. Experts consequently formulated a need for more reliable comparisons. We therefore formulate the following DR:

DR3 (Comparison): The system should increase the validity of the comparison of image collections.

Furthermore, the system must be able to handle large image collections efficiently. The experts described situations where 400 inserts have to be analyzed.

DR4 (Scalability): The system should enable scalability of image analysis.

Additionally, when dealing with data-heavy algorithms, like DL, it is necessary to keep the cost factor (like compute cost for model training and employee cost for data labeling) in mind. An interviewee in a management position specifically emphasized this point. We therefore formulate the following DR:

DR5 (Cost-effectiveness): The system should be cost-effective.

The inherent problem of image data is that it is usually subject to human interpretation (Patel & Sethi, 2007), which can lead to a huge variance in generated recommendations. Many domain experts stated a need for reproducibility in tool wear analysis. Additionally, the measurement of image features incorporates a significant variance that should be reduced. We therefore formulate the following DR:

DR6 (Reproducibility): The system should decrease the variance of feature measurement and human image interpretation while keeping the quality at least equal.

In industrial settings, a single image is often just a snapshot. To derive knowledge, the dispersion of the process must be displayed.

DR7 (Dispersion): The system should capture and display the dispersion of an image collection.

Lastly, data mining needs exploration and hypothesis generation potential (Gillies et al., 2016). The experts stressed the complexity of tool wear analysis and the exploratory nature thereof. We therefore formulate the following final DR:

DR8 (Exploration): The system should provide users with the possibility of exploring image collections interactively.

4 Rigor Cycle: Theories Informing the Design of Image-Mining-Based DSSs

With the results of the relevance cycle at hand, we describe the theoretical foundations of our research. First, we delve deeper into image mining. Next, we present related work of image-mining-based decision support systems by summarizing the results of a structured literature review (SLR). Finally, we provide insights on the analysis of tool wear.

4.1 Image Mining

Image mining is the extraction of knowledge from large image collections by utilizing techniques from image processing and data mining to improve decision-making in an image-rich domain (Hsu et al., 2002). Image mining is applied in multiple domains, such as medicine, where it is called radiomics (Gillies et al., 2016; Lambin et al., 2012).

The image mining process follows the data, information, knowledge, and wisdom (DIKW) hierarchy (Ackoff, 1989) by first transforming image data into information and subsequently into knowledge. Mishra and Silakari (2012) and Khodaskar and Ladhake (2014) outline the traditional process of image mining. In the first step, the images, meaning the *data*, must be stored in an image database. The next step is to preprocess the data – crop the images or improve their quality. Then the region of interest (ROI) should be segmented, meaning that each pixel is classified (compare Figure 7 on page 14 for an example). Thereafter, features like color or texture are extracted from the ROI, transforming the image data into processable *information*. These features can then be analyzed with data mining techniques to find patterns and generate *knowledge*.

A key step in the image mining process is the segmentation of ROIs because it is the basis of the feature extraction (Gillies et al., 2016). In the following part, we will present and discuss technical options for segmentation – manually, semi-automatically, or automatically. If performed manually, an expert defines and segments each ROI, using image labeling tools like Fisher and Mackiewicz (2020). Manual segmentation has the disadvantage of significant intra- and interobserver variability (Louie et al., 2010), and requires considerable manual effort. Automatic segmentation can be differentiated in traditional computer vision approaches and machine-learning-based approaches. Semi-automatic approaches combine both techniques, for example by pre-segmenting the images automatically and refining the segmentation by experts. Here we would like to emphasize that computer vision is an important step in image mining – but only one step.

For computer vision tasks, DL techniques have been shown to be the most suitable machine learning approach. DL is a subset of machine learning and is based on artificial neural networks, which simulate functionalities of the human brain (O'Mahony et al., 2019). In contrast to traditional approaches, DL does not require extensive feature engineering and therefore increases the scalability of image analysis (O'Mahony et al., 2019). DL has also outperformed humans in image classification tasks (He et al., 2015). Lastly, DL provides superior flexibility, because the models can be retrained for similar tasks (Pan & Yang, 2010).

After the segmentation, image mining techniques need to be applied to generate knowledge from the information – in contrast to single image analysis the information extracted from an image collection needs to be aggregated. Hsu et al. (2002) give a holistic overview of the most common image mining techniques. Traditional techniques comprise image retrieval, image classification, and clustering or association rule mining.

4.2 Image-Mining-Based Decision Support Systems

"Decision support systems is a general term for any computer application that enhances a person or group's ability to make decisions" (Power, 2008). Our work synthesizes the long-lasting knowledge body of decision support systems with image mining. This provides new potential by transforming image data into knowledge, thereby creating competitive advantages out of image collections. To ensure a rigorous overview of former research, we conducted an SLR following Cooper (1988) and Vom Brocke et al. (2009) by searching all fields in the Scopus database with the following search string:

(radiomics OR "image mining" OR "image data mining") AND "decision support system"

The papers were included if their abstract and/or title focused on DSSs and image mining. We applied three exclusion criteria: First, we excluded the paper if the researchers focused purely on object recognition or image classification because these are just preprocessing steps of image mining and focus on single images. Second, because we are interested in DSS design, we excluded papers that focused purely on algorithms. Lastly, we excluded research about image retrieval because this is a specialized subfield (Hsu et al., 2002) and we are interested in holistic solution approaches.

We conducted the SLR in November 2022 and identified 1328 potentially relevant papers, mostly from the field of radiomics research. Most research in radiomics outlined the development of a DSS as potential future research but did not develop one. This led to a significant reduction in the number of papers. After evaluating all abstracts, 18 papers remained. After reading the papers in detail, we excluded eleven more based on our previously defined exclusion criteria (research that focused purely on object recognition or image classification because these are just preprocessing steps of image mining, research that focused purely on algorithms, and research that focused on single images). The seven relevant papers are outlined below.

Most previous research was conducted in medical application areas. Exceptions are the works of Zaiane, Han, Li, Chee, and Chiang (1998) and Koh and Cui (2022). Zaiane et al. (1998) developed a DSS for multimedia mining. Koh and Cui (2022) developed an IM-DSS to analyze the impact of visual attributes of thumbnails on the view-through of videos. In terms of medical research, Berlage (2007) reviewed image mining for biomedical imaging experiments. Barnathan, Zhang, and Megalooikonomou (2008) designed a framework for image mining and instantiated it in a web application for mining medical image data. Foran et al. (2011) developed software for the image mining of tissue micro-arrays. It comprises image processing, segmentation, feature extraction, and classification. Gatta et al. (2019) built a holistic IM-DSS and evaluated it on two data sets of cancer patients. Cheng et al. (2019) developed a clinical decision support system aimed at weight loss prediction after head and neck cancer.

Our SLR shows that the existing research on IM-DSSs largely focuses on the medical field. Our study allows generalization by conducting research in the machining industry. Furthermore, previous research on IM-DSSs did not leverage the advantages of DL. On a more general level, existing research on IM-DSSs did not aim to develop generalized design knowledge. In addition, none of them used real-world expert input to derive requirements. Lastly, except for Cheng et al. (2019), the impact of the system was not evaluated with potential end users.

Besides the results of the SLR we would like to mention two works that we are aware of that derive design knowledge for similar classes: Landwehr et al. (2022) and Zschech, Walk, Heinrich, Vössing, and Kühl (2021). Landwehr et al. (2022) derive design knowledge for image-based DSSs, this is the first study developing and evaluating design knowledge for image-based DSSs. For their study, they conduct a case study on power line infrastructure maintenance. They use images from unmanned aerial vehicles to analyze the wear state of power line infrastructure. Their DSS provides decision support in scoping and planning maintenance orders through improved data and information quality. However, they focus on decision support based on single images. Hence, no image mining is required. Zschech et al. (2021) develop design knowledge for computer vision-based hybrid intelligence systems. They focus on design knowledge that facilitates hybrid intelligence (i.e., the combination of human and artificial intelligence) in any information system based on computer vision. The design knowledge we develop in this work differs in two ways. First, we target concrete design knowledge for DSSs while focusing less on the hybrid intelligence part. Second, we address design knowledge for DSSs that rely on image mining in particular.

4.3 Machining and Tool Wear

The design knowledge is instantiated and evaluated in the machining industry. Machining is applied in various industries, such as aerospace (Nabhani, 2001), automotive (Dasch et al., 2005), and medicine (Kreiss et al., 1996).

The machining process unavoidably results in tool wear. In the following part, we shortly describe the three most common types of tool wear (compare Figure 3, Figure 4 and Figure 5 for exemplary images). The first type is abrasive wear on the flank, called *flank wear* (Kuram & Ozcelik, 2014). Flank wear is unavoidable and the most frequent wear characteristic (Siddhpura & Paurobally, 2013). For this reason, it is the most commonly used criterion for evaluating tool life, meaning deciding when to change a tool (ISO, 1993). Another wear characteristic that frequently and heavily impacts product quality is *chipping*. This refers to particles of the cutting edge breaking off or thermal cracking (ISO, 1993). Lastly, high temperature and pressure can lead to *built-up edges* (Kuram & Ozcelik, 2014). Both chipping and built-up edge lead to deformations of the cutting edge, this may lead to insufficient product quality, increased scrap, and high costs.

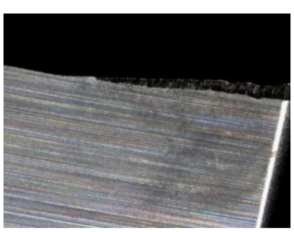

Figure 3. Flank Wear

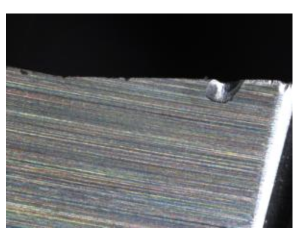

Figure 4. Chipping

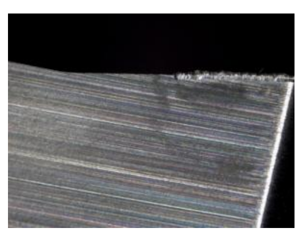

Figure 5. Built-up Edge

5 First Design Cycle: Effective and Scalable Image Mining

Our overall goal in the DCs is to derive design knowledge for IM-DSSs and evaluate it with the help of a developed artifact. Figure 6 summarizes our DRs, as well as our suggested DPs and DFs, iteratively derived over three DCs.

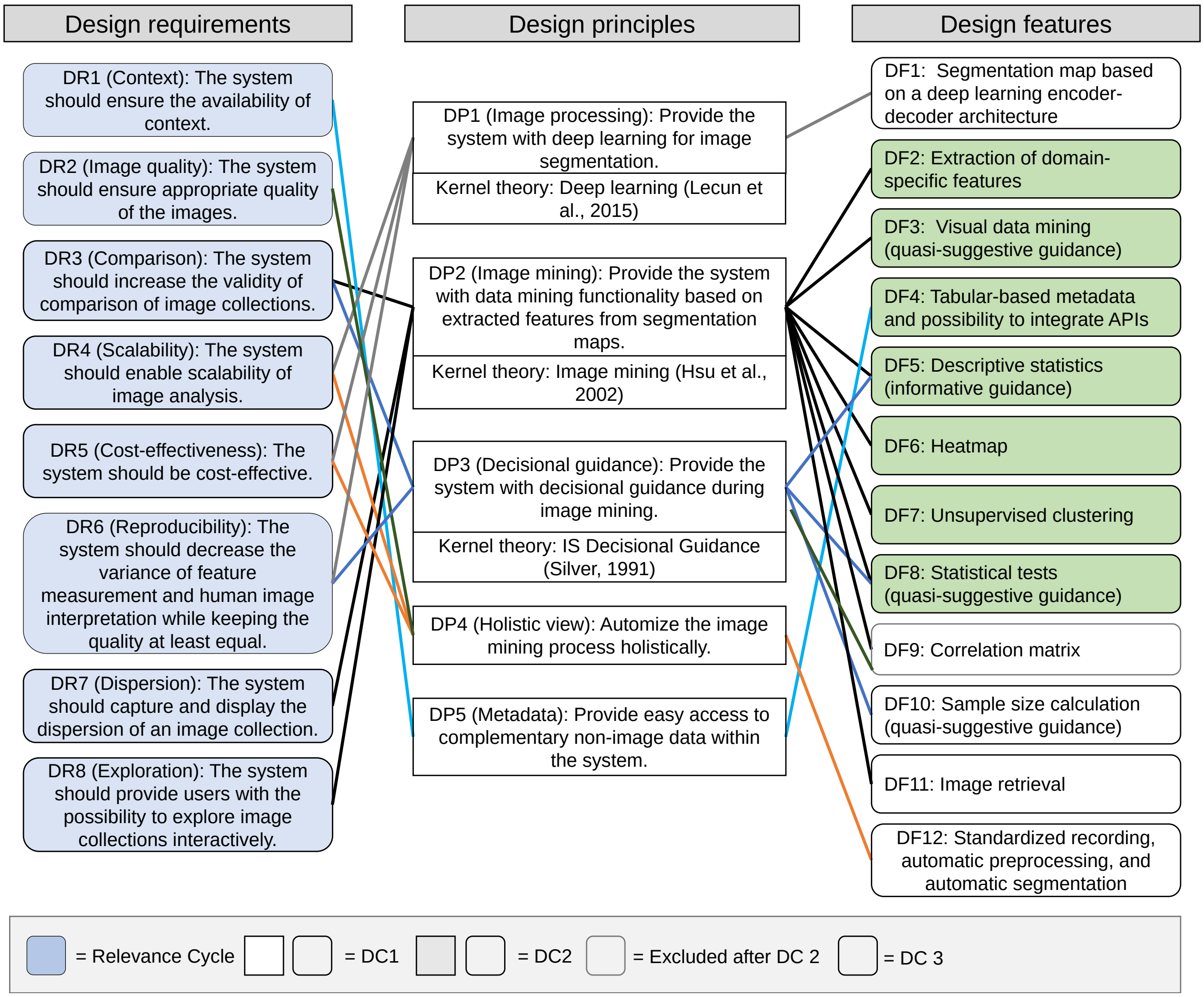

Figure 6. Design Knowledge for Image-Mining-Based Decision Support Systems (colors of lines indicate connections from DRs to DPs to DFs). DCs Refer to Design Cycles, DRs to Design Requirements, DPs to Design Principles, and DFs to Design Features

As explained in Section 4.1, image mining starts with preprocessing and segmenting the images, enabling a transformation of the image data into processable information. For this reason, the goal of the first DC is to derive design knowledge for this initial transformation process.

5.1 Suggestion and Development

Deep learning models, in particular convolutional neural networks, have shown significant performance in computer vision tasks, outperforming humans in image classification (He et al., 2015). This shows the potential of automated image analysis of human-level quality and addresses the fourth DR (scalability) and the fifth DR (cost-effectiveness). We therefore formulate the following DP:

> **DP1 (Information extraction): Provide the system with deep learning capabilities for image segmentation.**

To implement the DP, we propose the following DF. We used the U-Net[2] (Ronneberger et al., 2015), an encoder-decoder CNN architecture, as a basis for our DL model because it has shown high performances with low amounts of labeled images, addressing DR7 (cost-effectiveness).

We used an open-source web application to label images on a pixel level (Fisher & Mackiewicz, 2020). Our image data set comprises 213 labeled cutting tool inserts that customers of our case company used in real machining processes. The labeling covers four classes: The three major tool wear characteristics (flank wear, chipping, and built-up edge – see Section 4.3) as well as the remaining background pixels. The data set has two major sources of imbalance. First, because flank wear is the major wear type, there is an imbalance on an image level. Second, on a pixel level, even after cropping, the images contain more background pixels than wear pixels. To address the imbalance of the dataset, we used a weighted cross entropy as a loss function that equalizes the weighting of the classes (Ronneberger et al., 2015). Figure 7 shows an exemplary image, label, and output.

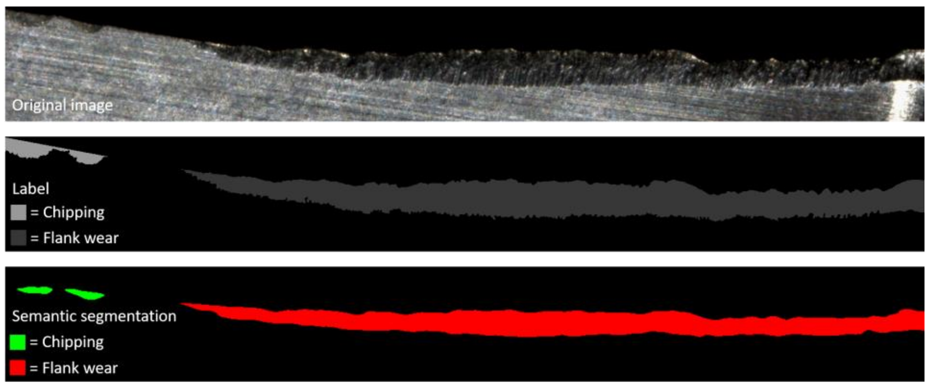

Figure 7. Transformation from Original Recorded Images to Segmentation Maps

5.2 Evaluation and Discussion

The test data set incorporated 51 images and was hand-selected by experts to ensure representativeness. To evaluate the results of the segmentation, we used pixel accuracy and the mean dice coefficient. We did so because accuracy is easy to interpret, and the mean dice coefficient takes the class imbalance into account. As Garcia-Garcia et al. (2017) proposed for the related Jaccard-coefficient, the mean dice coefficient is a function of precision and recall calculated for all classes and averaged (see Equation (1)).

$$\boldsymbol{Mean\ dice\ coefficient} = \frac{\mathbf{2}}{\mathbf{C}}\sum_{c=1}^{C} * \frac{\boldsymbol{precision_c * recall_c}}{\boldsymbol{precision_c + recall_c}} \qquad (1)$$

We reached a pixel accuracy of 0.977 and a mean dice coefficient of 0.631. Table 2 shows the specific results of all four classes.

Table 2. Performance Results of the Semantic Segmentation

Class	**Dice coefficient**
Background	0.991
Flank wear	0.695
Chipping	0.244
Built-up edge	0.596
Mean dice coefficient	0.631

Flank wear and built-up edge can be predicted with sufficient consistency. Due to a low number of chipping labels, the algorithm has difficulties predicting this class. Overall, we trained with a relatively

[2] Because deep learning has developed rapidly in recent years, we have made some changes to the original model, addressing overfitting and performance issues. In particular, we added batch normalization (Ioffe & Szegedy, 2015) and L2-regularization (Cortes et al., 2012). We implemented the U-Net for semantic segmentation in Keras (Chollet, 2015). We trained the model for 200 epochs, using an Adam optimizer and a learning rate of 0.00001.

small data set. Research has shown that DL performance grows logarithmically with increasing data-volume (C. Sun et al., 2017). Increasing training data, especially when starting with small datasets, therefore has a significant influence on the performance. We conclude that the results show the general technical feasibility of DL as a preprocessing step for image mining in tool wear analysis.

6 Second Design Cycle: General Desirability of Proposed Image-Mining-Based DSS

The second DC is conducted to derive DPs to process the transformed image data further. Subsequently, the DPs are instantiated in an artifact and evaluated to show the general desirability and collect formative feedback for refinement.

6.1 Suggestion

Image segmentation is just one step in the image mining process (Lambin et al., 2017). To extract knowledge, the resulting segmentation maps need to be processed further. A common step in literature is the derivation of features from the segmentation map (Lambin et al., 2017). We propose to calculate domain features based on the segmentation maps to enable data mining and aggregation of the information contained in image collections. This addresses **DR3** (Comparison), as unprocessed image data is unstructured, and statistical comparisons can only be applied to structured numerical or categorical data (Müller et al., 2016). Furthermore, the DP addresses **DR7** (Dispersion) and **DR8** (Exploration). We therefore formulate the following initial DP:

> **DP2 (Image Mining): Provide the system with the ability to aggregate features extracted from segmentation maps.**

Tool wear analysis has dispersion on two levels (i.e., the result of tool wear analysis performed by two different domain experts can differ due to two reasons), as summarized in **DR6** (Reproducibility). First, the tool wear measurements they make can vary. Second, even if their tool wear measurements are equal their resulting decisions might differ. Automatic segmentation and feature extraction reduce the dispersion on the first level. To address the second level, we propose to use design knowledge from the decision support system body of knowledge and implement features of decisional guidance. Decisional guidance is defined as structuring and guiding the user's decision-making process (Silver, 1991). Silver (1991) differentiates three forms of guidance – informative, suggestive, and quasi-suggestive. Informative guidance provides decision-makers only with decision-relevant information, whereas suggestive guidance makes judgmental recommendations (Silver, 1991). Quasi-suggestive guidance is guidance “that does not explicitly make a recommendation but from which one can directly infer a recommendation or direction” (Silver, 2006). Research has shown that decisional guidance can improve decision-making and decrease the variance in generated decisions (Sharda et al., 1988). We therefore derive the following DP:

> **DP3 (Decisional guidance): Provide the system with decisional guidance during image mining.**

To address **DR4** (Scalability) and **DR5** (Cost-effectiveness) further, the process of tool wear analysis needs to be viewed holistically (i.e., end-to-end). This starts with item collection, followed by recording, segmenting, and finally analyzing. A bottleneck of image mining can occur in each step. In terms of tool wear analysis, the bottleneck is the creation of tool wear. Depending on the material used, the process of wear creation can be protracted. We therefore formulate the following DP:

> **DP4 (Holistic view): Automize the image mining process holistically.**

Lastly, to address **DR1** (Context), research in radiomics has shown that the system needs to provide metadata to enable image mining (Bannach et al., 2017). Especially in exploratory analysis, image features, and non-image features should be combined in a single dataset to enable the investigation of relationships (Lambin et al., 2017). Image mining frameworks like the one by Hsu et al. (2002) define metadata as a key element to extract knowledge. We therefore formulate:

> **DP5 (Metadata): Provide easy access to complementary non-image data within the system.**

6.2 Development

To test the DPs in practice, we translated them into concrete DFs that address the specific project environment. For the second DP, image mining, many specific DFs are possible. Figure 8 visualizes different options on a high level. Our chosen DFs are highlighted. The options can be clustered into three major categories: segmentation, feature engineering, and data mining.

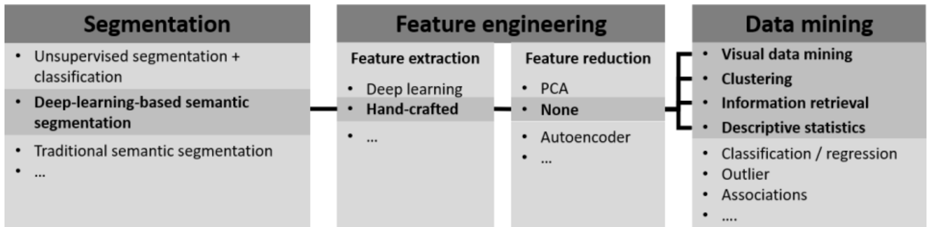

Figure 8. Possible Design Features for Image Mining

As stated in Section 5, an algorithm to derive the segmentation map is necessary **(DF1)**. Due to the reasons explained in Section 5, the segmentation map is created by using DL and in particular CNNs. Next, based on that segmentation map, domain features are extracted. There are multiple ways to derive the domain features, the two major methods being the traditional approach and automatic feature generation (Afshar et al., 2019). The traditional approach uses handcrafted features, while the automatic approach derives these features by utilizing techniques like DL (Afshar et al., 2019). We decided to use the traditional approach to use domain knowledge and enable data exploration **(DF2)**. Because some of the features were familiar to end-users, we aimed to create trust. The relevant domain measures were defined with experts and based on the corresponding ISO standard (ISO, 1993). Table 3 summarizes the handcrafted features.

Table 3. Domain Features Extracted from Segmented Tool Wear Images

Feature	Description
Maximum flank wear	Maximal height of the flank wear
Flank wear	Average height of the flank wear
Flank wear length	Length of the flank wear
Flank wear area	Size of the ROI of the flank wear
Homogeneity	Dispersion in the distribution of flank wear height
Number of chippings	Number of chippings
Chipping area	Size of the ROI of the chipping
Built-up edge area	Size of the ROI of the built-up edges

The third category includes data mining techniques chosen with respect to the application domain and the use case. Tool wear analysis is exploratory, seeking to understand tool wear mechanisms. We therefore draw mainly from the knowledge base of exploratory data mining. In general, the development of the graphical user interface was guided by principles of visual data mining (Keim, 2002) **(DF3)**. Visual data mining aims to facilitate human machine collaboration in the data exploration process (Keim, 2002).

The first and second DF were developed in Python and are the input for the DSS. The data mining techniques were prototyped in Tableau. The prototype allowed us to engage in further discussions with the experts. We called the resulting artifact the **automatic tool wear analyzer (ATWA)**. ATWA supports the analysis of tool wear experiments as well as the subsequent decision-making. In the experiments, the domain experts vary parameters of the machining process and measure target values like the flank wear.

The data mining techniques were implemented in four views – the aggregated, detailed, comparison, and exploration view. Their design was guided by the information seeking mantra: *Overview first, zoom and filter, and then details-on-demand* (Shneiderman, 1996). Each view needs relevant metadata **(DF4)** to enable filtering and facilitating the analysis of the tool wear mechanisms. Together with the experts, necessary metadata was defined.

In the following part, we describe each view and the corresponding DF in detail: To enable the overview, the aggregated view summarizes the image collection of a selected tool wear experiment. For summarizing, we apply multiple methods from the descriptive statistics knowledge base **(DF5)**. We use histograms, boxplots, and summary statistics such as the mean and the variance. Additionally, we implement statistics from the domain, in particular the C_pk value. The C_pk value has its origins in the manufacturing industry and measures whether a process is capable of reproducing items within specification limits (Pearn & Lin, 2004). Inverted, the C_pk value can also be used to calculate the process capability of tools used to produce the items, as shown for example by Nabil and Mabrouk (2006). To add quasi-suggestive guidance, we use knowledge from the existing body of tool wear analysis. As explained in Section 4.3, the common criterion for tool life evaluation is the maximum flank wear. The ISO standard defines a maximum flank wear of 0.3 as an end-of-life criterion (ISO, 1993). This criterion is implemented as default value for an upper limit that is graphically highlighted in the histogram and boxplot.

Additionally, to enable a visual exploration of the image collection, we propose using heatmaps **(DF6)**. The heatmap is the aggregation of all selected segmentation maps. This means that the numerical array representation of the segmentation maps is summarized. The heatmap allows users to visually explore the tool wear characteristics. Figure 9 shows the approach for the tool wear characteristic chipping for 95 images. Blue values imply low occurrence of the characteristics in that area, and red values point to high occurrence. As the image shows, mainly chippings in the left side of the cutting edge are occurring, pointing out the mechanism of chip breaking. These two DFs, descriptive statistics **(DF5)** and heatmaps **(DF6)**, allow the user to get a fast overview.

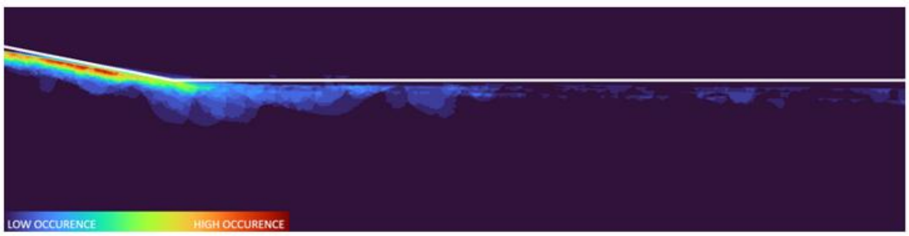

Figure 9. Heatmap for Chippings

The second view is the detailed view that enables the analysis on an image level. The view provides access to the original images and depicts the segmentation maps. An additional feature is the clustering of the flank wear pixels **(DF7)**. To guide the experts, we colored small groups red to draw the attention of the experts to regions in the image that may have been ignored otherwise.

A major task in tool wear analysis is the comparison of the performance of tools, therefore the third view enables statistical comparison of two machining experiments **(DF8)**. Quasi-suggestive guidance was implemented by utilizing symbolic guidance in the form of traffic lights.

The last view, the exploration view, incorporates a correlation matrix **(DF9)** where the user can interactively explore correlations of image mining features and corresponding metadata. By selecting two features, the correlation matrix changes into a scatterplot visualizing all data points and a trend line. The prototype can be viewed here: https://youtu.be/1UdqHV35lkc.

6.3 Evaluation and Discussion

To evaluate the prototype, we use exploratory focus groups (Tremblay et al., 2010a). The goal is to discuss the usefulness of the proposed DFs and generate feedback for the refinement of the artifact. Seven interviewees and two researchers participated in two sessions. We conducted two focus groups, one with a focus on the application engineer perspective and one with a focus on the developer perspective to control the homogeneity of discussion (Tremblay et al., 2010b). We chose a small sample size because we knew from the interview phase that the participants are experts in their area and have much to contribute. During the session, we showed the participants the prototype and led them through the different options by using a click-through approach. The focus groups were recorded and afterwards transcribed using MaxQDA. Following the data analysis approach from Tremblay et al. (2010a), we used template coding (King, 2004). Our template included the following codes: each DF, new requirements, usability requirements as well as evidence and counterevidence of usefulness. Template analysis is especially useful for hierarchical coding (King, 2004), allowing us to label each DF as evidence/counterevidence of usefulness or requirement. To ensure intercoder reliability, the transcripts were coded by two researchers, and deviations were discussed.

Finally, we summarized the feedback, visualized in Table 4. The overall feedback was very positive. We can conclude that IM-DSSs are desired by the experts and have the potential for tool wear analysis. Additionally, we could generate feedback for the artifact refinement.

Table 4. Evidence and Counterevidence of Usefulness Gathered from Exploratory Focus Groups

Design feature	Evidence of usefulness	Counterevidence of usefulness
DF1: Segmentation map	Supports inexperienced users	None
DF2: Domain features	Reduces variance in feature measurement.	None
DF3: Visual data mining	Interactivity is perceived as useful	None
DF4: Metadata	Necessary to interpret the image data	None
DF5: Descriptive statistics	ATWA facilitates statistical grounding, this is lacking in current tool wear analysis	None
DF6: Heatmap	The heatmap is perceived as useful, especially for generating hypotheses	Blurring due to rotational or zooming errors
DF7: Unsupervised clustering	Helps to notice irregularities of tool wear	None
DF8: Statistical tests	Perceived as useful for customer discussions	Redundancy with tests for profitability
DF9: Correlation matrix	Enables assessment of own hypotheses	Requires huge data sources; comparison of experiments is difficult due to varying metadata

7 Third Design Cycle: Effectiveness, Efficiency, and Usefulness of Proposed Image-Mining-Based DSS

In the third design cycle, we used the feedback from the second design cycle's EE to refine the artifact. Finally, we evaluated our design knowledge with the help of the final artifact in a summative way, conducting four EEs.

7.1 Suggestion and Development

In the following part, we discuss the feedback from DC2 and the implications. We clustered the implications into four categories: usability requirements (requirements of the end users regarding usability of the artifact), exclusions (elements that were excluded from the final artifact), refinements (changes of DFs), and new DFs. In terms of usability, the experts expressed the need for explanations because they were unfamiliar with statistical techniques. We implemented these explanations in the artifact as tool tips, meaning that an explanation is shown when the mouse is positioned over certain elements of the web application. While most descriptive statistics **(DF5)** were easy to understand for the end users, some were unfamiliar and led to confusion, for example the boxplot. We therefore excluded these elements from the artifact. We also excluded **DF9** (Correlation matrix) for this iteration of the artifact. Even though it is perceived as useful and is a necessary step for large-scale exploration, it needs a magnitude of data that is currently not available. Referring to the refinements, the experts articulated the need for additional descriptive statistics **(DF4)**. One expert explained the current practical approach for determining a limit value of tool wear in machining processes: *"[...] then there are 10 inserts and then the best and the worst are deleted. And the worst of the remaining eight sets the limit."* In other words, the experts were interested in quantile information. We developed a more rigorous approach to calculate these and implemented an interactive-value-at-risk-based approach (Pflug, 2000).

Additionally, we derived three new DFs. First, a major theme of the exploratory focus groups was the dispersion in machining processes. To make accurate decisions from tool wear experiments, the dispersion must be taken into account. Therefore, a sufficient sample size needs to be chosen before conducting a tool wear experiment. To add suggestive guidance we added a sample size calculation function **(DF10)**.

Additionally, because the experts emphasized the value of the original images, we implemented image retrieval **(DF11)**. In particular, we implemented shape-based image retrieval (Burl et al., 1999). For example, the domain experts can filter for all images having two chippings or search for outliers with the maximum number of chippings of the dataset.

Lastly, we automatize the service process holistically **(DF12)**. Based on the exploratory study of Section 3, we first analyzed and then generalized the process. For generalization, we searched for key elements in each expert's process description. Recurring key elements were then interpreted as part of the generalized process. Afterward, we conceptualized an adapted process, including automation potentials. Figure 10 depicts the adapted process. It starts with an external customer order (0). Afterward, the customer collects a predefined number of inserts (1). These inserts are shipped to the analyzing center and in parallel metadata of the process is collected (2). Subsequently, the inserts are cleaned (3) and recorded (4). Cleaning and recording need to be done in an automatized way to address **DR6** (Scaling). Currently, we are working on automizing this step with a robot. Afterward, a batch job is triggered that segments the images (5) and calculates features (6). These features are the basis for the DSS. In step (7), the domain experts use the DSS to find important features and build recommendations for the customer. These recommendations are implemented (8) and afterward evaluated with respect to optimization criteria (9). Figure 11 shows the final artifact's graphical user interface. A video of the artifact can be found here: https://youtu.be/OdZZBRXchyE.

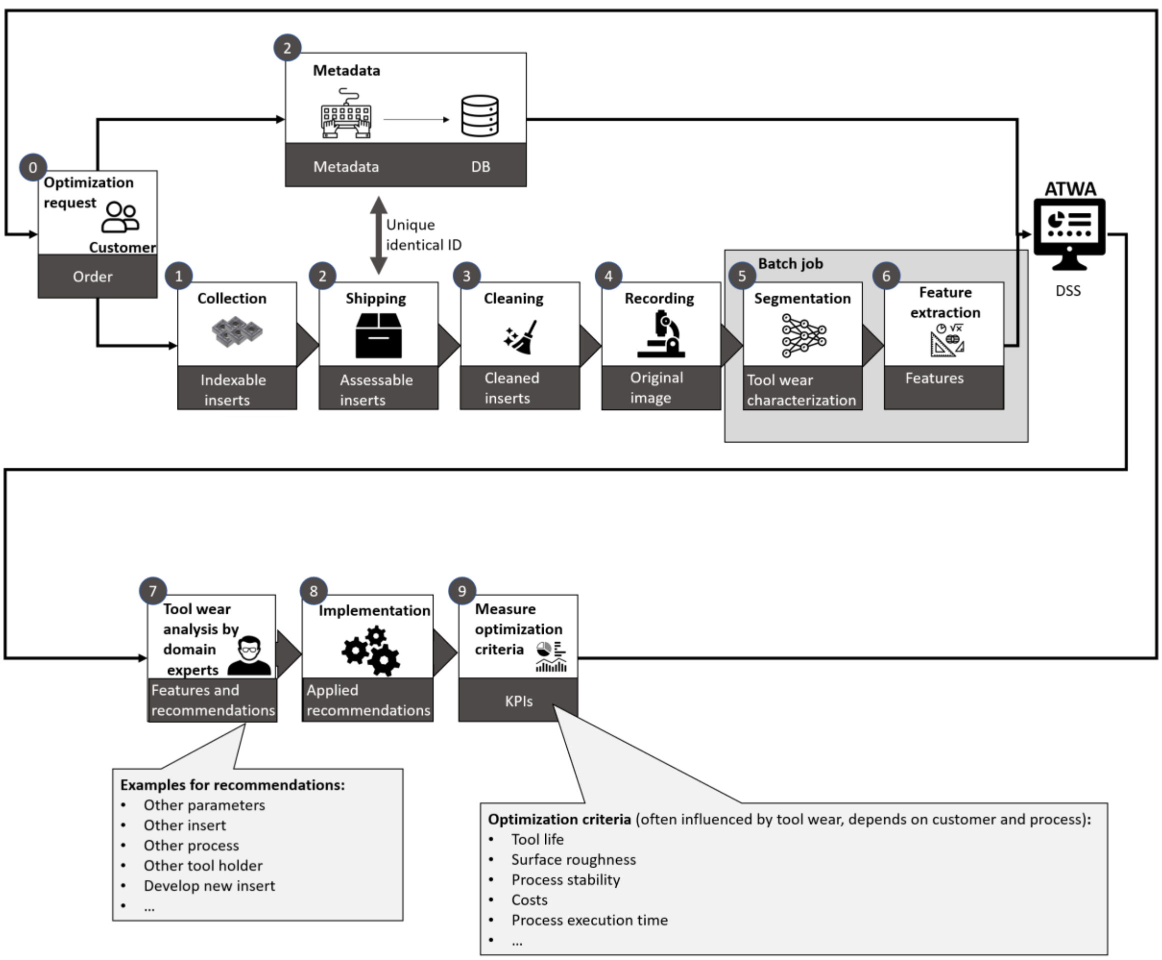

Figure 10. Data Flow and Architecture of ATWA

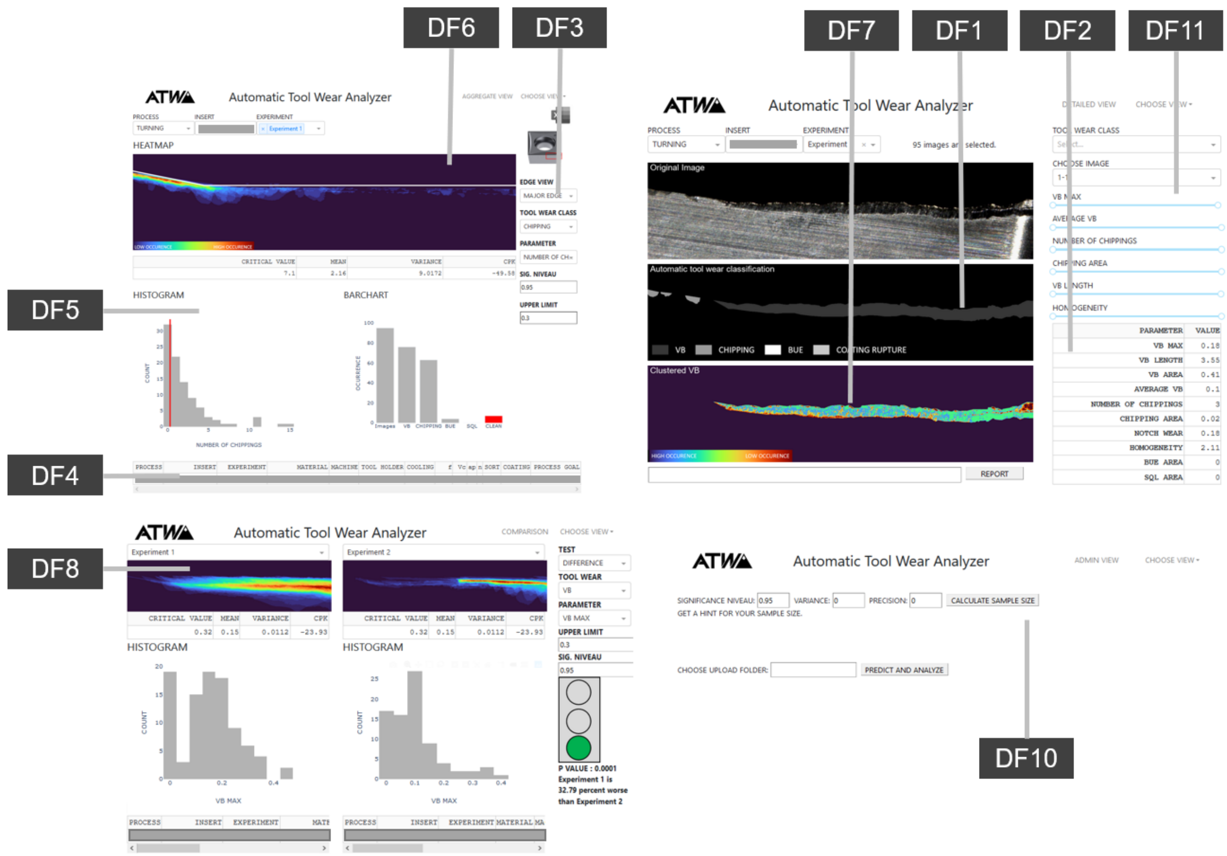

Figure 11. The Interface of ATWA (Top left: Aggregated View; Top right: Detailed View; Bottom left: Comparison View; Bottom Right: Admin View). Note, BUE Stands for Built-Up Edge, VB for Flank Wear

7.2 Evaluation

To evaluate the artifact in a summative way, we conducted four EEs. As Gregor and Jones (2007) recommended, we evaluate the design using testable propositions:

> **Proposition 7.1: The transformation of image data into information is more effective using IM-DSSs than manual information extraction.**

> **Proposition 7.2: The transformation of image data into information is more efficient using IM-DSSs than manual information extraction.**

> **Proposition 7.3: Domain experts perceive the application of IM-DSSs as useful.**

Testing Proposition 7.1: To measure the effectiveness of the feature measurement, we chose a key feature in machining, the maximum height of the flank wear, to compare the human and ATWA error rates. Lutz et al. (2019) use the same approach to evaluate the effectiveness of a tool wear monitoring system. We define the error rate as the mean absolute error (MAE). The manual measurements were conducted by a domain expert familiar with microscopy and tool wear analysis. The ground truth was based on the created labels for Section 5. Conducting ten measurements led to a human MAE of 0.025mm with a standard deviation of 0.021mm and ATWA's MAE of 0.049mm with a standard deviation of 0.026mm. The MAE shows that there is a small difference between human and ATWA's feature measurement. Even though we could not verify Proposition 7.1, we believe that, as explained in Section 5, additional data should further improve the automatic semantic segmentation and increase the effectiveness of the feature measurement. Discussions with domain experts have shown that they already perceive our current results as sufficiently effective.

Testing Proposition 7.2 In the second EE, we measured the efficiency of ATWA. We defined efficiency in tool wear analysis as the savings in human working time. Human work in tool wear analysis is mainly performed during image recording and tool wear measurement. By observing and tracking an expert in tool wear analysis, we found that the recording step takes an average of 24.4 seconds and the measurement step on average 28.2 seconds, leading to a total of 52.6 seconds per image. Under the assumption that the recording is automated as well, the saving for a sample size of 100 would be more than one hour of human working time for a single tool wear experiment. We conclude that ATWA enables significant efficiency enhancements.

Testing Proposition 7.3 In the third EE, we conducted two confirmatory focus groups, with 12 experts and two researchers participating. Similar to the second DC, the focus groups were audio-recorded and transcribed, and the second researcher took observational notes. The transcripts and field notes were combined and afterward coded with template analysis by two researchers independently. The results were discussed and merged afterward.

Table 5 visualizes the results of the confirmatory focus groups. In summary, the results show more evidence than counterevidence of usefulness, pointing towards ATWA being a useful artifact. Furthermore, we conducted a survey based on questions of the well-known TAM (Venkatesh & Davis, 2000). This survey was handed out after each experiment and after each focus group. The items were chosen based on literature (Venkatesh & Bala, 2008; Venkatesh et al., 2003). Each item was measured on a 5-point Likert scale.

Table 5. Evidence and Counterevidence of Usefulness Gathered from Confirmatory Focus Groups

Design feature	Evidence of usefulness	Counterevidence of usefulness
DF1: Segmentation map	Supports inexperienced users and acts as a control function	None
DF2: Domain features	Measurement of new features like areas	None
DF3: Visual data mining	Interactivity is perceived as useful	None
DF4: Metadata	Perceived as the most critical element in tool wear analysis.	None
DF5: Descriptive statistics	Enables statistics for the customer	Raw data needs to be well-prepared to enable useful data mining
DF6: Heatmap	Provides an overview	None
DF7: Unsupervised clustering	Inspires to find new wear patterns	None
DF8: Statistical tests	Increases the validity of image collection comparisons.	Difficult to compare different geometrics; difficult to get the necessary number of images
DF9: Correlation matrix	Excluded	
DF10: Sample size	Provides guidance	None
DF11: Image retrieval	Good to detect outliers	None
DF12: Standardized recording, automated preprocessing, and segmentation	Relief of expenses for the customer	None

We calculated the mean and standard deviation (Std.) of each item. The results are shown in Table 6. A total of 17 experts completed the survey. Overall, with a mean of 4.28, the participants perceived the tool as very useful for tool wear analysis.

Table 6. Results of the Survey Regarding the Perceived Usefulness, Automatic Tool Wear Analyzer (ATWA)

Construct	Mean	Question	Mean per item	Std.
Perceived usefulness	4.28	Using ATWA would improve tool wear analysis.	4.41	0.8
		Using ATWA in my job would increase the productivity of tool wear analysis.	4.18	0.73
		Using ATWA would enhance my effectiveness in tool wear analysis.	4.06	0.83
		I find ATWA to be useful in tool wear analysis.	4.47	0.62

The evaluation shows that the instantiated design knowledge is sufficiently effective, efficient, and perceived as useful by domain experts.

8 Discussion

Having developed and evaluated our design knowledge, we now discuss our contributions and limitations. First, in Section 8.1 we discuss the EEs in relation to the developed design knowledge and summarize the nascent design knowledge. Subsequently, in Section 8.2 we highlight similarities and differences in our design knowledge with regard to the one by Landwehr et al. (2022) for a similar class of systems of image-based DSSs. Afterward, in Section 8.3., we discuss the applicability of our design knowledge to other use cases. In Section 8.4, we reflect on our contributions beyond the design knowledge and lastly, in Section 8.5, we describe the limitations of our study and present opportunities for future work.

8.1 Summarization and Contribution of Our Design Knowledge

The first two EEs (effectiveness and efficiency) illustrate the automation potential of tool wear characterization and address the issue of **DR6** (Reproducibility) and **DR4** (Scalability). Utilizing DL for image processing addresses the cost-effectiveness **(DR7)** of the system (O'Mahony et al., 2019). By providing statistical tests, we address **DR3** (Comparison). The recording's standardization addresses **DR2** (Image quality) and easily accessible interfaces for metadata address **DR1** (Context). Implemented descriptive statistics provide information about the dispersion of the image collection **(DR7)**. Lastly, techniques from the image mining knowledge base target **DR8** (Exploration). As stated in DC1, Figure 6 on page 13 shows the matching of DRs, DPs, and DFs.

Table 7. Nascent Design Knowledge for IM-DSSs

Component	Description
Purpose and scope	Prescriptive knowledge for developing image-mining-based DSSs to improve information and knowledge extraction from collections of images.
Key Constructs	We defined three levels of output-specific constructs (Offermann et al., 2010): segmentation, feature measurement, and decision quality. Technical metrics measure the segmentation quality. Feature measurement is evaluated domain specifically. Lastly, the decision quality is measured by the process outcome quality.
Principles of form and function	Drawing on the body of knowledge, we derived five tentative design principles and evaluated the design in four evaluation episodes through 12 design features.
Justificatory knowledge	We conceptualize our design principles based on the kernel theories of Hsu et al. (2002), Lecun et al. (2015), and Silver (1991).
Testable propositions	We formulated and tested three testable propositions: Proposition 7.1, Proposition 7.2, and Proposition 7.3.

Artifact mutability	We discuss the mutability of image-mining-based DSSs due to advances in image processing techniques, as well as the instantiation of the design.
Principles of implementation	We derived design features as a concrete instantiation of the design principles.
Expository instantiation	We built an artifact, the **automatic tool wear analyzer** (**ATWA**), to support the experts in conducting tool wear experiments and evaluating these.

Furthermore, the confirmatory focus groups (EE3) and the survey (EE4) indicate the general usefulness of our developed nascent design knowledge. We use the framework for the core components of IS design theory from Gregor and Jones (2007) to structure and present our overall developed nascent design knowledge for IM-DSSs. Table 7 summarizes the design knowledge.

We build on top of existing work of image mining and DSSs and synthesize both into a novel design class, image-mining-based DSSs (IM-DSSs). We see IM-DSSs as an extension of the knowledge base of intelligent DSSs (i.e., systems that involve the application of artificial intelligence (AI) (Arnott & Pervan, 2012)). Due to the complexity of image data, novel processing and aggregation techniques need to be developed. Image mining research provides a lot of technical work on segmentation or feature selection (Gillies et al., 2016; Mishra & Silakari, 2012), but no generalized design knowledge. Our design requirements and design principles can guide researchers and practitioners to develop efficient and useful IM-DSSs. We, thereby, increase the general design knowledge available for IM-DSS to a new problem class and thereby contextualize the available design knowledge of intelligent DSSs (Vom Brocke et al., 2020).

8.2 Design Knowledge for Image-Mining-Based DSSs and Image-Based DSSs: Similarities and Differences

As described in Section 4.2, Landwehr et al. (2022) develop design knowledge for image-based DSSs (IB-DSSs) by using a similar methodology. In the following, we discuss the differences between the design principles developed and evaluated in the manuscript at hand and the ones by Landwehr et al. (2022). Generally, we understand IM-DSSs as a specialized subclass of IB-DSSs. In IM-DSSs decisions are made based on image collections instead of based on single images (as in IB-DSSs). Both studies are the first ones to develop and evaluate design knowledge for the respective class of systems.

Overall, the design knowledge for IB-DSSs developed by Landwehr et al. (2022) and the design knowledge for IM-DSSs developed in the manuscript at hand are similar to a certain extent since both deal with DSSs based on image data. For instance, both the design knowledge for IM-DSSs and IB-DSSs contain a design principle addressing the need for metadata in the DSS. At first glance, there are also several differences. But, while formulations are different, upon closer inspection it becomes clear that the design knowledge developed in the two manuscripts is similar. For example, Landwehr et al. (2022), describe an explicit DP5 (Visual Data Exploration). In contrast, we view this as part of DP2 (Image Mining) which results in many DFs enabling visual data exploration. Another example is that Landwehr et al. (2022) describe the additional DP4 (Interpretability), realized by, for example, bounding boxes and segmentation masks. This did not become an explicit DP in our design knowledge for IM-DSSs since we used semantic segmentation as a computer vision technique as described in DP1 (Image processing). The output of a semantic segmentation model, the segmentation mask, assigns a class label to each pixel in a given input image. This automatically provides the level of interpretability Landwehr et al. (2022) achieve by the explicit DP4.

Consequently, the design knowledge developed and evaluated in the two studies reinforces each other. While the two studies were conducted in different industries, for different use cases, and with different domain experts, still a considerable overlap in generated design knowledge emerges. From a methodological perspective, this demonstrates the benefit of running a larger research program with several design science research projects in parallel. It allows the involved researchers to compare and possibly confirm their design knowledge once it is formulated—and, thus, perform first validity and plausibility checks. Of course, this may well offer support to contrast and possibly confirm design knowledge, not to alter it, though.

The major differences are additional design requirements and design features for IM-DSSs that are linked to the design principle of DP2 (Image Mining) of our work. These are due to the different scenarios in which IB-DSSs and IM-DSSs are used for decision support—based on single images and based on image collections, respectively. While having only one additional DP seems like an almost minor difference, this DP has a major impact on the overall design knowledge. Four of the total eleven DFs in our work are highly relevant for IM-DSSs and not applicable for IB-DSSs.

Because in both manuscripts the design knowledge is based on a single use case, we describe the resulting design knowledge as nascent. Hence, the design knowledge for both IB-DSS and IM-DSS needs to be confirmed and possibly refined by future studies.

8.3 Applicability of Our Design Knowledge to Other Use Cases

We are convinced the design knowledge developed and evaluated in this manuscript is directly applicable to many other use cases. For example, we can think of more manufacturing use cases like an IM-DSS used for quality control based on images that are taken in the production line. Also, for example, we think the design knowledge can directly be applied to an IM-DSS in the sports domain in which images of worn running shoes are analyzed with the goal of developing more durable running shoes.

Additionally, we see many use cases where our design knowledge is applicable but can be slightly extended. For example, after DC2 we excluded the DF9. Correlation matrix due to a lack of sufficient data in our use case. In use cases with more data available, this DF will probably be relevant. Also, we see other domain-specific extensions. In IM-DSSs for medical applications (compare Sollini et al. (2019)) data privacy will most likely be an additional DP. Lastly, there are use cases where the dimension's space and time are relevant and hence lead to additional design knowledge affecting the data processing and visualization. For example, in an IM-DSS for monitoring wind turbine blade degradation and subsequent maintenance planning and product improvements both time and space are relevant. In conclusion, we believe that our design knowledge is directly applicable in many use cases and serves as a useful starting point for slight extensions in many other use cases.

8.4 Contributions Beyond Design Knowledge

Beyond our contribution of nascent design knowledge for IM-DSSs, we developed an artifact to facilitate human-machine collaboration and evaluated it in practice. The goal of human-machine collaboration is to leverage the advances of AI and human intelligence to enable synergy effects, for example, free employees' time for higher-level tasks (Wilson & Daugherty, 2018). This approach also tackles one of the inherent and often discussed challenges of data mining, namely the difference between correlation and causation (Smith, 2020). While image processing and the consequent image mining might be not enough to extract valuable knowledge from images in an autonomous way, incorporating the functions into DSSs allows humans to intervene and add a notion of causality.

The AI part of our artifact is the semantic segmentation of the images through DL which enables an efficient and effective transformation of image data into information. The DSS interface provides access to human intelligence and enables human-AI synergy. With our work, we could show the usefulness of an artifact using human-machine collaboration.

8.5 Limitations and Future Work

Besides the aforementioned contributions, our research also has limitations. A first limitation of our study is that our design knowledge is developed and evaluated based on a single use case. As discussed in Section 8.2 a major part of the design knowledge is confirmed by another manuscript in our research program on DSSs based on images (Landwehr et al., 2022). Still, future studies should evaluate the design knowledge for IM-DSSs developed and evaluated in the manuscript at hand in other application areas and evaluate it accordingly. Once more such studies exist, we believe a meta-study consolidating design knowledge for IM-DSSs (similar to Zschech et al. (2021) for computer vision-based hybrid intelligence systems) would be particularly helpful.

Another potential limitation that is worth discussing is the effectiveness of ATWA. As described in Section 7.2, the effectiveness of ATWA could not be shown. Here, we use a very strict understanding of effectiveness. We test the DL model for feature measurement for better-than-human performance. In fact, there are many other applications where the performance of a ML/DL model is worse than human performance but still the application is useful (Serre, 2019), in particular when properly combined with human intelligence (Vössing et al., 2022) as in our IM-DSS. This is also the case for the work at hand–as described in Section 7.2 domain experts perceive the DL model as sufficiently effective. Consequently, the design knowledge developed and evaluated in this manuscript is not affected by this effectiveness analysis. Additionally, as described in Section 5.2, we trained our DL model on a relatively small dataset. The performance of DL models grows logarithmically with the size of the training dataset (C. Sun et al., 2017). Hence, we are confident that it is possible to build DL models that surpass human performance also for this use case.

Another limitation regarding the evaluation is that the technology acceptance model aims to measure potential users' intended usage behavior. A further study should assess the artifact's long-term effects.

9 Conclusion

The purpose of this study is to develop design knowledge for image-mining-based decision support systems. We initiated the design science research project by conducting an exploratory study (relevance cycle). Subsequently, we analyzed the existing body of knowledge of image mining, deep learning, and decision support systems to inform our research (rigor cycle). We then conducted a first design cycle to derive design knowledge for image processing. In the second design cycle, we suggested initial design principles for image-mining-based decision support systems. These were mapped into specific design features that were implemented in a prototype and qualitatively evaluated using exploratory focus groups. The focus groups indicated the general desirability of the artifact and consequently of the design knowledge. In the third design cycle, we used the results of the second design cycle as input and refined our design knowledge and the artifact. Following that, we evaluated our design knowledge with the help of the developed artifact and conducted four summative evaluation episodes that indicated sufficient effectiveness, efficiency, and usefulness of our nascent design knowledge.

Our research contributes to theory and practice. Regarding practical contributions, we translated the design principles in specific design features and instantiated them in an artifact, the automatic tool wear analyzer. This instantiation solves a real-world problem at our case company by removing manual work and supporting the knowledge generation process. Our evaluation episodes confirm the usefulness of the artifact for the domain experts.

Regarding theoretical contributions, we shed first light on a problem class that we defined as image-mining-based decision support systems. We developed preliminary design requirements and design principles that could guide the future development of such systems.

We see potential for the application of our design knowledge in several other domains, such as medicine, sports, or biology. The domains should use the advances in image processing to extract previously inaccessible knowledge from large image collections to create competitive advantages.

We invite researchers and practitioners to instantiate, evaluate, and extend the proposed nascent design knowledge for image-mining-based decision support systems.

Acknowledgements

We would like to thank Ceratizit Austria GmbH, in particular Jonathan Schäfer and Adrian Weber, for facilitating this research.

References

Abbasi, A., & Chen, H. (2008). CyberGate: A design framework and system for text analysis of computer-mediated communication. *MIS Quarterly*, *32*(4), 811–837.

Ackoff, R. L. (1989). From data to wisdom. *Journal of Applied Systems Analysis*, *16*(1), 3–9.

Afshar, P., Mohammadi, A., Plataniotis, K. N., Oikonomou, A., & Benali, H. (2019). From handcrafted to deep-learning-based cancer radiomics: Challenges and opportunities. *IEEE Signal Processing Magazine*, *36*(4), 132–160.

Alberti, M., Ciurana, J., Rodriguez, C. A., & Özel, T. (2011). Design of a decision support system for machine tool selection based on machine characteristics and performance tests. *Journal of Intelligent Manufacturing*, *22*(2), 263–277.

Arnott, D., & Pervan, G. (2005). A critical analysis of decision support systems research. *Journal of Information Technology*, *20*(2), 67–87.

Arnott, D., & Pervan, G. (2012). Design science in decision support systems research: An assessment using the Hevner, March, Park, and Ram guidelines. *Journal of the Association for Information Systems*, *13*(11), 1.

Arrazola, P. J., Özel, T., Umbrello, D., Davies, M., & Jawahir, I. S. (2013). Recent advances in modelling of metal machining processes. *CIRP Annals*, *62*(2), 695–718.

Athanasopoulos, G., Riba, C. R., & Athanasopoulou, C. (2009). A decision support system for coating selection based on fuzzy logic and multi-criteria decision making. *Expert Systems with Applications*, *36*(8), 10848–10853.

Bannach, A., Bernard, J., Jung, F., Kohlhammer, J., May, T., Scheckenbach, K., & Wesarg, S. (2017). Visual analytics for radiomics: Combining medical imaging with patient data for clinical research. *2017 IEEE Workshop on Visual Analytics in Healthcare* (VAHC) (pp. 84–91). IEEE.

Barnathan, M., Zhang, J., & Megalooikonomou, V. (2008). A web-accessible framework for the automated storage and texture analysis of biomedical images. *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, Proceedings, ISBI* (pp. 257–259).

Berlage, T. (2007). Analyzing and mining automated imaging experiments. *Expert Opinion on Drug Discovery*, *2*(4), 561–569.

Bhatt, C. A., & Kankanhalli, M. S. (2011). Multimedia data mining: State of the art and challenges. *Multimedia Tools and Applications*, *51*(1), 35–76.

Black, J. T. (1989). Introduction to machining processes. In *ASM handbook, machining, vol. 16*. *ASM International, 1*.

Brendel, A. B., Lembcke, T.-B., Muntermann, J., & Kolbe, L. M. (2021). Toward replication study types for design science research. *Journal of Information Technology*, *36*(3), 198-215.

Burl, M. C., Fowlkes, C., & Roden, J. (1999). Mining for image content. *Systemics, Cybernetics, and Informatics/Information Systems: Analysis and Synthesis*. https://ics.uci.edu/~fowlkes/papers/SCI-ISAS99.pdf

Castejón, M., Alegre, E., Barreiro, J., & Hernández, L. K. (2007). On-line tool wear monitoring using geometric descriptors from digital images. *International Journal of Machine Tools and Manufacture*, *47*(12–13), 1847–1853.

Cheng, Z., Nakatsugawa, M., Zhou, X., Hu, C., Greco, S., Kiess, A.P., Page, B.R., Alcorn, S.R., Haller, J.W., Utsunomiya, K., Sugiyama, S., Fu, W., Wong, J., Lee, J., McNutt, T., & Quon, H. (2019). Utility of a clinical decision support system in weight loss prediction after head and neck cancer radiotherapy. *JCO Clinical Cancer Informatics*, (3), 1–11.

Chollet, F. (2015). *Keras*. https://blog.keras.io/author/francois-chollet.html

Coenen, F., & Dittakan, K. (2016). Image representation for image mining: A study focusing on mining satellite images for census data collection. *International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management* (pp. 3–27).

Cooper, H. M. (1988). Organizing knowledge syntheses: A taxonomy of literature reviews. *Knowledge in Society*, *1*(1), 104–126.

Cortes, C., Mohri, M., & Rostamizadeh, A. (2012). L2 regularization for learning kernels. *ArXiv Preprint ArXiv:1205.2653*.

Coyne, I. T. (1997). Sampling in qualitative research. Purposeful and theoretical sampling; merging or clear boundaries? *Journal of Advanced Nursing*, *26*(3), 623–630.

Dasch, J., D'Arcy, J., Gundrum, A., Sutherland, J., Johnson, J., & Carlson, D. (2005). Characterization of fine particles from machining in automotive plants. *Journal of Occupational and Environmental Hygiene*, *2*(12), 609–625.

Dey, N., Karâa, W. B. A., Chakraborty, S., Banerjee, S., Salem, M. A. M., & Azar, A. T. (2015). Image mining framework and techniques: A review. *International Journal of Image Mining*, *1*(1), 45–64.

Fisher, M., & Mackiewicz, M. (2020). *Image labelling tool. Bitbucket*. GitHub. https://bitbucket.org/ueacomputervision/image-labelling-tool

Foran, D. J., Yang, L., Chen, W., Hu, J., Goodell, L. A., Reiss, M., Wang, F., Kurc, T., Pan, T., Sharma, A., & Saltz, J. H. (2011). ImageMiner: A software system for comparative analysis of tissue microarrays using content-based image retrieval, high-performance computing, and grid technology. *Journal of the American Medical Informatics Association*, *18*(4), 403–415.

Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., & Garcia-Rodriguez, J. (2017). A review on deep learning techniques applied to semantic segmentation. *ArXiv Preprint ArXiv:1704.06857*.

Gatta, R., Vallati, M., Dinapoli, N., Masciocchi, C., Lenkowicz, J., Cusumano, D., Casá, C., Farchione, A., Damiani, A., van Soest, J., Dekker, A., & Valentini, V. (2019). Towards a modular decision support system for radiomics: A case study on rectal cancer. *Artificial Intelligence in Medicine*, *96*, 145–153.

Gillies, R. J., Kinahan, P. E., & Hricak, H. (2016). Radiomics: Images are more than pictures, they are data. *Radiology*, *278*(2), 563–577.

Gregor, S., & Jones, D. (2007). The anatomy of a design theory. *Journal of the Association for Information Systems*, *8*(5), 312–335.

Grove, O., Berglund, A. E., Schabath, M. B., Aerts, H. J., Dekker, A., Wang, H., Velazquez, E. R., Lambin, P., Gu, Y., Balagurunathan, Y., Eikman, E., Gatenby, R. A., Eschrich, S., & Gillies, R. J. (2015). Quantitative computed tomographic descriptors associate tumor shape complexity and intratumor heterogeneity with prognosis in lung adenocarcinoma. *PloS One*, *10*(3).

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques*. Elsevier.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1026–1034).

Hevner, A. R. (2007). A three cycle view of design science research. *Scandinavian Journal of Information Systems*, *19*(2), 4.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, *28*(1), 75–105.

Hsu, W., Lee, M. L., & Zhang, J. (2002). Image mining: Trends and developments. *Journal of Intelligent Information Systems*, *19*(1), 7–23.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *ArXiv Preprint ArXiv:1502.03167*.

ISO. (1993). *ISO 3685:1993. Tool-life testing with single-point turning tools, 2nd*. International Standard.

Jourdan, N., Biegel, T., Knauthe, V., von Buelow, M., Guthe, S., & Metternich, J. (2021). A computer vision system for saw blade condition monitoring. *Procedia CIRP*, *104*, 1107-1112.

Keim, D. A. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, *8*(1), 1–8.

Khodaskar, A. A., & Ladhake, S. A. (2014). Image mining: An overview of current research. *Proceedings - 2014 4th International Conference on Communication Systems and Network Technologies, CSNT 2014*, (pp. 433–438).

King, N. (2004). Using templates in the thematic analysis of text. In *Essential guide to qualitative methods in organizational research* (pp. 257 - 270). Sage.

Koh, B., & Cui, F. (2022). An exploration of the relation between the visual attributes of thumbnails and the view-through of videos: The case of branded video content. *Decision Support Systems*, *160*, 113820.

Kohli, R., & Devaraj, S. (2004). Contribution of institutional DSS to organizational performance: Evidence from a longitudinal study. *Decision Support Systems*, *37*(1), 103–118.

Kreiss, K., Mroz, M. M., Newman, L. S., Martyny, J., & Zhen, B. (1996). Machining risk of beryllium disease and sensitization with median exposures below 2 μg/m 3. *American Journal of Industrial Medicine*, *30*(1), 16–25.

Kuckartz, U. (2012). Qualitative inhaltsanalyse. methoden, praxis, computerunterstützung. Weinheim und Basel: Beltz Juventa, 188 S. [Rezension]. *Journal for Educational Research Online, 6*, 141-145.

Kuechler, B., & Vaishnavi, V. (2008). On theory development in design science research: Anatomy of a research project. *European Journal of Information Systems*, *17*(5), 489–504.

Kumar, V., Gu, Y., Basu, S., Berglund, A., Eschrich, S. A., Schabath, M. B., Forster, K., Aerts, H. J., Dekker, A., Fenstermacher, D., Goldgof, D. B., Hall, L. O., Lambin, P., Balagurunathan, Y., Gatenby, R. A., & Gillies, R. J. (2012). Radiomics: The process and the challenges. *Magnetic Resonance Imaging*, *30*(9), 1234–1248.

Kurada, S., & Bradley, C. (1997). A review of machine vision sensors for tool condition monitoring. *Computers in Industry*, *34*(1), 55–72.

Kuram, E., & Ozcelik, B. (2014). Modern mechanical engineering. In J. Oaulo Davim (Ed.), *Modern Mechanical Engineering Research, Development and Education.* Springer.

Lambin, P., Leijenaar, R.T.H., Deist, T.M., Peerlings, J., de Jong, E.E.C., van Timmeren, J., Sanduleanu, S., Larue, R.T.H.M., Even, A.J.G., Jochems, A., van Wijk, Y., Woodruff, H., van Soest, J., Lustberg, T., Roelofs, E., van Elmpt, W., Dekker, A., Mottaghy, F.M., Wildberger, J.E., & Walsh, S. (2017). Radiomics: The bridge between medical imaging and personalized medicine. *Nature Reviews Clinical Oncology*, *14*(12), 749–762.

Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., van Stiphout, R.G., Granton, P., Zegers, C. M., Gillies, R., Boellard, R., Dekker, A., Aerts, H. J. (2012). Radiomics: Extracting more information from medical images using advanced feature analysis. *European Journal of Cancer*, *48*(4), 441–446.

Landwehr, J. P., Kühl, N., Walk, J., & Gnädig, M. (2022). Design knowledge for deep-learning-enabled image-based decision support systems: Evidence from power line maintenance decision-making. *Business and Information Systems Engineering*, *64*(6), 707–728.

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lee, A. S. (2010). Retrospect and prospect: Information systems research in the last and next 25 years. *Journal of Information Technology*, *25*(4), 336–348.

Liu, M., Zhang, J., Nie, D., Yap, P.-T., & Shen, D. (2018). Anatomical landmark based deep feature representation for MR images in brain disease diagnosis. *IEEE Journal of Biomedical and Health Informatics*, *22*(5), 1476–1485.

Louie, A. V., Rodrigues, G., Olsthoorn, J., Palma, D., Yu, E., Yaremko, B., Ahmad, B., Aivas, I., & Gaede, S. Inter-observer and intra-observer reliability for lung cancer target volume delineation in the 4D-CT era. *Radiotherapy and Oncology*, *95*(2), 166–171.

Lukić, L., Polajnar, D., & Šolaja, V. (1991). A Yugoslav approach to decision support for optimization of FMS technologies. *CIRP Annals*, *40*(1), 99–102.

Lutz, B., Kisskalt, D., Regulin, D., Reisch, R., Schiffler, A., & Franke, J. (2019). Evaluation of deep learning for semantic image segmentation in tool condition monitoring. *Proceedings - 18th IEEE International Conference on Machine Learning and Applications, ICMLA 2019*.

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems*, *15*, 251–266.

Miah, S. J., Gammack, J. G., & McKay, J. (2019). A metadesign theory for tailorable decision support. *Journal of the Association for Information Systems*, *20*(5), 570–603.

Mishra, N., & Silakari, S. (2012). Image mining in the context of content based image retrieval: A perspective. *International Journal of Computer Science Issues (IJCSI)*, *9*(4), 69.

Müller, O., Junglas, I., Brocke, J. vom, & Debortoli, S. (2016). Utilizing big data analytics for information systems research: Challenges, promises and guidelines. *European Journal of Information Systems*, *25*(4), 289–302.

Nabhani, F. (2001). Machining of aerospace titanium alloys. *Robotics and Computer-Integrated Manufacturing*, *17*(1–2), 99–106.

Nabil, B. F., & Mabrouk, M. (2006). Effects of random aspects of cutting tool wear on surface roughness and tool life. *Journal of Materials Engineering and Performance*, *15*(5), 519–524.

Offermann, P., Blom, S., Levina, O., & Bub, U. (2010). Proposal for components of method design theories. *Business & Information Systems Engineering*, *2*(5), 295–304.

Ofori, M., El-Gayar, O., O'Brien, A., & Noteboom, C. (2022). A deep learning model compression and ensemble approach for weed detection.

O'Mahony, N., Campbell, S., Carvalho, A., Harapanahalli, S., Hernandez, G. V., Krpalkova, L., Riordan, D., & Walsh, J. (2019). Deep learning vs. traditional computer vision. *Science and Information Conference*, 128–144.

Palinkas, L. A., Horwitz, S. M., Green, C. A., Wisdom, J. P., Duan, N., & Hoagwood, K. (2015). Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Administration and Policy in Mental Health and Mental Health Services Research*, *42*(5), 533–544.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359.

Patel, N., & Sethi, I. (2007). Multimedia data mining: An overview. In *Multimedia Data Mining and Knowledge Discovery* (pp. 14–41). Springer.

Pearn, W. L., & Lin, P. C. (2004). Testing process performance based on capability index Cpk with critical values. *Computers \& Industrial Engineering*, *47*(4), 351–369.

Peffers, K., Rothenberger, M., Tuunanen, T., & Vaezi, R. (2012). Design science research evaluation. *Proceedings of the 7th International Conference on Design Science Research in Information Systems: Advances in Theory and Practice* (pp. 398–410).

Peffers, K., Tuunanen, T., Rothenberger, M., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77.

Pflug, G. C. (2000). Some remarks on the value-at-risk and the conditional value-at-risk. In *Probabilistic Constrained Optimization* (pp. 272–281). Springer.

Power, D. J. (2008). Understanding data-driven decision support systems. *Information Systems Management*, *25*(2), 149–154.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, *9351*, 234–241.

Serre, T. (2019). Deep learning: The good, the bad, and the ugly. *Annual Review of Vision Science*, *5*, 399-426.

Sharda, R., Barr, S. H., & McDonnell, J. C. (1988). Decision support system effectiveness: A review and an empirical test. *Management Science*, *34*(2), 139–159.

Shihavuddin, A., Chen, X., Fedorov, V., Christensen, A. N., Riis, N. A., Branner, K., Dahl, A. B., & Paulsen, R. R. (2019). Wind turbine surface damage detection by deep learning aided drone inspection analysis. *Energies*, *12*(4), 676.

Shin, D., He, S., Lee, G. M., Whinston, A. B., Cetintas, S., & Lee, K. C. (2020). *Enhancing social media analysis with visual data analytics: A deep learning approach* (pp. 1459-1492). SSRN.

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. *Proceedings 1996 IEEE Symposium on Visual Languages* (pp. 336–343). IEEE.

Siddhpura, A., & Paurobally, R. (2013). A review of flank wear prediction methods for tool condition monitoring in a turning process. *The International Journal of Advanced Manufacturing Technology*, *65*(1–4), 371–393.

Silver, M. S. (1991). Decisional guidance for computer-based decision support. *MIS Quarterly*, *15*(1), 105–122.

Silver, M. S. (2006). Decisional guidance: Broadening the scope. In *Human-Computer Interaction and Management Information Systems: Foundations* (pp. 104–133).

Smith, G. (2020). Data mining fool's gold. *Journal of Information Technology*, *35*(3), 182–194.

Sollini, M., Antunovic, L., Chiti, A., & Kirienko, M. (2019). Towards clinical application of image mining: A systematic review on artificial intelligence and radiomics. *European Journal of Nuclear Medicine and Molecular Imaging* (pp. 1–17).

Spangler, W. E., May, J. H., & Vargas, L. G. (1999). Choosing data-mining methods for multiple classification: Representational and performance measurement implications for decision support. *Journal of Management Information Systems*, *16*(1), 37–62.

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. *Proceedings of the IEEE International Conference on Computer Vision* (pp. 843–852).

Sun, W., Taniar, D., & Torabi, T. (2008). Image mining: A case for clustering shoe prints. *International Journal of Information Technology and Web Engineering*, *3*(1), 70–84.

Thomas, D. R. (2006). A general inductive approach for analyzing qualitative evaluation data. *American Journal of Evaluation*, *27*(2), 237–246.

Tremblay, M. C., Hevner, A. R., & Berndt, D. J. (2010a). Focus groups for artifact refinement and evaluation in design research. *Communications of the Association for Information Systems*, *26*(27), 599–618.

Tremblay, M. C., Hevner, A. R., & Berndt, D. J. (2010b). The use of focus groups in design science research. In *Design Research in Information Systems* (pp. 121–143). Springer.

Trinks, S., & Felden, C. (2019). Image mining for real time fault detection within the smart factory. *2019 IEEE 21st Conference on Business Informatics (CBI)* (pp. 584–593).

Turban, E., Sharda, R. E., & Delen, D. (2011). *Decision support and business intelligence systems*. Pearson Prentice Hall.

Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: A framework for evaluation in design science research. *European Journal of Information Systems*, *25*(1), 77–89.

Venkatesh, V., & Bala, H. (2008). Technology acceptance model 3 and a research agenda on interventions. *Decision Sciences*, *39*(2), 273–315.

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science*, *46*(2), 186–204.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, *27*(3), 425–478.

Villarroel Ordenes, F., & Zhang, S. (2019). From words to pixels: Text and image mining methods for service research. *Journal of Service Management*, *30*(5), 593–620.

Vom Brocke, J., Simons, A., Niehaves, B., Riemer, K., Plattfaut, R., & Cleven, A. (2009). Reconstructing the giant: On the importance of rigour in documenting the literature search process. *Proceedings of the 17th European Conference on Information Systems (ECIS 2009)* (pp. 2206–2217).

Vom Brocke, J., Winter, R., Hevner, A., & Maedche, A. (2020). Special issue editorial–accumulation and evolution of design knowledge in design science research: A journey through time and space. *Journal of the Association for Information Systems*, *21*(3), 9.

Vössing, M., Kühl, N., Lind, M., & Satzger, G. (2022). Designing transparency for effective human-AI collaboration. *Information Systems Frontiers*, *24*(3), 877-895.

Wang, J., Ma, Y., Zhang, L., Gao, R. X., & Wu, D. (2018). Deep learning for smart manufacturing: Methods and applications. *Journal of Manufacturing Systems*, *48*, 144–156.

Wang, K., Zhang, D., Li, Y., Zhang, R., & Lin, L. (2016). Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, *27*(12), 2591–2600.

Whiting, L. S. (2008). Semi-structured interviews: Guidance for novice researchers. *Nursing Standard*, *22*(23), 35–41.

Wilson, H. J., & Daugherty, P. R. (2018). Collaborative intelligence: Humans and AI are joining forces. *Harvard Business Review*, (August), 114–123.

Winter, R. (2008). Design science research in Europe. *European Journal of Information Systems*, *17*(5), 470–475.

Zaiane, O. R., Han, J., Li, Z.-N., Chee, S. H., & Chiang, J. Y. (1998). MultiMediaMiner: A system prototype for multimedia data mining. *ACM SIGMOD Record*, *27*(2), 581–583.

Zschech, P., Walk, J., Heinrich, K., Vössing, M., & Kühl, N. (2021). A Picture is Worth a Collaboration: Accumulating Design Knowledge for Computer-Vision-based Hybrid Intelligence Systems. *Proceedings of the 29th European Conference on Information Systems*.

About the Authors

Jannis Walk is Data Engineer/Scientist at cellcentric GmbH & Co. KG. Previously, he obtained his Ph.D. while working in the Applied AI in Services Lab at the Karlsruhe Service Research Institute (KSRI) at KIT. He has more than seven years of experience in research dealing with the practical application of machine learning and decision support systems in cooperation with several companies. He has published his work in leading journals (Business and Information Systems Engineering and Journal of Cleaner Production) and conferences (ECML and ECIS etc.).

Max Schemmer is a research associate at the Karlsruhe Institute of Technology (KIT). He conducts research in the area of applied AI with a focus on human-AI complementarity and explainable AI. He has published in international journals such as the European Conference on Information Systems (ECIS) and the ACM CHI Conference on Human Factors in Computing Systems (CHI).

Niklas Kühl received a Ph.D. in information systems with a focus on applied machine learning and his Habilitation in applied computer science. He is a Full Professor of information systems and human-centric AI at the University of Bayreuth. He is also a Group Lead with Fraunhofer FIT for business analytics and a Senior Expert in artificial intelligence with IBM. In the past, he was a Managing Consultant for Data Science with IBM, which complemented his theoretical knowledge with practical insights from the field. In his studies, he is working on conceptualizing, designing, and implementing artificial intelligence (AI) artifacts, focusing on inter-organizational learning and fair and effective collaboration within human–AI teams. He collaborates internationally with institutions such as the University of Texas at Austin, Carnegie Mellon University, and the MIT–IBM Watson AI Lab.

Gerhard Satzger heads the research group “Digital Service Innovation” at the Karlsruhe Institute of Technology (KIT) and is co-founder of the Karlsruhe Digital Service Research and Innovation Hub (KSRI), an “industry-on-campus” initiative driving practical impact of academic research with a variety of industry partners. His research interests include the use of data and artificial intelligence in information systems to develop and transform service business models. He draws on multi-year industry experience in various national and international roles within IBM, among others as head of a global consulting team focusing on analytics-based transformation and as CFO of IBM’s technology services business in Central Europe. He holds a postdoctoral lecturer qualification (“Habilitation”) from the University of Augsburg, a PhD from the University of Giessen, an MBA from Oregon State University, and a diploma in Industrial Engineering and Management from KIT.