Image-based Decision Support Systems: Technical Concepts, Design Knowledge, and Applications for Sustainability

Zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften
(Dr.-Ing.)

von der KIT-Fakultät für Wirtschaftswissenschaften
des Karlsruher Instituts für Technologie (KIT)

genehmigte

Dissertation

von

M.Sc. Jannis Walk

Tag der mündlichen Prüfung:	11.04.2023
Referent:	Prof. Dr. Gerhard Satzger
Korreferent:	Vertretungsprof. Dr. Michael Färber

Abstract

Unstructured data accounts for 80-90% of all data generated, with image data contributing its largest portion. In recent years, the field of computer vision, fueled by deep learning techniques, has made significant advances in exploiting this data to generate value. However, often computer vision models are not sufficient for value creation. In these cases, image-based decision support systems (IB-DSSs), i.e., decision support systems that rely on images and computer vision, can be used to create value by combining human and artificial intelligence. Despite its potential, there is only little work on IB-DSSs so far.

In this thesis, we develop technical foundations and design knowledge for IB-DSSs and demonstrate the possible positive effect of IB-DSSs on environmental sustainability. The theoretical contributions of this work are based on and evaluated in a series of artifacts in practical use cases: First, we use technical experiments to demonstrate the feasibility of innovative approaches to exploit images for IB-DSSs. We show the feasibility of deep-learning-based computer vision and identify future research opportunities based on one of our practical use cases. Building on this, we develop and evaluate a novel approach for combining human and artificial intelligence for value creation from image data. Second, we develop design knowledge that can serve as a blueprint for future IB-DSSs. We perform two design science research studies to formulate generalizable principles for purposeful design — one for IB-DSSs and one for the subclass of image-mining-based decision support systems (IM-DSSs). While IB-DSSs can provide decision support based on single images, IM-DSSs are suitable when large amounts of image data are available and required for decision-making. Third, we demonstrate the viability of applying IB-DSSs to enhance environmental sustainability by performing life cycle assessments for two practical use cases — one in which the IB-DSS enables a prolonged product lifetime and one in which the IB-DSS facilitates an improvement of manufacturing processes.

We hope this thesis will contribute to expand the use and effectiveness of image-based decision support systems in practice and will provide directions for future research.

Acknowledgement

During my work on this dissertation thesis, I could count on the support of many great people around me. First, I would like to thank Prof. Dr. Gerhard Satzger for creating such a suitable setup for me and for supervising me over the years both professionally and personally. Moreover, I would like to thank Dr. Michael Färber for reviewing my dissertation thesis. Additionally, I would like to thank Prof. Dr. Frank Schultmann and Prof. Dr. Andreas Geyer-Schulz for serving on the dissertation committee.

Inspiring colleagues around me made my PhD time a productive and fun one. In this regard, I would like to particularly thank Prof. Dr. Niklas Kühl for his support of my academic work as well as my professional and personal development in his role as my post-doctoral supervisor. I am especially grateful for his ability to frame almost every challenge as an opportunity and for giving me a great deal of freedom with regard to both the focus and organization of my work. Also, I am very thankful to my colleagues in the Digital Service Innovation group — discussing ideas with you, working on research together, and spending fun times apart from work was great! Regarding this final document, I especially want to thank Jakob & Max for reviewing and proofreading. Additionally, I am grateful for the many motivated and inspiring students I could work with: Florian W., Alexander, Max, Sebastian, Florian G., Julius, Ulrike, Jasmin, Maximilian, and Magnus. Lastly, I would like to thank the researchers from other institutions I got to collaborate with: Patrick Zschech, Kai Heinrich, and Michael Saidani - working with you has been fun and a great learning experience.

In addition to my colleagues at the university, I had the pleasure to work with great colleagues in two partner firms: Ceratizit Austria GmbH and Plansee SE. Thank you for enabling this project, introducing me to your world, and supporting me!

Besides my colleagues, I want to thank my family very much. Thanks to my parents Gudrun & Karl and my brother Mario for supporting me in my work and all other aspects of life. Above all, I would like to thank my wife Tanja: for your moral support, for the good time we had during my work on my dissertation, and for finding the balance between motivating me to keep working and reminding me of other important things in life.

Contents

List of Abbreviations

4R	Re-design, Remanufacturing, Reuse, And Recycling
AI	Artificial Intelligence
ANN	Artificial Neural Network
AP	Average Precision
ARC	Architectural Floor Plans
ATWA	Automatic Tool Wear Analyzer
AUROC	Area Under The Receiver Operating Characteristic Curve
BI	Business Intelligence
BNN	Bayesian Neural Network
BUE	Built-up Edge
CAR	Car Configuration
CFG	Confirmatory Focus Group
CNN	Convolutional Neural Network
CUT	Cutting Tools
CV	Computer Vision
CV-HIS	Computer-vision-based Hybrid Intelligence System
DC	Design Cycle
DF	Design Feature
DIKW	Data, Information, Knowledge, And Wisdom
DL	Deep Learning
DNN	Deep Neural Network
DP	Design Principle
DR	Design Requirement
DSC	Dice Similarity Coefficient
DSR	Design Science Research
DSS	Decision Support System
DW	Data Warehousing
EE	Evaluation Episode
EFG	Exploratory Focus Group
EIS	Executive Information System
FN	False Negative
FP	False Positive

FPR	False Positive Rate
FU	Functional Unit
Grad-CAM	Gradient-weighted Class Activation Mapping
GSS	Group Support System
HI	Hybrid Intelligence
IB-DSS	Image-based Decision Support System
IDSS	Intelligent Decision Support System
IM-DSS	Image-mining-based Decision Support System
IoU	Intersection Over Union
IS	Information Systems
KMDSS	Knowledge Management-based Decision Support System
LCA	Life Cycle Assessment
LCI	Life Cycle Inventory
LCIA	Life Cycle Impact Assessment
LRP	Layer-wise Relevance Propagation
MAE	Mean Absolute Error
mAP	Mean Average Precision
MC-dropout	Monte Carlo Dropout
MCC	Matthews Correlation Coefficient
ME	Maintenance Engineer
ML	Machine Learning
MR	Meta-requirement
MSCD	Multistage Component Detection Pipeline
NRG	Energy Infrastructure
NSS	Negotiation Support Systems
PDSS	Personal Decision Support System
PLC	Power Line Component
PLM	Power Line Maintenance
PSS	Product-service System
ReLU	Rectified Linear Unit
ROC	Receiver Operating Characteristic
ROI	Region Of Interest
RQ	Research Question
SLR	Structured Literature Review
SOL	Solar Panels
SSCD	Single-stage Component Detection Pipeline
TAM	Technology Acceptance Model
TeP	Testable Proposition
TN	True Negative

TP	True Positive
TPR	True Positive Rate
UAV	Unmanned Aerial Vehicle
UIC	User Interface Component
UTAUT	Unified Theory Of Acceptance And Use Of Technology
VB	Flank Wear
VIT	Viticulture

Part I

Fundamentals

1 Introduction

"A picture is worth a thousand words."[1] Consider a photograph of a house to understand the meaning of this saying: observers see many details like the roof, doors, windows, and the colors of walls. Furthermore, they can see the location and size of different rooms. They might even make assumptions about the people living in the house and their lifestyle. Also, the image conveys additional information about, for example, the surrounding of the house, its age, and its architectural style. This example illustrates that it requires a large number of words to only approximate the amount of information contained in a single image. From a data perspective, image data is considered unstructured data as opposed to structured data, which is stored, for example, in spreadsheets and relational databases, and can be directly processed automatically. It is estimated that 80 to 90% of all data is unstructured (CIO.com, 2019). Hence, it is particularly promising to further exploit this type of data for value creation.

A crucial step for value creation based on image data is the application of CV techniques that allow to automatically extract information from images (Szeliski, 2010) and thus enable subsequent automated processing of image data. In recent years, the field of CV has made significant advances because of machine learning (ML) techniques (LeCun et al., 2015). Already in 1959, Arthur Samuel defined ML as giving computers the ability to learn without being explicitly programmed (Samuel, 1959). Nowadays, mainly deep learning (DL) is used for CV tasks. DL is a subfield of ML which relies on deep neural networks (Janiesch et al., 2021). In broader terms, ML and DL are subfields of artificial intelligence (AI) (Kühl et al., 2019).

While these technological advances facilitate novel forms of value creation, companies are currently facing challenges due to fast-changing environments. The average lifespan of companies in the S&P 500 stock index, comprising the largest 500 stock-indexed companies in the United States, has been reduced from around 35 years in 1976 to approximately 22 years in 2020 (INNOSIGHT, 2019). More recently, the consequences of current events like the Covid-19 pandemic, the Ever Given cargo ship blocking the Suez Channel, and the Russian invasion of Ukraine

[1] derived from "One look is worth a thousand words" by Frederick R. Barnard in Printer's Ink, 1921.

show how rapidly changing, globalized, and interconnected our modern world is. This threatens many established forms of value creation. Therefore, companies must adopt novel forms of value creation to stay competitive.

A popular, novel way of value creation is to combine data and analytical methods to make better decisions and solve complex problems (Hunke et al., 2022).

For CV, in particular, the application in various areas has economic importance and can bring value to society: e.g., in agriculture (Tian et al., 2020), manufacturing (J. Wang et al., 2018) and autonomous driving (Grigorescu et al., 2020). However, contrary to past predictions (e.g., Faggella (2020)), many of these applications are not fully automated yet and require certain human inputs. For example, in the field of autonomous driving, the level of autonomy is classified from level 0 (no driving automation) to level 5 (full driving automation) (On-Road Automated Driving (ORAD) committee, 2021). Currently, the cars equipped with the highest degree of autonomy are in level 3 (Nedelea, 2021) — conditional driving automation — according to the classification mentioned above. That means the driver is still required to take over when the automated driving system requests this.

Considered with a more abstract lens, CV alone is often just an intermediary step. In terms of the *data, information, knowledge, and wisdom (DIKW)* pyramid based on Ackoff (1989), CV techniques can be used to convert images (data) to information automatically. Yet, to reach higher levels of abstraction (knowledge and wisdom), combining human and artificial intelligence is often necessary. Decision support

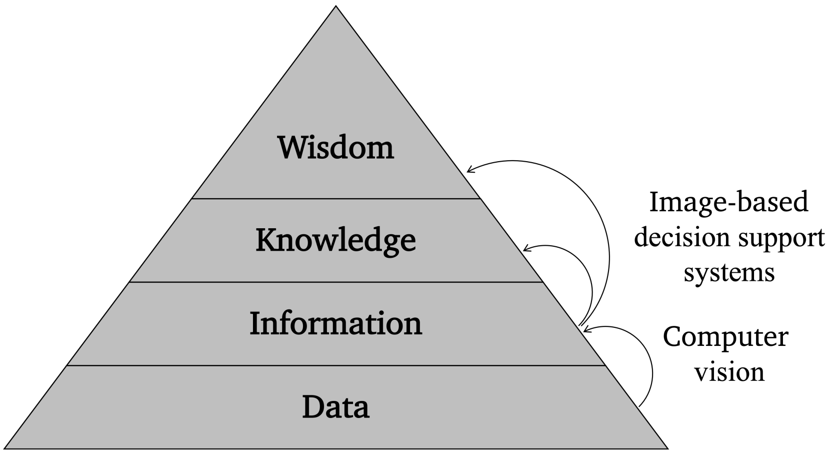


Fig. 1.1.: Data, information, knowledge, and wisdom pyramid for image-based decision support systems based on Ackoff (1989).

systems (DSSs) are a proven tool to generate value (Kohli & Devaraj, 2004) by combining human and artificial intelligence (Power, 2002, p. 149). DSS "is a general term for any computer application that enhances a person or group's ability to make decisions" (Power, 2008, p. 149). While existing literature already addresses similar data types like text (Abbasi & Chen, 2008), there is only little work on DSSs that rely on images and CV. We call this novel class of systems image-based decision support systems (IB-DSSs). Compare Figure 1.1 on page 4 for a graphical depiction of the DIKW pyramid for image data and the role of CV and IB-DSSs.

This thesis contributes to making use of the information value of images. It demonstrates how IB-DSSs can be designed and used to be of value to society — supporting people in their work life, creating monetary value, and lessening environmental impact.

1.1 Essential Terminology

Before introducing the structure of this work and our research design, we briefly present terminology that is essential for understanding the remainder of this thesis. Central terms for this thesis are image-mining-based decision support systems (IM-DSSs), image-based decision support systems (IB-DSSs), and computer-vision-based hybrid intelligence systems (CV-HISs). Their relation is depicted in the Venn diagram in Figure 1.2 and described in the following.

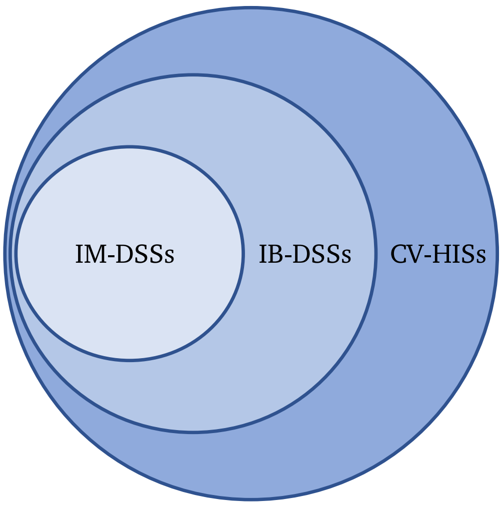


Fig. 1.2.: Venn diagram showing the relationship of image-mining-based decision support systems (IM-DSSs), image-based decision support systems (IB-DSSs), and computer-vision-based hybrid intelligence systems (CV-HISs).

We understand the most general class of the systems under consideration, CV-HISs, as all information systems (IS) that rely on CV and hybrid intelligence (HI) — a combination of human and artificial intelligence (Dellermann, Ebel, et al., 2019).

For example, CV-HISs comprise active learning systems (Settles, 2009); i.e., systems for developing accurate ML/DL models for CV efficiently. In active learning systems, AI in the form of ML/DL models signals for which data instances human input would probably be most helpful for model improvements. Subsequently, human intelligence is used for the actual labeling.

In contrast, we understand DSSs as IS that support decisions that are directly relevant to the business and rely on “finished” ML/DL models. As a result, we understand IB-DSSs as a subclass of CV-HISs. Conversely, we understand every IB-DSS as CV-HIS because DSSs are intended to support decision-makers and not replace them (Power, 2002). As a consequence, all IB-DSSs rely on hybrid intelligence because artificial intelligence in the form of CV models based on DL is always combined with human intelligence for decision-making. An example of an IB-DSS is a DSS for the detection of maintenance needs in infrastructure like wind turbines: a single image is sufficient to decide whether a component needs to be serviced or exchanged.

This is in contrast to the subclass of IM-DSSs: they provide decision-makers with information based on large amounts of image data, e.g., multiple images describing the same real-world phenomenon. Consequently, the information extracted from the images with CV needs to be aggregated further to facilitate human decision-making. This aggregation is achieved with image mining, which “deals with the extraction of implicit knowledge, image data relationship, or other patterns not explicitly stored in the image databases” (Hsu et al., 2002, p. 1). Image mining is most widely used for medical applications; in this domain, it is called radiomics (Gillies et al., 2016; Lambin et al., 2012). An industrial example of an IM-DSS is the monitoring of process stability in a production line. Inspecting several images of produced goods is necessary to estimate the process stability and derive subsequent actions.

1.2 Structure of Work & Research Design

Having presented essential terminology for this thesis, we now introduce the structure of this work and our research design. This thesis comprises five parts, each consisting of one or more chapters. Figure 1.3 provides an overview of the structure of this thesis. Also, it indicates which research question (RQ) is addressed in which chapter.

In Part I, we lay the foundations for the rest of the thesis. First, in Chapter 1, we

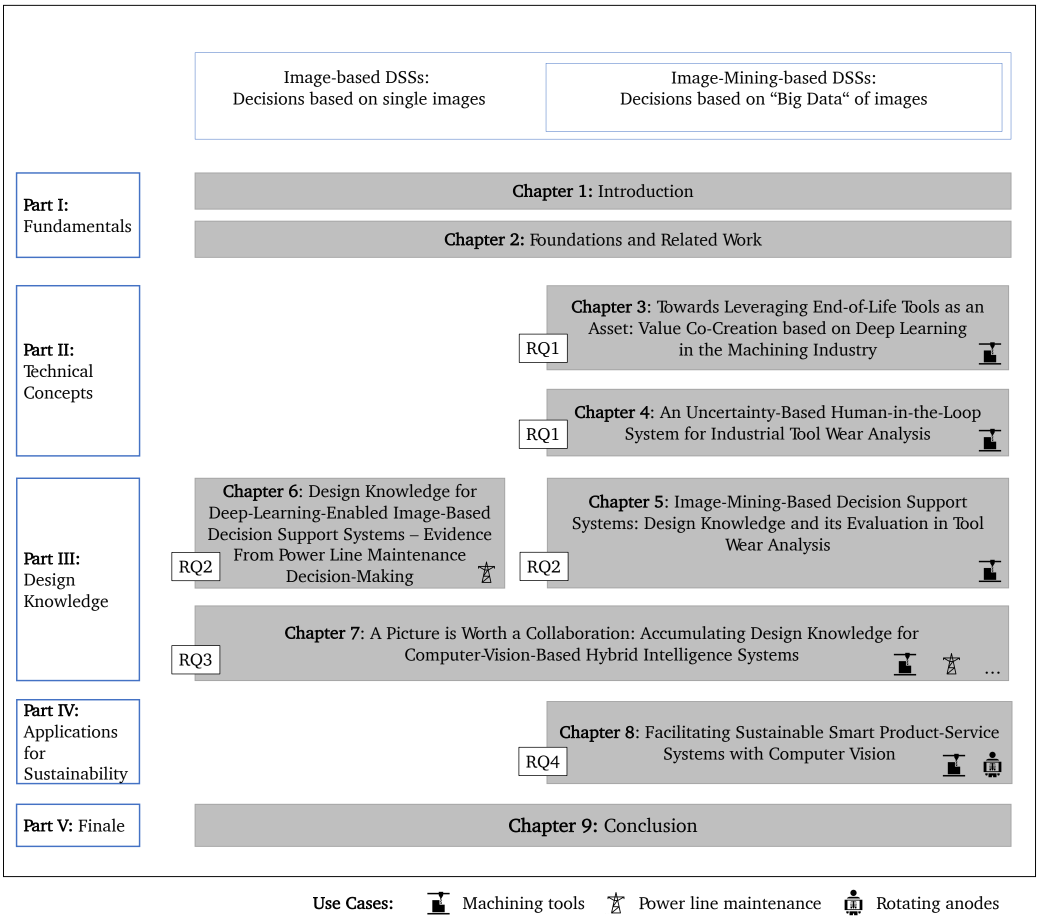


Fig. 1.3.: Structure of this thesis.

introduce the motivation and essential terminology for this thesis. Furthermore, we describe the structure and research design of our work. Then, we present foundations and related work in Chapter 2. We start with foundations regarding design science research (DSR), CV, and DL. Afterwards, we describe related work regarding DSSs, and finally regarding IB-DSSs and IM-DSSs in particular.

Tab. 1.1.: Use case overview.

	Machining tools	**Power line maintenance**	**Rotating anodes**
Business challenge	Identify wear on machining tools to derive improvement options for machining processes and tools	Identify maintenance needs for power line infrastructure	Identify wear on rotating anodes to derive remanufacturing options
Computer vision task	Image classification and semantic segmentation	Object detection	Semantic segmentation
Relating to research questions	1, 2, 3, 4	1, 2, 3	1, 4

Parts II-IV constitute the core part of this thesis — each part comprises one to three studies as separate chapters. We are working with several real-world use cases in different industry domains to ensure our work's real-world validity and usefulness. A brief overview of the use cases can be found in Table 1.1. Also, the industry use cases described in Table 1.1 are shown as pictograms in the lower right corner of the chapters in Figure 1.3 on page 7. In the following, we describe our RQs, the corresponding research design, and the individual chapters in Part II-IV.

A first, crucial step for enabling IB-DSSs is to convert image data into information that computers can process (compare Figure 1.1 on page 4). To this end, in Part II, we explore, develop and evaluate established and novel options for turning image data into valuable information based on mainly artificial and also human intelligence when needed. Thus, we ask:

> **Research Question 1 (RQ1)**
> How can image data be converted into valuable information by combining artificial and human intelligence?

We address RQ1 with two technical studies. The first study, in Chapter 3, relies on DSR[2] — DSR has proven to be an essential paradigm for the development of IS (Gregor & Hevner, 2013) as it allows to examine the theoretical and practical

[2] For readers unfamiliar with design science research, we recommend reading Section 2.1 already now.

tasks required for designing and building IS. We apply DL-based CV in a particular scenario — and turn image data into information in the machining tools use case. This serves not only to demonstrate feasibility in this particular domain, but also to explore opportunities for future research. To this end, we first build a dataset of images of worn tools from an actual machining process. Subsequently, we train and evaluate a DL-based CV model for classifying if a given microscopic image of a machining tool shows a specific type of wear. The results show that it is possible to classify different wear types with a DL-based CV model.

In Chapter 4, we realize the first of the future research possibilities identified and described in Chapter 3. Domain experts state that a DL-based CV model for detecting wear on machining tools in a pixel-accurate manner would be even more valuable than image classification. Consequently, we train and evaluate DL-based CV models for pixel-accurate wear detection on microscopic images of machining tools. Having shown the feasibility of this wear detection, we develop and assess an approach for estimating the uncertainty of predictions of the applied DL-based CV model. Images for which the DL-based CV model indicates high uncertainty are deferred to a human expert. We show that this selected involvement of humans in a human-in-the-loop system can improve the overall system's performance. To ensure that the findings of this study are valid not only for our specific use case, we also evaluate the approach on the publicly available Cityscapes dataset for urban scene understanding (Cordts et al., 2016).

Subsequently, in Part III, we develop and evaluate design knowledge for IB-DSSs, IM-DSSs, and CV-HISs. As described previously in Chapter 1 and depicted in Figure 1.1 on page 4, CV (turning data into information) alone is just an intermediary step for generating value in terms of knowledge and wisdom from images. DSSs are well-established tools for generating knowledge and wisdom based on a combination of artificial and human intelligence. However, despite this potential for value creation, there is little previous research regarding the design of IB-DSSs. Consequently, the next step in this thesis is to address the design of IB-DSSs. We rely on DSR for addressing this RQ since it has proven to be a particularly important paradigm for developing DSSs (Arnott & Pervan, 2012). One of the main goals of DSR is the formulation of design theories (Beck et al., 2013). Therefore, we ask:

Research Question 2 (RQ2)
What design knowledge should guide the development of image-based decision support systems?

We perform two separate studies to address this RQ — one regarding the subclass of IM-DSSs and the other for IB-DSSs. In each study, we use a different use case to develop and evaluate the design knowledge.

First, in Chapter 5, we develop and evaluate design knowledge for IM-DSSs. As described previously, IM-DSSs are suitable when a big amount of image data is to be analyzed for decision-making. To ensure the practical relevance of the developed design knowledge, we conduct the study based on the machining tools use case. This is suitable since the wear on machining tools from an identical process is subject to variations. Consequently, there is a need for images of many worn machining tools from a given process. Image mining allows for a holistic view of the wear on machining tools from a machining process and consequently leads to a better decision basis for process improvements. In the first step of this DSR study, we obtain design requirements for IM-DSSs from literature and interviews with domain experts. Then, we formulate design principles addressing these design requirements based on appropriate literature. Subsequently, over three design cycles, we instantiate and evaluate design features derived from these design principles in an IM-DSS. This IM-DSS supports the analysis of machining processes and the identification of process improvement options. Depending on the design principles and features, the design cycles, and their goals, we apply different evaluation methods: two technical experiments, one exploratory focus group, one confirmatory focus group, and one logical argument. The evaluation confirms our nascent design knowledge's sufficient effectiveness, efficiency, and usefulness. Also, the evaluation confirms the usefulness of the artifact itself.

In Chapter 6, we develop and evaluate design knowledge for IB-DSSs. Also for this study, we use a practical industry use case to ensure the practical relevance and usefulness of the developed design knowledge. In the power line maintenance use case, it is possible to derive the necessary information for one business-relevant decision from one single image. Based on a single image of, e.g., an insulator, it is possible to decide if maintenance is required for this part. Hence, this use case is suitable for a IB-DSS. The methodology is similar to the previous study: Initially, we formulate design requirements based on a structured literature review and interviews with domain experts. Then, we describe design principles addressing the design requirements. Building on this conceptualization, design features derived from the design principles are instantiated in and evaluated through an IB-DSS that supports power line maintenance decision-making based on images captured by unmanned aerial vehicles.
This study's structure follows the partition of a DSS of Turban et al. (2010) into the model component and the user interface component. First, we address the model

component of the DSS. As we cannot build on previous work for this, we assess the technical feasibility of converting image data into valuable information with a DL-based CV model (compare RQ1). To this end, we build DL-based CV models for detecting the state and localization of power line infrastructure components. We then evaluate the models with a technical experiment and interviews with domain experts. The technical experiment confirms that it is possible to use DL-based CV models to convert images of power line components into valuable information. Additionally, the interviews confirm that domain experts perceive the model component as useful. Building on this model component of the DSS, we design, implement and evaluate the user interface component. We assess the user interface component with nine one-on-one confirmatory workshops. The results of these workshops show the suitability, usefulness, and effectiveness of the developed design knowledge and the corresponding artifact.

Having developed and evaluated design knowledge for IB-DSSs and IM-DSSs in the studies addressing RQ2, we aim to reach a higher degree of generalizability with addressing RQ3. Here, we regard design knowledge for CV-HISs — a superordinate class of IB-DSSs.

> **Research Question 3 (RQ3)**
> What design knowledge should guide the development of computer-vision-based hybrid intelligence systems?

The study in Chapter 7 addressing this RQ contributes in a more theoretical manner due to two reasons. First, this study abstracts across four real-world CV use cases in addition to the machining tools and the power line maintenance use case. Second, as described previously, we understand the focus of CV-HISs as broader than that of IB-DSSs. Furthermore, the design knowledge in this study complements the design knowledge for IB-DSSs and IM-DSSs by addressing another perspective. The main focus of the design knowledge developed in Chapter 7 is how IS can be designed such that hybrid intelligence (i.e., the combination of human and artificial intelligence) is facilitated. In contrast, the design knowledge described in Chapter 5 and Chapter 6 is focused on the concrete design of IB-DSSs and IM-DSSs.

We also apply DSR for this study — however, we rely on a different DSR strategy. As described in more detail in Section 2.1 there are two strategies for DSR studies (Iivari, 2015). The studies we described so far rely on *strategy I* if they rely on DSR. In these studies, we start with a problem class: we build an artifact as a general solution concept to address this problem class. In contrast, in studies relying on *strategy II*, first, an artifact that solves a specific client problem is created. Afterwards, design knowledge is formulated by abstracting from the concrete artifact (Möller et al., 2020). In the study in Chapter 7, we aim to derive design knowledge for CV-HISs by employing the reflective DSR *strategy II*. We gather a focus group of six experts with relevant expertise in developing CV-HISs to participate in a series of workshops. First, we conceptualize the CV-HIS as a collaboration of human and computer to solve a vision-based task. As a result, we identify four design-related mechanisms: automation, signaling, modification, and collaboration. These mechanisms inform our derived meta-requirements and design principles. Then, we describe meta-requirements that are derived from case-specific requirements and literature. Subsequently, we formulate design principles based on specific design features implemented in the six different real-world CV use cases. This study can help practitioners design CV-HISs and lays the foundation for many future research directions.

The studies described so far focus on technical aspects and the design of IB-DSSs. Additionally, we evaluate the real-world impact of IB-DSSs. In light of climate change being a serious threat to humanity (Pörtner et al., 2022), we perform this evaluation in terms of possible improvements in environmental sustainability. In Part IV, we aim to answer the call by Zeiss et al. (2021, p.148) for IS research helping "to intensify and extend use of products and components and recycle waste materials".

Research Question 4 (RQ4)
How can image-based decision support systems be applied to improve environmental sustainability in the industry?

In the study in Chapter 8 that addresses RQ4, we rely on two use cases: the machining tools case frequently mentioned above and the rotating anode case. For the rotating anode case, we first show in a technical experiment in this study that DL-based CV models are suitable for pixel-accurate wear detection on microscopic images of rotating anodes (compare RQ1) since this has not been shown previously. To assess the sustainability impact of the different improvement scenarios enabled by the IB-DSS for the respective use case, we perform life cycle assessments (LCAs). LCA is a widely used, standardized method for assessing the quantitative environmental impact of products, processes, services, and systems throughout their life cycles (Finkbeiner et al., 2006). The LCAs in this study are based on real data from our case companies when possible and assumptions by domain experts when necessary.

The results of the LCAs show that the improvement scenarios enabled by the IB-DSSs facilitate a reduction of 12% (machining tools) and 44% (rotating anodes) of emission of CO_2 equivalents. We are convinced that IB-DSSs can be employed for sustainability improvements for many more challenges. Therefore, at the end of this study, we conceptualize our approach and describe the prerequisites for applying it to other use cases.

The last Part V first summarizes this thesis and then presents contributions, limitations, and directions for future research.

1.3 Development of Work

As mentioned previously, Part II-IV each contain one to three studies. Four of the six studies in total have already been published in academic outlets; the remaining two are included as working papers. Table 1.2 on page 14 provides an overview of the studies in this thesis. When available, we report the VHB-Jourqual3 (Verband der Hochschullehrerinnen und Hochschullehrer für Betriebswirtschaft e.V., 2022) ranking for all published studies. The outlet for the study in Chapter 4 is not ranked in the VHB-Jourqual3; therefore, we report the CORE ranking (The Computing Research and Education Association of Australasia, CORE Inc, 2022).

Tab. 1.2.: Overview of studies included in this thesis. Status as of December 15, 2022.

Chapter	Title	Authors	Outlet	Year	Ranking
3	Towards Leveraging End-of-Life Tools as an Asset: Value Co-Creation based on Deep Learning in the Machining Industry	Walk, J.; Kühl, N.; Schäfer, J.	Proceedings of the 53rd Hawaii International Conference on System Sciences (HICSS-53)	2020	C
4	An Uncertainty-Based Human-in-the-Loop System for Industrial Tool Wear Analysis	Treiss, A.; Walk, J.; Kühl, N.	Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2020)	2020	A
5	Image-Mining-Based Decision Support Systems: Design Knowledge and its Evaluation in Tool Wear Analysis	Walk, J.; Schemmer, M.; Kühl, N.; Satzger, G.	Working Paper	2022	N/A
6	Design Knowledge for Deep-Learning-Enabled Image-Based Decision Support Systems	Landwehr, J.; Kühl, N.; Walk, J.; Gnädig, M.	Business & information systems engineering (BISE)	2022	B
7	A Picture is Worth a Collaboration: Accumulating Design Knowledge for Computer-Vision-based Hybrid Intelligence Systems	Zschech, P.; Walk, J.; Heinrich, K.; Vössing, M.; Kühl, N.	Proceedings of the 29th European Conference on Information Systems (ECIS 2021)	2021	B
8	Facilitating Sustainable Smart Product-Service Systems with Computer Vision	Walk, J.; Kühl, N.; Saidani, M.; Schatte, J.	Working Paper	2022	N/A

2 Foundations and Related Work

In this chapter, we present relevant foundations and related work. First, in Section 2.1, we describe foundations regarding design science research (DSR). After, in Section 2.2, we describe basics regarding computer vision (CV) and deep learning (DL). Subsequently, in Section 2.3 we introduce decision support systems (DSSs). Finally, we discuss related work regarding image-based decision support systems (IB-DSSs) and image-mining-based decision support systems (IM-DSSs) in Section 2.4.

2.1 Design Science Research

In this section, we briefly present key concepts, terminology, and strategies of DSR that are relevant for the remainder of this thesis.

DSR is a research paradigm aiming to provide design knowledge and construct innovative artifacts for real-world problems (vom Brocke et al., 2020). It is an important research paradigm for IS research in general (Gregor & Hevner, 2013) and the development of DSSs in particular (Arnott & Pervan, 2012).

The formulation of design knowledge is a major goal of DSR. Design knowledge describes “how things can and should be constructed or arranged (i.e., designed), usually by human agency, to achieve a desired set of goals” (vom Brocke et al., 2020, p. 2). Thus, design knowledge can serve as a blueprint for practitioners and researchers that build similar systems or address similar problems. Generally, design knowledge is formulated on such a level of abstraction that it is valid for a problem class and not only a single problem (Hevner et al., 2004). A typical way to describe design knowledge, which we also rely on in several chapters of this thesis, is along the categories of design requirements, design principles, and design features. In the following, we describe these key terms.
Design requirements are abstract requirements for a problem class. While the terms “design requirement” and “meta-requirement” are often used interchangeably, some authors use the term meta-requirement to indicate a higher level of abstraction

(Maedche, Gregor, et al., 2019). Importantly, both design requirements and meta-requirements are to be distinguished from requirements in software engineering; software requirements are less abstract (Maedche, Gregor, et al., 2019). **Design principles** are “prescriptive statements that indicate how to do something to achieve a goal” (Gregor et al., 2020, p. 1622); they are formulated such that they address the design requirements. **Design features** are derived from the design principles and are concrete enough that they can be instantiated in an artifact (Meth et al., 2015).

Artifacts in DSR studies can be of many different types and serve various purposes. Exemplary artifacts realized in DSR studies include blockchain-based smart contracts for shipping documents (Nærland et al., 2017), an ML-based classification model for detecting fraudulent documents in the context of stock market manipulations (Siering et al., 2021), and a phase model for the development and application of maturity assessments for social and technical systems (Mettler, 2011). Further examples are a conversational agent (Gnewuch et al., 2017) and a system for semi-automated requirements elicitation from natural language (Meth et al., 2015). Also, many DSSs serve as artifacts in DSR studies. For example, Koornneef et al. (2020) develop a DSS for aircraft dispatch assessment and Gottschlich and Hinz (2014) propose a DSS for stock investment recommendations using collective wisdom in their DSR study. The artifact developed by Ferro et al. (2020) addresses a problem in the health care sector; their DSS predicts patient no-show behavior.

There are two DSR strategies for deriving generalizable design knowledge (Iivari, 2015). The currently predominant strategy of DSR studies is referred to as *strategy I* (Iivari, 2015). Studies employing this strategy start with a problem class: they build their artifact as a general solution concept to address this problem class. There are many different research frameworks and guidelines for structuring DSR studies following *strategy I* (e.g., Kuechler and Vaishnavi (2008), Peffers et al. (2007), and Venable et al. (2016)). The three-cycle guidelines by Hevner (2007) are a frequent choice for structuring DSR studies following *strategy I*. These studies then comprise a rigor cycle, a relevance cycle, and one or more build-and-evaluate cycles. In the relevance cycle, it is ensured that a real-world problem is addressed (vom Brocke et al., 2020). This can be achieved, for example, through a literature review or exploratory focus groups with domain experts (Tremblay et al., 2010a). Furthermore, conducting a rigor cycle makes sure that the research is “standing on the shoulder of giants” by using appropriate, existing foundations and methodologies (Hevner, 2007). Often, a literature review is performed to this end. Building on the relevance and rigor cycle, one or more build-and-evaluate cycles are performed (vom Brocke et al., 2020); often, these cycles are also called design cycles (Hevner, 2007). In these

cycles, a real-world artifact is created and subsequently evaluated. For example, in a technical experiment or with a survey with potential users. These design cycles are the central activity in DSR studies following *strategy I*.
On the other hand, studies employing DSR *strategy II* first build a concrete artifact that aims to solve a specific client problem. Based on this, generic knowledge addressing the respective problem class is formulated in a reflective manner by abstracting from the specific implementation (Möller et al., 2020). According to Iivari (2015), there was little practical experience with *strategy II* in 2015; this is still true today. Therefore, we cannot refer to established research frameworks and guidelines for *strategy II*.

2.2 Computer Vision and Deep Learning

In this section, we will first define computer vision (CV) and briefly describe its history and the basic working principles of deep learning (DL)-based CV models. Then, we present different CV tasks and appropriate models. Lastly, we describe suitable evaluation measures that are relevant for the remainder of this thesis.

CV aims to equip computers with human-like visual perception abilities (Szeliski, 2010). Originally, CV relied on techniques like edge detection and filters (Szeliski, 2010). For those techniques, a CV engineer needs to define, e.g., the filters, which are then applied to images. These techniques are now referred to as *traditional* CV techniques (O'Mahony et al., 2019). Lately, it has been shown that learning entire CV models, including the filters from data with ML leads to CV models that produce more accurate outputs. For specific tasks, even human performance was surpassed by CV models based on ML (He et al., 2015). ML is a relatively old field of research — already in 1959, Arthur Samuel defined ML as giving computers the ability to learn without being explicitly programmed (Samuel, 1959). Current CV models are based on DL, a subfield of ML that relies on deep neural networks (Janiesch et al., 2021). In particular, convolutional neural networks (CNNs) are used for CV tasks. They are well suited for data with spatial relationships like images and can be trained with less data than fully-connected neural networks since the number of parameters to be trained is lower.

In the following, based on LeCun et al. (2015), it is briefly described how CNNs work. Similar to other types of neural networks, CNNs consist of multiple processing layers. In each layer, the input data is represented with different degrees of abstraction. The outputs of a certain layer that serve as inputs for the next layer are called feature

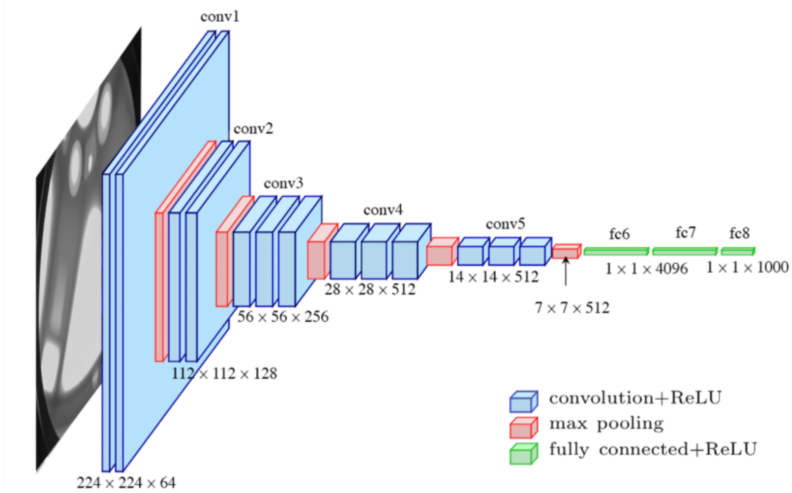


Fig. 2.1.: VGG16 network architecture (Simonyan & Zisserman, 2014), figure from Ferguson et al. (2017).

maps. In the first layers, CNNs use mainly convolutions, pooling operations, and activation functions. In the convolutional layers, filters are applied for detecting features — as opposed to traditional CV, these filters are not selected by humans but learned from the data. Pooling layers combine semantically similar features. Activation functions allow neural networks to learn non-linear relationships.

Figure 2.1 shows an exemplary CNN architecture; we will now explain the different operations applied in the first layers. The network receives an image with a size of 224x224x1 pixels as input (we are considering a grayscale image here). In the first two convolutional layers, 64 filters are applied per layer. The filters are typically much smaller than the input feature maps and are thus applied to the different regions of the input feature map in a sliding window fashion. Since 64 filters are applied, the resulting feature maps have a dimensionality of 224x224x64.[1] After each convolutional layer, an activation function is applied to introduce non-linearity in the CNN. In Figure 2.1, a rectified linear unit (ReLU) is applied as an activation function. Following the convolutional layer and the ReLU, 2x2 maximum pooling is applied — in a sliding window fashion, the maximum value of a square of four adjacent entries from the feature maps is computed. For example, we regard a feature map resulting from applying a filter that detects edges. By applying maximum pooling, it becomes irrelevant which of the four adjacent entries in the feature map detected the edge. In Figure 2.1, this maximum pooling is applied to all non-overlapping squares containing four pixels. This halves the dimensionality of the resulting feature maps

[1]Technical details like filter size, stride, and padding that ensure that the first two output dimensions of a convolutional layer are equal to the first two input dimensions are left out here for simplicity.

and allows the detection of features with a certain independence of their localization in the image. Hence, the resulting feature map has a dimensionality of 112x112x64. After applying two more convolutional layers, the dimensionality is increased to 112x112x128.
As described above, each filter is applied to all inputs from the previous layer in a sliding window fashion. Thereby the number of weights is reduced considerably compared to fully-connected networks where the weight is distinct for each connection of two neurons. As a consequence, CNNs can be successfully trained with drastically less data and computing power than a fully-connected neural network for the same task (LeCun, Bengio, et al., 1995). The nature of the last layers depends on the concrete task: Usually, the output is computed either directly by a convolutional layer (Ronneberger et al., 2015) or by a series of fully-connected layers (Krizhevsky et al., 2012) (cf. Figure 2.1 on page 18).

The following CV tasks are the most typical ones on static images (cf. Figure 2.2): Image classification, object detection, and semantic segmentation (Griebel, Dürr, et al., 2019). A CNN for image classification produces just one or more class labels for the entire image as output, while a CNN for object detection locates objects of interest within an image — its outputs are bounding boxes around the objects alongside the class labels. CNNs for semantic segmentation yield an even more detailed output by assigning a class label to each individual pixel.

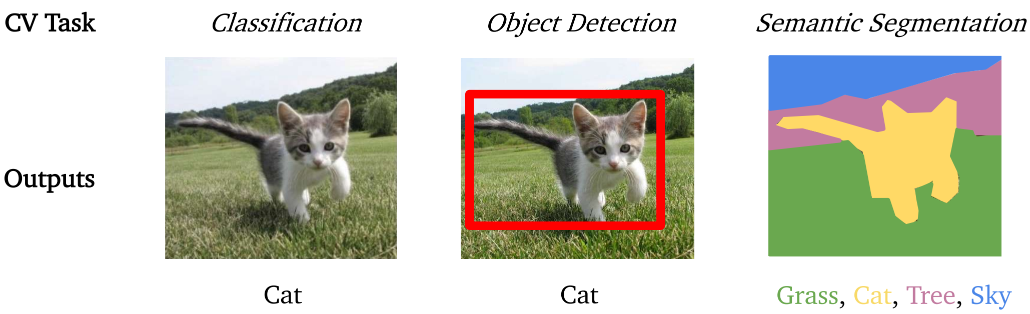


Fig. 2.2.: Outputs of typical CV tasks for an image of a cat. Own representation based on F.-F. Li et al. (2017) and Kosson and Marklund (2018).

In the following, we briefly describe typical architectures for the CV tasks depicted in Figure 2.2. An early CNN for image classification, in particular recognizing handwritten zip codes for the U.S. postal service, was developed in 1989 by LeCun et al. With increasing computer power, and in particular graphical processing units being used for training CNNs, the architectures got “deeper” in terms of the number of layers and were able to solve more and more complex computer vision tasks. Proven architectures for image classification include the VGG16 network (Simonyan

& Zisserman, 2014) and the ResNet (He et al., 2016). More recent architectures enable two main types of improvement: reduced need for resources and higher classification accuracy. To use DL-based CV models on devices with less computational power (mobile devices, edge devices, etc.), novel architectures like the MobileNetV2 (Sandler et al., 2018) that are significantly less resource-hungry were created. The MobileNetV2 network consists of 3.5 million parameters that have to be learned during training and stored for inference. In contrast, e.g., the VGG16 network mentioned above consists of 138.4 million parameters. Also, this reduced number of parameters leads to lower inference times which is particularly important for near real-time applications. All these enhancements were possible while the accuracy on the ImageNet dataset (J. Deng et al., 2009), which is widely used for benchmarking, is on par with the one of the VGG16 (Keras, 2022). In case classification accuracy is more important than the reduced need for resources, novel architectures like the EfficientNetV2 are available (Tan & Le, 2019).

Building on these advances regarding image classification, there were also innovations regarding CNNs for object detection and semantic segmentation. CNNs for object detection have a more complex structure since they perform classification and localization. Early approaches like the one by Sermanet et al. (2013) rely on image classification for various regions of an input image in a sliding window fashion. Current architectures reduce the computational effort by using generated region proposals instead of the sliding window approach. The current architectures can be categorized into one-stage detectors, CNNs performing classification and localization at once (e.g., YOLO by Redmon et al. (2016) and SSD by W. Liu et al. (2016)), and two-stage detectors, CNNs predicting localization and class label separately (e.g., Faster R-CNN by S. Ren et al. (2015)).

CNNs for semantic segmentation compute a class label for each input pixel as shown in Figure 2.2 on page 19. To achieve this, they consist of a downsampling path and an upsampling path. The downsampling path is similar to the first layers of an image classification network (cf. Figure 2.1 on page 18) — the dimensionality is reduced from layer to layer. At the same time, the number of features is increased. In the upsampling path, the number of features is reduced, and the dimensionality is increased again until the original image's dimensions are reached. Typical architectures for semantic segmentation include the U-Net (Ronneberger et al., 2015), Fully Convolutional Neural Networks (Long et al., 2015), Mask R-CNN (He et al., 2017), and the Segnet (Badrinarayanan et al., 2017).

Tab. 2.1.: Confusion matrix.

	Predicted positive	**Predicted negative**
Ground truth positive	True positive	False negative
Ground truth negative	False positive	True negative

In the following, we describe relevant metrics for evaluating the accuracy of DL models throughout this work; the metrics are sorted according to the CV tasks described in Figure 2.2 on page 18.

Metrics for image classification are built on the so-called confusion matrix. It compares ground truth labels and predictions. There are four options as illustrated in Table 2.1 (Ting, 2017a): true positives (TPs) and true negatives (TNs) are correct predictions, while false positives (FPs) and false negatives (FNs) describe the two error types.

The most straightforward evaluation metric for image classification is the accuracy — it divides all correct predictions by the total number of predictions issued (Sammut & Webb, 2017):

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}. \tag{2.1}$$

However, in reality, the accuracy is often not suitable due to the class imbalance problem: consider an image classifier distinguishing cats and dogs, trained and tested with datasets consisting of 95% cat and 5% dog images. A simple classifier that predicts the majority class for each image, in this case, cat, will yield an accuracy of 95%. Despite 95% sounding like a good result, the classifier cannot distinguish cats and dogs and hence does not provide actual value for this task.

A common evaluation metric for classification tasks with imbalanced classes is the matthews correlation coefficient (MCC) (Matthews, 1975). The MCC considers class imbalance — an MCC of “0” corresponds to random guessing based on the relative size of the classes. Perfect predictions yield an MCC of “1”, “-1” indicates that the predictions are inverse to the actual labels. According to Chicco and Jurman (2020), the MCC is defined as

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}. \tag{2.2}$$

Precision and recall are two evaluation metrics stemming from document retrieval (Ting, 2017b). In ML as well as document retrieval tasks, there is often a high share of TNs, and in many application cases, the TNs are not of particular interest — hence, they are ignored by precision and recall. Precision is defined as

$$Precision = \frac{TP}{TP + FP}. \tag{2.3}$$

Recall, also called true positive rate (TPR), is defined as

$$Recall/TPR = \frac{TP}{TP + FN}. \tag{2.4}$$

Based on precision and recall, the F1-score is defined as the harmonic mean of precision and recall (Ting, 2017b):

$$F1\text{-}score = \frac{2 * Recall * Precision}{(Recall + Precision)}. \tag{2.5}$$

For tasks with class imbalance, it is possible and reasonable to weight precision, recall, and F1-score according to the relative class frequencies.
Another popular evaluation metric for classification tasks is the receiver operating characteristic (ROC) curve (Melo, 2013): it shows the trade-off relationship between the TPR and the false positive rate (FPR). The latter is defined as

$$FPR = \frac{FP}{FP + TN}. \tag{2.6}$$

AUROC stands for area under the receiver operating characteristic curve: "1" indicates a perfect image classification model while "0.5" corresponds to random guessing. It is considered to be particularly robust as it takes into account all possible classification thresholds.

Evaluation metrics for object detection consider the accuracy of the object localization in addition to the accuracy of the classification. In object detection, object localization is usually defined by the coordinates of a rectangular bounding box containing the object of interest (cf. Figure 2.2 on page 19). In the following, we explain common metrics for object detection relevant for this work based on Padilla et al. (2020). A standard metric for evaluating the localization error of an object detection model is the intersection over union (IoU). It is computed based on the area of the bounding box depicting the ground truth and the area inside the bounding box that represents the prediction. Precisely, it is defined as:

$$IoU = \frac{area\ of\ intersection}{area\ of\ union} = \frac{ground\ truth\ area\ \cap\ predicted\ area}{ground\ truth\ area\ \cup\ predicted\ area}. \tag{2.7}$$

The IoU takes values between “0” (no overlap) and “1” (perfect overlap). To convert the IoU values to binary TP, FP, and FN values, a threshold between “0” and “1” is applied. Note that TNs are not of interest for object detection as there is an high number of TNs in each image. Increasing the IoU threshold leads to higher precision because there will be fewer FPs and lower recall because there will be more FNs. Conversely, a lower threshold leads to a lower precision since there will be more FPs and a higher recall since there will be fewer FNs. Ideally, we want a model to have high precision and recall. A common metric for assessing this trade-off relationship is the average precision (AP). The AP is calculated by averaging the precision values evaluated at multiple recall points of the precision-recall curve. A popular implementation is the 11-point interpolation:

$$AP_{11} = \frac{1}{11} \sum_{R \in \{0, 0.1, \ldots, 0.9, 1\}} P_{\text{interp}}(R) \tag{2.8}$$

with

$$P_{\text{interp}}(\tilde{R}) = \max_{\tilde{R}:\tilde{R} \geq R} P(\tilde{R}). \tag{2.9}$$

In case an object detection task has more than two classes, the APs of each class are aggregated to the mean average precision (mAP):

$$mAP = \frac{1}{C} \sum_{c=1}^{C} AP_{\text{c}} \tag{2.10}$$

with C denoting the number of classes.

Evaluation metrics for semantic segmentation are conceptually simpler again since we can directly compare prediction and ground truth on a pixel basis. A widely used metric is the dice similarity coefficient (Setiawan, 2020), also called the dice coefficient. It assesses the overlap, or intersection, between the model’s outputs as binarized one-hot encoded predictions $\hat{y}_{i,c}$ and the one-hot-encoded ground truth labels $g_{i,c}$. It is defined as:

$$Mean\ Dice\ Coefficient = \frac{2}{C} \sum_{c=1}^{C} \frac{\sum_{i=1}^{N} \hat{y}_{i,c}\, g_{i,c}}{\sum_{i=1}^{N} \hat{y}_{i,c} + \sum_{i=1}^{N} g_{i,c}}. \tag{2.11}$$

C denotes the possible classes, and N the number of pixels in an input image. A value of “1” represents a complete overlap between a prediction and a ground truth label. If there is no overlap, the dice coefficient returns “0”.

2.3 Decision Support Systems

In the following, we introduce decision support systems (DSSs). First, we provide a definition of the term DSS. Then, we describe the history and different classes of DSSs. Lastly, we present exemplary application areas of DSSs.

Power (2008, p. 149) defines a DSS as "a general term for any computer application that enhances a person or group's ability to make decisions". They can serve various purposes, e.g., improve consistency in decision-making, enforcement of policies, or distribution of expertise to non-expert staff (Power, 2002). An important characteristic is that in contrast to automated decision systems (Harris & Davenport, 2005), DSSs are intended to support skilled decision makers instead of replacing them (Power, 2002).

In the following, we briefly describe the history and different classes of DSSs. A graphical representation can be found in Figure 2.3 on page 25. The history of DSSs goes back to the late 1960s, with the first DSSs being implemented at that time (Power, 2007). Gorry and Scott Morton (1971) coined the term DSS in 1971 (Arnott & Pervan, 2005). The earliest DSSs are personal decision support systems (PDSSs) — designed and implemented to support the decision-making of a single manager or a small group of managers (Arnott & Pervan, 2005).
Over time, three types of DSSs evolved from PDSSs: intelligent decision support systems (IDSSs), executive information systems (EISs), and group support systems (GSSs) (Arnott & Pervan, 2014). IDSSs involve artificial intelligence techniques — either rule-based expert systems or data-based approaches like neural networks (Power, 2002). EISs are targeted at all levels of management despite the *executive* in the name; they provide an overview of information concerning the organizational goals (Arnott & Pervan, 2005). GSSs aim to support a group of people in their decision-making process; nowadays, they are typically deployed on the web (Turban et al., 2010). Negotiation support systemss (NSSs) are a special type of GSS, their goal is to enhance negotiation between two opposite parties (Arnott & Pervan, 2014).
Knowledge management-based decision support systems (KMDSSs) are successors of IDSSs — they are designed to assist knowledge storage, retrieval, transfer, and application (Arnott & Pervan, 2012).
EISs require large-scale data for decision support. Consequently, data warehousing (DW) was developed as an appropriate infrastructure. Business intelligence (BI) systems are often built on top of DW to facilitate decision-making (Turban et al., 2010). However, the term BI can be confusing since it lacks a clear definition and

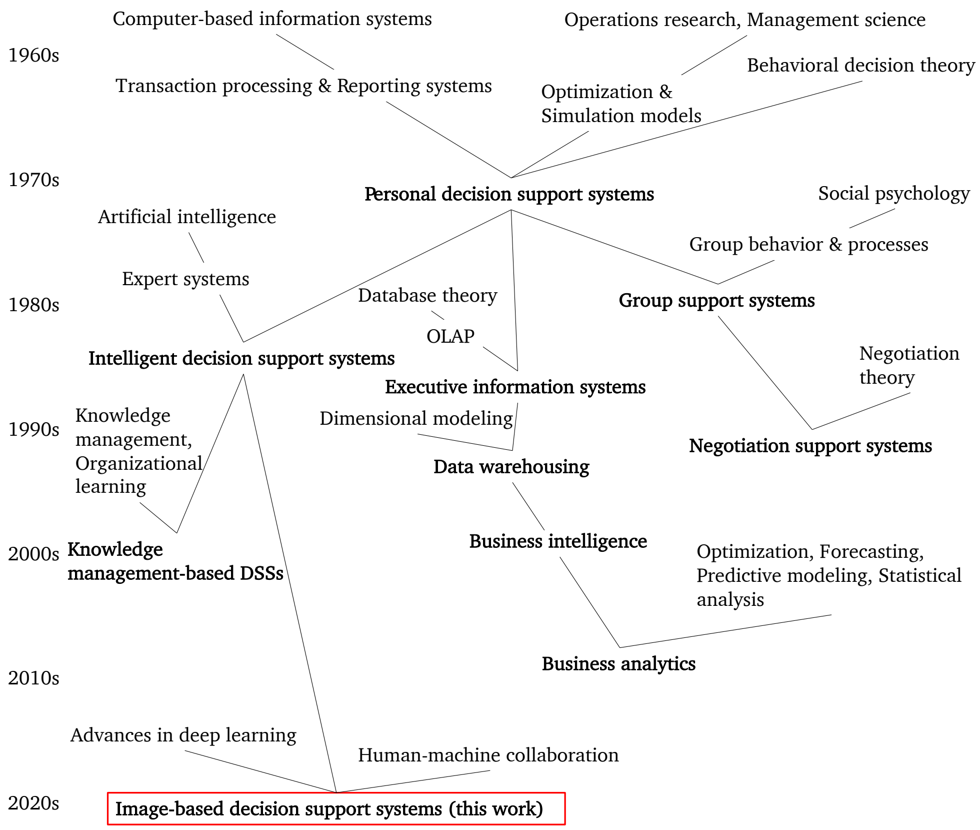


Fig. 2.3.: DSS genealogy. Own representation based on Arnott and Pervan (2014) and Schemmer (2020).

many companies use it for different commercial offerings in the field of decision support (Arnott & Pervan, 2005). (Business) Analytics, an umbrella term for data analysis applications, evolved from BI (Watson, 2014). Typically, analytics applications are distinguished into descriptive, predictive, and prescriptive ones (Davenport, 2013). Descriptive analytics report on the past, while predictive analytics is about predicting the future based on data from the past. Prescriptive analytics goes one step further by recommending optimal actions for the future.
This thesis introduces the novel class of *image-based decision support systems*; we will present them in Section 2.4.

DSSs are now developed and applied for many application areas. For example, in the banking industry, there are DSSs for credit decisions (Ignatius et al., 2018; Sachan et al., 2020). Also, DSSs support several financial fraud detection use cases. For example, Nasir et al. (2021) develop a DSS for a bank to detect and prevent misstatements and fraudulent actions of bank employees. Craja et al. (2020) develop a DSS for potential investors, auditing companies, and state regulators to detect financial statement fraud. In the health care domain, various clinical decision support systems exist, e.g., for severity risk prediction and triage of Covid-19 patients (Wu et al., 2020) and for diagnosing the heart disease status of patients (Fitriyani et al., 2020). Disaster prevention and management also benefits from DSSs — Ahmad and Simonovic (2006) develop a DSS for flood forecasting and management, Chang et al. (2022) develop a DSS for post-earthquake pedestrian evacuation, and Schätter et al. (2019) develop a DSS for supply chain risk management after disasters. Another interesting application area of DSSs is agriculture and forestry. Navarro-Hellín et al. (2016) develop a DSS to manage irrigation in agriculture by estimating the weekly irrigation needs of plantations. In the related field of forestry, the DSS of Wikström et al. (2011) enables, e.g., long-term forest level planning under different climate scenarios and estimation of recreation values. DSSs also play a major role for different logistics tasks like port logistics (Irannezhad et al., 2020) and ambulance relocation (Hajiali et al., 2022). Finally, DSSs are also used in manufacturing. Marcos et al. (2020) develop a DSS for improving the usage and maintenance of a heat-exchanger network used in the process industry. Guo et al. (2015) propose the usage of radio frequency identification in a DSS for production monitoring and scheduling.

2.4 Image-Based and Image-Mining-Based Decision Support Systems

In case the DSSs described above rely on data, they are primarily based on structured, tabular data. In contrast, image data is considered unstructured data. As described in Section 2.2, there have been significant advances in the field of DL for CV — as a result, accurate, automated processing of images is now possible. This facilitates the usage of images as a data source for DSSs. In the work at hand, this novel class of DSSs — image-based DSSs — plays a significant role. As displayed in Figure 2.3 on page 25, we see them as a subclass of IDSSs since artificial intelligence techniques, in particular, DL models, are used to convert images into processable information.
We distinguish between image-based decision support systems (IB-DSSs) and image-mining-based decision support systems (IM-DSSs) as described in Section 1.1 — IB-DSSs provide decision support based on single images while IM-DSSs rely on large amounts of data like multiple images and image mining. In recent years, the first works on IB-DSSs have been published. Chatterjee et al. (2018) design and implement a "vision-based DSS for road crack detection". It can be used to enhance

road infrastructure monitoring and maintenance. Chaudhuri and Bose (2020) utilize DL to identify earthquake survivors on image data from social media. Also, several IB-DSSs are developed in the medical field. For example, Ben-Cohen et al. (2017) develop an IB-DSS to assist human experts in the localization of the primary cancer sites in patients with liver metastasis.
Extant research on IM-DSSs is more scarce and even more focused on medical applications. Zaiane et al. (1998) is the only published IM-DSS outside the medical domain we are aware of. In their work, a domain-independent prototype for mining knowledge in image and video databases is developed. An example from the medical domain is Gatta et al. (2019): they develop an IM-DSS supporting the entire process of cancer patient treatment based on images.

In conclusion, to the best of our knowledge, none of the existing research provides guidance for the design of IB-DSSs and IM-DSSs. Hence, we see a research gap due to the great potential of IB-DSSs and IM-DSSs for value creation and the lack of existing research in this field. We hope this work advances the field by targeting this gap, that is, developing and evaluating generalizable design knowledge for IB-DSSs and IM-DSSs.

Part II

Technical Concepts

3 Towards Leveraging End-of-Life Tools as an Asset: Value Co-Creation based on Deep Learning in the Machining Industry[1]

3.1 Introduction

Sustainability is the key concept regarding the management of products having reached their end-of-life. Various approaches have been developed which suggest to implement sustainable end-of-life strategies already in the product development phase (Chan & Tong, 2007; Gehin et al., 2008; Rose et al., 1999). Such exemplary strategies range from refurbishing over remanufacturing to direct resale.

We argue that products having reached their end-of-life have additional value, which exceeds the material value, for provider and customer. Thus, these products should be considered an asset. They can be leveraged to gain insights into their usage. This, in turn, can be utilized to positively impact earlier stages of the value chain through value co-creation which involves manufacturer and customers.

Precisely, we propose to use worn tools from machining processes as a basis for easier and more objective optimization of customer's production processes. To this end, images of worn tools are automatically turned into valuable information by a deep-learning-based computer vision system. Information about occurrence, extent, and frequency of wear phenomena on the tools is usually the basis for understanding and improving machining processes. Due to the complexity of process optimization, tool manufacturers typically have dedicated teams of application engineers responsible

[1]This chapter comprises an article that was published as: Walk, J., Kühl, N., & Schäfer, J. (2020). Towards Leveraging End-of-Life Tools as an Asset: Value Co-Creation based on Deep Learning in the Machining Industry. *Proceedings of the 53rd Hawaii International Conference on System Sciences (HICSS-53)*. https://aisel.aisnet.org/hicss-53/da/bi_applications/2/. Note: The abstract has been removed. Minor edits have been made and tables and figures were reformatted, and newly referenced to fit the structure of the thesis. Chapter, section and research question numbering and respective cross-references were modified. Formatting and reference style was adapted and references were integrated into the overall references section of this thesis.

for supporting the customers in the optimization of their processes. They often rely on the visual inspection of worn tools to understand potential problems in a machining process. This, however, is usually done manually with small and non-representative samples. Our proposed system enables the automatic characterization of a large quantity of worn tools. This leads to more reliable and information-rich results and thus facilitates an easier and more objective process optimization. To maximize the real-world impact, scalability and generalizability of our proposed system, we formulate the following requirements:

Labelled training data should be the only required human input. As a consequence, the system can easily be trained for other tools or wear mechanisms. Also, the images for the testing and development of the system should be from real production processes.

In addition to enhancing process optimization, the insights based on our proposed system can also support the development of new tools. First, the development process itself can be accelerated since wear characterization is a frequent task in tool development and executed manually so far. Second, and more important, our system enables profound insights into potential problems of certain tools. So far, testing is mainly done internally and with standardized, simplified processes. With our proposed system it will be possible to analyze the wear mechanisms on a large quantity of tools used by customers in different real-world processes. This supports identifying promising directions for the development of new tools in the machining industry.

The remainder of this work is structured as follows: in Section 3.2, we present our research design. Subsequently, related work from different domains is introduced in Section 3.3. Based on this we then present our first, already completed, design cycle in detail in Section 3.4. In Section 3.5 we then present our agenda for future research. Afterwards, in Section 3.6, we summarize our work and describe limitations.

3.2 Research Design

As an overall research design, we choose design science research (DSR), as it allows to consider the theoretical and practical tasks necessary when designing IT artifacts (March & Smith, 1995) and has proven to be an important and legitimate paradigm in information systems research (Gregor & Hevner, 2013). For the design of the artifact, we follow the DSR process methodology and its individual phases according to Kuechler and Vaishnavi (2008), as we favor a clear differentiation between an abstract "suggestion" and a concrete, more programming-specific "development".

The work at hand presents the first DSR cycle as part of a larger research endeavor. Our overall goal is to assess the following general research question:

> **General Research Question A**
> How can we utilize end-of-life tools to improve processes at the interface of tool manufacturer and customer?

In the work at hand, we complete the first cycle with the individual phases as illustrated in Figure 3.1.

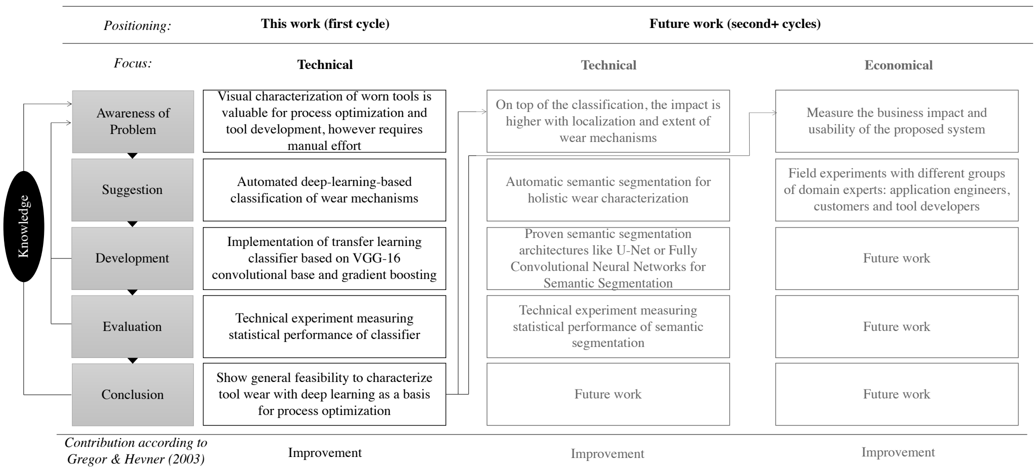


Fig. 3.1.: Overview of DSR cycles of the work at hand and the overall research endevour.

We ask the following specific research question:

> **Research Question A.1**
> How can we design a system for deep-learning-based computer vision to automatically classify worn tools regarding their wear phenomena?

This research question forms the basis of our overall research endeavor and allows to draw conclusions regarding the future steps of our reseach project.
In terms of knowledge contribution, the presented work of the first cycle depicts an “improvement” according to Gregor and Hevner (2013), since we apply a novel method, i.e., supervised machine learning with deep neural networks (LeCun et al., 2015), to the existing problem of worn tool classification. In order to evaluate the resulting artifact, we use a technical experiment as proposed by Peffers et al. (2012). We evaluate the statistical classification performances of the identified

models. Figure 3.1 on page 33 presents the activities of this first DSR cycle, as well as the future research activities containing two additional cycles, separated into the steps of problem awareness, suggestion, development, evaluation and conclusion. After elaborating on the state of the art of relevant fields for the research at hand in a designated rigor cycle (Hevner, 2007), we present all aforementioned steps for the first design cycle. Subsequently, we describe our research agenda for the second and third design cycle in Section 3.5.

3.3 Rigor Cycle and Related Work

To set a foundation for the remainder of this work, we review relevant literature from the body of knowledge. Several fields of research are of relevance, which we elaborate on in the following subsections: machining and wear mechanisms, deep learning and computer vision as well as value co-creation.

3.3.1 Machining and Wear Mechanisms

Machining is "one of the most important of the basic manufacturing processes" (Black, 1995, p. VI). It is applied in a variety of industries like aerospace, automotive, and the electro and energy industry. In general, machining describes the process of removing unwanted material from a workpiece (Black, 1995). The removal of unwanted material is generated by a relative motion between the cutting tool and the workpiece (Boothroyd & Knight, 1989). In regards to the different types of material, metallic workpieces are most widespread (Black, 1995). The tools used for machining can be regarded as consumables, as the occurrence of wear which ultimately results in a tool that can not be used anymore is inherent. For the first design cycle, we aim to show the general feasibility of our proposed system, therefore, we concentrate on the two main wear mechanisms we observed in our dataset: flank wear (82.56%) and chipping (55.40%) of the cutting edge. We will briefly describe those in the following.
Flank wear occurs due to friction between tool flank surface and workpiece (Altintas, 2012). It is unavoidable and thus the most commonly observed wear mechanism (Siddhpura & Paurobally, 2013). As a consequence it is regarded as good criterion for tool-life, i.e. for deciding when to change a tool (ISO - International Organization for Standardization, 1991). An exemplary image with flank wear is depicted in Figure 3.2 on page 35.

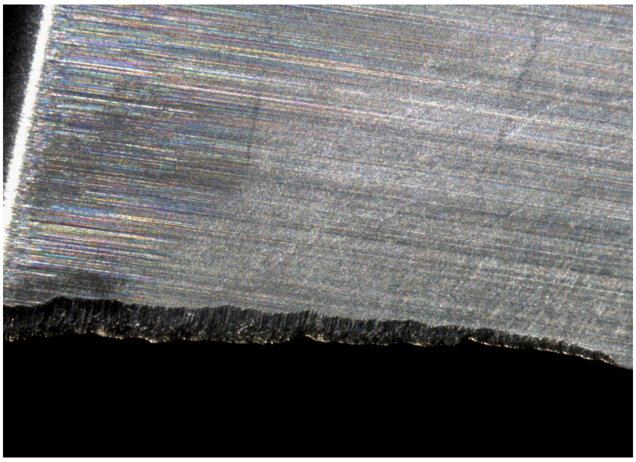


Fig. 3.2.: Example of flank wear.

Chipping refers to particles of the cutting edge breaking off and thermal cracking (ISO - International Organization for Standardization, 1991). This is less common and also less desirable since it suddenly deforms the cutting edge and leads to poor surface quality on the workpiece. Figure 3.3 shows an exemplary image with chipping.

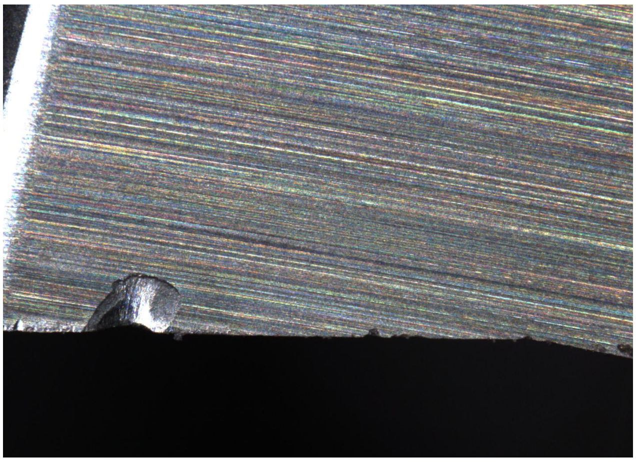


Fig. 3.3.: Example of chipping.

For the research at hand the application of image processing techniques for tool condition monitoring is the most related field of machining research. Tool condition monitoring based on image processing techniques means that an automatic visual inspection is used to determine the wear state of cutting tools. This enables to decide whether a tool can still be used or not. In the following we briefly present research from this field.
Dutta et al. (2013) provide a comprehensive review of the field of wear classification and measurement based on image processing. We briefly describe the papers

most relevant for our research: First, there is a multitude of research developing approaches for automatic wear measurement. Several articles describe systems for flank wear measurement for drills, which are based on traditional computer vision approaches (Duan et al., 2010; Y.-T. Liang & Chiou, 2006; Su et al., 2006). Traditional computer vision refers to approaches like texture-based image segmentation and edge detection for which the user needs to fine-tune a multitude of parameters (O'Mahony et al., 2019). Another common approach is to classify the extent of wear into different classes. For example, Alegre et al. (2009) use traditional computer vision algorithms (preprocessing like filtering and then automatic segmentation) to classify the flank wear on cutting inserts into low and high. Castejón et al. (2007) extract geometrical descriptors with traditional computer vision approaches and then use machine learning to classify if the wear on a given image is low, medium or high. Another stream of research works on classifying which wear mechanisms are visible on a given image. For instance, Schmitt et al. (2012) use traditional computer vision features (image statistics, surface texture, canny analysis, histogram and fourier coefficients) as input for a neural network which decides if the wear mechanism on the image is flank wear or tool breakage. Subsequently, they also apply an active contour algorithm to extract the wear region. In addition to the wear region, they compute the maximum and average wear perpendicular to the cutting edge. Lanzetta (2001) proposes a system to detect all types of wear on cutting inserts. Depending on the concrete tool, several parameters have to be chosen by the user of the system. Interestingly, this is the only identified article stating that the images are from cutting tools that were used in a real production environment—other articles either describe how they used the tools in their laboratory or do not elaborate on the environment.

Overall, we conclude that extant literature provides meaningful ideas for the development of an automated tool characterization system. However, none of the regarded research described above satisfies the requirements we formulated for our system in Section 3.1. Namely, that only labelled images are necessary as human input and that all images for testing and development of the system should be from real production processes. Furthermore, critically viewed, the performance of the systems developed in the articles above is often intransparent and hard to reproduce. Most datasets are rather small, e.g. Schmitt et al. (2012) use 15 images for the training of their neural network and 25 for testing. Also, they do not describe their dataset in detail—it is unclear on how many images which wear mechanisms are visible. Thus, it is not clear if they also worked with images where more than one wear mechanism is visible. Other papers rely on a purely visual evaluation based on concrete image examples (Duan et al., 2010; Lanzetta, 2001; Y.-T. Liang & Chiou, 2006).

3.3.2 Deep Learning and Computer Vision

Some of the systems for wear measurement and classification we just presented already use machine learning. However, they all rely on traditional computer vision approaches like edge detection to extract features from the raw images (Alegre et al., 2009; Castejón et al., 2007; Schmitt et al., 2012). With deep learning algorithms this becomes obsolete. Deep learning algorithms implement representation learning, i.e. they are able to directly process raw data and learn the relevant features for the task themselves (LeCun et al., 2015). Even more importantly, deep learning algorithms have been proven to achieve far better results than the previous state-of-the-art techniques in many computer vision applications (Krizhevsky et al., 2012; Voulodimos et al., 2018). Specifically, convolutional neural networks are applied for computer vision tasks. The first and main part of these networks consists of a series of convolutional and pooling layers (LeCun et al., 2015). In the convolutional layers filters are applied. These filters are learned from the data by backpropagation. Pooling layers "merge semantically similar features into one" (LeCun et al., 2015, p. 439), a typical application is to compute the maximum over e.g., nine pixels. In a given layer, the respective operations (convolution or pooling) are applied to all inputs from the previous layer. This drastically reduces the amount of weights to be learned compared to fully-connected networks where the weight is distinct for each connection of two neurons. Depending on the concrete computer vision application, the output is computed directly by a convolutional layer (Ronneberger et al., 2015) or by a series of fully-connected layers (Krizhevsky et al., 2012).

3.3.3 Value Co-Creation

With the relevant research from a technical perspective at hand, we now regard related work from a business perspective. Especially in the machining industry, the understanding of value has been mainly influenced by the goods-dominant logic: value is created (manufactured) by one firm and distributed in the market, usually through exchange of goods and money (Vargo et al., 2008, p. 146). Other industries like the software industry, in contrast, have already adopted the idea of service-dominant logic where "the roles of producers and consumers are not distinct, meaning that value is always co-created, jointly and reciprocally, in interactions among providers and beneficiaries through the integration of resources and application of competences" (Vargo et al., 2008, p. 146).

Several studies show that this value co-creation can be beneficial. For instance Nike, formerly also a product-centric company, successfully used a social networking site for co-creation with their customers. Among other benefits they also use the social networking site to learn about their customers' needs and preferences. Overall, they used the internet engagement platform "to establish customer relationships on a scale and scope as never before" (Ramaswamy, 2008, p. 10). On a more general level, Kale and Singh (2009) show that partnerships between companies generally help increasing firm value. In the remainder of this work, we take the perspective of service-dominant logic as well, as our general research question A refers to the creation of value at the interface between provider and customer.

The subfield of "reverse use of customer data" is even more closely related to our research. Saarijärvi et al. (2014) describe three cases how customer data can be turned into information that directly supports customers' value creation. We build on this research and extend it since the cases of Saarijärvi et al. (2014) rely on usage data as a basis for value creation. We, however rely on products without any usage data. In that sense our analysis is forensic. We do not have access to any usage data and can only rely on the tool having reached its end-of-life and the observations we can make directly from it.

3.3.4 Summary and Delineation

Based on the related work described above we believe we can contribute to the body of knowledge on different levels: Our proposed system addresses the lack of reproducibility and generalizability in existing research on automatic wear characterization and measurement based on image processing. First, we aim to ensure reproducibility by a detailed description of both our datasets and the computer vision systems. To the same end, we will use acknowledged machine learning evaluation techniques.

In regards to the missing generalizability, which we encountered in existing literature, we aim to utilize flexible, modern approaches. Existing research is based on traditional computer vision approaches, thus, a multitude of parameters need to be fine-tuned by the user. The recent developments in the area of deep learning facilitate end-to-end learning. Consequently, labelled training data is the only required human input for our proposed system based on deep learning. Thus, the system can be trained for other cutting tools or wear mechanisms without the need to fine-tune parameters.

From a business perspective our research contributes to the field of value co-creation and reverse use of customer data since it shows that these value creation mechanisms are also feasible based on forensic analyses.

3.4 First Design Cycle: Wear Mechanism Classification based on Deep Learning

So far, we completed the first cycle of our research endeavor, which we present in this chapter.

3.4.1 Awareness of Problem and Data Set

Visual characterization of worn tools is an essential part of the optimization of machining processes. We conducted interviews with domain experts to better understand their general approach for this optimization. Usually, visual characterization is done manually. The necessary effort leads to small and non-representative samples of worn tools. Due to the advances in the deep learning field described in Section 3.3.2, it seems plausible to apply deep learning for characterizing images of worn tools. Therefore, in the first cycle, we assess the feasibility of characterizing worn tools with deep learning. To be precise, we implement and evaluate two classification models: one for each of the two most prevalent wear mechanisms. We consider this a reasonable feasibility study since it gives an indication if and how deep learning algorithms are able to extract relevant features directly from the images in our dataset.

Our dataset consists of 648 images of worn cutting inserts from real production

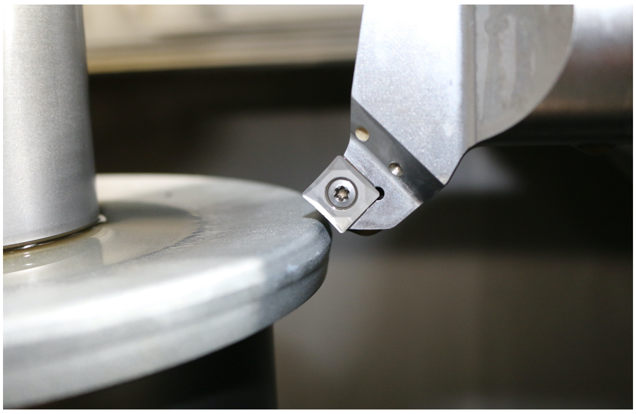


Fig. 3.4.: One of the two production processes.

processes on two different machines. The type of cutting insert is always the same for this first design cycle. Figure 3.4 shows one of the two production processes. The workpiece to the left rotates at high speed such that the cutting insert to the right removes unwanted material; during production the workpiece and cutting insert are

Tab. 3.1.: Frequency of wear mechanisms in our dataset.

Wear mechanism	Frequency (relative)
Flank wear	536 (82.72%)
Chipping	359 (55.40%)
No wear	96 (14.81%)
Built-up edge (Black, 1995)	90 (13.89%)

in direct contact. The images in our dataset show the flank side, i.e. the back side of the cutting edge.

To train and evaluate a classification algorithm we labelled the images manually. The first 60 images were labelled jointly by three domain experts. Afterwards labels were assigned individually, unclear cases were discussed by the three domain experts.

Several wear mechanisms are present on the images. Table 3.1 shows the absolute and relative frequency of different wear mechanisms. A cutting edge could show no wear, if e.g., wear on other parts of the cutting insert prevent a utilization. Note, that the data implies the presence of more than one wear mechanism on many pictures. Due to the scarcity of data for all other wear mechanisms, we only consider flank wear and chipping as wear mechanisms for our first design cycle.

Regarding the data, it is important to understand that the images depicted in Figure 3.2 and Figure 3.3 on page 35 are abnormally easy cases compared to the rest of the data set. Figure 3.5 shows a more representative image: both flank wear and chipping are present and the areas of chipping are relatively small.

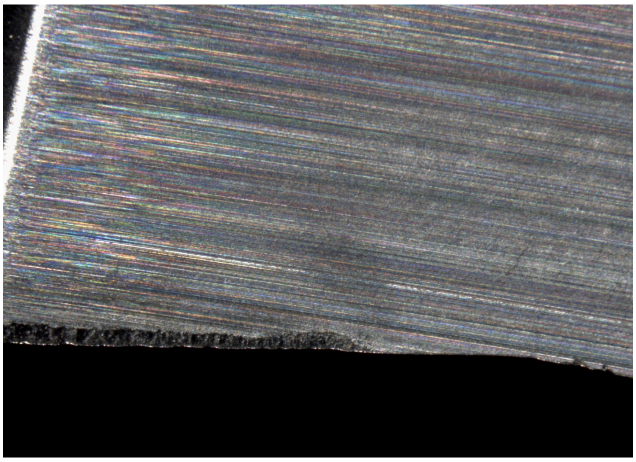


Fig. 3.5.: Example of combination of flank wear (left to middle) and chipping (right).

3.4.2 Suggestion

In this subsection, we address the classification tasks described above and explain the selection of certain options for the machine learning approach.

In general, deep neural networks require a large amount of data (e.g. millions of samples) to be trained (Oquab et al., 2014). However, in our case, relatively little data is available. For such cases, transfer learning has proven to be successful for other image classification tasks (Oquab et al., 2014; Shin et al., 2016). Transfer learning with deep neural networks refers to reusing the first part of a network which has been trained on a different task with a big data set. To be precise, one uses the already trained weights of the first layers of a neural network. These first layers perform feature extraction. Research has shown that these learned features can often be successfully transferred from one task to another (Oquab et al., 2014; Shin et al., 2016).

Consequently, we apply transfer learning in our first cycle. First, we use an already trained deep neural network to extract features from our images. Since this outputs a large quantity of features we need a feature selection mechanism. Thus, we apply a gradient boosting classifier that automatically performs feature selection as part of the classification (Friedman, 2011). To be precise, we apply the gradient boosting classifier to perform the classification regarding chipping and flank wear.

As evaluation strategy, we choose 3-fold cross-validation. Cross-validation is applied to use the whole data set and since it gives a good estimate for the error on unseen data (Friedman, 2011, p. 241). We choose only three folds due to the high runtime of the algorithms. As evaluation metric, we report the matthews correlation coefficient (MCC) (Matthews, 1975) since our dataset is imbalanced for both classification tasks. The MCC takes class imbalance into account — a MCC of “0” corresponds to random guessing based on the relative size of the classes. Perfect predictions yield an MCC of “1”, “-1” indicates that the predictions are inverse to the actual labels. Also, contrary to other popular evaluation measures like precision, recall and f-Measure the MCC also takes the true negatives into account. Thus, it gives a more holistic assessment of the classifier’s performance (Powers, 2007). Additionally, we report the confusion matrix to enable a more in-depth evaluation of the different types of correct and false predictions.

3.4.3 Development

In this subsection, we describe the implementation of our approach in the python programming language.
The images in our dataset have a size of 1600x1200 pixels. Since so far our computations are performed on a standard laptop we resized the images to 640x480 pixels to speed up computation. For the classification, we apply the transfer learning approach described in Section 3.4.2. Figure 3.6 shows an overview of the pipeline. We use the convolutional base of a VGG-16 network (Simonyan & Zisserman, 2014) to extract features from the raw images. The convolutional base comprises the first layers of a convolutional neural network, i.e. all layers apart from the fully-connected ones and the last softmax layer. The VGG-16 network which we apply, is pretrained on the `ImageNet` dataset (J. Deng et al., 2009). Specifically, we use the implementation from the `Keras` package (Chollet et al., 2015). At this point, the 153,600 features per image are saved to disk since the computation of these features is time consuming and independent of the concrete classification task. These

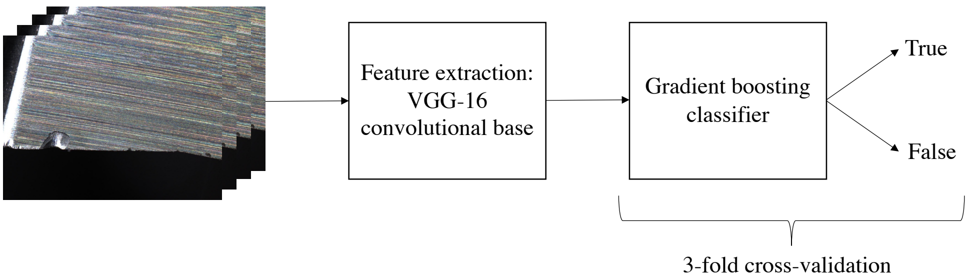


Fig. 3.6.: Overview of classification pipeline.

features are then used as input for two classification models. One with chipping and one with flank wear as binary target variable. The gradient boosting classifier implemented in the `scikit-learn` package (Pedregosa et al., 2011) is applied for these classification tasks.

3.4.4 Evaluation

After describing the implementation of the classification pipeline, we now present the results of the two classification models in terms of matthews correlation coefficient and confusion matrix.
The matthews correlation coefficient resulting from the flank wear classifier is 0.878. Table 3.2 on page 43 contains the corresponding confusion matrix.

Tab. 3.2.: Confusion matrix for the flank wear classifier.

	Predicted flank wear	**Predicted no flank wear**
Actually flank wear	532	4
Actually no flank wear	18	94

Tab. 3.3.: Confusion matrix for the chipping classifier.

	Predicted chipping	**Predicted no chipping**
Actually chipping	292	67
Actually no chipping	48	241

For the chipping classifier the matthews correlation coefficient is 0.644. The corresponding confusion matrix is shown in Table 3.3.

3.4.5 Conclusion

Our results show that it is possible to use deep learning to extract relevant features and perform classification regarding wear mechanisms based on our raw images. Keeping in mind, that a matthews correlation coefficient of "0" corresponds to random guessing based on the class sizes our results are significantly better. Discussions with domain experts confirmed that the approach is promising. The usefulness, however, can be increased when the location and extent of wear mechanisms are determined as well.

3.5 Future Cycles: Semantic Segmentation and Business Impact & Usability

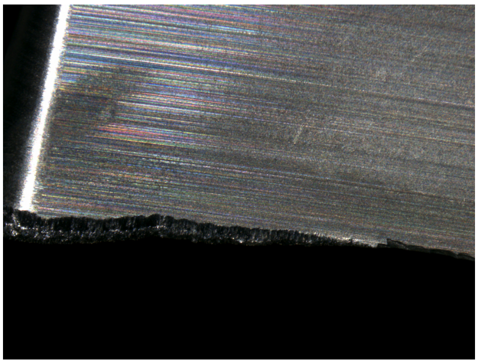
In this section we present the currently planned second and third cycle on a conceptual level. For easier reading we refrain from using the dedicated steps in the DSR cycle (awareness, suggestion etc.) in this section.

3.5.1 Second Design Cycle: Semantic Segmentation

In the first design cycle, we have shown that it is possible to use deep learning to extract relevant features for a classification regarding wear mechanisms based on our raw images. Discussing the results with domain experts, we learned that a more detailed characterization of images from worn tools would be beneficial. In detail, an exact identification of the location as well as the extent of wear phenomena would significantly increase the impact of our system. First, this enables statistics over certain wear phenomena. For instance, flank wear is a widespread tool life characteristic—the corresponding ISO Norm for turning (ISO - International Organization for Standardization, 1991) recommends 0.3 mm as tool life criterion. Measurements of flank wear on many tools from one process give an indication if the tools are changed too late, too early or just right. Second, heatmaps can be generated which show the locations of frequent wear. Accordingly, our research question for the second cycle is:

> **Research Question A.2**
> How can we design a system for deep-learning-based computer vision to automatically determine the location and extent of wear phenomena on images from worn tools?

In the following, we present how we propose to address this research question in the future. To extract the location and extent of wear mechanisms from the images, we propose a system for automatic semantic segmentation. The goal of semantic segmentation is to classify each pixel in a given image into a fixed set of categories (He et al., 2017). Figure 3.7 illustrates this by depicting an original image and the corresponding labels. So far, we are manually generating this pixelwise labelling as input for supervised learning. The goal of the second cycle is to automatically generate such labels.



(a) Raw image.

(b) Segmentation map: flank wear in light grey and chipping in white.

Fig. 3.7.: Example of raw image and corresponding segmentation map.

Research shows that deep convolutional neural networks are the best known approach for semantic segmentation; compare Ronneberger et al. (2015) for the U-Net architecture and Long et al. (2015) for the so-called fully convolutional networks for semantic segmentation. Accordingly, we will implement our system for the second cycle based on these network architectures. Since the pixelwise labelling is labor-intensive, we will explore if data augmentation techniques can help to reduce the number of required labelled images. Data augmentation refers to generating more training images by applying operations like shifting, rotating, flipping, distorting etc. to the original images (Perez & Wang, 2017). Previous research shows mixed results: Long et al. (2015) note that data augmentation does not help for their task, Ronneberger et al. (2015) describe data augmentation as an essential part of their approach.

We plan to evaluate this system for automatic semantic segmentation as follows. Training deep neural networks involves the optimization of hyperparameters such as the learning rate. Accordingly, we will split our data into three disjoint sets. The training set is used to learn the weights of the neural network, the validation set is used to find optimal hyperparameters, and the test set gives an estimate for performance on unseen data (Goodfellow et al., 2016).

Since our images contain a lot of background and unworn tool surface the choice of a proper evaluation measure is crucial. A popular and well-suited choice is the intersection over union (IoU) (Rahman & Wang, 2016). It is defined as

$$IoU = \frac{True\ Positives}{True\ Positives + False\ Positives + False\ Negatives}.$$

Thus, the intersection between the labelled area and the area predicted by the algorithm is divided by the union of these two areas, hence the name. Depending on the use case this measure can then be aggregated, e.g., over all pixels or over different classes of wear mechanisms.

3.5.2 Third Design Cycle: Business Impact & Usability

Whilst the first two design cycles focus on technical feasibility, implementation and statistical performance, the third cycle will focus on business impact and usability. Thus, the research question we address in the third cycle is:

> **Research Question A.3**
> How is the business impact and usability of the system for semantic segmentation perceived by the users?

In order to investigate this research question, we envision to examine two different scenarios which we describe in the following: First, the application of the system to improve process optimization. Second, the application of the system to optimize tool development.

Currently, customers of a tool manufacturer request an inspection of an application engineer in case they see optimization potential regarding their production process. Then the application engineer visits their production line and works on optimizing the production process. Thus, usually he[2] just looks at a small number of worn tools which are obtained during his visit or shortly before. Our proposed system enables an improved scenario: Again, a customer assuming optimization potential in a machining process requests a visit of an application engineer. He is then asked to collect all worn tools from the respective process for the next days/weeks. These worn tools are then sent to the application engineers and automatically analyzed by our proposed system. This has two major advantages. First, the application engineer receives the results of the wear characterization already before visiting the production line. This enables him to prepare better and to focus on the actual problem to be solved. Second, he gets deeper insights since the sample of worn tools is bigger and more representative which is even more important than the first advantage.

Thus, application engineers of a tool manufacturer are an important user group of our proposed system. Of course, we will also consider the customers of the application engineers. To ensure real-world impact, we will assess the business impact and usability of our proposed system in a field experiment: application engineers use the system for a certain time in their daily work. Afterwards, we interview both the application engineers and their customers regarding the business impact and usability of our proposed system.

[2]To ensure a steady reading flow in this work, we use only one gender and use male pronouns (he, his, him) when necessary. This always includes all genders.

Supporting tool development is another promising application of the system for semantic segmentation. When used with images from many different customers the proposed system can be utilized to understand inherent problems of certain tools and needs of the market. For example, if a certain tool suffers from severe chipping even though utilized at different customers on different material and with different process parameters, this is an important indication for the next generation of tools. Such an analysis can be another example for successful value co-creation: the development of tools tailored to the most prevalent problems in the market is only possible when information is shared between tool users and manufacturer. This is particularly promising since according to domain experts there is relatively little communication between tool manufacturers and companies using the tools. The proposed system could alleviate this problem. Often, the exchange of data or information between companies is restricted due to data confidentiality concerns (Hirt & Kühl, 2018). Worn tools are already sold from the companies using them to special recycling companies, thus they are not considered as confidential information. Consequently, tool developers are another important user group of our proposed system. We will perform a field experiment with this group to ensure real-world impact: they use our proposed system for a certain time and then we will interview them regarding the business impact and usability.

3.6 Conclusion

So far, research regarding the management of products having reached their end-of-life focuses on facilitating sustainable solutions like refurbishing and recycling instead of e.g., landfilling. In the work at hand, we propose an approach to generate additional value from products having reached their end-of-life. An exemplary use case in the machining industry illustrates how an automatic characterization of worn tools can foster value co-creation between tool manufacturer and the users of the tool. Both parties can benefit from easier and better process optimization and tool development.

There are four main contributions of this work: First, we summarize the state-of-the-art in automatic wear characterization on machining tools and show how such systems can be used beneficially apart from tool condition monitoring. Second, we show the feasibility of a deep-learning-based classification approach for different wear phenomena. With first results at hand, we, thirdly, present our agenda for future research. From a technical point of view, it will enable a complete characterization of worn tools including details like the exact location and extent of each wear phenomena. From a business point of view, it will evaluate the actual impact of the system. As fourth and more general contribution, we describe an example of how deep learning and products which have reached their end-of-life can be leveraged to positively impact earlier stages of the value chain. Thus, we argue that in certain cases products having reached their end-of-life should be considered an asset. This approach can be promising for further applications: experts in the respective domain confirm the potential usefulness of analyzing worn industrial seals. Furthermore, several applications in the business-to-consumer setting seem feasible: for instance, worn shoes could be analyzed to improve future generations of shoes.

Besides these contributions, this work has limitations. On a general level, parts of this paper are still conceptual. A more specific limitation regarding the process optimization and tool development use cases is that the survivorship bias (Brown et al., 1992) has to be kept in mind: in extreme cases, machining tools can break completely and customers will (probably) not send back these tools. Consequently, our proposed system cannot generate a complete overview of the wear mechanisms in real production processes. Another technical limitation is our (so far) limited consideration of only the flank of a worn cutting edge. Analyzing the other side (called face (ISO - International Organization for Standardization, 1991)) can also provide valuable information. However, domain experts confirmed the usefulness of an automatic characterization of the flank side. Thus, we believe this is a reasonable scope for now and leave this aspect for future work.

Overall, we believe this "cutting edge" research is a promising field of research. It has potential real-world impact and extends the research on value co-creation by showing possibilities based on forensic analyses of products having reached their end-of-life.

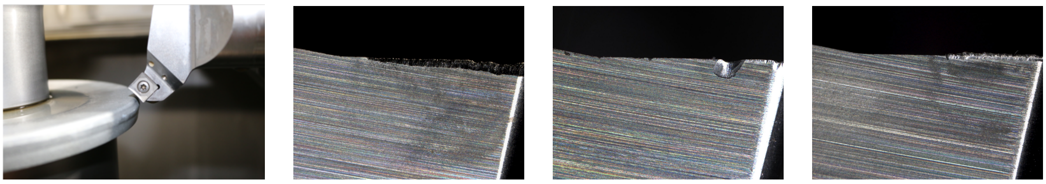
4 An Uncertainty-Based Human-in-the-Loop System for Industrial Tool Wear Analysis[1]

4.1 Introduction

Machining is an essential manufacturing process (Altintas, 2012), which is applied in many industries, such as aerospace, automotive, and the energy and electronics industry. In general, machining describes the process of removing unwanted material from a workpiece (Black, 1995). Thereby, a cutting tool is moved in a relative motion to the workpiece to produce the desired shape (Boothroyd & Knight, 1989). Figure 4.1a on page 50 displays an exemplary image of a machining process applying a *cutting tool insert*. Cutting tools are consumables because of the occurrence of wear on the tools, which ultimately results in unusable tools. In the following, we will briefly describe three common wear mechanisms, compare Figures 4.1b – 4.1d on page 50 for exemplary images. *Flank wear* occurs due to friction between the tools flank surface and the workpiece (Altintas, 2012). It is unavoidable and thus the most commonly observed wear mechanism (Siddhpura & Paurobally, 2013). *Chipping* refers to a set of particles breaking off from the tool's cutting edge (Altintas, 2012). A *built-up edge* arises when workpiece material deposits on the cutting edge due to localized high temperatures and extreme pressures (Black, 1995). Chipping and built-up edge are less desirable than flank wear since they induce a more severe and sudden deformation of the tool's cutting edge, leading to a reduced surface quality on the workpiece. Ultimately, this can lead to an increase of scrap components. A visual inspection of cutting tools enables an analysis of different wear mechanisms

[1] This chapter comprises an article that was published as: Treiss, A., Walk, J., and Kühl, N. (2020). An Uncertainty-Based Human-in-the-Loop System for Industrial Tool Wear Analysis. *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD 2020), (Part V), pp. 85–100.* https://link.springer.com/chapter/10.1007/978-3-030-67670-4_6. Note: The abstract has been removed. Minor edits have been made and tables and figures were reformatted, and newly referenced to fit the structure of the thesis. Chapter, section and research question numbering and respective cross-references were modified. Formatting and reference style was adapted and references were integrated into the overall references section of this thesis.

and provides insights into the usage behavior of cutting tools. Tool manufacturers, as well as tool end-users, can later leverage these insights to optimize the utilization of tools and to identify promising directions for the development of the next tool generations. Furthermore, an automated visual inspection enables the application of tool condition monitoring within manufacturing processes (Dutta et al., 2013). These analytics-based services possess a high economic value. Research suggests that tool failures are responsible for 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). Due to the relevance of these analytics-based sercives, our industry partner Ceratizit Austria GmbH, a tool manufacturer, agreed to closely collaborate within the research and implementation of an automated visual inspection for tool wear analysis.



(a) Process **(b)** Flank wear **(c)** Chipping **(d)** Built-up edge

Fig. 4.1.: Machining process and common wear mechanisms.

Numerous studies have examined wear analysis in the machining industry to address the need for automated wear analysis. However, the majority of the research to date focuses on traditional computer vision techniques (Walk et al., 2020). Since traditional computer vision approaches require the user to fine-tune a multitude of parameters (O'Mahony et al., 2019), scalability often becomes an issue. Contrary to traditional computer vision approaches, deep-learning-based approaches learn the required features themselves and can, therefore, be applied to different wear problems more efficiently. Another critical advantage of deep neural networks (DNNs) is their performance. In particular, the exploitation of convolutional neural networks (CNNs) has contributed to a performance increase in several computer vision tasks, e.g., classification and segmentation. CNNs are even able to surpass human-level performance in some of these settings (He et al., 2015), which demonstrates that a variety of visual tasks, previously performed by humans, can be automated using CNNs. Recently, Lutz et al. (2019) published the first work utilizing a CNN for wear analysis on cutting tool inserts, reporting promising results. However, CNNs, functioning as black-box systems, generally do not provide a reliable measure about the confidence of their decisions. This shortcoming is critical because the trustworthiness of a model's output remains unclear for a human supervisor. Two scenarios (J. D. Lee & See, 2004) can unfold: First, the model's capabilities can be underestimated, resulting in a disuse of the system. Secondly, the human supervisor can overestimate the model's capabilities, leading to a misuse. The correct balance between trust and distrust constitutes one of the main barriers for a successful adoption of CNNs in many real-world use cases (Dellermann, Ebel, et al., 2019). In particular, a measure of confidence is essential in safety-critical applications and in scenarios for which data is limited (Kendall & Gal, 2017). Limited amounts of data often occur in industrial problems, where resources and knowledge required to label and retrieve data are frequently restricted. In these settings, CNNs can occasionally produce sub-optimal results because they usually require a substantial amount of training data. Moreover, while performance can be high on average, within safety-critical applications, it is crucial to filter out any erroneous output. Lastly, due to the black-box property, CNNs are currently non-compliant with the ethics guidelines for trustworthy artificial intelligence by the european union (AI HLEG, 2019). In the future, these guidelines could translate into legislation that would limit the application of CNN-based systems in some industrial settings.

In this work, we address the need for CNN-based systems to output a confidence measure in an industrial environment. We consider the task of tool wear analysis using a unique, real-world data set from our industry partner Ceratizit. We employ an image segmentation algorithm based on the U-Net (Ronneberger et al., 2015) for the pixel-wise classification of three different wear mechanisms on cutting tool inserts. To increase transparency and performance, we further enhance the tool wear analysis system with capabilities to function as an uncertainty-based human-in-the-loop system. The suggested system aims at classifying the quality of a prediction, enabling the incorporation of a human expert. In particular, we estimate the quality of a prediction using the model's uncertainty. As a foundation for uncertainty, we apply monte carlo dropout (MC-dropout), which approximates a bayesian neural network (BNN) (Gal & Ghahramani, 2016). The approximated BNN outputs a probability distribution over the outputs. Based on the probability distribution, we apply multiple measures that aim at capturing the uncertainty of an output. Subsequently, we show that for the use case of wear analysis, there exists a significant linear correlation between the uncertainty measures and the performance metric, the dice coefficient. The linear relationship enables the utilization of the uncertainty measures as explanatory variables to predict the quality of a prediction in the absence of ground truth. We utilize these quality estimations in the following way: Predictions, which are estimated to be of high quality, are marked as successful by the system. These predictions are then passed on for automated analysis without any further human involvement. Otherwise, if an output is marked as failed, a human expert is requested to annotate the image manually. Hence, the system is introducing transparency by measuring the confidence of the predictions and is

furthermore increasing performance with the selective use of a human expert.
Overall, we see the following contributions of our work. While research is carried out within the field of medical imaging, no previous study, to the best of our knowledge, has investigated the use of uncertainty estimates in order to predict segmentation quality in an industrial setting. We contribute by showing how an uncertainty-based assessment of segmentation quality can be utilized in an industrial task of tool wear analysis. While Lutz et al. (2019) implement a CNN for tool wear analysis, we are additionally able to generate and leverage confidence estimates of the predictions. Besides the industrial relevance, our work also contributes on a more technical level. Most studies derive uncertainty estimates for binary classification problems, we are only aware of a study of Roy et al. (2018), which focuses on the task of deriving uncertainty measures in a multi-class image segmentation problem. Therefore, we contribute by deriving uncertainty measures for two further multi-class image segmentation problems. Additionally, we demonstrate, that a multiple linear regression can be applied to estimate segmentation quality in these multi-class segmentation problems. Regarding the challenge of estimating segmentation quality using uncertainty measures, we are only aware of DeVries and Taylor (2018), who use a DNN to predict segmentation quality. Within our use case, we rely on a multiple linear regression model, as it is interpretable, and also can be used in scarce data settings. Additionally, researchers press for more insights on human-in-the-loop systems (Brynjolfsson & McAfee, 2011), as successful designs are still scarce. Especially the allocation of (labeling) tasks between humans and machines is under-researched (Dellermann, Ebel, et al., 2019). We contribute by implementing our human-in-the-loop system and evaluating it by a simulation study. To ensure generalizability, we assess the use of our approach on the publicly available Cityscapes (Cordts et al., 2016) dataset for urban scene understanding.

4.2 Foundations and Related Work

First, we shortly introduce the motivation to use MC-dropout as an approach to estimate uncertainty. Subsequently, we present selected related studies, which focus on assessing the quality of a predicted segmentation by the use of uncertainty estimates.
There is a considerable body of literature growing around the theme of uncertainty estimation in DNNs. In classification tasks, a softmax output displays the probability of an output belonging to a particular class. Thus, softmax outputs are occasionally used to represent model uncertainty (Hendrycks & Gimpel, 2016). However, as

illustrated by Gal and Ghahramani (2016), a model can be uncertain even with a high softmax output, indicating that softmax outputs do not represent model uncertainty accurately. Contrary to traditional machine learning approaches, a bayesian perspective provides a more intuitive way of modeling uncertainty by generating a probability distribution over the outputs. However, inference in BNNs is challenging because the marginal probability can not be evaluated analytically, and, therefore, inference in BNNs is computationally intractable. Nevertheless, in a recent advance, Gal and Ghahramani (2016) show that taking monte carlo samples from a DNN in which dropout is applied at inference time approximates a BNN. In a study by Kendall and Gal (2017), the authors show that these approximated BNNs lead to an improvement in uncertainty calibration compared to non-bayesian approaches. Since dropout exists already in many architectures for regularization purposes, MC-dropout presents a scalable and straightforward way of performing approximated bayesian inference using DNNs without the need to change the training paradigm.
For the human-in-the-loop system, we are particularly interested in the prediction of segmentation quality based on uncertainty estimates. To date, several studies on this particular topic have been conducted within the field of medical imaging. DeVries and Taylor (2018) use MC-dropout as a source of uncertainty to predict segmentation quality within the task of skin lesions segmentation. A separate DNN is trained to predict segmentation quality based on the original input image, the prediction output, and the uncertainty estimate. While the subsequent DNN achieves promising results in predicting segmentation quality, the subsequent model lacks transparency and explainability itself. In particular, the subsequent model does not provide any information why a prediction failed. Furthermore, a DNN is only applicable if a considerable amount of data is available. Nair et al. (2020) utilize MC-dropout to explore the use of different uncertainty measures for multiple sclerosis lesion segmentation in 3D MRI sequences. The authors show that for small lesion detection, performance increases by filtering out regions of high uncertainty. While the majority of studies focus on pixel-wise uncertainties, there is a need to aggregate uncertainty on whole segments of an image. These aggregated structure-wise uncertainty measures allow an uncertainty assessment on an image-level. The work of Roy et al. (2018) introduces three structure-wise uncertainty measures, also based on MC-dropout, for brain segmentation. While the authors show that these uncertainty measures correlate with prediction accuracy, the work does not display how these uncertainty measures are applicable in a broader context, e.g., in a human-in-the-loop system.

4.3 Methodology

On the basis of the previously presented foundations, we now introduce our applied methodology. In Section 4.3.1, we shortly introduce foundations regarding image segmentation. Subsequently, in Section 4.3.2, we present the modified U-Net architecture, which we use to approximate a BNN. Section 4.3.3 then describes the loss function and the evaluation metric, which we use to train and evaluate the modified U-Net. Lastly, Section 4.3.4 depicts the computation of the uncertainty measures on which our human-in-the-loop system relies.

4.3.1 Image Segmentation

For our use case of wear analysis in the machining industry, information must be available on a pixel level to facilitate the assessment of location and size of wear for a given input image. An approach that provides this detailed information is called *image segmentation*. While image segmentation can be performed unsupervised, it is often exercised as a supervised learning problem (Garcia-Garcia et al., 2017). In supervised learning problems, the task is generally to learn a function $f: X \to Y$ mapping some input values (X) to output values (Y). In the case of image segmentation, the concrete task is to approximate a function f, which takes an image x as input and produces a segmentation $\hat{y}$. The predicted segmentation assigns a category label $c \in C$ to each pixel $i \in N$ in the input image, where C denotes the possible classes and N the number of pixels in an input image. Therefore, the task of image segmentation is also referred to as pixel-wise classification. Consequently, the outputs must have the same height and width as the input image, the depth is defined by the number of possible classes.

4.3.2 Model: Dropout U-Net

We apply a modified U-Net architecture due to its ability to produce good results even with a small amount of labeled images (Ronneberger et al., 2015). To avoid overfitting and increase performance, we implement the following adaptions to the original U-Net architecture: We use an L2-regularization in every convolutional layer, and additionally, we reduce the number of feature maps in the model's architecture, starting with 32 feature maps instead of 64 feature maps in the first layer (Bishop, 1995). The number of feature maps in the remaining layers follows the suggested approach from the original U-Net architecture, which doubles the number of feature maps in the contracting path and then analogically halves the number of feature maps in the expansive path. To accelerate the learning process, we add batch normalization between each pair of convolutional layers (Ioffe & Szegedy, 2015). We use a softmax activation function in the last layer of the model to obtain predictions in the range [0,1]. Therefore, the modified U-Net takes an input image x and produces softmax probabilities $p_{i,c} \in [0,1]$ for each pixel $i \in N$ and class $c \in C$, compare equation (1).

$$p_{i,c} = f(x) \qquad \forall i \in N, \forall c \in C \tag{4.1}$$

As a source of uncertainty, we realize the human-in-the-loop system based on MC-dropout because of its implementation simplicity, while still being able to generate reasonable uncertainty estimates (Gal & Ghahramani, 2016). We employ dropout layers in the modified U-Net to enhance the model with the ability to approximate a BNN (Gal & Ghahramani, 2016). Units within the dropout layers have a probability of 0.5 to be multiplied with zero and therefore, to drop out. We follow the suggested approach by Kendall et al. (2015) in the context of the Bayesian SegNet to use dropout layers at the five most inner decoder-encoder blocks. Hereinafter, we will refer to the modified U-Net, which applies dropout at inference time, as the *Dropout U-Net*. By the application of dropout at inference time, the Dropout U-Net constitutes a stochastic function f. For multiple stochastic forward passes T of an input image x, the Dropout U-Net generates a probability distribution for each $p_{i,c}$. To obtain a segmentation $\hat{y}$, we calculate the mean softmax probability first, as described in equation (4.2). Then, the segmentation $\hat{y}$ is derived by applying the argmax function over the possible classes of the softmax probabilities $p_{i,c}$, compare equation (4.3).

$$p_{i,c} = \frac{1}{T}\sum_{t=1}^{T} p_{i,c,t} \qquad \forall i \in N, \forall c \in C \tag{4.2}$$

$$\hat{y}_i = \underset{c \in C}{\operatorname{argmax}}(p_{i,c}) \qquad \forall i \in N \tag{4.3}$$

4.3.3 Loss Function and Performance Evaluation Metric

As a loss function, we use a *weighted category cross-entropy* loss. We weight each class with the inverse of its occurrence (pixels) in the training data due to class imbalance. This leads to an equal weighting between the classes in the loss function (Crum et al., 2006). The weighted categorical cross-entropy loss is defined in equation (4.4); $g_{i,c}$ denotes the one-hot-encoded ground truth label, and w_c the computed weights for each class.

$$\mathcal{L}(p, g) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} w_c \; g_{i,c} \; \log \; p_{i,c} \tag{4.4}$$

To reflect the quality of a predicted segmentation, we rely on the *dice similarity coefficient (DSC)* (Dice, 1945) as a performance evaluation metric. The dice coefficient assesses the overlap, or intersection, between the model's outputs and the one-hot-encoded ground truth labels $g_{i,c}$. A full overlap between a prediction and a label is represented by a value of one. If there is no overlap, the dice coefficient returns zero. Representing the outputs, one could use the softmax probabilities $p_{i,c}$ or the binarized one-hot encoded predictions $\hat{y}_{i,c}$. We use the binarized predictions $\hat{y}_{i,c}$ since these predictions represent the foundation for the subsequent tool wear analysis. As suggested by Garcia-Garcia et al. (2017) for the related jaccard-coefficient, we compute the dice coefficient for each class separately. To assess the segmentation quality of an input image, we compute the averaged dice coefficient across all classes, leading to a *mean dice coefficient*, defined in equation (4.5). To evaluate a model on the test set, we calculate the average across the mean dice coefficients per image. Next to the dice coefficient, we also compute the pixel accuracy as a performance measure. It defines the percentage of correctly classified pixels.

$$Mean\ Dice\ Coefficient = \frac{2}{C}\sum_{c=1}^{C}\frac{\sum_{i=1}^{N}\hat{y}_{i,c}\; g_{i,c}}{\sum_{i=1}^{N}\hat{y}_{i,c} + \sum_{i=1}^{N} g_{i,c}} \tag{4.5}$$

4.3.4 Uncertainty Estimation

As a next step, we describe how uncertainty measures can be calculated using the probability distribution outputs of the Dropout U-Net. In general, the information-theory concept of entropy (Shannon, 1948) displays the expected amount of information contained in the possible realizations of a probability distribution. Following previous work (Kendall & Gal, 2017), we utilize the entropy as an uncertainty measure to reflect the uncertainty of each pixel in a predicted segmentation:

$$H(p_i) = -\sum_{c=1}^{C} p_{i,c} \, \log \, p_{i,c} \qquad \forall i \in N \tag{4.6}$$

The entropy displays its maximum if all classes have equal softmax probability and it reaches its minimum of zero if one class holds a probability of 1 while the other classes have a probability of 0. Therefore, the entropy reflects the uncertainty of a final output $\hat{y}_i$ by considering the model's outputs $p_{i,c}$ over all classes C. For the task of image segmentation, the entropy is available per pixel. However, for several applications, it is necessary to derive an uncertainty estimate on a higher aggregation level. For example, within the tool wear analysis, we want to decide on an image basis, whether a segmentation is successful or failed. One approach is to average the pixel-wise entropy values over an image to come up with an image-wise uncertainty estimate. Roy et al. (2018) propose to calculate the average pixel uncertainty for each predicted class in a segmentation. We utilize this idea in the context of wear analysis and define the entropy per predicted class U_c in equation (4.7). Equation (4.8) defines the number of pixels for each predicted class. The entropy per predicted class provides information on an aggregated class-level and can be used to estimate uncertainty for each class in an input image. Since only the predictions are used, this uncertainty estimate is applicable in the absence of ground truth.

$$U_c = \frac{1}{S_c} \sum_{i=1}^{S_c} H(p_i) \qquad \forall c \in C \tag{4.7}$$

$$S_c = \{ i \in N \mid \hat{y}_i = c \} \qquad \forall c \in C \tag{4.8}$$

4.4 Experiments

With the methodology at hand, the upcoming Section 4.4.1 provides information about the two datasets, the preprocessing and training procedures. Subsequently, Section 4.4.2 briefly describes the performance results in terms of the dice coefficient.

In Section 4.4.3, we evaluate the uncertainty-based human-in-the-loop system in the following way: First, we illustrate the relation between uncertainty and segmentation quality using two exemplary predictions. Then, we quantitatively assess the relation between uncertainty and segmentation quality using the bravais-pearson correlation coefficient. Next, we fit a multiple linear regression on each test set, which uses the uncertainty measures as independent variables to explain segmentation quality. Lastly, we simulate the performance of the uncertainty-based human-in-the-loop system based on the multiple regression and compare it against a random-based human-in-the-loop system.

4.4.1 Datasets, Preprocessing and Training Procedure

The unique **Tool wear dataset** consists of 213 pixel-wise annotated images of cutting tool inserts, which were previously used by Ceratizit's customers in real manufacturing processes until their end-of-life. The labels are created as follows: The first 20 images are labeled jointly by two domain experts from Ceratizit. Afterwards, labels are assigned individually, whereas at least two domain experts discuss unclear cases. The recording of the images is standardized to reduce the required amount of generalization of the learning algorithm. The images initially have a resolution of 1600×1200 pixels. As a preprocessing step, we cut the image to a shape of 1600×300, which lets us focus on the cutting edge where the wear occurs. For computational efficiency and as a requirement of the U-Net architecture, we resize the images to a shape of 1280×160 using a bilinear interpolation. We randomly split the dataset into 152 training, 10 validation and 51 test images. The model is trained for 200 epochs, using an Adam optimizer, a learning rate of 0.00001, an L2-regularization with an alpha of 0.01, and a batch size of 1. We choose the hyperparameters after running a brief hyperparameter search. The training is conducted on a Tesla V100-SXM2 GPU, with a training duration of approximately two hours. In the literature, the number of conducted forward passes ranges from 10 to 100 (DeVries & Taylor, 2018), we use 30 monte carlo forward passes to create the probability distribution over the outputs.

The **Cityscapes dataset** (Cordts et al., 2016) is a large-scale dataset, which contains images of urban street scenes from 50 different cities. It can be used to assess the performance of vision-based approaches for urban scene understanding on a pixel-level. The Cityscapes dataset initially consists of 3475 pixel-wise annotated images for training and validation. Performance is usually specified through 1525 test set images for which ground truth labels are only available on the Cityscapes website (Cordts et al., 2016). Since we need ground truth labels to assess the uncertainty estimates, we use the 500 proposed validation images as the test set and randomly split the remaining 2975 images into 2725 training and 250 validation images. There are initially 30 different classes which belong to eight categories. We group classes belonging to the same category together, to create a similar problem setting between the Cityscapes dataset and the tool wear dataset. Furthermore, we combine the categories 'void', 'object', 'human', and 'nature' to one class, which we consider in the following as the background class. The remaining categories are 'flat', 'construction', 'sky' and 'vehicle', ultimately resulting in five different classes. The original images have a resolution of 2048×1024 pixels. During training, an augmentation step flips the images horizontally with a probability of 0.5. Then, a subsequent computation randomly crops the images with a probability of 0.15 to an input size of 1024×512. Otherwise, the images are resized using a bilinear interpolation to the desired input shape of 1024x512. Following a brief hyperparameter search, we train the model for 15 epochs with a learning rate of 0.0001 using an Adam optimizer, a batch size of 10, and an L2 weight regularization with an alpha of 0.02. We train the model on a Tesla V100-SXM2 GPU for approximately 2.5 hours. The Dropout U-Net uses five monte carlo forward passes, considering the higher computational complexity due to the more extensive test set and the large image size.

4.4.2 Performance Results

Next, we assess the quality of a predicted segmentation in terms of the dice coefficient. Table 4.1 on page 60 displays the performance results on each respective test set. The background class has the largest proportion of pixels (93.9%) in the tool wear dataset, followed by flank wear (4.7%), built-up edge (0.9%) and lastly, chipping (0.4%). We assume that the amount of labeled pixels of a specific class is closely related to prediction performance. We find, that the model has particular difficulties segmenting chipping phenomena, which is expressed by a dice coefficient of 0.244. We explain this lack of prediction performance by the punctual and minor occurrence of chipping phenomena within images, compare Figure 4.2 on page 61, and the characteristic of the dice coefficient. In particular, the dice coefficient per class drops to zero, if the model produces a false negative, meaning the model falsely predicts a wear mechanism, and if there is no wear for the corresponding class labeled in the image. This characteristic and the challenging task of classifying small chipping phenomena in an input image causes the dice coefficient of the chipping class to drop to zero for several images. In contrast to chipping, the Dropout U-Net recognizes the background class well and is classifying flank wear and built-up edge

Tab. 4.1.: Performance results.

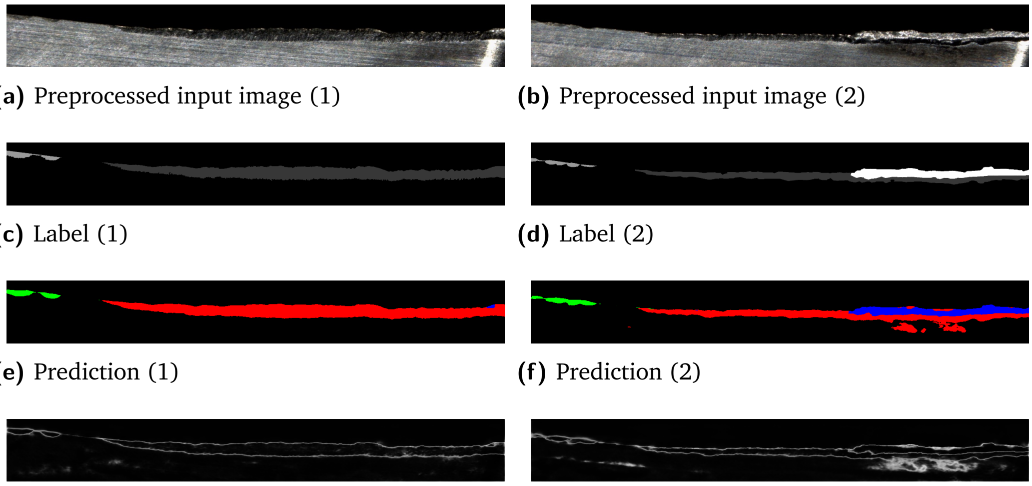
Tool wear dataset		Cityscapes dataset	
Class	Dice coefficient	Class	Dice coefficient
Background	0.991	Background	0.929
Flank wear	0.695	Flat	0.693
Chipping	0.244	Construction	0.830
Built-up edge	0.596	Sky	0.769
		Vehicle	0.773
Mean DSC	0.631	Mean DSC	0.799
Pixel accuracy	0.977	Pixel accuracy	0.875

considerably well. Compare the label (Figure 4.2c) and the prediction (Figure 4.2e) on page 61 for an illustration. The performance results of the Cityscapes dataset also indicate that the Dropout U-Net can generate a predicted segmentation for each class considerably well.

4.4.3 Evaluation: Uncertainty-Based Human-in-the-Loop System

With the performance results at hand, we focus on assessing the uncertainty-based human-in-the-loop system. Figure 4.2 on page 61 presents two preprocessed images, their corresponding human labels, their predictions, and the generated uncertainty maps for the tool wear dataset. The uncertainty maps are generated using pixel-wise uncertainties based on the entropy, compare equation (4.6). Within the uncertainty maps, brighter pixels represent uncertain outputs, and darker pixels represent certain predictions. The uncertainty map (g) of the left input image (a) displays uncertain outputs, indicated by brighter pixels, at the edge between classes. This behavior is often noticed within uncertainty observation in image segmentation tasks (Kendall et al., 2015), as it reflects the ambiguity of defining precise class regions on a pixel-level. The most interesting aspect of Figure 4.2 on page 61 is the prediction (f) and the corresponding uncertainty map (h). The model falsely predicted several areas on the right hand side as flank wear (red). However, the model also indicates high uncertainty for this particular area, indicated by brighter pixels in the corresponding uncertainty map (h).

This relationship is essentially the foundation of the uncertainty-based human-in-the-loop system. The relationship between a pixel's uncertainty and the probability of being classified correctly enables the system to distinguish between images which the model segments successfully and images which the model segments poorly.



(a) Preprocessed input image (1) **(b)** Preprocessed input image (2)

(c) Label (1) **(d)** Label (2)

(e) Prediction (1) **(f)** Prediction (2)

(g) Entropy per pixel (1) **(h)** Entropy per pixel (2)

Fig. 4.2.: Illustration of two images of the test set with their corresponding labels, predictions and uncertainty maps (best viewed in color).
Color coding: Flank wear = dark grey/red, Chipping = light grey/green, and Built-up edge = white/blue.

As the goal is to distinguish segmentation quality on an image level, the average uncertainty of a predicted segmentation can be used as an uncertainty measure for an image. However, we find that for the task of tool wear, there does not exist a significant correlation (-0.34) between the averaged entropy per prediction and the mean dice coefficient on the 51 images of the test set. The same analysis yields a correlation of -0.57 for the Cityscapes dataset. However, we find that the mean entropy of a predicted class, U_c, is highly correlated with the corresponding dice coefficient per class. Table 4.2 on page 62 displays the bravais-pearson correlation between the entropy of a predicted class and the corresponding dice coefficient per class on the respective test sets. In the case of the tool wear dataset, the linear relationship is especially strong for the classes flank wear, chipping and built-up edge, while it is slightly weaker for the background class. These results are reproducible on the Cityscapes dataset. As can be seen from Table 4.2 on page 62, correlations, besides the background class, range from -0.751 to -0.878, indicating a strong negative linear relationship. We use ordinary least squares to fit a multiple linear regression on the test set for both datasets using the uncertainties per predicted class U_c as independent variables and the mean dice coefficient as the dependent variable. Subsequently, the linear regression aims at quantifying the prediction quality on an image-level in the absence of ground truth. We find that, for both datasets, the independent variable 'uncertainty per background class' is statistically not significant at the 0.05 value. Therefore, we discard it as an independent variable from the multiple regression model. The remaining independent variables are significant at

Tab. 4.2.: Correlation coefficients between the uncertainty per predicted class (U_c) and the dice coefficients per class.

Tool wear dataset	
Class	Correlation
Background	-0.656
Flank wear	-0.911
Chipping	-0.818
Built-up edge	-0.932

Cityscapes dataset	
Class	Correlation
Background	-0.208
Flat	-0.878
Construction	-0.754
Sky	-0.751
Vehicle	-0.858

the 0.01 level for both datasets. The regression results yield a $R^2 = 0.718$ for the Tool wear dataset and a $R^2 = 0.655$ for the Cityscapes dataset. This indicates that the multiple linear regression can explain a substantial amount of variation of the mean dice coefficient. The full regression results can be found in Table 4.3.

In the following paragraph, we assess the use of the multiple linear regression in the

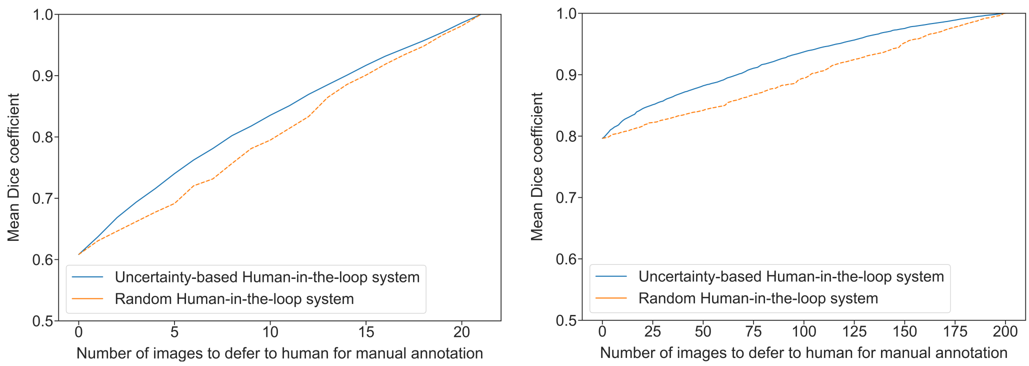
Tab. 4.3.: Regression results.

Tool wear dataset	
Dependent variable: Averaged Dice	
Const	0.922***
Background	
Flank wear	-0.165***
Chipping	-0.099***
Built-up edge	-0.169***
Observations	51
R2	0.718
Adjusted R2	0.7
Residual Std. Error	0.066
F Statistic	39.849***

Cityscapes dataset	
Dependent variable: Averaged Dice	
Const	1.206***
Background	
Flat	-0.172***
Construction	-0.203***
Sky	-0.131***
Vehicle	-0.132***
Observations	500
R2	0.655
Adjusted R2	0.653
Residual Std. Error	0.078
F Statistic	235.335***

Note: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

context of a human-in-the-loop system by running a simulation. The multiple linear regression predicts the quality of a predicted segmentation in terms of the mean dice coefficient, using the uncertainties per predicted class as independent variables. Then, in an iterative process, the input image, for which the prediction displays the lowest estimated mean dice coefficient is forwarded for human annotation. Images, displaying a higher estimated dice coefficient are retained by the system. Within the simulation, we assume a perfect human segmentation, and set the corresponding dice coefficient of the forwarded image to one. The performance of the system is then calculated by combining the mean dice coefficients of the retained images and the forwarded human annotated images. Figure 4.3 shows the simulation results for both datasets. The x-axis displays the number of images, which are forwarded to human annotation. The y-axis displays the performance of the system. We compare the performance of the human-in-the-loop system (blue line) against a random-based human-in-the-loop sytem (orange line). Contrary to the uncertainty-based system, the random-based system decides randomly, which images are forwarded to human annotation. To avoid overfitting, we use a split for each dataset as follows: For the tool wear dataset, the multiple linear regression is fitted on 30 images, the remaining 21 predictions are then used for simulation. Within the Cityscapes dataset, we use 300 test set images to fit the regression and 200 images for simulation. For both datasets, the uncertainty-based human-in-the-loop system is able to achieve a better mean dice coefficient using less human annotations than a random-based approach. This is due to the multiple regression, which identifies low-quality predictions and therefore enables the system to forward these predictions to human annotation first.



(a) Tool wear dataset

(b) Cityscapes dataset

Fig. 4.3.: Simulation results.

4.5 Discussion and Outlook

In this work, we show the applicability and usefulness of an uncertainty-based human-in-the-loop system for the task of industrial tool wear analysis. The human-in-the-loop system addresses critical challenges regarding the adoption of CNNs in industry. In particular, it increases transparency by providing uncertainty measures which are correlated with segmentation performance. Additionally, it improves performance by incorporating a human expert for the annotation of images that are estimated to be of low quality.

Within the use case of tool wear analysis, we consider the task of segmenting three different wear mechanisms on cutting tool inserts. We apply and train a modified U-Net architecture on a real-world dataset of our industry partner Ceratizit, achieving good performance results. For the human-in-the-loop system, we enhance the existing tool wear analysis with the following capabilities: We implement MC-dropout and use the information-theory concept of entropy to compute pixel-wise uncertainties. Furthermore, we aggregate the pixel-wise uncertainties to compute class-wise uncertainty measures on an image-level. A multiple linear regression reveals that the class-wise uncertainties can be used as independent variables to explain a substantial amount of the mean dice coefficient of an image. The multiple linear regression is then leveraged within the human-in-the-loop system to decide, whether a given segmentation should be forwarded to a human expert, or be retained in the system as a successful prediction. A simulation study demonstrates that the performance improves through the utilization of a human expert, which annotates estimated low-quality predictions. Furthermore, the system increases transparency by additionally issuing an estimate about the quality of a prediction.

We assess our system not only on our proprietary tool wear data set but also on the publicly available and substantially larger Cityscapes data set, confirming the generalizability of our approach to the task of urban scene understanding. Nevertheless, we consider the application to only two data sets as a limitation of the study. In the future, we aim at validating the system on additional datasets. Another promising avenue for future research is to further distinguish uncertainty into epistemic (model) and aleatoric (data) uncertainty (Kendall & Gal, 2017). While aleatoric uncertainty is due to inherent noise in the input data, e.g., a blurred image, epistemic uncertainty occurs due to model uncertainty, e.g., lack of training data. Within human-in-the-loop systems, this distinction can lead to more informed decisions, e.g., when images are forwarded to a human expert, a possible cause for a failed prediction can be provided. Regarding uncertainty estimation, further research is also needed on a more theoretical level, to establish a more profound understanding of uncertainty outputs of different approaches and their relation to prediction quality. This could include a structured comparison of different ways to calculate uncertainty across different use cases. Lastly, from a human-centric machine learning standpoint, further research should assess, if the increased transparency of the human-in-the-loop system leads to a more calibrated level of trust from the user. While we found several indications in the literature, few studies have investigated this relationship in a systematic way.

We see a broad applicability of the uncertainty-based human-in-the-loop system in industrial applications. While we consider the task of image segmentation, the general observations should be relevant in a variety of supervised learning problems. A human-in-the-loop system can be beneficial for all types of automation tasks, in which human experts display superior performance than automated systems, but in which the automated system is more cost efficient. An example for such a system would be an industrial quality control system. Otherwise, we perceive limited potential for tasks, in which the performance of human experts is inferior compared to the performance of automated systems. This scenario would include many applications of time series forecasting. In these tasks, the estimated prediction quality could only be used to issue warnings whenever an output is likely to be faulty. Altogether, we believe that the uncertainty-based human-in-the-loop system represents an essential building block for the facilitation of a more widespread adoption of CNN-based systems in the industry.

Part III

Design Knowledge

5 Image-Mining-Based Decision Support Systems: Design Knowledge and its Evaluation in Tool Wear Analysis[1]

5.1 Introduction

Many decision processes are based on image data, for example cancer diagnosis in medicine (Lambin et al., 2017), infrastructure maintenance (Chatterjee et al., 2018), user decision-making in e-commerce (Naumzik & Feuerriegel, 2020), real estate pricing (Kucklick & Müller, 2020), or industrial monitoring (J. Wang et al., 2018). Due to rapid and major advances in image processing techniques, the process of automatically transforming image data into information has improved heavily. Deep learning development in particular is often regarded as a breakthrough (LeCun et al., 2015) because respective algorithms are able to outperform humans in image classification tasks (He et al., 2015).

The common denominator in most existing research and practical implementations is the focus on analyzing single images to detect or classify objects. There is however large untapped potential in analyzing image collections. In manufacturing, for example, an inherent dispersion in processes makes them unreliable to be analyzed based on single images. Figuratively speaking, a single image does not differ from a single database row, therefore the analytical potential is limited. To retrieve knowledge from data collections, data mining techniques can be applied (Han et al., 2011; Spangler et al., 1999). When it comes to images, the inherent complexity leads to challenges that are not covered by standard data mining techniques. Image mining aims to discover patterns in image data, thereby generating value out of image collections (Bhatt & Kankanhalli, 2011).

[1]This chapter comprises an article that is currently under review as: Walk, J., Schemmer, M., Kühl, N., Satzger, G. (2022). Image-Mining-Based Decision Support Systems: Design Knowledge and its Evaluation in Tool Wear Analysis. *Working Paper*. Note: The abstract has been removed. Minor edits have been made and tables and figures were reformatted, and newly referenced to fit the structure of the thesis. Chapter, section and research question numbering and respective cross-references were modified. Formatting and reference style was adapted and references were integrated into the overall references section of this thesis.

Decision support systems are a proven tool to generate value out of data by providing a basis for making informed decisions (Kohli & Devaraj, 2004). Image data has high potential to be used in these systems because of its information richness (Lambin et al., 2017). A single image contains more information than text or numerical data — as the saying goes, *"A picture is worth a thousand words"*. For example, an image of a car can provide information about the manufacturer or the color, but also about the condition, the setting, and many more.

Due to the information richness of image data and recent advances in deep learning, we see great potential in combining the research of image mining and decision support systems. "Image-mining-based decision support systems" can create value in every domain where images need to be analyzed for decision-making. There is a magnitude of potential application areas, such as manufacturing (Trinks & Felden, 2019), medicine (Sollini et al., 2019), sport (W. Sun et al., 2008), geo-information (Coenen & Dittakan, 2016), or services (Villarroel Ordenes & Zhang, 2019). Most of these areas have conducted initial research about image mining, this confirms the potential for value creation. However, to the best of our knowledge, there is no design knowledge that supports the design of image-mining-based decision support systems. Such 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 (IS) knowledge body on designing decision support systems (DSSs) should be expanded to enable incorporation of this important type of data. While other special data types, e.g. text, are traditionally embodied in DSS research (Turban et al., 2010) 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 yet more information. We believe that in later times it was neglected because of the complexity of transforming image data into information. Due to recent advances in the field of deep learning this complexity has been reduced (LeCun et al., 2015). We therefore formulate the following research question:

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

In his outlook for IS research, A. S. 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 test 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, 2012; 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 a 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 possess a high economic importance: Research suggests that tool failures are responsible for 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).

Having outlined the importance of our case industry we now present our contributions. This research contributes to theory as well as practice. Our work provides design knowledge for image-mining-based decision support systems that contributes to information systems research because image-mining-based decision support systems are an important, but neglected class of design. 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 utilize 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 5.2, we introduce our design science research methodology. In Section 5.3, we derive tentative design requirements based on an exploratory interview study and literature. In Section 5.4, we present related work and the key concepts of our research. In Section 5.5, we derive and evaluate design knowledge for image processing. Based on this, in Section 5.6, we derive further initial design principles, following the image mining process and evaluate them. In Section 5.7, we refine our derived design knowledge and the artifact based on the prior feedback and evaluate it. In Section 5.8 we discuss our results and present our developed design theory. Lastly, in Section 5.9, we summarize this research, explain limitations of our study and provide an outlook on future work.

5.2 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 conduct three DCs that each refine the proposed design knowledge and the artifact. Each DC follows the steps 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 in design science research is the rigorous evaluation (Peffers et al., 2012; Prat et al., 2015). To structure the evaluation, we follow the guidelines of Venable et al. (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) propose. After targeting the technical viability of the image processing in DC1, in DC2 we derive further design knowledge for image-mining-based decision support systems 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 (EFGs) (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 (CFGs), as Tremblay et al. (2010a) propose, to gather qualitative feedback about the artifact's usefulness. Lastly, to validate perceived usefulness, we conduct a survey based on questions from the technology acceptance model (TAM) (Davis, 1989; Venkatesh & Bala, 2008). Figure 5.1 on page 74 visualizes the interplay of the three DCs and our EEs, including the objectives, the methods, and our results.

5.3 Relevance Cycle: Design Requirements for Image-Mining-Based Decision Support Systems

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

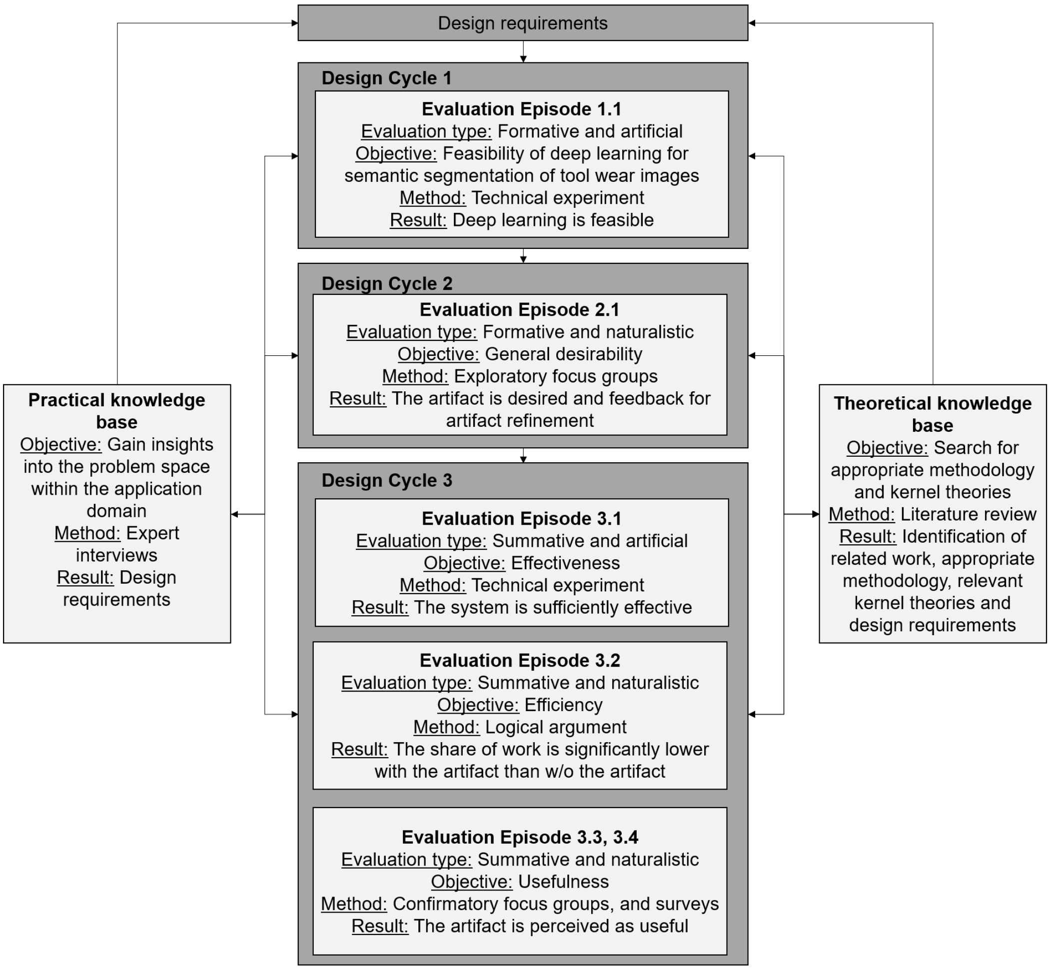


Fig. 5.1.: Design science research methodology based on Hevner (2007), Kuechler and Vaishnavi (2008), and Venable et al. (2016).

standardized images that facilitate image processing. The industry in general as well as the particular case company are therefore highly suitable for IM-DSSs 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, 1995). There are several types of cutting tools, for example drills and indexable inserts. We focus on the latter, shown in Figure 5.2 on page 75.

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 first preliminary DRs for IM-DSSs. In the following subsection, we describe the methodology and results.

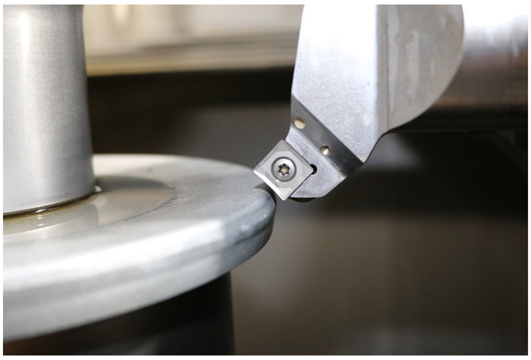


Fig. 5.2.: Exemplary machining process with an indexable insert.

5.3.1 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. Afterwards, 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 are 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 (D. R. 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 (D. R. Thomas, 2006). Based on the identified codes, we derived eight initial tentative DRs.

5.3.2 Results of the Exploratory Study

We found 12 image mining use cases in the case company[2]. For our research, we chose services for customers of our case company where experts (application engineers and developers) analyze machining processes and give recommendations on how to improve the process — optimization-as-a-service. Because image mining needs huge image collections to be useful (Hsu et al., 2002) and the customer services enable an efficient collection of relevant images, we identified the use case as highly suitable for an IM-DSS.

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 5.1 on page 78 shows exemplary quotes from the interviews done in the case company to illustrate the derivation. We also support the DRs with evidence from literature.

[2](1) Tool condition monitoring; (2) Single-image recommender systems; (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

According to the experts, context is needed to generate significant value from image data (Alpha). 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 sufficient image quality.*

A recurring theme of the experts was the comparison of image collections in benchmark tests (Beta). An image collection in the machining industry usually represents the wear on cutting tools that were utilized 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 image collection comparisons.*

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

> **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 in mind. An interviewee in a management position specifically emphasized this point (Zeta). We therefore formulate the following DR:

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

Tab. 5.1.: 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." (Alpha)	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." (Alpha)	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." (Beta)	M. 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." (Epsilon)	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." (Zeta)	K. Wang et al. (2016)
DR6 (Reproducibility)	"If you ask three experts now, you'll get five opinions." (Gamma)	Patel and 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." (Gamma)	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." (Delta)	Gillies et al. (2016)

The inherent problem of image data is that it is usually subject to human interpretation (Patel & Sethi, 2007), this 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.*

5.4 Rigor Cycle: Theories Informing the Design of Image-Mining-Based Decision Support Systems

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.

5.4.1 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 a 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. 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 presegmenting the images automatically and refining the segmentation by experts.

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. Hsu et al. (2002) give a holistic overview of the most common ones. Traditional techniques comprise image retrieval, image classification, and clustering or association rule mining.

5.4.2 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, p. 149). Our work synthesizes the long-lasting knowledge body of DSSs 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 H. M. Cooper (1988) and Vom Brocke et al. (2009) by searching the Scopus database with the following search string: *(radiomics OR "image mining" OR "image data mining") AND "decision support system"*.

The articles 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, as 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. The seven relevant papers are outlined below.

Most previous research was conducted in medical application areas. Exceptions are the works of Zaiane et al. (1998) and Koh and Cui (2022). Zaiane et al. (1998) developed an IM-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 et al. (2008) designed a framework for image mining and instantiate 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 consists of image processing, segmentation, feature extraction, and classification. Gatta et al. (2019) built a holistic IM-DSS and evaluate it on two data sets of cancer patients. Cheng et al. (2019) developed a clinical DSS 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) (compare Chapter 6) and Zschech, Walk, et al. (2021) (compare Chapter 7). Landwehr et al. (2022) derive design knowledge for image-based DSSs and conduct a case study in power line infrastructure maintenance. They use images from unmanned aerial vehicles to analyze the wear 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, Walk, 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.

5.4.3 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 Figures 5.3 to 5.5 on page 83 for exemplary images). The first type is abrasive wear on the flank, called *flank wear* or *VB* (Dutta et al., 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 - International Organization for Standardization, 1991). 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 - International Organization for Standardization, 1991). Lastly, high temperature and pressure can lead to a *built-up edge (BUE)* (Dutta et al., 2014). Chipping and BUE lead to deformations of the cutting edge, this may lead to insufficient product quality, increased scrap, and high costs.

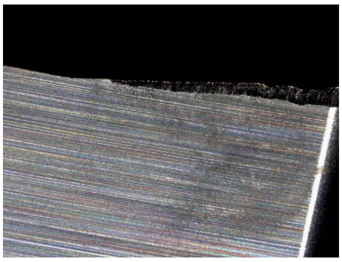


Fig. 5.3.: Flank wear.

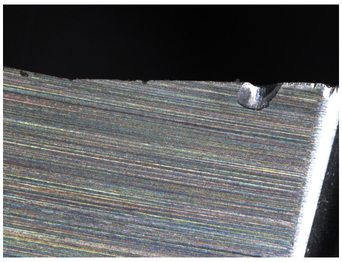


Fig. 5.4.: Chipping.

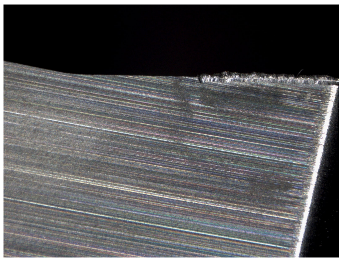


Fig. 5.5.: Built-up edge.

5.5 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 5.6 on page 84 summarizes our DRs, as well as our suggested DPs and DFs, iteratively derived over three DCs.

As explained in Section 5.4.1, 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.

5.5.1 Suggestion and Development

Deep learning, 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 (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). The implementation details can be found in the Appendix on page 100.

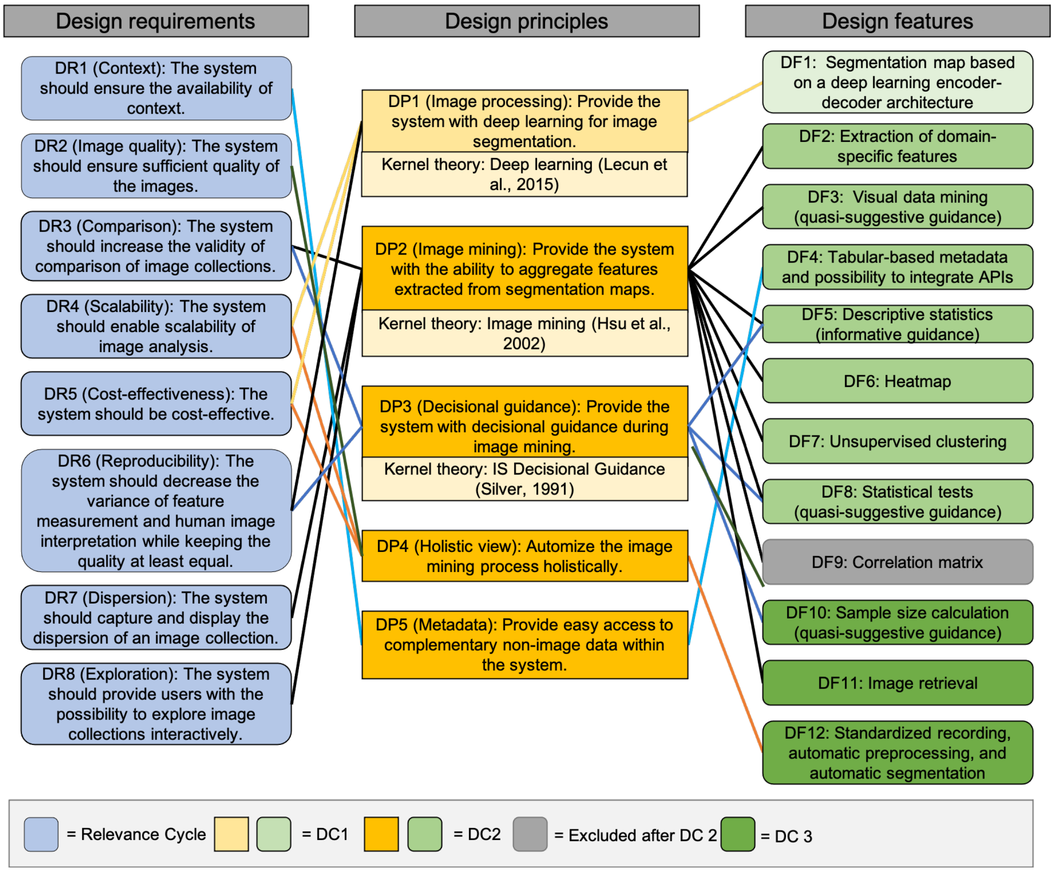


Fig. 5.6.: Design knowledge for image-mining-based decision support systems.

We used an open source web application to label images on a pixel level (Fisher & Mackiewicz, 2020). Our image data set consists of 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 BUE — see Section 5.4.3) as well as the remaining background pixels. The data set has two major sources of imbalance: First, because VB 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 5.7 on page 85 shows an exemplary image, label and output.

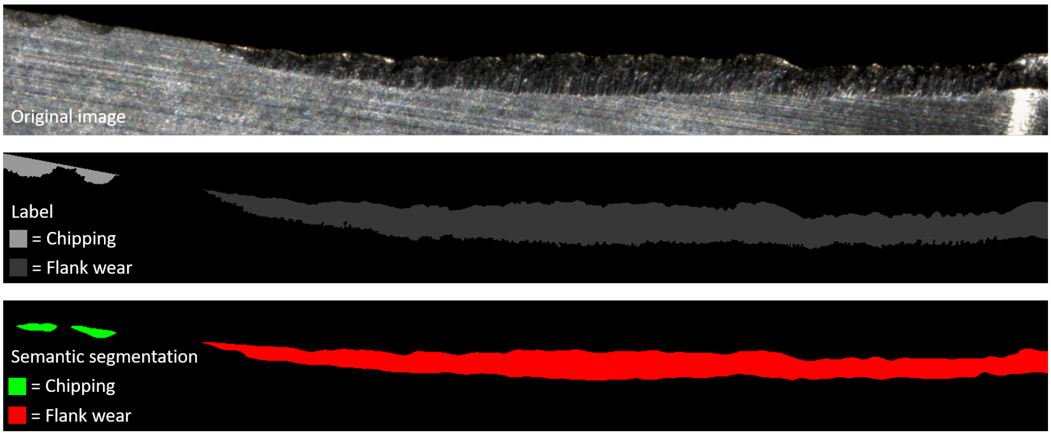


Fig. 5.7.: Transformation from original recorded images to segmentation maps.

5.5.2 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 (5.1)).

$$Mean\ dice\ coefficient = \frac{2}{C} \sum_{c=1}^{C} * \frac{precision_c * recall_c}{precision_c + recall_c} \tag{5.1}$$

We reached a pixel accuracy of 0.977 and a mean dice coefficient of 0.631. Table 5.2 shows the specific results of all four classes.

Tab. 5.2.: Performance results of the semantic segmentation.

Class	Dice coefficient
Background	0.991
Flank wear	0.695
Chipping	0.244
BUE	0.596
Mean dice coefficient	0.631

Flank wear and BUE 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 relative 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 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.

5.6 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.

5.6.1 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 an 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 on 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 (Information aggregation):** *Provide the system with the ability to aggregate features extracted from segmentation maps.*

Tool wear analysis has dispersion on two levels, summarized in **DR6** (Reproducibility) — first, on the level of tool wear measurements and second, on the decisional level. Automatic segmentation and feature extraction reduce the dispersion on the first level. To address the second level, we propose to utilize 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 (M. S. Silver, 1991). M. S. Silver (1991) differentiates three forms of guidance — informative, suggestive, and quasi-suggestive. Informative guidance provide decision makers only with decision relevant information, whereas suggestive guidance makes judgmental recommendation (M. S. 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" (M. S. Silver, 2006, p. 109). Research has shown that decisional guidance can improve decision-making and decreases 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. 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 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.*

5.6.2 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 5.8 on page 88 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.

As stated in Section 5.5, an algorithm to derive the segmentation map is necessary **(DF1)**. Due to the reasons explained in Section 5.5, the segmentation map is created by using DL and in particular CNNs. Next, based on that segmentation map, domain

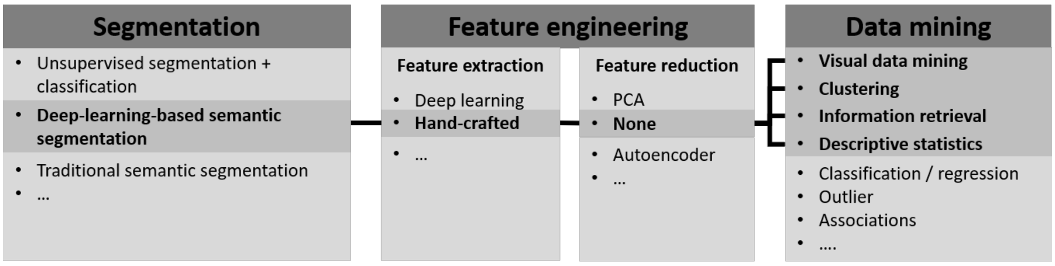


Fig. 5.8.: Possible design features for image mining.

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 utilize 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 - International Organization for Standardization, 1991). Table 5.3 summarizes the handcrafted features.

The third category are data mining techniques chosen with respect to the application domain and the use case. Tool wear analysis is exploratory, seeking to understand

Tab. 5.3.: Domain features extracted from segmented tool wear images.

Feature	Description
VBMax	Maximal height of the flank wear
VB	Average height of the flank wear
VB length	Length of the flank wear
VB 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 chippings
BUE area	Size of the ROI of the BUEs

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 5.4.3, the common criteria for tool life evaluation is the VBMax. The ISO standard defines a VBMax of 0.3 millimetres as end of life criterion (ISO - International Organization for Standardization, 1991). This criterion is implemented as default value for an upper limit which 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 5.9 on page 90 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.

Fig. 5.9.: 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 color 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 at https://youtu.be/1UdqHV35lkc.

5.6.3 Evaluation and Discussion

To evaluate the prototype, we use EFGs (Tremblay et al., 2010a). The goal is to discuss the usefulness of the proposed DFs and generate feedback for the refinement of the artifact. To select the participants, we used theoretical sampling (Coyne, 1997) to control the homogeneity of the groups and increase free-flowing discussion (Tremblay et al., 2010b). 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. 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, p. 258), 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 5.4. The overall feedback was very positive. We can conclude that IM-DSSs are desired by the experts and have potential for tool wear analysis. Additionally, we could generate feedback for the artifact refinement.

Tab. 5.4.: 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 the 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

5.7 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.

5.7.1 Suggestion and Development

In the following part, we discuss the feedback from DC2 and the implications. We clustered the implications in 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 which 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 EFGs 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 5.3, 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. Afterwards, we conceptualized an adapted process, including automation potentials. Figure 5.10 on page 94 depicts the adapted process. It starts with an external customer order (0). Afterwards, 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. Afterwards, 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 afterwards evaluated with respect to optimization criteria (9).
Figure 5.11 on page 95 shows the final artifact's graphical user interface. A video of the artifact can be found at https://youtu.be/OdZZBRXchyE.

5.7.2 Evaluation and Discussion

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 5.7.1. *The transformation of image data into information is more effective using IM-DSSs than manual information extraction.*

Proposition 5.7.2. *The transformation of image data into information is more efficient using IM-DSSs than manual information extraction.*

Proposition 5.7.3. *Domain experts perceive the application of IM-DSSs as useful.*

Testing Proposition 5.7.1 To measure the effectiveness of the feature measurement, we chose a key feature in machining, the height of the flank wear (VBMax), to compare the human and ATWA's error rate. 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

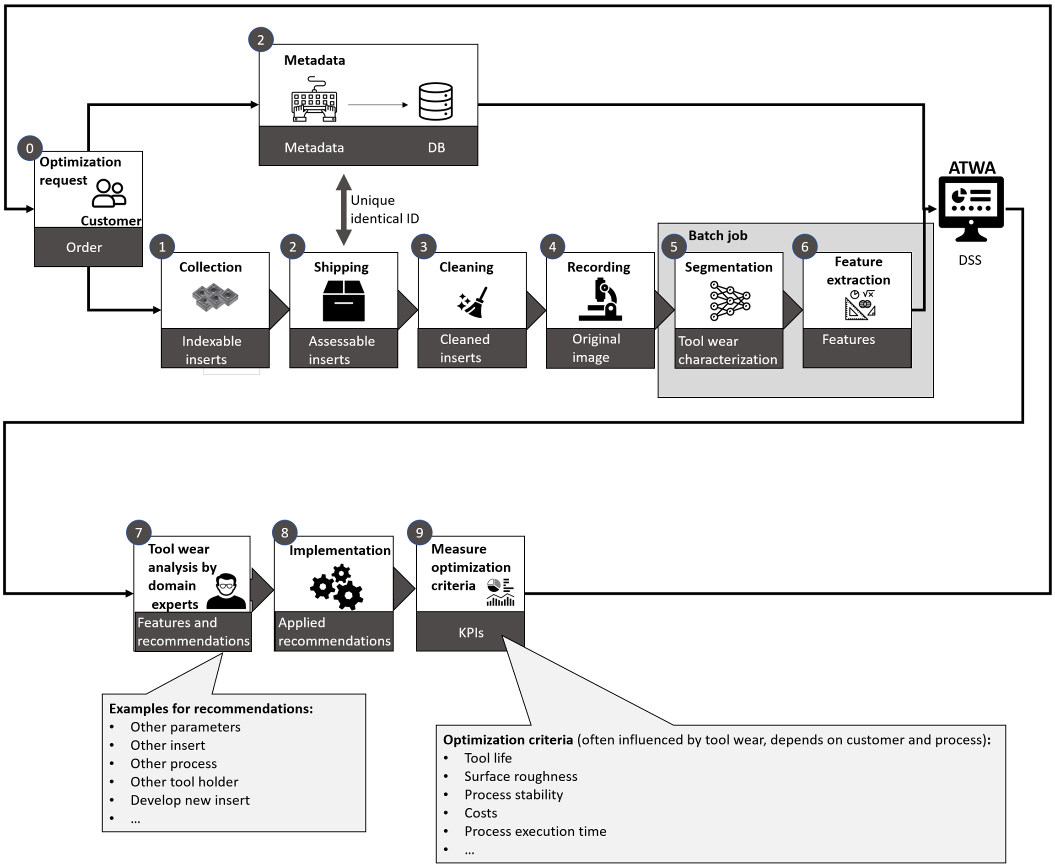


Fig. 5.10.: Data flow and architecture of ATWA.

was based on the created labels for Section 5.5. Conducting ten measurements led to a human MAE of 0.025 mm with a standard deviation of 0.021 mm and ATWA's MAE of 0.049 mm with a standard deviation of 0.026 mm. The MAE shows that there is a small difference between human and ATWA's feature measurement. Even though we could not verify Proposition 5.7.1, we believe that, as explained in Section 5.5, 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 5.7.2 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 on average 24.4 seconds and the

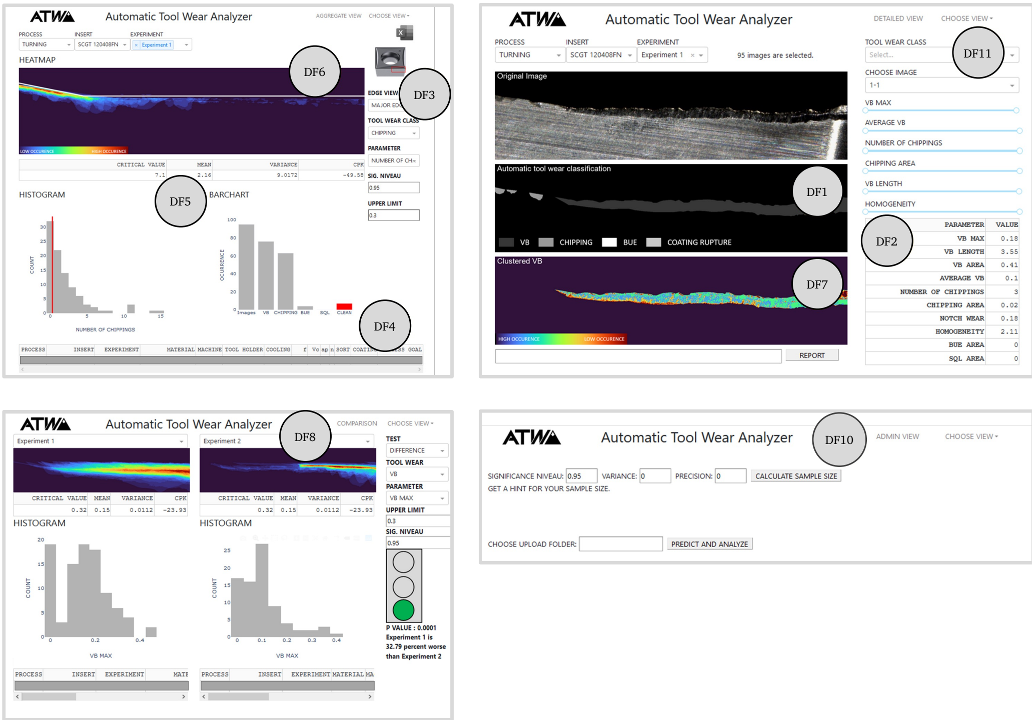


Fig. 5.11.: Interface of ATWA (top left: aggregated view; top right: detailed view; bottom left: comparison view; bottom right: admin view).

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 5.7.3 In the third EE, we conducted two CFGs, 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 afterwards coded with template analysis by two researchers independently. The results were discussed and merged afterwards. Table 5.5 on page 96 visualizes the results of the CFGs. 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

Tab. 5.5.: 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

item was measured on a 5-point likert scale. We calculated the mean and standard deviation (Std.) of each item. The results are shown in Table 5.6 on page 97. 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.

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

Tab. 5.6.: 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

5.8 Discussion

The first two EEs (effectiveness and efficiency) illustrate the automation potential of tool wear segmentation and 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 Section 5.5, Figure 5.6 on page 84 shows the matching of DR, DP, and DF.

The CFGs (EE3) and the survey (EE4) indicate the general usefulness of our developed nascent design knowledge. We use the core components of the IS design theory framework from Gregor and Jones (2007) to structure and present our overall developed nascent design knowledge for IM-DSSs. Table 5.7 on page 98 summarizes the design knowledge.

We build on top of existing work of image mining and DSSs and synthesize both into a novel design class, IM-DSSs. We see IM-DSSs as an extension of the knowledge base of intelligent decision support systems, 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. Our design requirements and design principles can guide researchers and practitioners to develop efficient and useful IM-DSSs.

Tab. 5.7.: Nascent design knowledge for IM-DSS.

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): The 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 from the body of knowledge, we derived five tentative design principles and evaluated the design in four evaluation episodes through twelve design features.
Justificatory knowledge	We conceptualize our design principles based on the kernel theories of Hsu et al. (2002), LeCun et al. (2015), and M. S. Silver (1991).
Testable propositions	We formulated and tested three testable propositions: Proposition 5.7.1, Proposition 5.7.2, and Proposition 5.7.3.
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 **a**utomatic **t**ool **w**ear **a**nalyzer (**ATWA**), to support the experts in conducting tool wear experiments and evaluating these.

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).

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 for human intelligence and enables human-AI synergy. With our work, we could show the usefulness of an artifact using human-machine collaboration.

5.9 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 system 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, which were implemented in a prototype and qualitatively evaluated using exploratory focus groups. The focus groups indicated a 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, which indicated sufficient effectiveness, efficiency, and usefulness of our nascent design knowledge.

Our research contributes to theory and practice. Regarding theoretical contributions, we shed first light on a problem class which 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.

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.

Besides the aforementioned contributions, our research also has limitations. First, regarding the evaluation, the technology acceptance model aims to measure potential users' intended usage behavior. A further study should assess the artifact's long-term effects. We therefore want to conduct a field test and already equipped the artifact with a logging functionality to access and analyze usage data. Second, we developed and evaluated the artifact at a single company. Future studies should apply the design knowledge in other domains and evaluate it.

We see potential 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.

5.10 Appendix for Chapter 5: Implementation Details

As 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 et al., 2015). We trained the model for 200 epochs, using an Adam optimizer and a learning rate of 0.00001.

6 Design Knowledge for Deep-Learning-Enabled Image-Based Decision Support Systems — Evidence From Power Line Maintenance Decision-Making[1]

6.1 Introduction

With modern-day societies increasingly relying on electrical power, the importance of continuous electricity supply cannot be overlooked. Continuous power supply has two central building blocks — the electricity generation as well as its transmission and distribution to the consumer. From the perspective of transmission or distribution system operators, the maintenance program of the power line infrastructure is crucial in avoiding unexpected disruptions. These system operators have typically adopted condition-based maintenance programs to minimize the probability of equipment breakdowns (Jalil et al., 2019; Pagnano et al., 2013). Condition-based maintenance is considered as a three-step process of data acquisition, data processing, and maintenance decision-making (Jardine et al., 2006), with the last step integrally including maintenance order planning (Gopalakrishnan et al., 2015).

Assessing the condition of the components in an electricity network includes inspecting towers or poles with their connected components, conducting power lines,

[1]This chapter comprises an article that was published as: Landwehr, J., Kühl, N., Walk, J. and Gnädig, M. (2022). Design Knowledge for Deep-Learning-Enabled Image-Based Decision Support Systems. *Business & Information Systems Engineering*. https://link.springer.com/article/10.1007/s12599-022-00745-z. Note: The abstract has been removed. Minor edits have been made and tables and figures were reformatted, and newly referenced to fit the structure of the thesis. Chapter, section and research question numbering and respective cross-references were modified. Formatting and reference style was adapted and references were integrated into the overall references section of this thesis.

and the surrounding vegetation of the two previous elements, as it can cause short circuits. Operators routinely examine these aspects regarding faults. Based on the operator's composed inspection reports, maintenance engineers need to compile situation-dependent, well-defined, complete, and prioritized maintenance orders. This requires the consideration of several other factors, such as infrastructure topology, available workforce and skill sets, scheduled infrastructure revision projects, and bundling of maintenance orders. Fast and accurate inspection as well as complete and exhaustive data and information dissemination are crucial for efficient maintenance decision-making and can reduce the risk of power outages due to component failures, increasing the reliability of electricity supply.

Traditionally, the inspection is performed through human visual observation by means of manual ground inspection, helicopter-based patrolling, and tower climbing. These inspection methods are costly, time consuming, partly hazardous, do not comprehensively capture data, and are hardly scalable. Recent technical advances in the fields of unmanned aerial vehicles and image processing or computer vision[2] have spurred the development of automated power line inspection. Specifically, deep learning has proven to boost the performance of image processing applications (LeCun et al., 2015) — converging towards human level performance or even surpassing humans (He et al., 2015). Researchers are therefore increasingly focusing on the automatic vision-based detection of components and the immediate diagnosis of faults in the inspection of power lines (X. Liu et al., 2020) leaving only the eventual maintenance decision-making for human handling.

Previous research on power line maintenance has been scattered, focusing on the technical building blocks. Today, the majority of studies either focus on performing unmanned aerial vehicle inspection flights autonomously (Hui et al., 2018) (data acquisition), on task-specific image processing approaches for component detection and fault diagnosis (Nguyen et al., 2018) (data processing), or on orchestrating the various technical components (Homma et al., 2017; Huang et al., 2018) (interplay between data acquisition and processing). So far, little effort has been devoted to holistic and end-to-end considerations establishing a relationship between the solely technical problems of automating the data acquisition and processing and the need for integrating and transferring the acquired data and extracted information into maintenance decision-making. To this end, we conduct a project to design and evaluate a suitable decision support system following the design science research paradigm (Hevner, 2007) and its common research guidelines (March & Smith, 1995; Winter, 2008). We address the ever-increasing need for maintaining the impeccable condition of power lines, and consequently the reliability of electricity supply.

[2]Note that we will use the term image processing and computer vision interchangeably, as there is no common agreement between the boundaries of the two terms (Gonzalez & Woods, 2018).

We do so by utilizing available technological possibilities for holistic vision-based applications to provide decision support in scoping and planning maintenance orders for maintenance engineers through improved data and information quality. We focus on addressing this need by answering the following research question (RQ):

> **Research Question D**
> How can an automated, efficient, and useful vision-based power line maintenance decision support system be designed?

By answering this question, we tap the still largely unregarded and nascent problem class of *image-based decision support systems*, which we believe to be the higher level abstraction for our specific vision-based power line maintenance decision support system. In particular, following the dual mission of design science research of developing usable artifacts for practice and generating theoretical knowledge for the knowledge base (Gregor & Jones, 2007), we initially explore the challenges and issues of power line maintenance to derive a number of design requirements for *image-based decision support systems*. Subsequently, we conceptualize design principles based on justificatory knowledge from image processing and deep learning. Based on these design principles, we obtain a number of design features as our application domain specific design for the *image-based decision support systems* for vision-based maintenance of power line components. We instantiate these design features into a concrete artifact that allows us to rigorously evaluate the proposed design knowledge in practice.
The remainder of this work is structured as follows: Section 6.2 summarizes the existing relevant literature. Next, in Section 6.3, we introduce the research methodology. In Section 6.4, we conceptualize our design knowledge for *image-based decision support systems*, before we introduce the developed artifact as well as its various evaluations in Section 6.5. Finally, in Section 6.6, we discuss our research findings, reflect on the limitations of our work, and provide an outlook for future studies.

6.2 Related Work

To determine the potential of extensively captured images of power line components (PLCs), we review related work and the literature background in several fields. First, we briefly introduce foundations regarding deep learning (DL) (Section 6.2.1). In Section 6.2.2 we present how computer vision (CV) is used for infrastructure inspection in different application domains. Subsequently, in Section 6.2.3 we present related work regarding automated vision-based power line inspection using UAV-captured images. Afterwards, in Section 6.2.4 we examine image-based decision support systems (IB-DSSs) as a way to harness images in efficient decision-making. We conclude this section by synthesizing the presented literature and depicting our research gap in Section 6.2.5.

6.2.1 Deep Learning

Within the past decade, machine learning has shown significant results solving complex problems — both in theory as well as in application within industry (Brynjolfsson & Mcafee, 2017). Especially in the field of DL[3], a family of algorithms solely based on artificial neural networks with multiple hidden layers, the developments grew rapidly (Bharati & Pramanik, 2020).

DL overcomes a general limitation of machine learning to handcraft appropriate features in order to find and learn patterns in input data. The advanced architecture gives DL the capability to automate feature learning and consequently reduce human effort (Janiesch et al., 2021). Hence, DL is able to better deal with large-scale, noisy, and unstructured data.

The exact amount and size of layers is a design choice such that the ideal architecture for a given problem and its data must be found through experimentation (Goodfellow et al., 2016). Each layer is subject to learning and computes non-linear input-output mappings which enables a DL model to represent extremely intricate functions of its input (LeCun et al., 2015).

Due to these capabilities, DL has brought breakthroughs in processing images, videos and audio like speech (LeCun et al., 2015). In particular, convolutional neural networks (CNNs) a class of DL algorithms which excel at learning hierarchical features (Janiesch et al., 2021), are especially suited for the application to feature-rich data — like images. Therefore, DL is a promising candidate for applications within the field of CV.

[3] For a general introduction into machine learning and deep learning, we refer the interested reader to Janiesch et al. (2021).

6.2.2 Computer-Vision-Based Infrastructure Inspection

CV aims to equip computers with visual perception skills similar to the human ones (Szeliski, 2010). CV models based on DL have led to a significant increase in performance — DL models have even been proven to surpass human-level performance for specific applications (He et al., 2015). Typically, four different CV tasks are distinguished on static images (Griebel, Dürr, et al., 2019): in image *classification* the whole image is assigned a class label. Object *detection* additionally outputs an approximate location of the object of interest. *Semantic segmentation* produces even more fine-granular information, as each pixel is assigned a class label. In the specific case of *instance segmentation*, neighboring objects of the same class are distinguished additionally.

In the past years, several specific architectures have been developed to allow for these different CV tasks. While the two main optimization criteria are the accuracy of the prediction and the time inferred to obtain the solution ever more tailored solutions building on CNNs are being developed recently. Architectures such as VGG16 (Simonyan & Zisserman, 2014) and ResNet (He et al., 2016) for image classification and Faster R-CNN (S. Ren et al., 2015) and SSD (W. Liu et al., 2016) for object detection have proven to provide good accuracy at reasonable inference time.

CV is utilized for infrastructure inspection in many application domains. The typical challenges addressed with CV in this area are cases where large amounts of physical objects are to be inspected and they are geographically remote and / or dispersed. Selected research articles are presented in the following and summarized in Table 6.1 on page 109.

A major application area is road surface inspection and maintenance. Roads in bad condition can ultimately result in more accidents and higher costs (Baladi et al., 2017; Gleave et al., 2014). Thus, CV is utilized to automatically assess road surface condition and derive necessary maintenance actions. Over the last years this became possible without expensive, specialized hardware (compare e.g. Quintana et al. (2016)). Chatterjee et al. (2018) show how machine learning-based CV can be used to detect road surface cracks and develop a "vision-based DSS for crack detection". They offer first insights into a nascent design theory for the application case of road crack detection on the basis of images.

Not only roads, but also railways need to be inspected periodically to ensure safe transports. Wei et al. (2019) employ a Faster R-CNN to detect defects of railway track fasteners. Gibert et al. (2017) propose a CNN-based multitask learning approach that detects railway track fasteners and crossties and classifies the state of these components.

Wind turbine blades are another physical object of interest for CV-based infrastructure inspection. Akhloufi and Benmesbah (2014) present a CV approach to identify ice accreation on wind turbine blades. Ice accreation can require a maintenance action since it can cause malfunction and premature wear and is a safety hazard for nearby people and infrastructure like roads and powerlines. Shihavuddin et al. (2019) show how faults like leading edge erosion can be detected with a Faster R-CNN on wind turbine blades.

6.2.3 Automated UAV Vision-Based Approaches for Power Line Inspection

In this work, we are particularly interested in CV solutions for power line inspection relying on UAV images. From a component-based view, power line inspection can be divided into four major categories: towers or poles, insulators, conductors, and fittings (X. Liu et al., 2020). Each of these categories contains several subcomponents (Nguyen et al., 2018) that typically vary in size, kind, and material according to the voltage level. For instance, some part of a distribution network with low voltage might have wooden poles, small standing insulators, and a single, relatively thin conductor. On the other hand, transmission networks usually have lattice steel towers, large suspending insulators, and thicker conductors. Several studies have been published that utilize various potential platforms (e.g. helicopter, satellite, and UAV) to collect different data types (e.g. optical images, laser scanner data, thermal images, and synthetic aperture radar images) and analyzed these through different processing techniques (Matikainen et al., 2016). The vision-based approach — with image data from the visible spectrum captured by UAVs and automatically analyzed through image processing capabilities — has gained the most attention and traction in the power line inspection research domain (X. Liu et al., 2020).
With a few exceptions, automated vision-based power line inspection based on UAV-captured images requires two inherently related tasks (X. Liu et al., 2020): component detection and localization as well as fault diagnosis. The exceptions relate to objects such as bird nests, whose detection already represents a fault. Previous research applying image processing for the detection and fault diagnosis of PLCs is numerous (Mirallès et al., 2014). X. Liu et al. (2020) identify several characteristics and shortcomings of previous studies using UAV-captured images in their exhaustive literature review. Most studies in the field of vision-based inspection of power lines focus on the insulator and its faults (X. Liu et al., 2020) — mainly missing caps (e.g. Sampedro Pérez et al. (2019), Y. Yang et al. (2019), Zhai et al. (2018)) — while little attention has been paid to other components. The safety pin that prevents

other components from loosening and falling is the smallest object in the power line and has, despite its importance, received little attention and has only been regarded in fault diagnosis but not in the detection step. Finally, both X. Liu et al. (2020) and Nguyen et al. (2018) conclude that the mediocre performance of task-specific approaches presented in the vast majority of studies has been superseded by DL approaches that have improved the performance of component detection as well as fault diagnosis.

To move towards the operationalization of automated vision-based inspection, we require approaches capable of detecting a wide variety of components and diagnosing their faults in order to integrate them into a valuable system. Although "the component detection is a relatively mature area" (X. Liu et al., 2020)[p.10], we found that only a few articles shed light on detecting several components in a single approach or pipeline. Besides the identified challenges, we therefore review all available DL-based approaches that consider more than one component in the detection step.

The first steps in this field were done by Zhu et al. (2018), who investigate the cascading of two Faster R-CNN architectures for high-voltage PLCs. While towers, spacers, vibration dampers, and insulators are directly detected from the input image on the first stage, the pixel coordinates of the tower are used to crop the input image and consequently feed it into the second stage to detect small objects — in their case bird nests and tower plates. Their results show that the cascaded architecture is able to detect small objects at better performance. Nguyen et al. (2019) propose a similar approach for low-voltage PLCs (pole, cross-arm, insulator, or top cap) with a large number of various subcomponents totaling 54 classes. The authors detect poles in the first stage, crop the respective image and detect other, smaller components in the second stage. In a third stage, the recropped components are fed into image classifiers to perform a fault diagnosis. This work shows the feasibility of designing a cascaded multistage detection and classification pipeline utilizing spatial relationships. However, it does so only for larger components in terms of pixel size. H. Liang et al. (2020) take a different approach. They do not follow the prevalent approach of separating detection and fault diagnosis, but skip the general detection of PLCs and directly detect only components that exhibit faults. While including a total number of ten fault types, the work naturally states the problem of the detection of intact components as defective components. It also does not try to achieve the detection or fault diagnosis of overly small components.

The aforementioned approaches can strongly facilitate inspection and thus the prioritization of subsequent maintenance operations. Additionally, the data that is acquired in an automatic and structured manner can serve as foundation for predictive maintenance (Selcuk, 2017). By utilizing the data to train detection models (as shown later in this work), continuous forecasts about the future occurrence of defects can be issued. A well-trained and deployed model can, therefore, support experts in indicating future maintenance needs early and prioritize potential work orders.

6.2.4 Image-Based Decision Support Systems

The access to increasing volumes of images and the capabilities of DL to process and extract information from images creates the potential to harness this rich data and DL methods to facilitate effective decision-making (Chaudhuri & Bose, 2020). Despite their capabilities, DL methods, particularly CNNs, have found limited adoption in extant research of IS in general (Kraus et al., 2020), and specifically DSS. Most research performed on image-based decision support focuses on the medical application domain (Ben-Cohen et al., 2017; Comaniciu et al., 1999). However, these works use highly specific medical scans rather than images from the visible spectrum. Some examples of the scarce literature on DL-enabled image-based decision support in non-medical contexts include vision-based maintenance and monitoring applications or pattern analysis (Chaudhuri & Bose, 2020; Jamshidi et al., 2018; Nazerdeylami et al., 2019; M. Ren et al., 2021; Schumann et al., 2019; Xie et al., 2020).

Despite the efficacy of DL methods for image processing in related decision support contexts, none of the previous work provides guidance on how to design IB-DSSs. Specifically, although all these studies aim for improved data and information availability, close to no insight is provided on how to bridge the gap between the sole image processing as well as consequent information extraction, and the respective efficient, high-quality decision-making.

Tab. 6.1.: Deep-learning-based power line inspection approaches to detect and diagnose multiple components and similar approaches from other application domains.

Articles	Application domain	Component detection		Fault diagnosis		Design focus
		Components	Method	Components	Method	
Chatterjee et al. (2018)	Road surface inspection	Road	Graph-based hierarchical clustering	Road cracks	Multiple machine learning classifiers	✓
Shihavuddin et al. (2019)	Wind turbine blade inspection	Leading edge erosion, vortex generator panel, vortex generator panel with missing teeth, lightning receptor	Faster R-CNN	*Fault diagnosis treated as detection task*		x
Wei et al. (2019)	Railway track inspection	Railway track fastener	Faster R-CNN	*Fault diagnosis treated as detection task*		x
Zhu et al. (2018)	Powerline inspection	Spacer, bird nest, insulator, damper, tower plate, tower	Cascaded Faster R-CNN	-	-	x
Nguyen et al. (2019)	Powerline inspection	Pole, cross-arm, insulators	Cascaded Faster R-CNN/ SSD/Yolo	Insulator, pole, top cap, cross-arm	ResNet50	x
H. Liang et al. (2020)	Powerline inspection	Defect tower foundation, insulator, grading ring, contact terminal, triple-plate, earth wire, bird thorn, bird nest, foreign body	Faster R-CNN	*Fault diagnosis treated as detection task*		x
Our work	Powerline inspection	Insulator, bird nest, fitting, safety pin	Cascaded Faster R-CNN/ SSD	Safety pin	ResNet50	✓

6.2.5 Synthesis and Research Gap

This work aims to interweave two research domains. It combines the applied research of image processing in power line maintenance (PLM) with the need for decision support in vision-based domains in general and in PLM in particular. This allows us to tap new potential through making previously unattainable data and information from individual images available.
We address this potential by investigating the environment of automated vision-based PLC maintenance, focusing on the design of a holistic image-based decision support solution. We develop design knowledge for IB-DSSs and evaluate it by instantiating a concrete artifact for PLC maintenance. We extend the reviewed existing works (cf. Table 6.1 on page 109) by managing to detect PLCs of extreme size difference (insulators and safety pins), which we believe is a crucial prerequisite for moving towards decision support in this domain.

6.3 Research Methodology

The research at hand develops design knowledge for IB-DSSs which supports the maintenance decision-making and planning of maintenance engineers (MEs) for power lines. Since design science research (DSR) has proven itself to be not only a suitable but also an important paradigm to develop IS in general (Gregor & Hevner, 2013) and DSS in particular (Arnott & Pervan, 2012), we follow its steps to develop and evaluate our artifact. At its core, DSR is a problem solving paradigm that involves two primary and distinct activities to design solutions to real-world problems: (1) the development of innovative artifacts in a series of design activities based on a deep understanding of the problem, justificatory knowledge, and the capabilities of the researcher and (2) the evaluation of the novel artifact to assess its ability and utility in solving the identified problem (Hevner et al., 2004). Following this "build-and-evaluate loop" (Hevner et al., 2004), we iteratively develop an artifact to extend the knowledge base regarding IB-DSSs.
Besides this loop — more precisely termed design cycle — Hevner (2007) describes the existence of two additional cycles: relevance and rigor. The three cycles are inherently related and part of any DSR project. The relevance cycle connects the environment, application domain, or case company of the research project to the design science activities by, for instance, incorporating input from expert practitioners. It does not only provide the requirements, problems, or challenges for the research, but also defines acceptance criteria (Hevner & Chatterjee, 2010). The rigor

cycle relates the design science activities to the existing knowledge base. It provides knowledge from scientific theories, engineering methods, experience, and expertise to the research project. The often repeatedly performed design cycle is the core of any DSR project and naturally builds on the insights from the two previous cycles. Specifically, during a design cycle the research iterates between construction and evaluation of an evolving artifact (Hevner & Chatterjee, 2010) to eventually deploy the artifact in the environment as well as distill insights and output the research's design knowledge contributions into the knowledge base.

In the general view of our research displayed in Figure 6.1 we start with study-

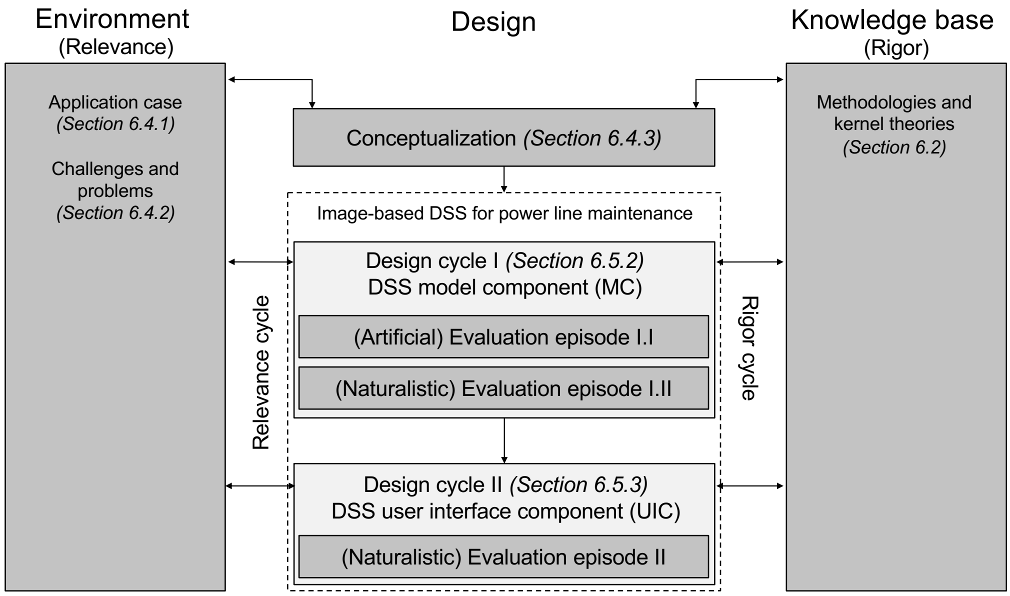


Fig. 6.1.: Overview of the research cycles and activities in the conducted study (based on Hevner, 2007).

ing the environment in which the research is embedded. We consequently state our application case (Section 6.4.1) and review related challenges and problems (Section 6.4.2). Joining these insights with knowledge from kernel theories we conceptualize principles and requirements for the problem class of IB-DSSs. We subsequently derive a concrete PLM artifact and, based on Turban et al.'s (2010) high-level notion of a DSS, first focus on the *model component* (MC) of our DSS artifact in the first design cycle (Section 6.5.2). Afterwards we move to the *user interface component (UIC)* in the second design cycle (Section 6.5.3). To orchestrate the evaluation of our artifact, we apply and follow the overarching *framework for evaluation in design science* (Venable et al., 2016) to rigorously demonstrate the utility and efficiency of the artifact and its underlying design knowledge. Figure 6.1 provides an overview of the performed evaluation episodes (EEs) in these design cycles. As it is our goal to indicate technical feasibility as well as utility of IB-DSSs enabled through DL, we start with a technical evaluation and then move to a naturalistic context within the application setting.

6.4 Application Case and Conceptualization

Our DSS artifact, built on images from the visible spectrum, intends to support MEs of power line infrastructure in their decision-making. More precisely, our system supports the planning and scoping of individual maintenance orders for the repair and replacement of components through improved data and information quality. Because the artifact is to intervene in an organizational context, it is considered "socio-technical" (Gregor & Hevner, 2013). To manage the complexity of the artifact construction in terms of size as well as social and technical components, Gregor and Hevner (2013) suggest the explicit extraction of design principles (DPs). We therefore conceptualize and suggest a number of tentative DPs for the design of artifacts of the problem class of IB-DSSs by first investigating challenges in power line maintenance (PLM). These are recast into a prescriptive mode with appropriate abstraction yielding preliminary design requirements (DRs), which then serve as a basis for deriving the DPs.

6.4.1 Application Case and Decision Process

As the largest distribution system operator in Baden-Württemberg, Netze BW supplies around 2.2 million customers and operates a network of almost 100,000 km. The distribution network, which is largely rural, poses challenges in the inspection of towers, poles, and overhead line routes. Every year, Netze BW operators routinely carry out around 7,000 scheduled inspections of high-voltage towers and lines, which include a visual inspection from the ground or by helicopter. For around 1,400 of these, towers must be climbed physically. Whenever operators identify an issue or defect on a tower during these inspections they manually create a report including the location, description, and if possible images. Based on these largely unstructured inspection reports, MEs need to subsequently compile situation-dependent, well-defined, complete, and prioritized maintenance orders. Accordingly, based on reported incidents MEs first scan the report and verify the priority of the incident. While the priority determines the processing order, for any incident several maintenance order specific details need to be compiled regardlessly. MEs will therefore check the topology surrounding an incident location as it determines which device and equipment can be used. Additionally, the incident and its preferred solution approach determine whether either internal operators can be dispatched or contractors are required. Another important aspect especially for incidents of lower priority is the consideration of forthcoming infrastructure revision projects. These can typically include the required maintenance order and, thus, avoid additional work. Finally, to avoid hazards during the maintenance work the respective circuits must be free of electrical current which requires routing the current flow to other power lines. Hence, MEs need to appropriately terminate these so called switches based on the incident priority and in close consultation with the grid control center as well as operators or contractors. Since the electrical grid often offers small margins for additional current flow such switches are often times difficult to set up. On this occasion MEs need to bundle incidents on the same power line to use such switches as efficiently as possible.

6.4.2 Challenges in the Power Line Maintenance

To understand PLM from a practitioner as well as a theoretical perspective, we started our research with a series of expert interviews among the case company's employees and a structured literature review (SLR) of domain-specific articles. The interviewees were chosen based on their work experience and affiliation to different departments dealing with the various aspects of the PLM process (cf. Table 6.2). This sampling allowed us to benefit from diverse viewpoints and nuanced perspectives on the challenges of PLM with today's manual inspection.

To guarantee a rigorous overview, we conducted the SLR following Webster and Watson (2002) and Vom Brocke et al. (2009) by querying various databases (cf. Table 6.3 on page 114). We harnessed a selection of search strings, as displayed in Table 6.3, to retrieve the initial set of relevant articles. To extract only relevant articles, we defined three exclusion criteria. If the paper examined or investigated only one specific solution approach for the automation of PLM, it was excluded. If a paper focused on constant monitoring of power lines rather than periodic inspection, it was also excluded. Finally, if on a thorough read of the paper no challenges regarding PLM were mentioned, the paper was ruled out. These exclusions allowed us to focus on review and survey contributions for the automation of PLM. The SLR conducted in January 2020 resulted in a large number of potentially relevant contributions as depicted in Table 6.3, with 22 papers remaining after the first exclusion and 18 survey and review papers mentioning challenges in today's PLM.

Tab. 6.2.: Overview of interview participants to determine challenges in power line maintenance.

ID	Role	Experience [years]
Alpha	Senior standardization engineer	10
Beta	Operator of high- and medium-voltage power lines	25
Gamma	Operations manager of high-voltage power lines	28
Delta	Asset manager	12

Tab. 6.3.: Search strings and respective results for the structured literature review.

Search strings	EBSCO	WoS	IEEE Xplore	Scopus
"Automat*" AND "Power line" AND "Inspection"	24	79	86	158
"Power line" AND "Quality control"	4	89	12	141
"Transmission line" AND "Automat*" AND ("Inspection" OR "Monitoring")	21	97	213	370
"Inspection" AND ("Power line" OR "Transmission line")	104	393	547	1301
("Power line" OR "Transmission line" OR "Overhead lines" OR "Overhead power lines") AND "Condition monitoring"	18	131	0	271
"Challenges" AND "Power line" AND "Inspection"	2	7	8	24

Statements from both the interview transcripts and scientific articles were then coded in an open coding process and combined in a qualitative content analysis as proposed by Mayring (1991) to derive a category system of today's PLM challenges. Table 6.4 on page 115 depicts a part of the identified challenges with the respective subchallenges and their sources. These three challenges (C1-3) appeared to be specific to our context of infrastructure inspection with its concrete characteristics being dependent on power line infrastructure and therefore inform the design of our artifact. Further identified challenges attributed to company and industry specifics can be found in Section 6.8.1 on page 137 within the Appendix.

Tab. 6.4.: Challenges in the maintenance of power lines based on expert interviews and a structured literature review.

ID	Challenge	Subchallenge	Source
C1	Complicating workplace characteristics	C1.1—Hazardous work environment	Pagnano et al., 2013; Nguyen et al., 2018; Jones, 2005; D. Li and Wang, 2019; Seok and Kim, 2016; Huang et al., 2018; Toth and Gilpin-Jackson, 2010; Alpha; Beta
		C1.2—Strenuous inspection activities	Alpha
		C1.3—Requirement for broad expertise	Takaya et al., 2019; Pernebayeva and James, 2020; Huang et al., 2018; Alpha; Beta; Gamma; Delta
		C1.4—Impact of subjectivity	Nguyen et al., 2018; Jones, 2005; Katrasnik et al., 2010; Toth and Gilpin-Jackson, 2010; Homma et al., 2017; Beta; Delta
C2	Inspectability challenges	C2.1—Inspection type related scope restrictions	Jones, 2005; Katrasnik et al., 2010; Beta; Gamma; Delta
		C2.2—Requirement for unscheduled inspections	Matikainen et al., 2016
C3	Infrastructure characteristics	C3.1—Age of power line infrastructure	Aggarwal et al., 2000; Toussaint et al., 2009; Alpha
		C3.2—Extent of power line infrastructure	Pagnano et al., 2013; Aggarwal et al., 2000; Pernebayeva and James, 2020; Huang et al., 2018; Homma et al., 2017; Alpha
		C3.3—Topography of infrastructure territory	Prasad and Rao, 2016; C. Deng et al., 2014; Aggarwal et al., 2000; Takaya et al., 2019; Pernebayeva and James, 2020; Matikainen et al., 2016; Seok and Kim, 2016; Huang et al., 2018; Toth and Gilpin-Jackson, 2010; Homma et al., 2017
		C3.4—Vast spectrum of inspection aspects	Nguyen et al., 2018; Prasad and Rao, 2016; Jones, 2005; Homma et al., 2017; Alpha; Gamma

6.4.3 Design Requirements

Our DSS artifact intends to support MEs of power line infrastructure in their planning and scoping of individual maintenance orders to repair and replace components. To accomplish this by systematically addressing the aforementioned uncovered challenges in PLM with a vision-based application, we cast these challenges into a prescriptive mode and derive DRs as depicted in Figure 6.2 on page 117. Consequently, we derive five DRs which describe our system objectives and confine to which objectives our subsequently derived design knowledge applies (Walls et al., 1992). Because we target developing generalized design knowledge for the problem class of IB-DSSs, we formulate the DRs on the relevant level of abstraction in the following.

The infrastructure characteristics (C3.1 - C3.4) pose challenges with regard to efficient data capturing as, for instance, power lines running across valleys or in mountainous areas complicate inspection and hinder data acquisition. In addition to this, the three inspection types used in today's PLM provide heterogeneous condition data of varying quality (C2.1). Together, these factors result in the need for an appropriate **image quality** relating to uniformly captured high-resolution image condition data regardless of infrastructure characteristics and with process consistency.

> **DR1 – Image quality:** *The system should uniformly capture condition image data of sufficient quality.*

Image data contains large amounts of unstructured information. However, the information contained in an image is typically of little use if its observer lacks contextual information. Context allows for a broader understanding of specific pieces of information and it places them in a bigger picture by for example providing temporal or geographical information. Images of the infrastructure and in particular of components therefore need to be contextualized in an appropriate way. On the other hand, the infrastructure characteristics (C3.1 - C3.4) pose the requirement for providing infrastructural context to enhance decision-making.

> **DR2 – Context:** *The system should capture and provide context.*

Today's inspection process of power line infrastructure is fully manual and labor-intensive. Above that, various human-factor-related challenges (C1.1, C1.2, & C1.4) influence the inspection's susceptibility to errors. Additionally, characteristics of the infrastructure, such as topography (C3.2) and extent (C3.3), result in an increased labor effort for maintenance. To mitigate the limitations of today's inspection process, both parts of the process — image acquisition and image processing — should be infused with **automation** capabilities.

> **DR3 – Automation:** *The system should allow for automatic image acquisition and provide automated image processing.*

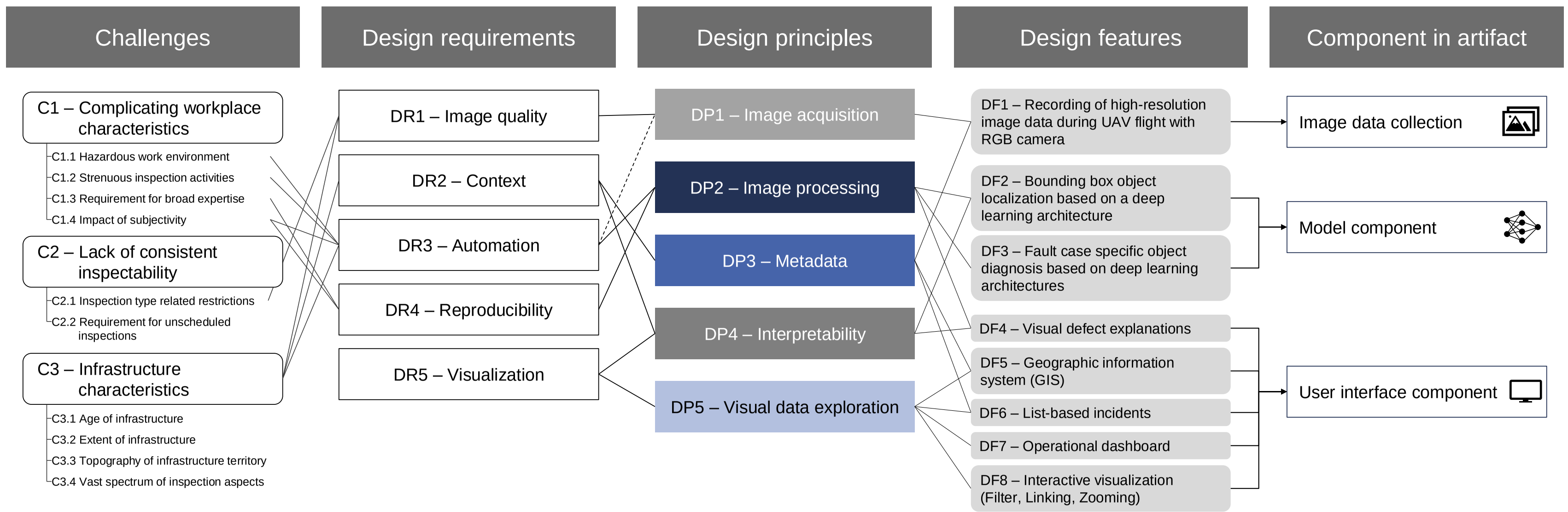


Fig. 6.2.: Design knowledge for image-based decision support systems with its respective instantiations.

To make adequate maintenance decisions in terms of repair or replacement prioritization, MEs require consistent condition data. However, just as in any human-based inspection, the fault diagnosis of power lines is characterized by the personal experience and expertise (C1.3) of the inspector, making the evaluation or judgment subjective (C1.4). To objectify the fault diagnosis and making it less subject to the experiences of a wide variety of inspectors, personal biases need to be eliminated or harmonized. Consequently, the analysis of the condition data needs to build upon equal decision parameters, achieving reproducible results. By **reproducibility** of results we refer to similar evaluation or fault diagnosis of a unique PLC within a range of potentially changing environmental conditions (e.g. lighting conditions).

> **DR4 – Reproducibility:** *The system should provide image processing in a reproducible manner.*

To draw inferences from the previously captured data and extract crucial information, proper visualization is required. Consequently, not only the quality but also the presentation of information regarding faults in the power line infrastructure are crucial. Specifically, it is important to integrate and transfer the entire collected data from the data acquisition and the extracted information from the data processing into the maintenance decision-making to enable the compilation of situation-dependent, well-defined, complete, and prioritized maintenance orders.

> **DR5 – Visualization:** *The system should support the process of decision-making with the visualization of the extracted information.*

6.4.4 Design Principles

In the following, we suggest several design principles (DPs) which prescribe how to develop the artifact in order to accomplish our predefined preliminary DRs (Chandra Kruse et al., 2015). The translation process from DRs into tentative DPs is displayed in Figure 6.2 on page 117. The DPs use the knowledge of several theories in order to meet the DRs. The main contributions originate from the domains of image processing, DL, DSS, as well as visual data exploration.

We have identified that the images of the PLCs need to be captured uniformly and with sufficient quality (DR1). Additionally, the system should capture context (DR2) of the images for unambiguity regarding their location and time. To address these design requirements, two considerations have to be made: the type and kind of data collected and the collection method, which we will refer to as platform. The primary type of collected data is predefined in our use case to be image data from the visual domain as it (1) provides enough information to detect a wide variety of common faults (Nguyen et al., 2018) — especially on PLCs — and (2) allows fast comprehension by MEs. On the other hand, the platform responsible for the data acquisition needs to be able to acquire uniform image data. In particular, the platform should be able to combine the advantages of today's inspection methods of helicopter-based, ground-based, and climbing-based inspection in a way that each of these methods that are specifically suitable for different components can be imitated. The platform is consequently able to capture images from above, below, and the front while maintaining a uniform viewing perspective per component type. The system should also allow data acquisition to happen in a potentially automated fashion (DR3) to further increase the scalability and reduce human involvement in the inspection process.

> **DP1 – Image acquisition:** *Provide the system with (automated) capabilities for uniform acquisition of images in context.*

The system relies on a vision-based approach with captured images containing information about the infrastructure condition. The image data should be processed in an automated and reproducible fashion (DR3 & DR4). Image processing is necessary to process and analyze the data in order to extract the desired information. Image processing has traditionally been implemented for industrial applications like quality control of manufactured parts, as they exhibit inherently less challenging lighting conditions and scene complexity than outdoor environments (Mirallès et al., 2014). Owing to the rapid growth and evolution of DL (LeCun et al., 2015; X. Liu et al., 2020) in general and CNNs in particular, there are many successful approaches that have improved the performance of visual recognition systems in application areas such as self-driving cars, face recognition, image search, and image understanding (Nguyen et al., 2018) despite the challenging conditions of outdoor application. CNNs provide a method for automatically learning features in images, which can drastically reduce the effort in hand-designing solutions and improve generalization. In summary, this makes its application promising for the analysis of images containing PLCs (Jalil et al., 2019; Prates et al., 2019; Sampedro Pérez et al., 2019). Consequently, based on the assumption that all relevant components are captured in images, they can be extracted using DL. In particular, the assessment of a component's condition features is determined by two factors. First, the component needs to be detected in the captured image, containing one or more component objects. Second, each detected component requires component-specific fault diagnosis. The system should therefore include these two tasks performed by a DL approach.

> **DP2 – Image processing:** *Provide the system with state-of-the art deep learning for the detection and fault diagnosis of components.*

Images containing PLCs form the basis of the IB-DSS for vision-based maintenance. However, without any additional information the images can hardly be seen as sufficient for a system designed for component maintenance. To enable MEs in their decision-making, *metadata* (Sen, 2004) regarding the images or contained components is required. The primary purpose of this metadata is to provide context (DR2) to the reported data and therefore provide enriching information that leads to knowledge creation (Kassam, 2002). It can describe both physical (e.g. towers and insulators) as well as digital objects (e.g. images and documents) through providing values or information for certain characteristics (Clobridge, 2010). The main purpose of attaching metadata to a data item is to uniquely identify it in a system and to find it by browsing or searching (Burgin, 2016). In the PLM, metadata can range from geographical and temporal image tags all the way to geographical location, age, history, et cetera of the individual infrastructure components. However, the main consideration to be taken here is that the physical objects, such as towers, insulators, or conductors, are to be considered the focal data as they represent the maintained infrastructure. The captured images contain information about these components and should therefore be appropriately linked, at best based on the individual component.

DP3 – Metadata: *Provide the system with metadata.*

The availability of context in the form of simple metadata such as the geographic location and a time stamp or advanced/processed metadata such as the object location, object type, and binary fault presence adds valuable information to an IB-DSS. However, in terms of context for the individual fault contained in an image, these details are of limited help. In the light of fault diagnosis, the required context (DR2) should be defined as parts of it that can be accessed to clarify and understand the fault. The combination of the contextualized fault diagnosis as well as visualization of the extracted information (DR5) directly results in necessary interpretability of the decision in the fault diagnosis. Consequently, the decision of the fault diagnosis should be interpretable for MEs such that they are able to comprehend why for instance an insulator was marked as faulty. Thereby, we adapt the definition of Miller (2019)[p.14] referring to interpretability as “the degree to which an observer can understand the cause of a decision”. The interpretability of the results of the fault diagnosis provides MEs with additional information (context) at a PLC level which in turn enhances their ability to make high-quality decisions.

DP4 – Interpretability: *Provide the system with interpretable fault diagnosis.*

To facilitate decision-making in PLM, we found that acquired and processed data should be visualized (DR5) to the respective users in order to determine a fault's existence, location, and significance. Because such a user interface can be considered as the "source of many of the power, flexibility, and ease of use" (Turban et al., 2010, p.100) of a DSS, it requires careful consideration. MEs face a situation where they need to compile well-defined, complete, and prioritized maintenance orders with a variety of details and latent information requiring their consideration. An appropriate interface should therefore harness visual data exploration (Keim, 2002) by integrating its user into the data exploration process by applying their perceptual abilities. It can help the personnel to answer the mission critical questions such as the required equipment and achieve high decision quality regarding maintenance prioritization.

> **DP5 – Visual data exploration:** *Provide the system with an interface for visual data exploration.*

6.5 Image-Based Decision Support System for Vision-Based Power Line Maintenance

To improve the planning and scoping of individual maintenance orders, enhanced data and information quality needs to be provided to MEs. By following the prescribed tentative DPs for an IB-DSS our designed and evaluated artifact provides evidence of achieving this objective. The artifact is integrated into our case company by deriving specific capabilities to satisfy the DPs , termed design features (DFs) (Meth et al., 2015). Accordingly, we present the image data collection, their subsequent processing and analysis through the MC, and the presentation of the results through the UIC along with their respective DFs depicted in Figure 6.2 (cf. page 117) in the following three subsections.

6.5.1 Image Data Collection

The platform responsible for the image data collection is required to capture images of sufficient quality. Consequently, it needs to be able to acquire uniform, standardized, and consistent image data in a potentially automated way (DP1). UAVs equipped with capabilities to capture optical images (DF1) meet these expectations (Matikainen et al., 2016; Nguyen et al., 2018; Spencer et al., 2019) for our specific use case. This is due to three main reasons. First, UAVs are able to capture images from above, below, and the front, combining the best aspects of today's helicopter, ground, and climbing inspection methods. Second, a UAV's ability to fly close to power lines allows it to take detailed images. Finally, although an approach for UAVs' autonomous navigation and image acquisition along power lines still has to be developed, the general feasibility of this automation step is undisputed (Nguyen et al., 2018).

6.5.2 Deep-Learning-Enabled Model Component

To build an efficient IB-DSS for infrastructure maintenance, images containing relevant components, meaning components that exhibit faults, need to be identified from the entire dataset. For this purpose, we present the preparation, instantiation, and evaluation of our MC below.

Data Description and Preparation

To build a DL vision-based MC, large quantities of data are required. We therefore collected images of PLCs, annotated them according to our desired component classes, prepared them for training through creation of several datasets, and finally used them for model training.
The images were collected by flying a UAV along high-voltage power lines in several selected areas in southern Germany and circling around power towers to take pictures of PLCs. The power line passages were selected so that the captured images would contain diverse background scenes and PLCs of varying age and type. For each power tower, around 70 images were captured. Images containing faulty safety pins were created artificially in collaboration with field experts. Accordingly, an insulator and fitting application was installed in the lower area of one power tower (see Figure 6.3 on page 123 — left image) and a sequence of 608 images was captured while modifying the splint itself as well as changing the respective image perspective.
After collecting the images, each one was annotated with bounding boxes (BB_{GT}) representing the ground truth. Each BB was associated with one of five PLC classes ($insulator$, $fitting_{top}$, $fitting_{bottom}$, $birdnest$, $safetypin$) that we chose for this project. These annotations and the respective images eventually constituted our root dataset DS_{Ro}, containing 1,424 $insulators$, 1,073 $fittings_{top}$, 1,438 $fittings_{bottom}$, 61 $birdnests$, and 5,186 $safetypins$. Two further datasets $DS1_{Co}$ and $DS2_{Fi}$ were

obtained through subsampling DS_{Ro} to train different aspects of the object detection as depicted in Table 6.5. Finally, $DS3_{Pi}$ was derived to train the classifier for $safetypins$, with 1,494 images of defective and 3,692 images of intact $safetypins$. The characteristics of the four datasets are summarized in Table 6.5 and sample images are shown in Figure 6.3 and Figure 6.4 on page 124.

Tab. 6.5.: Characteristics of the datasets.

Dataset	# Images	Image resolution	Volume	Annotation type	# Annotation	Objective
DS_{Ro}	1690	5280x3956	15.2 GB	BB + label	9182	Single-stage component detection ($insulator$, $fitting_{top}$, $fitting_{bottom}$, $birdnest$, $safetypin$); derive data set $DS1_{Co}$, $DS2_{Fi}$, and $DS3_{Pi}$
$DS1_{Co}$	1589	5280x3956	14.3 GB	BB + label	3996	Multistage large component ($insulator$, $fitting_{top}$, $fitting_{bottom}$, $birdnest$) detection
$DS2_{Fi}$	1820	1200x1200	1.2 GB	BB + label	5186	Multistage small component ($safetypin$) detection from cropped $fitting_{top}$ and $fitting_{bottom}$
$DS3_{Pi}$	5186	60x60	35.3 MB	Label	5186	$safetypin$ fault diagnosis

Fig. 6.3.: Exemplary images of the DS_{Ro} dataset containing $insulators$ (cyan), $fittings$ (blue and dark blue), $birdnest$ (not present), and $safetypins$ (pink). The images show various subcomponents of the component types, captured from varying perspectives to ensure the robustness of the model; the left image provides an impression of the artificial setup for capturing defective $safetypins$.

Instantiation of a Multistage Pipeline

Inspired by Nguyen et al. (2019) and X. Liu et al. (2020), we designed a DL-based multistage component detection pipeline (MSCD) and classification pipeline for high-resolution images containing multisized objects with spatial relationships (DF2 & DF3) to satisfy DP2. This addresses the requirement for automation (DR3) of

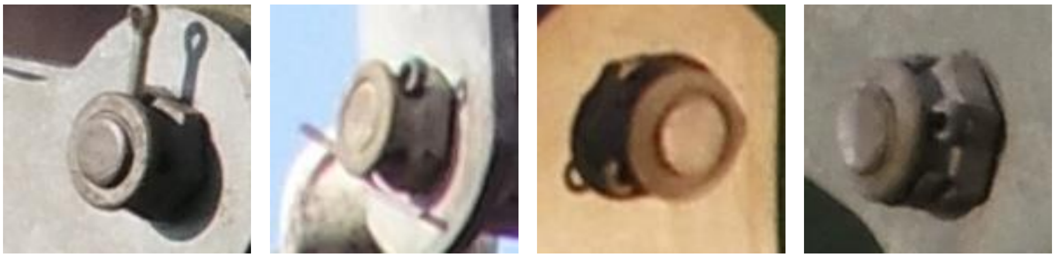


Fig. 6.4.: Exemplary images of the $safetypin$ component type from the $DS3_{Pi}$ dataset. The defective $safetypins$ (two to the left) are not completely bent, while the intact ones (two to the right) are completely bent and consequently prevent slipping out.

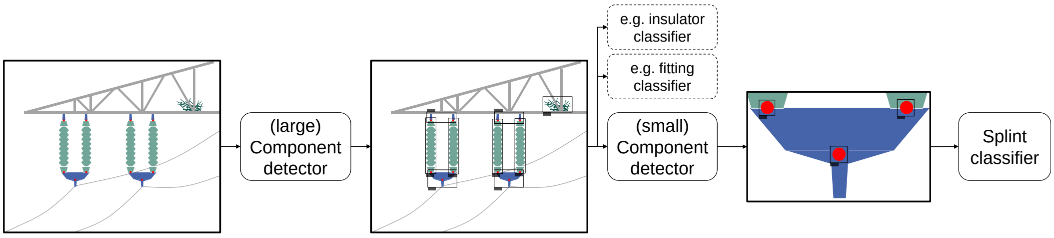


Fig. 6.5.: Structure of our multistage power line component detection and classification pipeline for high-resolution images.

infrastructure inspection (Katrasnik et al., 2010; Montambault et al., 2010) and reproducibility (DR4) of the derived results to mitigate subjective decisions (Katrasnik et al., 2010; Toth & Gilpin-Jackson, 2010). While our case company is interested in the fault diagnosis of a significantly larger number of components, for the purpose of this study we intend to only demonstrate the feasibility of detecting both the smallest components ($safetypins$), as well as the largest ones ($insulators$), in images taken of high-voltage power lines — a topic not yet considered in the automated inspection of power lines. The pipeline consists of three elements responsible for different detection and classification tasks, as displayed in Figure 6.5.

In the proposed MSCD pipeline, the *(large) component detector* first detects $insulator$, $fitting_{top}$, $fitting_{bottom}$, $birdnest$ from an input image. The detected $fittings$ are cropped from the input image and used as input for the subsequent *(small) component detector* to detect $safetypins$. The detected $safetypins$ are re-cropped and passed into the *pin classifier* for fault diagnosis.
For the implementation of the MSCD, we chose to compare two well-proven DL object detection architectures — SSD (W. Liu et al., 2016) and Faster R-CNN (S. Ren et al., 2015) — which we additionally benchmarked against a single-stage component detection pipeline (SSCD), meaning all components are detected in one step. We selected ResNet as the backbone CNN for the object detection architectures as well as our main classifier for the fault diagnosis of the $safetypins$. To compare

and benchmark the fault diagnosis, we chose the well-known VGG16 (Simonyan & Zisserman, 2014) architecture. In both tasks, image augmentation was used to improve the generalization of the models. For object detection the brightness of the images was randomly adjusted. For the classification task, where cropped images of $safetypins$ were classified, we applied horizontal and vertical flipping, random brightness adjustment, width as well as height range shifting, and random image blurring.
The component detectors were implemented using the Tensorflow[4] DL framework[5] (Abadi et al., 2016) with models pretrained on the MS COCO dataset (Lin et al., 2014). The image classifiers were realized using the Keras DL library[6] (Chollet et al., 2015) which provides image classification models pretrained on the ILSVRC dataset (Russakovsky et al., 2015).

Evaluation of the Instantiated Model Component (EE I.I & EE I.II)

For the evaluation of DF1-DF3 and DP1 and DP2 respectively, we conducted both an artificial evaluation to closely assess the pipeline's efficacy and efficiency as well as a naturalistic evaluation to generally judge the design's acceptance and usefulness. In accordance, the evaluation episodes were guided by the questions below:

> **EE I.I**
> How well does the proposed DL-based MC extract power line components of various sizes? How well does it diagnose component faults?

> **EE I.II**
> Do MEs regard the MC's capabilities as helpful?

Artificial evaluation of the model component (EE I.I)
The efficiency evaluation of the proposed pipelines required two considerations. First, the pipeline's ability to detect the chosen components needed to be evaluated. Second, the accuracy of the fault diagnosis — which we performed for detected $safetypins$ — had to be assessed.

[4]Version 1.15
[5]In particular, the tensorflow object detection API
[6]Version 2.3.1

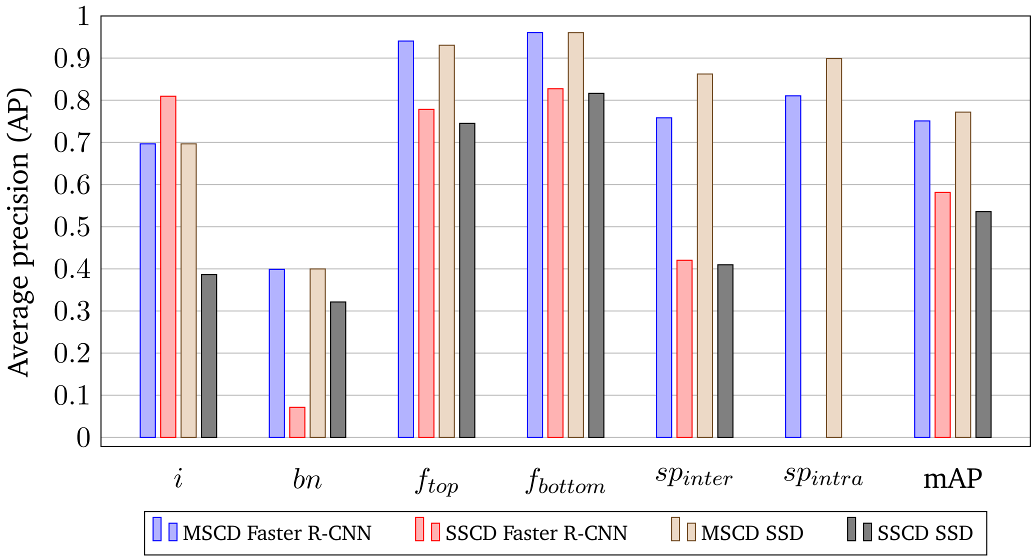


Fig. 6.6.: Average precision of the different pipelines using Faster R-CNN and SSD on the five selected component classes.

Evaluating the efficacy and efficiency of the detection task in terms of average precision (*AP*) and mean average precision (*mAP*) (Padilla et al., 2020), we compared our proposed MSCD to the SSCD pipeline. As we were working with our own proprietary dataset DS_{Ro}, the available images were split into a training set comprising 80% of the data, with the remaining 20% used for the evaluation set. To increase the evaluation's validity, images captured at one tower were held out from the random split and solely utilized for the evaluation dataset, while maintaining the split ratio. This image-level split was kept consistent across the derived datasets $DS1_{Co}$ and $DS2_{Fi}$. The SSCD pipeline was fine-tuned to detect the respective component classes using the DS_{Ro} dataset. Accordingly, both detection stages of the MSCD pipeline were fine-tuned on $DS1_{Co}$ and $DS2_{Fi}$ respectively. All models were trained using the stochastic gradient descent optimizer with 0.0003 (Faster R-CNN) and 0.001 (SSD) initial learning rate respectively, 0.9 momentum, and batch size 64. We determined the models by using early stopping on the validation loss with a patience of 100 for all models. The testing results of the different pipelines using the different architectures are shown in Figure 6.6. The performance for the $safetypin$ class is disclosed in terms of *inter* pipeline performance for both the SSCD and the MSCD pipeline as well as the *intra* pipeline performance for solely the MSCD pipeline.

We evaluated the fault diagnosis task performed for the $safetypins$ class in terms of weighted precision, weighted recall, and weighted F_1-score (Pedregosa et al., 2011) to account for class imbalance. We applied a 3-fold cross validated grid-search to identify the optimal combination of parameters. We chose to account for the following parameters: unfrozen convolutional layers, dense layer size, optimizer and its respective learning rate, dropout rate, and batch size. The images in dataset $DS3_{Pi}$ were shuffled, a hold out set containing 10% of the images was retained and the remaining images were split into 3 folds. Consequently, for each grid search configuration three models were trained with early stopping with patience 30. The best resulting model of the Resnet and VGG16 model were harnessed to be evaluated on the retained hold out set. The results of the evaluation of the cropped $safetypin$ classification task based on the test set are shown in Table 6.6. All details on the machine learning steps and choices are depicted within Section 6.8.2 on page 139 in the Appendix (Kühl et al., 2021).

Tab. 6.6.: $Safetypin$ crop classifier test results on the $DS3_{Pi}$ dataset.

Architecture	AUROC	Weighted precision	Weighted recall	Weighted F1-score
VGG16	0.8114	0.80	0.80	0.80
ResNet50	0.8080	0.76	0.75	0.75

Naturalistic evaluation of the model component (EE I.II)
To answer whether the detection and fault diagnosis of PLCs help MEs, we conducted nine purposefully sampled (Coyne, 1997) interviews with potential users of the IB-DSS from our case company. The interviewees included two senior MEs (Epsilon – Zeta) with a working experience of 34 and 41 years, five MEs (Eta – Lambda) with on average 27 years experience, one operations manager (My) with 28 years' working experience, and one senior standardization engineer (Ny) with 10 years' working experience. Each interviewee received a brief introduction to the DF1-DF3. Accordingly, the image data collection setup employing UAVs and the image analysis to detect and diagnose PLCs was introduced. Exemplary images (cf. Figure 6.3, page 123) were shown to clarify the use case. The interviewees were allowed to ask questions of comprehension. Subsequently, in a semi-structured interview fashion each participant was asked to evaluate the DFs. A detailed overview of the questionnaire can be found in Section 6.8.3 on page 141 in the Appendix. The question of whether each presented DF appropriately addresses its respective DPs served as the starting point. The interviewees opinion and attitude regarding all DFs was explored and probing questions were asked if necessary. This allowed us to assess the attitude of human expert workers towards the technology. This initial evaluation of part of the IB-DSS's tentative design serves as initial mediation to ensure that the final artifact can be designed as a useful and efficient instrument for solving our research question.
In accordance with Hevner and Chatterjee's (2010) suggestion for the analysis of confirmatory focus groups and King's (1998) general proposal of *template analysis* for textual data, we adapted the approach for the analysis of the interview transcripts. The artifact's DPs served as the initial coding categories.

In general, the interviewees confirmed the usefulness of the way the **image acquisition (DP1)** is performed and also acknowledged the **image processing (DP2)** to extract comparable, trustful, and helpful information. They specifically confirmed the usefulness of the vision-based approach for capturing a wide variety of different faults. More significantly, the ability to "[...] look into the detailed pictures is already of high value" (Iota) since it is easier to scope maintenance operations from component images rather than plain table entries. Additionally, the interviewees emphasized the good quality of the images as well as the improved perspective to view the PLCs and respective defects, due to the UAVs being able to fly close to the component of interest. Similarly, the functionality to automatically analyze the images for components and their faults was perceived as a major gain and precisely addressed the request of interviewee Zeta: "It would actually be quite interesting if someone or something evaluates these pictures that the drone captures and then just sends the damage." The interviewees stressed several particular factors. First and foremost, the prevention of subjectivity was mentioned, leading to a uniformity in fault diagnosis and consequently to a flawless comparability between faults. Second, besides the presented ability to detect $insulators$, $fittings$, $birdnests$ and $safetypins$, the interviewees assumed that several other components could be added easily. However, in more detail two participants raised doubts about the system's ability to recognize severe incidents such as completely broken and consequently dangling insulators. Finally, six out of the nine participants indicated, without being asked, that they felt there were benefits in using an automated process to extract defective components. They specifically mentioned benefits regarding timeliness, cost, and performance in comparison to the current manual inspection methods. However, although the proposed extraction of faults generated generally positive feedback, the need to "comprehend: how did this assessment come about" (Ny) was mentioned. Consequently, both the results and the reasoning of the fault diagnosis require visualization.

6.5.3 User Interface Component

Supporting MEs based on improved data and information quality requires making them accessible through a UIC. In the following, we describe the UICs' design and evaluation.

Instantiation of the User Interface Component

To create a UIC that accomplishes the preliminary DRs of visualization (DR5) of the network and related defects (Shakhatreh et al., 2019), we implemented the artifact based on the inferred DPs (cf. Figure 6.2 on page 117) using Tableau[7] and JavaScript. The artifact integrates two data sources: (1) UAV-captured image data (DP1) and its according metadata (DP3) as well as (2) metadata about the physical objects of the power line infrastructure (DP3) at our case company, such as geographical position or age. Information that is extracted as part of the image processing (cf. Section 6.5.2) is integrated into the artifact (DP2 & DP4). Finally, these building blocks are arranged in a meaningful way to support decision-making through visual data exploration (DP5). Figure 6.7 depicts the different views and their interactive links along with the respective DFs.

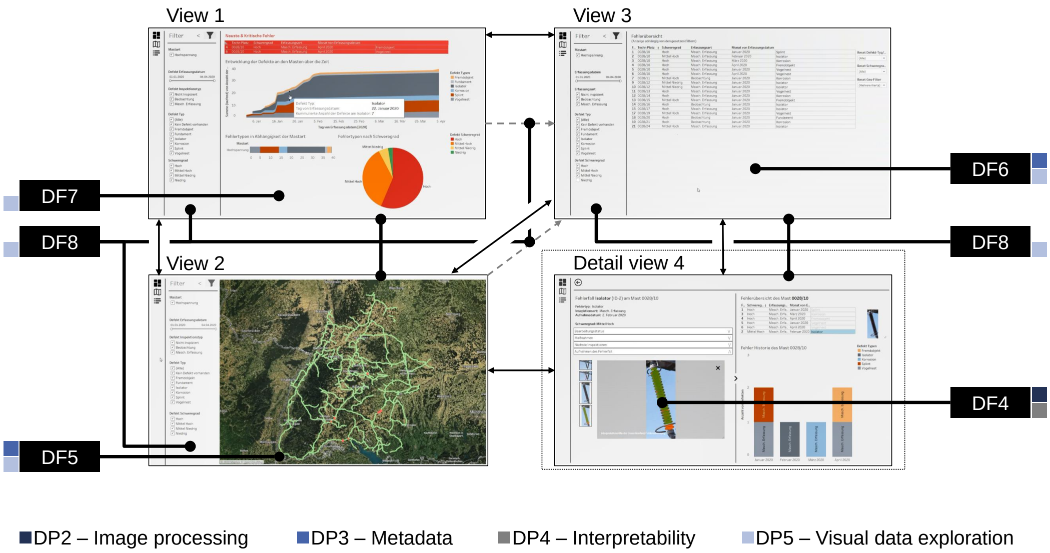


Fig. 6.7.: Structure of our user interface with its different views and the transitions between the views.

[7]https://www.tableau.com/, last accessed: 07.07.2022

To satisfy DP5, the general layout of the UIC should follow the visual exploration paradigm (Shneiderman, 1996) and provide overview first, allow for zoom and filter capabilities, and then accommodate details on demand. We base our UIC on four different views which emphasize different task properties in our multidimensional data and maximize the availability of explicit and latent information. View (1) provides an operational dashboard view (DF7) to get a quick and aggregated sense of the condition of the power line infrastructure. View (2) contains a geographical information system view (DF5) to find and inspect adjacent infrastructure items and faults. This allows MEs to explore both the incident location to determine maintenance order specifics as well as further incidents which can be bundled. View (3), a list-based view (DF6), enables MEs to examine a large number of faults regarding their attributes as well as to find specific faults. This may help in either bundling incidents, making sourcing decisions upon resource scarcity, or ordering replacement components. View (4) presents a fault detail view to inspect particular faults regarding the results of the fault diagnosis, including properties, specifics, and context. It consequently enables MEs to assess the faults priority, judge the skills required for the faults resolutions, and determine the affected circuit. The interactive visualization (DF8) allows MEs to directly interact with the visualizations to obtain and extract the relevant data at the right time. A persistent filter sidebar with domain-specific filters provides consistency across the first three views. While View 1 through 3 already provide different levels of zoom, the list-based view is the closest to viewing a single fault. Users are therefore able to filter subsamples of faults in View 1 as well as 2 and through interactive linking consequently invoke their display in the list-based View 3. Finally, detailed information on a particular fault identified either in View 2 or 3 can be examined. Images of the defective component are available in a gallery. To address DP4, the gallery provides the user with visual fault explanations (DF4) of the component for improved interpretability of the fault diagnosis. In particular, the detected defective component is framed by a bounding box for convenient localization. Additionally, based on the type of fault either a segmentation mask (for insulators) or a heat map (for splints) is visualized. Besides the image gallery, the user is able to expand related information showing other faults on the power tower and the fault timeline of the power tower. As a summary, a video demonstration of the user interface shows all described views in detail[8].

[8]https://youtu.be/Y3oIJghtRT4, last accessed: 07.07.2022

Evaluation of the User Interface Component (EE II)

For the evaluation of our UIC, we applied a qualitative evaluation to test the proof of applicability in the real-world context and to assess the usefulness as well as efficiency. In particular we aimed to answer the evaluation question:

> **EE II**
> Does the instantiated UIC support MEs in making improved decisions about planning and scoping individual maintenance orders?

To answer this question, we remotely[9] conducted nine one-on-one, confirmatory workshops with the same participants already questioned in EE I.II over the company's collaboration platform. This confirmatory evaluation approach was chosen for two reasons. First, the flexibility of the method enabled us to adapt the procedure if necessary. Second, each user was able to individually explore and use the prototype in their accustomed work setting, which allowed the integration into the user's working routine and ensured that the artifact and its capabilities were understood unambiguously.

For each workshop, we initially introduced the intent of the UIC. We subsequently started a screen sharing session and asked each participant to explore and use the UIC and verbalize their thoughts. Whenever appropriate, the researcher enriched the participant's experience by providing information about the DFs. Afterwards, each participant was asked to fill out a survey based on Davis's (1989) technology acceptance model (TAM). Finally, the participant was asked to evaluate whether the presented artifact addresses its decisive DPs during a semi-structured interview. The question of whether each presented DF appropriately addresses its respective design cycles (DCs) served as the starting point. The transcripts of the workshops were analyzed in analogy to E I.II, using template analysis by King (1998).

The survey results as well as the results from our qualitative evaluation indicate that our instantiated artifact is able to support MEs in their decision-making regarding PLCs. While our TAM survey comprising the nine interviewed experts cannot claim significance, it suggests the tool's usefulness as the perceived usefulness averaged 6.2 on a 7-point Likert scale. In accordance, the interviews revealed that the artifact would support the MEs in their everyday work by enhancing the availability of data and information of the power line infrastructure and the appropriate arrangement of the information. The confirmatory workshops therefore showed that the underlying design knowledge is suitable, useful, and effective for developing IB-DSSs artifacts aimed at the vision-based maintenance of infrastructure.

[9]due to COVID-19

In particular, the participants mentioned that the IB-DSS allows fast and convenient **visual data exploration (DP5)** while being helpful to experienced workers as well as (and especially) those in training. The interviewed experts mentioned that the artifacts' capabilities for overview, interactive zooming, and interactive filtering are the main facilitators for convenient exploration. The interactive zooming across the multiple views makes latent information, for example staggering faults on one passage or the circumstances around a tower, visually available. Finally, the filter capabilities support finding relevant faults, as "[one] can filter out the unimportant ones" (Eta). However, six participants requested additional filters based on further metadata concerning the components in the infrastructure. While the available **metadata (DP3)** regarding towers and their identified faults was perceived as a good starting point, all participants mentioned further data which could be integrated: fault-related workflow tracking metadata as well as component-related material and reordering metadata. The participants also recognized that the visual fault explanations could mainly help them localize faults significantly faster as well as develop a thorough comprehension and understanding of the fault. Specifically, it was mentioned that the easier localization could reduce the workload and accelerate the root cause analysis. On the downside, it could hinder independent examination of the images in the long run. The image augmentations consequently provide fault **interpretability (DP4)**. Most significantly, all participants acknowledged that the IB-DSS is especially suitable for improved maintenance decision-making, as they would be able to "work more efficiently, simply work more or even combine activities" (Epsilon). In fact, besides the planning and scoping of individual maintenance orders, the improved data and information availability and quality enhance four key decision-making tasks: finding and discovering systematic faults (Epsilon, Iota, Lambda), combining maintenance orders (Eta, Epsilon, Kappa), discussing maintenance budget (My, Iota, Ny), and scoping and planning long-term restoration projects (My, Kappa).

6.6 Discussion

In this section we depict the contributions and limitations of our work and present an outlook regarding PLC inspection and maintenance.

6.6.1 Contributions

Our results imply that our instantiated IB-DSS enables maintenance engineers to make better, more informed decisions about repairing or replacing PLCs through improved data and information quality.
More generally, this suggests that the rich information from uniformly acquired images extracted through deep-learning-based image processing capabilities combined with contextual information of metadata and interpretability provided through visual data exploration is a valuable solution to the information intensive context of maintenance and monitoring applications. Figure 6.8 depicts the schematic layout of these aspects. Consequently, we hypothesize that our derived knowledge provides a nascent design theory for the still underresearched class of IB-DSSs. This design knowledge might be particularly valuable to create automated decision support systems in information-intensive contexts where decision-makers largely rely on unstructured vision-based image data. This in turn would increase the quality of decision both in terms of efficiency and effectiveness (Kraus et al., 2020).
The schematic layout of our conceptualized design principles, as depicted in Figure 6.8, therefore provides prescriptive knowledge that may serve as a blueprint (Gregor & Jones, 2007) to develop similar systems for vision-based applications.

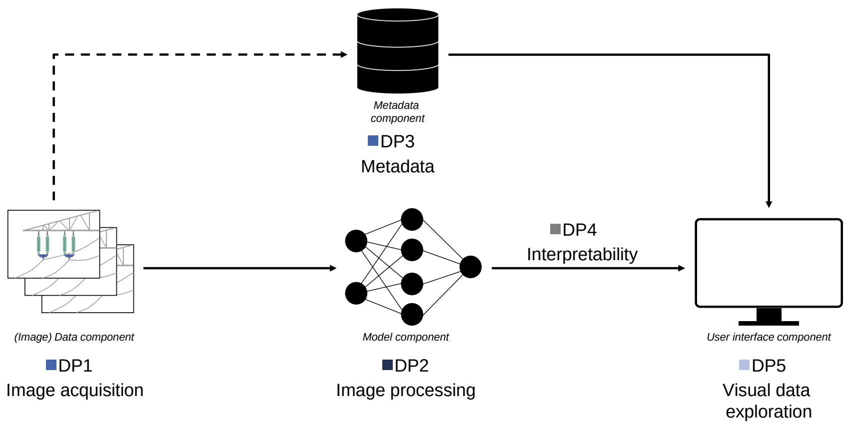


Fig. 6.8.: Schematic layout of the design principles of image-based decision support systems.

In our specific use case of PLC inspection, the proposed IB-DSS relying on UAV generated images can provide multiple benefits over the status quo. It can prevent accidents since hazardous inspection methods like tower-climbing are no longer necessary — as the inspection of the towers is now performed by unmanned UAVs. While no coherent numbers are available within Europe, recent reports from the US demonstrate that power line workers are listed among top 10 most dangerous jobs. Each year, over 40 power line workers receive fatal injuries resulting from falling or electrocution (Schwarz & Drudi, 2018). While certainly only a share of these workers die during inspection activities (rather than the repair activity itself), any saved life is desirable. The non-fatal injuries amount to 1,200 per year in the US (Schwarz & Drudi, 2018) and the typical reasons are falling, slipping and tripping. We also expect significant reduction of injuries in this area, once automation of inspection is implemented.

Currently, the data that MEs work with are tables of compiled inspection reports with heterogeneous assessments of a distributed workforce. The standardized data acquisition and processing results in (1) more reliable and (2) more structured data. Combined with the benefits of a unified interface that provides metadata and latent information maintenance decisions are fully comprehensible.

In total, the participants of the confirmatory workshop affirmed that the IB-DSS enhances their decision-making substantially. As mentioned by Epsilon, Theta, Kappa, and My even besides the pure planning of maintenance orders, moreover, the artifact could be utilized for other tasks, like the combination of maintenance orders or the planning of long-term restoration projects.

6.6.2 Limitations

While meeting Gregor and Jones's (2007) six common criteria for design theories, our design knowledge for IB-DSSs carries limitations that open opportunities for future research. Our research can only be generalized to a limited extent because it was conducted at one company in the power line infrastructure domain and focused on a selection of defect cases. While we can claim some generalization through supporting our design through kernel theories and other studies, further IB-DSSs should be developed for other use cases and in other domains to extend and consolidate the design theory. Furthermore, our research lacks the quantification of the effect on the field efficiency of the image processing. Quantitative studies in this regard could be conducted to benchmark the artifacts' effects in terms of performance of automated versus manual image processing.

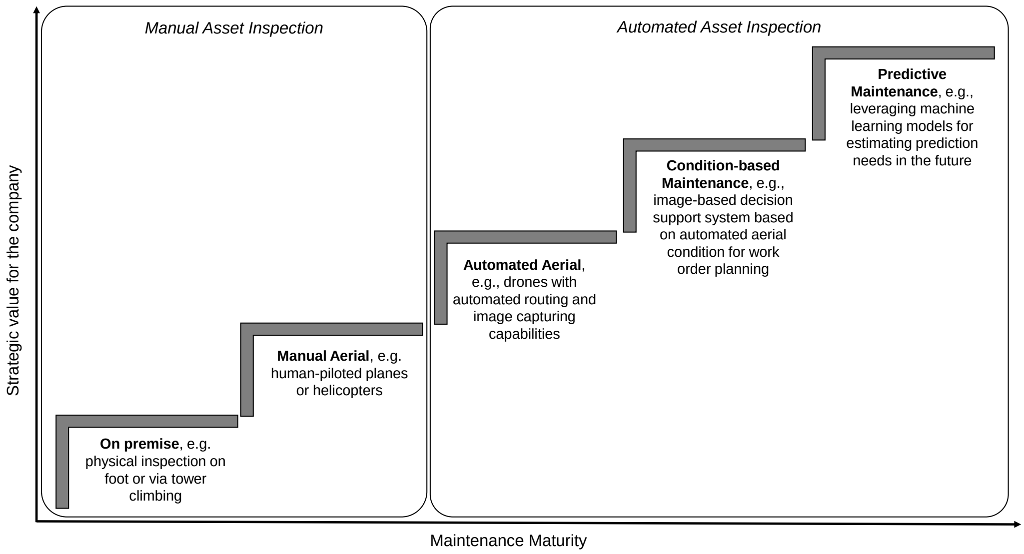


Fig. 6.9.: Road map towards predictive maintenance.

6.6.3 Future Design Activities

Within our presented research, we showed novel ways to design condition-based maintenance systems. More precisely, we utilized images captured by UAVs which are subsequently automatically analyzed and included within an image-based decision support system. Figure 6.9 shows a possible general road map demonstrating increasing maintenance maturity where the next evolutionary step is to use the data as well as the generated models not only as a basis for maintenance order planning, but moreover to predict maintenance needs for the (distant) future, i.e., predictive maintenance.

In regards to the practical aspects at the case company, the artifact is currently prepared for a broader implementation and deployment into the business. For these steps, the solution is containerized (Rufino et al., 2017) to allow for flexible and scalable applications. An expert team analyzes the different possibilities of automated UAV routing (Avellar et al., 2015) to allow for a continuous and correct gathering of the required data. Meanwhile, experts are being educated on the possibilities of integrating the tool in their current day-to-day processes, supported by an expert for change management of industrial business processes (Bokrantz et al., 2020). One remaining challenge is the aspect of data storage and management, e.g., within a data warehouse. On the basis of the required data volume shown in Table 6.5 on page 123, we estimate a total volume of images for a one-time acquisition of the complete network of our case company of 9 TB. How often this data has to be refreshed and how it is stored precisely (e.g. hybrid cloud) needs to be discussed for future iterations of the artifact.

In a broader context, the automated inspection of PLCs will be an important, yet only intermediate step for PLC maintenance in the future. The image data used in this work can be combined with multiple additional data sources such as weather and location characteristics (e.g., sun exposure and topology). The inclusion of additional sources of information can enable an accurate prediction of future maintenance needs which further facilitate effective planning and resource utilization.

6.7 Conclusion

Planning and preparing of maintenance orders in power line maintenance is a challenging task for maintenance engineers, as they must rely on human-created, heterogeneous, and largely unstructured information. These characteristics make the process both time-intensive and costly, which can adversely affect the continuous supply of electricity. As most research on power line maintenance focuses on automated inspection through UAV-captured images and deep learning, there is an apparent gap in literature for transferring the acquired data into maintenance decision-making.

Following the design science research guidelines, we designed, developed, and evaluated an artifact to address this research gap. Initially, we rigorously analyzed the challenges in power line maintenance. Building on these, we conceptualized design principles for an image-based decision support system that integrates the capabilities of deep learning to extract faulty components from a set of captured images and appropriately presents the information to relevant users. Accordingly, we instantiate our design principles into an exemplary artifact. The evaluation using a technical experiment as well as two qualitative evaluation episodes with long-standing experts indicates the utility of our design knowledge and can therefore inform future system designs of similar nature.

6.8 Appendix for Chapter 6

6.8.1 Further Challenges in the Power Line Maintenance

In Table 6.7 on page 138 we present further challenges in the maintenance of power lines. Challenge C4 attributes to organizational or administrative levels of introducing novel solutions which include the consideration of the exact purpose of the infrastructure (C4.1), the cost associated to their inspection (C4.2), and specific challenges that come with the culture, digitization maturity, and capabilities of an organization (C4.3). Moreover, the challenge C5 addresses the fact that power lines are considered as critical infrastructure which's operations, inspection, and maintenance is strictly regulated. Another challenge that applies to generally all infrastructure related inspections is of environmental kind (C6). Environmental challenges include limitations in the maintenance of power lines due to weather conditions and general seasonal circumstances.

Tab. 6.7.: Further challenges in the maintenance of power lines based on expert interviews and a structured literature review.

ID	Challenge	Subchallenge	Source
C4	Organizational challenges	C4.1—Significance of uninterrupted power supply	Pagnano et al., 2013; Nguyen et al., 2018; Prasad and Rao, 2016 D. Li and Wang, 2019; Matikainen et al., 2016; Toussaint et al., 2009; Katrasnik et al., 2010; Seok and Kim, 2016; Beta
		C4.2—Scale of inspection cost	Pagnano et al., 2013; Nguyen et al., 2018; Mirallès et al., 2014; Prasad and Rao, 2016; C. Deng et al., 2014; Jones, 2005; Aggarwal et al., 2000 D. Li and Wang, 2019; Takaya et al., 2019; Pernebayeva and James, 2020; Matikainen et al., 2016; Katrasnik et al., 2010; Seok and Kim, 2016; Huang et al., 2018; Ostendorp, 2000; Alpha; Beta; Gamma; Delta
		C4.3—Company-specific challenges	Alpha; Gamma; Delta
C5	Regulatory requirements	C5.1—Compliance with regulations	Pagnano et al., 2013; Prasad and Rao, 2016; Jones, 2005; Takaya et al., 2019; Matikainen et al., 2016; Toussaint et al., 2009; Gamma
C6	Impact of environmental conditions	C6.1—Dependence on seasonal circumstances	Delta
		C6.2—Dependence on climatic conditions	Nguyen et al., 2018; Pernebayeva and James, 2020; Seok and Kim, 2016; Homma et al., 2017; Beta

6.8.2 Supervised Machine Learning Report Card based on Kühl et al. (2021)

General Information	Problem statement	Detect objects of power line components (insulator, fitting, safety pin, birdnest) from an input image and classify whether the safety pin components are intact or defect.	
	Data gathering	The proprietary data set originates from the application case company Netze BW, a distribution system operator in Southern Germany. We harnessed UAVs to capture images of their high voltage power lines as part of a technology driven proof of concept.	
	Data distribution	After annotation of the images of the proprietary data set it contains (BB = Bounding Box):	
		Insulator (BB)	1,424
		fitting_top (BB)	1,073
		fitting_bottom (BB)	1,438
		Birdnests (BB)	61
		Safety Pins (BB)	(3,692 intact/1,494 defect) 5,186
	Data quality	High-quality images with a resolution of 5280x3956 pixels. Bounding boxes and labels generated by researchers who had received instructions and feedback from field experts.	
	Data preprocessing methods	Rescaling (1/255)	
	Performance Estimation		
Object Detection Task – General Information	Parameter optimization	None	
	Data split	Training data set 80%, Evaluation data set 20% *To increase the evaluation's validity, images captured at one tower were held out from the random split and solely utilized for the evaluation dataset, while maintaining the split ratio.*	
	Sampling/ Data augmentation	Random brightness adjustment	
	Performance metric	mean average precision (mAP) (Rafael Padilla & da Silva 2020)	
Faster R-CNN	Algorithm Parameters	CNN backbone	ResNet-50
		Early stopping patience (on validation loss)	100
		Optimizer	SGD (learning rate 0.0003 and 0.9 momentum)
		Batch size	1
		Performance evaluation	0.7510
SSD	Algorithm Parameters	CNN backbone	ResNet-50
		Early stopping patience (on validation loss)	100
		Optimizer	SGD (learning rate 0.001 and 0.9 momentum)
		Batch size	64
		Performance evaluation	0.7718

Classification Task – General Information	Parameter optimization	Yes	Search space	cf. *Table 6.8*
			Search algorithm	Grid search
	Data Split	10% Hold out set 90% train and validation set with 3-fold cross validation		
	Sampling/ Data augmentation	- Average blur [0,11] - Brightness range [0.2,1.5] Height shift range [0.1] - Width shift range [0.1] - Horizontal flip: true - Vertical flip: true		
	Performance metric	Weighted precision, weighted recall, and weighted F1-score (Pedregosa et al. 2011) to account for class imbalance		
ResNet -50	Final algorithm parameters	Dense layers	(512, 512)	
		Unfrozen layers	3	
		Dropout rate	0.1	
		Early stopping patience (on validation loss)	30	
		Optimizer	Adam (learning rate 0.0005)	
		Batch size	32	
	Performance evaluation	AUROC: 0.8080 Weighted precision: 0.76 Weighted recall: 0.76 Weighted F1-score: 0.71		
VGG16	Final algorithm parameters	Dense layers	(512, 512)	
		Unfrozen layers	8	
		Dropout rate	0.1	
		Early stopping patience (on validation loss)	30	
		Optimizer	Adam (learning rate 0.0005)	
		Batch size	32	
	Performance evaluation	AUROC: 0.8114 Weighted precision: 0.80 Weighted recall: 0.80 Weighted F1-score: 0.78		

Tab. 6.8.: Parameter set options for the training of convolutional neural network for both ResNet-50 and VGG16.

Parameters	Dense layers	Unfrozen layers	Optimizer	Learning rate	Batch size	Dropout rate
Ranges	((512, 512) (512, 1024) (512, 2046) (1024, 1024) (1024, 2046) (2046, 2046))	[3, 8]	Adam, SGD	(0.0005, 0.001, 0.0015, 0.002)	(32, 64, 128)	[0, .6]

6.8.3 Questionnaire (translated from German to English)

1. *General introduction:*
   How was the “application” of the the image-based decision support system (IB-DSS) for you?
   - Would you use the IB-DSS in practice?
   - What worked particularly well?
   - What did not work well?
   - What possibilities result from the application of the IB-DSS?
   - What problems could occur during the application?

   #### Model component:

2. **Design principle 1:**
   Unmanned aerial vehicle (UAV) & RGB images
   Was the quality of the RGB images of the UAVs sufficient and do they enable a good overview over the most important properties / the condition of the infrastructure?
   - Were you lacking important images / information / data?
   - Were the images from the UAVs well inspectable?
3. **Design principle 2:**
   Deep learning for computer vision
   Does the usage of machine / deep learning enable additional, helpful information (e.g., severity, defect type, etc.)?
   - What chances and risks arise through the IB-DSS?
   - What strengths and weaknesses does the IB-DSS possess regarding the recognition of faulty / defect components?

User Interface Component:

4. **Design principle 3:**
   Interpretability
   How did the accentuation of the condition of components influence the interpretability? Was it a good assistance to understand the result?

   - Is the IB-DSS a good tool to comprehend the condition of a component?
   - What are the advantages of the visualization?
   - What are the disadvantages of the visualization?
   - What problems can occur due to the visualization?

5. **Design principle 4:**
   Exploratory (data) visualization
   Does the IB-DSS enable an investigation / exploration of the data? Does it facilitate to gain information about the condition of power lines and a corresponding overview?

   - What are the strengths and weaknesses of the visualisation of the condition data?
   - How important is the availability of the data?
   - What long-term chances and risks do you see regarding the maintenance process?

6. **Finalization**
   Anything else you would like to share with me?

   - Where do you see room for improvement?
   - Is there any other feedback you would like to share?

7 A Picture Is Worth a Collaboration: Accumulating Design Knowledge for Computer-Vision-Based Hybrid Intelligence Systems[1]

7.1 Introduction

Intelligent information systems (IS) that incorporate artificial intelligence (AI), so-called AI-based IS (Maedche, Legner, et al., 2019), have a massive impact on our society and revolutionize how we live, act, and work together. Cars begin to drive autonomously in real traffic (Grigorescu et al., 2020); smart home systems recognize and adapt to individual user preferences (Fischer et al., 2020); and medical assistance systems support doctors in diagnosing hard-to-find diseases (McKinney et al., 2020). A key enabler for the realization of such scenarios is the capability of modern AI-based systems to automatically process high-dimensional data to identify useful patterns and relationships that can be utilized for decision support and business automation purposes (Brynjolfsson & Mcafee, 2017).
An important sub-field in this context is the area of computer vision (CV). It seeks to automatically extract useful information from images to mimic human capabilities of visual perception (Szeliski, 2010). On this basis, time-consuming and labor-intensive tasks like the recognition, detection, localization, tracking, and counting of objects can be supported more efficiently to save unnecessary resources and relieve the burden of human workers (Heinrich, Roth, et al., 2019).

[1]This chapter comprises an article that was published as: Zschech, P., Walk, J., Heinrich, K., Vössing, M., and 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*. https://aisel.aisnet.org/ecis2021_rp/127/. Note: The abstract has been removed. Minor edits have been made and tables and figures were reformatted, and newly referenced to fit the structure of the thesis. Chapter, section and research question numbering and respective cross-references were modified. Formatting and reference style was adapted and references were integrated into the overall references section of this thesis.

Fueled by the broad availability of huge online image databases and the broad access to necessary computing power, the field of CV is currently experiencing a considerable phase of scientific progress and dissemination expressed by manifold activities in research and practice. As such, we can observe a continuous development of advanced algorithms based on machine learning and especially artificial neural networks (ANNs) (Bharati & Pramanik, 2020); procedure models and tutorials for solution development are proposed (Griebel, Dürr, et al., 2019); and a global community of developers shares reusable software code and provides user-friendly programming frameworks (Chollet, 2017). As a result, more and more CV systems are being embedded into organizational and societal contexts across a wide range of domains, such as traffic surveillance (W. Liu et al., 2017), manufacturing (T. Wang et al., 2018), agriculture (Tian et al., 2020), and sports (G. Thomas et al., 2017).

However, past efforts in research and practice often exclusively focused on technical performance aspects when designing and developing CV systems (e.g., achieved accuracy, required computing resources), while neglecting socio-technical facets, such as transparency, control, and autonomy. Coming from an IS research perspective, such aspects are crucial, for example, to ensure that a technology is accepted by its users and that it is in line with the organization's objectives (Schaper & Pervan, 2005). Especially in the realm of designing and working with AI systems, such aspects play a fundamental role and therefore should be translated into an AI-based system design to aid the user in setting up, understanding and using autonomously operating AI systems (Thiebes et al., 2020). Dellermann, Ebel, et al. (2019) discuss the concept of hybrid intelligence (HI) that combines the complementary strengths of both sides in order to reach superior performance than would be achievable separately. Although this hybrid system design can bring out the best of both worlds, it is faced with challenges like algorithm aversion that occur due to the complexity of the AI system resulting in the distinctness of human and AI system in such a setting (Berger et al., 2021).

Against this background, this paper deals with the design of computer-vision-based hybrid intelligence systems (CV-HISs). The aim is to derive prescriptive design knowledge from a socio-technical view that should ultimately result in a (nascent) design theory (Gregor & Jones, 2007). To this end, we follow a design science research (DSR) approach (Hevner et al., 2004) and reflect on accumulated design knowledge generated in six comprehensive CV development projects. With this strategy, we pursue a reflective approach (Möller et al., 2020) based on real problem cases as encountered and solved in practical settings (Iivari, 2015). More specifically, we contribute in the following ways: (i) we conceptualize the HI collaboration in the realm of vision-based tasks to identify design-related mechanisms, (ii) we derive requirements and abstract them to meta-requirements in relation to central kernel theories in terms of justificatory knowledge, and (iii) we accumulate design principles by abstracting from specific design features which were implemented across the different CV projects.

Our paper is structured accordingly: In Section 7.2, we introduce the foundations and refer to related work. Subsequently, we depict our research approach in more detail in Section 7.3. In Section 7.4, we describe our selected CV cases, followed by the conceptualization of CV-HISs in Section 7.5. We then proceed in Section 7.6 to outline our derived design knowledge by distinguishing between four identified mechanisms as introduced by design. Finally, we summarize and discuss our contribution and present an outlook of further research opportunities in Section 7.7.

7.2 Foundations and Related Work

7.2.1 Computer Vision and Artificial Intelligence

The field of CV is concerned with the development of techniques for the acquisition, processing, analysis, and understanding of digital images to transform high-dimensional data into symbolic or numerical information (e.g., for automated decision-making). Just as humans use their eyes and brains to understand the world around them, CV attempts to produce the same effect so that computers can perceive and understand an image or a sequence of images and act accordingly in each situation. This understanding can be achieved by disentangling high-dimensional data from images using models built with the aid of geometry, statistics, physics, and learning theory (Forsyth & Ponce, 2002). Driven by personal or industrial motives, grand advances have been made in several areas such as optical character recognition, machine inspection, 3D model building, disease diagnostics, motion capture, or surveillance (Szeliski, 2010). Nowadays, CV tasks are increasingly performed by AI-based systems that rely on machine learning algorithms. Of particular interest are ANNs, which can be organized in deep network architectures consisting of multiple, hierarchical processing layers (Janiesch et al., 2021). This allows them to automatically process spatial information in raw image data and learn patterns that are relevant for prediction tasks, which is often also referred to as deep learning (DL) (LeCun et al., 2015). The automated learning of patterns by DL models is usually done in a supervised manner. This means that humans provide training data that are tagged with labels/annotations to specify the target of the learning task (Hastie et al., 2017; Sager et al., 2021). Computer vision is also gaining momentum in the field of IS. Extant publications range from use cases like road crack detection (Chatterjee et al., 2018) and automated fashion recommendations (Griebel, Welsch, et al., 2019) to guidelines for the technical aspects of CV projects (Griebel, Dürr, et al., 2019).

7.2.2 Hybrid Intelligence

Due to the rising capabilities of AI in the last decade (Russell, 2016), researchers are increasingly reconsidering much of the established design knowledge regarding how intelligent systems should be designed. As noted by Zheng et al. (2017), the development of AI is profoundly changing how humans interact with their environment and how they support their work processes. The authors introduce *hybrid-augmented intelligence* as a means to combine "human cognitive ability and the capabilities of computers". Similarly, Dellermann, Ebel, et al. (2019) define *hybrid intelligence* as "the ability to achieve complex goals by combining human and artificial intelligence, thereby reaching superior results to those each of them could have accomplished separately". The core concept of augmenting both sides, computers and humans, and thus creating a symbiotic relationship between them is also mentioned by Akata et al. (2020), Maedche, Legner, et al. (2019), and Seeber et al. (2019).
While the benefits of the relationships are depicted in the recent literature, it is also mentioned that to achieve beneficial utility; it is required to "[develop] novel interaction paradigms that exploit the strengths and overcome the weaknesses of both partners" (Terveen, 1995). However, so far, only a few researchers have formalized the required design knowledge.
Zheng et al. (2017) provide a framework that suggests a human-in-the-loop approach comprising a computer with AI capabilities that outputs a prediction along with confidence scores as an uncertainty assessment and a human decision-maker. Additionally, it is suggested that the human gives feedback to the AI system through additional data labeling. In comparison to the rather human-centric approach, Schwartz et al. (2016) provide the concept of an augmented human in the context of industrial cyber-physical systems. The human is equipped with sensors and augment VR tools like gloves and glasses to be able to act on the AI system's suggestions and provide additional feedback through collected sensor data. Dellermann, Ebel, et al. (2019) provide a structured overview of design knowledge for HI systems in the form of a taxonomy, including the dimensions task characteristics as well as human and machine learning paradigms. As a specific example, Dellermann, Lipusch, et al. (2019) outline design principles for business model validation by suggesting to combine crowd-sourced data and machine predictions. Table 7.1 on page 147 summarizes our findings of the related work of HI systems.

Tab. 7.1.: Overview of related literature and positioning of this work.

Reference	Approach	Focus	Application area	Derived design knowledge
Dellermann, Ebel, et al. (2019)	Definitional	Conceptualization	General	n/a
D. Wang et al. (2019)	Behavioral	Understanding of the field	AutoAI	n/a
Chakraborti and Kambhampati (2018)	Behavioral	Mental models	Urban search and rescue	n/a
Zheng et al. (2017)	Behavioral	Conceptualization	Multiple	n/a
Dellermann, Lipusch, et al. (2019)	DSR Strategy I	Prescriptive knowledge	Business model validation	Requirements and design principles
This work	DSR Strategy II	Conceptualization and prescriptive knowledge	Computer vision	Meta-requirements and design principles

7.3 Research Approach

DSR is a fundamental paradigm in IS research concerned with the construction of socio-technical artifacts to solve organizational problems and derive generalizable design knowledge (Gregor & Hevner, 2013). One of the ultimate goals of DSR is the formulation and consolidation of design theories (Beck et al., 2013). To communicate intermediate theoretical results of the theorizing process, so-called nascent design theories, Gregor and Jones (2007) describe important components, ranging from the purpose and scope (meta-requirements) and justificatory knowledge (kernel theories) to principles of form and function (design principles) and expository instantiations (implemented systems).

In order to derive generalizable design knowledge, two strategies are applicable (Iivari, 2015). The first strategy ("Strategy I") deals with the construction of IT meta-artifacts as a generic solution concept for a problem class in advance. In the second strategy ("Strategy II"), abstract knowledge is derived in a reflective manner, i.e., specific IT artifacts are first designed and implemented within a practical context. Subsequently, generalizable knowledge emerges during or after the design iterations of the artifact when abstracting from the specific implementation (Iivari, 2015; Möller et al., 2020). For the work at hand, we chose a reflective, practice-inspired approach in-line with Strategy II. More specifically, we analyzed CV development projects from industry and accumulated design knowledge from specific implementations while informing our findings with justificatory knowledge from well-established kernel theories to support the observed phenomena. In this way, it was intended to unveil implicit design knowledge concerning socio-technical mechanisms with respect to hybrid intelligence interactions, which so far have been little addressed in the literature on the design of CV systems.

To conduct our research, we organized a focus group with experts in the field (Morgan, 1996). For this purpose, we identified researchers in the IS community fulfilling the following four criteria: (i) active involvement in multiple CV system development projects, (ii) fundamental understanding of AI-based technologies, (iii) sufficient experience in conducting DSR projects, and (iv) solid understanding of IS-related theories. As a result, we recruited six researchers from six different institutions which were asked to participate in a series of workshops to reflect their collected design knowledge during their involvement in CV-related development projects.

Due to COVID-19, the workshops were conducted using video conferencing tools. Initially, there was no predefined structure for the full series of workshops as it only became apparent during the sessions which consecutive steps were necessary to derive generalizable design knowledge. Overall, this resulted in a total of seven workshop sessions over a period of three months, each lasting between 60 and 120 minutes. To mitigate the bias caused by opinion leaders within the group, there were several tasks to be performed by each researcher individually before presenting and reflecting the results in each workshop with the entirety of the group. The full series of workshops is summarized in Figure 7.1 and is briefly described below with regard to the performed activities and achieved results.

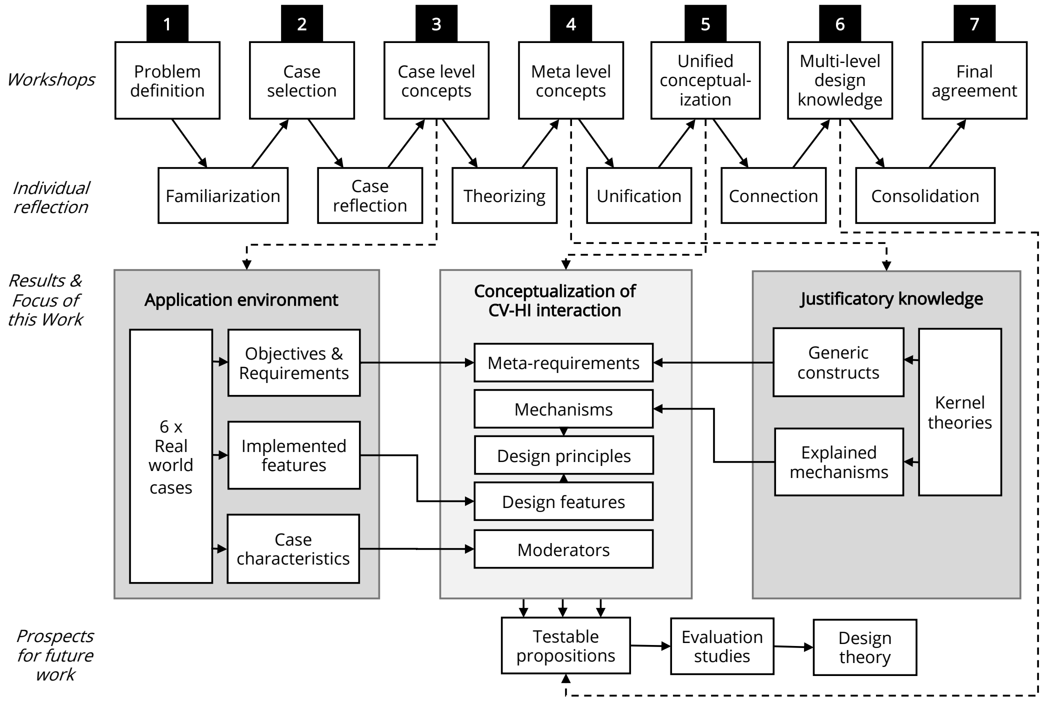


Fig. 7.1.: Applied research approach and focus of this paper.

In the **first** session, the project's initiator introduced the research idea and the group agreed upon a common understanding of relevant fundamentals (e.g., CV/AI/HI technology, understanding of generalizable design knowledge) to establish a shared problem definition. On this basis, all participants had to prepare initial proposals on how design knowledge could be derived and presented systematically by using their individual experiences from research and practice.

In the **second** session, the individual proposals were discussed, which finally resulted in the decision to pursue a reflective methodological strategy by starting from a selection of practical cases and then incrementally accumulating and generalizing design knowledge towards the development of a nascent design theory. Thus, a total of six CV development cases were chosen as a basis for knowledge generation, which subsequently had to be reflected and characterized by the individual researchers.

In the **third** session, the results were presented to the group to derive a collection of (i) case requirements, (ii) implemented features, and (iii) several more case characteristics (e.g., domain, criticality, expertise of users) to classify each specific case (cf. Section 7.4). By comparing similarities and differences across all projects, various case level concepts could be extracted and tentatively connected to gain a first level of abstraction towards generalizable design knowledge. This included known concepts like supervised learning, active learning, degree of automation, accuracy, speed, image complexity, labeling cost, black-box behavior, explanation, uncertainty, compliance, and others. On this basis, all researchers were asked to conduct an individual step of theorizing. To this end, each participant should prepare conceptualizations and schemata to link theoretical constructs and mechanisms with observations collected within the cases.

The results were jointly discussed in the **fourth** session. Several potential kernel theories were identified, such as technology acceptance theory (Venkatesh et al., 2012), principal-agent and signaling theory (Eisenhardt, 1989), and algorithm aversion (Dietvorst et al., 2015). In combination, they resulted in several meta-level concepts like performance, effort, system restrictiveness, information asymmetry, perceived control, trust, and others. However, the participants struggled to connect observations from the application environment with justificatory knowledge from theory. This led to a divergent picture within the individual conceptualizations, which were created with different emphases on different levels of abstraction. For example, some participants focused on HI interaction patterns between the human and the computer from different theoretical lenses, while others connected specific case properties, features, and requirements to theoretical constructs. As a result, three researchers agreed to translate the different views into a unified conceptualization to serve as a foundation for further discussion.

In the **fifth** session, the unified conceptualization was discussed and incrementally refined by the entire group. With this unified framing, three distinct design-related components could be distinguished from each other towards an adequate level of abstraction of derived design knowledge, precisely (i) *meta-requirements*, (ii) *designed mechanisms* in relation to the HI interaction, and (iii) *case characteristics* in terms of moderating effects related to the nature of the vision-based task (e.g., task criticality) (cf. Section 7.5). All researchers were asked to iterate another cycle of individual reflection and formulate testable propositions to reflect on how these three components are related to each other as observed within their cases. In this respect, the identified mechanisms had to be translated into corresponding design principles by abstracting from particular features implemented in the CV projects (Gregor et al., 2020).

In the **sixth** session, the design principles and propositions were evaluated and harmonized. As a result, an overview of the generated multi-level design knowledge was derived spanning from high-level mechanisms and meta-requirements to case-specific features and concrete requirements.
In the **seventh** session, the results were revalidated and minor adjustments on all partial results (i.e., meta-requirements, mechanisms, design principles, design features, moderators, and testable propositions) were implemented. The required steps for future work were discussed, precisely the setup of corresponding evaluation studies towards the development of a nascent design theory. At this point, the series of workshops was paused to prepare the results for presentation and obtain external feedback from a broader community.

7.4 Overview of Regarded Cases

For the selection of suitable CV development projects, the principles of Yin (2017) were followed to select cases that shared common properties (e.g., implementation of AI technology) while differing from each other to obtain the required variance (e.g., vision-based task, application domain). This non-probability sampling technique is similar to comparison focused sampling in which cases are selected to compare, contrast, and learn about characteristics that explain their similarities and differences Saunders et al. (2009). On this basis, a total of six different cases were selected, which we describe in the following. The reported characteristics were discussed (cf. Figure 7.1 on page 149) until the experts agreed that the characteristics both represent the individual cases appropriately and can be compared across cases (Table 7.2 on page 153). All cases were conducted in cooperation with industry partners. Some projects are completed, while others are still under continuous development.
Car configuration (CAR). Car manufacturers offer their customers car configurators to assess different combinations of characteristics like color, rims, and headlights. As the 3D car rendering process is highly complex, the rendering software can output virtual car models with black holes instead of the chosen part. Currently, these faulty virtual car models are identified manually. This is highly inefficient as there are billions of possible combinations. A CV-HIS was built to detect faulty virtual car models. The CV-HIS relies on active learning to reduce the labeling effort.
Energy infrastructure (NRG) (compare Chapter 6). At present, power line maintenance relies mainly on human inspection via manual ground visitation, helicopter-based patrolling, and tower climbing. This is costly, time consuming, and often

hazardous. A CV-HIS was developed that detects faults like bird nests or open safety pins on image data acquired by unmanned aerial vehicles. The detected faults are presented to human operators as a basis for decisions like prioritization and route optimization.

Solar panels (SOL). Manufacturers of solar panels must meet high quality standards when offering their products on the market. It is therefore important that quality impairments are detected at an early stage in the manufacturing process to treat them accordingly and avoid unnecessary costs. Thus, a CV-HIS was developed in which defects had to be detected automatically based on electroluminescence images. The challenge here was to separate defective solar cells from flawless ones and distinguish between specific types of defects while ensuring low inference times as determined by the rigid setting of the production environment.

Viticulture (VIT). Apart from planting crops, harvesting grapes, and producing vine, viticulture is faced with many tasks and obstacles, such as control of perfect planting positions, disease detection, and personnel allocation, especially in the harvesting process. To support these tasks with a low-cost structure, a CV-HIS was deployed that was integrated into the daily processes with minimal additional effort by mounting cameras on farm tractors to capture image data of vines and grapes to be subsequently used for disease detection and yield prognosis.

Cutting tools (CUT) (compare Chapter 5). In machining processes, unwanted material is removed from a workpiece by a cutting tool. Different types of wear occur on the tools due to friction and heat, so over time the tools are rendered unusable. A frequent task in the machining industry is the visual inspection of cutting tools. It serves as a decision basis for developing new generations of tools as well as for optimizing the parameters of machining processes. For the decisions, it is important to know precisely where which type of wear occurred. Here, a CV-HIS was developed that performs this visual analysis.

Architectural floor plans (ARC). The architecture, engineering, and construction industry often relies on floor plans only available as rasterized images or analog documents. For tasks such as pricing services and building operations, information such as symbols, room size, or space use must be extracted. A CV-HIS was developed to automate the digitization and analysis of floor plans. Domain experts were included to manage unknown symbols and uncertain predictions.

Tab. 7.2.: Overview of cases.

	CAR	NRG	SOL	VIT	CUT	ARC
Domain	Automotive	Energy	Manufacturing	Agriculture	Manufacturing	Architecture
Objective	Quality assurance	Fault detection	Quality assurance	Yield prediction and improvement	Understanding behavior in use of cutting tools	Digitizing and analyzing floor plans
Vision-based task	Detection	Detection	Classification, detection, segmentation	Detection, counting, object tracking	Segmentation	Detection, counting
CV experience of user(s)	Low	Medium	Medium	Low	Medium	Low
Task complexity	Medium	High	Low - medium	Low	Medium	Medium
Task criticality	Low	High	Medium	Low	Medium	Medium
Exemplary requirement	Reduction of highly repetitive manual work	Visualization of automatically extracted information	High accuracy and very low inference time	Model decisions modifiable by humans	Access to images and CV outputs as input for human decision-making	Flexibility regarding objects to be detected
Exemplary design feature	End-to-end automation	Interface for visual data exploration	Repository of models for different defect types	Object tracking for counting in image sequences	Human-centric decision support	Monte Carlo dropout as basis for uncertainty measures
References	Hemmer, Kühl, et al. (2021)	Landwehr et al. (2022)	Zschech, Sager, et al. (2021)	Heinrich, Zschech, et al. (2019)	Treiss et al. (2020) Walk et al. (2020)	Hemmer, Vössing, et al. (2021)

7.5 Conceptualization of CV-Based Hybrid Intelligence Systems

To aid the design process for CV-HIS, we integrated HI theory with findings from the cases to depict a conceptualization that forms the basis of our design knowledge. The CV-HIS is conceptualized as an interaction of the *human* and the *computer* to solve a vision-based task in alignment with granular case requirements. The conceptual model is depicted in Figure 7.2 on page 155 distinguishing between three components.

First, case requirements could be observed and linked to generic theoretical constructs to obtain *meta-requirements* determining the need, and likewise, the usefulness of any designed system functionalities.

Second, it was possible to observe several mechanisms within the CV-HIS interaction patterns between humans and computers as introduced by specifically implemented design choices, which could also be justified by mechanisms of related kernel theories. We can relate these mechanisms to three aspects that act as key resources in AI-based systems: *data*, *model*, and *decision* (Thiebes et al., 2020). The computer provides an *automation mechanism* that uses labeled image data to learn a model that generates a decision in the form of recognized objects (e.g., through the presentation of bounding box coordinates). Since the computer employs black-box models like deep neural networks, there exists a natural information asymmetry between the human and the computer. The human cannot fully comprehend the inner decision logic (i.e., the model) of the computer, and thus the computer is required to provide transparency and reduce information asymmetry by using a *signaling mechanism* (e.g., visualizing the relation between data and decisions). Additionally, the human wants to maintain some control within the decision process through a *modification mechanism*, which, for example, allows to modify data labels or manually change the computer's decision. Following the idea of a beneficial symbiosis in HI systems, the proposed mechanisms need to be coordinated by a *collaboration mechanism* that provides an interaction design via push and pull principles so that both sides can request resources (e.g., the computer can employ an active learning approach and signal the need for additional labels as a data modification from the human).

Third, several *case characteristics*, particularly related to the nature of the vision-based task, could be identified as potential moderating factors that are likely to have an effect on the relevance and need for any designed system functionality. This includes characteristics like the task's criticality and complexity or the CV experience of a user, as exemplarily summarized in Table 7.2 on page 153.

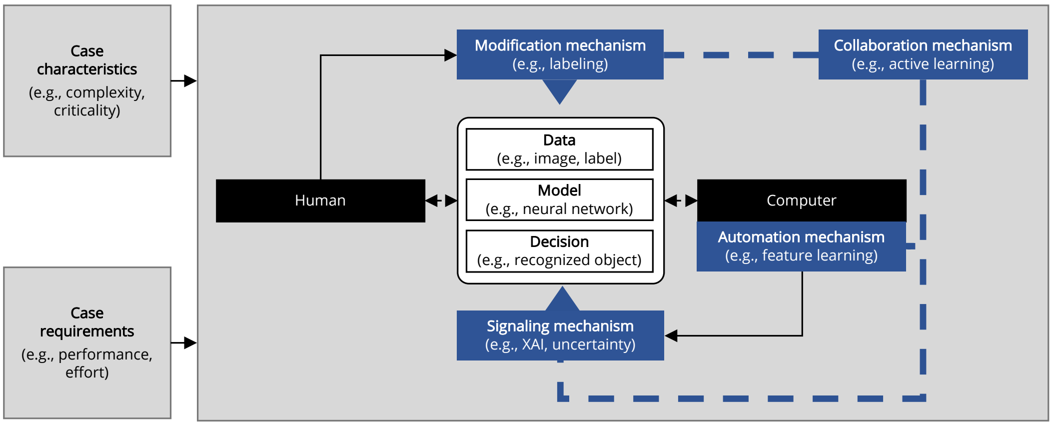


Fig. 7.2.: Conceptualization of CV-based hybrid intelligence systems.

7.6 Design Knowledge of CV-Based Hybrid Intelligence Systems

In this section, we present our results on the accumulated design knowledge of CV-HISs. Due to space limitations, we focus on the presentation of (i) meta-requirements and their connection to case requirements and related kernel theories, as well as (ii) mechanisms introduced by design in conjunction with design principles and exemplary features as implemented in the cases. A summary of the derived elements and their relationships is depicted in Figure 7.3 on page 156. In the following, we provide further details by organizing our findings in relation to the identified mechanisms. Moreover, we present some exemplary propositions and the role of observed moderators at the end of this section.

7.6.1 Automation Mechanism

When examining the extracted requirements from the individual cases, it became apparent that the main requirement of AI-based CV systems is to reduce the manual effort to perform the vision-based tasks. Depending on the baseline situation, *effort* can take on different dimensions. In the VIT case, for example, there was no system support before so that previously vineyard objects had to be inspected and counted manually. In other cases, effort refers to the operations and configurations during system usage to perform the vision-based task. In the SOL case, for example, electro-luminescence images were previously inspected by domain experts via techniques that required costly image engineering efforts that had to be reduced.

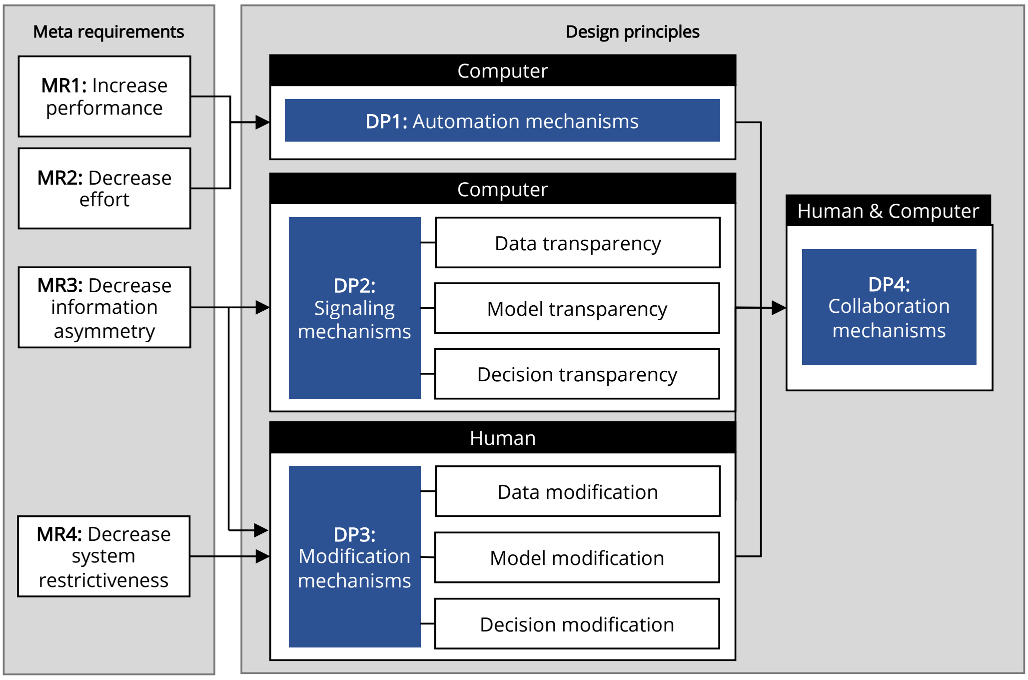


Fig. 7.3.: Meta-requirements and design principles for CV-based hybrid intelligence systems.

Another important requirement across all cases concerned the *performance* of the system. As such, it was necessary to automatically process and analyze visual objects while achieving a certain degree of quality, which was often required to be close to or even better than human performance. This requirement was expressed by several dimensions, such as model accuracy, inference time, and required resources. In the ARC case, for example, almost 100% detection accuracy was required for correct order generation and pricing. In the SOL case, a defect detection accuracy of 98% was demanded while simultaneously guaranteeing inference times of less than two seconds. In the CUT case, on the other hand, the users did not press for perfect predictions as long as they were good enough and cheaper than human experts.

Effort expectancy and performance expectancy are also pivotal factors within the unified theory of acceptance and use of technology (UTAUT). Among several other determinants, such as social influence and facilitating conditions, both factors play a significant role whether users adopt and use a new system (Venkatesh et al., 2012). Since this effect can also be assumed in the given context, we can derive the following two meta-requirements (MRs) in-line with theory and observations from practice:

> **MR1:** *The vision-based task should be supported by system functionalities that improve the overall task performance (e.g., detection accuracy).*

> **MR2:** *The vision-based task should be supported by system functionalities that decrease the human effort required to perform the task (e.g., manual inspections).*

The mechanism, introduced by design, which essentially addresses these two requirements and thus constitutes the main strength of the computer within the HI interaction is the *automation mechanism*. This is about automatically extracting and processing useful information from image data that can be exploited for the respective vision-based task. As concrete system features, implementing this mechanism, different types of DL models could be identified across the six cases that are basically all based on convolutional neural network (CNN) architectures. The nested design of CNNs allows them to be fed with high-dimensional raw data and then automatically discover internal representations at different levels of abstraction that are needed for visual recognition tasks (LeCun et al., 2015). On this basis, CV tasks can be executed with high quality results while outperforming conventional types of CV systems, such as statistical approaches or shallow ML. Moreover, in contrast to conventional systems, there is no need for extensive data preparation, especially with regard to manual feature engineering, thus minimizing undesired human effort. The only central prerequisite for DL models is the availability of sufficiently large training data with labeled instances so that they can automatically recognize relevant structures. In summary, we therefore formulate the following design principle (DP) addressing the automation mechanism.

> **DP1:** *Provide the system with the functionality to automatically extract visual features from image data and build a model that supports the vision-based task to minimize undesirable manual interventions.*

This design principle constitutes a core principle of today's CV systems, as it can also be observed in broader practice. It has a remarkable influence on all other design-related components, such as (meta) requirements, design features and their abstraction towards design principles. This includes, for example, that DL models generally show black-box characteristics limiting their interpretability, or that they are prone to biases induced into training data by undesired effects, which demand for further mechanisms to address such issues (Janiesch et al., 2021).

7.6.2 Signaling Mechanisms

Besides the main requirements of automation, the individual case requirements revealed additional needs concerning the reduction of information asymmetry between the human and the computer within the CV-HIS. In the VIT case, for example, a comparison between the detection of vines and grapes by the AI and the actual input data and bounding boxes was required to rely on the yield prognosis. Similarly, system users in the SOL case asked for explanations on which basis automatically detected errors were classified in one class or another. Other examples could be found in the SOL, NRG, and CUT cases, where users asked for input data visualizations to inspect training samples and labels to reduce uncertainty with regard to the actual labeling process.

Adopting principal agency theory, we can state that there is an information asymmetry between the human and the computer in CV-HISs that needs to be resolved or reduced (Vladeck, 2014). On the one hand, the human has the meta-knowledge of what constitutes a real-world object that should be detected from image data by the computer, whereas the computer itself does not have such knowledge due to the lack of superintelligence (Jebari & Lundborg, 2020). On the other hand, the trained system outputs predictions that are not comprehensible for the human due to the black-box-nature of the automation mechanism (Wanner et al., 2020). Failure to reduce this information asymmetry can result in decreased system adoption (Castelo et al., 2019; Miller, 2019; Oh et al., 2017). Thus, connecting the theory with the case requirements, we can derive the following meta-requirement:

> **MR3:** *The vision-based task should be supported by system functionalities that reduce information asymmetry between the human and the computer.*

Thus, to comply with the requirement, signaling mechanisms can be introduced by design that reduce uncertainty by creating explanations for the different aspects of the CV-HIS (i.e., data, model, and decision). The first set of explanations are aimed at increasing *data transparency*. In CV problems, the user is usually faced with detecting objects and classifying them into several available classes. The user needs to be able to determine if the input data along with its labeling is in alignment with his or her understanding of the problem domain to provide a correct input to the computer to be processed (e.g., only providing image counts as labels for criminal surveillance where specific persons are sought after will not suffice (Barbosa & Chen, 2019). Specific features related to this aspect are visualizations of input data together with suggested labels via labeling tools or providing metadata on input images. The second set of explanations covers *model transparency*. This may include to provide global explanations and meta-information regarding the trained model (e.g., configurations, hyperparameters) that are useful for a user to better comprehend the system's behavior. Lastly, local explanations are required to create *decision transparency*. The human should be able to compare the provided explanation with his or her decision logic to assess the quality of the system's decision. Methods from the field of explainable AI like LRP (layer-wise relevance propagation) or Grad-CAM (gradient-weighted class activation mapping) can be implemented for this purpose to generate pixel heat maps highlighting important parts of the input image that are responsible for the prediction. Enhancing the computer with the ability to express its confidence in single predictions is another option to reduce decision transparency by implementing features such as uncertainty measures based on, e.g., Monte Carlo dropout (Gal & Ghahramani, 2016). In summary, we can thus formulate the following design principle:

> **DP2:** *Provide the system with the functionality to generate and signal explanations about the computer's behavior in terms of data, model, and/or decision to increase transparency for the human.*

Figure 7.4 shows exemplary images, labels, predictions, and uncertainty maps from the CUT case. In the uncertainty maps, a higher degree of uncertainty is indicated by brighter pixels. The left side depicts a situation with high prediction quality. In the uncertainty map, only the pixels at the borders between classes are bright/uncertain - classifying these pixels is also difficult for humans (Kendall et al., 2015). The images on the right, on the other hand, depict a situation with low prediction quality where whole areas are bright. Consequently, the system indicates that it is uncertain about the situation, and thus, reaches its limits for correctly inspecting the cutting tool.

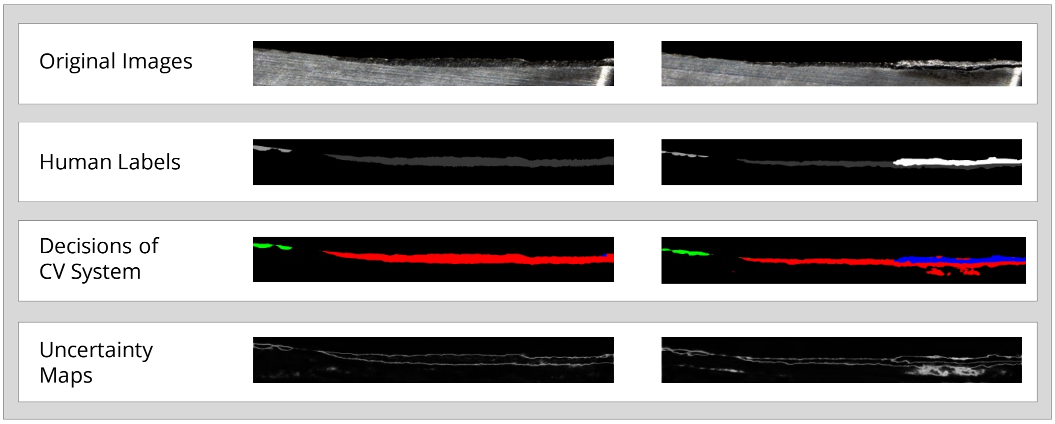


Fig. 7.4.: Uncertainty-based signaling in the CUT case (Treiss et al., 2020) (compare Chapter 4).

7.6.3 Modification Mechanisms

On top of the previously discussed mechanisms, we could observe that the human using the developed CV-HIS did not only ask for more transparency to better understand the systems' behavior but that they also wanted to be in control of the situation they were facing. For example, in the CUT and ARC cases, professionals asked for the explicit possibility to intervene and make decisions based on their own experience when uncertain outcomes are indicated by the system. Furthermore, in the SOL and VIT cases, we could observe the requirement to have the option to directly adjust faulty labels when inspecting input data in order to achieve better model quality. Another example is given in the SOL case where the users asked to have control over the selection and configuration of detection models to choose a suitable approach depending on the situation (e.g., available resources). With regard to informing theories, our observations can be related to two theoretical lenses; that is algorithm aversion and decision support system (DSS) theory. Algorithm aversion describes that individuals are hesitant to rely on predictions computed by algorithms and rather prefer human forecasters because humans tend to believe they have superior reasoning in comparison to algorithms. However, this effect is not always present. Instead, algorithm aversion can be reduced by giving humans (even a slight degree) of control over the algorithm to modify its prediction (Dietvorst et al., 2018). Similarly, DSS theory states that a system should not be too restrictive when preparing decisions in order to not negatively limit the users' decision strategy and thus allow sufficient control over the situation (Meth et al., 2015). To sum up, we can therefore derive the following meta-requirement:

> **MR4:** *The vision-based task should be supported by system functionalities with minimal restrictiveness facilitating a sufficient degree of flexibility for the human user.*

To address this requirement, it is therefore necessary to give the user sufficient flexibility and control over the system by providing appropriate *modification mechanisms*. In analogy to the manifestation of the signaling mechanisms, modification possibilities can also be designed at multiple touchpoints between the human and the computer (i.e., data, model, and decision components). Considering the *modification of data*, e.g., the designers of the SOL case developed an integrated labeling tool that was closely connected to the actual detection system. In this way, domain experts could quickly enter the image repository at any time to refine bounding boxes or modify class affiliations whenever it was necessary according to their expert knowledge. Referring to another example introduced above, it was a crucial design element within the CUT and ARC cases to integrate human knowledge in uncertain situations, which is related to the *modification of the decision*. As such, we can formulate the following design principle:

> **DP3:** *Provide the system with the functionality to modify the computer's behavior in terms of data, model, and/or decision to allow the human to contribute knowledge and to control the vision-based task.*

7.6.4 Collaboration Mechanism

As outlined by Dellermann, Ebel, et al. (2019), collaboration between human users and the computer is an important characteristic of HI. It was found from the CUT and ARC cases that a collaboration mechanism can be crucial for the overall task performance. The collaboration mechanism raises awareness about a collaborator's activity and subsequently enables input requests with regard to the different mechanisms and aspects (e.g., the human could request an explanation for a specific decision from the computer). Hence, we formulate the following design principle:

> **DP4:** *Provide the system with the functionality to facilitate a collaboration of human and computer-based mechanisms.*

The collaboration mechanism enables the interplay of the CV-HIS mechanisms by providing collaboration functionalities such as push and pull requests. For example, when individual decisions are uncertain (DP2 - decision transparency), the CV system can request the user to modify the provided decision where appropriate (DP3 - decision modification). This is an important aspect of the CV-HIS developed in the ARC case. The user can also support the learning process of the computer by utilizing his or her knowledge to label unknown data (DP3 - data modification) where the system expects the highest level of improvement (DP2 - model transparency). This pattern, frequently referred to as *active learning*, is a well-known example of dialogue-based collaboration mechanism and is a crucial design element within the AUT case. Another example of such a combination is given in the SOL case, where users can also choose between different levels of model complexity (DP3 - model modification) based on the task characteristics and the computer's signaling outcome (DP2 - model transparency).

7.6.5 Outlook: Exemplary Propositions and Moderating Effects

With the accumulated design knowledge outlined in the previous sections, it is possible to derive a set of testable propositions (TePs) constituting the relationships between meta-requirements and design principles. Since we cannot fully discuss all derived propositions due to space limitations, we only provide two examples by TeP1 and TeP2 to pave the further way towards a nascent design theory.

Example proposition 1: *Using a CV-HIS with automation mechanisms (DP1) will result in a higher degree of perceived performance than using a system without automation mechanisms.*

Example proposition 2: *Using a CV-HIS with automation mechanisms (DP1) and modification mechanisms (DP3) will result in a higher degree of perceived control than using a CV-HIS with automation mechanisms, but without modification mechanisms.*

Both propositions reflect generalized design knowledge as observed within the cases and place it in relation to a respective reference system. Thus, TeP1 describes the effect of employing a system with the core design principle DP1 in contrast to another type of CV system (e.g., conventional system based on statistical approaches), whereas TeP2 describes the effect of two different design configurations with regard to DP1 and DP3. Thus, with the different meta-requirements and design principles, a system of propositions can be obtained that describes the effects of the mechanisms introduced by design in their entirety. However, since not all mechanisms could be observed equally across all cases, the propositions are only tentative assumptions that need to be examined in larger evaluation studies in more controllable settings and with more users involved.

Furthermore, we assume that the contextual case characteristics, such as the criticality of the vision-based task or the CV experience of users, might have a significant influence on the intensity of the need for any design principle. Thus, we assume, for example, that there is a higher need for explanation (DP2) and modification (DP3) of generated decisions in cases where domain professionals are responsible for critical situations that can lead to high costs or even lives at risk. Similarly, we expect that CV cases with rather technically oriented users (e.g., SOL case) will presumably require a higher degree of control over models and configurations (DP3) so that they can bring their technical expertise into the processes than it will be the case, for example, in agricultural domains (e.g., VIT case). While some of these characteristics and their influences were already captured in this research project, such moderating effects need to be further examined in future studies.

7.7 Conclusion

In this paper, we provided insights from a research project with the goal to accumulate prescriptive design knowledge for computer-vision-based hybrid intelligence systems. To this end, we pursued a reflective DSR approach, introduced as "Strategy II" by (Iivari, 2015). We conducted a series of workshops with IS researchers involved in several industrial computer vision projects. As a result, we were able to derive generalizable design knowledge illustrated through meta-requirements, mechanisms, design principles as well as testable propositions. Even though our focus was on the application area of computer vision, we are confident that the results show a more generalizable character and can therefore be transferred to broader contexts in which hybrid intelligence systems need to be designed (e.g., in the realm of natural language processing).

However, the generalizability of these results is subject to certain limitations. For instance, we only regarded a total of six cases and future work should include additional examples to further facilitate the theorizing process. Furthermore, due to space restrictions, we could only elaborate on some excerpts of the current findings. For instance, we had to exclude design principles with a more technical focus (like scalability and robustness), and we could not discuss the potential moderating effects and testable propositions in sufficient detail, which will therefore be part of subsequent work.

Our work can help practitioners to design CV-HIS in a more human-centric manner by incorporating socio-technical considerations. From an academic point of view, this research contributes to the knowledge base by proposing generalizable design knowledge and laying the foundation for many future research directions in need of further investigation by considering factors beyond technical performance like restrictiveness and information asymmetry between human and computer. Future work needs to focus on the testable propositions and their translation into evaluation studies and experiments. A promising field of research lies ahead.

Part IV

Applications for Sustainability

8 Facilitating Sustainable Smart Product-Service Systems with Computer Vision[1]

8.1 Introduction

In addition to the agricultural and energy sectors, the manufacturing industry is one of the largest emitters of CO_2 (Edenhofer et al., 2014) and its demand is continuously rising (Hatfield-Dodds et al., 2017). Therefore, there is an urgent need to reduce the carbon footprint of the manufacturing industry (Park et al., 2009). An option to address this challenge is to substantially reduce waste generation through prevention, reduction, recycling, and reuse, as formulated by the United Nations (United Nations, 2015, p. 27). However, technical innovations enabling and supporting these activities are an imperative (Mohmmed et al., 2019).

In recent years, there have been major breakthroughs in the field of artificial intelligence (AI) systems based on deep learning (DL), which have enabled them to surpass human performance in specific tasks (He et al., 2015; D. Silver et al., 2017). As described by Vinuesa et al. (2020), AI can positively affect sustainable development and multiple frameworks help to structure these endeavours (S. Ren et al., 2019; Zhang et al., 2017). However, thus far, only a few real-world implementations of AI address sustainable development goals.

To address this research gap, this work applies DL-based computer vision (CV) to facilitate initiatives that address the aforementioned challenges. Specifically, we utilize CV to efficiently determine the wear states of products. As machine learning (ML) and DL has been shown to yield benefits in the related field of smart manufacturing (Flath & Stein, 2018; Miguéis, Borges, et al., 2022; J. Wang et al., 2018), it promises to be effective for the proposed approach.

[1]This chapter comprises an article that is currently under review as: Walk, J., Kühl, N., Saidani, M., Schatte, J (2022). Artificial Intelligence for Sustainability: How Computer Vision Can Facilitate Sustainable Smart Product-Service Systems — Evidence From Life Cycle Assessments Based on Two Case Studies. *Working Paper.* Note: The abstract has been removed. Minor edits have been made and tables and figures were reformatted, and newly referenced to fit the structure of the thesis. Chapter, section and research question numbering and respective cross-references were modified. Formatting and reference style was adapted and references were integrated into the overall references section of this thesis.

The products investigated in our case studies are machining tools used for machining processes and rotating anodes used in diagnostic imaging applications. These products show different types of wear over their life cycles, which can only be observed accurately with a microscope. We employ a DL-based CV model, which is a specialized AI technique, to determine the wear state of these products from microscopic images. For the associated use cases of our products, complete automation is neither possible nor desirable. Instead, we rely on combining the strengths of human and artificial intelligence — an approach called hybrid intelligence (Dellermann, Ebel, et al., 2019). The DL-based CV detects product wear from images in a reproducible and efficient manner. This task is tedious and difficult for humans to perform. However, final decisions, such as the adaptation of machining process parameters, are at the discretion of human experts. They can incorporate a plethora of additional information, such as operating conditions and years of domain expertise. At present, training an AI to incorporate this additional information is infeasible because a large amount of data is required to reflect all the nuances of real-world situations.

The results of the wear analysis performed through CV facilitate different types of product-service systems (PSSs) that support environmental sustainability in four ways. First, it can facilitate typical reversible strategies (4R): re-design, remanufacturing, reuse, and recycling (X. Li et al., 2021). Understanding the current wear state of a product is crucial for deciding which one of the 4Rs is most suitable from an economic and environmental sustainability perspective. Second, the product usage stages can be improved in terms of environmental sustainability. Data on the usage of products are scarce and generally of low quality because the products are typically owned by customers at this stage (Zhang et al., 2017). The proposed approach provides information about this important phase, and thus, it enables more data-driven product lifecycle management. Third, assessing the wear state of products using CV can improve the usage of the same product in future iterations. For instance, P. Wang et al. (2018) show for steel that altering the usage stage is one of the biggest levers of manufacturing companies regarding sustainability. Finally, result-oriented PSSs can yield sustainability benefits (Tukker, 2015). In result-oriented PSSs, the client and provider agree on an outcome, but the provider can choose how the outcome will be achieved — in particular, no predefined product is involved (Tukker, 2004). However, a detailed understanding of product usage is necessary to offer result-oriented PSSs because it is crucial to assess the risks and costs to make a competitive yet profitable offering (Tukker, 2004). An accurate assessment of the wear state of products using CV can facilitate a detailed understanding of product usage.

Conceptually, the proposed approach depicts a novel type of sustainable smart product-service system as proposed by X. Li et al. (2021). In contrast to previous studies by Zhang et al. (2017) and X. Li et al. (2021), the products under consideration do not have to be smart for the proposed approach; that is, no sensors are integrated in or placed on the product.

This work contributes to the existing literature by demonstrating the effectiveness of DL-based CV for extracting valuable information from non-smart products to improve environmental sustainability. We validate this approach using two products by demonstrating its technical feasibility through an experiment and environmental sustainability through a life cycle assessment (LCA). The technical feasibility for detecting wear on the two products is demonstrated, and the environmental sustainability benefit of the proposed approach is verified through LCAs. Furthermore, the requirements of this approach are conceptualized to provide researchers and practitioners with guidance regarding its applicability to other PSSs.

The remainder of this work is organized as follows. In Section 8.2, we introduce relevant foundations and the methodology employed. Subsequently, in Section 8.3, we present the evaluation and the results of the CV models and the LCAs. In Section 8.4, we discuss our work in a broader context and conceptualize it. Finally, in Section 8.5, we summarize our work and discuss its limitations and possible future research.

8.2 Materials and Methods

This chapter introduces the relevant foundations in the fields of PSSs, CV, and DL. We then present the methodology and the selected case studies.

8.2.1 Foundations

In the following section, we first discuss sustainable PSSs and then introduce the fundamentals of CV and DL.

Sustainable Smart Product-Service Systems

PSSs were first defined in 1999 by Goedkoop et al. (1999) as “a marketable set of products and services capable of jointly fulfilling a user’s need.” In the early stages of PSSs, sustainability was already an important concept (A. Q. Li et al., 2020). In 2016, based on their literature review, Annarelli et al. (2016) define PSSs as "a business model focused toward the provision of a marketable set of products and services, designed to be economically, socially, and environmentally sustainable, with the final aim of fulfilling a customer's needs." More recently, various authors extended the concept of sustainable PSSs to be smart based on technological innovations, such as AI or Internet of things-based connectivity (de Jesus Pacheco et al., 2019; X. Li et al., 2021; Sakao & Neramballi, 2020). Alcayaga et al. (2019) coined the term smart-circular systems, which they conceptualized as a combination of circular strategies, smart products, and PSSs. Based on Ellen MacArthur Foundation (2016), smart products are considered to possess the ability to sense, store, and communicate information about their environments and themselves. X. Li et al. (2021) propose a data-driven reversible framework for achieving sustainable smart PSSs. They illustrated this framework by sustainably developing a 3D printer. This work contributes to this area of research by providing empirical proof of sustainable smart PSSs with the support of AI in the form of DL-based CV models.

Fundamentals of Computer Vision and Deep Learning

CV aims to equip computers with the ability to visually perceive the world similar to humans (Szeliski, 2010). For decades, CV systems relied on techniques such as edge detection and filters (Szeliski, 2010), which are now referred to as "traditional" CV techniques (O'Mahony et al., 2019). Recently, it was shown that CV systems based on ML have the potential to produce more accurate outputs. In isolated cases, ML-based CV systems even surpassed human performance (He et al., 2015).

ML is a relatively old field of research, defined in 1959 by Arthur Samuel as giving computers the ability to learn without being explicitly programmed (Samuel, 1959). Current CV systems are based on DL, a subfield of ML that relies on deep neural networks (Janiesch et al., 2021). In particular, convolutional neural networks (CNNs) are typically applied to CV tasks, as they perform well on visual data (Kim, 2017).

In this paragraph, we briefly describe how CNNs work according to LeCun et al. (2015). Like other types of neural networks, CNNs consist of multiple processing layers that learn to represent data with different degrees of abstraction. However, as opposed to fully connected neural networks the same operations are applied to all inputs from the previous layer. Hence, the number of weights is reduced considerably in comparison to fully connected networks. Consequently, a CNN can be successfully trained with significantly less data and computing power than a classic fully connected neural network for the same task (LeCun, Bengio, et al., 1995).

The following CV tasks are most typical for static images (compare Figure 8.1): *image classification*, *object detection* , and *semantic segmentation* (Griebel, Dürr, et al., 2019). In image classification, the CNN output is a class label for the entire image. In object detection, a bounding box containing the object of interest is output along with the class label. The information produced by a CNN for semantic segmentation is even more fine-granular — each pixel in an input image is assigned a class label.

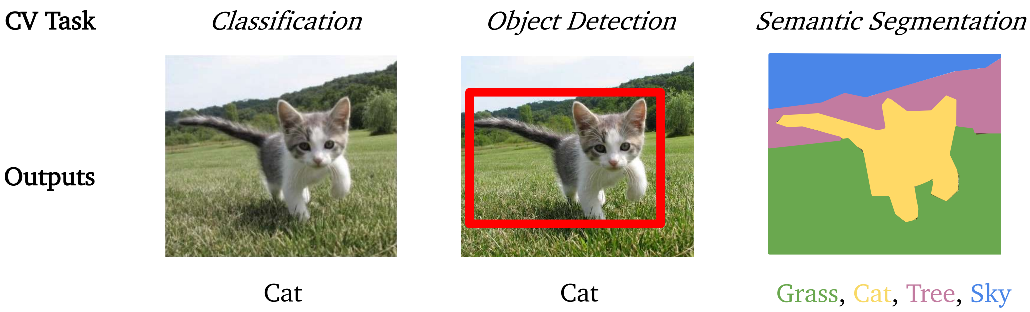


Fig. 8.1.: Outputs of typical CV tasks for an image of a cat. Own representation based on F.-F. Li et al. (2017) and Kosson and Marklund (2018).

8.2.2 Methodology

With the foundations of PSSs and DL-based CV at hand, we now introduce our methodology. Figure 8.2 on page 172 presents a high-level overview of the methodology. First, we describe various steps to assess the feasibility of extracting relevant information from microscopic images using DL-based CV. The last two steps describe the LCAs at a high level. As shown at the bottom of the figure, some steps are performed manually while others are performed by the DL-based CV model. The following subsection details the CV and LCA methodology.

Computer Vision Methodology

To allow for an accurate assessment of the wear state of products in 2D images, we train a DL-based CV model to perform semantic segmentation. Semantic segmentation provides pixel-wise information; thus, it enables a detailed assessment of the different wear types. According to the domain experts in the case companies, this level of detail is most helpful for the case studies considered. We use a CNN based on the U-Net architecture (Ronneberger et al., 2015). This architecture stems from

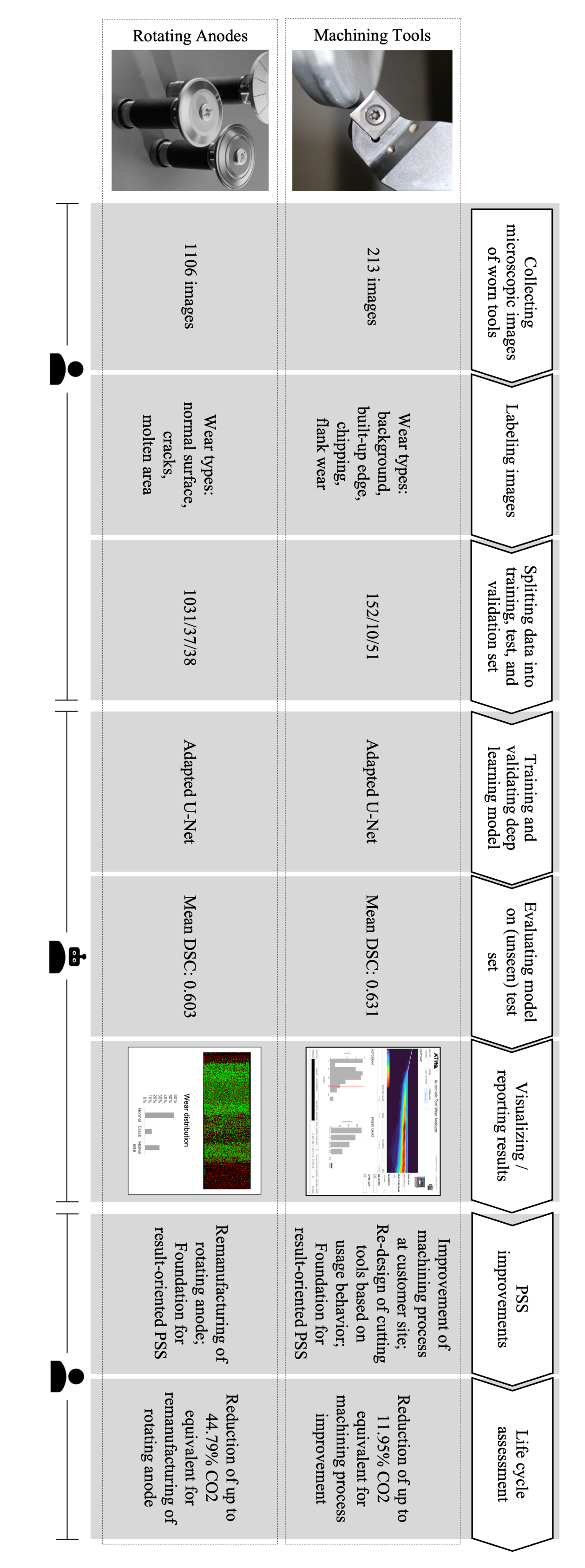


Fig. 8.2.: Overview of the methodology of this work.

the biomedical domain and has become a standard for different types of semantic segmentation tasks. Semantic segmentation models are typically trained in a supervised manner — the neural network learns to solve a task using the optimal results as a (to be predicted) target during training. For semantic segmentation, the target is specified in the form of a human-created label that defines the wear class for each pixel in the corresponding image. To train, tune, and evaluate the DL-based CV, we split the datasets into training, validation, and test sets (Hastie et al., 2017, p. 222). The training set is used to fit the models, the validation set is used to select a model configuration from different models with varying hyperparameters, and the test set is used to estimate the fraction of errors the DL-based CV will commit later in a — previously unseen — real-world setting. To assess the prediction quality, we compare the human-assigned labels and predictions and compute numerical evaluation measures. For the numerical assessment of the predictions, we rely on the pixel accuracy and mean dice similarity coefficient (mean DSC) (Dice, 1945), which is a common metric for evaluating the performance of semantic segmentation models (Setiawan, 2020). It is defined as follows:

$$Mean\ DSC = \frac{2}{C}\sum_{c=1}^{C}\frac{\sum_{i=1}^{N}\hat{y}_{i,c}\ g_{i,c}}{\sum_{i=1}^{N}\hat{y}_{i,c}+\sum_{i=1}^{N}g_{i,c}} \tag{8.1}$$

with the prediction $\hat{y}$ assigning a class label $c \in C$ to each pixel $i \in N$. C represents the total number of classes and N denotes the number of pixels in the input image. $g_{i,c}$ denotes the one-hot-encoded human labels used as the ground truth. We implement, train, and evaluate our model using the Python library *Keras* in version 2.1.6. (Chollet et al., 2015).

Life Cycle Assessment Methodology

Once we are aware of the CV results, we integrate them into a user-centric artifact that supports domain experts in decision-making and facilitates the different PSSs described previously. Additionally, we discuss the CV results with domain experts from the case companies. Based on this, we examine the impact on the environmental sustainability of different PSSs together with domain experts. To assess the environmental impacts of the selected PSSs, we perform LCAs using the methodology described in the following.

LCA is an internationally standardized method used for the quantitative environmental impact assessment of products, processes, services, and systems throughout their life cycles (Finkbeiner et al., 2006). LCA can be particularly useful for comparing alternative strategies and understanding the trade-offs between the benefits and

impacts of different systems, which can help in making informed decisions (Niero et al., 2014). LCA can be deployed as a quantitative decision-support tool in sustainable design engineering or green manufacturing (Saidani et al., 2021).
According to ISO standards 14040 (ISO - International Organization for Standardization, 2006a) and 14044 (ISO - International Organization for Standardization, 2006b), an LCA comprises four major steps:

1. Goal and scope definition: The goal phase defines the overall objectives of the study. The scoping phase sets the boundaries of the system studied, sources of data, and functional unit to which the results refer.

2. Life cycle inventory (LCI): A detailed compilation of all the inputs (e.g., material and energy) and outputs (e.g., pollutants) at each stage of the life cycle is performed.

3. Life cycle impact assessment (LCIA): It aims to quantify the relative importance of all environmental burdens obtained in the LCI by analyzing their influence on the selected environmental impact categories.

4. Interpretation of results: The outcomes of the LCI and LCIA stages are interpreted to find hotspots and compare alternative scenarios.

A key aspect to consider in the goal and scope definition is the functional unit (FU). It provides a reference to which the inputs and outputs of the LCA can be related (J. S. Cooper, 2003). According to ISO 14044 (ISO - International Organization for Standardization, 2006b), the FU should be clearly defined and measurable. This enables a scientifically sound (i.e., consistent and unbiased) comparison between different product systems and scenarios. Joint Research Centre (2010) recommends including the following aspects in the definition of the FU: (i) verb (functional analysis); (ii) what (form of the output); (iii) how much? (magnitude), how well? (performance), and how long? (duration).
For the LCIA, the OpenLCA software (2021 version 1.10.3), developed by GreenDelta (Ciroth et al., 2014), was used to model the PSSs and conduct comparative LCAs. Within OpenLCA, the Ecoinvent database (2021 version 3.7) (Wernet et al., 2016) and the ReCiPe 2016 Midpoint (H) method were used to perform the environmental evaluation. Ecoinvent is one of the most comprehensive and acknowledged databases providing the necessary data for impact calculations, including region-specific production and manufacturing data for numerous commodities across multiple industries (Frischknecht & Rebitzer, 2005). ReCiPe is a scientifically sound and acknowledged impact calculation method that provides characterization factors and normalization methods for calculating the impact (Huijbregts et al., 2017).

8.2.3 Case Studies

In this subsection, we introduce our case studies: the products, necessary domain knowledge, and respective industries. In each case study, we introduce DL-based CV as a tool to better assess wear and, subsequently, demonstrate how the results impact environmental sustainability.

Machining Tools

Machining is an important manufacturing process (Black, 1995, p. VI). It is utilized in numerous industries, such as healthcare (Churi et al., 2009), aerospace (Ezugwu, 2005), and automotive (Tai et al., 2014). Figure 8.3 on page 177 shows an exemplary turning process, a specific type of machining process. During the turning process, the workpiece rotates at a high speed; consequently, there is a relative motion between the workpiece (left) and the cutting tool (right). This results in the removal of unwanted material from the workpiece (Black, 1995, p. VI). Owing to the powerful forces and elevated temperature, the cutting tool is subject to wear and must be changed regularly (Bergs et al., 2020). Analyzing the wear of the cutting tool is essential for understanding the improvement potential of the machining process. Because the cutting tools are small and frequently changed, it is economically unviable to equip the tools themselves with sensors for connectivity, as is the case with several other small tools with low unit prices (Martin & Kühl, 2019). Consequently, the visual inspection of worn cutting tools is an important building block for understanding the wear and, therefore, the improvement potential.

Machining processes are influenced by several interdependent components such as cutting tools, tool holders, workpieces, workpiece holders, engines, and cutting fluids. The interplay of these components leads to an inherent variance in the machining processes. Particularly, the wear on two tools used in an identical process can vary significantly. Analyzing a single tool only provides a snapshot of the entire process. By contrast, analyzing several tools provides a more holistic overview of a given machining process. However, manual visual inspection of cutting tools requires considerable effort and, therefore, is currently not performed on a large scale. Visual inspection by CV facilitates efficient analysis of a large number of cutting tools and consequently allows for more reliable conclusions regarding the machining process. As described in Section 8.2.2, we utilize a DL-based CV to detect the wear on worn cutting tools.

The possible sustainable smart PSSs resulting from this can be grouped into *improvement of machining processes at customer sites*, *re-design of cutting tools based on behavior in use*, and *establishing the foundation for result-oriented PSSs*. These three approaches will be discussed in greater detail in the following section in terms of the improvement of environmental sustainability.

Tool producers typically have teams of specialized application engineers who support customers in improving their machining processes. The visual inspection of worn cutting tools is an important part of their job. Domain experts describe the current process for machining process improvements as follows: Typically, an application engineer visits the production site of a customer and inspects a small number of worn cutting tools (e.g., three) with a magnifier. Based on this, as well as additional information, such as machining parameters and domain expertise, a recommendation for process improvement is provided and implemented. Our approach enables the inspection of many worn cutting tools as a decision basis. Owing to its efficiency, it is possible to assess the wear state of a large number of worn cutting tools, for example, 200. This provides more reliable insights into the process improvement potential. Consequently, better results are expected from the process improvements. Additionally, the efficiency of the DL-based CV enables more process improvements. In Section 8.3.2, we present the results of an LCA for machining process improvements based on the proposed DL-based CV for wear assessment.

These insights can also be utilized in a more strategic way — understanding the usage behavior of machining tools can aid the design of the next generation of machining tools. Currently, the development of new generations of machining tools relies mainly on internal tests in controlled environments, which do not necessarily reflect real usage and the possible variances therein.

Ultimately, understanding the usage behavior of machining tools can support a change in the business model. Tool manufacturers typically sell their products to companies that use them in their production. Research suggests that PSSs, particularly result-oriented ones, can yield both environmental sustainability (Tukker, 2015) and economic benefits (Annarelli et al., 2016; M. Yang & Evans, 2019). A major hurdle for the adoption of PSSs is the difficulty of the offering party in estimating expected costs and risks (Erkoyuncu et al., 2011). Visual inspection of worn cutting tools using CV can facilitate a quick and accurate estimation of the risks and costs for a given machining process.

For this case study, we collected 213 worn cutting inserts from a real production process and captured microscopic images of the cutting edges. The wear mechanisms relevant to this case study are shown in Figure 8.4 on page 177. *Flank wear* results from friction between the cutting tool and workpiece (Altintas, 2012). This is the preferable wear mechanism because it occurs continuously. Also, it is

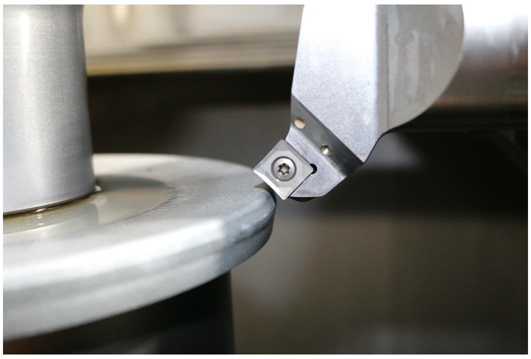
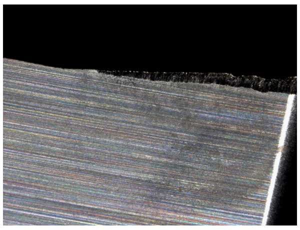
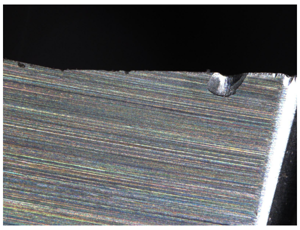


Fig. 8.3.: Exemplary turning process.

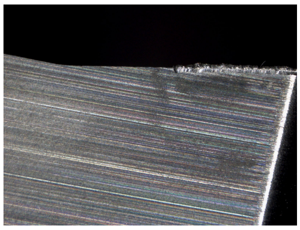
the most frequent wear mechanism (Siddhpura & Paurobally, 2013). In contrast, *chipping* and *built-up edge* are less desirable because they occur suddenly, leading to a significant deformation of the cutting edge. The cutting edge deformation can lead to the workpiece being considered scrap owing to its poor surface quality. *Chipping* describes the phenomenon of the cutting edge particles breaking off. A *built-up edge* occurs because of stress in the form of heat and pressure, resulting in the deposition of workpiece material on the cutting edge. In our case study, we trained a DL-based CV model to detect these wear mechanisms using microscopic images of worn cutting edges. A detailed description of the ML-related technical details of this case study can be found in Treiss et al. (2020).



(a) Flank wear



(b) Chipping



(c) Built-up edge

Fig. 8.4.: Common wear mechanisms in machining processes.

We closely cooperated with the domain experts of the case company to validate the results of the CV model and LCA. Additionally, we were in frequent contact with them to obtain real data and assumptions as inputs for LCAs. Table 8.1 on page 178 describes the domain experts as well as the IDs used for the remainder of this work.

Tab. 8.1.: Domain experts for the machining tools case.

ID	Role	experience in years
Alpha	development engineer	5-10
Beta	development engineer	5-10
Gamma	application engineer	<5
Delta	application engineer	>10

Rotating Anodes

Modern medicine relies on X-rays for the diagnosis of injuries and illnesses, for example, broken bones (Behling, 2015, p. 110)[2], breast cancer (Behling, 2015, p. 139), and cardiac and vascular diseases (Oppelt, 2005, p. 479). To produce X-rays, a cathode emits electrons via thermal emission. These electrons are accelerated towards an anode, where they are decelerated, and X-rays emerge as a result (Behling, 2015, p. 180). Rotating anodes are typically used for high-intensity X-rays (Behling, 2015, p. 18). The rotation counteracts overheating because the electron beam from the cathode hits different spots along the circumference of the anode (Oppelt, 2005, p. 283). The rotating anode can be considered the most important part of an X-ray tube, and is usually one of the most expensive parts. It determines the performance of the overall system (Behling, 2015, p. 233, Oppelt, 2005, p. 280). During operation, temperatures of rotating anodes reach up to 1,500 °C (2,732 °F) (Behling, 2015, p. 240) with microsecond-long pulses of up to 2,500 °C (4,532 °F) (Mehranian et al., 2010). Consequently, in the area where the electron beam hits the rotating anode, the focal track erodes due to extreme thermal cycling (Behling, 2015, p. 235, 247). Focal tracks show two types of wear (Behling, 2015, p. 248): *Cracks* occur because of repeated heating and cooling, they provide stress relief. *Molten areas* appear because the grains of the focal track are isolated due to cracks, and, consequently, heat conduction to the surrounding material is limited. The focal track wear often limits the lifespan of a rotating anode (Erdélyi et al., 2009). Examples of rotating anodes are shown in Figure 8.5 on page 179. Figure 8.6 on page 179 shows a microscopic image of an eroded focal track.

The state of the focal track often limits the lifespan of the rotating anode. We utilize a DL-based CV model to detect and quantify the wear state of the focal tracks of the worn rotating anodes. In this case, it is not only more scalable and reproducible to characterize the wear state of the product. Owing to the large size of the rotating

[2]Please note that rotating anodes are a niche and specialized topic. The reference books (and largely the only ones) for this topic are Behling (2015) and Oppelt (2005), which is why they are frequently used in this section. The rights for the reuse of images (Figure 8.5 and Figure 8.6) are granted.

Fig. 8.5.: Rotating anodes based on Behling (2015).

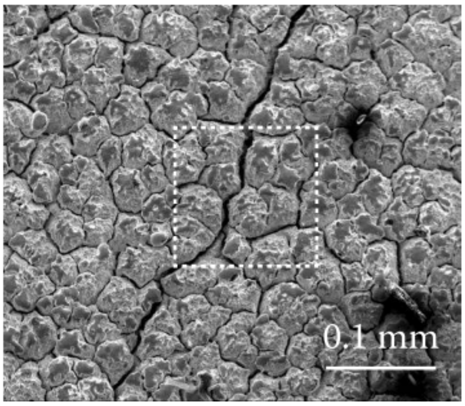


Fig. 8.6.: Microscopic image of eroded rotating anode focal track from Behling (2015).

anodes and focal tracks and the small size of the wear mechanisms, it is almost impossible for humans to assess the wear state in detail. During our project, we captured microscopic images of the focal tracks, which enable an assessment of the wear state. A single microscopic image is more than 19,000 pixels high and 5,000 pixels wide. This microscopic image shows less than 0.5% of the entire focal track. Hence, a detailed manual assessment of the wear state of an entire focal track is tedious.

The possible sustainable smart PSSs resulting from this detailed assessment of the wear state by CV can be grouped into *remanufacturing and recycling decisions, re-design of rotating anodes based on actual wear*, and *establishing the foundation for result-oriented PSSs*. In this section, we further describe these three approaches for improving environmental sustainability.

Tab. 8.2.: Domain experts for the rotating anode case.

ID	Role	experience in years
Epsilon	development engineer	5-10
Zeta	development engineer	>10
Eta	sales manager	>10
Theta	sales manager	<5

An assessment of the wear state of the focal track of a rotating anode using CV enables making an efficient and reproducible decision regarding the options remanufacturing and recycling. There are different options for remanufacturing depending on the type and severity of the wear. If none of the remanufacturing options is applicable, the rotating anode must be recycled. As described by Fang et al. (2015), uncertainty regarding the state of returned products is a major hurdle in remanufacturing endeavors, which is addressed by our approach. In Section 8.3.2, we present the results of an LCA that compares a baseline scenario with a remanufacturing scenario.

Similar to machining tools, the development of new generations of rotating anodes can be aided by actual usage behavior and wear instead of internal laboratory settings.

The rationale behind wear information serving as the foundation for establishing result-oriented PSSs is also in accordance with that regarding the machining tools described in Section 8.2.3.

For this case study, we collected images of focal tracks of several worn rotating anodes that were sampled for a realistic wear distribution. As described before, the images captured by the microscope are large; therefore, labeling an entire image manually would be extremely tedious. Instead, we choose 1,106 representative small patches of microscopic images together with domain experts, labeled them, and used them for the training and evaluation of our DL-based CV.

As stated previously, we cooperated closely with the domain experts of the respective case company. In addition to selecting representative image patches, they provided input data for the LCA and validated the results of the CV model and LCA. Table 8.2 describes the domain experts as well as the IDs we use for the remainder of this work.

8.3 Results

In this section, we report the application results of our previously described methodology. More precisely, we first describe the CV results in Section 8.3.1 and, subsequently, in Section 8.3.2, the results of the LCAs.

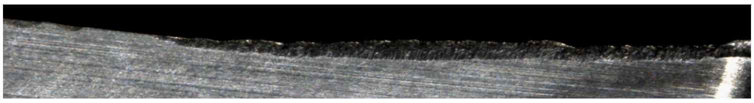
8.3.1 Computer Vision Results

In the following, we describe the results of the DL-based CV models for the machining tools and rotating anode case study.

Machining Tools

We split our dataset containing 213 microscopic images of worn cutting tools into 152 training images, 10 validation images, and 51 test images. Figure 8.7 shows an exemplary original image, human label, and prediction from the test set.

Table 8.3 on page 182 lists the performance results of our trained U-Net on the



(a) Preprocessed input image.

(b) Human label.

(c) Prediction.

Fig. 8.7.: Image from test set with the corresponding human label and prediction of the neural network (best viewed in color).
Color coding: flank wear = dark grey/red, chipping = light grey/green, and built-up edge = white/blue.

unseen test set. Theoretically, the DSC can reach values between zero (no overlap between prediction and ground truth) and one (perfect overlap). In our case, depending on the wear type, we obtained results between 0.244 (chipping) and 0.991 (background). We attribute the relatively poor prediction performance for chipping

Tab. 8.3.: Performance results for the machining tools dataset.

Class and Dice coefficients	
Background	0.991
Flank wear	0.695
Chipping	0.244
Built-up edge	0.596
Mean DSC	0.631
Pixel accuracy	0.977

to two phenomena: first, chipping seldom occurs in our dataset. Consequently, there is relatively little data from which the model can learn. Second, the DSC drops to zero in the case of a false positive, that is, the respective wear class is predicted but not present in the ground truth, which occurs several times for chipping in our dataset. We obtained a mean DSC value of 0.631. Although there is no unified scale to judge the results, they are in accordance with related work in the medical domain (Zou et al., 2004). Guindon and Zhang (2017) [p.51] state that a DSC of 0.7 is "deemed to be indicative of an excellent match between the segmentation result and human expert delineation". However, to further validate the results, we engaged in a dialogue with domain experts from our case company. During the discussions, they confirmed the feasibility of our approach (Alpha and Beta).

Although an efficient assessment of the wear state of machining tools is already helpful for domain experts, the utility is further increased by the integration of the outputs in a user-centric artifact. A screenshot showing an exemplary view of the artifact is shown in Figure 8.8 on page 183. This user-centric artifact was developed iteratively with seven domain experts as end users. It visually and statistically aggregates a dataset consisting of images of worn cutting tools from the same process assessed using CV. Overall, it enables domain experts to interactively explore the dataset. Consequently, they can combine the information extracted through CV with their domain expertise to make data-driven decisions. For example, an application engineer can use the results of the DL-based CV model for wear detection in the user-centric artifact to optimize the machining processes of customers.

Rotating Anodes

We split our dataset of 1,106 microscopic image patches into 1,031 training images, 37 validation images, and 38 test images. The results reach performance levels comparable to those of the machining tools case study. These results are listed in Table 8.4 on page 184. They demonstrate the feasibility of extracting relevant

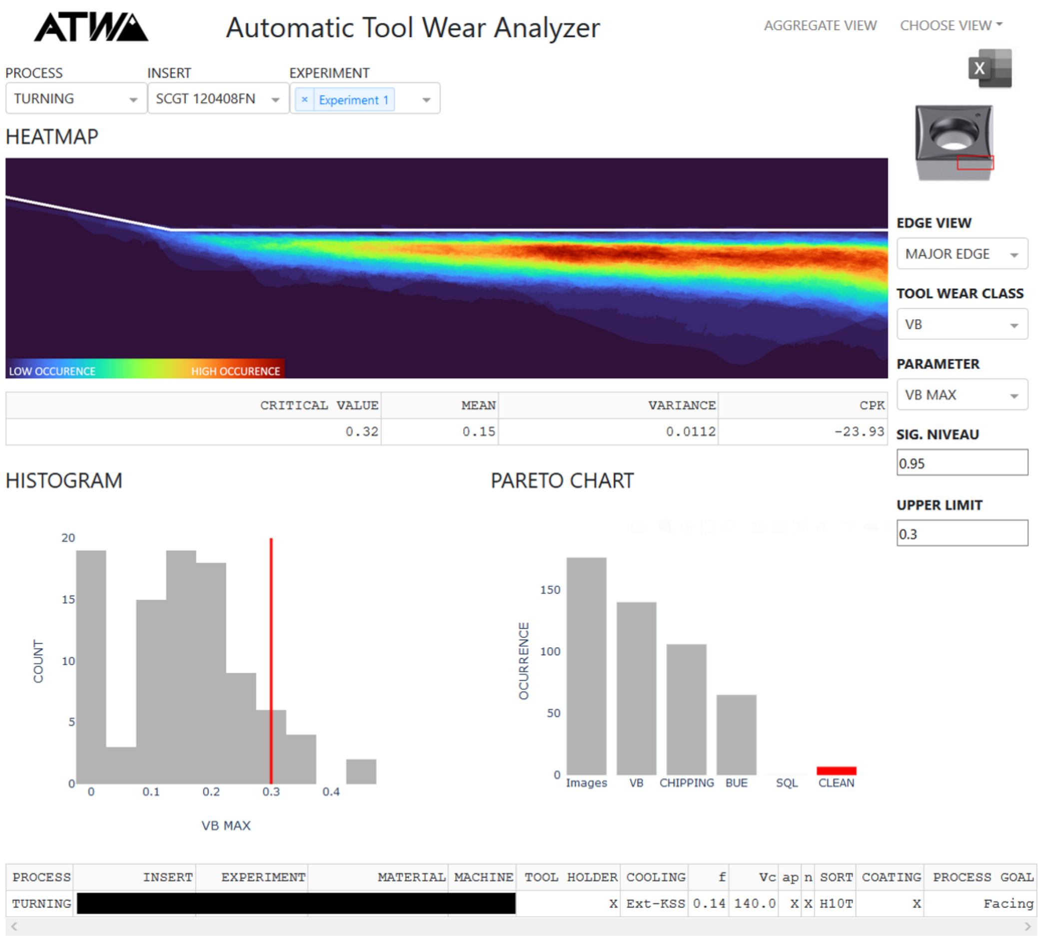


Fig. 8.8.: Exemplary screenshot of user-centric artifact for machining tool wear interpretation.

information from images using CV according to domain experts (Epsilon and Zeta). In direct comparison with the machining tools, the somewhat lower values for the evaluation metrics can be explained by the fact that the images of focal tracks of the rotating anodes do not contain any black background, which is particularly easy to detect.

A similar user-centric artifact is conceptualized for rotating anodes, which allows domain experts to explore the wear state as detected by CV. Domain experts can use it to better understand the wear of rotating anodes — a prerequisite for designing new generations of rotating anodes based on the usage behavior and for offering result-oriented PSSs.

Tab. 8.4.: Performance results for the rotating anodes dataset.

Class and Dice coefficients	
Normal surface	0.485
Cracks	0.634
Molten area	0.690
Mean DSC	0.603
Pixel accuracy	0.737

8.3.2 Life Cycle Assessments

Using the performance metrics of the CV models, the potential impact savings and environmental effects owing to their usage within the processes at our case companies can be calculated. We followed the LCA terminology depicted in Section 8.2.2.
Goal and scope definition: The main objective of the current LCAs is to assess the potential impact savings facilitated by the DL-based CV models to enhance the usage and lifespan of manufactured products through PSSs. For the two case studies, the status-quo situation is compared with improvement scenarios that are supported by the DL-based CV models:

- For the machining tools case, the FU is defined as follows: Manufacture 100 unit shafts (42CrMo4, 800 grams) per hour with tungsten carbide cobalt cutting tools (WC-Co, 9.06 grams) according to predefined specifications.
- For the rotating anode case, the FU is defined as follows: Provide two X-Ray rotating anodes according to predefined specifications for assumed usage of five years.

The system boundaries for the two case studies are as follows:

- For the machining tools case, the cutting tool, electricity for the machining center, cutting fluid, and CV model for wear assessment are considered within the scope of the LCA.
- For the rotating anode case, two rotating anodes are within the scope of the LCA. Depending on the scenario, either both are produced from scratch, or the second is remanufactured. Additionally, transportation from the production site to customers (and back if necessary) is considered. Additionally, the CV model for wear assessment is accounted for.

Life cycle inventory (LCI): Product factsheets and real measurements provided by our case companies have been used for the LCI whenever possible, e.g., for the exact material composition of the cutting tool, and the energy and resource consumption of several production steps of rotating anodes. If this was not possible, assumptions made by the domain experts described in Table 8.1 on page 178 and Table 8.2 on page 180 were used in combination with the scientific literature on the respective topics.

For the machining tools case, the expected lifespan under the baseline scenario of the cutting tool is estimated to be 30 minutes of operation (Gamma and Delta). The manufacturing of one unit shaft is estimated to take 30 seconds. Considering breaks for the employee and setup times, we assume that 100 unit shafts can be produced in one hour (Gamma and Delta). The energy consumption of the machining center is estimated to be 12.5 kWh (Gamma and Delta) on average, using the German electricity mix, that is, where the pieces are manufactured in the present case. Regarding the lubrication system, the consumption of cutting fluid is 0.0155 liters per hour, based on real data from our case company's customer. Finally, the actual surplus energy consumption due to the training of the DL-based CV model for wear assessment is considered and detailed in the improvement scenarios to account for potential impact transfers (Bonvoisin et al., 2014).

The considered improvement scenarios are improvements in machining processes at customer sites by application engineers using the DL-based CV model for wear assessment. We make the (rather conservative) assumption that one trained model can be used to manufacture 1000 unit shafts. In theory, the CV model can be used an infinite number of times for the same type of machining process. The electricity usage to train and run the CV model and run the user-centric artifact is estimated to be 2.4 kWh, as detailed in the Appendix on page 195.

The improvement of machining processes usually aims at increasing both the lifespan of the cutting tool and the process speed, while still meeting the predefined specifications. In the following, typical machining process improvement scenarios for our FU are described based on the experience of our case company's application engineers. According to the domain experts (Alpha, Beta, Gamma, and Delta), an efficient assessment of the wear state can support these process improvements by providing accurate and fine granular information, and can hence lead to better outcomes. Additionally, the domain experts (Alpha, Beta, Gamma, and Delta) confirmed that the wear assessment by CV is highly efficient because little manual effort is required. Consequently, it is possible to provide additional process improvements to customers. A typical machining process improvement for our FU enables an enhanced lifespan of 20% on average for the present cutting tool, and increases the speed by 20% on average with a maximum increase of up to 50%. Note that there exists a clear trade-off between the cutting speed and lifespan. Higher cutting speeds result in shorter tool lifespans. For this process, it is estimated that increasing the cutting speed by 20% decreases the tool lifespan to 70% of its original lifespan, and increasing the cutting speed by 50% decreases the tool lifespan to 30%, based on experience from internal tests (Beta, Gamma, and Delta) and existing literature (e.g., Klocke and König (2008)). Despite this trade-off relationship, domain experts (Gamma and Delta) confirm that it is often possible to achieve a longer lifespan and higher cutting speed. Based on this, the improvement scenarios are computed as follows:

- Lifespan increased by 20%
- Speed increased by 20% (implies more wear and tear on the tool as aforementioned, i.e., the tool needs to be replaced more often but less electricity and cutting fluid consumption for the same FU)
- Speed increased by 50%
- Lifespan increased by 20% and speed increased by 20%
- Lifespan increased by 20% and speed increased by 50%

For the rotating anode case, we work with a typical rotating anode that weighs 1.9 kilograms and comprises 12.5% tungsten-rhenium alloy (a typical 95% tungsten and 5% rhenium mix (Oppelt, 2005, p. 284)) for the focal track, 12.5% graphite for the metallic disc, and 75% molybdenum for the cup. For this LCA, we explicitly consider the energy and resource consumption of the production steps because they have a high impact.

The lifespan of the rotating anode is estimated by experts (Eta and Theta) to be on average 2.5 years. Because the FU refers to five years of operation, two new rotating anodes are required in the baseline case, which corresponds to the current predominant usage in this industry. As described in Section 8.2.3, the state of the focal track often limits the lifespan of the rotating anode (Erdélyi et al., 2009). Different remanufacturing strategies can be applied to restore the focal track depending on its wear state. Consequently, the rotating anode has another 2.5 years of estimated lifespan (Eta and Theta). The CV model for wear assessment is crucial for determining whether remanufacturing is possible and the suitable strategy.

For the LCAs of the different scenarios, we almost always consider real measurements of, for example, the energy consumption of different production steps. If such values are not available, we rely on the assumptions made by the domain experts of our case company. Additionally, for this LCA, we considered the electricity consumption of the DL-based CV model, which is estimated to be 2.875 kWh (description of this can be found in the Appendix on page 195).

To factor in transportation impacts, we consider distances from our case company in Austria to representative customers in Europe (874 km on trucks), and Asia and the United States (124 km on trucks and on average 8930.5 km by plane). Rotating anodes are typically recycled close to their last usage sites. Consequently, in the baseline case, we assume only a one-way trip from the production site to the respective customers. In the remanufacturing scenario, a round trip from the customer to the production site and back is considered. The transportation phase is represented by ton-kilometers (tkm), which is defined as the transport of 1 ton of material over a distance of 1 km (Goedkoop et al., 2008).
Note that the impacts from the infrastructure needed to support the manufacturing facilities are beyond the scope of this study and therefore not included in the LCAs of both case studies.

Life Cycle Assessment: Machining Tools

For the machining tools case study, we first compare the carbon footprint of the baseline with the DL-based CV-supported improvement scenarios. We then expand our analysis to the 18 ReCiPe midpoint indicators to consolidate our interpretation and/or fine-tune our recommendations, for example, in the case of impact transfers. The baseline scenario has a carbon footprint of 8.013 kg CO_2 eq. per hour to manufacture 100 unit shafts, as described in the FU. Enabled by process improvement with the DL-based CV model for wear assessment, the combination of keeping the cutting tool in use closer to its maximum lifespan (original +20%) and increasing the process speed by 50% allows a reduction in the global warming potential of almost 1 kg CO_2 eq. per hour (around 12% of baseline scenario), as illustrated in Figure 8.9 on page 188. Considering complementary environmental indicators (see Figure 8.10 on page 189), the lifespan +20% and speed +20% improvement scenario leads to the most significant mitigation of environmental damage. The reduction in the carbon footprint is slightly lower than in the lifespan +20% and speed +50% scenario. However, there is considerably less transfer to other impact categories. Note that only increasing the lifespan of the cutting tool by 20%, with the support of the DL-based CV model for wear assessment, is not a relevant strategy in terms of the carbon footprint because of the surplus of impact allocated to the training of the CV model. Consequently, a dedicated LCA is necessary to ensure the environmental benefits of other machining process improvements.
A limitation of this LCA is that we quantify the impact of the cutting tool based on material data from the Ecoinvent database and a literature value (Furberg et al., 2019) for hard metal sintering of 11 kWh/kg. However, the impact of an individual tool can vary significantly (>100%) depending on the raw materials, production technologies, and energy sources (Alpha).

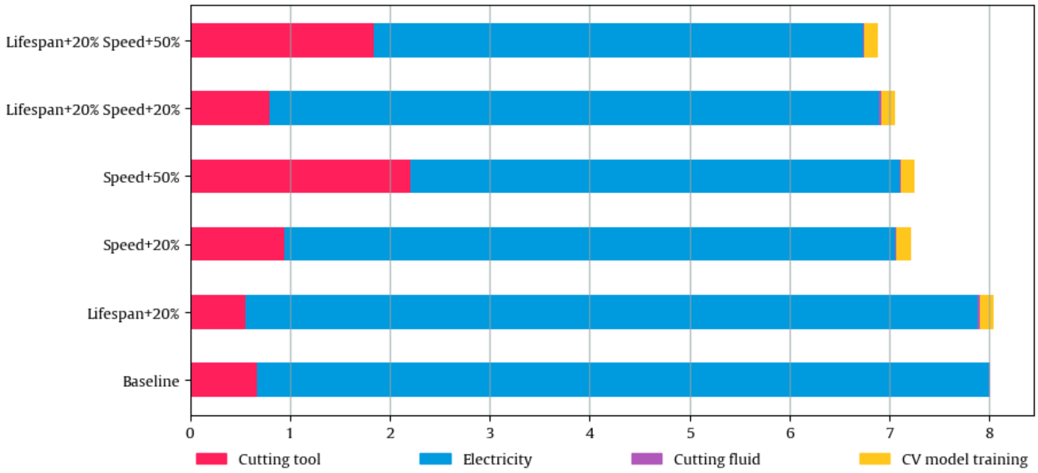


Fig. 8.9.: Carbon footprint for multiple scenarios in the machining case.

Life Cycle Assessment: Rotating Anodes

In the rotating anode case, the DL-based CV-supported remanufacturing scenario leads to significantly increased environmental sustainability (compare Figure 8.11 on page 190 and Figure 8.12 on page 191 for European and non-European customers, respectively). This is possible because many energy- and resource-intensive processes for producing rotating anodes do not have to be repeated. The carbon footprint is reduced by 44.79% / 39.26% in the remanufacturing scenario (European/non-European customers). Figure 8.13 on page 192 illustrates that no impact transfers occur for the improvement scenario for the rotating anode case.

8.4 Discussion

In this section, we interpret the obtained results, relate them to similar literature, conceptualize our approach, and explore possible implications. It is our hope that other researchers and practitioners can transfer it to similar scenarios.

8.4.1 Interpretation of Results

Regarding the CV results, it is essential to note that the detection performance of DL-based CV models will further improve with a higher amount of training data (C. Sun et al., 2017). In terms of the LCAs, several considerations should be remembered. Generally, we want to highlight again that the sustainability improvements cannot

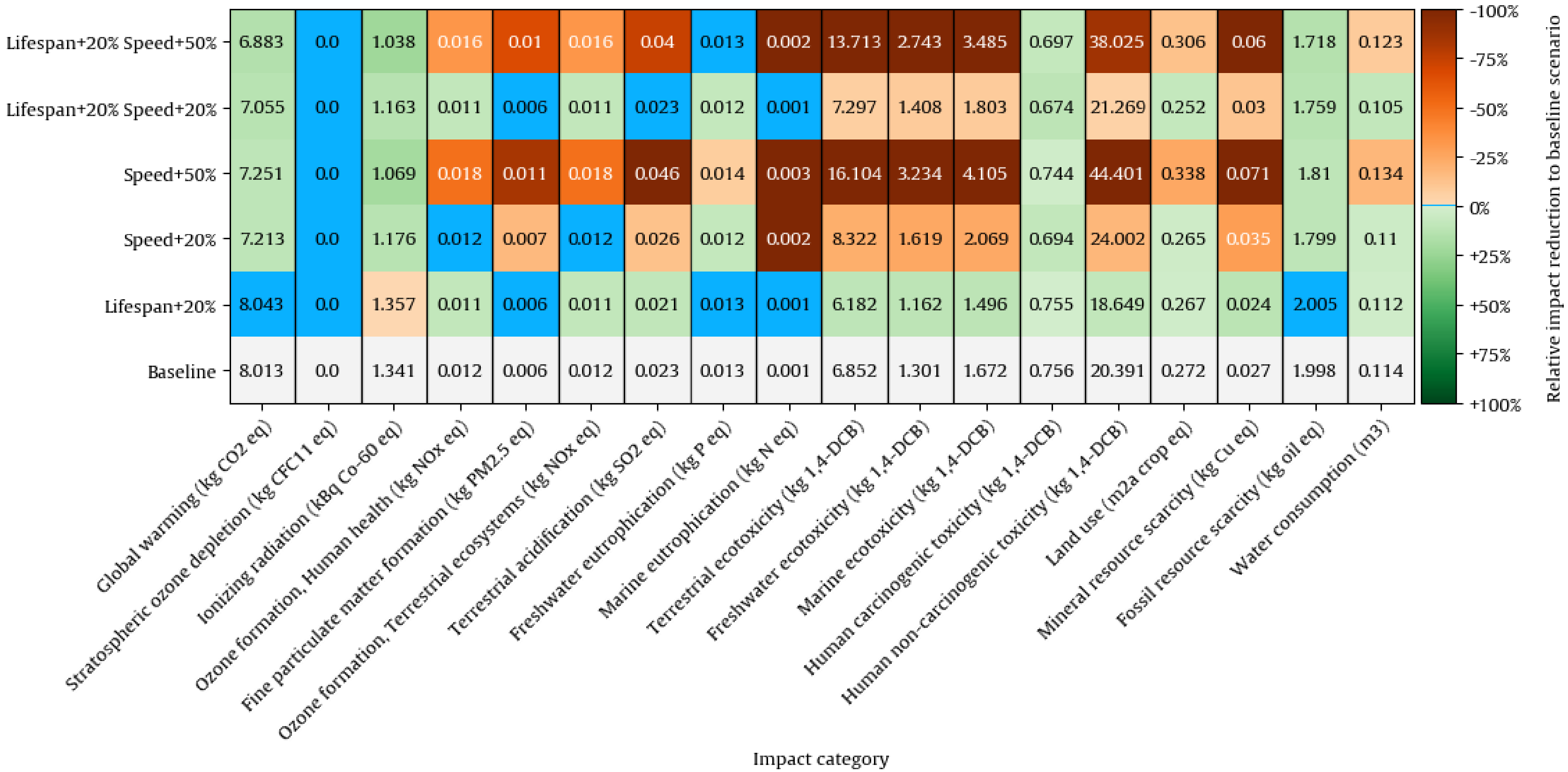


Fig. 8.10.: Comparative results for the machining tools case between the baseline and the five DL-based CV-enabled improvement scenarios, for the 18 ReCiPe midpoint indicators.

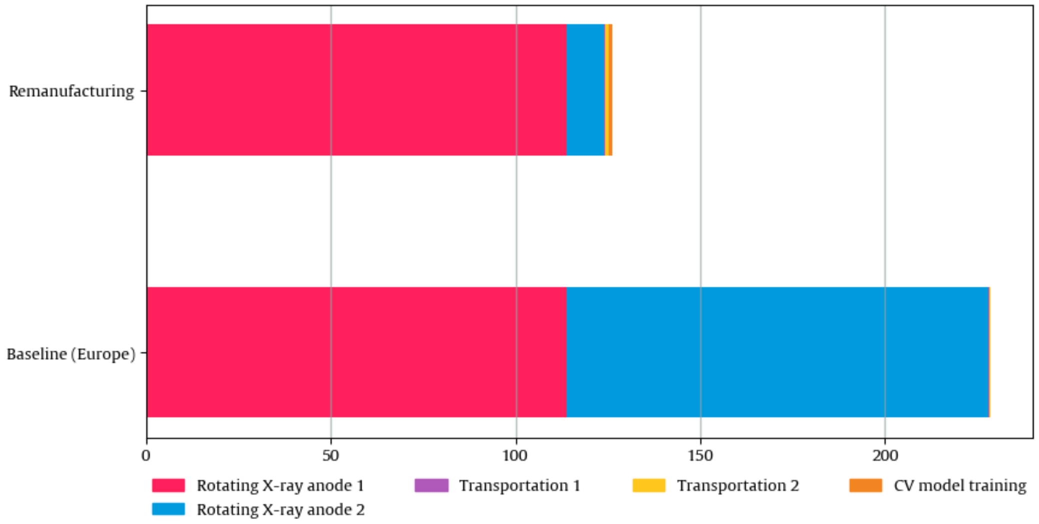


Fig. 8.11.: Carbon footprint in the rotating anode case for multiple scenarios for the European market.

be attributed only to the DL-based CV models. To a certain extent, the described improvement scenarios are also possible with a selective manual visual inspection of the wear states of the products. For example, machining processes of customers are already being improved by application engineers. However, the DL-based CV model for wear assessment facilitates further process improvements owing to the efficiency of the CV system. Additionally, domain experts confirm that more effective machining process improvements are possible because of the large number of images that can be assessed efficiently. Consequently, decisions are made on a better statistical basis. Similarly, in the rotating anode case, it is possible to realize the described improvement scenario without a DL-based CV model for wear assessment. However, for this type of product, manual visual inspection would be highly inefficient owing to the image size and the amount of wear to be detected. Consequently, the improvement scenario for the rotating anode case is only partially implemented in the respective industry.

The LCA results indicate a clear improvement in terms of environmental sustainability in the rotating anode case. The remanufacturing of the focal track leads to significant environmental savings (up to 44.79%) because many resource- and energy-intensive processes necessary to produce rotating anodes do not have to be repeated. In comparison, improvements in terms of environmental sustainability in the machining case study seem minor. However, considering the improvements over a more extended period of time, they become considerable. The reduction of almost 1 kg of CO_2 equivalent in the improvement scenario with a 20% increase in both lifespan and process speed is for a single hour of production on a single machining

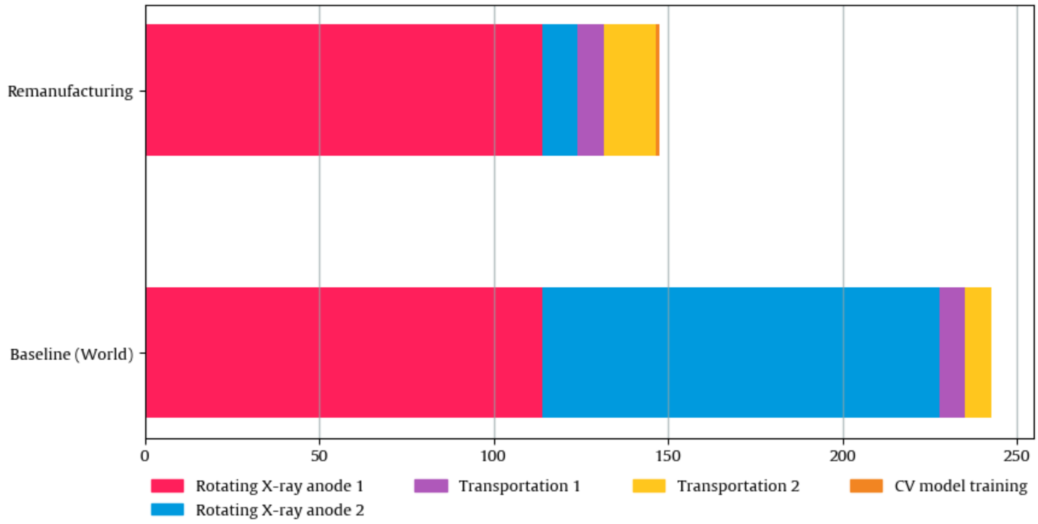


Fig. 8.12.: Carbon footprint in the rotating anode case for multiple scenarios for the non-European market.

center. Scaling this to an entire shift of eight hours and five working days for the duration of a year leads to a saving of 2080 kg CO_2 equivalent for one machining center. Additionally, machining is an extremely widespread manufacturing process; consequently, these savings can be scaled up to numerous production sites and processes.

For both LCAs, it is interesting to consider the current trend towards an emission-free electricity mix in many countries (IEA - International Energy Agency, 2021). For the rotating anode case, the emissions resulting from transportation will become more relevant. For the machining tools case, the lifespan of the tool will play an even more significant role than it currently does.

8.4.2 Related Work: Similarities and Differences

Our approach is a type of sustainable smart PSS described by X. Li et al. (2021). As proposed by them, we do not focus only on the sustainability of physical materials and components but also consider the information value that can be extracted from physical products. However, in contrast to previous studies (e.g., X. Li et al. (2021) and Zhang et al. (2017)), our approach does not rely on smart products in the sense of connected products. Consequently, we extend X. Li et al. (2021)'s conceptualization to a wider range of products. There is a multitude of reasons why a product is not equipped with sensors or radio frequency identification tags: it might not be economically viable, technically not possible, or not desirable from

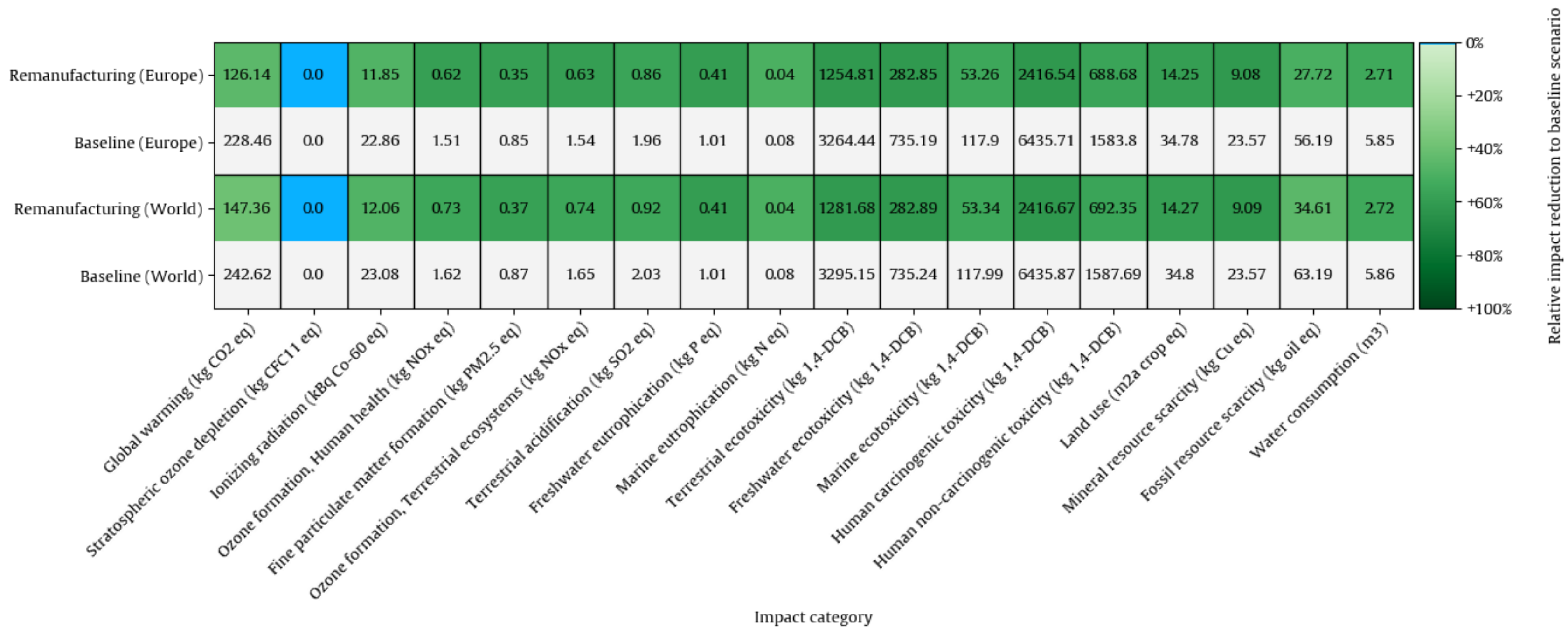


Fig. 8.13.: Comparative results for the rotating anode case for the European and non-European market between the baseline and the respective two DL-based CV-enabled improvement scenarios, for the 18 ReCiPe midpoint indicators.

a data privacy point of view. One might also argue that the camera capturing the pictures is our sensor (Martin et al., 2021). Then one could consider our approach as a type of smart retrofitting, which Jaspert et al. (2021) describe as ”a sustainable approach of transforming the current state of legacy equipment into smart and connected assets.“ Our approach differs from previously documented ones because a static image provides information about the usage stage. In that sense, our analysis is forensic, as we do not have live data about the usage stage or several points in time.

8.4.3 Conceptualization of Our Approach

In the following section, we describe the prerequisites that need to be fulfilled to make the proposed approach possible. First, it must be reasonable to implement potential improvements in environmental sustainability, such as re-design, remanufacturing, reuse, and recycling (4R), or a result-oriented PSS based on wear assessment by CV. Hence, the product under consideration must be produced and used in sufficient quantity now and in the future. Second, it must be possible to obtain images wherein the wear state of the product can be visually assessed. In this regard, it is also necessary to have sufficient information about the product or usage process to counterbalance the real-life variance in terms of usage and observable wear. For the case of machining tools, this is given because we analyze images of many tools from the same production process. For the rotating anodes, this is given by the large image size. In addition to these technical aspects, organizational aspects are also important. The change from a linear business model to a more circular one requires a willingness to transform of both parties involved — customer and provider (Ceschin, 2013). Additionally, co-creation between the provider of the product and the user is helpful, as described by Arnold (2017). In our cases, the providing company can perform the wear state assessment and draw relevant insights based on their domain knowledge. However, most business decisions benefit from additional metadata such as usage process parameters, which typically belong to the company using the product.

8.4.4 Implications

As demonstrated in this work, DL-based CV can facilitate sustainable smart PSSs. This can yield numerous benefits for the manufacturing industry. Primarily, providers and consumers can reduce the environmental impact of their respective overall systems.

Additionally, economic benefits can be expected. Domain experts confirm for both case studies that economic benefits are anticipated for providers and customers. Additionally, existing literature (Annarelli et al., 2016; M. Yang & Evans, 2019) confirms that result-oriented PSSs can yield economic advantages. As outlined before, DL-based CV can be used to address uncertainties regarding risks and costs that often hinder the formation of result-oriented PSSs (Erkoyuncu et al., 2011).

8.5 Conclusion

In this work, we demonstrate the effectiveness of deep-learning-based computer vision, a special type of artificial intelligence, for facilitating sustainable smart product-service systems. To this end, we perform two case studies: one on machining tools and another on X-ray rotating anodes. For both case studies, we first demonstrate the feasibility of detecting the wear state with deep-learning-based computer vision as an input for sustainable smart product-service systems. Subsequently, we perform life cycle assessments based on real data and the inputs of domain experts. The results demonstrate the possible improvements in environmental sustainability resulting from sustainable smart product-service systems based on deep-learning-based computer vision.

A limitation of our work is its focus on the environmental dimension of sustainability. The concept of sustainability typically consists of three pillars: economic, environmental, and social. Economic sustainability was not evaluated explicitly in this work; however, as described previously, benefits are expected according to domain experts and the existing literature. Although we did not explicitly evaluate social sustainability, a direct effect on this dimension is not expected. Particularly, the proposed approach aims not to fully automate jobs but to complement human experts in tedious jobs and free their capacity for jobs that are more suited to their skill levels.

We hope that this work will inspire researchers and practitioners to conduct similar studies and look forward to studies extending sustainable smart product-service systems to products that are not inherently smart. This can help the manufacturing industry reduce its environmental impact while increasing its competitiveness. More broadly, we hope that this work accelerates the implementation of novel ideas and artificial-intelligence-based innovations that have a positive environmental impact.

8.6 Appendix for Chapter 8: Energy Consumption for Training of Computer Vision Models

In our improvement scenarios, there are several computer-supported phases: Training of the DL-based CV model, predictions of the DL-based CV model, and execution of the user-centric artifact. The energy consumption of the latter two is negligible because these applications can run on a standard laptop in parallel to other tasks. Training of the DL-based CV model is the most energy-intensive step because it must be performed on specialized hardware. For both machining tools and rotating anodes, approximately five training runs are required to find a suitable model. For the machining tools case, a training run takes 25 minutes, and for the rotating anodes, it takes 30 minutes. The IBM AC922 machine with two graphical processing units we used consumes a maximum of 1.15 kW. Multiplied by five training runs and 25/30 minutes per training, we obtain 2.395 kWh and 2.875 kWh for the machining tools and the rotating anode cases, respectively.

Part V

Finale

9 Conclusion

Technological advances in the field of computer vision based on deep learning enable efficient, automated processing of image data. This offers great potential for value creation — Data Bridge Market Research (2022) valued the global computer vision market at 12.78 billion US$ in 2021 and expects it to grow to 24.19 billion US$ by 2029. But for many application cases, a combination of artificial and human intelligence is required — image-based decision support systems can enable this combination. However, thus far, there is a lack of research regarding image-based decision support systems. Therefore, this thesis investigates image-based decision support systems from various perspectives. We address four research questions related to image-based decision support systems with the help of several use cases from different industry domains.

9.1 Summary and Contributions

Combining data and analytical methods has become a popular way of value creation by solving complex problems and making better decisions (Hunke et al., 2022). In this thesis, we focus on value creation based on image data. The automated extraction of information from image data with computer vision has made significant advances due to the development of novel neural network architectures, increased computing power, and a growing number of freely available image data sets. For many use cases based on image data, a combination of artificial and human intelligence (Dellermann, Ebel, et al., 2019; Hemmer, Schemmer, et al., 2021) in the form of image-based decision support systems seems most beneficial.

This thesis contributes on several levels to enable value creation with image-based decision support systems. To ensure that the findings can be generalized, all studies were performed with real-world data and incorporated domain experts when reasonable. First, we show the technological feasibility of automatically extracting valuable information from image data for our industry use cases. Building on this, we develop, implement and assess a novel approach that enables the beneficial combination of artificial and human intelligence for extracting valuable information from images in a human-in-the-loop system. Second, building on this, we contribute design knowledge for image-based decision support systems developed in two design science research studies — each performed and evaluated based on a different use case in a different industry domain to ensure the practical relevance of our design knowledge. Third, we contribute design knowledge for the superordinate class of computer-vision-based hybrid intelligence systems. To this end, we perform a study following the reflective design science research *strategy II* based on the previous two and four additional use cases from different domains. Lastly, we contribute by demonstrating the possible positive impact of image-based decision support systems on environmental sustainability for two selected industry use cases and conceptualizing the approach.

In this thesis, we address four research questions with different research methods like technical experiments, interview studies, structured literature reviews, and life cycle assessments. In the remainder of this section, we discuss and summarize the contributions of this thesis. The structure follows the research questions described in Section 1.2.

As described in Chapter 1 and illustrated in Figure 1.1 on page 4, the first step for enabling image-based decision support systems is to convert image data into information that computers can process automatically.

> **Research Question 1 (RQ1)**
> How can image data be converted into valuable information by combining artificial and human intelligence?

We address RQ1 with two studies (Chapter 3 and Chapter 4) utilizing the machining tools use case; each study contains a technical experiment. As described in Table 1.1 on page 8 and the respective chapters, the business challenge in the machining tools use case is to identify wear on machining tools to derive improvement options for machining processes and tools. First, in Chapter 3, we develop a deep-learning-based computer vision model to classify microscopic images of worn machining tools regarding the occurrence of relevant wear classes. In this study, we demonstrate that performing image classification for the given use case is possible and explore opportunities for future research. At the time of publication of this study, there was no published work on utilizing only a deep-learning-based computer vision model for the characterization of wear on microscopic images of machining tools. Hence, we contributed by showing the feasibility of this more modern and flexible approach. According to domain experts, this is already helpful, but pixel-granular information would be even more valuable for their use cases.

Building on this, in Chapter 4, we first develop a deep-learning-based computer vision model for semantic segmentation. The model decides for each pixel of a given microscopic image of a worn machining tool which type of wear is present. After showing the feasibility of this approach, we develop and evaluate a human-in-the-loop approach for increasing the accuracy of the overall system. To this end, we develop, implement and evaluate an approach to assess the uncertainty of a given prediction issued by the deep learning model. We show that the chosen uncertainty measure is correlated with the probability of the prediction being faulty. Hence, the uncertainty measure enables the identification of potentially incorrect outputs of the deep learning model without requiring ground truth labels. Therefore, in cases of high uncertainty, a human expert can be incorporated to make the final decision. This approach leads to a higher accuracy with little human effort. As pointed out in Section 1.2 and the study in Chapter 4, we also evaluate this approach on the publicly available Cityscapes dataset (Cordts et al., 2016) to ensure the generalizability of our approach.

Having addressed RQ1, we then consider image-based decision support systems. As described in Section 1.2, turning image data into information is not yet sufficient for value creation in many cases. Image-based decision support systems are a promising approach for achieving this value creation by combining human and artificial intelligence. In terms of the *data, information, knowledge, and wisdom* pyramid (Ackoff, 1989), RQ1 is about turning image data into information. RQ2 is about turning information into knowledge and wisdom (compare Figure 1.1 on page 4 for a graphical depiction). A. S. Lee (2010) states that theory for action and design should become the predominant form of theory. We respond to this call by developing and evaluating design knowledge for image-based decision support systems in our studies addressing RQ2:

Research Question 2 (RQ2)
What design knowledge should guide the development of image-based decision support systems?

In the first study addressing RQ2 (Chapter 5), we develop and evaluate design knowledge for image-mining-based decision support systems, a special class of image-based decision support systems. Image-mining-based decision support systems are suited for use cases where large amounts of image data, e.g., many images describing the same real-world phenomenon, are used as a basis for decision-making. In such use cases, image mining is applied to extract patterns, relationships, and implicit knowledge from these images (Hsu et al., 2002). To ensure the practical relevance and usefulness of our findings, we use the machining tools use case for the development and evaluation of the design knowledge. An image-mining-based decision support system is suitable for the machining tools use case since the wear on machining tools from an identical machining process is subject to variations. Hence, to support decision-making for a given machining process, it is necessary to identify wear on multiple images of worn machining tools and aggregate this information. In this design science research study, we develop design knowledge in the form of design requirements and design principles based on literature and interviews with domain experts. Subsequently, over three design cycles, we implement an artifact based on concrete design features derived from our design principles. We then evaluate the design knowledge and the artifact with different suitable evaluation techniques like technical experiments and focus groups. These evaluation results attest that the design knowledge, as well as the artifact itself, are useful, effective, and efficient.

In addition, we address RQ2 with a study (Chapter 6) developing design knowledge for image-based decision support systems. As described in Section 1.1, image-based decision support systems are suitable for use cases in which the analysis of a single image is sufficient to support business-relevant decisions. Again, we use an industry use case to develop and evaluate the design knowledge; for image-based decision support systems the power line maintenance use case is suitable. In this design science research study, we apply a methodology similar to the one in Chapter 5. The main methodological difference is that as opposed to Chapter 5, the preceding step of extracting information relevant for the use case from image data is covered more extensively. This is due to the lack of respective previous work for the power line maintenance use case. First, we formulate design requirements based on interviews with domain experts and a structured literature review about challenges in power line maintenance. Subsequently, we develop design principles. Based on these design principles, we then instantiate concrete design features in an artifact. The artifact and design knowledge are then evaluated through a technical experiment and interviews, and confirmatory workshops with domain experts. These evaluation episodes indicate the utility of the design knowledge for image-based decision support systems and the artifact itself.

Building on the previous two and four additional use cases, we then address the development of design knowledge for computer-vision-based hybrid intelligence systems. As described in Section 1.2, we understand computer-vision-based hybrid intelligence systems as the superordinate class of image-based decision support systems. Specifically, we ask:

Research Question 3 (RQ3)
What design knowledge should guide the development of computer-vision-based hybrid intelligence systems?

To address RQ3, in Chapter 7, we rely on six real-world computer vision use cases. Building on these use cases, we extract design knowledge according to the reflective design science research *strategy II* (compare Section 2.1 on page 15 for an explanation of this strategy). To this end, we gather six expert researchers for participation in several workshops. Over the course of these workshops, we conceptualize computer-vision-based hybrid intelligence systems and identify four design-related mechanisms: automation, signaling, modification, and collaboration. Based on this, we derive meta-requirements and design principles based on specific design requirements and design features from the real-world computer vision cases. The resulting design knowledge complements the design knowledge for image-based decision support systems and image-mining-based decision support systems since it focuses on facilitating hybrid intelligence. Also, by relying on six use cases and addressing the superordinate class of computer-vision-based hybrid intelligence systems, we understand the contribution as even more generalizable. This study can be the basis for many future research endeavors and informs practitioners regarding the design of computer-vision-based hybrid intelligence systems.

Having addressed the conversion of image data into valuable information (RQ1) and design knowledge for image-based decision support systems (RQ2) and computer-vision-based hybrid intelligence systems (RQ3), we then assess the real-world impact of image-based decision support systems. Due to climate change being one of the most pressing challenges for humanity (Pörtner et al., 2022), the last research question (RQ4) is concerned with the real-world impact of image-based decision support systems in terms of environmental sustainability.

Research Question 4 (RQ4)
How can image-based decision support systems be applied to improve environmental sustainability in the industry?

With the study addressing RQ4, we contribute as described by Gholami et al. (2016) — they call for solution-oriented information systems studies mitigating negative impacts with regard to environmental sustainability. In this study (Chapter 8), we rely on two use cases: the machining tools case used in four of the previous studies and the rotating anode case that is introduced first at this point. Since the rotating anode case is novel at this point, we first show in a technical experiment that it is possible to automatically extract valuable information from image data of rotating anodes. To be precise, we build and assess a deep-learning-based computer vision model to detect wear on the surface of rotating anodes in a pixel-accurate fashion. This has not been shown before to the best of our knowledge. Building on this successful extraction of relevant information from image data for both use cases, we perform a life cycle assessment for each use case. These life cycle assessments are a quantitative evaluation of improvement scenarios in terms of impact on environmental sustainability. The improvement scenarios in this study are based on image-based decision support systems. We rely primarily on real data from our case companies as input for the life cycle assessments. When lacking such real data, we use assumptions by domain experts. The results of these life cycle assessments show that image-based decision support systems can facilitate noteworthy improvements in terms of environmental sustainability: Remanufacturing allows for 44% less emission of CO_2 equivalents in the rotating anode case. For the machining tools case, the emission of CO_2 equivalents can be reduced by 12% through improvements of machining processes. We strongly believe that image-based decision support systems can support sustainability improvements also for other challenges. Therefore, we then conceptualize our approach and depict what is necessary for an application to different use cases.

In summary, we see the following contributions of our work. In Chapter 3, we show the feasibility of utilizing deep learning to characterize wear on microscopic images of machining tools and explore opportunities for future research. More generally, we illustrate how deep learning can be used for novel use cases based on image data. In Chapter 4, we first show how the wear on microscopic images of machining tools can be characterized in a pixel-granular fashion. Building on this, we show how human and artificial intelligence can be purposefully combined for this computer vision task to achieve better results with little human effort in a human-in-the-loop system.

Building on these technical studies, we perform two design science research studies (Chapter 5 and Chapter 6) to contribute design knowledge for image-based decision support systems and the subclass of image-mining-based decision support systems. This can inform the design of similar decision support systems. In Chapter 7, we provide design knowledge for the overarching class of computer-vision-based hybrid intelligence systems. This study serves as a research agenda for future research. Also, the design knowledge can support academics and practitioners in designing their computer-vision-based hybrid intelligence systems in a more human-centric manner. Lastly, in Chapter 8, we contribute by demonstrating how image-based decision support systems for different use cases can support environmental sustainability. Overall, we hope that our contributions inspire practitioners and researchers alike to pursue similar projects.

9.2 Practical Implications

In the following, we shed light on the practical implications of our work. First, we describe practical implications for the industries in which we performed our use cases. Then, we describe more general practical implications independent of concrete use cases and industries. Lastly, we illustrate the practical implications by a hypothetical example of the support provided by this work for the development of an image-mining-based decision support system.

The studies within this thesis have direct implications for the respective industry domains that are used to evaluate feasibility and design knowledge by solving concrete real-world problems. For the machining tools use case, we first show how computer vision models can be used to characterize wear on microscopic images of worn machining tools in a pixel-granular way in Chapter 4. This wear characterization is of high practical relevance in the machining industry. In addition to our previously described use case, it can be used for tool condition monitoring (Dutta et al., 2013). Tool condition monitoring is the frequent inspection of the wear state of a machining tool currently in use. This prevents unnecessary downtime in production and enables full utilization of the tool's lifetime (Ambhore et al., 2015). This is of high economic importance — Kurada and Bradley (1997) suggest that failures of tools are responsible for 20% of production downtime in machining processes. Also, according to Castejón et al. (2007), machining tools and their replacement account for 3-12% of total production cost. Building on the computer vision models, we describe the development, implementation, and evaluation of the automatic tool wear analyzer — an image-mining-based decision support system for the machining industry in Chapter 5. The automatic tool wear analyzer can support the work of two main user groups: First, application engineers can compare different machining processes to choose the best machining process parameters and tools. With the automatic tool wear analyzer, they can make decisions based on the wear state of many automatically analyzed machining tools. This allows for more reliable insights into improvement potentials than the currently predominant manual visual inspection of a few worn machining tools. Also, due to its efficiency, the automatic tool wear analyzer enables more machining process improvements. Second, tool developers can use insights from the automatic tool wear analyzer to develop the next generations of machining tools. Currently, tool development relies mainly on internal, standardized tests in controlled environments. This does not necessarily reflect the actual usage of the machining tools considered. The automatic tool wear analyzer enables tool developers to understand the wear state of many machining tools used in different real machining processes. Consequently, they can develop future generations of machining tools explicitly tailored to the needs and potentials observed in real machining processes. A possible design goal is increased longevity — this translates into increased cost efficiency and environmental sustainability of machining processes. Additionally, in Chapter 8, we demonstrate how computer vision and image-mining-based decision support systems can be used to improve environmental sustainability in the machining tools use case (up to 12% less emission of CO_2 equivalents).

For the power line maintenance use case, we describe the development, implementation, and evaluation of an image-based decision support system in Chapter 6. This decision support system is an important building block for enabling power line inspections with unmanned aerial vehicles. Power line inspection with unmanned aerial vehicles can substitute current hazardous inspection methods like tower-climbing. According to Schwarz and Drudi (2018), this offers great potential for avoiding injuries — they report over 40 fatal and 1200 non-fatal injuries per year among power line workers. Of course, only a share of these accidents occurs during the inspection (rather than during repair activities). Still, an inspection based on unmanned aerial vehicles and our decision support system has the potential to reduce injuries among power line workers by replacing tower-climbing-based inspection. Additionally, this decision support system supports maintenance engineers in their decision-making — for example, through a geographical information system view that enables the identification of adjacent infrastructure faults.

For the rotating anode use case, we first show how computer vision models can be used to identify wear on microscopic images of rotating anodes to derive remanufacturing options — to the best of our knowledge this has not been shown previously. Subsequently, we demonstrate how these computer vision models can improve environmental sustainability in combination with an image-mining-based decision support system (up to 44% less emission of CO_2 equivalents).

In general, computer vision is expected to become even more relevant in industry — according to the 2021 hype cycle for artificial intelligence of the technology research firm Gartner, Inc. computer vision has already passed the “peak of inflated expectations” and is expected to reach the “plateau of productivity” between 2023 and 2027 (Gartner, Inc., 2022). Consequently, the research in this thesis will also gain more relevance because it relies on computer vision and extends the application. We believe that, in particular, the studies in Chapter 3 and Chapter 4 can facilitate this expected, more widespread adoption of computer vision in industry. In Chapter 3, we exemplify how the usage of computer vision can be explored with a relatively low entry barrier in terms of effort by relying on pre-trained computer vision models. The study and findings in Chapter 4 are particularly relevant for industry use cases where finding a highly accurate computer vision model is not feasible. A frequent reason for insufficient model accuracy is a too small amount of data — a problem most prevalent in small and medium-sized enterprises. Approaches to combine human and artificial intelligence for computer vision tasks, like the one described in Chapter 4, can alleviate this problem and consequently lead to more widespread adoption of computer vision in industry.
As described previously, for many use cases, computer vision models alone are insufficient for actual value creation. Image-based decision support systems and the superordinate class of computer-vision-based hybrid intelligence systems enable value creation by combining human and artificial intelligence. While the first image-based and image-mining-based decision support systems and computer-vision-based hybrid intelligence systems exist, the studies in Chapter 5 - 7 are the first ones in which explicit design knowledge for the respective class of system is developed and evaluated. This design knowledge can guide practitioners in the design of computer-vision-based hybrid intelligence systems and image-based and image-mining-based decision support systems for their use cases.

In the following, we describe an additional, hypothetical image-mining-based decision support system to illustrate how practitioners can benefit from the studies in this thesis regarding design, development, and evaluation. The assumed business goal in this use case is to design running shoes with an increased useful lifetime. The image-mining-based decision support system relies on many images of worn running shoes. First, computer vision models are used to identify wear (e.g., abrasion on the sole and holes on the top) on the images of the shoes. A first feasibility study, as described in Chapter 3, can reveal if the images are suited for classifying different wear types with deep-learning-based computer vision. Also for this use case, it appears relevant to characterize wear in a pixel-granular fashion — this enables insights regarding the exact location and extent of wear. Chapter 4 can guide the development of this computer vision model for semantic segmentation. Additionally, in Chapter 4, we demonstrate how human and artificial intelligence can be combined for this task in case the accuracy of the computer vision model is not yet sufficient. To enable faster improvements of the computer vision model, an active learning system can be employed — it can be designed based on the guidelines in Chapter 7. Having built a sufficiently accurate computer vision model to identify wear on running shoes, the next step is to create an image-mining-based decision support system for wear analysis. For this, the design knowledge formulated in Chapter 5 and chapter 6 is helpful — e.g., *image mining* and *metadata* are highly relevant design principles for this use case. Image mining is used to aggregate the wear identified on different images of worn running shoes. Relevant metadata (total distance ran in the respective shoe, distance per different surfaces ran on, etc.) enables an understanding of different wear modes and the underlying reasons. Overall, by combining computer vision, image mining, and human expert intelligence through an image-mining-based decision support system, future generations of running shoes can be designed to address these wear patterns. This leads to longer-lasting running shoes and hence could increase environmental sustainability. The expected effect on environmental sustainability can be measured as shown in Chapter 8.

9.3 Limitations and Future Research

The research in this thesis has several limitations. We already describe the limitations of the studies in this thesis in the individual chapters. In the following, in contrast, we discuss three general limitations of the research on image-based decision support systems in this thesis. Also, we highlight the resulting potential for future research.

The first potential we see for future research is the creation of additional use cases. The research in this thesis already allows for reasonable generalizability of the findings for the following reasons. First, we rely on several use cases with different characteristics. Second, in many studies, we apply appropriate design science research strategies that allow for findings independent of the concrete use cases. Still, future work based on other use cases could further validate and refine the findings. The studies in this thesis all focus on manufactured goods sold from business to business. A study considering goods sold from business to consumer (like the running shoe case depicted before) would be interesting. Additionally, studies working with use cases from entirely different domains, like medicine or biology, could lead to interesting additional findings.

Another overarching avenue for future research is a long-term evaluation of the findings in this thesis. For example, the design knowledge formulated in Chapter 5 and Chapter 6 could be evaluated by a long-term study in which employees of the partner companies use the respective image-(mining-)based decision support system for their daily work over several months. An analysis of the users' interactions with the image-(mining-)based decision support systems and a subsequent focus group with the users could further refine and validate the design knowledge.

Lastly, the image-(mining-)based decision support systems described in this thesis can be classified as descriptive according to the categorization by Davenport (2013). Future research could extend image-(mining-)based decision support systems such that they can be used for predictive and prescriptive tasks. For example, for the power line maintenance case, the image-based decision support system could then predict infrastructure maintenance needs based on location characteristics, infrastructure type, etc., combined with a data history of past maintenance needs. Similarly, the image-mining-based decision support system for the machining tools case could be extended by predicting the expected lifetime of a given tool in a particular machining process. These extensions require suitable prediction models based on a sufficient data history obtained by a more extended usage period of the image-(mining-)based decision support systems. Later, the data accumulated over time could be used for prescriptive components. For example, the image-mining-based decision support system for the machining tools case could prescribe suitable parameters and tools for a given, novel machining process.

Overall, we hope this thesis contributes to extending the meaning of "a picture is worth a thousand words". In its original sense, this famous saying refers to the high information value of images. This thesis contributes to leveraging the information value of images. It shows how image-based decision support systems can be designed and used such that they are of value to society — supporting people in their jobs, creating monetary value, and reducing environmental impact.

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* (cit. on p. 125).

Abbasi, A., & Chen, H. (2008). Cybergate: A design framework and system for text analysis of computer-mediated communication. *MIS Quarterly*, (4), 811–837 (cit. on pp. 5, 70).

Ackoff, R. L. (1989). From data to wisdom. *Journal of applied systems analysis*, *16*(1), 3–9 (cit. on pp. 4, 80, 201, 243).

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 (cit. on pp. 78, 88).

Aggarwal, R., Johns, A., Jayasinghe, J., & Su, W. (2000). An overview of the condition monitoring of overhead lines. *Electric Power Systems Research*, *53*(1), 15–22 (cit. on pp. 115, 138).

Ahmad, S., & Simonovic, S. P. (2006). An intelligent decision support system for management of floods. *Water resources management*, *20*(3), 391–410 (cit. on p. 26).

AI HLEG. (2019). Ethics Guidelines for Trustworthy AI. *B-1049 Brussels* (cit. on p. 51).

Akata, Z., Balliet, D., De Rijke, M., Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K., & Hoos, H. (2020). A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *Computer*, *53*(8), 18–28 (cit. on p. 146).

Akhloufi, M., & Benmesbah, N. (2014). Outdoor ice accretion estimation of wind turbine blades using computer vision. *2014 Canadian Conference on Computer and Robot Vision*, 246–253 (cit. on p. 106).

Alberti, M., Ciurana, J., Rodríguez, 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 (cit. on p. 71).

Alcayaga, A., Wiener, M., & Hansen, E. G. (2019). Towards a framework of smart-circular systems: An integrative literature review. *Journal of Cleaner Production*, *221*, 622–634 (cit. on p. 170).

Alegre, E., Alaiz-Rodríguez, R., Barreiro, J., & Ruiz, J. (2009). Use of contour signatures and classification methods to optimize the tool life in metal machining. *Estonian Journal of Engineering*, *15*(1), 3–12 (cit. on pp. 36, 37).

Altintas, Y. (2012). *Manufacturing automation*. Cambridge University Press. (Cit. on pp. 34, 49, 176).

Ambhore, N., Kamble, D., Chinchanikar, S., & Wayal, V. (2015). Tool condition monitoring system: A review. *Materials Today: Proceedings*, *2*(4-5), 3419–3428 (cit. on p. 205).

Annarelli, A., Battistella, C., & Nonino, F. (2016). Product service system: A conceptual framework from a systematic review. *Journal of Cleaner Production*, *139*, 1011–1032 (cit. on pp. 170, 176, 194).

Arnold, M. (2017). Fostering sustainability by linking co-creation and relationship management concepts. *Journal of Cleaner Production*, *140*, 179–188 (cit. on p. 193).

Arnott, D., & Pervan, G. (2005). A critical analysis of decision support systems research. *Journal of information technology*, *20*(2), 67–87 (cit. on pp. 24, 25).

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) (cit. on pp. 9, 15, 24, 71, 97, 110).

Arnott, D., & Pervan, G. (2014). A critical analysis of decision support systems research revisited: The rise of design science. *Journal of Information Technology*, *29*(4), 269–293 (cit. on pp. 24, 25, 243).

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 (cit. on pp. 71, 74).

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 (cit. on p. 71).

Avellar, G. S., Pereira, G. A., Pimenta, L. C., & Iscold, P. (2015). Multi-UAV routing for area coverage and remote sensing with minimum time. *Sensors*, *15*(11), 27783–27803 (cit. on p. 135).

Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, *39*(12), 2481–2495 (cit. on p. 20).

Baladi, G. Y., Dawson, T., Musunuru, G., Prohaska, M., Thomas, K., et al. (2017). *Pavement performance measures and forecasting and the effects of maintenance and rehabilitation strategy on treatment effectiveness* (tech. rep.). United States. Federal Highway Administration. (Cit. on p. 105).

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)*, 84–91 (cit. on p. 87).

Barbosa, N. M., & Chen, M. (2019). Rehumanized crowdsourcing: a labeling framework addressing bias and ethics in machine learning, 1–12 (cit. on p. 158).

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*, 257–259 (cit. on p. 81).

Beck, R., Weber, S., & Gregory, R. W. (2013). Theory-generating design science research. *Information Systems Frontiers*, *15*(4), 637–651 (cit. on pp. 9, 148).

Behling, R. (2015). *Modern diagnostic x-ray sources: Technology, manufacturing, reliability*. CRC Press. (Cit. on pp. 178, 179, 245).

Ben-Cohen, A., Klang, E., Diamant, I., Rozendorn, N., Raskin, S. P., Konen, E., Amitai, M. M., & Greenspan, H. (2017). CT image-based decision support system for categorization of liver metastases into primary cancer sites: Initial results. *Academic Radiology*, *24*(12), 1501–1509 (cit. on pp. 27, 108).

Berger, B., Adam, M., Rühr, A., & Benlian, A. (2021). Watch me improve—algorithm aversion and demonstrating the ability to learn. *Business & Information Systems Engineering*, *63*(1), 55–68 (cit. on p. 144).

Bergs, T., Holst, C., Gupta, P., & Augspurger, T. (2020). Digital image processing with deep learning for automated cutting tool wear detection. *Procedia Manufacturing*, *48*, 947–958 (cit. on p. 175).

Berlage, T. (2007). Analyzing and mining automated imaging experiments. *Expert Opinion on Drug Discovery*, *2*(4), 561–569 (cit. on p. 81).

Bharati, P., & Pramanik, A. (2020). Deep Learning Techniques—R-CNN to Mask R-CNN: A Survey. In A. K. Das, J. Nayak, B. Naik, S. Dutta, & D. Pelusi (Eds.), *Computational intelligence in pattern recognition* (pp. 657–668). Springer. (Cit. on pp. 104, 144).

Bhatt, C. A., & Kankanhalli, M. S. (2011). Multimedia data mining: State of the art and challenges. *Multimedia Tools and Applications*, *51*(1), 35–76 (cit. on p. 69).

Bishop, C. M. (1995). *Regularization and complexity control in feed-forward networks* (tech. rep.). (Cit. on p. 54).

Black, J. T. (1995). Introduction to Machining Processes. In *ASM Handbooks: Volume 16: Machining* (2nd). ASM International. (Cit. on pp. 34, 40, 49, 74, 175).

Bokrantz, J., Skoogh, A., Berlin, C., Wuest, T., & Stahre, J. (2020). Smart maintenance: A research agenda for industrial maintenance management. *International Journal of Production Economics*, *224*, 107547 (cit. on p. 135).

Bonvoisin, J., Lelah, A., Mathieux, F., & Brissaud, D. (2014). An integrated method for environmental assessment and ecodesign of ICT-based optimization services. *Journal of Cleaner Production*, *68*, 144–154 (cit. on p. 185).

Boothroyd, G., & Knight, W. A. (1989). Fundamentals of machining and machine tools. Marcel Dekker. (Cit. on pp. 34, 49).

Brown, S. J., Goetzmann, W., Ibbotson, R. G., & Ross, S. A. (1992). Survivorshop Bias in Performance Studies. *The Review of Financial Studies*, *5*(4), 553–580 (cit. on p. 48).

Brynjolfsson, E., & McAfee, A. (2011). *Race against the machine: How the digital revolution is accelerating innovation, driving productivity, and irreversibly transforming employment and the economy*. Digital Frontier Press. (Cit. on p. 52).

Brynjolfsson, E., & Mcafee, A. (2017). The business of artificial intelligence. *Harvard Business Review* (cit. on pp. 104, 143).

Burgin, M. (2016). *Theory of knowledge: Structures and processes* (Vol. 5). World scientific. (Cit. on p. 120).

Burl, M. C., Fowlkes, C., & Roden, J. (1999). Mining for image content. *Systemics, cybernetics, and informatics/information systems: analysis and synthesis* (cit. on p. 92).

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 (cit. on pp. 36, 37, 50, 71, 205).

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-Dependent Algorithm Aversion. *Journal of Marketing Research*, *56*(5), 809–825 (cit. on p. 158).

Ceschin, F. (2013). Critical factors for implementing and diffusing sustainable product-service systems: Insights from innovation studies and companies' experiences. *Journal of Cleaner Production*, *45*, 74–88 (cit. on p. 193).

Chakraborti, T., & Kambhampati, S. (2018). Algorithms for the Greater Good! On Mental Modeling and Acceptable Symbiosis in Human-AI Collaboration. *arXiv preprint arXiv:1801.09854* (cit. on p. 147).

Chan, J. W., & Tong, T. K. (2007). Multi-criteria material selections and end-of-life product strategy: Grey relational analysis approach. *Materials and Design*, *28*(5), 1539–1546 (cit. on p. 31).

Chandra Kruse, L., Seidel, S., & Gregor, S. (2015). Prescriptive Knowledge in IS Research: Conceptualizing Design Principles in Terms of Materiality, Action, and Boundary Conditions. *Proceedings of the Annual Hawaii International Conference on System Sciences*, *2015*, 4039–4048 (cit. on p. 118).

Chang, K.-H., Wu, Y.-Z., & Ke, S.-S. (2022). A simulation-based decision support tool for dynamic post-disaster pedestrian evacuation. *Decision Support Systems*, *157*, 113743 (cit. on p. 26).

Chatterjee, S., Brendel, A. B., & Lichtenberg, S. (2018). Smart infrastructure monitoring: Development of a decision support system for vision-based road crack detection. *International Conference on Information Systems 2018, ICIS 2018*, (October) (cit. on pp. 26, 69, 105, 109, 146).

Chaudhuri, N., & Bose, I. (2020). Exploring the role of deep neural networks for post-disaster decision support. *Decision Support Systems*, *130*, 113234 (cit. on pp. 27, 108).

Cheng, Z., Nakatsugawa, M., Zhou, X. C., Hu, C., Greco, S., Kiess, A., Page, B., Alcorn, S., Haller, J., 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 (cit. on pp. 81, 82).

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*, *21*(1), 1–13 (cit. on p. 21).

Chollet, F. (2017). *Deep learning with Python*. Manning Publications. (Cit. on p. 144).

Chollet, F., et al. (2015). Keras. Retrieved August 10, 2022, from https://keras.io. (Cit. on pp. 42, 100, 125, 173)

Churi, N., Pei, Z., Shorter, D., & Treadwell, C. (2009). Rotary ultrasonic machining of dental ceramics. *International Journal of Machining and Machinability of Materials*, *6*(3/4), 270–284 (cit. on p. 175).

CIO.com. (2019). AI unleashes the Power of Unstructured Data. Retrieved August 10, 2022, from https://www.cio.com/article/3406806/ai-unleashes-the-power-of-unstructured-data.html. (Cit. on p. 3)

Ciroth, A., Winter, S., & Berlin, G. (2014). Openlca 1.4 overview and first steps. *GreenDelta, Berlin* (cit. on p. 174).

Clobridge, A. (2010). 5 - metadata. In *Building a digital repository program with limited resources* (pp. 85–109). Chandos Publishing. (Cit. on p. 120).

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*, 3–27 (cit. on p. 70).

Comaniciu, D., Meer, P., & Foran, D. J. (1999). Image-guided decision support system for pathology. *Machine Vision and Applications*, *11*(4), 213–224 (cit. on p. 108).

Cooper, H. M. (1988). Organizing knowledge syntheses: A taxonomy of literature reviews. *Knowledge in Society*, *1*(1), 104–126 (cit. on p. 81).

Cooper, J. S. (2003). Specifying functional units and reference flows for comparable alternatives. *The International Journal of Life Cycle Assessment*, *8*(6), 337–349 (cit. on p. 174).

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3213–3223 (cit. on pp. 9, 52, 58, 201).

Cortes, C., Mohri, M., & Rostamizadeh, A. (2012). L2 regularization for learning kernels. *arXiv preprint arXiv:1205.2653* (cit. on p. 100).

Coyne, I. T. (1997). Sampling in qualitative research. Purposeful and theoretical sampling; merging or clear boundaries? *Journal of Advanced Nursing*, *26*(3), 623–630 (cit. on pp. 75, 90, 127).

Craja, P., Kim, A., & Lessmann, S. (2020). Deep learning for detecting financial statement fraud. *Decision Support Systems*, *139*, 113421 (cit. on p. 25).

Crum, W. R., Camara, O., & Hill, D. L. (2006). Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE transactions on medical imaging*, *25*(11), 1451–1461 (cit. on p. 56).

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 (cit. on pp. 71, 82).

Data Bridge Market Research. (2022). Global computer vision market — industry trends and forecast to 2029. Retrieved August 11, 2022, from https://www.databridgemarketresearch.com/reports/global-computer-vision-market. (Cit. on p. 199)

Davenport, T. H. (2013). Analytics 3.0. *Harvard Business Review*, *91*(12), 64–72 (cit. on pp. 25, 209).

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly: Management Information Systems*, *13*(3), 319–340 (cit. on pp. 73, 131).

de Jesus Pacheco, D. A., ten Caten, C. S., Jung, C. F., Sassanelli, C., & Terzi, S. (2019). Overcoming barriers towards sustainable product-service systems in small and medium-sized enterprises: State of the art and a novel decision matrix. *Journal of Cleaner Production*, *222*, 903–921 (cit. on p. 170).

Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid Intelligence. *Business & Information Systems Engineering*, *61*(5), 637–643 (cit. on pp. 5, 51, 52, 144, 146, 147, 161, 168, 199).

Dellermann, D., Lipusch, N., Ebel, P., & Leimeister, J. M. (2019). Design principles for a hybrid intelligence decision support system for business model validation. *Electronic markets*, *29*(3), 423–441 (cit. on pp. 146, 147).

Deng, C., Wang, S., Huang, Z., Tan, Z., & Liu, J. (2014). Unmanned aerial vehicles for power line inspection: A cooperative way in platforms and communications. *Journal of Communications*, *9*, 687–692 (cit. on pp. 115, 138).

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Li, F.-F. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, 248–255 (cit. on pp. 20, 42).

DeVries, T., & Taylor, G. W. (2018). Leveraging uncertainty estimates for predicting segmentation quality. *arXiv preprint arXiv:1807.00502* (cit. on pp. 52, 53, 58).

Dey, N., Karâa, W. B. A., Chakraborty, S., Banerjee, S., Salem, M. A., & Azar, A. T. (2015). Image mining framework and techniques: A review. *International Journal of Image Mining*, *1*(1), 45–64 (cit. on p. 70).

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, *26*(3), 297–302 (cit. on pp. 56, 173).

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126 (cit. on p. 150).

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155–1170 (cit. on p. 160).

Duan, G., Chen, Y. W., & Sukegawa, T. (2010). Automatic optical flank wear measurement of microdrills using level set for cutting plane segmentation. *Machine Vision and Applications*, *21*(5), 667–676 (cit. on p. 36).

Dutta, S., Pal, S. K., Mukhopadhyay, S., & Sen, R. (2013). Application of digital image processing in tool condition monitoring: A review. *CIRP Journal of Manufacturing Science and Technology*, *6*(3), 212–232 (cit. on pp. 35, 50, 205).

Dutta, S., Pal, S. K., & Sen, R. (2014). Digital image processing in machining. In J. P. Davim (Ed.), *Modern mechanical engineering* (pp. 367–410). Springer. (Cit. on pp. 82, 83).

Edenhofer, O., Pichs-Madruga, R., Sokona, Y., Agrawala, S., Bashmakov, I. A., Blanco, G., Broome, J., Bruckner, T., Brunner, S., Bustamante, M., et al. (2014). IPCC, 2014: Summary for policymakers. *Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.* (cit. on p. 167).

Eisenhardt, K. M. (1989). Agency theory: An assessment and review. *Academy of management review*, *14*(1), 57–74 (cit. on p. 150).

Ellen MacArthur Foundation. (2016). Intelligent assets: Unlocking the circular economy potential. Retrieved April 29, 2022, from https://ellenmacarthurfoundation.org/intelligent-assets-unlocking-the-circular-economy-potential. (Cit. on p. 170)

Erdélyi, M., Lajkó, M., Kákonyi, R., & Szabó, G. (2009). Measurement of the x-ray tube anodes' surface profile and its effects on the x-ray spectra. *Medical Physics*, *36*(2), 587–593 (cit. on pp. 178, 186).

Erkoyuncu, J. A., Roy, R., Shehab, E., & Cheruvu, K. (2011). Understanding service uncertainties in industrial product-service system cost estimation. *International Journal of Advanced Manufacturing Technology*, *52*(9), 1223–1238 (cit. on pp. 176, 194).

Ezugwu, E. (2005). Key improvements in the machining of difficult-to-cut aerospace superalloys. *International Journal of Machine Tools and Manufacture*, *45*(12-13), 1353–1367 (cit. on p. 175).

Faggella, D. (2020). The self-driving car timeline – predictions from the top 11 global automakers. Retrieved August 10, 2022, from https://emerj.com/ai-adoption-timelines/self-driving-car-timeline-themselves-top-11-automakers/. (Cit. on p. 4)

Fang, H., Ong, S., & Nee, A. (2015). Use of embedded smart sensors in products to facilitate remanufacturing. *Handbook of Manufacturing Engineering and Technology*, 3265–3290 (cit. on p. 180).

Ferguson, M., Ak, R., Lee, Y.-T. T., & Law, K. H. (2017). Automatic localization of casting defects with convolutional neural networks. *2017 IEEE International Conference on Big Data (Big Data)*, 1726–1735 (cit. on pp. 18, 243).

Ferro, D. B., Brailsford, S., Bravo, C., & Smith, H. (2020). Improving healthcare access management by predicting patient no-show behaviour. *Decision Support Systems*, *138*, 113398 (cit. on p. 16).

Finkbeiner, M., Inaba, A., Tan, R., Christiansen, K., & Klüppel, H.-J. (2006). The New International Standards for Life Cycle Assessment: ISO 14040 and ISO 14044. *The International Journal of Life Cycle Assessment*, *11*(2), 80–85 (cit. on pp. 12, 173).

Fischer, M., Heim, D., Hofmann, A., Janiesch, C., Klima, C., & Winkelmann, A. (2020). A taxonomy and archetypes of smart services for smart living. *Electronic Markets*, 1–19 (cit. on p. 143).

Fisher, M., & Mackiewicz, M. (2020). Image labelling tool. Retrieved August 10, 2022, from https://bitbucket.org/ueacomputervision/image-labelling-tool/src/master/. (Cit. on pp. 80, 84)

Fitriyani, N. L., Syafrudin, M., Alfian, G., & Rhee, J. (2020). HDPM: an effective heart disease prediction model for a clinical decision support system. *IEEE Access*, *8*, 133034–133050 (cit. on p. 26).

Flath, C. M., & Stein, N. (2018). Towards a data science toolbox for industrial analytics applications. *Computers in Industry*, *94*, 16–25 (cit. on p. 167).

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 (cit. on p. 81).

Forsyth, D. A., & Ponce, J. (2002). *Computer vision: a modern approach*. Prentice Hall. (Cit. on p. 145).

Friedman, J. H. (2011). Greedy Function Approximation : A Gradient Boosting Machine. *The Annals of Statistics*, *29*(5), 1189–1232 (cit. on p. 41).

Frischknecht, R., & Rebitzer, G. (2005). The ecoinvent database system: a comprehensive web-based LCA database. *Journal of Cleaner Production*, *13*(13-14), 1337–1343 (cit. on p. 174).

Furberg, A., Arvidsson, R., & Molander, S. (2019). Environmental life cycle assessment of cemented carbide (WC-Co) production. *Journal of Cleaner Production*, *209*, 1126–1138 (cit. on p. 187).

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *Proceedings of the 33rd International Conference on Machine Learning*, *48*, 1050–1059 (cit. on pp. 51, 53, 55, 159).

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* (cit. on pp. 54, 56, 85).

Gartner, Inc. (2022). The 4 Trends That Prevail on the Gartner Hype Cycle for AI, 2021. Retrieved August 11, 2022, from https://www.gartner.com/en/articles/the-4-trends-that-prevail-on-the-gartner-hype-cycle-for-ai-2021. (Cit. on p. 207)

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 (cit. on pp. 27, 81).

Gehin, A., Zwolinski, P., & Brissaud, D. (2008). A tool to implement sustainable end-of-life strategies in the product development phase. *Journal of Cleaner Production*, *16*(5), 566–576 (cit. on p. 31).

Gholami, R., Watson, R. T., Molla, A., Hasan, H., & Bjorn-Andersen, N. (2016). Information systems solutions for environmental sustainability: How can we do more? *Journal of the Association for Information Systems*, *17*(8), 521–536 (cit. on p. 204).

Gibert, X., Patel, V. M., & Chellappa, R. (2017). Deep Multitask Learning for Railway Track Inspection. *IEEE Transactions on Intelligent Transportation Systems*, *18*(1), 153–164 (cit. on p. 105).

Gillies, R. J., Kinahan, P. E., & Hricak, H. (2016). Radiomics: Images are more than pictures, they are data. *Radiology*, *278*(2), 563–577 (cit. on pp. 6, 78–80).

Gleave, S. D., Frisoni, R., Dionori, F., Casullo, L., Vollath, C., Devenish, L., Spano, F., Sawicki, T., Carl, S., Lidia, R., Neri, J., Silaghi, R., & Stanghellini, A. (2014). *EU Road Surfaces: Economic and Safety Impact of the lack of Regular Road Maintenance–Study* (tech. rep.). (Cit. on p. 105).

Gnewuch, U., Morana, S., & Maedche, A. (2017). Towards designing cooperative and social conversational agents for customer service. *Proceedings of the International Conference on Information Systems (ICIS) 2017* (cit. on p. 16).

Goedkoop, M. J., Heijungs, R., Huijbregts, M., De Schryver, A., Struijs, J., & van Zelm, R. (2008). ReCiPe 2008, A life cycle impact assessment method which comprises harmonised category indicators at the midpoint and the endpoint level. (Cit. on p. 187).

Goedkoop, M. J., van Halen, C. J. G., te Riele, H. R., & Rommens, P. J. (1999). Product service systems, ecological and economic basis. (Cit. on p. 169).

Gonzalez, R., & Woods, R. (2018). *Digital image processing* (4th ed.). Pearson. (Cit. on p. 102).

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org. (Cit. on pp. 45, 104)

Gopalakrishnan, M., Bokrantz, J., Ylipää, T., & Skoogh, A. (2015). Planning of maintenance activities – a current state mapping in industry. *Procedia CIRP*, *30*, 480–485 (cit. on p. 101).

Gorry, G. A., & Scott Morton, M. S. (1971). A framework for management information systems. *Working Paper — Alfred P. Sloan School of Management* (cit. on p. 24).

Gottschlich, J., & Hinz, O. (2014). A decision support system for stock investment recommendations using collective wisdom. *Decision support systems*, *59*, 52–62 (cit. on p. 16).

Gregor, S., Chandra Kruse, L., & Seidel, S. (2020). Research Perspectives: The Anatomy of a Design Principle. *Journal of the Association for Information Systems*, *21*(6), 1622–1652 (cit. on pp. 16, 150).

Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS quarterly*, *37*(2), 337–355 (cit. on pp. 8, 15, 32, 33, 110, 112, 148).

Gregor, S., & Jones, D. (2007). The anatomy of a design theory. *Journal of the Association for Information systems*, *8*(5), 312–335 (cit. on pp. 71, 93, 97, 103, 133, 134, 144, 148).

Griebel, M., Dürr, A., & Stein, N. (2019). Applied image recognition: guidelines for using deep learning models in practice. *14th International Conference on Wirtschaftsinformatik*, 393–406 (cit. on pp. 19, 105, 144, 146, 171).

Griebel, M., Welsch, G., Greif, T., & Flath, C. (2019). A picture is worth more than a thousand purchases: Designing an image-based fashion curation system. *Proceedings of the 27th European Conference on Information Systems (ECIS)* (cit. on p. 146).

Grigorescu, S., Trasnea, B., Cocias, T., & Macesanu, G. (2020). A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, *37*(3), 362–386 (cit. on pp. 4, 143).

Grove, O., Berglund, A. E., Schabath, M. B., Aerts, H. J., Dekker, A., Wang, H., Velazquez, E. R., Lambin, P., Gu, Y., Balagurunathan, Y., et al. (2015). Quantitative computed tomographic descriptors associate tumor shape complexity and intratumor heterogeneity with prognosis in lung adenocarcinoma. *PLoS ONE*, *10*(3) (cit. on p. 78).

Guindon, B., & Zhang, Y. (2017). Application of the dice coefficient to accuracy assessment of object-based image classification. *Canadian Journal of Remote Sensing*, *43*(1), 48–61 (cit. on p. 182).

Guo, Z., Ngai, E., Yang, C., & Liang, X. (2015). An RFID-based intelligent decision support system architecture for production monitoring and scheduling in a distributed manufacturing environment. *International Journal of Production Economics*, *159*, 16–28 (cit. on p. 26).

Hajiali, M., Teimoury, E., Rabiee, M., & Delen, D. (2022). An interactive decision support system for real-time ambulance relocation with priority guidelines. *Decision Support Systems*, *155*, 113712 (cit. on p. 26).

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Morgan Kauffmann. (Cit. on p. 69).

Harris, J. G., & Davenport, T. H. (2005). Automated decision making comes of age. *MIT Sloan Management Review*, *46*(4), 2–10 (cit. on p. 24).

Hastie, T., Tibshirani, R., & Friedman, J. H. (2017). *The elements of statistical learning: data mining, inference and prediction* (2nd ed.). Springer. (Cit. on pp. 145, 173).

Hatfield-Dodds, S., Schandl, H., Newth, D., Obersteiner, M., Cai, Y., Baynes, T., West, J., & Havlik, P. (2017). Assessing global resource use and greenhouse emissions to 2050, with ambitious resource efficiency and climate mitigation policies. *Journal of Cleaner Production*, *144*, 403–414 (cit. on p. 167).

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. *Proceedings of the IEEE international conference on computer vision*, 2961–2969 (cit. on pp. 20, 44).

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*, 1026–1034 (cit. on pp. 17, 50, 69, 80, 83, 102, 105, 167, 170).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (cit. on pp. 20, 105).

Heinrich, K., Roth, A., & Zschech, P. (2019). Everything Counts: A Taxonomy of Deep Learning Approaches for Object Counting. *ECIS 2019* (cit. on p. 143).

Heinrich, K., Zschech, P., Möller, B., Breithaupt, L., & Maresch, J. (2019). Objekterkennung im Weinanbau — Eine Fallstudie zur Unterstützung von Winzertätigkeiten mithilfe von Deep Learning. *HMD Praxis der Wirtschaftsinformatik*, *56*(5), 964–985 (cit. on p. 153).

Hemmer, P., Kühl, N., & Schöffer, J. (2021). Utilizing Active Machine Learning for Quality Assurance: A case study of virtual car renderings in the automotive industry. *Karlsruhe Institute of Technology Working Paper* (cit. on p. 153).

Hemmer, P., Schemmer, M., Vössing, M., & Kühl, N. (2021). Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review. *28th Pacific Asia Conference on Information Systems*, 78 (cit. on p. 199).

Hemmer, P., Vössing, M., & Blumenstiel, B. (2021). Object Detection in Floor Plans: Lessons From Designing a Human-in-the-Loop System. *Artificial Intelligence in Architecture, Engineering and Construction* (cit. on p. 153).

Hendrycks, D., & Gimpel, K. (2016). A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136* (cit. on p. 52).

Hevner, A. R. (2007). A three cycle view of design science research. *Scandinavian journal of information systems*, *19*(2), 4 (cit. on pp. 16, 34, 72, 74, 102, 110, 111, 243).

Hevner, A. R., & Chatterjee, S. (2010). Design science research in information systems. In A. R. Hevner & S. Chatterjee (Eds.), *Design research in information systems* (pp. 9–22). Springer. (Cit. on pp. 110, 111, 127).

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS quarterly*, *28*(1), 75–105 (cit. on pp. 15, 71, 72, 110, 144).

Hirt, R., & Kühl, N. (2018). Cognition in the Era of Smart Service Systems: Inter-organizational Analytics through Meta and Transfer Learning. *39th International Conference on Information Systems*, 1–9 (cit. on p. 47).

Homma, R. Z., Cosentino, A., & Szymanski, C. (2017). Autonomous inspection in transmission and distribution power lines – methodology for image acquisition by means of unmanned aircraft system and its treatment and storage. *24th International Conference & Exhibition on Electricity Distribution (CIRED)*, *2017*(1), 965–967 (cit. on pp. 102, 115, 138).

Hsu, W., Lee, M. L., & Zhang, J. (2002). Image mining: Trends and developments. *Journal of Intelligent Information Systems*, *19*(1), 7–23 (cit. on pp. 6, 76, 79–81, 87, 98, 201).

Huang, L., Xu, D., & Zhai, D. (2018). Research and design of space-sky-ground integrated transmission line inspection platform based on artificial intelligence. *IEEE Conference on Energy Internet and Energy System Integration (EI2)*, 1–4 (cit. on pp. 102, 115, 138).

Hui, X., Bian, J., Zhao, X., & Tan, M. (2018). Vision-based autonomous navigation approach for unmanned aerial vehicle transmission-line inspection. *International Journal of Advanced Robotic Systems*, *15*(1) (cit. on p. 102).

Huijbregts, M. A., Steinmann, Z. J., Elshout, P. M., Stam, G., Verones, F., Vieira, M., Zijp, M., Hollander, A., & Van Zelm, R. (2017). Recipe2016: A harmonised life cycle impact assessment method at midpoint and endpoint level. *The International Journal of Life Cycle Assessment*, *22*(2), 138–147 (cit. on p. 174).

Hunke, F., Heinz, D., & Satzger, G. (2022). Creating customer value from data: Foundations and archetypes of analytics-based services. *Electronic Markets*, *32*(2), 503–521 (cit. on pp. 4, 199).

IEA - International Energy Agency. (2021). Net zero by 2050. Retrieved April 29, 2021, from https://www.iea.org/reports/net-zero-by-2050. (Cit. on p. 191)

Ignatius, J., Hatami-Marbini, A., Rahman, A., Dhamotharan, L., & Khoshnevis, P. (2018). A fuzzy decision support system for credit scoring. *Neural Computing and Applications*, *29*(10), 921–937 (cit. on p. 25).

Iivari, J. (2015). Distinguishing and contrasting two strategies for design science research. *European Journal of Information Systems*, *24*(1), 107–115 (cit. on pp. 11, 16, 17, 144, 148, 163).

INNOSIGHT. (2019). 2021 corporate longevity forecast. Retrieved August 10, 2022, from https://www.innosight.com/insight/creative-destruction/. (Cit. on p. 3)

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International conference on machine learning*, 448–456 (cit. on pp. 55, 100).

Irannezhad, E., Prato, C. G., & Hickman, M. (2020). An intelligent decision support system prototype for hinterland port logistics. *Decision Support Systems*, *130*, 113227 (cit. on p. 26).

ISO - International Organization for Standardization. (1991). ISO 3685:1993 Tool-life testing with single point turning tools. (Cit. on pp. 34, 35, 44, 48, 82, 88, 89).

ISO - International Organization for Standardization. (2006a). Iso 14040:2006 environmental management - life cycle assessment - principles and framework. (Cit. on p. 174).

ISO - International Organization for Standardization. (2006b). Iso 14044:2006 environmental management - life cycle assessment - requirements and guidelines. (Cit. on p. 174).

Jalil, B., Leone, G., Martinelli, M., Moroni, D., Pascali, M., & Berton, A. (2019). Fault detection in power equipment via an unmanned aerial system using multi modal data. *Sensors*, *19*, 3014 (cit. on pp. 101, 119).

Jamshidi, A., Hajizadeh, S., Su, Z., Naeimi, M., Núñez, A., Dollevoet, R., Schutter, B. D., & Li, Z. (2018). A decision support approach for condition-based maintenance of rails based on big data analysis. *Transportation Research Part C: Emerging Technologies*, *95*, 185–206 (cit. on p. 108).

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, *31*(3), 685–695 (cit. on pp. 3, 17, 104, 145, 157, 170).

Jardine, A. K., Lin, D., & Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical Systems and Signal Processing*, *20*(7), 1483–1510 (cit. on p. 101).

Jaspert, D., Ebel, M., Eckhardt, A., & Poeppelbuss, J. (2021). Smart retrofitting in manufacturing: A systematic review. *Journal of Cleaner Production*, 127555 (cit. on p. 193).

Jebari, K., & Lundborg, J. (2020). Artificial superintelligence and its limits: why AlphaZero cannot become a general agent. *AI & SOCIETY*, *36*(3), 807–815 (cit. on p. 158).

Joint Research Centre. (2010). *International Reference Life Cycle Data System (ILCD) Handbook - General guide for Life Cycle Assessment - Provisions and Action Steps*. Publications Office of the European Union. (Cit. on p. 174).

Jones, D. (2005). Power line inspection - a UAV concept. *2005 The IEE Forum on Autonomous Systems*, 8 (cit. on pp. 115, 138).

Kale, P., & Singh, H. (2009). Managing Strategic Alliances: What Do We Know Now, and Where Do We Go From Here? *Academy of Management Perspectives*, *23*(3), 45–62 (cit. on p. 38).

Kassam, S. (2002). Freedom of information. *Intelligent Enterprise*, *5*(7) (cit. on p. 120).

Katrasnik, J., Pernus, F., & Likar, B. (2010). A survey of mobile robots for distribution power line inspection. *IEEE Transactions on Power Delivery*, *25*(1), 485–493 (cit. on pp. 115, 124, 138).

Keim, D. A. (2002). Information visualization and visual data mining. *IEEE transactions on Visualization and Computer Graphics*, *8*(1), 1–8 (cit. on pp. 88, 121).

Kendall, A., Badrinarayanan, V., & Cipolla, R. (2015). Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680* (cit. on pp. 55, 60, 159).

Kendall, A., & Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 5574–5584 (cit. on pp. 51, 53, 57, 64).

Keras. (2022). Keras applications. Retrieved August 10, 2022, from https://keras.io/api/applications/. (Cit. on p. 20)

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*, 433–438 (cit. on p. 80).

Kim, P. (2017). Convolutional neural network. In *Matlab deep learning* (pp. 121–147). Springer. (Cit. on p. 170).

King, N. (1998). Template analysis. In G. Symon & C. Cassell (Eds.), *Qualitative methods and analysis in organizational research: A practical guide* (pp. 118–134). (Cit. on pp. 127, 131).

King, N. (2004). Using Templates in the Thematic Analysis of Text. *Essential guide to qualitative methods in organizational research*, 256 (cit. on p. 91).

Klocke, F., & König, W. (2008). *Fertigungsverfahren 1: Drehen, fräsen, bohren*. Springer-Verlag. (Cit. on p. 186).

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 (cit. on p. 81).

Kohli, R., & Devaraj, S. (2004). Contribution of institutional DSS to organizational performance: evidence from a longitudinal study. *Decision Support Systems*, *37*(1), 103–118 (cit. on pp. 4, 70).

Koornneef, H., Verhagen, W. J., & Curran, R. (2020). A decision support framework and prototype for aircraft dispatch assessment. *Decision Support Systems*, *135*, 113338 (cit. on p. 16).

Kosson, A., & Marklund, H. (2018). It's a no-brainer! Deep learning for brain MR images. Retrieved August 11, 2022, from https://medium.com/stanford-ai-for-healthcare/its-a-no-brainer-deep-learning-for-brain-mr-images-f60116397472. (Cit. on pp. 19, 171, 243, 244)

Kraus, M., Feuerriegel, S., & Oztekin, A. (2020). Deep learning in business analytics and operations research: Models, applications and managerial implications. *European Journal of Operational Research*, *281*(3), 628–641 (cit. on pp. 108, 133).

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/m3. *American Journal of Industrial Medicine*, *30*(1), 16–25 (cit. on pp. 71, 82).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems*, *1*, 1097–1105 (cit. on pp. 19, 37).

Kuckartz, U. (2012). *Qualitative Inhaltsanalyse*. Beltz Juventa. (Cit. on p. 76).

Kucklick, J.-P., & Müller, O. (2020). Location, location, location: Satellite image-based real-estate appraisal. *arXiv preprint arXiv:2006.11406* (cit. on p. 69).

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 (cit. on pp. 16, 32, 72, 74, 243).

Kühl, N., Goutier, M., Hirt, R., & Satzger, G. (2019). Machine Learning in Artificial Intelligence: Towards a Common Understanding. *Proceedings of the 52nd Hawaii International Conference on System Sciences 2019*, 5236–5245 (cit. on p. 3).

Kühl, N., Hirt, R., Baier, L., Schmitz, B., & Satzger, G. (2021). How to conduct rigorous supervised machine learning in information systems research: The supervised machine learning report card. *Communications of the Association for Information Systems*, *48*(1), 46 (cit. on pp. 127, 139).

Kumar, V., Gu, Y., Basu, S., Berglund, A., Eschrich, S. A., Schabath, M. B., Forster, K., Aerts, H. J. W. L., Dekker, A., & Fenstermacher, D. (2012). Radiomics: the process and the challenges. *Magnetic Resonance Imaging*, *30*(9), 1234–1248 (cit. on p. 78).

Kurada, S., & Bradley, C. (1997). A review of machine vision sensors for tool condition monitoring. *Computers in industry*, *34*(1), 55–72 (cit. on pp. 50, 71, 205).

Lambin, P., Leijenaar, R. T. H., Deist, T. M., Peerlings, J., De Jong, E. E. C., Van Timmeren, J., Sanduleanu, S., Larue, R., 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 (cit. on pp. 69, 70, 78, 86, 87).

Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., Van Stiphout, R. G. P. M., Granton, P., Zegers, C. M. L., Gillies, R., Boellard, R., & Dekker, A. (2012). Radiomics: extracting more information from medical images using advanced feature analysis. *European Journal of Cancer*, *48*(4), 441–446 (cit. on pp. 6, 79).

Landwehr, J. P., Kühl, N., Walk, J., & Gnädig, M. (2022). Design Knowledge for Deep-Learning-Enabled Image-Based Decision Support Systems. *Business & Information Systems Engineering* (cit. on pp. 82, 153).

Lanzetta, M. (2001). A new flexible high-resolution vision sensor for tool condition monitoring. *Journal of Materials Processing Technology*, *119*(1-3), 73–82 (cit. on p. 36).

LeCun, Y., Bengio, Y., et al. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, *3361*(10), 1995 (cit. on pp. 19, 170).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444 (cit. on pp. 3, 17, 33, 37, 69, 70, 98, 102, 104, 119, 145, 157, 170).

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, *1*(4), 541–551 (cit. on p. 19).

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 (cit. on pp. 71, 201).

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, *46*(1), 50–80 (cit. on p. 50).

Li, A. Q., Kumar, M., Claes, B., & Found, P. (2020). The state-of-the-art of the theory on product-service systems. *International Journal of Production Economics*, *222*, 107491 (cit. on p. 169).

Li, D., & Wang, X. (2019). The future application of transmission line automatic monitoring and deep learning technology based on vision. *2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, 131–137 (cit. on pp. 115, 138).

Li, F.-F., Johnson, J., & Yeung, S. (2017). Lecture 11: Detection and Segmentation - Stanford University. Retrieved August 10, 2022, from http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf. (Cit. on pp. 19, 171, 243, 244)

Li, X., Wang, Z., Chen, C. H., & Zheng, P. (2021). A data-driven reversible framework for achieving Sustainable Smart product-service systems. *Journal of Cleaner Production*, *279*(August), 123618 (cit. on pp. 168–170, 191).

Liang, H., Zuo, C., & Wei, W. (2020). Detection and evaluation method of transmission line defects based on deep learning. *IEEE Access*, *8*, 38448–38458 (cit. on pp. 107, 109).

Liang, Y.-T., & Chiou, Y.-C. (2006). An Effective Drilling Wear Measurement based on Visual Inspection Technique. *Proceedings of the 2006 Joint Conference on Information Sciences* (cit. on p. 36).

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. *European conference on computer vision*, 740–755 (cit. on p. 125).

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 (cit. on p. 78).

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). SSD: Single Shot MultiBox Detector. *Lecture Notes in Computer Science*, 21–37 (cit. on pp. 20, 105, 124).

Liu, W., Zhang, M., Luo, Z., & Cai, Y. (2017). An ensemble deep learning method for vehicle type classification on visual traffic surveillance sensors. *IEEE Access*, *5*, 24417–24425 (cit. on p. 144).

Liu, X., Miao, X., Jiang, H., & Chen, J. (2020). Review of data analysis in vision inspection of power lines with an in-depth discussion of deep learning technology. *arXiv preprint arXiv:2003.09802* (cit. on pp. 102, 106, 107, 119, 123).

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(4), 640–651 (cit. on pp. 20, 45).

Louie, A. V., Rodrigues, G., Olsthoorn, J., Palma, D., Yu, E., Yaremko, B., Ahmad, B., Aivas, I., & Gaede, S. (2010). Inter-observer and intra-observer reliability for lung cancer target volume delineation in the 4D-CT era. *Radiotherapy and Oncology*, *95*(2), 166–171 (cit. on p. 80).

Lukić, L., Polajnar, D., & Šolaja, V. (1991). A Yugoslav Approach to Decision Support for Optimization of FMS Technologies. *CIRP annals*, *40*(1), 99–102 (cit. on p. 71).

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*, 2008–2013 (cit. on pp. 50, 52, 93).

Maedche, A., Gregor, S., Morana, S., & Feine, J. (2019). Conceptualization of the problem space in design science research. In B. Tulu, S. Djamasbi, & G. Leroy. (Eds.), *Extending the boundaries of design science theory and practice* (pp. 18–31). (Cit. on p. 16).

Maedche, A., Legner, C., Benlian, A., Berger, B., Gimpel, H., Hess, T., Hinz, O., Morana, S., & Söllner, M. (2019). AI-based digital assistants. *Business & Information Systems Engineering*, *61*(4), 535–544 (cit. on pp. 143, 146).

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision support systems*, *15*(4), 251–266 (cit. on pp. 32, 71, 72, 102).

Marcos, M. P., Pitarch, J. L., & de Prada, C. (2020). Decision support system for a heat-recovery section with equipment degradation. *Decision Support Systems*, *137*, 113380 (cit. on p. 26).

Martin, D., & Kühl, N. (2019). Holistic system-analytics as an alternative to isolated sensor technology: A condition monitoring use case. *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 1005–1012 (cit. on p. 175).

Martin, D., Kühl, N., & Satzger, G. (2021). Virtual sensors. *Business & Information Systems Engineering*, *63*(3), 315–323 (cit. on p. 193).

Matikainen, L., Lehtomäki, M., Ahokas, E., Hyyppä, J., Karjalainen, M., Jaakkola, A., Kukko, A., & Heinonen, T. (2016). Remote sensing methods for power line corridor surveys. *ISPRS Journal of Photogrammetry and Remote Sensing*, *119*, 10–31 (cit. on pp. 106, 115, 121, 138).

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta*, *405*, 442–451 (cit. on pp. 21, 41).

Mayring, P. (1991). Qualitative inhaltsanalyse. In U. Flick, E. v. Kardoff, H. Keupp, L. v. Rosenstiel, & S. Wolff (Eds.), *Handbuch qualitative Forschung : Grundlagen, Konzepte, Methoden und Anwendungen* (pp. 209–213). (Cit. on p. 114).

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. C., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., … Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, *577*(7788), 89–94 (cit. on p. 143).

Mehranian, A., Ay, M. R., Alam, N. R., & Zaidi, H. (2010). Quantifying the effect of anode surface roughness on diagnostic x-ray spectra using Monte Carlo simulation. *Medical Physics*, *37*(2), 742–752 (cit. on p. 178).

Melo, F. (2013). Area under the ROC Curve. In W. Dubitzky, O. Wolkenhauer, K.-H. Cho, & H. Yokota (Eds.), *Encyclopedia of systems biology* (pp. 38–39). Springer New York. (Cit. on p. 22).

Meth, H., Mueller, B., & Maedche, A. (2015). Designing a Requirement Mining System. *Journal of the Association for Information Systems*, *16*(9), 799–837 (cit. on pp. 16, 121, 160).

Mettler, T. (2011). Maturity assessment models: A design science research approach. *International Journal of Society Systems Science (IJSSS)*, *3*(1/2), 81–98 (cit. on p. 16).

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 (cit. on p. 71).

Miguéis, V. L., Borges, J. L., et al. (2022). Automatic root cause analysis in manufacturing: An overview & conceptualization. *Journal of Intelligent Manufacturing*, 1–18 (cit. on p. 167).

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38 (cit. on pp. 120, 158).

Mirallès, F., Pouliot, N., & Montambault, S. (2014). State-of-the-art review of computer vision for the management of power transmission lines. *Proceedings of the 2014 3rd International Conference on Applied Robotics for the Power Industry*, 1–6 (cit. on pp. 106, 119, 138).

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 (cit. on p. 80).

Mohmmed, A., Li, Z., Arowolo, A. O., Su, H., Deng, X., Najmuddin, O., & Zhang, Y. (2019). Driving factors of CO2 emissions and nexus with economic growth, development and human health in the Top Ten emitting countries. *Resources, Conservation and Recycling*, *148*, 157–169 (cit. on p. 167).

Möller, F., Guggenberger, T. M., & Otto, B. (2020). Towards a Method for Design Principle Development in Information Systems. *International conference on design science research in information systems and technology*, 208–220 (cit. on pp. 12, 17, 144, 148).

Montambault, S., Beaudry, J., Toussaint, K., & Pouliot, N. (2010). On the application of VTOL UAVs to the inspection of power utility assets. *2010 1st International Conference on Applied Robotics for the Power Industry*, 1–7 (cit. on p. 124).

Morgan, D. L. (1996). *Focus groups as qualitative research* (Vol. 16). Sage publications. (Cit. on p. 148).

Müller, O., Junglas, I., Brocke, J. v., & Debortoli, S. (2016). Utilizing big data analytics for information systems research: Challenges, promises and guidelines. *European Journal of Information Systems*, *25*(4), 289–302 (cit. on p. 86).

Nabhani, F. (2001). Machining of aerospace titanium alloys. *Robotics and Computer-Integrated Manufacturing*, *17*(1-2), 99–106 (cit. on pp. 71, 82).

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 (cit. on p. 89).

Nærland, K., Müller-Bloch, C., Beck, R., & Palmund, S. (2017). Blockchain to rule the waves - nascent design principles for reducing risk and uncertainty in decentralized environments. *International Conference on Information Systems 2017, ICIS 2017* (cit. on p. 16).

Nair, T., Precup, D., Arnold, D. L., & Arbel, T. (2020). Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Medical image analysis*, *59*, 101557 (cit. on p. 53).

Nasir, M., Simsek, S., Cornelsen, E., Ragothaman, S., & Dag, A. (2021). Developing a decision support system to detect material weaknesses in internal control. *Decision Support Systems*, *151*, 113631 (cit. on p. 25).

Naumzik, C., & Feuerriegel, S. (2020). One Picture Is Worth a Thousand Words? The Pricing Power of Images in e-Commerce. *Proceedings of The Web Conference 2020*, 3119–3125 (cit. on p. 69).

Navarro-Hellín, H., Martinez-del-Rincon, J., Domingo-Miguel, R., Soto-Valles, F., & Torres-Sánchez, R. (2016). A decision support system for managing irrigation in agriculture. *Computers and Electronics in Agriculture*, *124*, 121–131 (cit. on p. 26).

Nazerdeylami, A., Majidi, B., & Movaghar, A. (2019). Smart coastline environment management using deep detection of manmade pollution and hazards. *2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI)*, 332–337 (cit. on p. 108).

Nedelea, A. (2021). Mercedes is first to sell a level 3 autonomous vehicle in 2022. Retrieved August 10, 2022, from https://insideevs.com/news/553659/mercedes-level3-autonomous-driving-2022/. (Cit. on p. 4)

Nguyen, V. N., Jenssen, R., & Roverso, D. (2018). Automatic autonomous vision-based power line inspection: A review of current status and the potential role of deep learning. *International Journal of Electrical Power and Energy Systems*, *99*, 107–120 (cit. on pp. 102, 106, 107, 115, 119, 121, 122, 138).

Nguyen, V. N., Jenssen, R., & Roverso, D. (2019). Intelligent Monitoring and Inspection of Power Line Components Powered by UAVs and Deep Learning. *IEEE Power and Energy Technology Systems Journal*, *6*(1), 11–21 (cit. on pp. 107, 109, 123).

Niero, M., Pizzol, M., Bruun, H. G., & Thomsen, M. (2014). Comparative life cycle assessment of wastewater treatment in denmark including sensitivity and uncertainty analysis. *Journal of Cleaner Production*, *68*, 25–35 (cit. on p. 174).

Offermann, P., Blom, S., Levina, O., & Bub, U. (2010). Proposal for components of method design theories. *Business & Information Systems Engineering*, *2*(5), 295–304 (cit. on p. 98).

Oh, C., Lee, T., Kim, Y., Park, S., Kwon, S., & Suh, B. (2017). Us vs. them: Understanding artificial intelligence technophobia over the google deepmind challenge match. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2523–2534 (cit. on p. 158).

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 (cit. on pp. 17, 36, 50, 80, 97, 170).

On-Road Automated Driving (ORAD) committee. (2021). *Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles* (tech. rep.). Society of Automative Engineers. (Cit. on p. 4).

Oppelt, A. (2005). *Imaging systems for medical diagnostics: Fundamentals, technical solutions and applications for systems applying ionizing radiation, nuclear magnetic resonance and ultrasound* (2nd ed.). Publicis Publishing. (Cit. on pp. 178, 186).

Oquab, M., Bottou, L., Laptev, I., & Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1717–1724 (cit. on p. 41).

Ostendorp, M. (2000). Innovative airborne inventory and inspection technology for electric power line condition assessments and defect reporting. *2000 IEEE 9th International Conference on Transmission and Distribution Construction, Operation and Live-Line Maintenance Proceedings*, 123–128 (cit. on p. 138).

Padilla, R., Netto, S. L., & da Silva, E. A. B. (2020). Survey on performance metrics for object-detection algorithms. *2020 international conference on systems, signals and image processing (IWSSIP)*, 237–242 (cit. on pp. 22, 126).

Pagnano, A., Höpf, M., & Teti, R. (2013). A roadmap for automated power line inspection. maintenance and repair. *Procedia CIRP*, *12*, 234–239 (cit. on pp. 101, 115, 138).

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 (cit. on p. 75).

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359 (cit. on p. 80).

Park, C.-W., Kwon, K.-S., Kim, W.-B., Min, B.-K., Park, S.-J., Sung, I.-H., Yoon, Y. S., Lee, K.-S., Lee, J.-H., & Seok, J. (2009). Energy consumption reduction technology in manufacturing—a selective review of policies, standards, and research. *International journal of precision engineering and manufacturing*, *10*(5), 151–173 (cit. on p. 167).

Patel, N., & Sethi, I. (2007). Multimedia data mining: An overview. In V. A. Petrushin & L. Khan (Eds.), *Multimedia Data Mining and Knowledge Discovery* (pp. 14–41). Springer. (Cit. on pp. 78, 79).

Pearn, W., & Lin, P. (2004). Testing process performance based on capability index cpk with critical values. *Computers & Industrial Engineering*, *47*(4), 351–369 (cit. on p. 89).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830 (cit. on pp. 42, 126).

Peffers, K., Rothenberger, M., Tuunanen, T., & Vaezi, R. (2012). Design science research evaluation. *International Conference on Design Science Research in Information Systems*, 398–410 (cit. on pp. 33, 72, 73).

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems*, *24*(3), 45–77 (cit. on pp. 16, 71).

Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621* (cit. on p. 45).

Pernebayeva, D., & James, A. P. (2020). Deep-learning-based approach for outdoor electrical insulator inspection. In A. P. James (Ed.), *Deep learning classifiers with memristive networks: Theory and applications* (pp. 81–88). (Cit. on pp. 115, 138).

Pflug, G. C. (2000). Some remarks on the value-at-risk and the conditional value-at-risk. In S. P. Uryasev (Ed.), *Probabilistic constrained optimization* (pp. 272–281). Springer. (Cit. on p. 92).

Pörtner, H.-O., Roberts, D. C., Tignor, M. M. B., Poloczanska, E., Mintenbeck, K., Alegría, A., Craig, M., Langsdorf, S., Löschke, S., Möller, V., Okem, A., & Rama, B. (2022). IPCC, 2022: Summary for Policymaker [In Press]. *Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (cit. on pp. 12, 203).

Power, D. J. (2002). *Decision support systems: Concepts and resources for managers*. Greenwood Publishing Group. (Cit. on pp. 4, 6, 24).

Power, D. J. (2007). A brief history of decision support systems. DSSResources.com. (Cit. on p. 24)

Power, D. J. (2008). Understanding data-driven decision support systems. *Information Systems Management*, *25*(2), 149–154 (cit. on pp. 5, 24, 81).

Powers, D. (2007). *Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation* (Technical Report SIE-07-001). School of Informatics and Engineering, Flinders University. Adelaide, Australia. (Cit. on p. 41).

Prasad, P. S., & Rao, B. P. (2016). Review on machine vision based insulator inspection systems for power distribution system. *Journal of Engineering Science & Technology Review*, *9*(5) (cit. on pp. 115, 138).

Prat, N., Comyn-Wattiau, I., & Akoka, J. (2015). A taxonomy of evaluation methods for information systems artifacts. *Journal of Management Information Systems*, *32*(3), 229–267 (cit. on p. 72).

Prates, R. M., Cruz, R., Marotta, A. P., Ramos, R. P., Simas Filho, E. F., & Cardoso, J. S. (2019). Insulator visual non-conformity detection in overhead power distribution lines using deep learning. *Computers & Electrical Engineering*, *78*, 343–355 (cit. on p. 119).

Quintana, M., Torres, J., & Menéndez, J. M. (2016). A simplified computer vision system for road surface inspection and maintenance. *IEEE Transactions on Intelligent Transportation Systems*, *17*(3), 608–619 (cit. on p. 105).

Rahman, M. A., & Wang, Y. (2016). Optimizing intersection-over-union in deep neural networks for image segmentation. *International symposium on visual computing*, 234–244 (cit. on p. 45).

Ramaswamy, V. (2008). Co-creating value through customers' experiences: The Nike case. *Strategy and Leadership*, *36*(5), 9–14 (cit. on p. 38).

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788 (cit. on p. 20).

Ren, M., Vu, H. Q., Li, G., & Law, R. (2021). Large-scale comparative analyses of hotel photo content posted by managers and customers to review platforms based on deep learning: Implications for hospitality marketers. *Journal of Hospitality Marketing & Management*, *30*(1), 96–119 (cit. on p. 108).

Ren, S., Zhang, Y., Liu, Y., Sakao, T., Huisingh, D., & Almeida, C. M. (2019). A comprehensive review of big data analytics throughout product lifecycle to support sustainable smart manufacturing: A framework, challenges and future research directions. *Journal of Cleaner Production*, *210*, 1343–1365 (cit. on p. 167).

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. (Cit. on pp. 20, 105, 124).

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*, 234–241 (cit. on pp. 19, 20, 37, 45, 51, 54, 83, 84, 171).

Rose, C., Beiter, K., & Ishii, K. (1999). Determining end-of-life strategies as a part of product definition. *Proceedings of the 1999 IEEE International Symposium on Electronics and the Environment*, 219–224 (cit. on p. 31).

Roy, A. G., Conjeti, S., Navab, N., & Wachinger, C. (2018). Inherent Brain Segmentation Quality Control from Fully ConvNet Monte Carlo Sampling. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 664–672 (cit. on pp. 52, 53, 57).

Rufino, J., Alam, M., Ferreira, J., Rehman, A., & Tsang, K. F. (2017). Orchestration of containerized microservices for IIoT using Docker. *2017 IEEE International Conference on Industrial Technology (ICIT)*, 1532–1536 (cit. on p. 135).

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252 (cit. on p. 125).

Russell, S. (2016). Rationality and intelligence: A brief update. In V. C. Müller (Ed.), *Fundamental issues of artificial intelligence* (pp. 7–28). Springer. (Cit. on p. 146).

Saarijärvi, H., Grönroos, C., & Kuusela, H. (2014). Reverse use of customer data: Implications for service-based business models. *Journal of Services Marketing*, *28*(7), 529–537 (cit. on p. 38).

Sachan, S., Yang, J.-B., Xu, D.-L., Benavides, D. E., & Li, Y. (2020). An explainable AI decision-support-system to automate loan underwriting. *Expert Systems with Applications*, *144*, 113100 (cit. on p. 25).

Sager, C., Janiesch, C., & Zschech, P. (2021). A Survey of Image Labeling for Machine Learning Applications. *Journal of Business Analytics*, *4*(2), 91–110 (cit. on p. 145).

Saidani, M., Pan, Z., Kim, H., Wattonville, J., Greenlee, A., Shannon, T., Yannou, B., Leroy, Y., & Cluzel, F. (2021). Comparative life cycle assessment and costing of an autonomous lawn mowing system with human-operated alternatives: Implication for sustainable design improvements. *International Journal of Sustainable Engineering*, *14*(4), 704–724 (cit. on p. 174).

Sakao, T., & Neramballi, A. (2020). A product/service system design schema: Application to big data analytics. *Sustainability*, *12*(8), 3484 (cit. on p. 170).

Sammut, C., & Webb, G. I. (Eds.). (2017). Accuracy. In *Encyclopedia of machine learning and data mining* (p. 8). Springer US. (Cit. on p. 21).

Sampedro Pérez, C., Rodriguez-Vazquez, J., Rodríguez Ramos, A., Carrio, A., & Campoy, P. (2019). Deep learning-based system for automatic recognition and diagnosis of electrical insulator strings. *IEEE Access*, *7*, 101283–101308 (cit. on pp. 106, 119).

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, *3*(3), 210–229 (cit. on pp. 3, 17, 170).

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4510–4520 (cit. on p. 20).

Saunders, M., Lewis, P., & Thornhill, A. (2009). *Research methods for business students* (5th ed.). Pearson education. (Cit. on p. 151).

Schaper, L., & Pervan, G. (2005). Exploring the links between technology acceptance and use and the attainment of individual and organisational goals: a case study in the community health sector. *Proceedings of the 11th Americas Conference on Information Systems*, 1145–1153 (cit. on p. 144).

Schätter, F., Hansen, O., Wiens, M., & Schultmann, F. (2019). A decision support methodology for a disaster-caused business continuity management. *Decision Support Systems*, *118*, 10–20 (cit. on p. 26).

Schemmer, M. (2020). *Design knowledge for image mining-based decision support systems: A case study in automated tool wear analysis* (Master's thesis). Karlsruhe University of Technology. Germany. (Cit. on pp. 25, 243).

Schmitt, R., Cai, Y., & Pavim, A. (2012). Machine Vision System for Inspecting Flank Wear on Cutting Tools. *ACEEE International Journal on Control System and Instrumentation*, *03*(01), 27–31 (cit. on pp. 36, 37).

Schumann, A., Mood, N., Mungofa, P., MacEachern, C., Zaman, Q., & Esau, T. (2019). Detection of three fruit maturity stages in wild blueberry fields using deep learning artificial neural networks. *2019 American Society of Agricultural and Biological Engineers Annual International Meeting* (cit. on p. 108).

Schwartz, T., Zinnikus, I., Krieger, H.-U., Bürckert, C., Folz, J., Kiefer, B., Hevesi, P., Lüth, C., Pirkl, G., Spieldenner, T., Schmitz, N., Wirkus, M., & Straube, S. (2016). Hybrid Teams: Flexible Collaboration Between Humans, Robots and Virtual Agents. In M. Klusch, R. Unland, O. Shehory, A. Pokahr, & S. Ahrndt (Eds.), *Multiagent system technologies* (pp. 131–146). Springer International Publishing. (Cit. on p. 146).

Schwarz, M., & Drudi, D. (2018). Workplace hazards facing line installers and repairers. *Monthly Labor Review*, 1–13 (cit. on pp. 134, 206).

Seeber, I., Bittner, E., Briggs, R. O., de Vreede, T., de Vreede, G.-J., Elkins, A., Maier, R., Merz, A., Oeste-Reiß, S., Randrup, N., Schwabe, G., & Söllner, M. (2019). Machines as Teammates: A Research Agenda on AI in Team Collaboration. *Information & Management*, *57*(2), 103174 (cit. on p. 146).

Selcuk, S. (2017). Predictive maintenance, its implementation and latest trends. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, *231*(9), 1670–1679 (cit. on p. 107).

Sen, A. (2004). Metadata management: Past, present and future. *Decision Support Systems*, *37*(1), 151–173 (cit. on p. 120).

Seok, K.-H., & Kim, Y. S. (2016). A state of the art of power transmission line maintenance robots. *Journal of Electrical Engineering & Technology*, *11*, 1412–1422 (cit. on pp. 115, 138).

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & LeCun, Y. (2013). Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229* (cit. on p. 20).

Setiawan, A. W. (2020). Image Segmentation Metrics in Skin Lesion: Accuracy, Sensitivity, Specificity, Dice Coefficient, Jaccard Index, and Matthews Correlation Coefficient. *2020 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM)*, 97–102 (cit. on pp. 23, 173).

Settles, B. (2009). *Active Learning Literature Survey* (tech. rep.). Retrieved November 21, 2022, from https://minds.wisconsin.edu/bitstream/handle/1793/60660/TR1648.pdf?sequence=1. (Cit. on p. 6)

Shakhatreh, H., Sawalmeh, A. H., Al-Fuqaha, A., Dou, Z., Almaita, E., Khalil, I., Othman, N. S., Khreishah, A., & Guizani, M. (2019). Unmanned Aerial Vehicles (UAVs): A Survey on Civil Applications and Key Research Challenges. *IEEE Access*, *7*, 48572–48634 (cit. on p. 129).

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379–423 (cit. on p. 57).

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 (cit. on p. 87).

Shihavuddin, A. S., Chen, X., Fedorov, V., Christensen, A. N., Riis, N. A. B., Branner, K., Dahl, A. B., & Paulsen, R. R. (2019). Wind turbine surface damage detection by deep learning aided drone inspection analysis. *Energies*, *12*(4), 1–15 (cit. on pp. 106, 109).

Shin, H.-C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., & Summers, R. M. (2016). Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE transactions on medical imaging*, *35*(5), 1285–98 (cit. on p. 41).

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. *Proceedings 1996 IEEE symposium on visual languages*, 336–343 (cit. on pp. 89, 130).

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 (cit. on pp. 34, 49, 82, 177).

Siering, M., Muntermann, J., & Grčar, M. (2021). Design principles for robust fraud detection: The case of stock market manipulations. *Journal of the Association for Information Systems*, *22*(1), 156–178 (cit. on p. 16).

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the Game of Go without Human Knowledge. *Nature*, *550*(7676), 354–359 (cit. on p. 167).

Silver, M. S. (1991). Decisional guidance for computer-based decision support. *MIS Quarterly*, *15*(1), 105–122 (cit. on pp. 86, 87, 98).

Silver, M. S. (2006). Decisional guidance: Broadening the scope. In P. Zhang & D. Galletta (Eds.), *Human-Computer Interaction and Management Information Systems: Foundations* (pp. 104–133). (Cit. on p. 87).

Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556* (cit. on pp. 18, 19, 42, 105, 125, 243).

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*, *46*(13), 2656–2672 (cit. on p. 70).

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 (cit. on p. 69).

Spencer, B. F., Hoskere, V., & Narazaki, Y. (2019). Advances in computer vision-based civil infrastructure inspection and monitoring. *Engineering*, *5*(2), 199–222 (cit. on p. 121).

Su, J. C., Huang, C. K., & Tarng, Y. S. (2006). An automated flank wear measurement of microdrills using machine vision. *Journal of Materials Processing Technology*, *180*(1-3), 328–335 (cit. on p. 36).

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*, 843–852 (cit. on pp. 86, 188).

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 (cit. on p. 70).

Szeliski, R. (2010). *Computer vision: Algorithms and Applications*. Springer Science & Business Media. (Cit. on pp. 3, 17, 105, 143, 145, 170).

Tai, B. L., Stephenson, D. A., Furness, R. J., & Shih, A. J. (2014). Minimum quantity lubrication (MQL) in automotive powertrain machining. *Procedia CIRP*, *14*, 523–528 (cit. on p. 175).

Takaya, K., Ohta, H., Kroumov, V., Shibayama, K., & Nakamura, M. (2019). Development of UAV System for Autonomous Power Line Inspection. *International Conference on System Theory, Control and Computing*, 762–767 (cit. on pp. 115, 138).

Tan, M., & Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *International conference on machine learning*, 6105–6114 (cit. on p. 20).

Terveen, L. G. (1995). Overview of human-computer collaboration. *Knowledge-Based Systems*, *8*(2-3), 67–81 (cit. on p. 146).

The Computing Research and Education Association of Australasia, CORE Inc. (2022). Core conference portal. Retrieved August 17, 2022, from http://portal.core.edu.au/conf-ranks/. (Cit. on p. 13)

Thiebes, S., Lins, S., & Sunyaev, A. (2020). Trustworthy artificial intelligence. *Electronic Markets*, *31*(2), 447–464 (cit. on pp. 144, 154).

Thomas, D. R. (2006). A general inductive approach for analyzing qualitative evaluation data. *American Journal of Evaluation*, *27*(2), 237–246 (cit. on p. 76).

Thomas, G., Gade, R., Moeslund, T. B., Carr, P., & Hilton, A. (2017). Computer vision for sports: Current applications and research topics. *Computer Vision and Image Understanding*, *159*, 3–18 (cit. on p. 144).

Tian, H., Wang, T., Liu, Y., Qiao, X., & Li, Y. (2020). Computer vision technology in agricultural automation—A review. *Information Processing in Agriculture*, *7*(1), 1–19 (cit. on pp. 4, 144).

Ting, K. M. (2017a). Confusion matrix. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of machine learning and data mining* (p. 260). Springer US. (Cit. on p. 21).

Ting, K. M. (2017b). Precision and recall. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of machine learning and data mining* (pp. 990–991). Springer US. (Cit. on p. 22).

Toth, J., & Gilpin-Jackson, A. (2010). Smart view for a smart grid — unmanned aerial vehicles for transmission lines. *2010 1st International Conference on Applied Robotics for the Power Industry*, 1–6 (cit. on pp. 115, 124).

Toussaint, K., Pouliot, N., & Montambault, S. (2009). Transmission line maintenance robots capable of crossing obstacles: State-of-the-art review and challenges ahead. *Journal of Field Robotics*, *26*(5), 477–499 (cit. on pp. 115, 138).

Treiss, A., Walk, J., & Kühl, N. (2020). An Uncertainty-based Human-in-the-loop System for Industrial Tool Wear Analysis. *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD 2020), Part V*, 85–100 (cit. on pp. 153, 159, 177, 244).

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 (cit. on pp. 16, 73, 90).

Tremblay, M. C., Hevner, A. R., & Berndt, D. J. (2010b). The use of focus groups in design science research. In A. Hevner & S. Chatterjee (Eds.), *Design Research in Information Systems* (pp. 121–143). Springer. (Cit. on p. 90).

Trinks, S., & Felden, C. (2019). Image mining for real time fault detection within the smart factory. *2019 IEEE 21st Conference on Business Informatics (CBI)*, *1*, 584–593 (cit. on p. 70).

Tukker, A. (2004). Eight types of product–service system: eight ways to sustainability? Experiences from SusProNet. *Business strategy and the environment*, *13*(4), 246–260 (cit. on p. 168).

Tukker, A. (2015). Product services for a resource-efficient and circular economy - A review. *Journal of Cleaner Production*, *97*, 76–91 (cit. on pp. 168, 176).

Turban, E., Sharda, R. E., & Delen, D. (2010). *Decision support and business intelligence systems*. Pearson Prentice Hall. (Cit. on pp. 10, 24, 70, 111, 121).

United Nations. (2015). Transforming Our World: The 2030 Agenda for Sustainable Development. Retrieved June 16, 2021, from https://sustainabledevelopment.un.org/post2015/transformingourworld/publication. (Cit. on p. 167)

Vargo, S. L., Maglio, P. P., & Akaka, M. A. (2008). On value and value co-creation: A service systems and service logic perspective. *European Management Journal*, *26*(3), 145–152 (cit. on p. 37).

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 (cit. on pp. 16, 72–74, 111, 243).

Venkatesh, V., & Bala, H. (2008). Technology acceptance model 3 and a research agenda on interventions. *Decision Sciences*, *39*(2), 273–315 (cit. on pp. 73, 95).

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science*, *46*(2), 186–204 (cit. on p. 95).

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, (3), 425–478 (cit. on p. 95).

Venkatesh, V., Thong, J. Y. L., & Xu, X. (2012). Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. *MIS Quarterly, Vol. 36*, *36*(1), 157–178 (cit. on pp. 150, 156).

Verband der Hochschullehrerinnen und Hochschullehrer für Betriebswirtschaft e.V. (2022). Liste der fachzeitschriften in vhb-jourqual3. Retrieved August 17, 2022, from https://vhbonline.org/vhb4you/vhb-jourqual/vhb-jourqual-3/gesamtliste. (Cit. on p. 13)

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 (cit. on p. 70).

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., & Fuso Nerini, F. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications*, *11*(1), 1–10 (cit. on p. 167).

Vladeck, D. C. (2014). Machines without principals: liability rules and artificial intelligence. *Washington Law Review*, *89*, 117 (cit. on p. 158).

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)*, *9*, 2206–2217 (cit. on pp. 81, 113).

vom Brocke, J., Hevner, A., & Maedche, A. (2020). Introduction to design science research. In *Design science research. cases* (pp. 1–13). Springer. (Cit. on pp. 15, 16).

Voulodimos, A., Doulamis, N., Doulamis, A., & Protopapadakis, E. (2018). Deep Learning for Computer Vision: A Brief Review. *Computational Intelligence and Neuroscience*, *2018*, 1–13 (cit. on p. 37).

Walk, J., Kühl, N., & Schäfer, J. (2020). Towards Leveraging End-of-Life Tools as an Asset: Value Co-Creation based on Deep Learning in the Machining Industry. *Proceedings of the 53rd Hawaii International Conference on System Sciences*, *3*, 995–1004 (cit. on pp. 50, 153).

Walls, J. G., Widmeyer, G. R., & El Sawy, O. A. (1992). Building an Information System Design Theory for Vigilant EIS. *Information Systems Research*, *3*(1) (cit. on p. 116).

Wang, D., Weisz, J. D., Muller, M., Ram, P., Geyer, W., Dugan, C., Tausczik, Y., Samulowitz, H., & Gray, A. (2019). Human-AI Collaboration in Data Science: Exploring Data Scientists' Perceptions of Automated AI. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–24 (cit. on p. 147).

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 (cit. on pp. 4, 69, 167).

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 (cit. on p. 78).

Wang, P., Kara, S., & Hauschild, M. Z. (2018). Role of manufacturing towards achieving circular economy: The steel case. *CIRP Annals*, *67*(1), 21–24 (cit. on p. 168).

Wang, T., Chen, Y., Qiao, M., & Snoussi, H. (2018). A fast and robust convolutional neural network-based defect detection model in product quality control. *The International Journal of Advanced Manufacturing Technology*, *94*(9), 3465–3471 (cit. on p. 144).

Wanner, J., Herm, L.-V., Heinrich, K., Janiesch, C., & Zschech, P. (2020). White, Grey, Black: Effects of XAI Augmentation on the Confidence in AI-based Decision Support Systems. *Proceedings of the 41st International Conference on Information Systems* (cit. on p. 158).

Watson, H. J. (2014). Tutorial: Big data analytics: Concepts, technologies, and applications. *Communications of the Association for Information Systems*, *34*(1), 65 (cit. on p. 25).

Webster, J., & Watson, R. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly*, *26*(2), xiii–xxiii (cit. on p. 113).

Wei, X., Yang, Z., Liu, Y., Wei, D., Jia, L., & Li, Y. (2019). Railway track fastener defect detection based on image processing and deep learning techniques: A comparative study. *Engineering Applications of Artificial Intelligence*, *80*, 66–81 (cit. on pp. 105, 109).

Wernet, G., Bauer, C., Steubing, B., Reinhard, J., Moreno-Ruiz, E., & Weidema, B. (2016). The ecoinvent database version 3 (part I): overview and methodology. *The International Journal of Life Cycle Assessment*, *21*(9), 1218–1230 (cit. on p. 174).

Whiting, L. S. (2008). Semi-structured interviews: guidance for novice researchers. *Nursing Standard*, *22*(23), 35–41 (cit. on p. 75).

Wikström, P., Edenius, L., Elfving, B., Eriksson, L. O., Lämås, T., Sonesson, J., Öhman, K., Wallerman, J., Waller, C., & Klintebäck, F. (2011). The heureka forestry decision support system: An overview. *Mathematical and Computational Forestry & Natural-Resource Sciences (MCFNS)*, *3*(2), 87–94 (cit. on p. 26).

Wilson, H. J., & Daugherty, P. R. (2018). Collaborative Intelligence: Humans and AI Are Joining Forces. *Harvard Business Review*, *96*(4), 114–123 (cit. on p. 98).

Winter, R. (2008). Design science research in Europe. *European Journal of Information Systems*, *17*(5), 470–475 (cit. on pp. 71, 102).

Wu, G., Yang, P., Xie, Y., Woodruff, H. C., Rao, X., Guiot, J., Frix, A.-N., Louis, R., Moutschen, M., Li, J., et al. (2020). Development of a clinical decision support system for severity risk prediction and triage of COVID-19 patients at hospital admission: an international multicentre study. *European Respiratory Journal*, *56*(2) (cit. on p. 26).

Xie, Y., Noguchi, R., & Ahamed, T. (2020). Deep learning and multiple sensors data acquisition system for real-time decision analysis in agriculture using unmanned aerial vehicle. *EasyChair Preprint*, (2690) (cit. on p. 108).

Yang, M., & Evans, S. (2019). Product-service system business model archetypes and sustainability. *Journal of Cleaner Production*, *220*, 1156–1166 (cit. on pp. 176, 194).

Yang, Y., Wang, L., Wang, Y., & Mei, X. (2019). Insulator self-shattering detection: A deep convolutional neural network approach. *Multimedia Tools and Applications*, *78*(8), 10097–10112 (cit. on p. 106).

Yin, R. K. (2017). *Case study research and applications: Design and methods* (Vol. 5). Sage publications. (Cit. on p. 151).

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 (cit. on pp. 27, 81).

Zeiss, R., Ixmeier, A., Recker, J., & Kranz, J. (2021). Mobilising information systems scholarship for a circular economy: Review, synthesis, and directions for future research. *Information Systems Journal*, *31*(1), 148–183 (cit. on p. 12).

Zhai, Y., Chen, R., Yang, Q., Li, X., & Zhao, Z. (2018). Insulator fault detection based on spatial morphological features of aerial images. *IEEE Access*, *6*, 35316–35326 (cit. on p. 106).

Zhang, Y., Ren, S., Liu, Y., Sakao, T., & Huisingh, D. (2017). A framework for big data driven product lifecycle management. *Journal of Cleaner Production*, *159*, 229–240 (cit. on pp. 167–169, 191).

Zheng, N.-n., Liu, Z.-y., Ren, P.-j., Ma, Y.-q., Chen, S.-t., Yu, S.-y., Xue, J.-r., Chen, B.-d., & Wang, F.-y. (2017). Hybrid-augmented intelligence: collaboration and cognition. *Frontiers of Information Technology & Electronic Engineering*, *18*(2), 153–179 (cit. on pp. 146, 147).

Zhu, X., Kong, L., Wang, G., Hu, Z., & Li, S. (2018). Multi-size object detection assisting fault diagnosis of power systems based on improved cascaded faster R-CNNs. In X. Jiang & J.-N. Hwang (Eds.), *10th International Conference on Digital Image Processing (ICDIP 2018)* (pp. 342–351). (Cit. on pp. 107, 109).

Zou, K. H., Warfield, S. K., Bharatha, A., Tempany, C. M., Kaus, M. R., Haker, S. J., Wells III, W. M., Jolesz, F. A., & Kikinis, R. (2004). Statistical Validation of Image Segmentation Quality Based on a Spatial Overlap Index. *Academic radiology*, *11*(2), 178–189 (cit. on p. 182).

Zschech, P., Sager, C., Siebers, P., & Pertermann, M. (2021). Mit Computer Vision zur automatisierten Qualitätssicherung in der industriellen Fertigung: Eine Fallstudie zur Klassifizierung von Fehlern in Solarzellen mittels Elektrolumineszenz-Bildern. *HMD Praxis der Wirtschaftsinformatik*, *58*(2), 321–342 (cit. on p. 153).

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* (cit. on p. 82).

List of Figures

List of Tables

Copyright for Figures

CNC Milling Machine by Toke Frello from Noun Project (https://thenounproject.com/icon/cnc-milling-machine-189666/) used on page 7 and page 8.

Tower by icon 54 from Noun Project (https://thenounproject.com/icon/tower-199891/) used on page 7 and page 8.