Editorial


Christian Müller

Liebe Leserinnen und Leser,

vor Ihnen liegt nunmehr die bereits sechzehnte Ausgabe des E-Journals **Anwendungen und Konzepte in der Wirtschaftsinformatik (AKWI)**.

Die Zeitschrift wurde vor einigen Monaten auf eine neue Hosting-Plattform umgezogen, welches aus unserer Sicht für uns alle eine Reihe von Vorteilen bringt: Einerseits konnte die Reichweite erhöht werden und es wird nun auch ein automatischer Abgleich mit weiteren Open Access Verzeichnissen ermöglicht. Für uns Herausgeber hat sich zudem die Pflege der Journalsoftware vereinfacht, da eine Reihe von manuellen Tätigkeiten, die i.d.R. vom Kollegen Marfurt organisiert wurden, nicht mehr notwendig sind und die Herausgeber sich nun noch stärker auf die Inhalte der Zeitschrift konzentrieren können.


Konrad Marfurt

Teil dieser Ausgabe sind wieder Zweitveröffentlichungen von Artikeln aus dem Bereich der Modellierung und Simulation, die von den Herausgebern als besonders interessant für die Leser dieses Journals empfunden werden. Diese Artikel liegen in englischer Sprache vor und stammen von der ECMS 2022. Die Artikel dieser Konferenz umfassen ein breites Spektrum an Themenbereichen: einerseits einen BPMN-basierten Simulationsansatz für die interne Logistikoptimierung, sowie einen Simulationsansatz, welcher mit Lithiumbatterien bestückte, fahrerlose Transportsysteme, mit welchen Spitzenlasten im Energiebedarf einer Fertigung abgedeckt werden sollen. Schließlich eine Untersuchung aus dem Bereich der Mikrobiologie – Modellierung und Simulation ist letztlich eine Methode, um Fragestellungen aus verschiedensten Domänen zu unterstützen.


Norbert Ketterer

Die originären Artikel aus dem Bereich der Wirtschaftsinformatik gliedern sich in gewohnter Weise wieder in eigentliche Zeitschriftenartikel sowie Kurzübersichten einiger Abschlussarbeiten.

Die Artikel stammen diesmal schwerpunktmäßig aus den Bereichen des Geschäftsprozess-managements, der KI mit CRM- und Produktions-Bezug, dem Management der Transformation von Anwendungssystemen, der Datenverarbeitung in der Cloud sowie der Software-Qualitätssicherung.

Im Detail behandelt ein Artikel die noch recht junge Disziplin des Process Minings zur Auswahl und Anwendung von Robotic Process Automation (RPA) Lösungen. Ein weiterer Artikel behandelt die Fragestellung der automatisierten Analyse von Kundenfeedbacks mit KI-Methoden. Hierbei wird die Analyse der Daten mit Hilfe von Grafikkarten beschleunigt. Eine weitere Arbeit behandelt Fragestellungen der KI basierten Mustererkennung, um mit Hilfe eines CNNs Fertigungsfehler im Rahmen eines Eloxierungsprozesses sehr hochwertiger Produkte erkennen zu können. Die Transformation von Anwendungssystemen wird in einer sehr praxisnahen Arbeit unterstützt, indem mittels In-Memory Technologie und Virtualisierung zumindest die wichtigsten Ergebnisse der Transformation bereits auf den bestehenden Datenmodellen in Echtzeit generiert werden, um genügend Zeit zu gewinnen, die eigentliche Transformation der IT-Landschaft durchführen zu können. Durch ein Cockpit in S/4 Hana werden die Ergebnisse einer späteren tatsächlichen Transformation der zugrundeliegenden Systeme in Teilaspekten vorweggenommen, um einen Blick auf die Zukunft zu ermöglichen. Mögliche Nachteile aufgrund der eingeschränkten Kontrolle einer serverlosen Datenverarbeitung in der Cloud werden in einem Beitrag ebenso behandelt, wie in einem anderen Beitrag Fragen aus dem Bereich der Software-Qualitätssicherung durch ein paralleles Mining von C# Code. In letzterem Beitrag wurden Arbeiten des maschinellen Lernens (ML) evaluiert, um aus bereits eingetretenen Fehlern automatisiert dazulernen zu können; Ziel ist die bessere Unterstützung von statischen Code-Analysen zur Software-Qualitätssicherung.


Frank Herrmann

Bei den Abschlussarbeiten behandeln zwei Arbeiten die für die WI immer relevanter werdenden Anwendungen von KI-Methoden, einerseits im Umfeld des automatisierten Handels am Finanzmarkt und andererseits ein Anwendungsbeispiel aus dem Bereich der Medizin. Weitere Abschlussarbeiten behandeln die Fragestellung der Organisation eines Process Minings über mehrere verteilte Systeme sowie die Erstellung einer Datenpipeline zur automatisieren Validierung der Qualität und Konsistenz von Bankdaten. Eine Arbeit mit einem stärkeren Hochschulbezug behandelt schließlich die Erstellung eines Chatbots zur Unterstützung der Programmierausbildung.

Über Ihr Interesse an der Zeitschrift freuen wir uns und wünschen Ihnen Freude bei der Lektüre.

Regensburg, Fulda, Luzern und Wildau, im Dezember 2022.

Frank Herrmann, Norbert Ketterer, Konrad Marfurt und Christian Müller

MICROBIAL GROWTH OF *LACTOBACILLUS DELBRUECKII* SSP. *BULGARICUS* B1 IN A COMPLEX NUTRIENT MEDIUM (MRS-BROTH)

Georgi Kostov*, Rositsa Denkova-Kostova**, Vesela Shopska*, Bogdan Goranov***, Zapryana Denkova***
*Department of Wine and Beer ** Department of Biochemistry and Molecular Biology, *** Department of Microbiology
University of Food Technologies, 4002, 26 Maritza Blvd., Plovdiv, Bulgaria
E-mail: george_kostov2@abv.bg; rositsa_denkova@mail.bg; vesi_nevelinova@abv.bg; dr.eng.bgoranov@gmail.com; zdenkova@abv.bg

KEYWORDS
Probiotics, growth kinetics, modeling, optimization, complex nutrient medium, MRS-broth.

ABSTRACT

The microbial growth of the probiotic strain *Lactobacillus delbrueckii* ssp. *bulgaricus* B1, cultivated in a complex nutrient medium (MRS-broth), was studied in the present work. The complex nutrient medium provides not only the carbon source necessary for the growth of biomass, but also all the additional sources of nitrogen, phosphorus and other components that the biomass needs for its growth. The use of non-structural mathematical dependences determines the optimal conditions (substrate concentration) for the accumulation of biomass or lactic acid, depending on the needs of the specific production.

INTRODUCTION

In recent years, there has been increased interest in the use of lactic acid bacteria of the species *Lactobacillus*, *Enterococcus*, *Pediococcus*, *Streptococcus*, *Lactococcus* and *Leuconostoc* for the development of probiotic and synbiotic preparations (Gibson, 2004). In order for a strain of these species to be classified as probiotic, it must meet a number of requirements, one of which is to allow the conduction of industrial cultivation (Saarela et al., 2002; Kostov et al., 2021).

To evaluate this property it is necessary to apply microbial kinetics, which can be used to assess parameters such as: specific growth rate (maximum and current value), specific rate of product accumulation (maximum and current value), different types of constants. saturation, inhibition, etc.). The combination of these parameters makes it possible to assess the growth of microorganism biomass (in particular lactic acid bacteria biomass), to determine the optimal growth conditions that can ensure the accumulation of biomass and/or metabolic products (Bouguettoucha et al., 2011).

In their classical work Baily and Ollis, 1986, proposed different types of models to describe the microbial growth kinetics. These dependencies are based on the S-shaped nature of microbial growth and are divided into four main groups: non-structural non-segregated models; non-structural segregated models; structural non-segregated models and structural segregated models (Bailey and Ollis, 1986). Nonstructural models view the growth of the microbial population as a whole. When applied, it is assumed that the microbial population grows in the conditions of unlimited food sources, unlimited space and lack of factors related to the vital activity of microorganisms. These models follow from the so-called equation of exponential growth (1) and the well-known Monod dependence (2):

$$\frac{dX}{d\tau} = \mu X \quad (1)$$

$$\mu = \mu_m \frac{S}{K_s + S} \quad (2)$$

where: μ_m - maximum specific growth rate, h^{-1}; X - biomass concentration, g/dm^3; S - substrate concentration, g/dm^3; KS - saturation constant, g/dm^3.

A number of non-structural models have been developed based on the Monod equation and they have been usually named after the researcher who proposed them. Such examples are the Tiessier model, the Andrews and Noack model, the Hinshelwood model, the Aiba model, the Ghose and Tyagi model and others, that try to solve various aspects of microbial growth (substrate inhibition; product inhibition; product and substrate inhibition, etc.) (Bailey and Ollis, 1986; Bouguettoucha et al., 2011; Kostov, 2015; Muloiwa et al, 2020). In the present paper we are going to consider other examples as well (see Materials and methods).

Non-structural models describe only the amount of biomass and/or the amount of metabolites accumulated. Thus, they do not reflect the qualitative characteristics of the cell population and the changes that occur in it during cultivation. These changes can only be described by structural models. They are based on the material balance equations, but in their construction it is necessary to select the key changes taking place in the population. In this model type one works with the concentration of the corresponding variable in a volume unit of biophase, taking into account the cell density, the rate of component formation, the cell mass and more. These models are usually quite complex and include a large number of variables that do not always have a clear and precise biological meaning (Bailey and Ollis, 1986; Kostov, 2015; Shopska et al., 2019).

One of the most well known species of lactic acid bacteria is *Lactobacillus delbrueckii* ssp. *bulgaricus*. Representatives of this species are included as starter cultures for the production of various types of food, as

well as for the production of probiotic preparations (Arena et al., 2015; Maisto et al., 2021; Ivanov etl al., 2021).

The ability to accumulate large amounts of biomass in the cultivation of lactic acid bacteria is very important for the production of probiotics. Complex nutrient media are usually used for the cultivation process. One of the most frequently used media for the cultivation of lactic acid media is MRS-broth medium (de Man, Rogosa and Sharpe). MRS-broth medium has been developed primarily for the cultivation of lactobacilli from various sources with the intention of producing a defined medium as a substitute for tomato juice agar. It is used for the cultivation of the whole group of lactic acid bacteria. The medium shows good productivity for nearly all lactic acid bacteria, but the original version is not selective. It was made selective for lactic acid bacteria by lowering the pH to 5.7 and the addition of 0.14% sorbic acid. Some strains from dairy sources show reduced growth rates in MRS. MRS agar is composed of tryptic digest of casein, beef extract, yeast extract, glucose, sorbitan monooleate, di-potassium hydrogen orthophosphate, magnesium sulfate, manganese (II) sulfate, ammonium citrate, sodium acetate, agar, and distilled or deionized water (Corry et al., 2003).

The main metabolite of lactic acid fermentation is lactic acid. It is known that its increasing concentration during fermentation has an inhibitory effect on the growth of the microbial population. The sensitivity to the accumulating lactic acid is strain-specific (Bouguettoucha et al., 2011; Gordeev et al., 2017).

The aim of the present work was to study the growth characteristics of the probiotic strain *Lactobacillus delbrueckii* ssp. *bulgaricus* when cultivated in a complex nutrient medium such as MRS-broth. The strain has demonstrated a number of probiotic characteristics and had been isolated from homemade yogurt (Goranov et al, 2015; Teneva et al., 2015). As already commented, in some cases, strains isolated from dairy products show reduced growth in MRS-broth medium. Six non-structural mathematical models (see Materials and methods) based on the Monod equation were used to model the microbial growth. The obtained data were used to determine the optimal concentrations of the complex food source in order to improve the accumulation of biomass or lactic acid.

MATERIALS AND METHODS

Microorganisms and cultivation conditions

The study was conducted with *Lactobacillus delbrueckii* ssp. *bulgaricus* B1 isolated from home made yougurt (Goranov et al., 2015; Teneva et al., 2015).

The strain was cultivated in MRS-broth, produced by Merck, with the following qualitative and quantitative composition (g/dm^3): peptone from casein - 10.0; meat extract - 8.0; yeast extract - 4.0; D(+)-glucose - 20.0; dipotassium hydrogen phosphate - 2.0; Tween® 80 - 1.0; di-ammonium hydrogen citrate - 2.0; sodium acetate - 5.0; magnesium sulfate - 0.2; manganese sulfate – 0.05.

The cultivation was performed in a bioreactor with mechanical stirring, shown in Fig.1.The apparatus has a geometric volume of 2 dm^3 and a working volume of 1.5 dm^3 and is equipped with a Sartorius A2 control device, which includes all the measuring instruments for the fermentation process: temperature, pH, dissolved oxygen, etc. The fermentation process was carried out at a stirring speed of 150 rpm at 37±1°C.

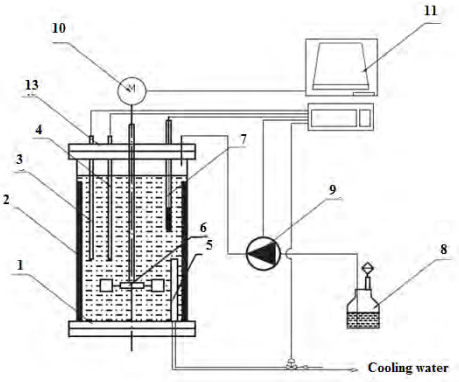


Figure 1: Laboratory Bioreactor
1 - vessel with a geometric volume of 2 dm^3; 2 - baffles; 3 - temperature electrode (thermometer); 4 - cooling/heating device (water jacket); 5 - an additional cooling/heating device; 6 - turbine stirrer; 7 - pH/Eh electrode; 8 - fermentation medium/inoculum/pH adjustment medium; 9 - peristaltic pump; 10 - stirrer drive; 11 – Sartorius A2 control device;

Methods of analysis and nutrient medium for analysis

- Determination of titratable acidity (ISO/TS 11869:2012);
- Determination of number of viable lactobacilli cells (ISO 7889:2005).
- Nutrient media (ISO 7889:2005)
  - MRS-broth;
  - MRS-agar;
  - Saline solution.

Modeling of microbial growth and identification of parameters in kinetic models

The following system of differential equations was used to model the kinetics of microbial growth:

$$\begin{cases} \dfrac{dX}{d\tau} = \mu(\tau) X(\tau) \\ \dfrac{dP}{d\tau} = q(\tau) X(\tau) \\ \dfrac{dS}{d\tau} = -\dfrac{1}{Y_{X/S}} \dfrac{dX}{\tau} - \dfrac{1}{Y_{P/S}} \dfrac{dP}{d\tau} \end{cases} \quad (3)$$

where: X – biomass concentration, g/dm^3; P – lactic acid concentration, g/dm^3; S – substrate concentration, g/dm^3; $Y_{P/S}$, $Y_{X/S}$ – yield coefficients; μ – specific growth rate, h^{-1}; q – specific lactic acid accumulation rate, g/(g.h).

The following dependences were used to model the biomass specific growth rate and the specific rate of lactic acid accumulation (Bailey and Ollis, 1986;

Bouguettoucha et al., 2011; Kostov, 2015; Muloiwa et al, 2020):

- Monod model

$$\mu = \mu_{\max} \frac{S}{K_{SX} + S}$$
$$q = q_{p\max} \frac{S}{K_{SP} + S} \quad (4)$$

- Haldane model

$$\mu = \mu_{\max} \frac{S}{K_{SX} + S + \frac{S^2}{K_{Xi}}}$$
$$q = q_{p\max} \frac{S}{K_{SP} + S + \frac{S^2}{K_{Pi}}} \quad (5)$$

- Aiba model

$$\mu = \mu_{\max} \frac{S}{K_{SX} + S} \exp(-K_{PX} P)$$
$$q = q_{p\max} \frac{S}{K_{SP} + S} \exp(-K_{PP} P) \quad (6)$$

- Haldane-Aiba model

$$\mu = \mu_{\max} \frac{S}{K_{SX} + S + \frac{S^2}{K_{Xi}}} \exp(-K_{PX} P)$$
$$q = q_{p\max} \frac{S}{K_{SP} + S + \frac{S^2}{K_{Pi}}} \exp(-K_{PP} P) \quad (7)$$

- Haldane model for product inhibition

$$\mu = \mu_{\max} \frac{S}{K_{SX} + S + \frac{S^2}{K_{Xi}}} \left(1 + \frac{P}{K_{PX}}\right)$$
$$q = q_{p\max} \frac{S}{K_{SP} + S + \frac{S^2}{K_{Pi}}} \left(1 + \frac{P}{K_{PP}}\right) \quad (8)$$

- Haldane-Jerusalimski model

$$\mu = \mu_{\max} \frac{S}{K_{SX} + S + \frac{S^2}{K_{Xi}}} \left(\frac{K_{PX}}{P + K_{PX}}\right)$$
$$q = q_{p\max} \frac{S}{K_{SP} + S + \frac{S^2}{K_{Pi}}} \left(\frac{K_{PP}}{P + K_{PP}}\right) \quad (9)$$

where: μ_{max} – maximum specific growth rate, h^{-1}; q_{pmax} – maximum specific rate of lactic acid formation, h^{-1}; K_{sx} и K_{sp} – Monod constants for saturation of biomass and product by substrate, g/dm^3; K_{xi} и K_{pi} – substrate inhibition constants for biomass and product, g/dm^3; K_{px} и K_{pp} – product inhibition constants for biomass and product, g/dm^3.

The parameters in the kinetic equations are calculated by solving the system of differential equations using the Runge-Kuta method of the 4th row, by minimizing the sum of the squares of the difference between the experimental and model data. The software used was Microsoft Excel 2013 (Choi et al., 2014).

RESULTS AND DISCUSSION

The dynamics of the studied fermentation process in the complex nutrient medium MRS-broth is presented in Fig. 2.

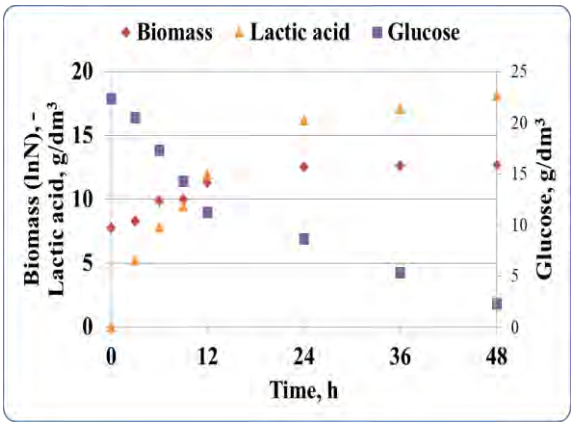


Figure 2: Dynamics of Lactic Acid Fermentation in Cultivation of *Lactobacillus delbrueckii* ssp. *bulgaricus* B1 in MRS-broth in Bioreactor with Stirring

The data show that the fermentation process developed according to the trends for this process type. The duration of the lag phase was about 3 hours, after which the culture entered the exponential growth phase. The exponential growth phase lasted about 24 hours, and at the end of this phase a high number of viable cells was reached - 12.57 logarithmic units. In the next 24 hours, the culture was in the stationary phase, and it retained the high number of viable cells. In this phase, the substrate continued to be utilized at a high rate and lactic acid was constantly accumulating. At the end of the process, the lactic acid concentration reached 18.09 g/dm^3 and the substrate concentration decreased to 2.3 g/dm^3. The unutilized substrate was due to the influence of the product and the substrate inhibition processes, which should be taken into account. This means that in order to optimize the process, opportunities for the complete utilization of the substrate should be sought, which is achieved by optimizing the concentration of the substrate. It is known that the process of cultivation of lactic acid bacteria is substrate and product dependent (inhibited) (Bouguettoucha et al., 2011; Kostov, 2015).
The first two models we are going to discuss at are the classic Monod model and the Haldane model. The data for the kinetic parameters and the errors of the models are shown in Table 1, and the convergence of the models to the experimental data is given in Fig. 3 and Fig. 4.

Table 1: Kinetic Constants in the Different Models Used to Describe the Microbial Growth Kinetics of *Lactobacillus delbrueckii* ssp. *bulgaricus* B1

	Monod	Haldane	Aiba	Haldane-Aiba	Haldane for product inhibition	Haldane-Jerusalimski
μ_{max}, h^{-1}	0.072	0.038	0.576	0.049	0.113	0.011
K_{SX}, g/dm^3	43.05	32.51	213.78	47.59	65.69	17.55
K_{Xi}, g/dm^3	-	85.45	-	156.67	163.23	283.82
q_{pmax}, h^{-1}	0.065	0.3982	0.266	0.063	0.256	0.065
K_{SP}, g/dm^3	11.13	132.36	29.71	0.005	29.12	41.39
K_{SPi}, g/dm^3	-	441.11		180.72	96.60	169.39
K_{PX}, g/dm^3	-	-	0.16	0.249	22.38	70.09
K_{PP}, g/dm^3	-	-	0.11	0.138	18.53	3.36
$1/Y_{x/s}$	0.3416	0.3055	0.6181	0.3318	0.3190	0.6680
$1/Y_{p/s}$	12	0.9820	2.2487	1.9694	0.0400	0.0400
$Y_{x/s}$	2.9277	3.2733	1.6176	3.0139	3.1348	1.4970
$Y_{p/s}$	0.0833	1.0183	0.4021	0.5078	25	25
R^2 (biomass)	0.7895	0.8473	0.8860	0.8712	0.9378	0.871
Error (biomass)	0.84	0.69	0.33	0.27	0.38	0.48
R^2 (product)	0.9507	0.9720	0.9763	0.983	0.9814	0.8895
Error (product)	2.63	2.42	1.42	1.38	1.42	1.80
R^2 (substrate)	0.9194	0.9361	0.9946	0.9907	0.9943	0.9697
Error (substrate)	2.75	2.25	1.27	1.23	1.27	1.12
S_{opt} (biomass), g/dm^3	-	52.71	-	86.35	103.55	70.58
S_{opt} (product)), g/dm^3	-	241.63	-	0.96	53.04	82.37

The data from Fig. 3 and Fig. 4, as well as those in Table 1, show that the Monod model and the Haldane model agree very well with the experimental data.

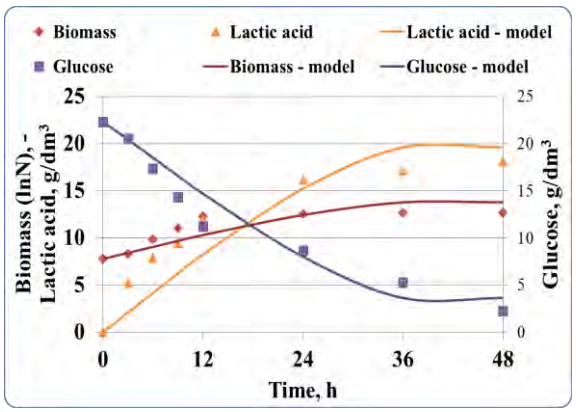


Figure 3: Kinetics of Lactic Acid Fermentation in Cultivation of *Lactobacillus delbrueckii* ssp. *bulgaricus* B1 in MRS-broth Described with the Monod Model

The correlation coefficients range from 0.7895 to 0.9720. The values of the calculated errors vary in the range of 0.69 to 2.25. The Monod model gives almost twice the maximum specific growth rate (0.072 h^{-1}) compared to the Haldane model - 0.038 h^{-1}, due to the fact that the Haldane model takes into account the substrate inhibition of the lactic acid fermentation process. In both models there is an increased saturation constant of the substrate - 43.05 g/dm^3 and 32.51 g/dm^3, respectively, (in this case the substrate is equated to the concentration of the carbon source - glucose 20 g/dm^3), which confirms the observation that glucose is the substrate limiting the fermentation process. The two models also give different values with respect to the maximum specific lactic acid accumulation rate. The fact that the Haldane model gives about 6 times higher rate of acid formation (0.3982 h^{-1}) than the Monod model (0.065 h^{-1}) is very interesting. This also leads to significant differences in the saturation constants by product - 11.13 g/dm^3 for the Monod model and 132.36 g/dm^3 for the Haldane model. This difference shows that according to the Haldane model the rate of acid formation depends to a greater extent on the concentration of the limiting substrate.

It is interesting to determine theoretically at what substrate concentrations the culture will undergo substrate inhibition. Information on this is given by the substrate inhibition constants for biomass and product in the Haldane model - K_{Xi} and K_{SPi}, respectively. From the data presented in Table 1 it can be seen that the inhibitory effect of the substrate on cell proliferation and growth will begin to be observed at K_{Xi} = 85.45 g/dm^3 and K_{SPi} = 441.11 g/dm^3, which is 8.545% and 44.11% substrate (glucose) in the nutrient medium, respectively. K_{Xi} is close to the experimental results of various authors who found that at concentrations of the substrate (glucose) in the nutrient medium higher than 10%, its inhibitory effect on the specific growth rate of lactic acid bacteria becomes noticeable.

However, the K_{SPi} calculated by the Haldane model has an abnormally high value, which deviates greatly from the K_{Xi}. In this case, the value of K_{SPi} is real from mathematical point of view, but not from biological point of view, because such high glucose concentration in the medium will not allow the growth and accumulation of high numbers of viable cells that actively produce lactic acid. It should be noted that the inhibition may be due not only to the carbon source, but also to some of the complex components in the nutrient medium. However, this is difficult to account for with

non-structural models that are usually used to describe the fermentation process.

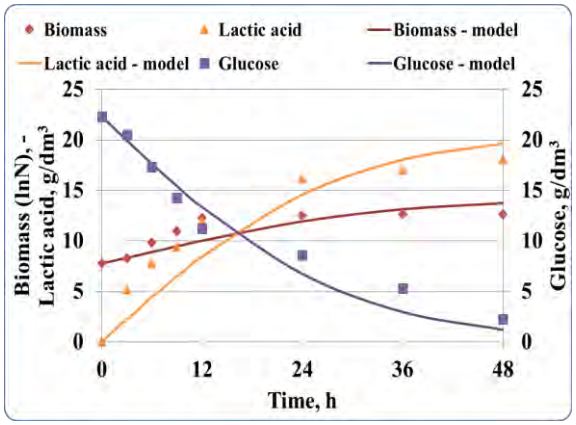


Figure 4: Kinetics of Lactic Acid Fermentation in Cultivation of *Lactobacillus delbrueckii* ssp. *bulgaricus* B1 in MRS-broth Described with the Haldane Model

It is also interesting to determine the metabolic (trophic) coefficients - $1/Y_{X/S}$ and $1/Y_{P/S}$, showing the consumption of substrate for biomass growth and for product synthesis. Since they are the reciprocal values of the respective economic coefficients, the latter can also be determined. From the data presented in Table 1 it is evident that both models show higher substrate consumption for biomass formation – the trophic coefficients being 0.3416 and 0.3055, respectively, while the economic coefficients being 2.9277 and 3.2733, respectively, and a smaller part of the substrate goes to lactic acid synthesis ($1/Y_{P/S}$ - 12 and 0.9820, respectively, and $Y_{P/S}$ - 0.083 and 1.0183, respectively).
The Haldane mathematical model has an advantage over the Monod model - one can determine theoretically what the optimal substrate concentration will be at the optimal (maximum) specific growth rate:

$$\mu = \mu_{\max}^{opt} \Rightarrow S_{opt} = \sqrt{K_{SX} K_{SXi}} \qquad (10)$$

Similarly, the optimal substrate concentration at which the rate of lactic acid synthesis will be optimal (maximum value) can be determined:

$$q_p = q_{p\max}^{opt} \Rightarrow S_{opt} = \sqrt{K_{SP} K_{SPi}} \qquad (11)$$

Then, according to the Haldane model, S_{opt} for the growth and reproduction of the strain will be 52.71 g/dm^3, which is 5.271% glucose in the nutrient medium. This value is also close to experimentally determined values by other authors (Bouguettoucha et al., 2011). Here, too, it should be noted that the concept of substrate should be considered as a balanced set of components that are necessary for cell growth. In this case, the result obtained is very close to the total amount of components in the used complex nutrient medium. For the acid formation rate S_{opt} is 241.63 g/dm^3. According to these data obtained from the Haldane model, it can be concluded that for *Lactobacillus delbrueckii* ssp. *bulgaricus* B1 the substrate concentration should be in the range from 52.71 g/dm^3 to 241.63 g/dm^3, and if the substrate concentration is above 241.63 g/dm^3 there will be complete inhibition of both the growth of the strain and its biosynthetic ability.
Lactic acid fermentation is also a product-inhibited process (Bouguettoucha et al., 2011;), which is why the growth kinetics and lactic acid biosynthesis of *Lactobacillus delbrueckii* ssp. *bulgaricus* B1 with the Aiba model have been modeled. The results are shown in Fig.5, and the parameters of the model are presented in Table. 1.

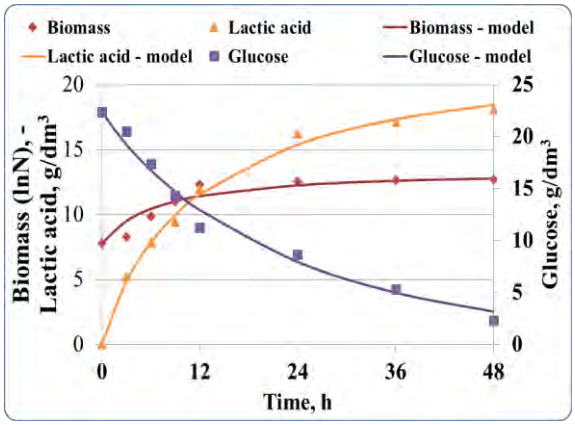


Figure 5: Kinetics of Lactic Acid Fermentation in Cultivation of *Lactobacillus delbrueckii* ssp. *bulgaricus* B1 in MRS-broth Described with the Aiba Model

The Aiba model is characterized by high correlation values ranging from 0.8860 to 0.9946 and low identification errors. It gives relatively high rates of the specific growth rate - 0.576 h^{-1} and 0.266 h^{-1}, but also confirms product inhibition. This is evidenced by the relatively close values of the constants K_{PX} and K_{SP} - 0.16 g/dm^3 and 0.11 g/dm^3. This in turn confirms that the cultivation process of *Lactobacillus delbrueckii* ssp. *bulgaricus* B1 must be carried out with neutralization of the lactic acid produced in order to achieve high maximum growth rate of the culture and high concentration of active cells. Low values of the product inhibition constants can also be considered as an indirect indicator of the sensitivity (resistance) of the strain to the acidic pH of the stomach, which means that this strain is good to be used in encapsulated form, as a probiotic strain. The presence of product inhibition increases the saturation constant value for biomass (213.78 g/dm^3). The Aiba model again shows that most of the substrate is used for biomass formation. The values of the trophic coefficients - $1/Y_{X/S}$=0.6182 and $1/Y_{P/S}$=2.2487, and therefore the values of the economic coefficients $Y_{X/S}$=1.6176 and $Y_{P/S}$=0.4021 serve as a proof of this conclusion.
The data described so far show that lactic acid fermentation is both a substrate- and a product-inhibited process, which should be taken into account in its modeling. Combined models such as the Haldane-Aiba model (equation 7), the Haldane model for product inhibition (equation 8) and the Haldane-Jerusalimski model (equation 9) can be used for this purpose. The results of the modeling of the cultivation process of

Lactobacillus delbrueckii ssp. *bulgaricus* B1 are reflected in Fig. 6 to Fig. 8, as well as in Table 1.

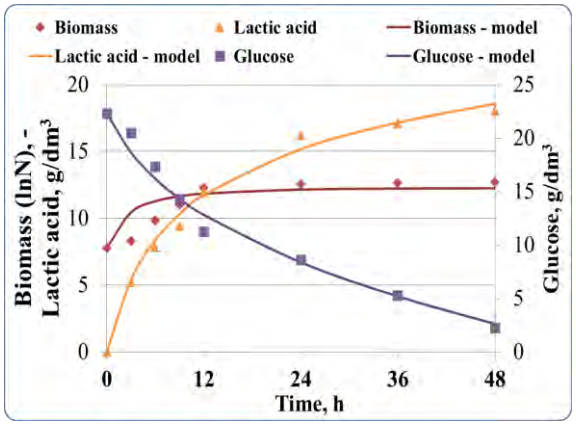


Figure 6: Kinetics of Lactic Acid Fermentation in Cultivation of *Lactobacillus delbrueckii* ssp. *bulgaricus* B1 in MRS-broth Described with the Haldane-Aiba Model

The data in Table. 1 show that all three models, including product and substrate inhibition, describe experimental data with high accuracy. The correlation coefficients vary in the range of 0.87 to 0.9943, the errors - in the range of 0.27 to 1.8 for the individual indicators.

The data in Table 1 show that the Haldane-Aiba model (Equation 7) and the Haldane-Jerusalimski model (Equation 9) predict lower biomass specific growth rate of 0.049 h^{-1} and 0.011 h^{-1}, respectively. The Haldane model for product inhibition (Equation 8) predicts growth rate of 0.113 h^{-1}. This difference is due to the approach used by the three models to describe the processes of product and substrate inhibition. The Haldane-Jerusalimski model determines lower value of the saturation constant - 17.55 g/dm^3 compared to the Haldane-Aiba model - 47.59 g/dm^3 and the Haldane model for product inhibition - 65.69 g/dm^3. Despite this difference, the results are within the range expected in the cultivation of lactic acid bacteria. The data in Table 1 show that the three models predict growth inhibition by the substrate at concentrations above 15.67%, i.e. well above the current value of glucose in the MRS-broth medium. The models show that complete inhibition of growth by the substrate will occur only at glucose concentrations in the medium above 28.38%, and such values are not typical for nutrient media designed for the cultivation of lactic acid bacteria.

Data on the specific rate of lactic acid formation are of great interest in the enlisted models. The Haldane-Aiba model and the Haldane-Jerusalimski model give quite low values of q_{pmax} - 0.063 h^{-1} - 0.065 h^{-1}, which means that the MRS-broth medium is designed to allow enhanced synthesis of biomass at the expense of lactic acid production. The Haldane model for product inhibition predicts more intense process of lactic acid biosynthesis, as evidenced by the significantly higher maximum specific rate of acid formation (Table 1).

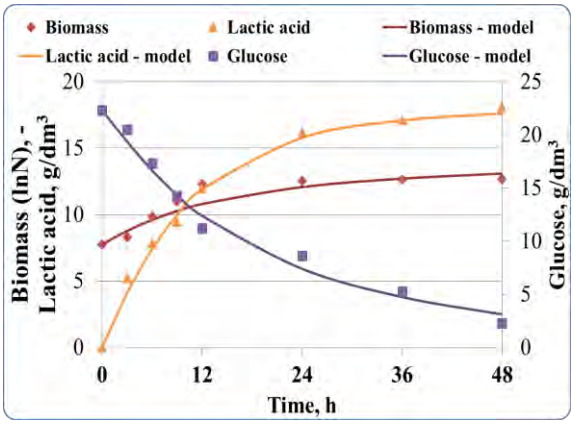


Figure 7: Kinetics of Lactic Acid Fermentation in Cultivation of *Lactobacillus delbrueckii* ssp. *bulgaricus* B1 in MRS-broth Described with the Haldane model for Product Inhibition

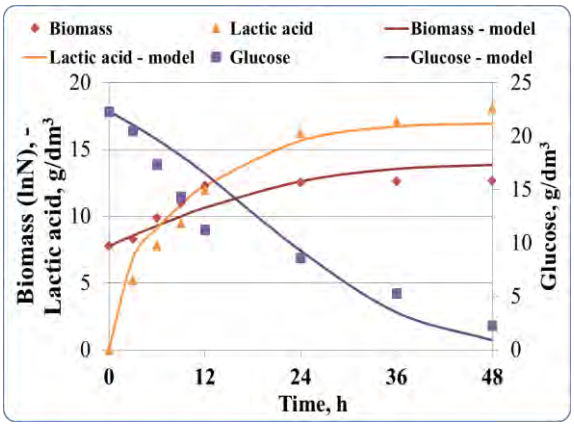


Figure 8: Kinetics of Lactic Acid Fermentation in Cultivation of *Lactobacillus delbrueckii* ssp. *bulgaricus* B1 in MRS-broth Described with the Haldane-Jerusalimski Model

However, the three models give different degrees of influence of lactic acid formation on cell growth. The Haldane-Aiba model shows relatively strong product inhibition due to low K_{PX} and K_{PP} values of 0.249 g/dm^3 and 0.138 g/dm^3, respectively. The Haldane model for product inhibition gives higher K_{PX} and K_{PP} values - 22.38 g/dm^3 and 18.53 g/dm^3, which according to this model means that the process is not product inhibited so strongly. The value of the product inhibition constant for the biomass given by the Haldane-Jerusalimski model is abnormally high. From a mathematical point of view, this value is correct and brings the values calculated by the model closer to the experimental ones, but its biological meaning is doubtful, as this is too high a concentration of lactic acid at which the specific growth rate would be half the maximum value. For K_{PP} this model gives a biologically realistic value - 3.36 g/dm^3.

The three models, including product and substrate inhibition, also allow the determination of the optimal substrate concentrations to provide maximum (optimal) specific growth rate or acid formation (Table 1). The data presented in the table show that the Haldane-Aiba model and the Haldane-Jerusalimski model give close

values of the optimal glucose concentration - 86.35 g/dm^3 and 70.58 g/dm^3, respectively. Once again, one must recall that the models determine the optimal concentration of the balanced set of components, rather than just the concentration of the carbon source in the medium. The Haldane model for product inhibition rather sets the substrate concentration limit (103.55 g/dm^3) at which inhibition of *Lactobacillus delbrueckii* ssp. *bulgaricus* B1 growth will begin.

Similar conclusions can be made for S_{opt} for lactic acid synthesis. This time, however, the Haldane model for product inhibition and the Haldane-Jerusalimski model predict substrate concentrations that are characteristic of lactic acid bacteria. Only the Haldane-Aiba model showed an abnormally low value for S_{opt} - 0.96 g/dm^3.

The results obtained (Fig. 2 to Fig. 8 and Table 1) do not support the statement cited in the introduction that the MRS-broth medium may not be suitable for the cultivation of strains originating from dairy products. In the cultivation of *Lactobacillus delbrueckii* ssp. *bulgaricus* B1 in MRS-broth data show that the strain grew at relatively high growth rates and lower rates of lactic acid accumulation. The effects of substrate and product inhibition are normal for this type of fermentation, which allows relatively higher concentrations of viable cells to accumulate at the end of fermentation.

CONCLUSION

An important requirement for selection of strains to be included in the production of probiotic foods and preparations is the ability of the selected strains to be cultivated in industrial conditions and to accumulate high concentrations of viable cells. In the present work, the cultivation of *Lactobacillus delbrueckii* ssp. *bulgaricus* B1, isolated from home-made yoghurt, cultivated in a complex culture medium (MRS-broth) was studied. Complex media provide a balanced set of components - carbon, nitrogen and phosphorus source, micro- and macroelements. These media are usually designed to provide optimal growth, but in some cases are unsuitable for certain strains. The obtained results show that the selected probiotic strain *Lactobacillus delbrueckii* ssp. *bulgaricus* B1 grew at relatively high specific growth rates and accumulated moderate amounts of lactic acid, determined by moderate specific acid formation rates. The data show that the strain may be sensitive to lactic acid, which is why pH adjustment and neutralization of lactic acid accumulated in the medium can be applied in industrial cultivation. This will ensure complete absorption of the substrate by the cells and the accumulation of maximum cell numbers in the medium.

The data obtained show that, although of dairy origin, *Lactobacillus delbrueckii* ssp. *bulgaricus* B1 can be cultivated in the complex nutrient medium MRS-broth.

REFERENCES

Arena, M.P., G. Caggianiello, P. Russo, M. Albenzio, S. Massa, D Fiocco, V. Capozzi, G. Spano. 2015. "Functional Starters for Functional Yogurt" *Foods*, 4 (1), 15-33. https://doi.org/10.3390/foods4010015

Bailey, J. E. and Ollis, D. F. (1986). Biochemical Engineering Fundamentals, 2nd edition. McGraw-Hill.

Bouguettoucha, A., B. Balannec and A. Amrane. 2011. "Unstructured Models for Lactic Acid Fermentation: A Review." *Food Technol. Biotechnol.*, 49 (1), 3–12.

Choi, M., M. Saeed Al-Zahrani, and S. Y. Lee. 2014. "Kinetic model-based feed-forward controlled fed-batch fermentation of *Lactobacillus rhamnosus* for the production of lactic acid from Arabic date juice." *Bioprocess Biosyst Eng.*, 37, 1007–1015. https://doi.org/10.1007/s00449-013-1071-7

Corry, J., G Curtis, R., Baird. 2003. "Handbook of culture media for food microbiology", (Chapter de Man, Rogosa and Sharpe (MRS) agar), *Progress in Industrial Microbiology*, 37, 511-513. https://doi.org/10.1016/S0079-6352(03)80066-8

Gibson, G. R. 2004. "From probiotics to prebiotics and a healthy digestive system". *J. Food Science*, 69 (5), M141- M143. https://doi.org/10.1111/j.1365-2621.2004.tb10724.x

Goranov, B., V. Shopska, R. Denkova, G. Kostov, Georgi. 2015. "Kinetics of batch fermentation in the cultivation of a probiotic strain *Lactobacillus delbrueckii* ssp. *bulgaricus* B1" *Acta Universitatis Cibiniensis. Series E: Food Technology*, 19 (1), 61-72. https://doi.org/10.1515/aucft-2015-0006

Gordeev, L., A. Koznov, A. Skichko, and Y. Gordeeva. 2017. "Unstructured mathematical models of the lactic acid biosynthesis kinetics: A Review." *Theoretical Foundations of Chemical Engineering*, 51 (2), 175-190.

ISO/TS 11869:2012. Fermented milks — Determination of titratable acidity — Potentiometric method

ISO 7889:2005. Yogurt — Enumeration of characteristic microorganisms — Colony-count technique at 37 degrees C

Ivanov, I., K. Petrov, V. Lozanov, I. Hristov, Zh. Wu, Zh. Liu, P. Petrova. 2021. "Bioactive compounds produced by the accompanying microflora in Bulgarian yoghurt" Processes 9 (1), 114. https://doi.org/10.3390/pr9010114

Kostov, G. 2015. "Intensivication of fermentation processes by immobilized biocatalysis". DSc Thesis, University of Food Technologies, Plovdiv. p. 307. (in Bulgarian).

Kostov, G., R. Denkova-Kostova, V. Shopska, B. Goranov, Z. Denkova. 2021. „Comparative evaluation of *Lactobacillus plantarum* strains through microbial growth kinetics", In *ECMS 2021 Proceedings* Kh. Al-Begain, M. Iacono, L. Campanile, A. Bargiela (Eds.), European Council for Modeling and Simulation. https://doi.org/10.7148/2021-0165

Maisto, M., G. Annunziata, E. Schiano, V. Piccolo, F. Iannuzzo, R. Santangelo, R. Ciampaglia, G. C. Tenore, E. Novellino, P. Grieco. 2021. "Potential

Functional Snacks: Date Fruit Bars Supplemented by Different Species of *Lactobacillus* spp." *Foods*, 10 (8), 1760. https://doi.org/10.3390/foods10081760

Muloiwa, M., S. Nyende-Byakika, M. Dinka. 2020. "Comparison of unstructured kinetic bacterial growth models". *South African Journal of Chemical Engineering*, 33, 141-150. https://doi.org/10.1016/j.sajce.2020.07.006

Saarela, M., L. Zahteenmaki, R. Crittenden, S. Salminen, and T. Mattila-Sandholm. 2002. "Gut bacteria and health foods – the European perspective." *Int. J. Food Microbiol*. 78, 99-117.

Shopska, V., R. Denkova, V. Lyubenova, G. Kostov. 2019. "Kinetic Characteristics of Alcohol Fermentation in Brewing: State of art and control of the fermentation process". In *Fermented Beverages*; Grumezescu, A.M., Holban, A.M., Eds.; Woodhead Publishing: Cambridge, UK, 2019; pp. 529–575. https://doi.org/10.1016/B978-0-12-815271-3.00013-0

Teneva, D., B. Goranov, R. Denkova, Z. Denkova. 2015. "Antimicrobial activity of *Lactobacillus delbrueckii* ssp. *bulgaricus* strains against Candida albicans NBIMCC 74". Scientific works of the University of Ruse, 54, series 10.4, 20-25.

ACKNOWLEDGEMENTS

This work were supported by the Bulgarian Ministry of Education and Science under the National Research Programme "Healthy Foods for a Strong Bio-Economy and Quality of Life" approved by DCM № 577/17.08.2018 and by the project "*Strengthening the research excellence and innovation capacity of University of Food Technologies - Plovdiv, through the sustainable development of tailor-made food systems with programmable properties*", part of the European Scientific Networks National Programme funded by the Ministry of Education and Science of the Republic of Bulgaria (agreement № Д01-288/07.10.2020).

AUTHOR BIOGRAPHIES

GEORGI KOSTOV is a full professor at the Department of Wine and Beer Technology at the University of Food Technologies, Plovdiv. He received his MSc degree in Mechanical Engineering in 2007, a PhD degreein Mechanical Engineering in the Food and Flavor Industry (Technological Equipment in the Biotechnology Industry) in 2007 at the University of Food Technologies, Plovdiv, and holds a DSc degree in Intensification of Fermentation Processes with Immobilized Biocatalysts from 2015. His research interests are in the area of bioreactor construction, biotechnology, microbial population investigation and modeling, hydrodynamics and mass transfer problems, fermentation kinetics, and beer production.

VESELA SHOPSKA is an associated professor at the Department of Wine and Beer Technology at the University of Food Technologies, Plovdiv. She received her MSc degree in Wine-making and Brewing Technology in 2006 at the University of Food Technologies, Plovdiv. She received her PhD in Technology of Alcoholic and Non-alcoholic Beverages (Brewing Technology) in 2014. Her research interests are in the area of beer fermentation with free and immobilized cells, yeast and bacteria metabolism and fermentation activity.

ROSITSA DENKOVA-KOSTOVA is an associated professor at the Department of Biochemistry and Molecular Biology at the University of Food Technologies, Plovdiv. She received her MSc degree in Industrial Biotechnologies in 2011 and a PhD degree in Biotechnology (Technology of Biologically Active Substances) in 2014. Her research interests are in the area of isolation, biochemical and molecular-genetic identification and selection of probiotic strains and development of starters for functional foods.

BOGDAN GORANOV is a chief assistant professor at the department of Microbiology at the University of Food Technologies, Plovdiv. He received his PhD in 2015 from the University of Food Technologies, Plovdiv. The theme of his thesis was "Production of Lactic Acid with Free and Immobilized Lactic Acid Bacteria and its Application in the Food Industry". His research interests are in the area of bioreactor construction, biotechnology, microbial population investigation and modeling, hydrodynamics and mass transfer problems, and fermentation kinetics.

ZAPRYANA DENKOVA is a full professor at the department of Microbiology at the University of Food Technologies, Plovdiv. She received her MSc in "Technology of microbial products" in 1982, PhD in „Technology of biologically active substances" in 1994 and DSc on "Production and application of probiotics" in 2006. Her research interests are in the area of selection of probiotic strains and development of starters for food production, genetics of microorganisms, and development of functional foods.

This publication has been specially selected for reprinting by the "Simulation and Optimization" track of the 36th ECMS International Conference on Modelling and Simulation.

ENTWICKLUNG EINES VORGEHENSMODELLS ZUR NUTZUNG VON PROCESS MINING FÜR DIE AUSWAHL UND ANWENDUNG VON RPA-LÖSUNGEN ZUR OPTIMIERUNG VON GESCHÄFTSPROZESSEN

Fabian Karkos
HS Pforzheim
Tiefenbronnerstr. 65,
75175 Pforzheim
karkosfa@hs-pforzheim.de

Frank Morelli
HS Pforzheim
Tiefenbronnerstr. 65,
75175 Pforzheim
frank.morelli@hs-pforzheim.de

SCHLÜSSELWÖRTER
Process Mining, Robotic Process Automation, Automatisierung, Geschäftsprozesse, Vorgehensmodell

ABSTRACT

Der Einsatz von Robotic Process Automation (RPA) kann helfen, Geschäftsprozesse zu optimieren. Die Kombination mit dem Process Mining (PM) Ansatz erweist sich geeignet, um diese Chancenpotenziale systematisch zu nutzen. Dies vollzieht sich über die Verknüpfung zur Selektion, Implementierung und Steuerung von RPA-Lösungen. Der vorliegende Artikel beschreibt Voraussetzungen, Einsatzbedingungen und Inhalte eines zugehörigen Vorgehensmodells. Das Konzept lehnt sich am BPM-Lebenszyklusmodell der einschlägigen Literatur an und erweitert dieses um die Aspekte Datenextraktion, RPA-Bot-Entwicklung und -Test. Die Schritte werden durch Beispiele illustriert.

EINLEITUNG

Geschäftsprozesse fungieren als zentraler Bestandteil in Unternehmen (Dumas et al. 2013). Sie geben Richtlinien und Anleitungen, wie Leistungen und Dienste möglichst effizient und effektiv erbracht werden sollen (Gadatsch 2020). Ziel des Geschäftsprozessmanagements (englisch Business Process Management bzw. BPM) ist es, diese Abläufe zu analysieren und kontinuierlich zu verbessern (Gadatsch 2020). Gleichzeitig vollzieht sich in vielen Unternehmen eine digitale Transformation, welche die Chancen zur Optimierung von Geschäftsprozessen erhöht, beispielsweise durch Automatisierung in Form von RPA. Das starke Interesse vieler Unternehmen an zugehörigen RPA-Lösungen dokumentiert das weltweite Marktwachstum um über 19% zwischen 2020 und 2021 (Gartner 2020). In der Theorie ermöglicht RPA eine schnelle, skalierbare Optimierung hinsichtlich Kosten, Durchlaufzeiten und Fehleranfälligkeit von Geschäftsprozessen, ohne dabei die bestehende IT-System- und Anwendungslandschaft zu verändern (Hofmann et al. 2020). Aus Untersuchungen geht jedoch hervor, dass RPA nicht für jeden Geschäftsprozess geeignet ist und ggfs. zusätzliche Fehlerquellen eröffnet. (Hofmann et al. 2020). Aufgrund der Problematik des RPA-Einsatzes wurde in der Literatur bereits der Zusammenhang zum PM beleuchtet (Peters und Nauroth 2019; Van der Aalst 2021).

PM nutzt die aufgezeichneten Daten aus Informationssystemen und anderen Quellen zur Abbildung von Abläufen, um diese über einen Algorithmus als Prozessmodelle realitätsnah zu visualisieren (Van der Aalst 2016). Oftmals beschränkt sich die Verknüpfung von PM und RPA auf das Identifizieren und Visualisieren von relevanten Geschäftsprozessen. Der vorliegende Artikel verfolgt das Ziel, weiterführende Möglichkeiten zur Verbindung beider Technologien im Sinne von Synergieeffekten aufzuzeigen. Aus Sicht der Autoren ist PM nicht ausschließlich für Prozessidentifikation und Visualisierung geeignet und der Erfolg von RPA nicht nur an dessen Implementierung gebunden (Smeets et al. 2019; Van der Aalst 2016).

Um diesen Zusammenhang zu untersuchen, werden zwei Forschungsfragen (FF) aufgestellt:

FF 1: Welchen Nutzen erbringt die Anwendung von PM für die Identifikation, Ausgestaltung und Überwachung von RPA-Lösungen im Rahmen des BPM?

FF 2: Welche Phasen lassen sich für einen PM-unterstützen RPA-Bot-Einführung definieren?

Der Aufbau des Artikels untergliedert sich in zwei Bestandteile. Im ersten Abschnitt wird die theoretische Grundlage zur späteren Beantwortung der Forschungsfragen gesetzt. Im zweiten Teil des Artikels erfolgt die explizite Behandlung der Forschungsfragen. Der Einsatz von PM zur RPA-Bot-Einführung und späteren Verwaltung wird aus verschiedenen Perspektiven beleuchtet. Dies erfolgt zum einen durch die Würdigung der zugehörigen einschlägigen Literatur. Zum anderen wird ein Vorgehensmodell erarbeitet, das auf den Ablauf einer RPA-Bot-Einführung mit unterstützendem PM abzielt. Dabei dient die Kombination des BPM-Lebenszyklusmodells mit der Vorgehensweise einer generischen RPA-Bot-Einführung aus der Literatur als Grundlage für den eigenen konzeptionellen Ansatz. Abb. 1 fasst die Vorgehensweise im vorliegenden Artikel zusammen.

Abb. 1: Vorgehensweise des Artikels

(Quelle: Eigene Darstellung)

Process Mining

PM wird der Process-Science-Disziplin zugeschrieben. Diese Domäne hat zum Ziel, das Wissen von Informationstechnologien mit dem Managementwissen eines Unternehmens zu verbinden, um eine Prozessoptimierung zu fördern. Dafür werden im Rahmen des PM, Daten aus einem Event-Log gezogen, das sich wiederum aus verschiedenen IT-Systemquellen im Unternehmen speist (Van der Aalst 2016). Ein Event-Log speichert Daten zu ausgeführten Arbeitsschritten. Die Menge an unterschiedlichen Daten, die sich im Event-Log nutzen lassen, hängt wesentlich von den aufgezeichneten Daten der zugrunde liegenden IT-Systeme ab. Die hieraus entstehende Datenvariabilität führt zu einer Vielfalt an auswertbaren Perspektiven eines Prozesses. Die Prozessdaten werden zur Entdeckung, Überprüfung und Erweiterung real vorkommender Prozesse verwendet, welche die Säulen bzw. grundlegenden PM-Ansätze widerspiegeln (Van der Aalst 2016). PM generiert aus den Daten durch einen Algorithmus Ablaufvarianten eines Prozesses, wodurch diese auf tatsächlich durchlaufenen Prozessinstanzen basieren (Van der Aalst 2016). Ziel ist die Objektivierung bzw. Transparenz von Ist-Prozessen, die im Gegensatz zur klassischen Vorgehensweise innerhalb des BPM nicht mehr auf der Basis von Interviews oder Workshops erstellt werden müssen (Grisold et al. 2021; Van der Aalst 2016).

Bei der *Prozessentdeckung* erfolgt die Datenübernahme automatisch durch einen Process-Mining-Algorithmus aus dem Event Log. Für die Visualisierung gibt es unterschiedliche Lösungsalternativen. Dieser Schritt benötigt kein menschliches Expertenwissen über den Ist-Prozess (Rozinat et al. 2007; Van der Aalst 2012; Van der Aalst 2016). Der Fokus dieser Prozessmodelle ist häufig eine kontrollflussorientierte Darstellung. Je nach Anwendungszweck kann es sich als sinnvoll erweisen, dass man eine andere Perspektive betrachtet (Van der Aalst 2016). Im Rahmen der Prozessentdeckung lassen sich die entstehenden Prozessmodelle beispielsweise auf organisatorische Strukturen oder zeitorientierte Faktoren in Abhängigkeit von einzelnen Prozessinstanzen untersuchen (Van der Aalst et al. 2011).

Die *Konformitätsprüfung* repräsentiert eine weiterführende Funktionalität des PM. Ziel ist es, die bereits im Unternehmen dokumentierten Soll-Prozessmodelle mit der tatsächlichen Handhabung dieser Abläufe zu vergleichen, um so mögliche Abweichungen zu ermitteln. (Van der Aalst 2016).

Das Ziel der *Prozesserweiterung* besteht darin, basierend auf der Datengrundlage der Event-Logs und der vorhandenen Prozessvarianten, das Prozessgefüge zu optimieren (Van der Aalst 2016). Dabei werden Daten über den tatsächlichen Prozessverlauf durch einen menschlichen Benutzer in das bestehende Soll-Prozessmodell eingepflegt, beispielsweise mithilfe der Modellierungssprache BPMN 2.0.

Es lassen sich vier Dimensionen definieren, nach denen man das Prozessmodell erstmals erstellen bzw. das Prozessmodell und die Datenlage später vergleichen kann:

- Passgenauigkeit: Je höher die Passgenauigkeit, desto mehr lassen sich im Event-Log identifizierte Prozessinstanzen durch das Prozessmodell erklären (Rozinat et al. 2007; Van der Aalst 2012).
- Einfachheit: Entsprechend „Ockhams Rasiermesser“ ist das einfachste Prozessmodell, welches das Event-Log zutreffend beschreibt, zu präferieren (Van der Aalst 2016).
- Präzision: Voraussetzung für ein präzises Prozessmodell ist die Abbildung aller Sequenzen im Event-Log. Entsteht hingegen ein Prozessmodell, das Ausprägungen enthält, die aus dem Event-Log nicht direkt hervorgehen, gilt es als unpräzise (Rozinat et al. 2007).
- Generalisierbarkeit: Im Sinne einer Verallgemeinerung wird bewertet, inwieweit das Prozessmodell in der Lage ist, zukünftiges Verhalten zu reproduzieren. Da Event-Logs lediglich einen Ausschnitt der Realität repräsentieren, sollte das Prozessmodell nicht nur exakt diese Sequenzen repräsentieren können (Rozinat et al. 2007; Van der Aalst 2012).

Aus den Ausprägungen der Dimensionen geht ein Spannungsverhältnis hervor: So steht bspw. ein verallgemeinerndes Modell im Widerspruch zu einem präzisen Modell, da hierbei ein entgegengesetzter Fokus gesetzt wird (Van der Aalst 2016). Die Dimensionen, nach welchen ein Prozessmodell erstellt oder verglichen wird, sollte man deshalb an den Zweck des Modells anpassen.

Die angestrebte Integration von PM und BPM wird nachfolgend durch eine Erweiterung des BPM-Lebenszyklusmodells realisiert. Hierzu erfolgt eine Beschreibung der einzelnen Phasen und des jeweiligen Nutzens, der durch die Anreicherung mit dem PM-Gedankengut generiert werden soll.

- Prozessidentifikation und Entdeckungsphase: Das erfahrungsbasierte, häufig auf Basis von qualitativen Erhebungen gewonnene, Ist-Prozessdesign wird durch eine automatisierte Prozessmodellierung auf Basis von faktischen Daten und Algorithmen ersetzt. Dies ermöglicht eine perspektivische Erweiterung um latente Prozessvarianten (Peters und Nauroth 2019; Van der Aalst et al. 2011; Van der Aalst 2016).
- Analyse- und Restrukturierungsphase: Der Einsatz verschiedener Perspektiven lässt sich durch Anreicherung des Event-Logs zugehöriger Attribute erreichen. Dies ermöglicht ein methodisches und hypothesengestütztes Vorgehen bei der Untersuchung: Ursache-Wirkungsbeziehungen und auftretende Muster bei Schwachstellen lassen durch das Verfolgen von Prozessinstanzen detailliert diagnostizieren (Peters und Nauroth 2019).
- Implementierungsphase: Diese entspricht der Vorgehensweise innerhalb des BPM-Lebenszyklusmodells und erfährt keine zusätzliche Unterstützung durch das PM (Van der Aalst et al. 2011).
- Performance – und Überwachungsphase: Die Bewertung von Prozessen basiert auf der Messung von Kennzahlen und der Konformität durchgeführter Aktivitäten. PM-basiertes Monitoring ermöglicht es, diese Vorgehensweise Echtzeit-nah und nicht erst ex post anzuwenden (Van der Aalst 2016).

Die aufgezeigte Verknüpfung von PM und BPM bildet die Basis für die Konzeption des Vorgehensmodell in Abb. 2.

Robotic Process Automation

Bei RPA handelt es sich um eine Softwarelösung zur Automatisierung von Geschäftsprozessen durch Software-Roboter, oftmals auch RPA-Bots genannt (Santos et al. 2020). RPA ist ein weitreichender Begriff für verschiedene Werkzeuge, welche auf den grafischen Oberflächen von IT-Systemen arbeiten, u.a. durch die Eingabe von digital strukturierten Daten (Santos et al. 2020; Syed et al. 2020; Van der Aalst et al. 2018). RPA baut auf den in Unternehmen verwendeten IT-Systemen auf, ohne diese zu verändern. Hierin unterscheidet sich diese Technologie von klassischen Prozessautomatisierungen durch ein Business Process Management System (BPMS) oder durch die Programmierung von Schnittstellen (Hofmann et al. 2020).

Bei den RPA-Bots differenziert man zwischen „Attended Robots“ (Desktop-RPA) und „Unattended Robots“ (sog. RPA-Plattformen):

• „Attended Robots“ bzw. überwachte RPA-Bots repräsentieren (häufig von Mitarbeitern oder dem verantwortlichen RPA-Team) programmierte Neuentwicklungen, die auf einem ausgewählten Computer oder mobilen Gerät lokal ausgeführt werden (Langmann und Turi 2020).

• „Unattended Robots“ bzw. unüberwachte RPA-Bots werden i.d.R. durch ein Projektteam entwickelt und zentral über virtuelle Maschinen verwaltet (Koch und Fedtke 2020; Langmann und Turi 2020; Smeets et al. 2019).

Durch Module, die Anweisungen zur Ausführung der Arbeitsschritte beinhalten, können Prozessanweisungen programmiert und bei Bedarf beliebig angeordnet, gelöscht oder ergänzt werden (Hofmann et al. 2020). RPA lässt sich in seiner Interaktion mit IT-Systemen als Imitation menschlichen Handelns interpretieren und im direkten Vergleich zur bisherigen Ausführung durch eine schnellere, weniger fehleranfällige, skalierbare und nachverfolgbare Bearbeitung der Arbeitsschritte charakterisieren. Zielsetzung ist es, durch das Ersetzen manueller Arbeit, Kosten zu reduzieren (Aguirre und Rodriguez 2017; Hofmann et al. 2020; Wewerka und Reichert 2020).

RPA wird im vorliegenden Artikel als Erweiterung bzw. Ergänzung von bereits bestehenden Automatisierungslösungen gesehen. Ein Einsatz erfolgt dann, wenn traditionelle Automatisierung wirtschaftlich nicht von Vorteil ist (Van der Aalst 2021). Je nach vorliegendem Prozess kann der Grad der Automatisierung durch RPA unterschiedlich sein: RPA lässt sich nutzen, um Prozesse durchgängig oder teilweise zu automatisieren. Im zweiten Fall geht es um die Automatisierung einzelner, nicht verbundener Aktivitäten eines Prozesses (Hofmann et al. 2020). Eine Automatisierung einzelner Aktivitäten wird typischerweise durch überwachte RPA-Bots realisiert (Hofmann et al. 2020; Smeets et al. 2019).

RPA lässt sich nicht universell einsetzen. Für Prozesse erweist es sich deshalb als sinnvoll, Merkmale zu definieren, die für die Auswahl von RPA-Bot-Implementierungsprojekten zu berücksichtigen sind:

- Häufigkeit der Ausführung: Je öfter man einen Prozess ausführt, desto mehr eignet sich dieser für eine RPA-Implementierung. Selten ausgeführte Prozesse sind c.p. aufgrund ihrer geringeren Wirtschaftlichkeit weniger relevant (Santos et al. 2020; Syed et al. 2020; Van der Aalst et al. 2018).
- Standardisierbarkeit des Prozesses: Je mehr Varianten eines Prozesses vorhanden sind, desto aufwändiger und komplexer ist die Realisierung der RPA-Lösung (Santos et al. 2020; Syed et al. 2020). Entsprechend erweisen sich einfachere Prozesse als vorteilhaft für die Implementierung.
- Ausprägung des Prozessverlaufs: Die Daten müssen einer klaren Struktur und Semantik folgen. Die RPA-Technologie ermöglicht eine Automatisierung von regelbasierten (Wenn-Dann-)Prozessen.

Unscharfe Prozesse, die flexibel gehandhabt werden müssen, eignen sich hingegen weniger (Hofmann et al. 2020; Santos et al. 2020; Schmitz et al. 2019; Syed et al. 2020).

- Grad der Automatisierung: Im Vergleich zu einem menschlichen Bearbeiter führt ein RPA-Bot die Aufgabe schneller und mit konstanter Qualität durch. Prozesse bzw. Aktivitäten, welche ein hohes Maß an menschlicher Arbeit aufweisen, sind deshalb für einen RPA-Einsatz höher zu priorisieren (Santos et al. 2020; Syed et al. 2020).
- Anzahl benötigter IT-Systeme: RPA bietet sich vor allem an, wenn auf viele verschiedene IT-Systeme zugegriffen wird, für die keine fertigen Schnittstellen bestehen. Durch RPA ist keine Veränderung der IT-Systeme notwendig. Traditionelle Prozessautomatisierung wird in solchen Fällen demgegenüber oft als zu teuer angesehen (Santos et al. 2020; Syed et al. 2020).
- Digitalisierte Prozesse und Daten: Daten und Prozess-Schritte in digitaler Form bilden eine notwendige Voraussetzung für den RPA-Einsatz.

Weiterhin erweisen sich nachfolgende Aspekte für die Entscheidungsfindung als relevant:

- Je besser der Prozess und seine Varianten dokumentiert sind, desto effektiver lassen sich die RPA-Bedingungen untersuchen (Syed et al. 2020).

- Die Verteilung der Prozessdurchläufe eines Prozesses folgt oftmals dem Pareto-Prinzip, wobei 80% der Durchläufe ca. 20% der Prozessvarianten zugeschrieben werden können (Van der Aalst et al. 2018; Van der Aalst 2021). Bestehen innerhalb der Dokumentation keine Angaben zur Verteilung der Prozessverläufe, kann dies zu einem verfälschten Prozessbild führen.

Die zuvor definierten Bedingungen für einen RPA-Einsatz werden im Verlauf der RPA-Bot-Einführung aufgegriffen und um weitere Schritte ergänzt. Der nachfolgende Absatz behandelt die generischen Schritte einer RPA-Bot-Einführung und deren Auswirkungen:

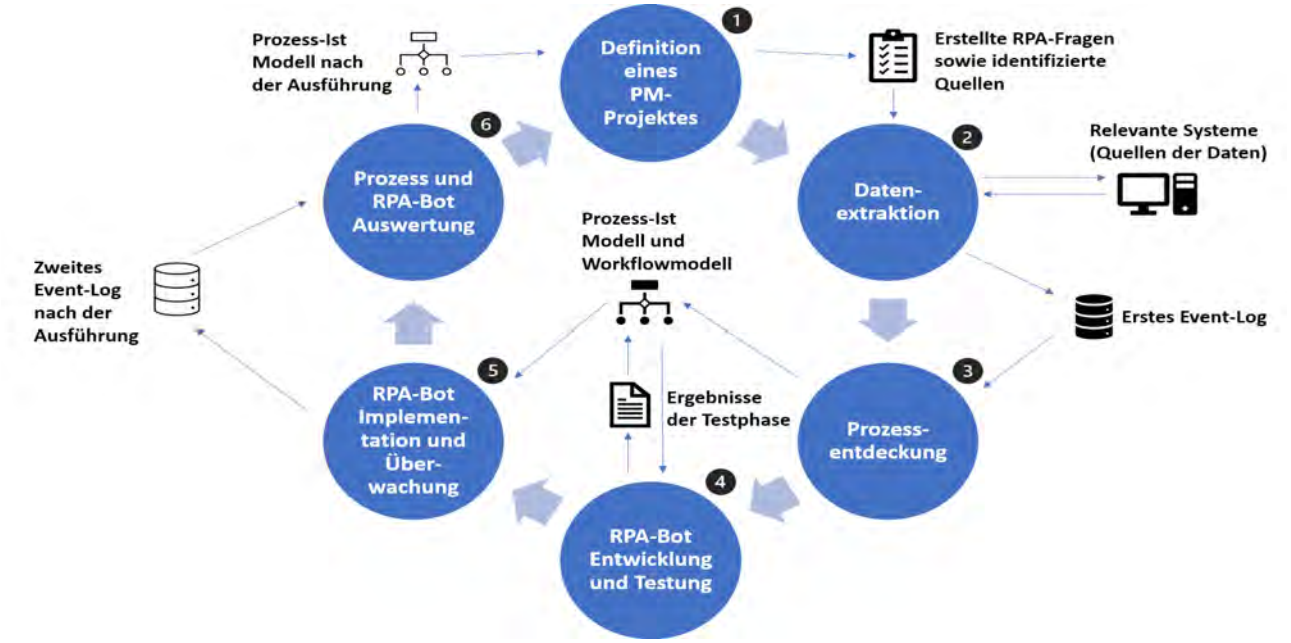
- Prozessselektion: Hierbei erfolgt die Auswahl eines oder mehrerer Prozesse für eine RPA-Bot-Einführung gemäß einer vorhandenen Zielsetzung. Dabei sind u.a. relevante Prozesse in Modellform zu visualisieren (Flechsig et al. 2019; Smeets et al. 2019; Soybir und Schmidt 2021).
- Prozessanalyse: In dieser Phase werden die zuvor visualisierten Prozesse entsprechend den genannten Bedingungen auf ihre Tauglichkeit zur RPA-Bot-Unterstützung untersucht und ggf. restrukturiert. Die Prozessanalyse endet mit der binären Entscheidung, einen Prozess durch RPA zu automatisieren oder nicht (Flechsig et al. 2019; König et al. 2020; Smeets et al. 2019; Soybir und Schmidt 2021).
- RPA-Bot Entwicklungsphase: Für einen ausgewählten Prozess erfolgt das Erstellen eines RPA-Bots. Es lassen sich komplette Prozessverläufe oder auch einfache Aktivitäten des Prozesses durch RPA-Module implementieren (Flechsig et al. 2019; Smeets et al. 2019; Soybir und Schmidt 2021).
- Testphase: Die zuvor entwickelten RPA-Module werden in einer Testumgebung auf ihre Funktionalität und Fehler überprüft. Danach erfolgen die gleichen Tests erneut mit Echtdaten. Die Ergebnisse der Tests sind dabei jeweils zu dokumentieren. (Smeets et al. 2019).
- Implementierungsphase: In diesem Zeitraum führt man die zuvor entwickelten RPA-Bots in die Unternehmensumgebung bzw. IT-Architektur ein. Dies kann gesamtheitlich oder bei weitreichenderen Implementierungen auch sukzessiv durchgeführt werden (Smeets et al. 2019; Soybir und Schmidt 2021).
- Performance- und Überwachungsphase: Typischerweise ist anfänglich mit einer hohen Fehleranfälligkeit der RPA-Bots zu rechnen, weshalb diesem Schritt eine hohe Bedeutung zukommt (Smeets et al. 2019). Das Hauptaugenmerk liegt hierbei auf Konformitätsprüfungen und dem Beheben geringfügiger Fehler (Flechsig et al. 2019; Soybir und Schmidt 2021).
- Auswertungsphase: Die im Durchlauf der Performance- und Überwachungsphase gesammelten Daten zur Laufzeit der RPA-Prozesse werden hinsichtlich gesetzter Ziele analysiert. Im Falle einer Änderungsnotwendigkeit erfolgt eine Anpassung (Flechsig et al. 2019; König et al. 2019).

Innerhalb der Performance- und Überwachungsphase benötigt man eine Referenz bzw. einen Maßstab zur Erfolgsmessung. Zur Quantifizierung lassen sich die RPA-Charakteristika im Sinne von erwarteten Ergebnissen durch die Implementierung heranziehen. Hierzu zählen beispielsweise die realisierten Kostensenkungspotenziale, die Fehlersenkungsrate sowie die durchschnittliche Verkürzung der Ausführungsdauer. Voraussetzung hierfür ist eine hinreichende Prozessdokumentation, die als Vergleichsmaßstab für die beobachteten Veränderungen fungiert (Wewerka und Reichert 2020).

Verknüpfung von PM und RPA zur Selektion, Implementierung und Steuerung von RPA-Lösungen

Zur Ausgestaltung eines generischen Vorgehensmodells für ein ganzheitliches RPA-Management lässt sich der BPM-Lebenszyklus aus der einschlägigen Literatur als Grundlage nutzen. Im generierten Vorgehensmodell der Autoren (vgl. Abb. 2) wird impliziert, dass vorab bereits die Auswahl eines Software-Anbieters erfolgt ist. Auf die Notwendigkeit von Mitarbeiterschulungen wird in diesem Vorgehensmodell ebenfalls nicht explizit eingegangen. Eine weitere Prämisse besteht darin, dass die Integration der PM- und RPA-Software in die bestehende IT-Systemlandschaft sowie deren Dokumentation bereits stattgefunden hat.

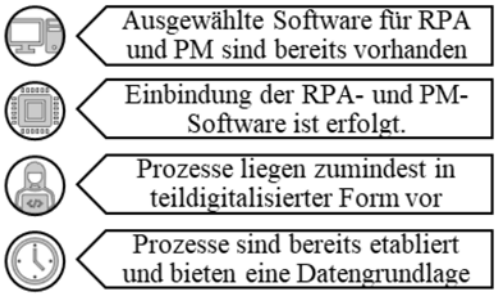
Abb. 2: Vorgehensmodell der RPA-Bot-Einführung mit PM



(Quelle: Eigene Darstellung)

Die Verknüpfung von PM und RPA benötigt als Voraussetzung teilweise digitalisierte Prozesse, d.h. dass nicht alle Schritte von vornherein durch IT-Systeme abgewickelt werden. Es erweist sich ferner als sinnvoll, einen etablierten Prozess auszuwählen, damit man auf eine hinreichende Datengrundlage zurückgreifen kann. Prozesse, die erstmalig eingeführt werden und somit keine Datengrundlage aufweisen, sind demgegenüber auszuschließen. Abb. 3 fasst die beschriebenen Implikationen zusammen.

Abb. 3: Ausgewählte Voraussetzungen und Annahmen für das Vorgehensmodell



(Quelle: Eigene Darstellung)

Definition eines PM-Projektes

Die Definition eines PM-Projekts steht am Anfang des Vorgehensmodells. Ziel ist es, vorab bereits zentrale Fragestellungen zu definieren, welche man mittels der Daten des Event-Logs beantworten will. Damit wird ein grundlegendes Verständnis für die Nutzung der Daten im Event-Log geschaffen (Van der Aalst et al. 2011).

Die aufgelisteten RPA-Bedingungen fungieren als Grundlage für die Definition der relevanten Fragen zur RPA-Bot-Einführung. Der Inhalt eines Event-Logs sollte entsprechend genutzt werden (vgl. Tab. 1), um die RPA-Bedingungen zu untersuchen.

Tab. 1: PM-Lösungsaspekte zur Identifikation eines RPA-Einsatzes

RPA-Bedingungen	PM-Lösungsaspekt
Häufigkeit der Ausführung	Die Häufigkeit, wie oft der Prozess oder eine einzelne Aktivität tatsächlich durchgeführt wurde, ist im Prozessmodell auszuweisen (Celonis Cloud Academia Software 2022).
Standardisierbarkeit des Prozesses	Basierend auf den visualisierten Prozessinstanzen lässt sich ein standardisierter Prozess formen, indem man Gründe für abweichende Verläufe identifiziert und sich auf die am häufigsten vorkommenden Varianten fokussiert (Geyer-Klingeberg et al. 2018).
Ausprägung des Prozessverlaufs	Regeln und Entscheidungen für den Verlauf des Soll-Prozesses lassen sich auf der Grundlage verschiedener Prozessvarianten ableiten (Geyer-Klingeberg et al. 2018).
Menschliche Involvierung	Event-Logs, die als Attribute Mitarbeiter-Rollen oder Stellen ausweisen, ermöglichen die Analyse eines geeigneten Prozessautomatisierungsgrads (Geyer-Klingeberg et al. 2018; Van der Aalst 2016).
Anzahl benötigter IT-Systeme	Analog zur menschlichen Involvierung ermöglicht die Abbildung involvierter IT-Systeme im Event-Log die Evaluierung der Architekturlösung (Geyer- Klingeberg et al. 2018; Van der Aalst 2016).

Nachfolgend muss man die Quellen für die (automatische) Repräsentation der relevanten Daten im Event-Log identifizieren. Es ist möglich, dass mit der Definition der benötigten Daten einzelne Lücken auftreten - im Hinblick darauf, was bereits verfügbar ist und was zukünftig zusätzlich erfasst werden muss. Aus Sicht der Autoren ist es deshalb sinnvoll, zwischen der Identifizierung der relevanten Daten und der Datenextraktion einen zeitlichen Abstand einzubauen. Ziel dieser zeitlichen Pufferung ist es, weitere relevante Daten zu generieren. Abb. 4 zeigt die entsprechende Vorgehensweise.

Abb. 4: Erweiterung des Event-Logs durch relevante RPA-Fragen

(Quelle: Eigene Darstellung)

Datenextraktion

Als Ausgangspunkt für den Beginn der Datenextraktion fungieren die inhaltliche Definition des Event-Logs aus fachlicher Perspektive und die Identifikation der relevanten Quellen. Aus den zugehörigen IT-Systemen werden im Anschluss die relevanten Daten, für die Erstellung eines Event-Logs, in Form eines „Extraction-Transformation-Loading“ (ETL)-Prozesses extrahiert.

Datengewinnung und Datenqualität erweisen sich als elementar: Sind die Daten nicht repräsentativ, unvollständig oder geben einen verzerrten Einblick in die Realität, so hat dies einen negativen Einfluss auf die nachfolgenden Phasen (Van der Aalst 2016). Im Kontext einer RPA-Bot-Einführung kann dieses verfälschte Bild des Prozesses zu einer Fehlentscheidung bzgl. der Tauglichkeit führen. Durch die Implementierung von RPA-Bots in einem ungeeigneten Prozess würden sich die Fehler des Prozesses weiter verstärken und zusätzlich einen zu hohen Aufwand in der Entwicklungsphase der RPA-Bots bewirken (Hofmann et al. 2020; Syed et al. 2020). Um dem Problem einer inadäquaten Datenextraktion entgegenzuwirken, lassen sich kritische Faktoren für die Datengewinnung aus der einschlägigen Literatur heranziehen (Van der Aalst 2016). Tab. 2 beschreibt und konkretisiert diese für den Sachverhalt einer RPA-Bot-Einführung.

Tab. 2: Kritische Faktoren einer Event-Log Datenextraktion

Kritische Faktoren	**RPA-Auswirkung**
Beziehungszusammenhang zwischen Aktivitäten: Sind Aktivitäten über mehrere IT-Systeme verteilt, muss der Zusammenhang der Aktivitäten reflektiert und in einem Event-Log verschmolzen werden (Van der Aalst 2016).	Wird der Zusammenhang der Aktivitäten nicht erkannt, führt dies zu lückenhaften Prozessmodellen.

Zeitbezogene Informationen: Diese nutzt man u.a. um Aktivitäten aus verschiedenen IT-Systemen, in eine Reihenfolge zu bringen (Van der Aalst 2016).	Erfolgt die Erfassung zeitbezogener Aktivitäten vom jeweiligen IT-System nicht oder zu grob (z.B. i.S. einer fehlenden Uhrzeit), kann dies die Richtigkeit des Ablaufs gefährden. (Van der Aalst 2016).
Gewählter Realitätsausschnitt: Event-Logs repräsentieren lediglich einen Ausschnitt der Realität. In Abhängigkeit vom zeitlichen Horizont kann ein unterschiedliches Gesamtbild des Prozesses entstehen (Van der Aalst 2016).	Für RPA sollte ein möglichst großer Zeitraum gewählt werden, um zu überprüfen, ob der Prozessverlauf über einen längeren Zeitraum konstant bleibt (Van der Aalst 2016). Weist ein Prozess häufiger fundamentale Änderungen auf, so ist er weniger für RPA geeignet.
Granularität: IT-Systeme zeichnen Aktivitäten teilweise in Form spezifischer Handlungen auf. Die Herausforderung besteht darin, diese Daten wieder zu übergreifenden Aktivitäten zu gruppieren (Van der Aalst 2016).	Grundsätzlich erweist sich eine spezifische Darstellung der Aktivitäten für die Entwicklung der RPA-Bots als geeignet. Für eine anfängliche Analyse der RPA-Bedingungen ist jedoch davon auszugehen, dass eine feingranulare Darstellung das Erkennen des Gesamtzusammenhangs für die Entscheidungsträger erschwert.

Das Ergebnis der Datenextraktionsphase ist ein vorhandenes, für PM und RPA geeignetes, Event-Log aus den Prozessdaten. Gleichzeitig wird dieses Event-Log den Ansprüchen bzgl. der Qualität der Daten aufgrund einer zugehörigen Validierung gerecht.

Prozessentdeckung

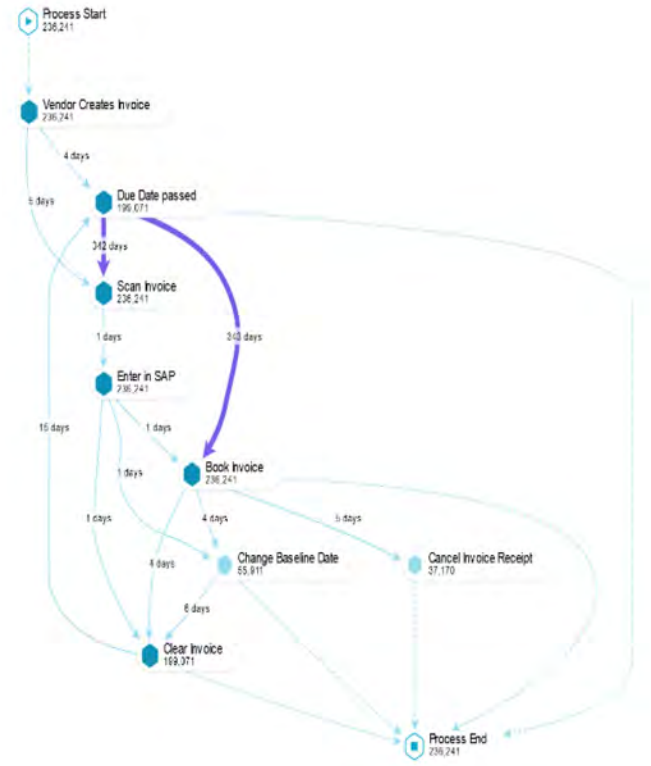
Bevor das Ist-Prozessmodell aus dem Event-Log erstellt wird, muss ein Algorithmus und eine Modellierungssprache für die Prozessentdeckung ausgewählt werden. Oftmals steht die Modellierungssprache bereits mit dem gewählten Algorithmus in Verbindung (Augusto et al. 2019). Der verwendete Algorithmus muss die im Event-Log enthaltenen Gegebenheiten möglichst passgenau darstellen, um das wahrscheinlichste zugrundeliegende Modell aufzuzeigen, welches die Prozessinstanzen des Event-Logs abdeckt. Hieraus erwächst jedoch die Gefahr, dass das Modell zu stark verallgemeinert und somit mehr als die vorkommenden Prozessinstanzen visualisiert, bzw. Ablauffolgen generiert, die nicht aus den Daten des Event-Logs resultieren. Als weitere maßgebende Dimension des erstellten Modells ist daher die Präzision zu definieren. Insbesondere das Gleichgewicht zwischen Präzision und Generalisierung erweist sich in dieser Phase als zentrale Aufgabenstellung (Van der Aalst 2016). Damit wird ein für RPA hinreichendes, analysierbares normatives Ist-Prozessmodell erzeugt.

Modellierungssprachen besitzen aufgrund ihrer Syntax Limitationen im Hinblick auf die Darstellbarkeit (Gadatsch 2020). Entsprechend muss man sicherstellen, dass die ausgewählte Modellierungssprache nicht nur verständlich für die Beteiligten ist, sondern auch den Inhalt hinreichend darstellen kann, um die Fragestellungen zu beantworten (Van der Aalst 2016).

Nach der Auswahl ist die Grundlage für die Erstellung eines kontrollflussorientierten Prozessmodells geschaffen. Abb. 5 zeigt ein Beispiel für ein generiertes Prozessmodell auf. Dabei wird als Annahme getroffen, dass die Aktivitäten im Prozessmodell atomar repräsentiert sind, d.h. dass jede Aktivität aus einem zeitlichen Event besteht, ohne dass Start und Ende einer Aktivität separat visualisiert werden (Van der Aalst 2016). Durch die atomare Repräsentation wird ein hinreichend detailliertes Prozessmodell erzeugt, wobei entweder der Prozess als Ganzes oder auch einzelne Aktivitäten sich auf RPA-Eignung analysieren lassen. Existieren bereits Prozessdokumentationen, kann man diese alternativ mit den Daten des Event-Logs vergleichen.

Um das vorliegende Prozessmodell zu validieren, eignet sich die Konformitätsprüfung. Bei dieser Konformitätsprüfung wird der erwartete Prozessablauf des Soll-Modells mit dem tatsächlichen Prozessablauf in den Varianten verglichen. Hierzu analysiert ein Process-Mining-Tool die Übereinstimmungen bzw. Abweichungen zwischen einem vorhandenen Soll-Modell und den Varianten im Ist-Zustand. Ausgegeben werden Diagnosedaten über die Diskrepanzen zwischen Event-Log und Soll-Modell (Van der Aalst 2011). Sind zwischen Soll-Modell und Event-Log Abweichungen vorhanden, gilt es zu untersuchen, wodurch diese entstehen. Die Unterschiede werden darauf hin im Soll-Modell angepasst und manuell, in Form einer Prozesserweiterung, eingepflegt. Das Ergebnis lässt sich mit der Prozessentdeckung eines normativen Ist-Prozessmodells gleichsetzen, welches eine hinreichende Analyse der RPA-Bedingungen ermöglicht.

Abb. 5: Darstellung eines kontrollflussorientierten Prozessmodells auf Basis einer Prozessentdeckung durch PM

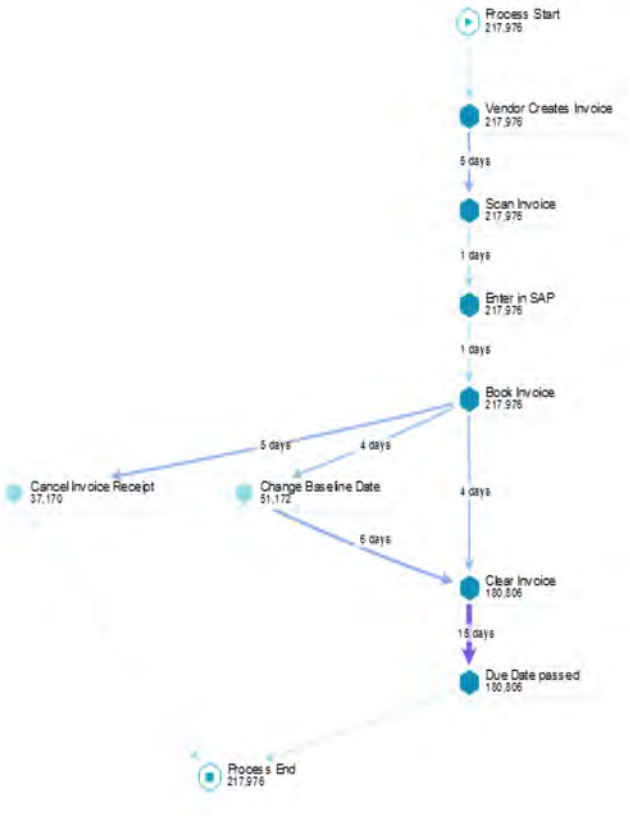


(Quelle: Celonis Cloud Academia Software (2022))

Das Beispiel in Abb. 5 zeigt ein Prozessmodell mit Varianten, welches das Resultat einer ersten Prozessentdeckung sein kann. Während mit RPA prinzipiell eine Automatisierungsmöglichkeit gegeben ist, um beliebig viele Prozessvarianten und Ausnahmen zu realisieren, liegt es aus kostentechnischen Gründen nahe, sich auf die Haupt-Prozessvarianten zu fokussieren (Syed et al. 2020). Zur Identifizierung von primäreren Prozessverläufen in Abb. 5 soll das zuvor vorgestellte Pareto-Prinzip aufgegriffen werden: Aus der Fokussierung auf die am häufigsten vorkommenden Prozessvarianten ergibt sich beispielsweise ein Prozessmodell wie in Abb. 6.

Abb. 6 deckt 92% aller im Event-Log identifizierten Fälle ab, stellt dabei allerdings lediglich 3 von 112 möglichen Prozessvarianten dar (Celonis Cloud Academia Software 2022). Diese Fokussierung der Prozessverläufe soll den ersten Schritt zur Prüfung des Prozesses, nach Erstellung des Modells, beispielhaft darstellen. Ziel ist es, einen standardisierten Prozessverlauf aus historischen Daten zu generieren, indem man abweichende Varianten isoliert. Im Beispiel der Abb. 6 ist der standardisierte Prozessverlauf durch dunkelblaue Hexagone gekennzeichnet. Parallel hierzu sind die Ursachen für die ungewünschten Prozessabweichungen zu ermitteln, um diese in Zukunft zu verhindern. Kann dies nicht erfolgen, sind diese Ausnahmefälle bei der RPA-Bot-Programmierung zu berücksichtigen. Durch den Fokus auf Standardisierung soll der Aufwand für eine potenziellen RPA-Bot-Programmierung reduziert und gleichzeitig die Analyse der verbleibenden Bedingungen eingegrenzt werden.

Abb. 6: Darstellung eines fokussierten kontrollflussorientieren Prozessmodells



(Quelle: Celonis Cloud Academia Software (2022))

Nach der Standardisierung des Prozesses gilt es zu untersuchen, auf welchen Regeln und Bedingungen die Aktivitäten in den verbleibenden Prozessvarianten basieren. Zuletzt erfolgt eine Analyse des verbleibenden Prozessmodells im Hinblick auf die restlichen identifizierten RPA-Bedingungen. Um diese Analyse zu erleichtern, lassen sich weitere Perspektiven des Prozessmodelles nutzen (vgl. hierzu Abb. 7).

Das Resultat besteht aus einer datenbasierten, standardisierten Darstellung des zu untersuchenden Prozesses, wodurch auch eine vereinheitlichte Prozessdokumentation entsteht. Je nach vorliegender Ausprägung des Prozesses lässt sich eine systematische Entscheidung für oder gegen die Entwicklung von zugehörigen RPA-Bots herbeiführen.

Abb. 7: Erweiterung des Ist-Prozessmodells um ausgewählte Perspektiven

(Quelle: Eigene Darstellung, angelehnt an Celonis Cloud Academia Software (2022))

RPA- Bot Entwicklung und Tests

Ist die Implementierung eines RPA-Bots beschlossen, lässt sich dieser z.B. durch die Aufnahme der in Abb. 6 visualisierten Aktivitäten erstellen. In einer kontrollierten Umgebung und mittels einer geeigneten Aufzeichnungssoftware kann man die notwendigen Schritte in UI-Logs sammeln. Diese beinhalten eine zeitlich geordnete Abfolge von Handlungen, die sich nicht weiter spezifizieren lassen und den Prozess auf einem Ein- und Ausgabelevel darstellen. Die im UI-Log aufgezeichneten Handlungen werden zunächst bereinigt und segmentiert, so dass erkennbar ist, welche Handlungen eine übergeordnete Aktivität ausmachen. Weiterhin erfolgt eine Unterteilung der Aktivitäten in Segmente, deren Kombination den gewünschten Prozess repräsentiert (Agostinelli et al. 2020; Leno et al. 2021).

Durch die Kombination des Prozessentdeckungsansatzes aus dem PM mit der Aufnahme im UI-Log lassen sich transparente Workflowmodelle generieren. Es empfiehlt sich, zunächst einen einfachen RPA-Bot zu erstellen, der den Standardfall abbildet, und in iterativen Schritten Ausnahmefälle hinzuzufügen. Ggf. ist die Anwendung des Modularisierungsprinzips vorteilhaft: Diese Vorgehensweise bietet sich insbesondere an, wenn der gesamte Prozess viele Ausnahmen beinhaltet oder man ihn nicht kontinuierlich automatisieren kann (Noppen et al. 2020; Smeets et al. 2019). Im Kontext von Abb. 6 würde dies bedeuten, dass man zunächst plant, den definierten und standardisierten Prozess vollständig und kontinuierlich durch RPA-Bots umzusetzen. Stellt sich jedoch heraus, dass der Prozess nicht kontinuierlich automatisierbar ist, setzt man einzelne Aktivitäten separat durch überwachte RPA-Bots um.

Vor der eigentlichen Implementierung von RPA-Bots in die IT-Systemlandschaft, empfiehlt es sich, eine vorausgehende Testphase durchzuführen. Damit soll sichergestellt werden, dass die RPA-Bots über die gewünschte Funktionalität verfügen und gleichzeitig fehlerfrei ausführbar sind (Smeets et al. 2019). Hierbei ist zwischen technischen Tests und Umgebungstests zu differenzieren: Technische Tests führen die Entwickler bereits während der RPA-Bot Erstellung aus. Dabei überprüfen sie, ob die verwendete Logik fehlerfrei funktioniert. Der Fokus von Umgebungstests liegt hingegen auf der Interaktion zwischen RPA-Bots und der späteren Umgebung im Rahmen der Datenverarbeitung (Smeets et al. 2019). Die Durchführung des Umgebungstests erfolgt zunächst mit Hilfe von Testdaten in einer gesicherten Umgebung. Echtdaten sind erst zu verwenden, sobald die Resultate den gesteckten Erwartungen entsprechen (Smeets et al. 2019). Die Ergebnisse der Tests sind im Hinblick auf die Performance der RPA-Bots zu dokumentieren.

RPA-Bot Implementierung und -Überwachung

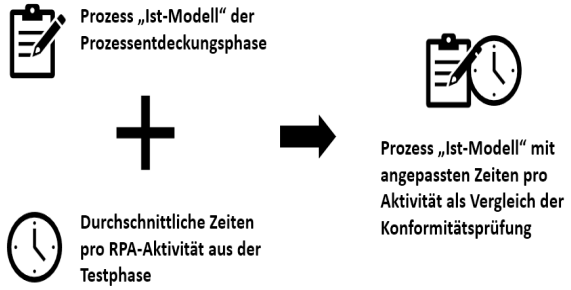
In dieser Phase werden die funktionalen, getesteten RPA-Module produktiv in die IT-Systemlandschaft eingeführt. Zuvor erfolgt eine Anpassung der RPA-Software an die verwendeten Anwendungssysteme, welche eine spätere Verknüpfung zwischen RPA-Bot Daten und Prozessdaten ermöglicht (Egger et al. 2020).

Für den Zeitraum direkt nach der Implementierung, ist es wahrscheinlich, dass unbekannte Fehler und unerwartete Ausnahmen auftreten, welche in den Testphasen nicht identifiziert wurden (Smeets et al. 2019). Dies erfordert eine strenge Überwachung der RPA-Bots (Smeets et al. 2019).

Durch PM lässt diese sich diese Aktivität unterstützen: Die eingesetzten IT-Systeme zeichnen Daten zur Ausführung der Bots in Echtzeit bzw. Echtzeit-nah auf (Van der Aalst 2016). Mit Hilfe der Konformitätsprüfung kann man eine automatische Echtzeitüberwachung der RPA-Bots realisieren. Das aus der Phase der Prozessentdeckung resultierende Prozessmodell fungiert in diesem Zusammenhang Referenzmodell. Ein reiner Kontrollflussabgleich ist dabei allerdings nicht ausreichend, um RPA-Fehler zu identifizieren: Dieser identifiziert lediglich primär abgebrochene oder auch angehaltene Prozesse (Behrens 2014; Deckard 2017; Smeets et al. 2019; Syed et al. 2020; Van der Aalst 2021). Um diese Problematik zu umgehen, ist das Referenzmodell um eine zeitliche Perspektive zu ergänzen. Fehler und Abweichungen eines RPA-Bots lassen sich dann durch eine Abweichung der durchschnittlichen Bearbeitungsdauer einer Aktivität identifizieren. Die durchschnittliche Bear-

beitungszeit ist daher als Grenzwert einer Aktivität zu definieren. Die ursprünglichen zeitbezogenen Informationen des Referenzmodells sind als jeweilige Grenzwerte i.d.R. jedoch nicht geeignet, da diese auf manueller Arbeit basieren. Um eine realistische Dauer zu bestimmen, erweist es sich vielmehr als empfehlenswert, auf die Dokumentation in der Testphase zurückzugreifen, da diese Zeiten auf Nutzung von RPA-Bots basieren. Die Ergänzung des Referenzmodells um die Zeiten der RPA-Ausführung soll im Sinne der Prozesserweiterung im PM stattfinden (vgl. Abb. 8)

Abb. 8: Grundlagen einer Echtzeit Konformitätsprüfung für RPA-Bots



(Quelle: Eigene Darstellung)

Weiterhin ist zu definieren, wie man die durchschnittliche Bearbeitungsdauer der RPA-Aktivitäten im ausführenden Prozess messen will. Es besteht die Möglichkeit, in einem Event-Log je Aktivität den Startzeitpunkt, den Endzeitpunkt oder beides gemeinsam zu verfolgen (Van der Aalst 2016). Für die Ermittlung einer durchschnittlichen Bearbeitungsdauer muss mindestens der Startzeitpunkt der ausgewählten Aktivität sowie der Startzeitpunkt der nachfolgenden vorliegen, um so aus der Differenz der beiden die Dauer der ausgewählten Aktivität zu bestimmen. Besser erscheint es jedoch, innerhalb einer Aktivität sowohl den Start- als auch den Endzeitpunkt zu bestimmen. Dies erweist sich vor allem mit Blick auf überwachte RPA-Bots, die nur einzelne Aktivitäten automatisieren, von Vorteil. Würde demgegenüber auf eine RPA-Aktivität eine manuelle Aktivität folgen, die nicht direkt gestartet wird, würde eine Berechnung der Dauer aus RPA-Aktivität und folgender manueller Aktivität einen unbrauchbaren Wert generieren. Am Ende dieser Phase sind die RPA-Bots erfolgreich in die Systemlandschaft eingegliedert und werden fortan durch eine PM-Lösung laufend überwacht.

Prozess- und RPA-Bot Auswertung

Mit der Ausführung der RPA-Bots werden weitere Prozessinstanzen generiert. Ziel dieser Phase ist es, die anwachsende Datenbasis zu nutzen, um die Auswirkungen der RPA-Bots auf den Prozess zu bewerten.

Als Grundlage für diese Auswertung fungiert die Erstellung eines neuen Prozessmodells, das man auf Basis der Prozessentdeckung im PM entwickelt hat und das verschiedene Prozessperspektiven aufzeigt. Um die Performance der RPA-Bots bewerten zu können, sind die gesetzten Erwartungen und Ziele zu Beginn der RPA-Einführung für die Bewertung heranzuziehen und in Form von Kennzahlen zu operationalisieren. Die nachfolgenden Beispiele sollen diese Vorgehensweise zur Aufstellung entsprechender Kennzahlen veranschaulichen:

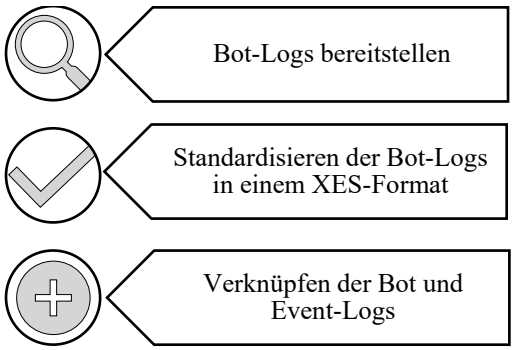
- Opportunitätserlöse (i.S.v. wegfallenden Kosten bei den menschlichen Bearbeitern): Durchschnittliche Prozessausführungszeit vor der RPA-Einführung mal der Anzahl an Ausführungen mal dem Kostensatz je Mitarbeiter.
- Fehlersenkung: Vergleich der prozentualen Anzahl an Fehlversuchen bei den Gesamtdurchläufen.
- Reduktion der Ausführungsdauer: Messung der durchschnittlichen Prozessausführungszeit vor und nach der RPA-Einführung.

Der Einsatz von RPA-Bots lässt sich auf der Basis sämtlicher Prozessinstanzen als Kombination des Event-Logs mit den Aktivitätsdaten von RPA-Bots sowohl im Einzelfall als auch auf aggregierter Ebene evaluieren: Eine Untersuchung des aggregierten Prozesses soll einen ersten Einblick in die allgemeinen Auswirkungen des RPA-Einsatzes ermöglichen. Die Untersuchung einzelner Prozessinstanzen vermag es aufzuzeigen, unter welchen spezifischen Umständen und Bedingungen der Einsatz eines RPA-Bots nicht funktioniert. Daraus lässt sich ableiten, welche Maßnahmen getroffen werden müssen, um ein solches Versagen zu verhindern. Zur Auswertung der Aktivität eines RPA-Bots in einzelnen Prozessinstanzen, sind die im Event-Log erfassten Prozessdaten um Daten über die Handlungen des RPA-Bots zu ergänzen. Die Daten der RPA-Bots werden dabei in Bot-Logs gespeichert, welche man hierfür extrahiert.

Bot-Logs lassen sich, ähnlich wie Event-Logs, als datenbasierte Darstellung historischer Prozessdurchläufe charakterisieren. Der zentrale Unterschied liegt in den gespeicherten Daten: Bot-Logs beinhalten Informationen zu den Handlungen der im Prozess eingesetzten RPA-Bots (Egger et al. 2020). Event-Logs und Bot-Logs sind daher in ihrem Inhalt identisch, stellen jedoch eine jeweils andere Perspektive auf den Prozess dar (Egger et al. 2020). Die Aufzeichnung der Bot-Logs geschieht meist automatisch durch die RPA-Software (Egger et al. 2020).

Im Anschluss an die Extraktion der relevanten Bot-Logs muss man diese, vor der Verknüpfung mit einem Event-Log, standardisieren. Mit der Standardisierung werden die Daten innerhalb des Bot-Logs in einem einheitlichen XES-Format definiert. Das Ziel der dann möglichen Verknüpfung besteht darin, ein Event-Log zu erstellen in dem jeder übergeordneten Aktivität eine Summe von komplementären RPA-Bot Handlungen zugeordnet ist (Egger et al. 2020). (vgl. Abb. 9).

Abb. 9: Erstellung eines Event-Logs mit RPA-Bot Handlungen



(Quelle: Eigene Darstellung, angelehnt an Egger at al. (2020) S.54 ff.)

Aus dem resultierenden Gesamt-Event-Log kann ein neues Prozessmodell erzeugt werden (Egger et al. 2020). Abb. 10 zeigt anhand des gewählten Beispiels das neu entstandene Soll-Modell.

Abb. 10: Erweiterung des Prozessmodells durch RPA-Bot Informationen

(Quelle: Eigene Darstellung, angelehnt an Celonis Cloud Academia Software (2022))

Abweichungen von den geplanten Zielen lassen sich so durchgängig und detailliert analysieren. Beispielsweise kann man eine Untersuchung einzelner Prozessinstanzen nutzen, um die Auswirkungen auftretender Probleme beim RPA-Bot Einsatz auf die zugehörige Dauer zu verstehen. Hierzu erfolgt der automatische Abgleich der tatsächlichen Dauer einer Prozessinstanz mit derjenigen aus dem Soll-Modell (Egger et al. 2020). Im Kontext einer überwachten RPA-Lösung kann das Zusammenspiel zwischen Mensch und RPA-Bot von einer Auswertung spezifischer Prozessinstanzen profitieren: Bei auftretenden Problemen lässt sich nach der Identifikation der entsprechenden Prozessinstanzen eine Ursache-Wirkungs-Analyse durchführen. Diese wiederum bildet die Grundlage für Optimierungsmaßnahmen (Egger et al. 2020).

Mit diesem Schritt endet das Vorgehensmodell. Da dieses Konzept von den Autoren als eine Weiterentwicklung des BPM-Lebenszyklus angesehen wird, impliziert dies eine iterative Bearbeitung: Basierend auf den zusätzlich gewonnenen Daten und daraus resultierenden offenen Fragen in der Auswertungsphase lassen sich im zeitlichen Verlauf weitere Verbesserungsmaßnahmen durch den Einsatz von RPA-Bots vornehmen. Die Auswertungsphase überlappt dabei mit den PM Projektphasen Definition, Datenextraktion und Prozessentdeckung. Entsprechend lässt sich, analog zum BPM-Vorgehensmodell, von einem Lebenszyklus-Konzept sprechen.

Zusammenfassung und Ausblick

Ziel des vorliegenden Artikels ist es, ein grundlegendes Verständnis und Konzept für den Einsatz und das Zusammenspiel von PM und RPA im Rahmen des BPM zu erstellen. Als Ausgangspunkt fungierten zwei Fragestellungen.

Zunächst geht es darum, welchen Nutzen die Anwendung von PM für die Identifikation, Ausgestaltung und Überwachung von RPA-Lösungen im Rahmen des BPM bringt. Der PM-Ansatz ist in der Lage, den BPM-Lebenszyklus zu ergänzen und in einer digitalen Umgebung zu optimieren. Der Einsatz der RPA-Technologie bietet die Möglichkeit, Geschäftsprozessoptimierung durch Automatisierung weiter auszubauen. Eine Kombination erweist sich aus Sicht der Verfasser sinnvoll innerhalb der Schritte Prozessentdeckung und -analyse, der RPA-Bot Entwicklung sowie der Überwachung und Auswertung: Hierdurch eröffnet sich die Möglichkeit, einen objektivierten Einblick in die vorliegende Prozesssituation zu gewinnen. Repetitive, regelbasierte Arbeiten lassen sich durch RPA-Bot Entwicklung minimieren oder zumindest reduzieren. Innerhalb der Auswertungsphase kann die Auswirkung der eingesetzten RPA-Bots detailliert auf der Ebene einzelner Prozessinstanzen analysiert werden.

Bei der zweiten Fragestellung geht es um die Definition von Phasen für eine PM-unterstützte RPA-Bot-Einführung. Abb. 2 gibt einen Überblick über das von den Autoren konzipierte Vorgehensmodell. Grundsätzlich entsprechen die Phasen der allgemeinen Vorgehensweise im BPM-Lebenszyklus. Die Datenextraktionsphase stellt eine Erweiterung dar, da sie für das PM von zentraler Bedeutung ist. Weiterhin erweitert das Vorgehensmodell den BPM-Lebenszyklus um die Schritte RPA-Bot-Entwicklung und -Tests auf Basis des vorgestellten Vorgehens zur RPA-Bot-Einführung.

Der vorgestellte Ansatz ist konzeptioneller Natur und deshalb in einem weiteren Schritt zu validieren. Erste Aktivitäten hierzu können prototypische Anwendungen auf der Basis von Fallstudien oder vergleichbare Aktivitäten sein. Aus Sicht der Autoren lohnt sich ein zugehöriger Diskurs in der angewandten Forschung, da die Kombination von RPA und PM auch aus praxeologischer Sicht in Zukunft Relevanz besitzen wird.

Abschließend wird deshalb auf die zu erwartende Weiterentwicklung von RPA und PM eingegangen. Ein Trend, der aktuell die Welt der IT-Lösungen im Bereich der Unternehmen prägt, ist der Einsatz von Künstlicher Intelligenz (KI) durch maschinelles Lernen (ML) (Gartner 2021). Auch für RPA und PM beinhaltet eine solche Kombination erhebliche Chancenpotenziale (Santos et al. 2020; Van der Aalst 2016). Verknüpft man RPA mit KI lassen sich die explizit formulierten RPA-Regeln erweitern. ML ermöglicht das Erkennen und Erlernen unterschiedlicher Muster, um so situationsabhängige Entscheidungen zu treffen, indem RPA-Bots sich selbst anpassen (Schmitz et al. 2019; Van der Aalst et al. 2018). Dies fördert eine flexiblere und ganzheitliche Prozessautomatisierung anstellen von Einzelmaßnahmen (Santos et al. 2020; Schmitz et al. 2018).

Eine solche Entwicklung bedeutet jedoch nicht, dass die hier vorgestellten Ergebnisse an Relevanz verlieren. Die Identifikation sinnvoll zu optimierender Prozesse lässt sich vielmehr durch eine Kombination aus RPA und KI ausweiten. Prozesse, die unregelmäßige Verläufe aufweisen und sich durch eine Vielzahl von Ausnahmen charakterisieren lassen, können zukünftig ebenfalls für die Evaluierung eines RPA-Einsatzes herangezogen werden.

Für die vorgestellte Echtzeit-Konformitätsprüfung erscheint es naheliegend, den Fokus des Referenzmodells von einem präzisen und passgenauen Referenzmodell hin zu einem allgemeineren Referenzmodell zu verschieben, um so die Möglichkeiten eines erweiterten Entscheidungsspielraums für die RPA-Bots adäquat zu nutzen. Weiterhin nehmen auch die Möglichkeiten der Konformitätsprüfung zu: Anstatt auf Probleme zu reagieren, lassen sich Annahmen über den weiteren Verlauf des Prozesses, basierend auf historischen Daten, treffen. Die Konformitätsprüfung soll auf dieser Basis bereits einschreiten, bevor ein Problem entsteht und Vorschläge bzgl. zu treffender Maßnahmen erstellen (Van der Aalst 2016). Die im Artikel vorgestellte Auswertung kann eine Kombination von RPA und KI unterstützen, indem sie Unsicherheiten im Umgang mit einer KI bzw. mit ML entgegensteuert: Durch die Visualisierung der Prozessdaten vergangener Prozesse lassen sich Prozessverläufe und KI-basierte Entscheidungen sowie deren Auswirkungen bis zu einzelnen Prozessschritten verfolgen und verstehen. Entsprechend weist die Zukunft auf eine flexiblere Anwendung der RPA-Technologie hin.

LITERATUR

- Aguirre, Santiago; Rodriguez, Alejandro (2017): Automation of a Business Process Using Robotic Process Automation (RPA): A Case Study. In: Workshop on Engineering Applications: Springer, Cham, S. 65–71.
- Agostinelli, Simone; Lupia, Marco; Marrella, Andrea; Mecella, Massimo (2020): Automated Generation of Executable RPA Scripts from User Interface Logs. In: International Conference on Business Process Management: Springer, Cham, S. 116–131.
- Dumas, Marlon; La Rosa, Marcello; Mendling, Jan; Reijers, Hajo A. Reijers (2013): Fundamentals of Business Process Management. Berlin, Heidelberg: Springer Berlin Heidelberg.
- Flechsig, Christian; Lohmer, Jacob; Lasch, Rainer (2019): Realizing the Full Potential of Robotic Process Automation Through a Combination with BPM. In: Logistics Management: Springer, Cham, S. 104–119.
- Gadatsch, Andreas (2020): Grundkurs Geschäftsprozess-Management. Wiesbaden: Springer Vieweg.
- Geyer-Klingeberg, Jerome; Nakladal, Janina; Baldauf, Fabian; Veit, Fabian (2018): Process Mining and Robotic Process Automation: A Perfect Match. In: Industry Track Session.
- Grisold, Thomas; Mendling, Jan; Otto, Markus; vom Brocke, Jan (2021): Adoption, use and management of process mining in practice. In: Business Process Management Journal 27 (2), S. 369–387.
- Hofmann, Peter; Samp, Caroline; Urbach, Nils (2020): Robotic process automation. In: Electron Markets 30 (1), S. 99–106.
- Koch, Christina; Fedtke Stephen (2020): Robotic Process Automation. Berlin, Heidelberg: Springer Vieweg.
- König, Maximilian; Bein, Leon; Nikaj, Adriatik; Weske, Mathias (2020): Integrating Robotic Process Automation into Business Process Management. In: Aleksandre Asatiani (Hg.): Business Process Management. BPM 2020 Blockchain and RPA Forum, Seville, Spain, September 13-18, 2020, Proceedings. Unter Mitarbeit von José María García, Nina Helander, Andrés Jiménez-Ramírez, Agnes Koschmider, Jan Mendling, Giovanni Meroni und Hajo A. Reijers. Cham: Springer International Publishing AG (Lecture Notes in Business Information Processing Ser, v.393), S. 132–146.

- Langmann, C. und D. Turi. 2020. "Robotic Process Automation (RPA) – Digitalisierung und Automatisierung von Prozessen: Voraussetzungen, Funktionsweise und Implementierung am Beispiel des Controllings und Rechnungswesens". Springer Gabler, Wiesbaden.
- Leno, Volodymyr; Polyvyanyy, Artem; Dumas, Marlon; La Rosa, Marcello; Maggi, Fabrizio Maria (2021): Robotic Process Mining: Vision and Challenges. In: Bus Inf Syst Eng 63 (3), S. 301–314.
- Noppen, Philip; Beerepoot, Iris; van Weerd, Inge de; Jonker, Mathieu; Reijers, Hajo A. (2020): How to Keep RPA Maintainable? In: International Conference on Business Process Management: Springer, Cham, S. 453–470.
- Peters, Ralf; Nauroth, Markus H. (2019): Process-mining. Geschäftsprozesse: smart, schnell und einfach. Wiesbaden: Springer Gabler (essentials).
- Rozinat, Anne; Medeiros, Ana Karla Alves de; Günther, Christian W.; Weijters, A. J. M. M.; van der Aalst, Wil M. P. (2007): The Need for a Process Mining Evaluation Framework in Research and Practice. In: International Conference on Business Process Management: Springer, Berlin, Heidelberg, S. 84–89.
- Santos, Filipa; Pereira, Rúben; Vasconcelos, José Braga (2020): Toward robotic process automation implementation: an end-to-end perspective. In: Business Process Management Journal 26 (2), S. 405–420.
- Schmitz, M., Stummer, C., Gerke, M. (2019). Smart Automation as Enabler of Digitalization? A Review of RPA/AI Potential and Barriers to Its Realization. In: Krüssel, P. (Hg.) Future Telco. Management for Professionals. Springer, Cham.
- Smeets, Mario; Erhard, Ralph; Kaußler, Thomas (Hg.) (2019): Robotic Process Automation (RPA) in der Finanzwirtschaft: Technologie – Implementierung – Erfolgsfaktoren für Entscheider und Anwender. Wiesbaden: Springer Fachmedien Wiesbaden.
- Soybir, Sefa; Schmidt, Christopher (2021): Project Management and RPA. In: The Digital Journey of Banking and Insurance, Volume I: Palgrave Macmillan, Cham, S. 289–305.
- Syed, Rehan; Suriadi, Suriadi; Adams, Michael; Bandara, Wasana; Leemans, Sander J.J.; Ouyang, Chun; Hofstede, Arthur H.M. ter; van de Weerd, Inge; Wynn, Moe Thandar; Reijers, Hajo A. (2020): Robotic Process Automation: Contemporary themes and challenges. In: Computers in Industry 115.
- van der Aalst, Wil (2012): Process Mining: Overview and Opportunities. In: ACM Trans. Manage. Inf. Syst. 3 (2), S. 1–17.
- van der Aalst, Wil (2016): Process Mining. Data Science in Action. 2. 2nd ed. 2016. Berlin, Heidelberg: Springer Berlin Heidelberg.
- van der Aalst, Wil (2021): 12 Process mining and RPA. In: Robotic Process Automation: De Gruyter, S. 223–240.
- van der Aalst, Wil; Adriansyah, Arya; Medeiros, Ana Karla Alves de; Arcieri, Franco; Baier, Thomas; Blickle, Tobias et al. (2011): Process Mining Manifesto. In: International Conference on Business Process Management: Springer, Berlin, Heidelberg, S. 169–194.
- van der Aalst, Wil M. P.; Bichler, Martin; Heinzl, Armin (2018): Robotic Process Automation. In: Bus Inf Syst Eng 60 (4), S. 269–272.
- Wewerka, Judith; Reichert, Manfred (2020): Towards Quantifying the Effects of Robotic Process Automation. In: 2020 IEEE 24th International Enterprise Distributed Object Computing Workshop (EDOCW): IEEE.

Weblinks

- Behrens, Katie (2014): Handling Errors: Can You Trust a Robot? URL: https://www.uipath.com/blog/rpa/handling-errors-can-you-trust-a-robot?utm_content=resources.base.vn/hr/ki-nguyen-moi-cua-nganh-nhan-su---hr-4.0-115?utm_content
- Celonis (2022): Celonis Cloud Academia Software – Accounts Payable Demo URL: https://eu-2.celonis.cloud/
- Deckard, Mina (2017): RPA In Our Own Words: Managing Unstructured Data. URL: https://www.uipath.com/blog/rpa/rpa-in-our-own-words-managing-unstructured-data
- Gartner (2020): Gartner Says Worldwide Robotic Process Automation Software Revenue to Reach Nearly $2 Billion in 2021. URL: https://www.gartner.com/en/newsroom/press-releases/2020-09-21-gartner-says-worldwide-robotic-process-automation-software-revenue-to-reach-nearly-2-billion-in-2021
- Gartner (2021): Gartner Forecasts Worldwide Artificial Intelligence Software Market to Reach $62 Billion in

2022. URL: https://www.gartner.com/en/news-room/press-releases/2021-11-22-gartner-forecasts-world-wide-artificial-intelligence-software-market-to-reach-62-billion-in-2022

OPTIMIZATION OF INTERNAL LOGISTICS USING A COMBINED BPMN AND SIMULATION APPROACH

Maximilian Wuennenberg
Benjamin Wegerich
Johannes Fottner
Chair of Materials Handling, Material Flow, Logistics
TUM School of Engineering and Design
Technical University of Munich
Boltzmannstraße 15, 85748 Garching bei Muenchen, Germany
E-mail: max.wuennenberg@tum.de
E-mail: benjamin.wegerich@tum.de
E-mail: j.fottner@tum.de

KEYWORDS

BPMN, Internal Logistics, Material Flow Systems, Model-based Systems Engineering, Planning, Simulation.

ABSTRACT

The optimization of material flow systems requires a profound understanding of the underlying processes. Business Process Model and Notation (BPMN) is an established way of creating a process model that allows an interdisciplinary analysis and optimization. Quantitative exploration of systems using discrete-event simulation can help to enrich these insights. For that reason, this paper introduces a combined BPMN-simulation approach that connects the advantages of both modeling frameworks. By synthesizing systems from generic modules, a comprehensive yet structured optimization process chain is developed. A case study evaluation based on key metrics for material flow operations proves the applicability of the methodology.

INTRODUCTION

Creating a virtual model of a material flow (MF) system promises higher availabilities and shorter throughput times. To achieve this, the model needs to contain data from various sources in the MF domain (e. g. stacker cranes or conveyor belts). However, the application of this approach in real-world systems is often impaired by complex and distributed processes in heterogenous organizations (Pires et. al. 2019). The necessary transparency can be generated by defining and implementing a proper process visualization. Business Process Model and Notation (BPMN) is a widespread standard for this task since the created models can be understood by experts from various domains. However, the focus of this modeling language primarily lies on administrative processes. (Muehlen and Recker 2013) The most common process type modeled with it is the information flow. In theory, BPMN can also be used to represent MFs as well as the movement of workers, forklift trucks, and other mobile resources. Common notations in this domain, such as flow charts or value stream mapping, often cannot meet the requirements for these particular use cases. Flow charts, for instance, suffer from a low level of standardization and development, and are unsuitable for depicting more complex process properties. Value stream mapping, on the other hand, does not depict sequence flows as accurately and in detail as BPMN, for example because events are not mapped (Garcia et. al. 2012). That makes it difficult to subsequently create a discrete-event simulation (DES). Additionally, this notation focuses heavily on manufacturing settings and is therefore difficult to understand for users outside of the domain of production management (Forno et. al. 2014).

But although BPMN possesses several properties that make it attractive to be applied to MF processes, there are only a few examples where it is actually used in this field (Zor and Leymann 2011). One potential reason for this is the lack of a scientific foundation for modeling strategies in the internal logistics domain (Robinson 2006). A generalized and well-structured approach that considers the specific characteristics of both modeling frameworks can offer guidance for practitioners and help them to model MF systems in a predictable amount of time.

Central Concepts & Related Research

When modeling administrative processes with BPMN, the sequence flow (SF) is used to show the chronological order in which events and activities take place. However, MF systems are characterized by a greater variety of different process types, which raises the question of how these can be mapped in BPMN. An intuitive option is to use the SF as a representation of the MF, resulting in a material-oriented model. Alternatively, the model can be resource-oriented, meaning that the SF represents the movements of mobile resources, e. g. workers. A third option is the addition of MF-specific elements to the BPMN syntax (Zor and Leymann 2011). While this somewhat reduces the ambiguity of the SF, it also makes the models more complex and limits the choice of modeling software. That is why the BPMN models shown in this article use the standard BPMN syntax and are either material- or resource-oriented.

Another challenge arises from the fact that the activities in MF systems are usually object-constrained, meaning that their execution requires the availability of certain objects (Wagner 2021). Depicting these relationships between objects and activities is essential for the modeling of MF systems, but it is also beyond the possibilities of the BPMN syntax. The different approaches discussed above, being exclusively visual, do not address this problem. That is why this article proposes the use of an additional tool in the form of DES. Being widely used in the MF domain, DES includes various possibilities to represent object-constrained activities. Since BPMN does not cover these activities, DES has the potential to work as a complement for BPMN. By synthesizing the two modeling approaches, the high variety of MF processes can be represented. The question of how those two tools can be combined has already been partly explored outside the manufacturing domain, e. g. by extending the DES framework to include tools for business process modeling (Wagner et. al. 2009). However, as of today, this is not yet supported by existing DES software. The reverse approach, adding DES elements to the BPMN syntax, is developed as a BPMN variation called "DPMN" (Guizzardi and Wagner 2011, Wagner 2018 & 2021). While this new modeling language can depict object-constrained activities in a qualitative way, existing DES software is still required to perform simulation experiments and generate quantitative results. Moreover, it is uncertain to what extent DPMN is applicable to MF systems.

Summing up, the ever-recurring goal of an MF operator to optimize the system could be met more effectively by creating support in the shape of a proper system model. It seems to be a promising approach to combine BPMN and DES using the existing modeling syntax and established software. However, there is no current research which sufficiently covers this topic.

Research Questions

Therefore, the first objective of this article is to assess if and to what degree BPMN is a suitable tool to model and illustrate MF systems, especially as a complement for DES. Based on that, a standardized methodology is proposed that combines both techniques to increase their usability and reduce the effort spent for modeling. This approach is expected to provide better insights into the process. Hence, the following two questions are investigated in this article:

1. How must a modeling approach for material flow systems be designed to ensure a sensible combination of BPMN process modeling and DES?
2. How can improvements for a material flow system be found based on its BPMN process model?

MATERIALS

Characteristics of Material Flow Systems

MF systems contain all operations which are necessary for the processing and distribution of goods within a defined area. Thus, they execute the physical component of an enterprise logistics process. MF processes lead to a transformation of transported goods regarding time, location, quantity, composition, and quality. For different types of transformations, MF systems contain different subsystems like conveying, storage, or handling. Although very different in terms of their technical design, these subsystems follow similar requirements from a process-overarching perspective. Key performance indicators (KPIs) for logistics contain, for instance, the throughput, which is the number of TUs processed in a certain timeframe. (Hompel et. al. 2018) An important concept from lean management are the seven forms of waste. They allow for a distinction between those activities which directly contribute to the value of goods and those which do not. Particularly in the field of MF, unnecessary transportation processes are one form of waste. (Guenthner and Boppert 2013) Those KPIs allow the controlling of how well an MF system can be optimized with a certain improvement. They therefore are an important aspect of an optimization methodology.

A typical challenge that needs to be dealt with in the MF domain is queueing systems. If a certain process can only handle one object at a time, but several objects require the availability of that process, all of these objects but one form a queue. Depending on whether – over a certain period of time – the number of arriving objects is smaller or larger than the number of processed objects, the queue either shrinks or grows. (Hompel et. al. 2018) Since neither arrival nor processing time are constant but rather can follow a complex distribution, analytical modelling of queueing system is a challenging task. (Arnold and Furmans 2019)

Business Process Model and Notation

The process models in this article follow the standard BPMN syntax and have been created with the software "Modelio" (SOFTEAM 2020, GitHub 2021).

In BPMN, a process is represented as a sequence of events and activities, connected by the SF. Various gateways allow for branching and merging of the SF to depict process variations. Additional elements like messages and data objects make BPMN useful for the modeling of information flow. (OMG 2011) Applied to MF systems, the SF can represent the movement of different objects, making the model material- or resource-oriented. However, BPMN does not offer any possibility of representing the scarcity of these objects resulting from object-constrained activities. While so-called pools and lanes can be used in BPMN to represent certain resources (e. g. the worker who is responsible for a task), the usefulness of this option is limited by the fact that each element can be part of only one lane (and each

lane can be part of only one pool) (OMG 2011). This does properly represent the complexity that is typical for more detailed models of MF systems.

Discrete-Event Simulation

DES on the other hand focuses more on the objects that a system consists of. How they are represented can vary depending on the simulation software, but common elements include TUs, containers, conveyors, workstations, assembly stations, resources, and submodels (JaamSim 2021a). A DES model also contains information about processes, but usually without an explicit visualization. However, in contrast to BPMN, visual representation is not the main purpose of a DES model anyway. Instead, the model is used to receive quantitative information about the system by performing simulation experiments. (VDI 2018) An example for this would be the simulation of a queuing system: Instead of trying to calculate dynamic queuing lengths and waiting times analytically, the system behavior is modeled and simulated using simple elements like entity generators, statistical distributions, queues, and servers. Based on these, various simulation experiments can be conducted to represent different operating scenarios and predict the system behavior quantitatively and in detail. In many cases, DES can deliver very accurate solutions for this type of problem while requiring less modeling effort than other tools (Arnold and Furmans 2019).

The simulation models for this article have been created with the DES software "JaamSim" (JaamSim 2021a & 2021b).

METHODOLOGY

Combining BPMN & DES

As discussed above, each of the two modeling tools focuses on its own aspects of an MF system. Combining them makes it possible to create a more comprehensive model by using their respective advantages, and to reduce the modeling effort by using similarities and synergies. Figure 1 shows a methodology for the modeling and simulation of MF systems in which the BPMN model is used both as a result in itself and as a starting point for the creation of the DES model. A key element of this methodology is the use of generic modules, which is illustrated with an example later in this article.

The purposes of the system analysis as the first step of the methodology are to document external requirements, to gather information about the system and to decide on a sensible substructure. Qualitative information covers the different process paths which simulation entities can follow, the order of conveying and processing operations and the connected paths of simulation entities in assembly processes. This kind of information is necessary for the process modeling. That is, for creating a BPMN model, the modeler needs to understand all different process variants in the system but without the necessity of specific values, e. g. process times. To generate the DES model; however, quantitative data is necessary as well. The parametrization of the simulation requires inputs for conveying times, process times, and inter-arrival times of entity generators. In addition to that, if downtimes of processes are supposed to be represented, statistical distributions of breakdowns and maintenance must be provided.

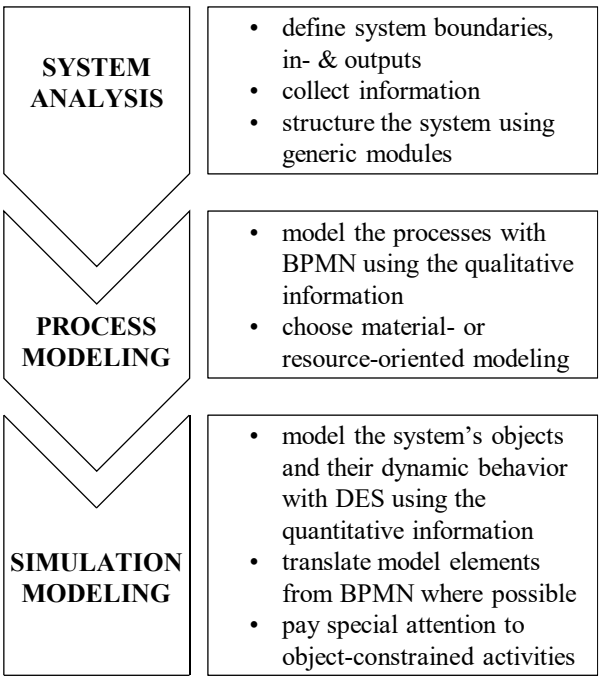


Figure 1: Systematic Modeling and Simulation of an MF System using both BPMN and DES

Regarding the second step of the methodology and the decision for material- or resource-oriented modeling in BPMN, a sensible approach is to use the SF to represent the more complicated (or more important) process type. Some examples for this are shown in this article.

Lastly, when creating the simulation model, the necessary effort can be reduced by translating between BPMN and DES. This is especially useful in material-oriented BPMN models, where the visualized sequence of activities shows strong similarities to the material flow in a DES model. Although the specific use always depends on the system, the modeling perspective, and other factors, it is generally possible to map certain elements between the two modeling tools. Table 1 shows selected examples. They are intended not only to show the underlying idea of translating between BPMN and DES, but also to clarify the content of figures 2 – 6. Although this table contains only a small subset of BPMN elements, it can be assumed that the vast majority of processes in MF systems can be mapped with it. This is firstly because only 20 % of the BPMN syntax is regularly used in practice (Muehlen and Recker 2013), and secondly because many qualitative and data-based relationships are mapped in BPMN via naming and comments, without requiring additional modeling elements.

Table 1: Comparison of Model Elements between BPMN & DES

BPMN	DES
Start Event	SubModelStart
End Event	SubModelEnd
Intermediate Event	Queue
Process TU	Server
Assemble TUs	Assemble
Add TUs to transport tool	AddTo
Remove TUs from transport tool	RemoveFrom
Transport TU	EntityConveyor
Subprocess ⊞	SubModel
Branching Gateway	Branch
Sequence Flow	Entity Flow
Message Flow between End and Start Event	Entity Flow

Generic Modules

When modeling complex systems in BPMN, it is recommended to identify subprocesses that are similar to each other and model them by creating and reusing a generic module, much as the source code of a computer program may define a function once and then call it multiple times (White 2004). Similarly, in DES, submodels are used to create a hierarchical structure, using generic modules here as well. It is therefore possible to translate not only individual elements, but even entire compound modules between BPMN and DES. In the logistics domain, most subsystems can be attributed to one of the basic functions mentioned above. This opens the possibility of creating a selection of generic modules that can be used for many different MF systems using different parameters, much like a software library. This, too, promises a reduction in the amount of work that must be invested in modeling an MF system with the described methodology. Generic modules provide support when structuring systems or modeling processes, objects, and their dynamic behavior.

An example for this is shown in the following. This module called “Deliver TUs” describes the movement of a vehicle transporting TUs between different workplaces in a recurring sequence. Figure 2 shows the depiction of this process in BPMN, where a resource-oriented model is used. Those activities that take place at the same location can be grouped into a subprocess (see Figure 3), which can then be translated into a DES submodel (see Figure 4).

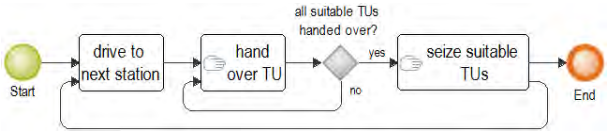


Figure 2: Module “Deliver TUs”: BPMN Process Model

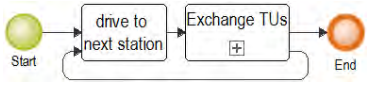


Figure 3: Module “Deliver TUs”: BPMN Process Model with subprocess “Exchange TUs”

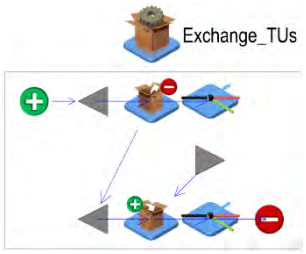


Figure 4: Module “Deliver TUs”, subprocess “Exchange TUs”: DES Model

When a simulation entity (representing a vehicle) enters the DES submodel, it will first pass through the upper part, releasing all suitable TUs to the Branch object, and then move on to the lower part, seizing all suitable TUs from the queue in the middle.

CASE STUDY

To evaluate the methodology described above, it has been applied to several real MF systems. One of these case studies, a production plant for motorcycles in Berlin, is shown in this section (BMW 2021, Welt 2017). Both qualitative and quantitative information for this system was gathered in workshops with process owners and cross-checked with specifications provided by developers of MF technology (e. g. conveying speed of belt conveyors). This MF system includes two assembly lines, "engine assembly" and "assembly", that each consist of several stations. The supply of components from the manufacturing department to these assembly stations is realized using a milk run delivery. This milk run is not used for the transport between the two assembly departments. Figure 5 and Figure 6 show parts of the BPMN and DES model for this system, respectively. The milk run delivery is a typical element of MF systems and can therefore be modeled generically, using the module "Deliver TUs" shown above, and additional EntityConveyors between the departments in the DES model, which represent the movement of the transport vehicles. In the BPMN model, activities labeled "[…]" are placeholders for additional processes, the illustration of which would go beyond the limits of this article.

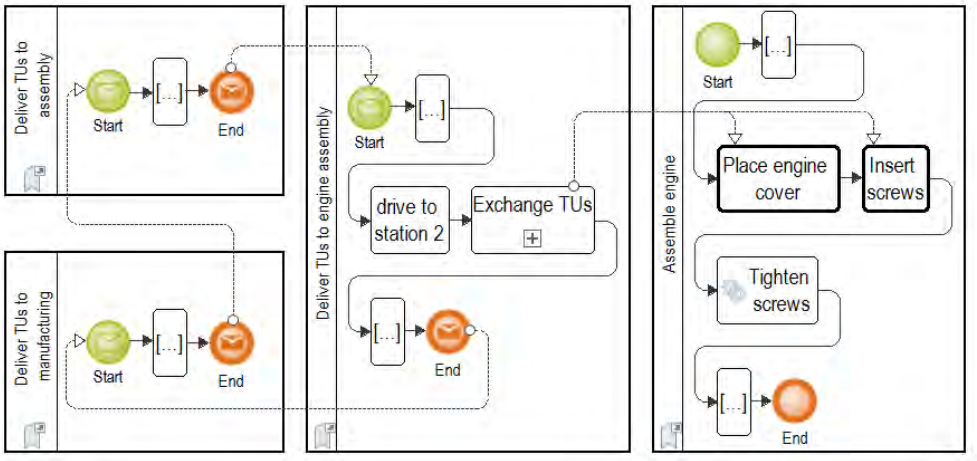


Figure 5: Motorcycle Production: Extract from the BPMN Process Model

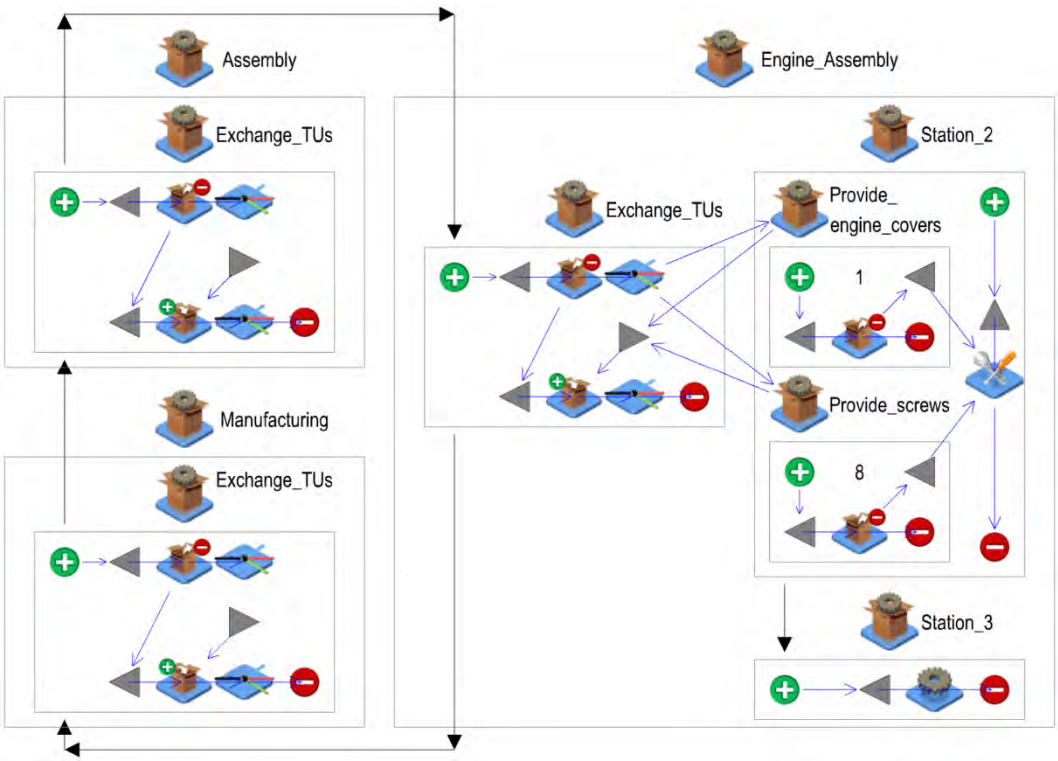


Figure 6: Motorcycle Production: Extract from the DES Model

As previously mentioned, modeling administrative processes with BPMN can be used to visualize weaknesses and help with improving the system. Using this case study, two examples show that the same is true for BPMN models of MF processes.

Firstly, Figure 5 shows that all the vehicles that deliver components pass through the departments of manufacturing, assembly, and engine assembly in the same order. There is no movement of individual components between assembly and engine assembly, so this procedure is inefficient. It can be improved by dividing the schedule so that the two assembly departments are supplied separately. The quantitative evaluation of this process change in the DES model shows that it could reduce the required number of vehicles by 25 %, thereby avoiding a waste of resources in the form of unnecessary WIP and transportation of material.

Secondly, a further consideration of the BPMN model reveals a potential improvement regarding equipment downtimes: In the engine assembly, a breakdown of the machine responsible for tightening the screws will result in a heavy accumulation of material right after station 2 or a shutdown of the assembly line. Depending on the reliability of this process, it could be sensible to implement a preventive measure, e. g. adding the activity "tighten screws" to the tasks of station 2 in case of a machine failure. This suggestion can be implemented into the BPMN model using additional gateways and intermediate events. Applying this change in the DES model shows that the "emergency plan" keeps the average throughput during a representative machine downtime at 60 % of its normal value. Without it, the machine failure would result in an increase in queue lengths and throughput times until station 3 is running again and the system can level off again.

DISCUSSION

Based on the results of the previous section, the two initial questions can be answered. For the combined modeling approach, it was first examined to what extent the BPMN syntax enables the mapping of different logistical processes. In this context, the principles of material- and resource-orientation were introduced. While the former is more suitable for linear connections of manufacturing and assembly steps, the latter should be used to depict transport processes in which material is picked up and delivered at several points. Especially when modeling in a material-oriented manner, most elements can be translated into DES without major problems. However, as the evaluation of important approaches from the literature showed, BPMN has proven to be largely unsuitable for mapping object-constrained activities and their quantitative effects, given that pools and lanes usually cannot represent the relationships in complex MF processes.

Lastly, it could be shown that BPMN models of MF systems are suitable to visualize typical improvement potentials of these systems as well as the qualitative advantages of the optimized processes. Changes in the MF can be modeled in BPMN by re-arranging the sequence flow between activities and events. Due to the lack of quantitative information in the BPMN models, DES is then used to test and evaluate the identified process improvements based on MF KPIs. It can be concluded that existing methodologies for systematic optimization are also applicable within this framework.

Although the presented methodology has a significant potential to increase the system performance while reducing the modeling effort, there are some limitations for its application. Modeling object-based activities with BPMN is hardly feasible and the scarcity of resources and other objects cannot be illustrated. Nonetheless, the approach can be applied to real-world industrial scenarios and process insights can be enhanced. The methodology clearly separates requirements and solutions and improves the usability, also by incorporating generic modules.

Compared to the related publications mentioned above, the methodology presented in this article applies BPMN specifically to MF systems without requiring any modifications of the existing syntax. Quite the reverse, combining BPMN and DES makes it possible to concentrate only on the common, useful, and well-understood modeling elements in each tool. This also results in a greater variety of eligible software, which is beneficial for applications both in industrial and academic settings.

SUMMARY

In this paper, a methodology for the modeling of MF systems using both BPMN and DES was presented. As a first part of it, the comparison between elements from both approaches allows for a translation from one framework into the other. Secondly, by synthesizing MF systems from generic modules, a plannable and time-saving modeling process could be created. The combination of standardized and well-known elements from both modeling languages allows for a methodology which is easy to understand and suitable for multidisciplinary teams. A case study at an MF process within a manufacturing system showed that the approach enables the identification and assessment of optimization potentials without the need for a costly real-world test run. Further research in this area could extend the "translations" to include rarer BPMN and DES elements, as well as expand the selection of generic modules to include other typical use cases in the MF domain. Regarding the detection and elimination of process weaknesses, the developed methodology could benefit from a more systematic approach specifically focused on the combination of BPMN and DES in the field of MF systems.

REFERENCES

Arnold, D.; and K. Furmans. 2019. *Materialfluss in Logistiksystemen*. Springer, Berlin / Heidelberg.

BMW Group. 2021. *BMW Group Werk Berlin*. bmwgroup-werke.com/berlin/de/unser-werk.html (last accessed on 2021-10-15).

Forno, A. J. dal; F. A. Pereira; F. A. Forcellini; and L. M. Kipper. 2014. "Value Stream Mapping: a study about the problems and challenges found in the literature from the past 15 years about application of Lean tools". *The International Journal of Advanced Manufacturing Technology* 72, 779-790.

García-Domínguez, A.; M. Marcos; and I. Medina. 2012. "A comparison of BPMN 2.0 with other notations for manufacturing processes". In *AIP Conference Proceedings* 1431 (Cadiz, Spain, 2011-09-21 – 23). 593-600.

GitHub (unknown authors). 2021. *Modelio User Documentation*. github.com/ModelioOpenSource/Modelio/wiki/Modelio-User-Documentation (last accessed on 2020-10-07).

Guizzardi, G.; and G. Wagner. 2011. "Can BPMN Be Used for Making Simulation Models?" In *Enterprise and Organizational Modeling and Simulation*, J. Barjis; T. Eldabi; and A. Gupta (Eds.). Springer, Berlin / Heidelberg, 100-115.

Guenthner, W. A.; and J. Boppert. 2013. *Lean Logistics – Methodisches Vorgehen und praktische Anwendung in der Automobilindustrie*. Springer, Berlin / Heidelberg.

Hompel, M. ten; T. Schmidt; and J. Dregger. 2018. *Materialflusssysteme*. Springer, Berlin / Heidelberg.

JaamSim Development Team. 2021a. *JaamSim – Discrete-Event Simulation Software*. Version 2021-05. jaamsim.com (last accessed on 2022-01-20).

JaamSim Development Team. 2021b. *JaamSim User Manual*. jaamsim.com (last accessed on 2022-01-10).

Muehlen, M. zur and J. Recker. 2013. "How Much Language Is Enough? Theoretical and Practical Use of the Business Process Modeling Notation". In *Seminal Contributions to Information Systems Engineering*, J. Bubenko et al. (Eds.). Springer, Berlin / Heidelberg, 429-443.

OMG (Object Management Group).2011. *BPMN (Business Process Model and Notation)*. Version 2.0.

Pires, F.; A. Cachada; J. Barbosa; A. P. Moreira; and P. Leitao. 2019. "Digital Twin in Industry 4.0: Technologies, Applications and Challenges". In *2019 IEEE 17th International Conference on Industrial Informatics (INDIN)*. IEEE, 721 – 726.

Robinson, S. 2006. "Conceptual Modeling for Simulation". In *Proceedings of the 2006 Winter Simulation Conference*, L. F. Perrone et al. (Eds.), 792-800.

SOFTEAM Group. 2020. *Modelio*.

VDI (Verein Deutscher Ingenieure). 2018. *Simulation von Logistik-, Materialfluss- und Produktionssystemen*. VDI Technical Rule No. 3633.

Wagner, G.; O. Nicolae; and J. Werner. 2009. "Extending Discrete Event Simulation by Adding an Activity Concept for Business Process Modeling and Simulation". In *Proceedings of the 2009 Winter Simulation Conference*, M. D. Rossetti (Ed.). IEEE et al., Piscataway, NJ 2951-2962.

Wagner, G. 2018. "Information and Process Modeling for Simulation – Part I: Objects and Events". *Journal of Simulation Engineering*, 1 (2018/2019).

Wagner, G. 2021. *Information and Process Modeling for Simulation – Part II: Activities and Processing Networks*.

WELT Nachrichtensender. 2017. *Die Motorradfabrik – Ein Superbike entsteht*. welt.de/mediathek/reportage/automobile/sendung171325367/Die-Motorradfabrik-Ein-Superbike-entsteht (last accessed on 2021-10-15).

White, S. 2004. *Introduction to BPMN*. IBM Corporation.

Zor, S.; D. Schumm; and F. Leymann. 2011. "A Proposal of BPMN Extensions for the Manufacturing Domain". In *New Worlds of Manufacturing*, CIRP (International Academy for Production Engineering) (Ed.).

AUTHOR BIOGRAPHIES

MAXIMILIAN WUENNENBERG is currently a research associate pursuing his Ph.D. at Technical University of Munich (TUM), Chair of Materials Handling, Material Flow, Logistics (fml), where he received his M.Sc. in Mechanical Engineering in 2020. He is responsible for the research project "Consistent Development of Material Flow Systems using a Model-based approach". His main research interests are Model-based Systems Engineering, Material Flow Systems and Data Analytics. His e-mail address is max.wuennenberg@tum.de and his web page can be found at https://www.mec.ed.tum.de/fml/ueber-den-lehrstuhl/mitarbeitende/maximilian-wuennenberg/

BENJAMIN WEGERICH received his B.Sc. in Mechanical Engineering in 2020. He is currently pursuing his M.Sc. in Mechanical Engineering at TUM, with his main research interests in manufacturing, logistics, digitalization, and Industry 4.0. His e-mail address is benjamin.wegerich@tum.de

JOHANNES FOTTNER is a professor for technical logistics at TUM, chair fml. His research areas are innovative identification technologies, digital planning of logistics systems and human factors in logistics. After obtaining his Ph.D. at TUM, chair fml in 2002, he worked in several management positions at Swisslog before becoming managing director of MIAS Group. Since 2015, he also has worked at the Association of German Engineers (Verein Deutscher Ingenieure, VDI) as chairman for Bavaria and vice-chairman for manufacturing and logistics. His e-mail address is j.fottner@tum.de and his web page can be found at https://www.mec.ed.tum.de/fml/ueber-den-lehrstuhl/mitarbeitende/prof-dr-ing-johannes-fottner/

MODELLING AGV OPERATION SIMULATION WITH LITHIUM BATTERIES IN MANUFACTURING

Ozan Yesilyurt
Marius Kurrle
Andreas Schlereth
Miriam Jäger
Fraunhofer Institute for Manufacturing Engineering and Automation IPA
Nobelstraße 12, 70569 Stuttgart, Germany

Alexander Sauer
Fraunhofer Institute for Manufacturing Engineering and Automation IPA & Institute for Energy Efficiency in Production, EEP, University of Stuttgart
Nobelstraße 12, 70569 Stuttgart, Germany

E-Mail: ozan.yesilyurt@ipa.fraunhofer.de

KEYWORDS

Simulation modeling, manufacturing, automated guided vehicles

ABSTRACT

This paper describes the development of a production simulation model with automated guided vehicle (AGV) operation to prepare relevant production data validating an approach using AGV batteries as energy storage to reduce peak loads in a manufacturing company. First, the definition of AGV and the simulation modeling approach are introduced. Then, the systematic literature review methodology is described to explore relevant existing simulation models with AGV operation. With the help of this information, a simulation model is designed and developed. The last sections include the experiments performed with the simulation model, analysis, and the following results. The results show that the developed simulation can be used to generate data to evaluate the above-described approach in production.

INTRODUCTION

Industrial manufacturing faces significant challenges due to the increasing importance of sustainability and the rise of complexity in markets. First, unlike conventional transportation systems, battery-driven AGVs produce little emissions and have high energy efficiency. Still, reducing energy consumption is essential to reach a firm's own or external greenhouse gas reduction goals and, at the same time, to benefit economically from lower energy costs and higher energy security (Roesch et al. 2019). Compared to conventional vehicles, battery-driven AGVs need more time to charge. Consequently, energy consumption and charging time can be very volatile, which has to be considered in real-life use cases to manage a possible decrease in costs and emissions (Ma et al. 2021; Pfeilsticker et al. 2019; Zhu et al. 2018). Second, the steady increase in the complexity of markets poses a challenge on manufacturing companies, and their value creation networks increasingly facing new challenges (Bauernhansl et al. 2014). The manufacturing companies are forced to respond to such events with non-automated operations, high stock levels, and significant lead times. As a result, decisions can be postponed, and remedial action can occur late. To solve these problems, simulations can forecast the planned changes in manufacturing because they acquire relevant results for practical implementations (VDI-Gesellschaft Produktion und Logistik 2014). With the new developments in the AGV field (Shihua Li et al. 2018), the relevance of using AGVs for companies is growing (Kunst 2018). To achieve the goal of optimal production despite the challenges of volatile power consumption and complex markets, simulation models can help companies to gather data for validating real-life use cases. This paper aims to generate realistic data for validating the approach of using AGV batteries as energy storage in production to minimize peak loads. Next, the problem statement and objective target of this paper are introduced.

PROBLEM STATEMENT AND OBJECTIVE TARGET

Manufacturing companies face a complex environment regarding current and future energy supply with their factories. The companies are charged for electricity based on two principles. One is for energy consumption and the other for peak demand. Besides the energy consumption contracts, the companies should agree with the electricity providers on contractual price models depending on the highest peak production load. The manufacturing companies experience these peak loads in their production and pay high amounts of money for generating them (Kurnik et al. 2017). Different concepts and applications of the energy storage systems such as stationary and electric vehicle (EV) batteries were studied and developed in manufacturing plants to minimalize peak loads and enhance savings for the companies.

In this paper, only the electrical energy storage devices of the AGV are considered to achieve the same goals. To validate this approach, some data such as AGV and energy consumption data is needed from a company. After contacting the different companies, one manufacturing company agreed to cooperate. The cooperated company sent the energy consumption data for one year. However, the company does not possess any

AGV. Therefore, it is planned that a simulation should generate the required AGV data. The goal is to develop a simulation model according to the company's logistics processes that simulate a production line with AGV. The simulation model should generate availability, position, state of the charge status of the AGVs, and availability of the charging stations. In summary, the scope of this paper is to develop a simulation model to generate and analyze the data on the energy consumption of AGVs.

STATE OF THE ART

In the following, first, AGVs are defined. Then, the definition of a simulation and a recommended approach to develop a simulation are introduced. Last, the systematic literature review identifies the existing AGV simulation concepts.

Automated Guided Vehicles

AGVs are in-plant, ground-based material flow systems consisting of automatically controlled vehicles whose main task is to transport the material (VDI-Gesellschaft Produktion und Logistik 2005). The most crucial flexibility criteria for material handling technology include the ability to be integrated into an existing production environment, transport a wide variety of goods, and adapt to productivity fluctuations.

AGVs offer the highest degree of flexibility among all automated material handling technologies. Other advantages of AGVs include minimal infrastructure measures, use of existing paths, and the possibility of easy replacement with another vehicle or a conventional forklift. Concerning energy flexibility, AGVs can also serve as storage units that establish resilience by creating buffers (Roesch et al. 2019). This is particularly important for enabling companies to create energy flexibility, which is the base for trading in dynamic energy markets (Pfeilsticker et al. 2019). Because of the wide range of possible applications, there are almost no restrictions on the design of the AGV (Steegmüller & Zürn 2014; Ullrich & Albrecht 2019).

Simulation Modelling

This section introduces the definition of a simulation, a recommended approach to generate a simulation, and simulation modeling methods. A simulation represents a physical system and its related processes in a model. Its goal is to obtain transferable results for practical applications (VDI-Gesellschaft Produktion und Logistik 2014). The terms system and model are related to the term simulation. The system is a collection of components and their properties, which are connected by interdependencies (Hall and Fragen 1956). A model is an abstract image of a system (Eley 2012). If computers are used for the necessary calculations in the simulation, this is called a computer simulation. For this purpose, the model must be available in a mathematical-logical form and implemented in a computer program. These computer programs are considered simulation tools (Eley 2012).

A simulation experiment is the reproduction of the behavior of a system with a model over a certain period of time (VDI-Gesellschaft Produktion und Logistik 2014). This particular period is called simulation time. On the other hand, the simulation time represents the time progressing in the existing system (Eley 2012). According to (VDI-Gesellschaft Produktion und Logistik 2014) the following approach is recommended to create a simulation:

1. formulation of problems,
2. test of simulation-worthiness,
3. formulation of targets,
4. data collection and data analysis,
5. modeling,
6. execution of simulation runs,
7. result analysis and
8. documentation.

The approach above is used to create a simulation model in this paper. The following section describes the systematic literature review to identify existing AGV simulation concepts.

RELATED WORK

The approach for finding related literature is based on the systematic approach according to (Jan Vom Brocke et al., 2009). It represents a patterned system for identifying and selecting relevant literature for the research field. (Jan Vom Brocke et al. 2009) identified five required steps for a literature review. In these steps, first, the research framework is established, second, the topic is conceptualized. Then, the literature review and literature analysis are conducted. Finally, a research agenda completes the research (Vom Brocke et al. 2009).

This literature search aims to find existing research approaches and implementations of AGV use in simulations. To conduct the literature search, search terms are created by using synonyms and closely related terms illustrated in Table 1.

Table 1: Search terms

Context Synonyms	Scope	Topic Synonyms
production	simulation	"AGV battery"
manufacturing	routing	
	modeling	

After that, Boolean operators are used to merge the search terms into a search string. Wild cards (*) are used to consider the plural of search terms and exclude forms of words in the literature search. The following search string is applied to find relevant literature in different databases.

("Production*" OR "Manufacturing*") AND ("simulation*" OR "routing*" OR "modeling*") AND ("AGV*" OR "AGV* battery*")

Five different databases are chosen to conduct the search string to find relevant papers. The databases' technical orientation and the search results' scientific relevance are considered in the selection of the databases. Figure 1 shows the selected databases and the methodology of the multi-stage filtering system of the search results.

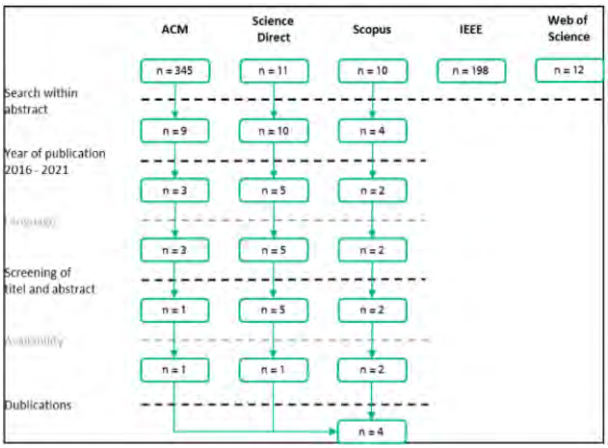


Figure 1: Results of the literature search with the multi-stage filtering system

Four relevant publications are found after conducting the literature search with the methodology of the multi-stage filtering system. In the next section, these publications are described in detail.

Existing AGV Simulation Concepts and Applications

The four different papers on AGV simulation concepts in production are introduced in this section.

The research paper by Ndiaye et al. 2016 introduces an AGV transportation system defined by a layout, several vehicles, several parking spaces, and a vehicle management policy. First, a simple formula determines the minimum required number of vehicles. Then, a discrete-event simulation model is used to evaluate different layouts and vehicle-dispatching policies. Initially, eight vehicles were required to meet the transportation demand, but with some optimizations, this number could be reduced to four vehicles. While these optimizations reduce the number of vehicles, they incur additional costs during the implementation phase. This means that savings achieved by reducing the number of vehicles are lost during the implementation phase in terms of software. The paper focuses on reducing the number of AGVs. However, generating the AGV and charging station data is required to verify the above-described approach.

Zhan et al. 2019 describe two-stage battery-charging strategies proposed for AGVs equipped with lithium-ion batteries to improve utilization. In stage 1, two routing decisions are developed. These are the nearest charging station (NCS) and the charging station with minimum delay (MDCS). In stage 2, the duration of each operation is reduced considering the charging characteristics of the lithium-ion battery. A real case is adopted to illustrate the applicability and effectiveness of the proposed approach. These methods help to improve manufacturing at a short-term capacity to meet the market demand. This paper shows that the loading strategy of MDCS performance is better than the loading strategy of NCS, in terms of AGV utilization and overall performance. This means that improving the utilization of AGV contributes to increasing the production of the system. The charging station with minimum delay methodology idea was taken from this paper and implemented in the new simulation. Nevertheless, the work of the researchers does not solve the problem statement described in the paper.

The research of Mousavi et al. 2017 used a fuzzy hybrid genetic algorithm (GA)-particle swarm optimization (PSO) algorithm with a comparison with three other algorithms (GA, PSO, and hybrid GA-PSO). Comparing the four algorithms results showed that the Fuzzy Hybrid-GA-PSO yields the lowest production time and AGV numbers. However, a difference was observed between the performance of Fuzzy-Hybrid-GA-PSO and Hybrid-GA-PSO. The only significant improvement over Hybrid-GA-PSO concerned the computation time. The AGV system simulation with Flexsim software proved the practicality of the developed model and the studied algorithms. The focus of this paper was to compare different optimization algorithms. Therefore, the results of this paper cannot be used for the described problem statement.

Mousavi et al. 2017 focused on multi-objective AGV scheduling in a flexible manufacturing system (FMS) using GA, PSO, and hybrid GA-PSO algorithms. A model for AGV task scheduling was developed. The comparison of the three algorithms shows that the hybrid GA-PSO provides the lowest production runtime and AGV numbers. It was found that after optimization, despite a slight increase in the total AGV running time (loaded and unloaded), reducing the idle time of the AGVs improved the operating efficiency of the AGVs. Consistent with the experimental results, FlexSim software has been used to prove the feasibility of the developed model and the suitability of the optimization algorithms for the scheduling problem. The developed model can be applied to any FMS. It can be applied to optimize the objectives separately or in a combinatorial way. Various algorithms are reviewed to enable and optimize the multi-scheduling of AGVs. This paper was out of scope because the problem statement could not be solved with the help of this paper.

The systematic literature review results show no existing AGV simulation to solve the above-described problem statement. Therefore, it is decided to develop a new AGV simulation to generate the required data that can be applied to a realistic situation. The next section describes the case study, including the concept of the simulated logistics process. After that, the simulation model is described.

CASE STUDY

Introduction of the company

The company for which the simulation was developed is a chemicals manufacturer in the synthetic leather and textile coating industry. The company's factory embraces 4000 m^2 and consists of a warehouse and two production lines. The company's warehouse is located between two production lines (with a diameter of 30 meters). Between 70-100 tons of products are transported per day. Two employees work three shifts per day in the company, and one employee transports a barrel with a forklift truck during one trip. With the help of this information, the assumptions of the concept are described in the next section.

Concept Description

The simulation model should be developed with the AGV operation. The following assumptions in the simulation model are made for the logistics processes of the company to realize a realistic simulation:

- The simulation duration is set to 1-year simulation time from 01.01.2021 to 31.12.2021, because, to calculate how many peak loads can be covered with AGV batteries, AGV data for one year should be generated.
- The transport processes of two products (chemicals) are considered.
- The interval of the production orders (min. in 3 minutes and max. in 27 minutes) is calculated for one year per day using the respective energy consumption data integrating it into the simulation.
- An AGV has an average speed of 1 m/s (approx. 4 km/h).
- The SoC limits of the AGV are predefined (20-90% for the lithium battery saving). If the lower SoC limit is exceeded, AGV should drive to the charging station, which has a minimum delay to charge the AGV battery.
- All AGVs have an initial value of 50% for the charge state of the battery.
- One AGV garage is located in the factory layout and consists of two AGVs. The employees have been replaced with AGVs.
- AGVs can transport the products to both production lines.
- Two charging stations are simulated so that AGV batteries are charged.

The data from Kuka KMP 1500 AGV (KUKA AG 2016) is used as the AGV data. It is shown in Table 2.

Table 2: Kuka AGV data

Feature	**Value**
maximum payload	1,500 kg
battery capacity	104 Ah (extended battery version)
charging current	52 A
charging voltage	96 V
battery energy density	9,984 Wh
charging time	2 hours (up to 100%)
driving consumption	13 A (min. 8 hours)

After the assumptions are determined, the simulation model is developed. The next chapter presents the model description.

MODEL DESCRIPTION

The simulator *Tecnomatix Plant Simulation* (SimPlan AG 2021) was chosen for its wide selection of objects used in production processes and pre-existing models using battery-driven AGVs. In addition, the simulation can be precisely adapted to the case study using custom-written *SimTalk* program language. The simulation model (see Figure 2) is based on Steffen Bangsow's training example which includes AGVs using tracks (Bangsow 2021).

At the beginning of the simulation, the AGVs launch in the garage below the transport routes. The number of spawned AGVs is defined through the variable *numAGV*, which, according to the concept, is set to two. The source of the products is connected to a buffer out that transfers the products to the AGVs. There are two products ("Part1", "Part2") in the simulated scenario that have the same characteristics but different destinations. According to the table *Workplan* that defines the routes of the products, one part must be transported to the first station on the right side, and, accordingly, the other part to the second station on the left side of the model. Therefore the AGVs operate on two separate paths which provide for collision avoidance. Besides the production sequence, *Workplan* defines the setup and processing times of the stations as well. These are set to avoid AGV queues before the source. If there would still occur, the following AGV would wait until the demanding one has received the product. This model is flexible and can be extended by adding more stations as long as the *Workplan* gets updated continuously.

The number of products generated by the source in each interval can be varied. The interval is set to one day in the case study. Thus, the daily production quantity can be

defined. This was implemented in the simulation by a generator that changes the time gap in which the source generates the products according to the table *ProductInterval*. The two endpoints of the transport routes are represented by two station objects in the simulation. In front of each station is a buffer that deals with the incoming products. In the simulation, the buffer serves the purpose of holding parts if components in the station cannot be processed in time. For the simulated case, a buffer is not crucial because capacities, setup, and processing times are calculated accordingly so that the products can be dispatched at the right time. However, since the simulation should be extendible in the future, the buffer was retained as a linking component. The buffer type is set to a queue. Consequently, products exit the buffer in the same order as they arrived (First-In-First-Out-principle).

All objects must be provided with the correct parameters to ensure that the correct production sequence can be created and automated later on. The simulation's most important object is the AGV itself. It is equipped with custom methods that define the logic of its operation. The most relevant method is *doJob*, which defines that an empty AGV should drive to the source to pick up a new part and afterward drive to the destination of the new job. The method also describes the procedure of battery charging. If the AGV's charge level goes below the defined battery reserve threshold, the vehicle checks the availability of the charging stations and drives to the next unoccupied station. This only occurs when the AGV is idle and therefore has no current job. After arrival, the AGV charges up to the specified capacity (90% of maximum capacity).

In *Plant Simulation,* properties of objects and process logic can be automated by so-called methods written in the programming language *SimTalk*. Methods can rely on tables and/or write new data in existing tables. The presented simulation consists of eight methods that ensure the correct execution of the aspired process. For this paper, the methods were adapted to the described conditions. In addition, the method *writeData* was created to save the data generated during the simulation in tables.

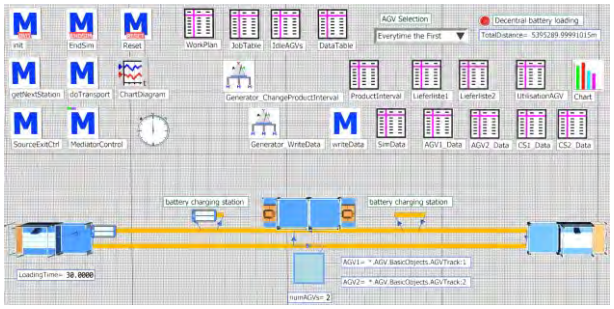


Figure 2: Production simulation model with AGV operation

RESULTS

The results of the developed simulation experiment are shown and interpreted in this section. The simulation time is set to one year for collecting data per minute so that the AGV batteries' approach can be validated. The simulation generated data for two AGVs and two charging stations. A total of 525,600 data was created per AGV and charging station in this simulation experiment. The simulated example charging station data is illustrated in Figure 3. They are current time, charging station ID, and charging station availability.

Identifier	CurrentTime	csID	Measurement	csAvailability
0	2.0100	1	ChargingStation	1
1	1:00.0000	1	ChargingStation	1
2	2:00.0000	1	ChargingStation	1
3	3:00.0000	1	ChargingStation	1
4	4:00.0000	1	ChargingStation	1
5	5:00.0000	1	ChargingStation	1
6	6:00.0000	1	ChargingStation	1
7	7:00.0000	1	ChargingStation	1
8	8:00.0000	1	ChargingStation	1
9	9:00.0000	1	ChargingStation	1
10	10:00.0000	1	ChargingStation	1

Figure 3: Charging station simulation data

The simulated example AGV data is shown in Figure 4. They are current time, AGV ID, AGV availability, state of charge of the AGV battery, AGV position on the X-axis, and the Y-axis, AGV velocity.

Identifier	CurrentTime	agvID	Measurement	agvAvailability	socAGV	agvLatPosition	agvLngPosition	velocityAGV
0	2.0100	1	AGV	0	50.00	13.50	0.00	1.00
1	1:00.0000	1	AGV	0	49.88	0.80	2.00	1.00
2	2:00.0000	1	AGV	0	49.82	0.80	2.00	1.00
3	3:00.0000	1	AGV	0	49.76	0.80	2.00	1.00
4	4:00.0000	1	AGV	0	49.69	0.80	2.00	1.00
5	5:00.0000	1	AGV	0	49.63	0.80	2.00	1.00
6	6:00.0000	1	AGV	0	49.56	0.80	2.00	1.00
7	7:00.0000	1	AGV	0	49.50	0.80	2.00	1.00
8	8:00.0000	1	AGV	0	49.44	0.80	2.00	1.00
9	9:00.0000	1	AGV	2	49.28	24.20	2.00	1.00
10	10:00.0000	1	AGV	0	49.20	27.20	2.00	1.00

Figure 4: AGV simulation data

After the simulation data was collected, the following AGV results were obtained and shown in Table 3. Both AGVs work approximately only 9% of their time. The simulation results show that they have over 81% idle time per year. The idle time allows the AGVs to be used not only as transport vehicles but also as energy storage in this company to reduce peak loads in production.

In future work, an economic analysis will be conducted to analyze different implementation strategies whether the AGVs' number should be reduced to save investment costs or whether the AGVs should be implemented in the simulation to reduce the peak load costs.

Table 3: AGV simulation results in one year

AGV status	AGV1	AGV2
working	8,9 %	8,5 %
idle	81,2 %	81,7 %
charging	9,9 %	9,9 %

Table 4 indicates the simulation results of the charging stations in one year. It is observed that the second charging station was nearly not used by AGVs.

Therefore, reducing the number of charging stations to one for this use case would be conceivable. However, it has to be investigated whether it is economical to have a second charging station when the AGVs discharge their batteries in peak times to support the company grid to reduce peak loads in production.

Table 4: Charging station simulation results in one year

CS status	CS 1	CS 2
working	20,5 %	0,2 %
idle	79,5 %	99,8 %

The simulation results of the AGVs on the highest (26.09.2021) and lowest energy consumption day (02.07.2021) are illustrated in Table 5. The results show that the AGVs on the day with the highest energy consumption work significantly longer than average working time and are charged more. For this day, it can be examined whether the AGVs have additional availability to minimize arising peak loads in the production. The results on the day with the least energy consumption highlight that the AGVs tend to have above-average idle time. It must be verified whether the charging times can be increased to enable the AGVs to pull more energy in off-peak times to use later.

Table 5: AGV simulation results during the highest and lowest energy consumption day

	Highest energy consumption day		Lowest energy consumption day	
AGV Status	AGV1	AGV2	AGV1	AGV2
working	15,9 %	17,2 %	6,7 %	5,5 %
idle	74,0 %	67,8 %	83,2 %	84,4 %
charging	10,1 %	15 %	10,1 %	10,1 %

The simulation results of Table 6 support the interpretation of Table 4. When it is cost-effective to utilize the second charging station for peak power reduction with AGV batteries, the charging station can be deployed. Otherwise, it is recommended to reduce the number of charging stations to one. However, (Weeber et al. 2020) show that machine availability is key to energy-efficient manufacturing. Because machine availability is dependent on the supply of materials by AGVs, in case of doubt, more charging stations than needed are appropriate.

Table 6: Charging station simulation results in the highest and lowest energy consumption day

	highest energy consumption day		lowest energy consumption day	
CS status	CS 1	CS 2	CS 1	CS 2
working	25,4 %	0 %	21,9 %	0 %
idle	74,6 %	100 %	78,1 %	100 %

CONCLUSION

A production simulation model has been developed with AGV operation to generate production data for the validation of the approach. It aims to apply AGV batteries as energy storage devices to mitigate peak loads in a manufacturing company. To realize this simulation model, first, the applied approach for the simulation modeling is presented. Then, the related work of different researchers is introduced, which created different AGV simulation concepts in production. After making sure that a simulation for the problem statement described has not yet been developed, a new concept of the simulation model with a case study is described. Considering the assumptions of the new simulation model concept, a simulation model is developed and presented in this paper. The simulation model results indicate that the generated data from the simulation model can be used to validate the described approach above. The results also show that the AGV's energy consumption can vary widely. In addition to the parameters of energy consumption and associated costs, the increasingly important energy flexibility issue must be considered. As a next step, the generated data will be entered into a software system to calculate whether using the AGV batteries as energy storage to minimize peak loads is cost-effective.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the financial support of the Kopernikus-Project "SynErgie" by the Federal Ministry of Education and Research of Germany (BMBF) and the project supervision by the project management organization Projektträger Jülich (PtJ).

REFERENCES

Bauernhansl, T.; A. Schatz; and J. Jäger. 2014. "Komplexität bewirtschaften – Industrie 4.0 und die Folgen". *Zeitschrift Für Wirtschaftlichen Fabrikbetrieb*, *109*(5), 347–350. https://doi.org/10.3139/104.111140

Eley, M. 2012. *Simulation in der Logistik: Einführung in Die Erstellung ereignisdiskreter Modelle unter Verwendung des Werkzeuges Plant Simulation* (1st ed.). *Springer-Lehrbuch Ser*. Springer Berlin / Heidelberg. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=971208

Hall, A. D. and R.E. Fragen. 1956. "Definition of System". *General Systems*, *1*(1), 18–28.

Vom Brocke, J.; A. Simons; B. Niehaves; K. Riemer; and A. Cleven. 2009. "Reconstructing the Giant: On the Importance of Rigour in Documenting the Literature Search Process". In *17th European Conference on Information Systems (ECIS)*. https://www.researchgate.net/publication/259440652_Reconstructing_the_Giant_On_the_Importance_of_Rigour_in_Documenting_the_Literature_Search_Process. Accessed 17.01.2022

KUKA AG. 2016, June 9. "KUKA Mobile Plattform 1500". https://www.kuka.com/de-de/produkte-leistungen/mobilit%C3%A4t/mobile-plattformen/kmp-1500. Accessed 19.01.2022

Kunst, A. 2018. *Relevanz von autonomen Transportsystemen in der Logistikbranche in Deutschland 2018 | Statista*. https://de.statista.com/prognosen/943349/expertenbefragung-zu-autonomen-transportsystemen -in-der-logistikbranche. Accessed 17.01.2022

Kurnik, C. W.; F. Stern; and J. Spencer. 2017. *Chapter 10: Peak Demand and Time-Differentiated Energy Savings Cross-Cutting Protocol. The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures.* https://doi.org/10.2172/1406991

Ndiaye, M. A.; S. Dauzère-Pérès; C. Yugma; L. Rullière; and G. Lamiable. 2016. "Automated transportation of auxiliary resources in a semiconductor manufacturing facility". In *2016 Winter Simulation Conference (WSC)*. Arlington, Virginia

Weeber, M.; J. Wanner; P. Schlegel; K. Birke; and A. Sauer. 2020. "Methodology for the Simulation based Energy Efficiency Assessment of Battery Cell Manufacturing Systems". https://www.semanticscholar.org/paper/Methodology-for-the-Simulation-based-Energy-of-Cell-Weeber-Wanner/cc821d0c2a93616456f81d68a678d872c259b13a

Ma, N.; C. Zhou; and A. Stephen. 2021. "Simulation model and performance evaluation of battery-powered AGV systems in automated container terminals". *Simulation Modelling Practice and Theory*, *106*, 102146. https://doi.org/10.1016/j.simpat.2020.102146

Mousavi, M.; H.J. Yap; and S.N. Musa. 2017. "A Fuzzy Hybrid GA-PSO Algorithm for Multi-Objective AGV Scheduling in FMS". *International Journal of Simulation Modelling*, *16*(1), 58–71. https://doi.org/10.2507/IJSIMM16(1)5.368

Pfeilsticker, L.; E. Colangelo; and A. Sauer. 2019. "Energy Flexibility – A new Target Dimension in Manufacturing System Design and Operation". *Procedia Manufacturing*, *33*, 51–58. https://doi.org/10.1016/j.promfg.2019.04.008

Roesch, M.; D. Bauer; L. Haupt; R. Keller; T. Bauernhansl; G. Fridgen; G. Reinhart; and A. Sauer. 2019. "Harnessing the Full Potential of Industrial Demand-Side Flexibility: An End-to-End Approach Connecting Machines with Markets through Service-Oriented IT Platforms". *Applied Sciences*, *9*(18), 3796. https://doi.org/10.3390/app9183796

Li, S.; J. Yan; and L. Li. 2018. "Automated Guided Vehicle: the Direction of Intelligent Logistics". *Undefined*. https://www.semanticscholar.org/paper/Automated-Guided-Vehicle%3A-the-Direction-of-Li-Yan/8b852eb668d7e5de0047ee6897c8176d695da4c2

SimPlan AG. 2021, September 2. *Simulation mit Plant Simulation*. https://plant-simulation.de/. Accessed 17.01.2022

Steegmüller, D. and M. Zürn. 2014. "Wandlungsfähige Produktionssysteme für den Automobilbau der Zukunft". In T. Bauernhansl, M. ten Hompel, & B. Vogel-Heuser (Eds.), *Industrie 4.0 in Produktion, Automatisierung und Logistik: Anwendung, Technologien, Migration* (pp. 103–119). Springer Vieweg. https://doi.org/10.1007/978-3-658-04682-8_5

Bangsow. S. 2021. *AGV modelling using tracks*. https://www.bangsow.eu/detail_en.php?id=851. Accessed 17.01.2022

Ullrich, G. and T. Albrecht. 2019. *Fahrerlose Transportsysteme: Eine Fibel - mit Praxisanwendungen - zur Technik - für die Planung* (3., vollständig überarbeitete Auflage). Springer Vieweg. https://doi.org/10.1007/978-3-658-27472-6

VDI-Gesellschaft Produktion und Logistik. *VDI 3633 Blatt 1:2014-12:* Simulation of systems in materials handling, logistics and production - Fundamentals (2014-12). https://www.beuth.de/en/technical-rule/vdi-3633-blatt-1/149034959?websource=vdin. Accessed 14.01.2022

VDI-Gesellschaft Produktion und Logistik. 2005-10. *VDI 2510:2005-10:* Automated Guided Vehicle Systems (AGVS) (VDI 2510). Beuth Verlag. https://www.beuth.de/de/technische-regel/vdi-2510/78228504. Accessed 14.01.2022

VDI-Gesellschaft Produktion und Logistik. 2014-12. *VDI 3633 Blatt 1:2014-12:* Simulation of systems in materials handling, logistics and production - Fundamentals. Beuth Verlag. https://www.beuth.de/en/technical-rule/vdi-3633-blatt-1/149034959?websource=vdin. Accessed 14.01.2022

Zhan, X.; L. Xu; J. Zhang; and A. Li. 2019. "Study on AGVs battery charging strategy for improving utilization". *Procedia CIRP*, *81*, 558–563. https://doi.org/10.1016/j.procir.2019.03.155

Zhu, Z.; Z. Gao; J. Zheng; and H. Du. 2018. "Charging Station Planning for Plug-In Electric Vehicles". *Journal of Systems Science and Systems Engineering*, *27*(1), 24–45. https://doi.org/10.1007/s11518-017-5352-6

AUTHOR BIOGRAPHIES

OZAN YESILYURT got his bachelor's and master's degree in electrical engineering and information technology at the Technical University of Munich. Since 2016, he has been working as a research fellow in the Competence Center DigITools at Fraunhofer IPA.

MARIUS KURRLE obtained his Bachelor of Science in Technical Business Administration, focusing on logistics and production systems at the University of Stuttgart in 2020. He is currently pursuing his Master of Science in the same study program and works as a student assistant in the Competence Center DigITools at the Fraunhofer IPA.

ANDREAS SCHLERETH received his bachelor's and master's degree in industrial engineering and management from Karlsruhe Institute of Technology. Since 2018, he has been working as a research fellow in the Competence Center DigITools at the Fraunhofer IPA.

ALEXANDER SAUER is director of the Fraunhofer Institute for Manufacturing Engineering and Automation IPA and head of the Institute for Energy Efficiency in Production EEP at the University of Stuttgart. He specializes in resource-efficient production and digitization for sustainable production.

MIRIAM JÄGER is currently pursuing her Bachelor of Engineering in Industrial Engineering - Product Engineering at the University Furtwangen. During her internship at Fraunhofer IPA, she worked on literature research.

Digitale Transformation in Echtzeit: Die Ziele von morgen basierend auf dem Datenmodell von gestern

Merveille Nangmo Wanko BSc.[1]
Bernhard Zeller[1]
Professor Dr. Frank Herrmann[2]
[1]Maschinenfabrik Reinhausen
[2]Ostbayerische Technische Hochschule Regensburg
E-Mail: [1]nwmerveille@yahoo.fr, [1]b.zeller@reinhausen.com und [2]Frank.Herrmann@OTH-Regensburg.de

SCHLÜSSELWÖRTER
Digitale Transformation, Strategie, ALV-GRID, Dynpro

ABSTRACT

Die Digitale Transformation fordert Unternehmen aller Couleur. Ironischer Weise sind es gerade die bisher verwendeten IT-Systeme mit ihren starren Strukturen, die Unternehmen in Ihrer digitalen Transformation oft ausbremsen. Auch wenn die Softwarehersteller längst reagiert haben und neue, flexiblere Versionen ihrer Produkte anbieten, so ist ein größerer Softwarewechsel immer noch eine Herausforderung für Unternehmen und ein Schritt der wohlüberlegt und geplant sein will.
In dieser Arbeit wird deshalb ein Vorgehen vorgestellt, um mittels In-Memory Technologie und Virtualisierung zumindest die wichtigsten Ergebnisse der Transformation bereits auf den bestehenden Datenmodellen in Echtzeit zu generieren. Dadurch wird genug Zeit gewonnen, um die eigentliche Transformation der IT-Landschaft geplant und mit der notwendigen Sorgfalt durchzuführen.

Digitale Transformation als Treiber

Die zunehmende Digitalisierung revolutioniert unseren Alltag – sowohl im privaten wie im geschäftlichen Umfeld. In immer kürzeren Zeitabständen tun sich Chancen auf und immer schneller tauchen neue Mitbewerber auf.
Um diese Digitale Transformation zu meistern reicht es nicht aus auf diese Veränderungen zu reagieren. Um Erfolg zu haben müssen die Unternehmen diesen Wandel aktiv mitgestalten.
Die Maschinenfabrik hat deshalb schon früh begonnen sich mit den neuen Technologien und den dafür notwendigen Prozessen und Organisationsformen zu beschäftigen und damit die Digitale Transformation einzuleiten.
Eine solche Transformation eines weltweit operierenden Unternehmens ist aber kräftezehrend und in der Natur der Sache bedingt immer zu langsam, zumal die Ziele zu Beginn oft noch unscharf sind.
Um hier aktiv zu gestalten und nicht Getriebener zu werden ist es notwendig das Heft des Handelns in der Hand zu behalten. Im Falle der Maschinenfabrik Reinhausen heißt das sehr wohl Diversifikation und Indeenvielfalt zu fördern, als wichtige Quellen für wertvollen Input. D.h. aber auch zum richtigen Zeitpunkt die Kräfte wieder zu sammeln und auf einige wesentliche Kernthemen zu fokussieren, um diese wichtigen Themen voranzutreiben.
Der technische, in der IT Umsetzung verwendete Begriff für diese Kernthemen ist bei der Maschinenfabrik Reinhausen „*Strategische Geschäftsfelder"*, weshalb in der späteren Beschreibung der techn. Umsetzung dieser Begriff verwendet wird. Strategische Geschäftsfelder ist aber eher ein technischer Arbeitstitel. Im Folgenden werden wir aber beim Begriff Kernthema bleiben, da er besser passt.
Kernthemen haben wenig mit der vorhanden Unternhemensstruktur zu tun haben. Es sind die Antworten der MR auf die sich ständig wandelnden Bedürfnisse, sowohl aus Marktsicht, wie auch aus technologischer Sicht. Dabei muss nicht immer die Umsatzgenerierung im Vordergrund stehen, sondern es kann auch der gezielte Erkenntnisgewinn in manchen Bereichen sein.
Auszeichnungen wie der Industrie 4.0 Award, der Red Dot Design Award und die Erfolge bei den Great Place to Work Umfragen zeigen, dass dieses Werkzeug auf allen Ebenen greift – Technologisch, wie auch Organisatorisch.
Zunehmend wird aber deutlich, dass die IT-technische Umsetzung dieses Konzeptes an Grenzen stößt. Um auch hier agieren zu können statt zu reagieren sind moderne Lösungen notwendig.

IT Schifffahrt: Schnellbote und Tanker

Die Digitalisierung beschleunigt alles und verlangt immer mehr Flexibilität und Umsetzungsgeschwindigkeit. Ironischer weise sind es gerade die IT-Systeme, die in den Unternehmen genau diese Digitalisierung ausbremsen.
Die in heutigen Unternehmen eingesetzten ERP Systeme wurden in den achtziger und neunziger Jahren entwickelt und waren dafür konzipiert ein stabiles Geschäftsmodell abzubilden und tausende, manchmal sogar millionen von Transkationen in kurzer Zeit abzuarbeiten. Es waren also eher Tanker, die vorgefertigte Routen befuhren und dabei Unmengen von Rohstoffen beförderten. Ein Richtungswechsel war aufwändig - vom Bau eines neuen Tankers ganz zu schweigen.
Heute sind eher Schnellbote gefragt, die schnell gebaut sind und schnell neu ausgerichtet. Sollte doch mal mehr Transportkapazität gefragt sein, so wird diese kurzfristig gemietet, z.B. als Cloud Service.
Die Softwarehersteller haben längst reagiert und bieten mittlerweile neue ERP Systeme an, die genau dieses

Bedürfnis nach Flexibilität befriedigen. Allerdings ist für ein Unternehmen der Umstieg von einer Tankerflotte auf eine Flotte von Schnellboten nicht mal eben schnell gemacht. Das Bedarf der Planung und Vorbereitung um den täglichen Betrieb nicht zu stören.
Um diesen gordischen Knoten zu durchschlagen hat die MR nach Wegen Gesucht die wichtigsten Aspekte der Kernthemen auf der bestehenden IT-Landschaft abzubilden und so Zeit für den Umbau der IT-Landschaft zu gewinnen.

Das virtuelle Produkt

Um zumindest die wichtigsten Aspekte der Kernthemen mit dem Datenmodell der bisher genutzten IT Systeme abbilden zu können ist eine Umformung der Daten notwendig. Diese Umformung muss beliebig änderbar sein, in Echtzeit geschehen und bis zu einem gewissen Grad robust gegen Änderungen des darunter liegenden Datenmodells sein.
Im Prinzip will man ähnlich wie beim Kochen verschiedenste Zutaten (Elemente des darunter liegenden Datenmodells) immer wieder zu neuen Gerichten (Aspekte der verschiedenen Kernthemen) kombinieren und vielleicht auch mal eine vorhandende Zutat ändern (geraspelt statt geschnitten), neue Zutat hinzufügen und auch mal die Zubereitungsart ändern. Im Vordergrund steht also das Rezept, das aus den Zutaten was Leckeres macht.
In der IT Welt ist dieser Gedanke des Rezeptes in Form der Konfiguration von System und Produkten weit verbreitet. Insbesondere in der Konfiguration von Produkten besitzt die MR tiefes Wissen. Es lag deshalb nahe auch die Umformung des zugrundeliegenden Datenmodells über die Bereitstellung einer entsprechenden Konfiguration zu steuern. Um flexibel zu bleiben, sollte das Ergebnis aber kein reales Produkt sein, sondern vielmehr eine rein virtuelle Abbildung.
Um all diese Umformungen in Echtzeit durchführen zu können beschäftigte sich die MR früh mit der In-Memory Technologie, insbesondere mit SAP HANA. Mit Hilfe dieser Technologie war es möglich die vorher in einer entsprechenden Konfigurationsmatrix festgehaltenen Umformungen in Echtzeit durchzuführen und so die Ergebnisse einer späteren tatsächlichen Transformation der zugrundeliegenden Systeme in Teilaspekten vorwegzunehmen und einen Blick auf die Zukunft zu ermöglichen. Diese Virtuelle Transformation dient als Grundlage für verschiedenste Steuerungsmaßnahmen (z.B. ein Plan/Ist-Reporting).
Auch wenn dieses Vorgehen bereits einen Erfolg darstellt, so ist es für die Bedürfnisse einer digitalisierten Welt immer noch nicht ausreichend, denn bisher müssen die Änderungen an der Konfigurationsmatrix durch die IT vorgenommen werden. Dieser Prozess ist immer noch zu langwierig und fehleranfällig. Deshalb wurde im Rahmen dieser Arbeit ein Tool entwickelt, dass es dem Fachbereich erlaubt die Konfiguration selbständig zu ändern. Dadurch wurde der Prozess der Änderung verkürzt und durch ein verifiziertes Tool abgesichert.

Die Herausforderung bei der Erstellung des Tools war, dass sich viele Rahmenbedingungen ändern können und sich das Tool eigenständig auf diese Änderungen reagieren soll. Deshalb wurde im Tool extensiv die Mittel der dynamischen Programmierung im ABAP Umfeld ausgenutzt. Anders ausgedrückt konnten wir die dem Thema zugrundeliegende Komplexität nicht reduzieren, aber wir haben zumindest das Wissen zur Beherrschung dieser Komplexität aus den Köpfen der Mitarbeiter in ein lauffähiges Programm transferiert. Damit ist das Wissen der Mitarbeiter ist konserviert und die Komplexität an einer zentralen Stelle adressiert.

Die Maschinenfabrik Reinhausen GmbH

Die Maschinenfabrik Reinhausen GmbH mit Hauptsitz in Regensburg ist ein mittelständisches Familienunternehmen. Das Unternehmen MR beschäftigt weltweit 3.300 Mitarbeiter. Im Geschäftsjahr 2016 erwirtschaftete das Unternehmen einen Umsatz von 750 Millionen Euro. Das Kerngeschäft der MR ist die Regelung und Steuerung von Leistungstransformatoren. Auf diesem Gebiet ist die MR Weltmarktführer - 50 % des weltweiten Stroms läuft durch ihre Produkte. Neben der Herstellung von Verbundhohlisolatoren, der Bearbeitung von Kunststoffzylindern beinhaltet das Produktportfolio der MR überwiegend Komplementärprodukte und zahlreiche Dienstleistungen rund um den Transformator [2]. Das macht das Unternehmen im Bereich Energietechnik zu einem kompetenten Ansprechpartner.

Bisherige Abbildung der strategischen Geschäftsfelder der MR mittels vorhandener Datenstrukturen

Um die einzelnen Strategisches Geschäftsfelder der MR steuern zu können ist ein entsprechendes Reporting notwendig. Dieses Reporting muss sich derzeit noch an den vorhanden Datenstrukturen orientieren. Dies sind vorallem Informationen über den genutzten Vertriebsweg, das verkaufte Produkt, die Unterscheidung nach Ersatzteil- und Produktgeschäft, die verkaufende Verkaufsorganisation und die Unterscheidung zwischen Produkt- und Projektgeschäft. Die Kombination aus diesen fünf Elementen bestimmt das SGF bei MR. D.h jede Kombination der möglichen Werte für diese fünf Merkmale wird einem SGF zugeordnet. Die so entsehende Matrix wurde bisher mittels MS Excel abgebildet. Diese Matrix unterstützt sowohl dem Fachbereich als auch der Geschäftsleitung bestmöglich bei der Marktbearbeitung – also dem Identifizieren potenzieller bzw. relevanter Märkte und wer zum Wettbewerbsumfeld gehört.
Die Matrix ermöglicht Umsatzauswertungen und COPA-Auswertung (Kosten, Deckungsbeiträge) nach SGF darzustellen..
Das bisherige Vorgehen zum erstellen der SGF Matrix war dabei bisher folgendes: Nach Angaben der Fachbereiche wurden folgenden Informationen erhoben:

Die SGF-Matrix wurde manuell mit Hilfe dem Microsoft Office Excel-Programm aufgestellt. Dies ist zeitaufwändig für die Mitarbeiter, die diese Matrix erstellen müssten.

Im Schnitt werden alle 2 - 3 Jahre neue Geschäftsfelder an der Matrix hinzugefügt. In 5 Jahren wurden im Schnitt 62 Änderungen durchgeführt. Bei jeder Änderung an der Matrix müssen wieder alle Informationen von Hand sowohl in der Matrix eingetragen werden als auch im Business Warehouse (BW)-System angepasst werden. Dieser Prozess war langwierig und fehleranfällig.

Die Matrix ist aufgrund neuer Geschäftsfelder mit der Zeit sehr komplex geworden. Um alle Informationen im Überblick behalten zu können, wurden Geschäftsfelder mit gemeinsamen Merkmalen in der Kopfzeile der Matrix gruppiert. Folglich verliert die Matrix an Ihrer Übersichtlichkeit, so dass sich die Auswertung der Umsätze und die Informationsgewinnung zu Kosten und Deckungsbeiträge anhand der in Microsoft Excel erstellten Matrix erschwert haben.

Die derzeitige Möglichkeit die SGF mittels Microsoft Excel Tabelle darzustellen und diese zu bearbeiten ist sehr komplex. Das verlangt zu viel Zeit und Ressourcen. Um aus dieser Komplexität rauszukommen, soll diese fehleranfälligen manuellen Tätigkeiten durch ein neues Berichtswesen-Tool ersetzt werden. Dafür wird ein Programm benötigt, das ein Cockpit – Benutzeroberfläche – implementiert und die Wünsche des Fachbereiches in das Cockpit bereitstellt.

Begriffsdefinitionen

In den folgenden Abschnitte können die ganze Syntax und alle Sprachelemente der Programmiersprache ABAP nicht erklärt, da das auch nicht möglich wäre. Stattdessen werden die wichtigsten Sprachelemente, die für die Implementierung des Cockpits nützlich sind, erläutert.

ABAP List Viewer

Laut [4] ist der *ABAP List Viewer* (ALV) ein Tool zur Anzeige von Massendaten, in Tabellen- und/oder Baumdarstellung. Der ALV bietet eine Reihe von interaktiven Standardfunktionen wie zum Beispiel: Drucken, Aufsteigend/Absteigend sortieren, Filter setzen, Exportieren und Grafik anzeigen. Diese Standardfunktionen können von den Entwicklern ausgeblendet, angepasst oder ergänzt werden, und brauchen nicht mehr implementiert zu werden. Zwei Programmiermodelle stehen zur Verfügung: das ALV Grid Control (verfügbar seit *SAP R/3 4.6*) und das ALVm Object Model (ab *SAP NetWeaver 2004*). Die Implementierung des Programmiermodells ALV Grid Control, welches in einen Container einzubetten ist, wird in dieser Bachelorarbeit ausgeführt.

Das zu implementierende Programmiermodell – ALV-Grid-Control – verwendet die Klasse *CL_GUI_ALV_GRID*. Diese Klasse hat eine Methode set-table-for-first-display, mit der die Listenausgabe formatiert wird und Daten an der Control übergeben werden. *Controls* sind Bereiche in einem Dynpro, die selbst als Grundlage für die Visualisierung des ALV-Grids dienen.
Die Methode *set_table_for_first_display* bietet drei wesentlichen Parameters:

- zu einem die Ausgabelisten bzw. die Ergebnisliste, die an den Parameter it_outtab übergeben werden,
- zu anderen ein Feldkatalog, der an den Parameter it_feldcatalogue übergeben wird,
- und eine Layout–Struktur – Voreinstellungen für die ALV-Ausgabe– , die an der Parameter is_layout übergeben wird.

Feldkatalog

Der Feldkatalog ist eine interne Tabelle mit Informationen über die darzustellenden Felder. Mit Hilfe dieser Tabelle wird zum Beispiel festgelegt, welche Felder beziehungsweise Spalten auf dem ALV-Grid anzuzeigen sind, welche Datentypen die Felder haben etc. Spalteneigenschaften der auszugebenen Liste kann über Felder des Katalogs beeinflusst werden.

Selektionsbild

Ein Selektionsbild ist ein spezielles Dynpro (Dynamisches Programm), das ohne Verwendung des Screen Painters mit den Anweisungen *SELECTION–SCREEN*, *PARAMETERS* und *SELECT–OPTIONS* im globalen Deklarationsteil von ausführbaren Programmen definiert werden kann [5]. Mit Selektionsbildern wird dem Anwender eine Eingabemaske für Wertemengen zur Verfügung gestellt, die zur Einschränkung bzw. Abgrenzung der Datenmenge genutzt wird, die von der Datenbank gelesen werden muss.

Dynpro

Dynpros sind frei definierbare Objekte, bei denen ABAP-Programmierer oder Benutzer mit Hilfe von Ein– und Ausgabefeldern, Listen usw. Informationen anzeigen oder eingeben können. Dynpros sind eine Form des Dialogs zwischen dem Benutzer und dem ABAP-Programm. Ein Dynpro wird im Programm durch eine eindeutige viertstellige Kennung – Dynpronummer genannt – identifiziert. Das Dynpro ermöglicht es dem Benutzer, Daten zu erfassen und anzuzeigen. Auf dem Dynpro können mit Hilfe von Drucktasten, Tabstrips, Table Controls und weiteren graphischen Elementen einen benutzerfreundlichen Dialog programmiert werden. Das Dynpro dient hauptsächlich als Träger weiterer Bildelemente [6]. Mit dem Screen Painter – Werkzeug der ABAP Workbench – können Dynpros zu einer Transaktion angelegt werden und mit Hilfe

von Bedien– und Anzeigeelementen definiert und gestaltet werden.
Ein Dynpro besteht aus einem Bildschirmbild und der zugrundeliegenden Ablauflogik. Die Ablauflogik ist ein Programm, das die Verarbeitung des Bildschirmbilds steuert. Die Ablauflogik eines Dynpro ist zumindest in zwei Ereignisblöcke aufgeteilt: Process Before Output (PBO) und Process After Input (PAI). Die Ereignisse PBO und PAI werden iterativ ausgelöst, d.h. nachdem das Ereignis PAI ausgelöst und die Funktion ausgeführt wurde, wird danach wieder automatisch das Ereignis PBO ausgelöst, so dass wieder der Programmcode in PBO ausgeführt wird.

Process Before Output (PBO)

Der Verarbeitungsblock zum PBO–Ereignis ist die Stelle, die abgearbeitet wird, bevor das Dynpro aufgerufen wird, bzw. auf dem Bildschirm erscheint. Der PBO–Verarbeitungsblock dient der Vorbereitung und Initialisierung von Programmvariablen und Bildschirmfeldern.

Process After Input (PAI)

Der Verarbeitungsblock zum PAI-Ereignis wird nach der Bearbeitung des Bildschirmbildes aufgerufen, um auf Eingabe – wie etwa das Klicken einer Drucktaste oder eines Buttons – zu reagieren. Der PAI-Verarbeitungsblock dient der Verarbeitung der Benutzereingaben. Jeder Benutzeraktion auf dem Bildschirm ist mit einem Funktionscode verknüpft. In die nachfolgende Grafik, wird der Zusammenhang zwischen Dynpro, PBO Und PAI aufgezeigt.

Funktionscode

Ein Funktionscode wird generell durch Betätigung eines Bedienelementes (wie z.B.: Schaltflache, Drucktaste) ausgelöst. Jede entsprechende Aktion auf dem Dynpro muss in seinen Attributen einen eindeutigen Funktionscode haben. Wenn ein Anwender auf die Drucktaste klickt, wird dies dem ABAP-Programm durch die Bereitstellung des Funktionscodes im Systemfeld syucomm signalisiert. Dies kann im Ereignisblock User-Command abgefragt werden und behandelt werden.

Container

Im ABAP.Programm wird der im Dynpro-Editor angelegte Bereich auf der Maske von einem so genannten Container verwaltet. Container sind spezielle Klasse im R/3System, deren Aufgabe ausschließlich darin liegt, derartige Bilschirmbereiche zu verwalten. Controls (Zum Beispiel: ALV-Grid, ALV-Tree etc.), die im System verwendet werden können, müssen über Container verwaltet werden. Da Container selbst auch Controls sind, ist es möglich, Container in Container einzufügen und somit den Custom–Bereich logisch zu unterteilen. [7].
Im SAP werden fünf verschieden Typen von Containern unterschieden, die jeweils durch eine ABAP–Klasse repräsentiertwerden. Die Klasse CL_GUI_CUSTOM_CONTAINER, welche in vorliegende Bachelorarbeit verwendet wird, verwaltet einen Custom–Control–Bereich auf einem Dynpro.
Bei der Gestaltung des Cockpits sollen Selektionsbild und ALV–Grid in einem selben Dynpro unterbringen. Diese Eigenschaft des zu implementierenden Cockpits ist auf Komfort der Benutzer gedacht. Hierfür kommen zwei Techniken im Einsatz: Container Control und Subscreen. *Subscreens* sind Bereiche auf dem Bildschirm, in denen andere Dynpros eingebettet werden können.
Container Control sind rechteckige Flächen auf dem Bildschirmbild eines Dynpros, die mit Namen versehen werden. Selektionsbilder sollen erst als Subscreen definiert werden und dann an ein Dynpro eingebunden werden. Das ALV–Grid soll mit Hilfe des Container Controls an dasselbe Dynpro gebunden werden.

Alle Änderungen in der Ergebnisliste sollen durch den Fachbereich ohne den IT-Eingriff erfolgen. Dafür muss die Ergebnisliste aus dem ALV-Grid zur Laufzeit generiert werden bzw. dynamisch gestaltet werden. Die Programmiersprache ABAP bietet schon Techniken zur dynamischen Programmierung, die diese Anforderung erfüllen. Datenreferenzen, Feldsymbole in Verbindung mit generischen Datentypen spielen eine große Rolle bei der dynamischen Programmierung.

Feldsymbole

Ein Feldsymbol ist ein symbolischer Name für ein Datenobjekt, dem zur Programmlaufzeit ein konkreter Speicherbereich zugewiesen werden kann. Über ein Feldsymbol wird auf ein zugewiesenes Datenobjekt durchgegriffen, das heißt alle Zugriffe werden mit dem zugewiesenen Datenobjekt durchgeführt [8]. Mit der Anweisung *FIELD-SYMBOLS* wird ein Feldsymbol vereinbart.

Datenreferenzen

Datenreferenzen sind Adressen von Datenobjekten. Um auf den Inhalt eines Datenobjekts zugreifen zu können, auf das eine Datenreferenz verweist, muss dieses zuerst dereferenziert werden. Die Dereferenzierung erfolgt über den Dereferenzierungsoperator (–>*): ASSIGN dref –>* TO <fs>. Diese Anweisung weist ein Feldsymbol <fs> das Datenobjekt zu, auf das die Datenreferenz in der Referenzvariable dref zeigt.

Interne Tabellen

Interne Tabellen bieten in ABAP die Funktionalität von Arrays. Ein wichtiges Einsatzgebiet ist z.B. die pro-

gramminterne Speicherung und Aufbereitung von Inhalten aus Datenbanktabellen. [9]. Interne Tabellen werden zeilenorientiert verarbeitet. Dabei werden die Daten über einer Schleife (LOOP…ENDLOOP) zeilenweise im Speicher abgelegt, wobei jede Zeile der internen Tabelle die gleiche Struktur bzw. den gleichen Arbeitsbereich hat.

In dieser Arbeit sollen weiterhin Funktionen implementiert werden, die dem Fachbereich die Möglichkeit geben, die Ergebnisliste zu manipulieren. Zudem werden Sprachelemente und Syntax zur Datenbankänderungen Programmierung verwendet. Für das Programmieren von Datenbankänderungen stehen dem Entwickler im ABAP die Befehle des Open SQL zur Verfügung: INSERT (Einfügen), UPDATE (Aktualisieren), MODIFY (Ändern), DELETE (Löschen).

Anforderungsdefinitionen

Um die Anforderungen an das zu implementierende SGF-Cockpit zu erfassen, wurden vor Ort Workshops mit dem Fachbereich durchgeführt. In diesen Workshops wurden die Wünsche an das zukünftigen Berichtswesen-Tool bzw. Cockpit erfasst. Daraus ergeben sich folgenden Anforderungen.

Das Cockpit soll:

- dem Anwender die Möglichkeit geben, die SGF in einer übersichtlichen und strukturieten Form ansehen zu können. Dafür soll ein ALV-Grid-Control implementiert werden.
- dem Anwender die Möglichkeit geben, die im ALV-Grid bestehenden Daten abfragen oder filtern zu können. Hierfür soll das SGF-Cockpit mit einem Selektionsbildbereich ausgestaltet werden.
- Dem Anwender die Möglichkeit geben, Einträge bzw. Zelleninhalte der Ergebnisliste ändern und speichern zu können. Dafür soll in der Ergebnisliste, die entsprechenden Zellen mit Dropdown-Listbox für das Auswählen der Daten ausgerichtet werden. Weiterhin soll das Diskette-Symbol – zum Sichern des geänderten Eintrags – in der Symbolleiste des Einstiegsdynpros aktiviert werden.
- Dem Anwender die Möglichkeit geben, die Ergebnisliste erweitern oder löschen zu können. Zudem sollen zwei weiteren Dynpros in das Einstiegsdynpro eingebettet werden. Diese Dynpros sollten dem Fachbereich Eingabefelder und Schaltflächen/Buttons für das Einfügen und Löschen von Datensätze bzw. Datenfelder in der Ergebnisliste anbieten.

Die durchzuführende Änderungen sollen nur auf der Ergebnisliste und nicht auf der Datenbank erfolgen!

Um die Akzeptanz des Cockpits gegenüber dem Anwender zu erhöhen, soll das zu implementierende SGF-Cockpit weiterhin folgenden Qualitäten-Anforderungen erfüllen:

Das Cockpit soll erweiterbar, zuverlässig und benutzbar sein.

Die nachstehenden Use-Case-Diagramm fasst das Cockpit in seinen Funktionalitäten zusammen.

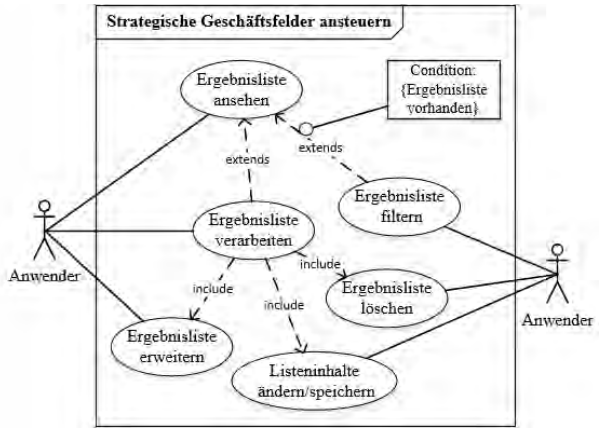


Abbildung 1: Use-Case Diagramm zum Ansteuern der SGF.

Um die Kommunikation bzw. Interaktion zwischen dem SGF-Cockpit und dem Fachbereich zu ermöglichen, werden Dynpros benötigt. Dafür müssen diese vorher zu einem Programm definiert werden. Dabei ist neben dem Layout bzw. Bildschirmmaske auch die sogenannte Ablauflogik des Dynpros zu realisieren.

- Die Layouts der anzulegenden Dynpros werden in der Entwurfsphase dieses Artikels vorgenommen.
- Die Ablauflogiken der entworfenen Dynpros werden in der Implementierungsphase realisiert.

Entwurfsphase

Ziel des Entwurfs ist es, die definierten Anforderungen an das Cockpit in eine „bildlich sichtbare“ [Benutzeroberfläche umzusetzen. In diesem Kapitel wird der Aufbau des Cockpits in seinen wesentlichen Elementen sowie dessen Architektur beschrieben und anhand diversen Screenshots ein erster Eindruck des SGF–Cockpits vermittelt. Zu Beginn werden die in diesem Projekt zu verwendenden Tabellen vorgestellt.

Benötige Datentabelle

Ein Ziel dieser Arbeit ist die SGF in Form eines ALV-Grid-Controls darzustellen. In dem ALV-Grid-Control müssen Daten aus drei Datenbanktabellen angezeigt werden. Diese sind: ZSGF_DATA, T179t und T25A7. Diese Datenbanktabellen sind alle miteinander über Fremdschlüssel verbunden.

Die wichtigste Datenbanktabelle ist ZSGF_DATA. Anhand dieser sollen Hauptdaten für die Darstellung des ALV-Grid-Controls beschafft werden. Bei der Transformation bzw. Verarbeitung der Ergebnisliste soll auf diese Datenbanktabelle verwiesen werden. Alle Daten der Datenbanktabelle ZSGF_DATA sind mit SAP-Schlüsseln gespeichert. Bei der Implementierung des ALV-Grid-Controls sollen aber einigen Daten mit entsprechenden Namen angezeigt werden. Um den entsprechenden Text bzw. Name zu diesen SAP-Schlüsseln zu haben, wird auf die Tabelle T25A7 und T179t verwiesen.

Die Datenbanktabelle T179 enthält alle mögliche Produkthierarchie, die es in der MR gibt. Sie wird für die Prüfung von Benutzereingabe verwendet.

Allgemeiner Aufbau

Damit das SGF-Cockpit einen einheitlichen und übersichtlichen Aufbau besitzt, empfiehlt es sich, dies in drei Benutzeroberflächen bzw. Dynpros zu untermauern. Jede Benutzeroberfläche im SAP-System wird über Dynpros aus realisiert. Dafür soll für jede Benutzeroberfläche eine eigene Dynpro angelegt werden. Somit sollen folgenden Dynpros in diesem Kapitel designt werden:

- Dynpro_0100 – Einstiegsdynpro – für das Einstiegsbild des SGF-Cockpits,
- Dynpro_9001 für das Bearbeiten der Datensätze des ALV-Grid-Outputs
- Dynpro_9002 für das Bearbeiten der Datenfelder des ALV-Grid-Outputs.

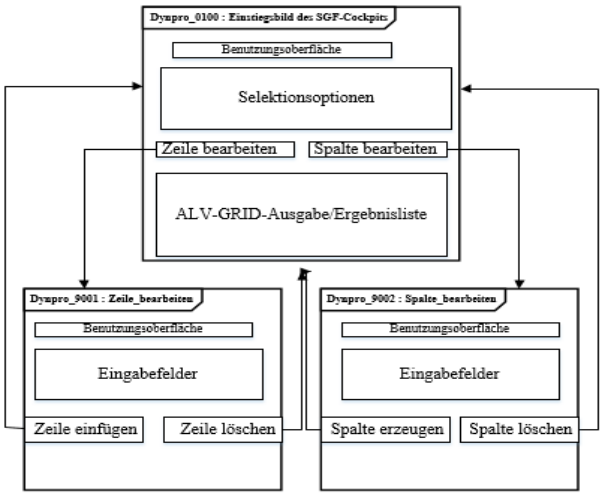


Abbildung 2: SGF-Cockpit-Architektur.

Die Abbildung 2 zeigt die Beziehung zwischen den Dynpros an. Die Pfeile Richtung verdeutlicht den Wechsel zwischen den jeweiligen Dynpros. z.B: klickt der Fachbereich auf die Taste “Zeile bearbeiten” in dem Dynpro_0100, wird er auf dem Dynpro_9001 gelandet und kann von dort aus einen Datensatz einfügen oder löschen. Das Verlassen eines Dynpro könnte auch über die GUI-Status erfolgen. Dafür müssen GUI-Status der jeweiligem Dynpro im Menü Painter festgelegt werden.

GUI-Status gestalten

Damit die dargestellten Dynpros verlassen werden können und die vom Benutzer geänderte Ergebnislisteninhalt gespeichert werden kann, sollte noch je Dynpro ein GUI-Status eingebaut werden, der bei Anzeige des Dynpros vorhanden ist. Dazu wird im PBO-Modul jedes Dynpro ein eigenes GUI-Status angelegt. In den Funktionstasten des Dynpro_0100 wurden nur nur die Speichern, Abbrechen, Zurück und Beenden Buttons aktiviert. Will das SGF-Cockpit auf diese Buttons reagieren so müssen diese Funktionen durch eigene Funktionscode belegt werden. Das ganze sollte dann folgendermaßen aussehen:

Abbildung 3: GUI-Status_0100: Funktionstastenleiste anlegen.

Einstiegsbild des SGF-Cockpits gestalten

Auf dem Dynpro_0100 (Vgl: Abb.2) sollen folgenden Bereiche reserviert werden: Ein Subscreen-Bereich für Selektionsbild, Drucktasten – Zeile bearbeiten und Spalte bearbeiten –, und ein Container-Bereich. Der Container-Bereich dient selbst als Grundlage für die Visualisierung des ALV-Grid-Controls.

Ansicht Spalte_bearbeiten

Die Ansicht Spalte_bearbeiten wird über das Dynpro_9002 gestaltet. Auf diesem Dynpro soll dem Fachbereich angeboten werden, Datenfelder zu der Ergebnisliste erweitern bzw. löschen zu können.

Um eine Spalte in der Ergebnisliste einzufügen oder zu löschen, werden folgenden Elemente benötigt:

- ein Eingabemaske für Spaltenangaben
- 2 Drucktasten: “Spalte einfügen” und “Spalte löschen”

Um die oben aufgelisteten Elementen in das Dynpro 9002 zu platzieren, muss in der Symbolleiste des Screen Painters auf der Drucktaste Dictionary/Programmfelder–Fenster doppelgeklickt werden.
Das Fenster Dict/Programmfelder wird geöffnet. Danach muss ein Tabellenname eingetragen werden und die Schaltfläche „*holen aus Dict*“ betätigen. Die Felder der Tabellenname erscheinen in einer Liste. Die benöti-

ge Felder müssen markiert und per Drag&Drop in die leere Bildschirmmaske platziert werden.

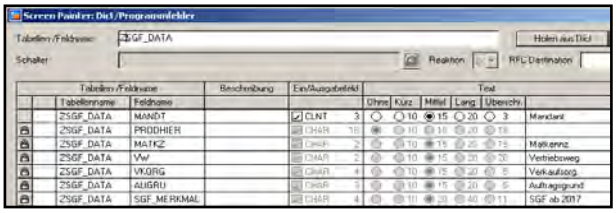


Abbildung 4: Auswahl von Eingabefeldern in das Dynpro 9002.

Das Platzieren von Drucktasten auf die Bearbeitungsfläche des Dynpro_9002 erfolgt über die Werkzeugleiste des Screen Painters. Dabei wichtig ist die Zuordnung von Funktionscoden an diesen Drucktasten. Somit können sie zum PAI-Ereignis und durch Benutzeraktion abgefangen und aufgelöst werden.
Die Drucktaste „Spalte einfügen“ wurde auf dem Dynpro mit dem Funktionscode „CREATE“ angelegt und wird zum PAI-Ereignis ementsprechen abgefragt. Die Drucktaste „Spalte löschen“ wird im Programm über den Funktionscode „DELETE“ identifiziert werden. Die folgende Graphik zeigt die Anordnung der Objekte des Dynpro_ 9002. Auf Komfort der Fachbereich wurde alle Objekte eingerahmt und betitelt.

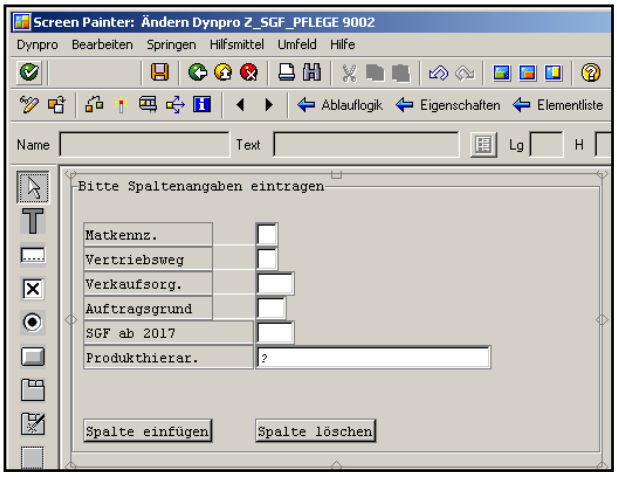


Abbildung 5: Platzieren von Eingabefelder und Drucktaste auf dem Dynpros 9002.

Die Ansicht Zeile_bearbeiten durch das Dynpro_9001 repräsentiert, wurde analog aufgebaut. Aber dort wird nur ein Eingabefeld “Produkthierarchie” benötigt. Da in der Datenbanktabelle ZSGF_ DATA wird eine Zeile durch das Feld Produkthierarchie eindeutig identifiziert.

Die Folgennde Abbildung zeigt das Ergebnis der Konfiguration des Ansichts Zeile_bearbeiten.

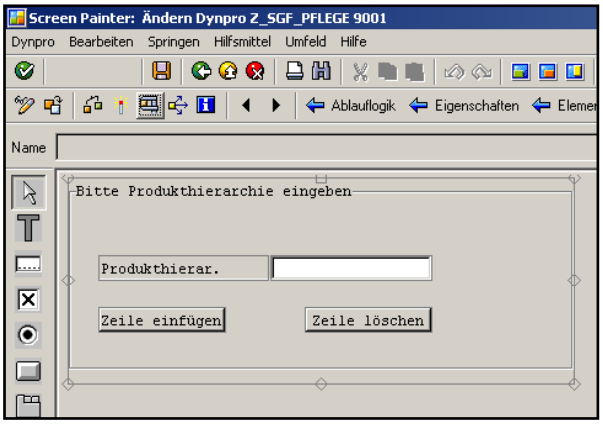


Abbildung 6: Konfiguration Ansicht Zeile_bearbeiten.

Nachdem die Layouts der Dynpros in der Entwurfphase realisiert wurden, werden in der Implementierungsphase die entsprechenden Ablauflogiken implementiert.

Implementierungsphase

Im folgenden Kapitel wird auf die Konkrete Umsetzung der definierte Anforderungen an das Cockpit eingegangen. Dabei wurden mehrere Form-Routinen auch Unterprogramme genannt implementiert. Für eine detaillierte Beschreibung der implementierung alle Form-Routinen wird auf die vollständige Bachelorarbeit verwiesen.

ALV-Grid-Control-Ausgabe

Zur ALV-Grid-Control-Ausgabe werden folgende Schritte durchlaufen:
und

Feldkatalog erstellen

Neben der Deklaration von Feldsymbole, interne Tabellen und Workareas – auch Struktur genannt – Typ für ALV-Ausgabetabelle, soll weiterhin ein Feldkatalog aufgebaut werden. Es wird bei dem Aufbau des Feldkataloges festgelegt, welche Spalten auf dem ALV anzuzeigen sind. Für die Erstellung des Feldkatalogs wurde eine Form-Routine Fieldcat_Build implementiert. Dabei erfolgte die Erstellung manuell. Der Feldkatalog wurde in Form einer Tabelle lt_fcat_temp vom Typ lvc_t_fcat aufgebaut und entsprechend der Wünsche von dem Fachbereich angepasst. Diese Tabelle wird später an die Methode *set_table_for_first_display* übergegeben
In der Abbildung 7 wird einen Codeabchnitt zum Feldkatalog Aufbau aufgezeigt.

```
*prodh TYPE c LENGTH 18,
 CLEAR LS_FIELDCAT.
 LS_FIELDCAT-COL_POS = 1.
 LS_FIELDCAT-FIELDNAME     = 'PRODH'.
 LS_FIELDCAT-KEY_SEL = 'X'.
 LS_FIELDCAT-EMPHASIZE = 'C500'.
 LS_FIELDCAT-DATATYPE = 'C40'.
 LS_FIELDCAT-OUTPUTLEN = 30.
 LS_FIELDCAT-FIX_COLUMN = 'X'.

 MOVE 'Produkthierarchie'(001) TO:
  LS_FIELDCAT-COLTEXT,
  LS_FIELDCAT-REPTEXT,
  LS_FIELDCAT-SCRTEXT_M,
  LS_FIELDCAT-SCRTEXT_S,
  LS_FIELDCAT-SCRTEXT_L.
 LS_FIELDCAT-STYLE = 32.
 APPEND LS_FIELDCAT    TO    LT_FCAT_TEMP.
```

Abbildung 7: Abchnittscode Feldkatalog aufbau.

Diese Abbildung zeigt, welche Ausgabeoptionen bzw. Eigenschaften das Feld ‚PRODH' in der Ergebnisliste haben soll. z.B das Feld ‚PRODH' soll als erste Spalte in dem ALV-Grid dargestellt werden, soll fixiert werden und soll grün gefärbt gefärbt. Über die Zeile der Feldkatalog-Tabelle wurden weitere Felder der Ergebnisliste beschrieben.

Weiterhin wurde über eine Layout-Struktur weitere Einstellungen für das Layout des ALV-Grid vorgenommen. Dabei wurden Feldnamen des Ausgabefelds für die Farben von Zeilen und Zellen, den Selketionsmodus und für die Optimierung der Spaltenbreite gesetzt.

Dynamischer Aufbau der Ergebnisliste

Es soll in dieser Arbeit dem Fachbereich die Möglichkeit geben, die Ergebnisliste aus dem ALV-Grid zu manipulieren. Hierfür ist es notwendig und sinvoll diese Liste zur Laufzeit zu generieren. Dafür wurde die Methode *cl_alv_table_create=>create_dynamic_table* verwendet (siehe Abb.Abbildung 8). Mit ihre wurde anhand des erstellten Feldkataloges die Ausgabeliste bzw. die Ergebnisliste, die später an den Parameter it_outtab der Methode *set_table_for_first_display* übergeben werden, aufgebaut. Die Bearbeitung dieser dynamischen erzeugten Tabelle erfolgt nicht mehr über Spaltenname sondern über Feldsymbole. Die einzelnen Felder des Arbeitsbereiches <dyn_wa> dieser Tabelle <gt_data> werden über den Namen der Felder mittels ASSIGN COMPONENT-Anweisung, angesprochen.

```
Create dynamic table
CALL METHOD CL_ALV_TABLE_CREATE=>CREATE_DYNAMIC_TABLE
  EXPORTING
    I_STYLE_TABLE             = 'X'
    IT_FIELDCATALOG           = LT_FCAT_TEMP
    I_LENGTH_IN_BYTE          =
  IMPORTING
    EP_TABLE                  = GT_DYN_ITAB.
ASSIGN GT_DYN_ITAB->* TO <GT_DATA>.
Workarea for the dyn. table
CREATE DATA D_LINE LIKE LINE OF <GT_DATA>.
ASSIGN D_LINE->* TO <DYN_WA>.
```

Abbildung 8: Dynamische interne Tabelle erzeugen.

ALV-Grid aufbauen

Im PBO-Modul des Dynpro_0100 wurden Referenzvariablen für das ALV Grid Control und den Custom_Container deklariert und instanziiert. Die Instanzierung erfolgte über das CREATE OBJECT-Befehl. Über den Exporting Parameter *parent* wurde das Container Control als Vater des ALV-Grid festgelegt. Dabei wird das Container Control an das Dynpro über den im Screen Painter angelegten Container gebunden.

Nachdem alle Einstellungen für die Ausgabe des ALV-Grid-Controls vorgenommen wurde, wurden die Ausgabetabelle bzw. die Ergebnisliste (gt_data) und die Strukturdaten (Layout und Feldkatalog) an das ALV-Grid über die Methode *set_table_for_first_display* übergegeben.

1. Selektionsbild aufbauen

Selektionsbild soll dem Benutzer erlauben, Parameters bzw. Selektionskriterien für die Datenselektion in der Ergebnisliste anzugeben. Dabei soll beachtet werden, dass Selektionsbild als Subscreen (Teildynpro) definiert werden. Abbildung 9 zeigt die Definition der Selektionskriterien im Selektionsbild. Diese Definition erfolgt über die SELECT-OPTIONS-Anweisung.

```
* Custom Selection Screen 1010
SELECTION-SCREEN BEGIN OF SCREEN 1010 AS SUBSCREEN.
SELECTION-SCREEN BEGIN OF BLOCK B1 WITH FRAME TITLE TEXT-004.
SELECT-OPTIONS   S_PRODH FOR ZSGF_DATA-PRODHIER .
SELECT-OPTIONS   S_MATKZ FOR ZSGF_DATA-MATKZ  .
SELECT-OPTIONS   S_VW FOR ZSGF_DATA-VW.
SELECT-OPTIONS   S_VKORG FOR ZSGF_DATA-VKORG.
SELECT-OPTIONS   S_AUGRU FOR ZSGF_DATA-AUGRU.
SELECTION-SCREEN END OF BLOCK B1.
SELECTION-SCREEN END OF SCREEN 1010.
```

Abbildung 9: Aufbau eines Selektionsbildes.

Wenn ein als Subscreen definiertes Selektionsbild in einem Dynpro **(**Dynamisches Programm**)** eingebunden werden soll, ist darauf zu achten, dass in der Ablauflogik des entsprechenden Dynpros sowohl zu PBO als auch zu PAI die Anweisung CALL SUBSCREEN ausgeführt werden muss, um die Daten zwischen Selektionsbild und ABAP-Programm zu transportieren.

2. Inhalte der Ergebnisliste ändern und Speichern

Inhalte der Ergebnisliste ändern

Um den Fachbereich das Ändern eines Zelleninhalts bzw. eines Eintrags in der Ergebnisliste zu erleichtern, wurde Dropdown-Listbox erstellt. um aus einer ALV-Zelle eine Listbox zu machen, soll zuerst im Feldkatalog an die entsprechenden Spalte das Feld DRDN_FIELD zugewiesen. In unserem Fall wurde die Spalte "Merkmal" dem Feld DRDN_FIELD zugewiesen.

Weiterhin sollte diese Listbox befüllt werden. Dazu wurde in einer Form-Routine set_dropdown_table einen Arbeitsbereich ls_dropdown mit Bezug auf die Dictionary-Struktur lvc_s_drop und eine interne Tabelle lt_dropdown mit Bezug auf die Dictionary-Tabelle lvc_t_drop angelegt und die relevanten Felder wurden gefüllt, wie etwa ls_dropdown-handle = '1' und ls_dropdown-value = '0070-Antriebe mit ISM'.

Danach wurden diese Felder über der Arbeitsbereich der interne Tabelle lt_dropdown hinzugefügt, etwa `APPEND LS_DROPDOWN TO LT_DROPDOWN.`

Abschließend wurde die Dropdown-Value dem ALV-Grid übergegeben werden.

Erfasste Änderung Speichern

Nachdem der Anwender aus der Dropdownl-Listbox einen Wert ausgewählt hat, muss er die *Enter-Taste* betätigen. Somit wird das Programm mitgeteilt, dass Änderung an einer Zelle der Ergebnisliste vorgenommen wurde. Zum abfangen bzw. zur Behandlung diese Event wurde eine Form-Routine Register_Edit implementiert. Zudem wurde die Methode Register_Edit_Event aufgerufen. Diese Methode stellt sicher, dass den ausgewählten Wert aus der Listbox im Programm vorgemerkt wird, nachdem die Enter-Taste gedrückt wurde.

Durch das Drücken des Sichern-Ikons (Funktionscode &SAVE_DATA) auf dem Dynpro_0100 sollen die geänderten Daten auf die Ergebnisliste geschrieben werden. Im PAI-Modul User_Command_0100 des Dynpro_0100 wird dazu das Unterprogramm Save_Change aufgerufen. Dabei soll zuerst die modifizierte Zelle in dem ALV-Grid erhaltet bzw. identifiziert werden. Zudem erfolgt eine Schleife über die dynamische interne Tabelle <*gt_data*>, um die geänderte Zelle in der Ergebnisliste zu suchen. Wird die entsprechende Zelle gefunden, wird sie über eine Read Table-Anweisung ausgelesen, dann in einer Struktur gespeichert und weiter an einer interne Tabelle angehängt. Der modifizierte Zelleninhalt ist nun bekannt und kann über der Modify-Befehl in der Ergebnisliste gestellt bzw. gespeichert werden.

3. Datensatz einfügen und Datenfeld löschen

In der Entwurfphase wurde den Aufbau der Ansicht Spalte_bearbeiten über das Dynpro_9002 realisiert. Für die Reaktion von Benutzereingaben, muss ein PAI-Modul entwickelt werden (siehe Abbildung 10). In diesem Modul werden die definierten Drucktasten und Funktionen für die Symbolleiste des Dynpro_9002 abgefangen bzw. behandelt.

```
MODULE USER_COMMAND_9002 INPUT.
  SAVE_OK = OK_CODE_SPALTE.
  CLEAR OK_CODE_SPALTE.
  CASE SAVE_OK.
    WHEN 'CREATE'.
      PERFORM INSERT_NEW_COLUMN.
      IF GV_EINGABE_KORREKT = 'X'.
        LEAVE TO SCREEN 100.
      ENDIF.
    WHEN 'DELETE'.
      PERFORM DELETE_COLUMN.
      LEAVE TO SCREEN 100.
    WHEN '&F03' OR '&F15' OR '&F12'.
      LEAVE TO SCREEN 0.
   WHEN OTHERS.
  ENDCASE.
ENDMODULE.
```

Abbildung 10: ABAP-Verarbeitungslogik-Dynpro_9002.

Das Feld save_ok dient der Übernahme von Funktionscodes aus dem Dynprofeld.

Datensatz einfügen

Damit der Benutzer einen Datensatz einfügt, steht ihm auf der Ansicht Zeile_bearbeiten ein Eingabefeld für das Lesen der Produkthierarchie-Nummer und die Drucktaste „Zeile *einfügen"* zur Verfügung. Trägt der Benutzer einen Wert in das Feld Produkthierarchie ein und betätigt die Drucktaste „Zeile *einfügen"* ruft das Programm die Form-Routine Insert_New_Row auf. Dabei wird zuerst geprüft, ob das Eingabefeld wirklich besetzt ist? Weiterhin wird geprüft, ob die eingegebene Produkthierarchie-Nummer in der Tabelle T179 vorhanden ist. Anschließend wird in der Ergebnisliste gesucht, ob der eingetragene Wert schon vorhanden ist. Falls nein, wird die eingegebene Produkthierarchie-Nummer in der Ergebnisliste eingefügt. Der Wert der Produkthierarchie-Nummer muss in der Ergebnisliste einmalig sein, da er jeden Datensatz eindeutig kennzeichnet. Über dem Insert-Befehl wird eine Zeile in der Ergebnisliste eingefügt. Hierfür wurde im Programm vor Aufruf des Insert-Befehls eine Struktur wa_data, die Daten der Datentabelle ZSGF_DATA einliest, angelegt. Danach wurde in dieser Struktur die einzufügende Zeile gestellt. Nachdem die Zeile eingefügt wurde, wurde mit dem Update-Befehl die Tabelle ZSGF_DATA aktualisiert. Wurde die Zeile erfolgreich in der Ergebnisliste eingefügt, wird den Benutzer über ein Pop-Up-Fenster informiert und beim Betätigen des Feldes "OK" verlässt der Benutzer der Ansicht Zeile_bearbeiten und landet er auf dem Einstiegsbild des Cockpits.

Das Unterstehende Ablaufdiagramm fasst die Form-Routine Insert_New_Row zusammen:

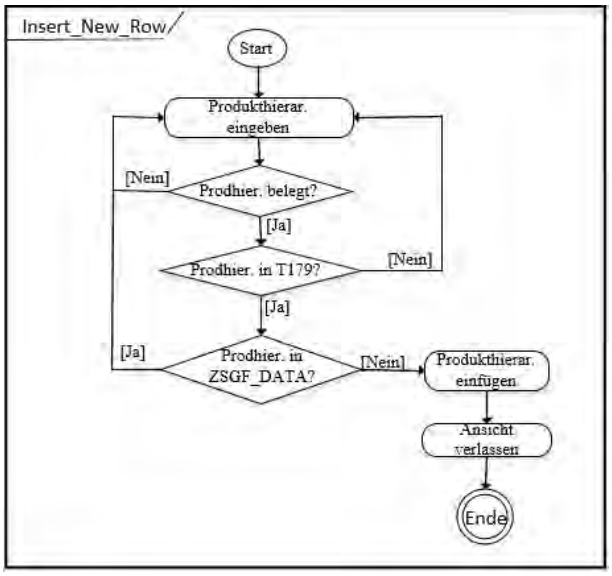


Abbildung 11: Ablaufsdiagramm Insert_new_Row.

Datenfeld löschen

Das Ereignis des Löschens eines Feldes wird durch Eingabe der zu löschenden Spalte und einen Klick auf die Schaltfläche „Spalte löschen“ im Bildschirmbild des Dynpro_9002 ausgelöst. Der dieser Schaltfläche zugeordnete Funktionscode „DELETE“ sorgt für den Aufruf der Form-Routine Delete_Column, welche wiederum eine Funktion „POPUP_TO_CONFIRM“ aufruft, um ein versehentliches Löschen einer Spalte oder eines Feldes zu verhindern. Dabei wird der Benutzer aufgefordert, seine gewünschte Aktion zu bestätigen. Drückt der Benutzer auf den *Yes*-Button, wird der Delete-Befehl ausgeführt. Hierbei wurde die zu löschende Spaltenangabe über eine Where-Klausel spezifiziert. Anschließend wird der Benutzer benachrichtigt über den Verlauf seiner Aktion.

```
IF LV_ANSWER_2 EQ '1'.   "If deletion is permitted
  "Delete data records from database table
  DELETE FROM ZSGF_DATA
  WHERE
   VW     = ZSGF_DATA-VW AND
   MATKZ  = ZSGF_DATA-MATKZ AND
   VKORG  = ZSGF_DATA-VKORG AND
   AUGRU  = ZSGF_DATA-AUGRU AND
   SGF_MERKMAL = ZSGF_DATA-SGF_MERKMAL.

  CLEAR: WA_DATA, ZSGF_DATA.
  MESSAGE 'Column has been removed' TYPE 'I'.
  LEAVE TO SCREEN 100.
ELSEIF   "If deletion is not permitted
  LV_ANSWER_2 EQ '2'.
  MESSAGE 'Column has not been removed' TYPE 'I'.
  LEAVE TO SCREEN 100.
ENDIF.
```

Abbildung 12: Datenfeld aus der Ergebnisliste löschen

Dynpro_9002 verlassen

Bei Betätigung der Cancel, Exit und Back Buttons auf der Symbolleiste wird das Dynpro_9002 verlassen und dem Anwender wird über die Anweisung – leave to screen 0 – in dem Object-Navigator Transaktion se80 – weitergeleitet.

Das Einfügen von Datenfeld und das Löschen von Datensatz in der Ergebnisliste wurden analog realisisert. Lediglich die Eingabefelder zum Einlesen aus der Datentabelle ZSGF_DATA unterscheiden sich in ihrer Anzahl.

Test-und Integrationsphase

Um die Qualität und Funktionalität des Programms zu gewährleisten, wurden sowohl Tests vom Entwickler, als auch von dem Fachbereich durchgeführt. Es wurden sowohl die funktionalen, als auch die nicht-funktionalen Anforderungen überprüft.

Überprüfen der nicht-funktionalen Anforderungen

Als erstes wurde getestet, ob das SGF-Cockpit sich erweitern bzw. modifizieren lässt. Durch die dynamische Programmierung sind Änderungen durch den Fachbereich einfach und schnell realisierbar.
Dieser dynamische Aspekt des Programms fördert auch die Wartbarkeit und die Performance. Als zweiten wurde die Applikation auf ihre Zuverlässigkeit geprüft. Das Cockpit reagiert zuverlässig auf alle möglichen Benutzereingaben Zum Schluss wurde die Anwendung auf ihre Benutzerfreundlichkeit getestet. Die Oberfläche unterstützt den Benutzer bei seinen Aufgaben. Sie ist einfach komfortabel bedienbar, mit Pop-up-Fenster ausgerichtet und mit wenigen Klicks lässt sich durch die Oberfläche navigieren. Dies wirkt sich positiv auf die Zufriedenheit der Benutzer im Umgang mit dem Cockpit aus.

Überprüfen der funktionalen Anforderungen

Die SGF wurde in Form eines ALV-Grids dargestellt. Die stehen SGF in übersichtlicher Form in das neue Cockpit zur Verfügung.
Des Weiteren sollte das Programm dem Anwender die Möglichkeit geben, die in das Cockpit bestehende SGF abfragen oder filtern zu können. Diese Anforderung wurde erfüllt, in dem Eingabefeldeder in das Einstiegsbild des SGF-Cockpits zu Verfügung gestellt wurden. Diese Eingabefelder dienen zu Selektionsoptionen oder Suchkriterien. Es Wurde auch überprüft, ob bei nicht vorhandenen Selektionsoptionen eine leere Ergebnisliste ausgegeben wurde.
Eine weitere Anforderung an das Programm war die Änderungen von Ergebnislisteinhalte und die Speicherung diese Änderungen zu ermöglichen. Um dies gerecht zu werden, wurde eine Dropdown-Listbox erstellt. Mithilfe dieser Listbox ist der Anwender in der Lage, einen Wert für strategisches Geschäftsfeld-Merkmal aus der erstellten Listbox auszuwählen. Nach Auswahl des gewünschten Wertes muss diese aber durch die Betätigung der Enter-Tatse bestätigt werden. Somit merkt das Programm diese Auswahl vor, um der geänderte Inhalt nach Betätigung des Speicherbuttons bzw. das Diskette-

Symbol in der Symbolleiste in das Cockpit zu speichern.
Bei den anderen Steuerungsfunktionen, wie das Einfügen und Löschen von Datensätzen bzw. Datenfeldern, wurden überprüft ob diese ordnungsgemäß funktioniert.
Zum abschließenden Test des neuen Berichtswesen-Tools wurde diese den Fachbereich zum Abnahme_Test zur Verfügung gestellt. Dabei soll der Fachbereich selbst alle Funktionen ausführen und auf Fehlermeldungen achten. Falls Fehler auftreten, sollen diese schnellstmöglich dem Entwickler des Cockpits mitgeteilt werden.
Fehler sind bei Testen nicht aufgefallen. Somit sind alle Steuerungsfunktionen geprüft und das Cockpit zur Pflege der strategischen Geschäftsfelder kann nun in das Produktivsystem transportiert und wird über eine angelegte Dialogtransaktion für alle involvierten Fachbereich zur Verfügung gestellt.

Fazit

Das Ziel war die Implementierung eines Cockpits für die Visualisierung und Verarbeitung der Strategischen Geschäftsfelder der Maschinenfabrik Reinhausen GmbH.
Das hier in dieser Arbeit implementiertes Cockpit befindet sich auf dem Stand der Anforderungen des Fachbereiches und wurde eingesetzt. Somit konnte den Nutzungsumfang betreffenden Vorgaben erfüllt werden. Das Cockpit bietet eine benutzerfreundliche und übersichtliche Ergebnisliste der strategischen Geschäftsfelder, mit der der Fachbereich alle relevanten Informationen jederzeit im Überblick hat, um wichtige Entscheidungen zu treffen, Umsätze auszuwerten, Kosten und Deckungsbeiträge zu ermitteln. Durch das erstellte Customizing-Cockpit werden der Zugang, die Handhabung und die Verarbeitung der SGF-Merkmale vereinfacht. Da die Ergebnisliste dynamisch programmiert wurde, kann sie nun ohne den Eingriff des IT–Bereiches durch den Fachbereich direkt transformiert beziehungsweise verarbeitet werden. Dadurch werden die Änderungen schneller, auf einem einzigen System – SAP R/3– erfasst und die IT wird parallel entlastet. Das erspart Zeit und Ressourcen gegenüber der fehleranfälligen manuellen Verarbeitung der strategischen Geschäftsfelder. Außerdem wurde durch ausgeführte Tests eine höhere Datenqualität gewährleistet.
Für die heutige Nutzung ist der Funktionsumfang des SGF-Cockpits ausreichend. Dennoch ist das Programmieren einer Benutzeroberfläche oder eine Schnittelle niemals komplett fertig. Es bestehen zukünftig immer Möglichkeiten das Programmieren dieser Schnittelle noch zu verfeinern. Die Möglichkeiten sind dabei vom Ideenreichtum des Programmierers abhängig.

Literatur

[1] Peyman Azhari, Nilufar Faraby, Alexander Rossmann, Bernhard Steimel, Kai S. Wichmann. Digital Transformation Report 2014. Köln: neuland GmbH & Co. KG (2014) (siehe S. 3-16).

[2] Maschinenfabrik Reinhausen GmbH: Über MR. Zugriff am 12.06.2017. http://www.reinhausen.com/de/desktopdefault.aspx/tabid-1449/1774_read-4521/

[3] Walsh Gianfranco; Deseniss Alexander; Kilian Thomas: Marketing. Eine Einführung auf der Grundlage von Case Studies. 2. Auflage. Berlin: Springer Gabler 2013 (siehe S. 126)

[4] BC405 ABAP-Reports programmieren SAP NetWeaver. SAP AG 2006/Q2. (Seite 2-13)

[5] Köble, Josef . Entwicklung barrierefreier Software mit SAP NetWeaver®. 1. Auflage. Bonn: Galileo Press (SAP PRESS) 2007 (siehe S. 234)

[6] BC410 Entwicklung dynprobasierter Benutzerdialoge. *SAP AG 2006.* (Seite 10-13)

[7] Rainer Riekert. ABAP – Programmierung, Fortgeschrittene Programmiertechniken für ABAP. Addison-Wesley Verlag, München, 2001 (siehe S. 132)

[8] Keller, Horst / Jacobitz Joachim / Hochlehnert, Bernhard (Hg.): ABAP Objects Referenz.1. Auflage, Bonn: Galileo Press 2002.

[9] Horst Keller und Sascha Krüger. ABAPObjects: Einführung in die SAP-Programmierung. 2. Auflage. Bonn: Galileo Press 2005.

Ein paralleler Algorithmus für API Mining von C# Code

Robert Horkovics-Kovats (M. Sc.)
Capgemini SE
Bahnhofstraße 30, 90402 Nürnberg
robert.horkovics-kovats@capgemini.com

Dr. Eldar Sultanow
Capgemini SE
Bahnhofstraße 30, 90402 Nürnberg
eldar.sultanow@capgemini.com

Prof. Dr.-Ing. Frank Herrmann
OTH Regensburg
Innovationszentrum für Produktionslogistik
und Fabrikplanung (IPF)
Galgenbergstraße 32, 93053 Regensburg
frank.herrmann@oth-regensburg.de

Zusammenfassung

Die Konformitätsanalyse ist eine Technik der statischen Code-Analyse (SCA) zur Software-Qualitätssicherung. Ihr Kernproblem ist, dass Werkzeuge nicht aus bereits eingetretenen Fehlern automatisiert dazulernen. Zur Lösung wurde in dieser Arbeit das maschinelle Lernen (ML) evaluiert, indem ein wissenschaftlich fundierter und praktisch erprobter Ansatz zur unüberwachten Lerntechnik angewandt und das Ergebnis analysiert wurde. Es wurde festgestellt, dass zur Anwendung auf verschiedene Programmiersprachen nur ein sprachspezifisches API Mining-Tool notwendig ist. Ein derartiges Tool durchsucht in parallelisierter Form Codezeilen und normalisiert sie für maschinelle Lernprozesse. Dieses System wurde für die Programmiersprache C# implementiert, da viele Industrieprojekte in dieser Sprache entwickelt werden. Zur funktionalen Validierung wurde in einer Fallstudie gezeigt, dass Regeln mit einem positiven Effekt auf Software-Qualität gelernt wurden. Konkret wurde der Wartungsaufwand eines Code-Smells in einem Beispielprojekt durch das Auslagern einer gelernten Assoziation in eine gemeinsame Methode um den Faktor 30 reduziert. Die Laufzeit des Algorithmus wurde empirisch in acht open-source Repositorys evaluiert. Durch Parallelisierung kann eine durchschnittliche Laufzeitverbesserung von 45,16% erwartet werden. Allerdings wurden bei der Anwendung auch Grenzen deutlich: Viele Assoziationen sind nutzlos, die Regelbewertung ist von einem subjektiven Faktor abhängig und die Wirtschaftlichkeit des Tools ist deshalb nicht transparent. Dennoch belegt diese Arbeit, dass ein ML-basiertes SCA-Tool als ergänzende Qualitätssicherungsmaßnahme im Software-Engineering möglich ist.

I. Einleitung

Eine Methode zur Qualitätssicherung im Software-Engineering ist die statische Code-Analyse (SCA). Dabei wird der Code durch Sichtung der Quelltexte untersucht, ohne diese in ein ausführbares Programm zu übersetzen. Eine konkrete Technik ist die Konformitätsanalyse, wo der Code auf vordefinierte Syntax-, Grammatik- und Semantikregeln auf Konformität geprüft wird. Ihr Kernproblem ist, dass sie nicht aus bereits eingetretenen Fehlern automatisiert dazulernt: Fehler müssen erst im Produktivbetrieb eintreten, bevor daraus manuell neue Regeln identifiziert werden können [Sult18]. Zur Lösung wurde in dieser Arbeit das maschinelle Lernen (ML) evaluiert. Ein wesentlicher Bestandteil besteht darin, die Codebasis in eine für ML-Prozesse verwertbare Datengrundlage zu normalisieren.

Aufbau des Manuskripts

Der Artikel ist wie folgt strukturiert: Zuerst wird das Problem anhand der Situation in einem Unternehmen präzisiert. Daraufhin werden die relevanten theoretischen Grundlagen zur Assoziationsanalyse dargelegt. Damit wird ein wissenschaftlich fundierter und praktisch erprobter ML-basierter SCA-Ansatz angewandt und das Ergebnis analysiert. Das Konzept eines parallelisierten API Miners für die Programmiersprache C# wird vorgestellt. Seine Funktionalität wird in einer Fallstudie validiert und die Laufzeit unter Parallelisierung empirisch untersucht. Die Ergebnisse werden als Nächstes kritisch hinterfragt und diskutiert. Abschließend wird ein Fazit mit Ausblick gezogen, dass ein ML-basiertes SCA-Tool als ergänzende Qualitätssicherungsmaßnahme im Software-Engineering möglich ist.

II. Problemstellung, Ziel und Lösungsansatz

Ein globaler Technologiekonzern, der vor allem auf die Elektrifizierung, Automatisierung und Digitalisierung insbesondere in den Bereichen Stromerzeugung und -übertragung, Medizintechnik, Infrastruktur und industrielle Anwendungen ausgerichtet ist, verfügt über viele in der Programmiersprache C# implementierte Projekte. In den Projekten werden klassische SCA-Werkzeuge wie SonarQube zur Software-Qualitätssicherung eingesetzt. Dabei definiert ein Architekt oder Software-Engineer Prüfregeln vorab, das Werkzeug validiert gegen diese Regeln und die Entwickler beheben, falls vorhanden, Regelverstöße. Treten neue Fehler im Produktivbetrieb auf, ergänzt der Architekt/Software-Engineer neue Regeln, um diese Fehler in Zukunft verhindern zu können. Die eingesetzte Methode weist die zuvor beschriebene Schwäche der Konformitätsanalyse auf: Neue Regeln sind erst a-posteriori, also nach dem mit Kosten verbundenem Eintritt des dazugehörigen Fehlers, bekannt. Die genannte Frage wurde auf diese Projekte angewandt. Sie lässt sich wie folgt präzisieren: Wie lässt sich mittels API-Mining C# Code so normalisieren, dass es für ML-basierte SCA verwertbar ist?

Das Ziel dieser Arbeit ist es, ein API-Mining-Prototyp zu entwickeln, das in parallelisierter Form C# Codezeilen durchsuchen und für maschinelle Lernprozesse aufbereiten (normalisieren) kann. Die normalisierten Daten enthalten Transaktionen mit Items, die eine Regelidentifikation anhand der unüberwachten Lerntechnik *Assoziationsanalyse* ermöglichen. Das Prinzip ist wissenschaftlich fundiert und praktisch erprobt [Sult18] und soll auf den Daten anwendbar sein, die aus der, mit der vorliegenden Arbeit angestrebten, Normalisierung resultieren.

III. Unüberwachte Lerntechnik und strukturierte Datenrepräsentation

Die Assoziationsanalyse (engl. Assocation Rule Mining) ist eine unüberwachte Lerntechnik des ML und wurde zuerst 1993 von Agrawal et al. als Warenkorbanalyse vorgestellt [Agra93]. Unüberwachte Ansätze zeichnen sich dadurch aus, dass sie auf einer Datenmenge ohne Labels operieren. Es wird in protokollierten Verkaufsdaten analysiert, welches Produkt häufig mit einem Anderen gekauft wird. Es werden Assoziationsregeln mit der Aussage „Wer Produkt A kauft, kauft häufig auch Produkt B“ gesucht.

Das Prinzip wird an folgendem Beispiel erklärt:

Tabelle 1: Warenkorbanalyse Beispiel

Transaktion	Smartphone	PC	PC-Bildschirm
1	x	x	x
2	x	x	x
3		x	x
4		x	

Hier sind vier Einkäufe protokolliert. Sie werden als Transaktionen bezeichnet. Jede Transaktion T_i beinhaltet ihre eingekauften Produkte, sogenannte Items. Die Transaktionsdatenbank zeigt, dass gemeinsam mit einem PC in 75% auch ein PC-Bildschirm gekauft wurde. Dieser Zusammenhang kann als Assoziationsregel R formuliert werden. Wird Item(-menge) A gekauft, besteht eine Wahrscheinlichkeit von x%, dass auch Item(-menge) B

gekauft wird. Diese Regel trifft auf y% aller Transaktionen zu. Formal notiert:
Sei I die Menge aller Items i_1, i_2, ..., i_n. Es gilt:

$$R : A \rightarrow B \ \ [support : y\%] \ \ [confidence : x\%],$$
$$\text{wobei } \ A \subset I, B \subset I \ \text{ und } \ A \cap B = \varnothing$$

WEKA ist ein open-source Data-Mining-Tool, das an der University of Wakato in Neuseeland entwickelt wird. Der Name steht für Wakato Environment for Knowledge Analysis. Es bietet erprobte und getestete ML-Software, auf die über eine grafische Oberfläche, über Kommandozeilenapplikationen oder über eine Java API zugegriffen werden kann. Es wird zum Lehren, zur Forschung und in industriellen Anwendungen eingesetzt und enthält eingebaute Tools für Standardproblemstellungen aus dem maschinellen Lernen [Weka20].

WEKA erwartet Input-Dateien in einem standardisierten Format, dem sogenannten Attribute-Relation File Format (ARFF). Es beschreibt eine Liste von Instanzen, die eine Menge von Attributen teilen. Die Struktur besteht aus den Sektionen Kopf (Header) und Daten (Data). Im Kopf werden Attribute mit Datentypen deklariert. Im darunter folgenden Körper werden Dateninstanzen aufgelistet. Das Format ermöglicht die strukturierte Repräsentation von zuvor unstrukturierten oder semistrukturierten Daten, wodurch die Anwendung von ML-Algorithmen vereinfacht wird.

```
1  @RELATION weather
2
3  @ATTRIBUTE temperature NUMERIC
4  @ATTRIBUTE humidity {high, normal, low}
5  @ATTRIBUTE note STRING
6  @ATTRIBUTE timestamp DATE "yyyy/MM/dd"
7
8  @DATA
9  25.3,high,'slightly cloudy',"2019/05/22"
10 31.1,normal,'cloudless and hot',
       "2019/06/26"
```

Listing 1: ARFF Beispiel

IV. Analyse eines ML-basierten SCA-Ansatzes

Die Assoziationsanalyse untersucht Items innerhalb eines Gruppierungskriteriums. Dieses Prinzip muss zur statischen Code-Analyse auf den Quelltext übertragen werden. Um Assoziationen über Schnittstellen zur Programmierung von Anwendungen (APIs) zu lernen, schlagen Sultanow et al. in [Sult18] vor, Methodeninvokationen (Calls) als Items zu betrachten. Es werden Beziehungen zwischen Methoden gesucht, die mit einer bestimmten Wahrscheinlichkeit zusammen aufgerufen werden. Als Gruppierungskriterium wird die dazugehörige aufrufende Methodendeklaration (Caller) festgelegt. Der Zusammenhang wird im folgenden Codebeispiel verdeutlicht:

```
1  int caller()           // Transaction
2  {
3    call1();             // Item 1
4    int x = call2();     // Item 2
5    x += 1;
6    call3(5);            // Item 3
7    return call4() + x; // Item 4
8  }
```

Listing 2: Methodendeklaration mit Methodeninvokationen

Assoziationsregeln werden aus einer verifizierten Codebasis gelernt. Neuer Code wird daraufhin gegen diese Regeln auf Verletzungen geprüft. Eine Transaktion erfüllt eine Regel $R : A \rightarrow B$, wenn sie die disjunkten Itemmengen A und B vollständig enthält. Fehlt auch nur ein Item aus diesen Mengen, liegt eine Regelverletzung vor. Im Kontext der Code-Analyse muss also eine erwartete Invokation in einer gelernten Regel fehlen, damit im untersuchten Code eine Verletzung gemeldet wird.

Weiterhin stellen Sultanow et. al in [Sult18] eine theoretische Ablaufsequenz auf, wie das vorgestellte Konzept in eine Continuous Integration Pipeline integriert werden kann (s. Abbildung 1). Darin arbeiten drei Instanzen zusammen. Erst holt ein *API Miner* den Quellcode aus einem Versionskontrollsystem wie Git und extrahiert die Inputdaten in ein Dateiformat, das für den ML-Vorgang verwertbar

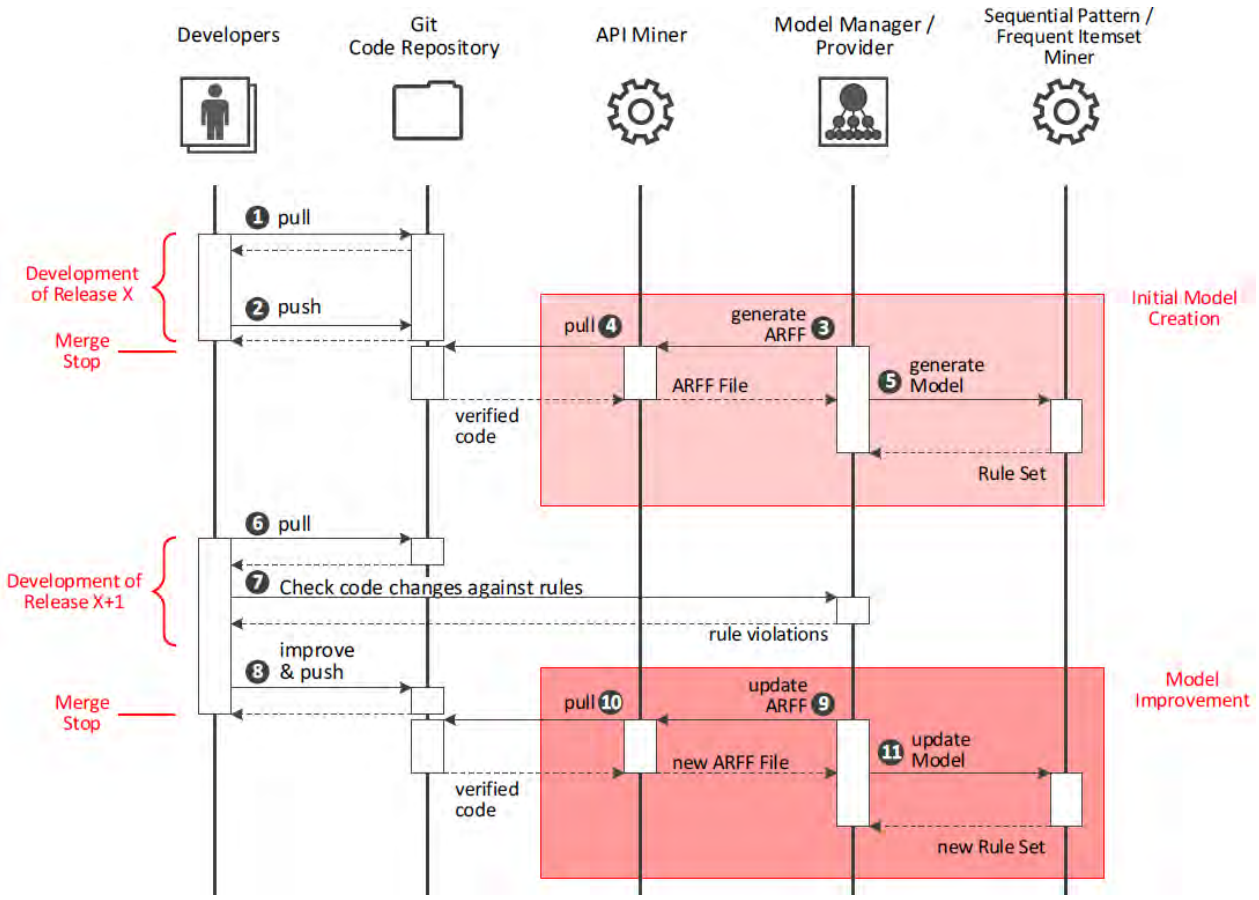


Abbildung 1: Theoretische Ablaufsequenz des ML-basierten SCA-Prototyps [Sult18]

ist. Ein *Model Manager* verwaltet das ML-Modell, indem der Manager es durch Aufruf der Algorithmen trainiert und Regeln auf Verletzungen prüft. Der *Pattern Miner* implementiert die Algorithmen zur Assoziationsanalyse. Da sowohl zweite als auch dritte Instanz mit Daten arbeiten, die unabhängig von der zugrundeliegenden Programmiersprache sind, ist nur der API Miner sprachspezifisch. Damit also eine statische Code-Analyse nach diesem Prinzip für C# möglich wird, muss nur ein API Miner entworfen werden, der C# Code extrahiert und in ein standardisiertes Datenformat transformiert.

V. Ein paralleler Algorithmus für API Mining von C# Code

Der C# API Miner wird als Konsolenapplikation entwickelt. Es verarbeitet Userinput und transformiert die geforderten Daten einer übergebenen API in eine ARFF-Datei.

Algorithmus 1: API Mining in C#

Eingabe:	C# API in Form einer Projektmappe (Solution), eines Projekts (Project), eines Unterverzeichnisses oder einer konkreten Quelldatei
Anweisungen:	Eine API wird zur weiteren Verarbeitung geladen. Es wird über ihre Elemente iteriert und für jedes Dokument der Syntaxbaum analysiert.

Aus diesem Objektmodell werden alle Methodendeklarationen des Dokuments gesammelt. Anschließend werden die Invokationen pro Deklaration gesucht und ihre Namen vollqualifiziert aufgelöst. Fehlerhaft aufgelöste Einträge werden mit „UNRESOLVED“ gekennzeichnet.

Ausgabe: ARFF-Datei mit extrahierten API-Daten

Zur Implementierung müssen Projektmappen mit ihren Abhängigkeiten kompiliert, Quelltexte in abstrakte Syntaxbäume geparst und Methodennamen zur eindeutigen Identifikation vollqualifiziert aufgelöst werden. Derartige Syntax- und Semantikanalysen sind in C# mithilfe des .NET Compiler Platform-SDK (Roslyn APIs) möglich [Wagn20]. Die extrahierten Daten müssen die aufgelösten Namen von Methodendeklarationen mit dazugehörigen Methodeninvokationen enthalten. Die Struktur der ARFF-Datei enthält somit zwei Spalten, die diese Caller und Calls als Zeichenketten halten. ARFF unterstützt bis dato keine Datenkollektionen [Univ08], weshalb die Auflistung der Invokationen als Text angegeben wird. Als Trennzeichen wird das Leerzeichen eingesetzt:

```
1 @RELATION myProject
2
3 @ATTRIBUTE caller STRING
4 @ATTRIBUTE calls STRING
5
6 @DATA
7 'dummy.MyClass.caller1()','dummy.MyClass.x
      dummy.MyClass.y'
8 'dummy.MyClass.caller2()','System.out.
      println'
```

Listing 3: ARFF-Inhalt bei ML-basierter SCA

Da die Genauigkeit bzw. Aussagekraft von ML-Algorithmen höher ist, wenn sie auf hohen Datenmengen angewandt wird [Hurw18, S. 8], wird die Parallelisierung des Algorithmus gefordert. Dabei nutzt das Programm die Multikernarchitektur von modernen Rechnern aus, um Laufzeitverbesserungen bei zunehmender Input-Größe zu erzielen. Im Algorithmus liegen dafür drei verschachtelte Iterationsstufen vor:

- Stufe 1: Iteration über Projekte einer Projektmappe
- Stufe 2: Iteration über Dokumente eines Projekts
- Stufe 3: Iteration über Methoden eines Dokuments

Jede dieser Stufen kann von Parallelisierung profitieren. Jedoch ist unklar, welche Ebenen zur höchsten Beschleunigung führen [Toub10, S. 28–29]. Deshalb kann die Parallelisierung über optionale Input-Parameter gesteuert werden, wodurch die Konstellation mit der höchsten Laufzeitverbesserung abhängig von der Eingabe erwählt werden kann. Um den generellen Effekt der Parallelisierung zu schätzen und Konfigurationsempfehlungen vorzuschlagen, wurde die Performance des API Miners empirisch untersucht.

VI. Funktionale Validierung

Zur funktionalen Validierung wird der C# API Miner dem ML-Prozess von Sultanow et al. aus [Sult18] (s. Abbildung 1) unterzogen. Der API Miner gilt als validiert, wenn damit mindestens eine Assoziationsregel in C# gelernt werden kann, die einen nachweisbar positiven Effekt auf die Software-Qualität der untersuchten Projektmappe hat. Um den abstrakten Begriff *Software-Qualität* zu präzisieren, wird mithilfe der Qualitätseigenschaften des standardisierten ISO-25010-Modells der Nutzen einer Regel begründet. Die Qualitätsmerkmale umfassen funktionelle Eignung, Performance Effizienz, Kompatibilität, Benutzbarkeit, Zuverlässigkeit, Sicherheit, Wartbarkeit und Portabilität [ISO20].

Als Fallbeispiel wird das open-source Repository mit der ID 4 (AI-Programmer) betrachtet.

Tabelle 2: Regelbefund aus dem Repository 4 mit positivem Effekt auf Software-Qualität

Repo	Rule	Supp.	Conf.
4	Math.Abs → Fitness-Base.IsFitnessAchieved	31	100%

Der Regelbefund weist einen absoluten Support-

Wert von 31 und eine Konfidenz von 100% auf. Das heißt, dass diese Invokationsreihe insgesamt 31-mal im Code auftritt und Antezedenz- und Konsequenz-Invokation stets gemeinsam als Paar vorkommen. Die beiden Methoden können deshalb in eine gemeinsame Methode ausgelagert werden, da sie einer logischen Einheit entsprechen. Dadurch konnten Wartungsaufwände an dieser Codestelle um den Faktor 30 reduziert werden.

Das kurze Beispiel beweist die funktionale Eignung des C# API Miners: Relevante API-Daten des Repositorys *AI-Programmer* wurden extrahiert, sodass eine Assoziationsanalyse darauf ausgeführt werden konnte. Dadurch wurden Assoziationsregeln gelernt, die der statischen Code-Analyse zur Qualitätssicherung dienen.

VII. Empirische Untersuchung der Performance

In der empirischen Laufzeituntersuchung wurde der API Miner neben sieben weiteren open-source Repositorys betrachtet (s. Tabelle 3). Insgesamt umfasst die Analyse knapp 400.000 Zeilen Code. Der Algorithmus wurde in jeder Konfiguration zehnmal auf einem sonst stillen System ohne Hintergrundprozesse ausgeführt, um mittlere Laufzeitwerte mit Standardabweichungen zu erhalten.

Tabelle 3: Acht untersuchte Repositorys mit Lines of Code

ID	Repo	LoC
1	Unitiy Reference	340733
2	ScriptCS	18597
3	Algorithms-in-CSharp	17057
4	AI-Programmer	8124
5	Easy-http	4172
6	APIMiner	4007
7	Domain-Driven-Design-Example	3353
8	Design-Patterns	3045

Es wurden fünf Code-Metriken erfasst, um zu bewerten, wie repräsentativ die Repositorys hinsichtlich ihrer Merkmalsausprägungen sind:

Tabelle 4: Erhobene Code-Metriken von acht C#-Projektmappen

Repo	LoC	P	D	M	M/D
1	340733	4	2870	37085	12,92
2	18597	10	274	1155	4,22
3	17057	5	125	290	2,32
4	8124	10	76	170	2,24
5	4172	2	71	278	3,92
6	4007	4	33	186	5,64
7	3353	3	124	189	1,52
8	3045	17	177	269	1,52

Legende:
LoC Lines of Code
P Anzahl Projekte
D Anzahl C#-Dokumente
M Anzahl Methoden
M/D Anzahl Methoden pro Dokument

Aus den gemessenen Daten kommt hervor, dass eine Kategorisierung nach Anzahl von Lines-Of-Code-Ziffern heuristisch wertvolle Vergleichswerte zur Gesamtlaufzeit bietet. So dauert der Algorithmus beispielsweise für Projektmappen mit vierstelligen Codezeilenzahlen ca. 5 bis 6 Sekunden:

Tabelle 5: Sequenzielle Laufzeit des C# API Miners in Sekunden

Repo	Load	Mine	Write	Total
1	02,932	20,080	07,790	30,908
2	04,037	04,392	00,270	08,790
3	03,112	03,830	00,077	07,114
4	02,860	02,821	00,047	05,824
5	02,605	02,801	00,080	05,577
6	03,168	03,095	00,060	06,415
7	02,648	02,702	00,057	05,502
8	03,603	02,456	00,058	06,211

Das Öffnen bzw. Laden von Projektmappen dauert in der Stichprobe zwischen 2 und 4 Sekunden. Diese Werte resultieren je nach Projektgröße in stark unterschiedlichen Anteilen bzgl. der Gesamtlaufzeit. 2 bis 4 Sekunden entsprechen bei kleinen Projektmappen ca. 50% und bei der größten untersuchten Projektmappe knapp 10%. Demnach ist für alle Re-

positorys das Öffnen der Projektmappe und damit der Einsatz von Roslyn eine laufzeitintensive Aktion.

Die Laufzeitdaten der tiefsten Iterationsstufe zeigen für sieben von acht Repositorys eine Laufzeitverbesserung (s. Spalte M in Tabelle 6). Es wird in der tiefsten Stufe somit genug Arbeit parallelisiert, sodass sich die Kosten dafür amortisieren. Es wird kein Antipattern begangen. Die Ausnahme dieser Beobachtung ist die kleinste Projektmappe. Der Anwendungsfall bestätigt, dass die alleinige Parallelisierung dieser Stufe nicht genügt, um für beliebige Projekte mit unterschiedlichsten Merkmalen Laufzeitverbesserungen zu erzielen. Parallelisierungsoptionen für die höheren Iterationsstufen sind dafür notwendig.

Tabelle 6 zeigt die prozentuale Laufzeitverbesserung des Code-Abschnitts *Mine* in den acht Repositorys. Da die Parallelisierung über drei optionale boolesche Parameter gesteuert wird, liegen $2^3 - 1 = 7$ Optionen vor. Sie werden mit ihren Initialen beschriftet. So steht M für die Parallelisierung der Methoden, D für Dokumente, P für Projekte und beispielsweise der Ausdruck D+P für die Kombination dieser Parameter.

Grün-markierte Zellen zeigen die Konstellation mit der höchsten Verbesserung pro Zeile. Die Konstellation mit den meisten grünen Einträgen ist D+P. In fünf von acht Repositorys wurden hier die höchsten Verbesserungen erzielt. Weiterhin beinhalten die effizientesten Konstellationen stets die Parallelisierung von Projekten (P). Die Parallelisierung auf der höchsten hierarchischen Programmstufe bietet somit den größten Nutzen. Durchschnittlich kann für die schnellste Konstellation eine Verbesserung von 45,16% erwartet werden. Die rot-markierten Zellen zeigen Fälle mit Laufzeitverlangsamungen. Außer dem Wert -1,33% vom Repository 8(M), finden sich diese Negativfälle erneut bei der D+P-Konstellation. Werden diese Repositorys mit ihren erhobenen Code-Metriken (s. Tabelle 4) gegenübergestellt, fällt ein Zusammenhang zwischen hoher Projektanzahl (>10) und Laufzeitverschlechterung auf. Daraus kann abgeleitet werden, dass für Projektmappen mit vielen Projekten ein geringerer Parallelisierungsgrad empfohlen wird. In diesen Fällen bietet die alleinige Parallelisierung der Projektverarbeitung (P) den größten Nutzen.

VIII. Limitationen der vorgestellten Lösung

Zwar wurde anhand des Fallbeispiels Nutzpotenzial im ML-basierten Ansatz demonstriert, muss dennoch die Sinnhaftigkeit des Werkzeugs kritisch hinterfragt werden. In vier der acht untersuchten Repositorys wurden bis zu 99,35% der Regeln aus Testcode gelernt. Dieses dominante Vorkommen lässt sich über die Art, wie Software-Tests geschrieben werden, begründen: Bei Komponententests o.ä. ist es üblich, mithilfe von Invokationen Daten vorzubereiten, Platzhalterobjekte (Mocks) zu erstellen und Behauptungen (Asserts) zu prüfen. Das führt zu vielen Test-Invokationen, die gemeinsam mit dem eigentlichen Geschäftscode aufgerufen werden. In Folge wurden deshalb überwiegend Assoziationen

Tabelle 6: Prozentuale Laufzeitverbesserung der Parameterkonstellationen

Repo	M	D	P	M+D	M+P	D+P	M+D+P
1	26,58%	39,33%	27,28%	38,55%	41,90%	49,38%	46,82%
2	8,75%	30,42%	45,28%	32,27%	46,11%	21,27%	15,30%
3	22,20%	33,92%	45,87%	29,71%	47,33%	48,66%	47,97%
4	16,50%	5,09%	46,93%	22,62%	44,98%	-0,57%	-20,27%
5	11,61%	37,68%	32,95%	36,14%	37,70%	46,92%	41,99%
6	12,90%	33,14%	39,32%	31,17%	43,11%	45,39%	39,50%
7	1,23%	23,26%	35,25%	22,45%	37,71%	37,94%	33,92%
8	-1,33%	23,04%	39,95%	19,67%	37,74%	-32,62%	-23,01%

aus diesem Segment gelernt. Diese Regeln können jedoch nicht als sinnvoll eingestuft werden: Eine Regelverletzung im Geschäftscode, die ein fehlendes *Assert.Equals* meldet, ist stets falsch-negativ, da diese Invokation nur in Testklassen relevant ist. Derartige Regeln werden als Noise (Lärm) bezeichnet, die die Suche nach nützlichen Regeln erschweren. Darüber hinaus erweist sich die Bewertung von gelernten Assoziationsregeln als schwierig, da Interessantheit subjektiv ist. Beispielsweise liefert eine Regel $R : Console.WriteLine() \rightarrow string.Format()$, die einen Zusammenhang zwischen Konsolenausgabe und Textformatierung ausdrückt, kaum fachlichen Mehrwert, auch wenn ihr absolutes Vorkommen hoch ist (Support) und die Regel oft Anwendung findet (Confidence). Es gibt keinen nachgewiesenen Zusammenhang zwischen objektiver und subjektiver Interessantheit [Geng06], weshalb eine Regel trotz hohem Support-, Confidence- oder Korrelationswert (Lift) dennoch nutzlos sein kann. Regeln müssen also aufwändig, beispielsweise mithilfe von Black- und Whitelists, verwaltet werden. Es ist unklar, ob der nachgewiesene Nutzen diesem Aufwand überwiegt. Die Wirtschaftlichkeit des vorgestellten Ansatzes ist nicht transparent und muss zur Analyse der Marktreife in einer weiterführenden Forschung untersucht werden.

IX. Fazit und Ausblick

Die statische Code-Analyse ist in ihrer Natur sprachspezifisch. Allerdings verarbeiten die eingesetzten Algorithmen zur Assoziationsanalyse Daten in einem standardisierten Datenformat. Somit muss je Sprache nur ein sprachspezifischer API Miner entwickelt werden, um den Ansatz für die korrespondierende Sprache zu ermöglichen.

Durch Analyse eines erprobten ML-basierten SCA-Ansatzes wurden die Anforderungen an ein API Mining Tool verstanden, das Regeln lernt, die der statischen Code-Analyse in C# dienen. Es normalisiert Methodendeklarationen zu Transaktionen und dazugehörige Methodeninvokationen zu Items und setzt das Datenformat ARFF zur Abbildung der Transaktionsdatenbank ein. Moderne Programmiertechniken zur Ausnutzung einer Multikernarchitektur und zum Zugriff auf Kompilierinformationen über Roslyn wurden eingesetzt, um die gestellten Anforderungen zu implementieren. Bei der Güteevaluation wurden Funktionalität und Performance der Lösung untersucht. Dabei wurde der Algorithmus auf acht Repositorys mit knapp 400.000 Codezeilen jeweils zehnfach ausgeführt, um mittlere Laufzeitergebnisse zu erhalten. Mithilfe des vorgestellten API Miners ist es möglich, Regeln zu lernen, die einen positiven Effekt auf die Software-Qualität haben. Die höchsten Performancegewinne können bei der Parallelisierung der obersten Ebene entlang der Codehierarchie (Parallelisierung der Projektverarbeitung) erwartet werden. Aus der Stichprobe kommt eine durchschnittliche Laufzeitverbesserung von ca. 45% in der optimalen Parameterkonstellation hervor. Aus einer kritischen Perspektive betrachtet, liefert die Assoziationsanalyse viel Noise und die Bewertung von gelernten Regeln ist aufgrund von subjektiver Interessantheit schwierig. Die Wirtschaftlichkeit der Gesamtlösung ist deshalb nicht transparent.

Zusammengefasst wurde in dieser Kurzpublikation unter anderem ein Ansatz vorgestellt, wie maschinelles Lernen mit statischer Code-Analyse zur Verbesserung von Software-Qualität kombiniert werden kann. Im Tool wurde Nutzpotenzial nachgewiesen, gleichzeitig wurden dabei seine Grenzen deutlich. Daraus kann abgeleitet werden, dass die innovativen ML-basierten Ansätze die klassische SCA nicht ersetzen werden, sondern als ergänzende Qualitätssicherungsmaßnahme in Erwägung gezogen werden können. Für die Zukunft wird ausgehend von den Ergebnissen dieser Arbeit prognostiziert, dass sowohl klassische als auch intelligente SCA gemeinsam genutzt werden, mit dem Ziel, die steigenden Ansprüche an Software-Qualität zu erfüllen und daraus einen Wettbewerbsvorteil zu generieren. Die intensive Forschung im maschinellen Lernen lässt vermuten, dass die genannten Schwierigkeiten überwunden werden können.

Literatur

[Agra93] Rakesh Agrawal, Tomasz Imieliński und Arun Swami: "Mining associati-

on rules between sets of items in large databases". In: *ACM SIGMOD Record* 22.2 (1993), S. 207–216.

[Geng06] Liqiang Geng und Howard J. Hamilton: "Interestingness measures for data mining". In: *ACM Computing Surveys* 38.3 (2006), 9–es.

[Hurw18] Judith Hurwitz und Daniel Kirsch: *Machine Learning: for dummies*. John Wiley & Sons, Inc., 2018.

[ISO20] *ISO 25010*. Zugriff am 04.03.2020. 4.3.2020. URL: https://iso25000.com/index.php/en/iso-25000-standards/iso-25010.

[Sult18] Eldar Sultanow, Stefan Konopik, André Ullrich und Gergana Vladova: "Machine Learning based Static Code Analysis for Software Quality Assurance". In: (2018). Hrsg. von IEEE.

[Toub10] Stephen Toub: *Patterns of Parallel Programming: Understanding and Applying Parallel Patterns with the .NET Framework 4 and Visual C#*. 2010.

[Univ08] University of Waikato: *Attribute-Relation File Format (ARFF)*. Zugriff am 18.02.2020. 1.11.2008. URL: https://www.cs.waikato.ac.nz/ml/weka/arff.html.

[Wagn20] Bill Wagner: *Das .NET Compiler Platform SDK (Roslyn APIs) | Microsoft Docs*. Zugriff am 28.02.2020. 28.2.2020. URL: https://docs.microsoft.com/de-de/dotnet/csharp/roslyn-sdk/.

[Weka20] *Weka 3 - Data Mining with Open Source Machine Learning Software in Java*. Zugriff am 18.02.2020. 2020. URL: https://www.cs.waikato.ac.nz/ml/weka/.

Ausgewählte Anwendungen der Künstlichen Intelligenz und deren Auswirkungen auf die beruflichen Tätigkeiten - Eine Momentaufnahme

Christoph Hauser
Hochschule Luzern - Wirtschaft
Zentralstrasse 9, 6002 Luzern
E-Mail: christoph.hauser@hslu.ch

Ute Klotz
Hochschule Luzern - Informatik
Suurstoffi 1, 6343 Rotkreuz
E-Mail: ute.klotz@hslu.ch

ABSTRACT
Die Anwendungen der Künstlichen Intelligenz sind immer noch neu, wenig sichtbar, werden in der Öffentlichkeit kontrovers diskutiert und sind für die Mitarbeitenden sowohl mit negativen und als auch mit positiven Szenarien verbunden. Das Forschungsprojekt, das diesem Beitrag zugrunde liegt, möchte die Auswirkungen von KI-Anwendungen auf das Anforderungsprofil und das Learning-on-the-Job von Mitarbeitenden feststellen, und auch klären, welche Auswirkungen dies auf die Arbeitsproduktivität und auf die arbeitende Person selbst hat. Dazu wurden narrative Interviews mit sieben Interviewpartner:innen aus sechs Branchen aus der Schweiz, mehrheitlich der Zentralschweiz, geführt. Die Gesprächsprotokolle der Interviews wurden in mehreren Stufen inhaltlich analysiert. Die Ergebnisse zeigen Auswirkungen auf die menschlichen Fähigkeiten, die Mensch-Maschine-Interaktion und die Unterstützung des Menschen bei der täglichen Arbeit. Es zeigen sich Vorteile, wie z.B. produktivere Prozesse, Herausforderungen, wie z.B. der geringere Kundenkontakt, aber auch kritische Aspekte in Bezug auf die Aus- und Weiterbildung.

SCHLÜSSELWÖRTER
Beschäftigung, Berufe, Tätigkeiten, Kompetenzen, Künstliche Intelligenz, Regionalentwicklung, Ökonomische Resilienz, Zukunft der Arbeit

EINFÜHRUNG

Die Diskussion über den Einfluss der Digitalisierung auf die Zukunft der Arbeit wird seit einigen Jahren geführt, insbesondere der Einfluss der Künstlichen Intelligenz (Arntz 2020). Als Gründe für den Einsatz von KI werden u.a. Kostensenkung, Qualitätssteigerung und die Einführung neuer bzw. die Anpassung bestehender Geschäftsmodelle genannt (Deckert und Meyer 2020). Die Verbreitung von KI-Verfahren in KMU wird aber dennoch und derzeit als gering angesehen, und das liegt einerseits an dem Mangel an Kompetenzen und der nicht ausreichenden Datenbasis (Deckert und Meyer 2020) aber andererseits auch an der als aufwändig zu erwartender Anpassung der Arbeitsorganisation (Frost et al 2020).
Die Publikationen, die sich eher mit den Auswirkungen des KI-Einsatzes auf die Arbeitswelt beschäftigen, sind vielfältig, können auf unterschiedliche Tätigkeiten fokussieren und trotzdem allgemein bleiben. Schön und wünschenswert wäre es, wenn man konkret nachschauen könnte, welche Berufe unter dem Einfluss der Digitalisierung, insbesondere der Künstlichen Intelligenz, wie verändert werden und welche Fortbildung wo und zu welchem Preis angeboten werden. Davon ist man noch etwas entfernt, aber durch die vermehrte Publikation von KI-Anwendungsszenarien (Barton und Müller 2021; Maris 2022, Miskolczi und Thingna 2022; OECD 2020), nimmt die Konkretisierung zu. Auch dieses Forschungsprojekt möchte dazu einen Beitrag leisten.

FORSCHUNGSFRAGE UND METHODIK

Im Rahmen des Projektes „Arbeitsmarktstudien“ an der Hochschule Luzern wollte man sich spezifische Arbeitsplätze anschauen, um herauszufinden, wie sich die spezifische Arbeit, das Anforderungsprofil und auch die Person selbst durch die Digitalisierung, insbesondere durch KI-Anwendungen, verändern.
Die übergeordnete Forschungsfrage lautete deshalb: *«Wie wirkt sich die Einführung von neuen Informationssystemen, insbesondere Anwendungen der Künstlichen Intelligenz, auf das Anforderungsprofil und den Learning-on-the-Job von Mitarbeitenden aus, und welche Auswirkung hat dies auf die Arbeitsproduktivität und auf die arbeitende Person selbst?»*
Es wurde ein mehrstufiges Vorgehen gewählt. Zuerst wurde die bestehende Literatur ausgewertet, um Fragestellungen für einen Interviewleitfaden zusammenzustellen. Dazu wurde eine sog. Longlist erstellt, und dann durch das interdisziplinär besetzte Projektteam auf einen konkreten Interviewleitfaden reduziert. Danach wurden mögliche Schweizer Interviewpartner identifiziert, und wenn möglich aus der Zentralschweiz, um den Stand der regionalen Entwicklung der Trägerkantone der Hochschule Luzern (Luzern, Uri, Schwyz, Obwalden, Nidwalden und Zug) (Systematische Rechtssammlung SRL - Kanton Luzern 2011) zu identifizieren. Hier wurde auf das Unternehmen- und Expertennetzwerk der Hochschule Luzern zurückgegriffen. Es wurden insgesamt sieben narrative Leitfadeninterviews geführt. Die Interviewteilnehmenden waren aus sechs verschiedenen Branchen. Der Interviewleitfaden wurde jeweils auf die aktuelle branchenspezifische Situation angepasst. Die Interviews wurden, wenn immer möglich, mit zwei Interviewenden aus zwei Departementen geführt, um die interdisziplinäre Sichtweise zu gewährleisten. Die Interviews dauerten zwischen 30 und 60 Minuten. Es wurden Gesprächsprotokolle erstellt, die von den Interviewenden miteinander abgeglichen wurden. Die Interviews wurden in mehreren Schritten analysiert (Rädiker und Kuckartz 2019). Zuerst wurden die Gesprächsprotokolle von einem Projektmitglied vercodet, d.h. Textstellen, die inhaltlich gleich waren, wurden mit den gleichen Codes versehen. Damit war

es möglich, einen Überblick über die verschiedenen Codes und deren Häufigkeit zu gewinnen.
Dabei wurden zwei Codierungsebenen übereinandergelegt. In einer ersten Ebene wurden Codes zugewiesen, die sich auf die Art einer Veränderung bezieht, nachdem eine KI-Anwendung eingeführt wurde. Die folgende Tabelle 1 zeigt die Kurzbezeichnung dieser veränderungsorientierten Codes und die Anzahl Sätze, die dem jeweiligen Code zugeordnet wurden.

Tabelle 1: Veränderungscodes

Nr.	*Kurzbezeichnung*	*Anzahl Sätze*
1	Statisch	185
2	Veränderung	85
3	Verbesserung exogen	47
4	Verbesserung endogen	142
5	Training	22
6	Kontrolle	59
7	Korrektur	2
8	Herausforderung	164
9	Verschlechterung	52

Daneben wurden zehn weitere Codes verwendet, die auf das beschriebene Objekt hindeuten. Die folgende Tabelle 2 zeigt die Kurzbezeichnung dieser objektorientierten Codes und die Anzahl Sätze, die diesem Code zugeordnet wurden.

Tabelle 2: Objektcodes

Nr.	*Kurzbezeichnung*	*Anzahl Sätze*
10	Menschliche Skills	125
20	Maschinen-Skills	21
30	Mensch-Maschine-Interaktion	49
40	Maschine-Mensch-Interaktion	80
50	Problem zu lösen	147
60	Problemlösung	123
70	Fehlerreaktion	93
80	Organisationsebene	66
90	Umweltebenen	55

In Kombination würden diese Codes 81 Kombination ergeben; einige sind jedoch von vorneherein als unwahrscheinlich ausgeschlossen worden. Von den so vorselektierten 36 Codes wurden drei nie zugeordnet. Die 33 verwendeten Codes wurden sodann inhaltsanalytisch zu Aussagegruppen verbunden. Zusammen fungierten die veränderungsorientierten Codes, die objektorientierten Codes, deren Kombination und die Aussagegruppen als Code-Memo bei der Codierung (Rädiker und Kuckartz 2019). Die derart gefundenen und sortierten Resultate sind im Resultatkapitel wiedergegeben.

BRANCHENSPEZIFISCHE SITUATION

Die Interviewpartner, die zum Stand ihrer KI-Anwendungen befragt wurden, repräsentieren folgende Branchen bzw. Aufgabenbereiche und werden in alphabetischer Reihenfolge aufgeführt: Eisenbahn, Kundendienst, Polizei, Radiologie, Steuerverwaltung, und Versicherung.
Zu allen Bereichen wurde die bestehende Literatur analysiert, um eine Beschreibung der berufsspezifischen Arbeitssituation zu erhalten und basierend darauf den Interviewleitfaden anzupassen. Der aktuelle Stand wird nachfolgend je Bereich wiedergegeben.

EISENBAHN

Die Europäische Kommission hatte das Jahr 2021 zum "European Year of Rail" erklärt, und wollte damit die Vorteile der Bahn als nachhaltiges, sicheres und «smartes» Verkehrsmittel hervorheben. U.a. sollte das Passagier- und Frachtvolumen erhöht, die Anstrengungen für den Einsatz von Hochgeschwindigkeitszügen verbessert und multimodaler Transport über europäische Landesgrenzen vereinfacht werden (European Commission 2020). Die Probleme und somit die Einsatzgebiete von Künstlicher Intelligenz zeigen sich aber oftmals in anderen Bereichen. Gemäss Balser (2022) möchte die Deutsche Bahn mithilfe von KI den Mangel an Gleisen besser managen, die Disponentinnen und Disponenten in den Leitstellen unterstützen, Live-Simulationen machen, um Verspätungen früh genug zu erkennen und Echtzeit-Fahrpläne einführen, um die Ressourcen und deren Kapazitäten besser managen zu können (siehe auch Deutsche Bahn 2021).
Die Schweizer Südostbahn mit der JustGo-App ein vereinfachtes Ticketsystem an, bei dem nach der Registrierung mittels Bluetooth das Ein- und Aussteigen bei den definierten Zügen und Bussen erkannt und am Tagesende zum günstigsten Preis abgerechnet wird (Scordamaglia 2019; Noser Engineering o.J.). Die Österreichischen Bundesbahnen (ÖBB) bieten in bestimmten Zügen verschiedene Medienprodukte zur Unterhaltung an, die Französische Bahn (SCNF) bietet automatische Ticketänderung bei verfrühten Ankunftszeiten und die Spanische Bahn (AVE) bietet via App Informationen zu Einkaufsmöglichkeiten in den Bahnhöfen an. Die Digitalisierung setzt sich aber auch im Bereich der Sicherheit fort. Während die installierten Kameras im Zug und auf Bahnhöfen fast schon zum Standard gehören, sind die ferngesteuerten Kameras, Sensoren und Drohnen zur synchronen oder asynchronen Überwachung der Züge und des Schienennetzes eher neu. So setzt der britische Infrastrukturbetreiber Network Rail Drohnen ein, um heikle Bahnübergänge oder beschädigte Infrastruktur zu überwachen und die Schweizerischen Bundesbahnen (SBB) haben ein Fahrerunterstützungssystem auf der Strecke Bern-Olten getestet, um die optimale Geschwindigkeit für den Fahrer zu berechnen und konsequenterweise den Energieverbrauch zu senken (Scordamaglia 2019). Im Bereich der vorausschauenden Instandhaltung (Predictive Maintenance) werden sowohl die Infrastruktur als auch das Rollmaterial mithilfe der erhaltenen Sensor- und Prozessdaten zur Fahrtzeit überwacht und am

Tagesende analysiert. Die darauf basierenden Diagnosesysteme ermöglichen es, Abweichungen festzustellen, die wiederum auf beschädigte Zug- oder Gleisfehler schliessen lassen. Wenn Defekte frühzeitig erkannt werden, können Instandhaltungsmassnahmen rechtzeitig geplant werden und helfen somit auch die Fahrzeugverfügbarkeit zu erhöhen. Es werden weiterhin Gleismesswagen eingesetzt, die die Linien alle 3-12 Monate überprüfen (Moshammer 2020). Die Kombination von Daten, die permanent und in Intervallen erhoben werden, stellen auch von der Datenmenge eine Herausforderung dar. So sendet eine moderne Lok circa 500 Millionen Datenpunkte pro Jahr (Siemens 2016).

KUNDENDIENST

Rainsberger (2021) meint, Kunden wollen alles sofort erledigt haben. Das gilt auch für die Kommunikation mit dem Unternehmen. Es muss in Echtzeit passieren und auf den gewohnten Geräten und Softwarelösungen stattfinden, ansonsten verliert der Kunde das Interesse.
Chatbots sind eine Technologie, die die Automatisierung der Kundenschnittstellen ermöglichen, und die oben genannten Anforderungen erfüllen. Sie können im Kundendienst, bei der Produktsuche oder bei Reservationsprozessen eingesetzt werden (Stucki, D'Onofrio & Portmann 2020). Die meisten Chatbots helfen zwar nur bei einfachen Fragen, aber trotzdem ermöglichen sie es dem Unternehmen personelle und finanzielle Ressourcen einzusparen (Fichter 2017). Grundsätzlich werden zwei Arten von Chatbots unterschieden: regelbasierte und KI-basierte. Bei ersteren sind die Fragen entweder vordefiniert und der Benutzer muss eine Auswahl aus den vorgegebenen Optionen treffen oder der Chatbot versucht die eingegebenen Schlüsselwörter zu erkennen und darauf basierend die Antworten zu geben. Die KI-basierten Chatbots versuchen das Gespräch menschenähnlich zu führen, indem sie entweder den Kontext aus den vorhandenen Informationen herstellen oder in natürlicher Sprache mit dem Nutzer kommunizieren (Rainsberger 2021). Unternehmen, die Chatbots einsetzen, sind z.B. im Bereich

- E-Commerce der Discounter Netto mit dem Netto-Online-Chatbot, der auf der Kontaktseite zur Verfügung steht und Auskunft zum Newsletter, zum Sortiment und zu Bestellung gibt (Stephanie 2021)
- Tourismus die deutsche Insel Norderney, mit dem Chatbot Leevke, der Fragen zur An- und Abreise und zu Unterkunftsmöglichkeiten beantwortet (Stephanie 2021)
- Support die Schweizerische Post, die im Informatik Service Desk als weiteren Kommunikations- und Informationskanal den textbasierten UHD Chatbot einsetzt, um einfache Fragen zu Softwareinstallationen beantworten oder Service Tickets erstellen zu können (Stucki, D'Onofrio & Portmann 2020).

Wenn die Chatbots textbasiert sind, dann muss der Nutzer das Problem verschriftlichen, was aber für manche Kunden eine Barriere darstellen kann. Die Konzeption von Chatbot-Dialogen erfolgt auf der Basis von bisherigen Kundendialogen, die entsprechend ausgewertet und von Fachpersonen überarbeitet wurden, um sicher zu gehen, dass die erwartete Qualität erreicht wird (Stucki, D'Onofrio & Portmann 2020). Es ist deshalb verständlich, dass die erstmalige Implementierung eines Chatbots sowie dessen Weiterentwicklung und Integration in die unternehmensinterne Systemlandschaft hohe Kosten verursachen kann (Schacker & Fuchs 2018).

POLIZEI

Der Entwicklungsstand der «Smart Criminal Justice» in der Schweiz, also dem Einsatz von (intelligenten) Algorithmen in der Polizeiarbeit und Strafrechtspflege, wird in der empirischen Studie von Simmler, Brunner & Schedler (2021) umfassend dargestellt. Eines der Ergebnisse ist, dass zwar in allen Schweizer Kantonen Algorithmen eingesetzt werden, aber nur vereinzelt sind es intelligente Algorithmen. Mit der Beschaffung von neuen Softwarelösungen hat sich einerseits der Schwerpunkt hin zu einer präventiven Polizeiarbeit verschoben, und andererseits helfen Algorithmen die Effizienz, die Qualität und die Transparenz bei Arbeits- und Entscheidungsprozessen zu steigern. Zur Polizeiarbeit gehören die Ermittlung (Kriminalarbeit), das personenbezogene Predictive Policing und das raumzeitbezogene Predictive Policing (ebd.). Beim personenbezogenen Predictive Policing geht es um die Identifikation von gefährlichen Personen mithilfe von Wahrscheinlichkeitsberechnungen und eine frühzeitige Intervention im Sinne einer Ansprache und Deeskalation. Beim raumzeitbezogenen Predictive Policing geht um die Identifikation von möglichen Tatorten und Tatzeiten mithilfe u.a. von Mustererkennung, wobei es mehrheitlich beim Wohnungseinbruchdiebstahl eingesetzt wird. Die Analyseverfahren basieren nur beschränkt auf Künstlicher Intelligenz und Maschinellem Lernen. Und dennoch ist die Software nicht unumstritten, da sich die Polizisten nicht mehr auf die eigene Erfahrung verlassen müssen und ein Arbeiten ohne Algorithmen fast nicht mehr möglich erscheint. Letztendlich liegt aber die Entscheidungskompetenz beim Menschen. Ein weiterer Kritikpunkt an der Softwarelösung ist der Fokus auf grossstädtische Gegebenheiten, die nicht immer die Realität in den kleineren Schweizer Städten abbilden. Gemäss Leese (2018) werden wesentliche Informationen bei der Polizeiarbeit im Rahmen eines Falles als Fliesstext erfasst. Das wiederum erschwert eine systematische Auswertung. Eine starke Arbeitsbelastung kann eine eventuell nicht medienbruchfreie Datenerfassung weiter negativ beeinflussen und somit auch die Analyseergebnisse.

RADIOLOGIE

Vereinfacht ausgedrückt schaut sich der Radiologe Bilder respektive Fotos an und analysiert diese. Dazu benötigt er Erfahrungswissen, denn er muss wissen, nach was er suchen muss und wie die vorhandenen Informationen zu interpretieren sind (Krupinski 2003). Softwarelösungen, die den Radiologen bei der Analyse der Bilddaten unterstützen, werden schon seit Jahrzehnten eingesetzt (Wittpahl 2019). Gemäss Castellino (2005) wird die Softwarelösung zur Mustererkennung eingesetzt, um erste Auffälligkeiten zu identifizieren. Der Radiologe überprüft diese dann, setzt die Softwarelösung nochmals ein und bewertet die auffälligen Bereiche ein zweites Mal. Wittpahl

(2019) meint, dass die Softwarelösungen zusätzlich zur Vermessung der Auffälligkeiten eingesetzt, d.h. u.a. zur Bestimmung der Grösse und des Volumens. Wenn zudem eine Elektronische Patientenakte vorhanden ist, können auch Vergleiche mit älteren Aufnahmen (Wittpahl 2019) und Vorhersagen über den Verlauf gemacht werden (Miotto et al 2018). Die Arbeit des Radiologen wird weiter vereinfacht, wenn sich die Softwarelösungen für die Bildanalyse zudem in den Prozessablauf des Krankenhauses integrieren lassen, d.h. ein Datenaustausch mit den vorhandenen Kommunikations- und Archivierungslösungen (an Tang et al 2018) vorhanden ist. Spracherkennungssysteme werden ebenfalls eingesetzt, um den Dokumentationsprozess zu vereinfachen. Sofern die Befundung strukturiert abgelegt wird, kann diese auch leichter automatisch ausgewertet werden (Jungmann et al 2018). Zusätzlich könnte man auch die Tätigkeit der Fallcodierung durchführen, um so die Abrechnung mit der Krankenkasse zu beschleunigen (Gödeke et al 2020). Haubold (2020) weist darauf hin, dass KI bisher nicht standardmässig in allen Krankenhäusern eingesetzt wird (siehe auch Forsting 2019).

STEUERVERWALTUNG

Seit 2016 ist in Deutschland die automatische Steuerprüfung rechtlich möglich. Die dafür eingesetzte Softwarelösung prüft, ob die Steuerklärung nachvollziehbar und schlüssig ist. Bis zum Jahr 2022 soll die Hälfte der Steuererklärungen auf diese Weise bearbeitet werden. Bislang werden noch keine Algorithmen der Künstlichen Intelligenz oder des Maschinellen Lernens verwendet. Die Arbeitsweise bei den Steuerbehörden hat sich aber dennoch verändert. Geprüft werden nur noch die Steuererklärungen, die u.a. nicht plausibel erscheinen und Freitexte enthalten, um sie dann priorisiert abzuarbeiten. Die Bearbeitung dieser Steuerklärung muss nachweislich dokumentiert werden. Schlussendlich bedeutet das, dass sich die Mitarbeitenden um weniger, aber dafür um komplexere Steuererklärungen kümmern müssen (Kleinz 2018). Es ist aber so, dass die beschriebene Automatisierung nicht bei allen Steuerbereichen gleich realisiert ist, sondern vor allem bei der Lohnsteuer (Bizer 2019). Der nächste Automatisierungsschritt wäre, dass die Steuerbelege, wie z.B. die für die jährlichen Sozialversicherungsbeiträge, direkt an die Steuerbehörde übermittelt (Bizer 2019; Kleinz 2018) oder definierte Angaben an andere Behörden weitergeben werden (Bizer 2019). Das Letztere würde dem Once-Only-Prinzip entsprechen, das heisst, die Angaben müssen nur noch einmal angegeben werden (Bizer 2019). Die Datenmengen sind gross und könnten noch grösser werden, weil durch die automatisierte Verarbeitung noch mehr Angaben vom Steuerpflichtigen verlangt werden könnten. Dafür müssten leistungsfähige Rechenzentren zur Verfügung stehen. In Deutschland haben sich deshalb mehrere Bundesländer zusammengeschlossen, um diese Infrastruktur arbeitsteilig zu nutzen (Bizer 2019).

In der Schweiz kommen die Initiativen, die sich mit dem Einsatz Künstlicher Intelligenz in der Steuerverwaltung auseinandersetzen, von einzelnen Kantonen. Während im Kanton St. Gallen (Genova 2018) seit dem Jahr 2016 fünf Prozent der jährlichen Steuererklärungen automatisiert verarbeitet, sind es im Kanton Bern im Jahr 2020 18 Prozent der natürlichen steuerpflichtigen Personen, die automatisiert veranlagt werden. Im Kanton Thurgau (Hämmerli 2019) überlegt man, ob man Algorithmen zur Erkennung von Steuerhinterziehung einsetzen soll, im Kanton Zürich (Staatskanzlei 2021) möchte man diese eventuell und u.a. auch für eine verbesserte Servicequalität und Kundenorientierung einsetzen, indem man den Bürgerinnen und Bürgern die Steuerklärung schon vorausgefüllt zur Prüfung schickt. Am weitesten fortgeschritten scheint der Einsatz von Künstlicher Intelligenz in der Steuerverwaltung beim Kanton Obwalden zu sein, hier werden die Steuererklärungen seit dem Jahr 2020 automatisiert geprüft (Nufer & Ackermann 2021).

VERSICHERUNG

Gemäss Gruhn (2018) können KI-Systeme gerade in Versicherungen ihre Leistungsfähigkeit zeigen. Von Busch (2019, S. 215) werden Versicherungen als «datengetriebene Unternehmen» bezeichnet, deren Versicherungsprodukte auf Informationen, mathematischen Modellen und Risikobewertungen bestehen, die letztendlich in einem Vertrag abgebildet werden. Die Verträge selbst können mittlerweile situativ bzw. adaptiv sein, d.h. die Verträge könnten an- und ausgeschaltet werden, wenn sie gebraucht werden. So wird z.B. die Diebstahl- bzw. Einbruchversicherung dann eingeschaltet, wenn die Wohnung verlassen wird oder das Reisegepäck wird abhängig vom Standort zu unterschiedlichen Tarifen versichert. Im Marketing & Sales Bereich könnten solche Angaben wie z.B. das Stornoverhalten des Kunden, der erwartete Zahlungsausfall (Zabel 2020) und das Cross-Selling Potenzial helfen (Busch 2019), Entscheidungen hinsichtlich dem Kundenbeziehungsmanagement vorzubereiten (Schmidt 2020).

Als sinnvoll wird auch die Automatisierung und somit die Dunkelverarbeitung, d.h. keine manuelle Bearbeitung (Hahn & Zwiesler 2018) der zahlreichen Prozesse gesehen (Mangei 2019). Ein Schadensregulierungsprozess könnte zukünftig so aussehen: ein Kunde meldet einen Wasserschaden durch Hochwasser. Daraufhin werden die Wetterdaten zum Zeitpunkt des Schadens überprüft. Eine Drohne macht Aufnahmen der Schadenssituation, eine Software vergleicht die Vorher-Nachher-Aufnahmen und schätzt die Kosten für die Reparatur. Der Auftrag an den zuständigen Handwerker wird automatisch erteilt (Reich & Braasch 2019).

Ein weiteres Einsatzgebiet von Künstlicher Intelligenz ist die Betrugsbekämpfung. Hier können Schadenbilder mit Schadenschilderungen verglichen werden und Abweichungen aufgezeigt werden, und durch die Automatisierung könnten auch kleine Schadensfälle effizient geprüft werden (Busch 2019).

Busch (2019) meint, dass die KI-Modelle und -Systeme nicht mit den derzeitigen Organisationsmodellen und heterogenen IT-Systemlandschaften in den Versicherungen zusammenpassen. Er schlägt deshalb vor, dass neue Unternehmen gegründet werden, weil damit der schwierige Transformationsprozess des bisherigen Unternehmens umgangen wird. Ein Beispiel dafür ist Smile Direct, ein

reiner Online-Anbieter der Helvetia Versicherungsgesellschaft (Zeier Röschmann & Erny 2019). Parallel dazu bieten aber auch InsurTech-Startups ihre Dienstleistungen an. *Snapsure* zum Beispiel ermöglicht dem Kunden, ein Foto von dem Gegenstand zu machen, den er versichern möchte. Das Foto wird analysiert und dem Kunden wird ein Angebot für eine Fahrrad- oder eine umfassendere Hausratsversicherung gemacht.
Konsequenterweise könnte der Engpass für die Erfassung der Daten nicht mehr Versicherungsmitarbeiter sein, sondern die Speicherkapazität und die Rechenleistung (Körzdörfer 2020).

RESULTATE

Im Folgenden werden die Resultate aus der empirischen Erhebung dargestellt. Diese Ausführungen resultieren aus der Inhaltsanalyse mithilfe der im Methodenteil vorgestellten Codes. Häufig werden typische, illustrative Zitate aus den Interviews wiedergegeben, die jeweils einen repräsentativen Ausschnitt aus den Interviews zeigen. Teilweise sind darin Aussagen so verallgemeinert worden, dass die Branche nicht erkennbar ist. Zum Beispiel würde das Wort «Versicherungsnehmer» ersetzt mit «Kunde». Geschweifte Klammern enthalten eine Referenznummer zur Quelle, immer die Nummer des Interviewpartners, im Sinne einer Anonymisierung, und die jeweilige Satznummer.
Als «Mitarbeitende» bezeichnet werden im Folgenden jene Mitarbeitende, die mit einem neuen oder erneuerten Informationssystem ihre Aufgabe erledigen. Mitarbeitende haben grundsätzlich in zwei Feldern Fähigkeiten mitzubringen: Einerseits die Anwendung des neu geschaffenen Tools an sich, anderseits die meist über Jahre erworbene Fachkompetenz, um die spezifischen Aufgaben kompetent zu lösen. Wenn zu betonen ist, dass die Mitarbeitenden über spezifisches Fachwissen verfügen, wird von Fachkräften gesprochen; wenn diese auf dem Arbeitsmarkt sehr selten sind, wird von Expertinnen und Experten gesprochen. IT-Fachkräfte werden explizit als solche bezeichnet. Der Begriff «Aufgabe» bedeutet hier der sogenannte «job-to-be-done», also die Aufgabe, welche die Mitarbeitenden zu erledigen haben; jetzt eben auch mit einem fortgeschrittenen Informationssystem. Mit «Informationssystem» ist das gemeint, was Mitarbeitende als neu geschaffenes Werkzeug einsetzen, um ihre Aufgabe zu erledigen. Das technische Design der Informationssysteme, um die sich die Interviews drehen, ist jeweils unterschiedlich den unterschiedlichen Aufgaben angepasst. Gemeinsam ist, dass ein KI-Element darin integriert ist.

MENSCHLICHE FÄHIGKEITEN

Veränderte Anforderungen an die menschlichen Fähigkeiten

Die vorliegend gestellte Forschungsfrage geht menschliche Fähigkeiten nach und wie diese den Anforderungen an die Menschen entsprechen, wenn sich Anforderungen durch Informationssysteme verschieben. Sämtliche Interviewpartner sehen in den Fähigkeiten der Mitarbeitenden grundsätzlich einen sehr wichtigen, wenn nicht den wichtigsten Schlüssel zur Bewältigung der gestellten Aufgaben. Die einen zollen dem menschlichen Können hohen Respekt, wie etwa Menschen bei ihrer Aufgabe teilweise auch mehrere Sinne einsetzen würden {6.019, 6.411}. Aber die meistgenannte Problematik im Thema der menschlichen Fähigkeiten sind die steigenden Anforderungen an die Mitarbeitenden. Über alle Interviews gesehen steigen die Anforderungen an die Fähigkeiten der Mitarbeitenden generell, sowohl was die Technikaffinität als auch die Fachkompetenz anbelangt.

Fehlende Usability konkurrenziert Fachkompetenz

In einigen Fällen explizit erwähnt und in sämtlichen Interviews mindestens implizit zu erkennen ist ein Tradeoff zwischen der Usability der Informationssysteme und der Technikaffinität der Mitarbeitenden. Gerade wenn die Mitarbeitenden auch noch hohe Fachkenntnisse mitbringen sollen, kann man Fachkräfte mit einer intuitiven, einfachen Bedienbarkeit der Informationssysteme besser motivieren, diese auch richtig einzusetzen. In mancher Hinsicht scheint es mitunter eine Generationenfrage zu sein, wenn «…es vor allem die älteren Leute vom Stuhl haut, [wenn man] über fünf Bildschirme hinweg ein paar Einstellungen machen» muss {7.177}. Entscheidend sei aber insbesondere die Neugierde und Offenheit für neue Wege, die ausdrücklich auch bei älteren Mitarbeitenden da sein könne {3.082}. Geduld brauche es zuweilen, diese dürfe jedoch nicht strapaziert werden, und «das Element der Usability ist das zentrale Element einer App-Entwicklung» {6.616}. Würden die Mitarbeitenden sich mit «Rohlingen» {6.618} herumschlagen müssen, erzeuge dies Frustrationen {6.633}. Kurzfristig würden die Mitarbeitenden von ihrer fachlichen Aufgabe abgelenkt, wenn die Usability nicht gegeben ist {5.038}, langfristig könne dies zu resignierten Abgängen von an sich fachkompetenten Mitarbeitenden führen, was zu Fach-Know-How-Verlusten führt {5.040}. «An der Ergonomie der Systeme mangelts. […] Da müssen wir schauen, dass wir zu vereinfachten Systemen kommen» {7.179}. Diese Beobachtungen bedeuten mit anderen Worten, dass eine fehlende Usability von Informationssystemen seinen Preis oft bei einer nachlassenden, fachlichen Professionalität der Mitarbeitenden hat.

Erhöhte fachliche Anforderungen

Der Verlust von Know-How durch Ablenkung vom Inhaltlichen oder durch Fluktuation von Fachkräften wiegt aber nicht zuletzt deswegen schwer, da die Interviews in vielen Fällen darauf hindeuten, dass durch den Einsatz von Informationssystemen das Anforderungsniveau bezüglich Fachkompetenz auf allen Stufen ansteigt. Für eine Jobpolarisation, also für eine Zunahme von entweder hoch oder tief qualifizierten Jobprofilen wurde insbesondere bezüglich des tieferen Segments allerdings keine Evidenz gefunden. Einige Interviewpartner beobachten explizit, dass die am tiefsten qualifizierten Stellen verschwinden. Nicht die mittel qualifizierten Jobs würden abgebaut, wie dies die Hypothese der Jobpolarisation eigentlich prophezeit, sondern die sogenannte «digitale Sklavenarbeit» wie das Abtippen von Dokumenten falle

immer mehr weg {7.107} und des gebe einen «steten Druck nach oben» {7.162}. Am oberen Ende der Qualifikationsskala brauche es «zukünftig mehr Experten, die umfangreiche Fachkenntnisse haben, um die komplexeren Einzelfälle zu erkennen und zu prüfen» {4.025}. Dies sei mithin für jene Mitarbeitenden ein Problem, deren Entwicklungsperspektiven durch «gewisse intellektuelle Grenzen» {7.050} limitiert seien. Die den Mitarbeitenden zugeteilten Aufgaben erforderten auch nach Einführung innovativer Informationssysteme Fachkenntnisse und Erfahrung. Aber «auch den Experten wird […] die Arbeit etwas vereinfacht – sie müssen sich jedoch auf diese Vereinfachung im Denken verändern» {4.039}. Die Mitarbeitenden müssen sich demnach auf die neuen Tools einlassen können. Gleichzeitig bleibt die Erfahrung im Berufsfeld eine wichtige Voraussetzung der Arbeit, um die Zusammenhänge zu kennen: An «den Bildschirmen möchten wir Personen, die Erfahrung haben, als würden sie nach draussen gehen» {6.294}.

Vorhandene oder zu entwickelnde Technik-Affinität

Auch wenn eine gute Usability der Informationssysteme Raum für mehr Fachkompetenz gibt, so bleibt die Technik-Affinität in den Interviews ebenfalls ein bestimmendes Thema. Mit nur einer Ausnahme {1.016} konstatieren alle, dass es «eine Verschiebung der Tätigkeiten» {6.236} gibt. Hier zeigt sich ein Generationen-Problem auch auf indirekte Weise, nämlich, dass sich das technikaffine Denken in der Aus- und Weiterbildung verzögert manifestiert. Interviewpartner:innen bedauern, dass «Informatikkenntnisse einen eher kleinen Teil der Ausbildung» {2.040} ausmache, oder dass das «Thema Algorithmen» in der Ausbildung fehle {5.035}. Das liege eben daran, dass die Lehrenden die aktuelle Bedeutung der Informatik für ihr Fachgebiet zu wenig kennen und entsprechend zu wenig in die Curricula der einschlägigen Lehrgänge einbauen würden. Ob sich allerdings so etwas wie eine generelle Technik-Affinität in eine Ausbildung einbauen liesse, wird allerdings auch angezweifelt {3.085}. Die konkreten Informationssysteme seien daher umso mehr «sicher on-the-job» zu erlernen {6.319}. Dazu dürfe man «keine Angst haben» und «man muss ein iPad in die Hand nehmen und damit umgehen können» {3.086}, aber «es wird nicht erwartet, dass jeder programmieren kann» {4.029}.

MENSCH-MASCHINE INTERAKTION

Damit rückt die gelingende Komplementarität von Mensch und Maschine ins Blickfeld. In den Interviews wird die Richtung der Kommunikation von Mensch zu Maschine etwa gleich oft erwähnt wie die umgekehrte Richtung von Maschine zu Mensch. Ein Schwerpunkt liegt jedoch interessanterweise bei der Kommunikation von Kund:innen direkt zu Informationssystemen.

Kommunikation - Kundschaft zu Maschine

Bei den interviewten Dienstleistern hat das neu eingeführte Informationssystem oft den Hauptzweck, die Kommunikation der Kund:innen abzukürzen und direkt in die Informationssysteme einzubinden. Dadurch wird den Mitarbeitenden dieser Unternehmen eine entsprechende «Übersetzungsarbeit» vom «Wunsch des Kunden» in die Sprache der Maschine abgenommen {3.029}. Versicherungsnehmer:innen, Patient:innen, Steuerzahler:innen oder Rechnungssteller:innen «sprechen» im Zuge der Dienstleistungserstellung in einem ersten Schritt mit dem Informationssystem, welches in einem weiteren Schritt und abhängig von Art und Komplexität des Einzelfalls die weitere Bearbeitung durch die Mitarbeitenden der Unternehmen vorbereitet und zuweist, manchmal auch automatisch erledigt. Wie weit jene Vorbereitung geht, ist vom Fall abhängig, und kann etwa das Zuteilen von Anfragen an bestimmte Mitarbeitende bis hin zum Entscheid führen, ob eine Kundeninteraktion überhaupt einen weiteren menschlichen Eingriff benötigt.

Die Informationssysteme kommen der Kundschaft dabei in Form und Sprache zunehmend entgegen, aber auch die Kundschaft gewöhnt sich zu guten Teilen an das Interagieren mit einem System. Das kann so weit gehen, dass «der Grund für die Einführung eines Chats, also einer schriftlichen Beratung anstelle einer telefonischen Hotline ist, dass die Kunden nicht mehr telefonieren wollen» {1.004}. Damit kann «viel Arbeit […] zum Kunden ausgelagert [werden], man stellt nur noch das System hin und es gibt ein paar, die hintenrum Kontrollen durchführen» {7.201}. So kann man alles «direkt mit der Datenbank abgleichen» {6.241}.

Kommunikation - Mitarbeitenden zu Maschine

Zwar wurde in den Interviews die Usability der Tools für die Mitarbeitenden oft diskutiert, doch wie genau die Kommunikation von Mitarbeitenden mit den Informationssystemen gelingt, wurde aber weniger konkret thematisiert. Dazu werden eher vage Visionen skizziert, wie «man kommt morgens [an den Arbeitsplatz], schaut sich elektronisch den Tagesplan [oder die Dringlichkeiten] an und überprüft die dazu gehörigen Handlungsempfehlungen» {2.001}. Die Visionen bleiben bei Unterstützung für den Menschen und gehen kaum weiter bis zu seinem Ersatz: «Die erwähnten Handlungsempfehlungen sollen aber immer Empfehlungen bleiben, die Entscheidungen werden von den verantwortlichen Personen getroffen» {2.008}. Aus heutiger Sicht seien die «zur Verfügung stehenden Technologien und Systeme […] lediglich unverzichtbare Hilfsmittel» {2.039}. Die Überlegenheit der Maschine gegenüber dem Menschen wird in einigen Bereichen erwartet, so «ist meine Hoffnung, dass in den [Rohdaten] noch Informationen von der Maschine erkennbar sind, über die der Mensch hinwegschaut» {6.397}.

Kommunikation - Maschine zu Kundschaft

Auch die Kommunikation in die andere Richtung – von Maschine zum Menschen – fokussierte in den Interviews weniger auf die Mitarbeitenden als auch hier direkt zur Kundschaft. Im Allgemeinen ist diese Richtung weniger herausforderungsreich, weil Text- und Sprachausgaben von Computern längst etabliert sind. Thema ist, wie ein System bei Unklarheiten bei Kunden nachhaken kann {3.027} und somit ein Dialog entsteht. Das Aufteilen von

Prozessen in einzelne Schritte ist punktuell auch in der Kommunikation zwischen Informationssystem und Mitarbeitenden angesprochen worden. Damit werde die Blackbox geöffnet respektive die Nachvollziehbarkeit auf Outputebene erhöht {5.076}. Dies erscheint insofern umso wichtiger, als die Maschine «nie den Menschen ersetzen kann» {1.021}, sondern es um «gegenseitige Unterstützung» {6.053} geht.

Weiterentwicklung der Algorithmen

Ein von den befragten Expertinnen und Experten vieldiskutiertes Thema ist das Training der Künstlichen Intelligenzen mit geeigneten Daten, die aus einer relevanten Erfahrung zu stammen haben. Hier kommen wieder die Mitarbeitenden stärker ins Spiel, denn «der Aufwand ist relativ gross, um das System, respektive die Algorithmen zu konfigurieren» {2.030}.

Es ist zunächst zu entscheiden, welche Daten als Trainingsdaten überhaupt geeignet sind. Klar ist: Mehr Daten bringen bessere Ergebnisse, aber die Daten müssen auch qualitativ stimmen. Wenn ein System darauf ausgelegt ist, Störungen zu erkennen, kann es besonders schwierig sein, genügend Daten zur Art dieser Störungen zu erhalten: «Unsere [Objekte] sind nicht in einem solch liederlichen Zustand, dass wir ganz viele Beispiele hätten für Nichtkonformitäten und für Abweichungen» {6.120} und anderseits «braucht es reproduzierbare Fehler statt dauernder Änderungen» {5.067}. Mit anderen Worten: Es ist insbesondere dort eine Herausforderung, eine Künstliche Intelligenz auf Fehlererkennung zu trainieren, wenn Fehler nicht entstehen sollten.

Es zeigen sich hier verschiedene weitere Schwierigkeiten und Dilemmata. Ein mehrfach genanntes Dilemma ist die Frage, wie weit in der Zeit zurückgegangen werden darf, soweit in der Vergangenheit Daten überhaupt gespeichert wurden. Geht man zu weit in die Vergangenheit zurück, können sich dort wichtige Umstände verändert haben und damit die Daten für die heutigen Umstände verzerren. So verändern sich etwa Rechtsgrundlagen {4.034} oder technische Prozesse {7.147} mit der Zeit. Konkret genannte Zeiträume, in denen auf vergangene Daten zurückgeschaut wird, reichen fünf oder zehn Jahre zurück, um noch hinreichend aktuell zu sein.

Von mehreren Interviewpartner:innen wurde unterstrichen, dass ein organisations- oder regionenübergreifendes Sammeln von Daten im Vordergrund steht, um auf hinreichend grosse Datenbasen zu kommen {5.006; 6.184}. Die Herausforderungen können dabei jedoch bei technischen Hürden wie unbekannten Dateiformaten und ähnlichem liegen. Privatwirtschaftliche Akteure können gar ein Interesse daran haben, durch technische Schranken das Teilen von Daten einzuschränken. «Da gibt es von diesen Firmen schon Tricks» {6.188}. Doch auch die eigenen Gesetze der Neuronalen Netzwerke können dazu führen, dass kompatibel erscheinende Datensätze eben doch nicht kompatibel sind. So «schauen» Neuronale Netzwerke ein Bild eines Gegenstandes bei unterschiedlichen Bildauflösungen völlig anders an {6.161}, da man nicht genau weiss, wie genau sich das Neuronale Netzwerk sein «Bildverständnis» auf Pixelebene antrainiert.

Wenn Daten strukturiert vorliegen, sehen sich Interviewpartner «sehr gut aufgestellt» {7.042}, aber in den untersuchten Beispielen sind strukturierte Daten nicht die Regel. Wegen zahlreichen derartigen technischen Herausforderungen wird das Daten-Training in den interviewten Fällen primär in IT-Abteilungen und nicht von den Mitarbeitenden in der Linie vorgenommen {1.010}. Die Mitarbeitenden und ihre realen, fachlichen Erfahrungen werden jedoch oft in gemischten, interdisziplinären Arbeitsgruppen oder ähnlich miteinbezogen, wo sich unter diesen verschiedenen Mitarbeitenden im idealen Fall gemeinsames Wissen akkumuliert {2.032; 3.016}. Dieser interdisziplinäre Austausch scheint nötig, denn «es ist so, dass der Software-Entwickler Annahmen trifft, und das ist gefährlich» {5.055}. Auch Hochschulen und andere Forschungspartner:innen spielen erhebliche Rollen bei der Verbesserung der KI-Systeme {6.343}. Oder auf den Punkt gebracht, «die wertvolle Arbeitskraft in Zukunft ist ziemlich matchentscheidend» {7.106} für Aufbau und Pflege von KI-Systemen.

MASCHINE HILFT DEM MENSCHEN – DIE VERANTWORTUNG BLEIBT BEIM MENSCHEN

Da die «wertvolle» Arbeitskraft generell eine knappe Ressource ist, besteht ein sehr wesentliches Ziel beim Einsatz von KI-Systemen oder ähnlichen digitalen Tools darin, ebendiese Ressource zu entlasten. «Die Maschine soll den Menschen unterstützen; sie entlastet [die Fachpersonen] und [jene Option] kann vorgeschlagen werden, welche die Beste ist» {5.086}. Dies heisst aber auch, der Mensch bleibt im Lead, insbesondere wenn es um finale Entscheidungen geht {2.008} respektive um das Bestätigen von maschinell getroffenen Entscheiden: «Man braucht immer noch Menschen wegen dem Vier-Augen-Prinzip» {4.077}. Die von der Maschine vorbereiteten Vorschläge «…gehen dann trotzdem durch die Hände von Personen» {6.264}, «das macht die Maschine noch nicht abschliessend» {6.270}.

Maschine hilft dem Menschen – in erheblichem Mass

Indem die Maschine die Arbeit des Menschen in erheblichem Mass vorbereitet, sind die Fachkräfte dadurch beispielsweise weniger im Feld, aber öfters hinter einem Bildschirm und beurteilen das, was die Maschine vorbereitet hat {6.275}. Die Entlastung von qualifizierten Mitarbeitenden verändert in einigen Fällen deren Aufgaben deutlich. «Es ist eine neue Art,…» wie manche hochqualifizierten Berufe zu interpretieren seien {5.009}. Wenn die Vorbereitungsarbeit der Maschine für den Menschen nicht hinreichend substanziell ist, dann wird die Maschine mitunter infrage gestellt oder gar ignoriert. «Wenn man sophisticated Methoden anwendet, und dann am Bildschirm aufleuchtet, ‘da stimmt etwas nicht’, dann müssen es dann schon wertvolle Hinweise sein, die da einem Mitarbeitenden gegeben werden» {7.045}. Wenn der Mensch sich in diesem Sinne nicht wirklich unterstützt fühlt, dann «fängt er an, [die maschinelle Hilfe] zu ignorieren» {7.047}.

In einigen Fällen ist eine maschinelle Unterstützung durch Deep Learning oder andere geeignete Tools aber schlicht notwendig. Die Fälle werden häufiger, in denen

wegen der schieren Menge an Daten, die vorhanden und zu interpretieren oder auf Unregelmässigkeiten zu prüfen sind, Expert:innen mengenmässig überfordert würden. In den Interviews sind zwei unterschiedliche Modi zu unterscheiden. Erstens die Vorinterpretation von Primärdaten: «Es sind zu viele Daten, die können nicht mehr verstanden werden» {5.080}. Dabei werden Daten zu besser verständlichen Informationen verdichtet oder gedeutet, aber es werden keine Auslassungen angestrebt. Davon zu unterscheiden ist ein zweiter Reduktionsmodus, bei dem die Maschine Fälle selektiert und gewisse Auslassungen angestrebt sind. Damit scheidet ein Teil der Daten als uninteressant aus dem weiteren Prozess aus, während ein anderer Teil als für den Menschen zu prüfend selektiert wird. Bei den zur weiteren Prüfung ausgewählten Daten haben die Fachpersonen umso mehr Zeit, sich diesen von der Maschine vorgeschlagenen Fällen zu widmen. «Und dann muss man halt auch von Hand genauer hinschauen und die Erfahrung spielen lassen {7.090}». Dieser Aufwand ist aber nicht mit allen Daten zu leisten. Der relevante «…Teil muss […] isoliert werden, […] der Rest muss dann nicht mehr angeschaut werden» {5.084}. Der ausgeschiedene «Rest» entgeht in der Folge dem menschlichen Auge, und was der Mensch «…nicht auf den Bildschirm bekommt, für das kann er keine Verantwortung übernehmen» {6.447}. Oft nicht explizit zu beantworten ist in solchen Fällen sodann, was bei Fehlern geschieht, denn «es gibt auch noch keinen Präzedenzfall» {6.448}.

Der Umgang mit Fehlern

Wer trägt die Verantwortung, wenn im Zusammenspiel von Mensch und Maschine fehlerhafte Entscheidungen herauskommen? Dies wird als «eine ganz delikate Frage» {6.421} bezeichnet. Denn einerseits ist «ein Problem […] auch, dass Systeme aus falschen Daten lernen, zum Teil» {4.043}, und anderseits gilt: «…der Maschine die Schuld zu geben ist juristisch nicht möglich… Das macht uns etwas Sorgen und wir müssen schauen…» {6.431}. Damit wird bei Aufbau und Pflege der Mensch-Maschine-Systeme insgesamt recht stark auf robuste, wenig fehlerbehaftete Resultate geachtet. Die Zuverlässigkeit wird mit verschiedenen Methoden überprüft, wobei auch «Zufriedenheitsmessungen bei Kunden und Leistungserbringern» {7.130} sowie Rückmeldungen der Mitarbeitenden selbst beachtet werden. «Wenn man mit Algorithmen kommt, dann will man auch den Mitarbeitenden die Möglichkeit geben, die Vorschläge zu prüfen. Das war jetzt eine gute Vorlage oder eben auch nicht, da gibt es dadurch Rückkopplungen. Rückweisungen werden dann auch wieder analysiert» {7.122}

Nebeneffekt: Der Mensch wird überwacht

Auch wenn jeweils das Informationssystem zu einem völlig anderen Zweck eingerichtet wurde: Es hat häufig den Nebeneffekt, dass die Leistung des Fachpersonals objektiv(-er) sichtbar wird. «Man kann im Rahmen des Testens […] feststellen, welche [Fälle] wurden früher korrekt gemacht und welche nicht. […] Das ist aber kein Aspekt, den wir bewusst verfolgen» {4.063}. Und doch wird dadurch die Performance von einzelnen Mitarbeitenden in einer zuvor nicht bestehenden Objektivität offengelegt. «Mit mehr Transparenz wurden die [Fachpersonen] dann aber auch vergleichbar, es ergab sich ein Benchmarking» 7.023}.

Wo früher «Aussage gegen Aussage» {3.094} stand, ist heute «…digital ein Riesenvorteil. Jetzt kann ich sauber nachverfolgen, was wann gesagt wurde» {3.097}. Als Nebenprodukt der Digitalisierung entsteht eine Datenspur, welche in umstrittenen Fällen als Beweismaterial dienen kann: «Wir können es euch beweisen, dass wir diese Daten haben, die wir analysiert haben» {6.473}. Dies bindet die Mitarbeitenden direkt und indirekt stärker an vorgegebene Regeln. Die Daten, welche die Handlungen oder Entscheidungen der Fachpersonen widerspiegeln, «gehen dann [zu den höheren Steuerungsebenen] – das lässt dann auch weniger die kreativen Lösungen zu» {3.105}. Der Ermessensspielraum wird enger – was je nach Umständen und Perspektive ein Vor- oder auch ein Nachteil für die Kundschaft oder für die Allgemeinheit, also «ein Fluch und Segen zugleich ist» {3.092}. Ein Nachteil wäre beispielsweise, wenn für einen speziellen Fall kein kulantes Augenmass mehr möglich ist. Vorteilhaft ist etwa, dass weniger Übervorteilungen oder gar korrupte Handlungen möglich sind.

VORTEILE VON MENSCH-MASCHINEN-SYSTEMEN (E)

Kriterienorientierte Prozesse

Die Objektivierung, welche die Maschinen mit ihrem algorithmischen Vorgehen und ihrer Datenspur in die Prozesse einbringen, wird insgesamt als ein Vorteil erachtet. Die Prozesse würden «klar prozessorientiert oder kriterienorientiert, objektiver Kriterien folgend» {6.043} und damit nachvollziehbarer.

Produktivere Prozesse

Ein grundsätzliches Ziel ist die Verbesserung der Produktivität. «Wir wollen natürlich immer mehr optimieren, automatisieren» {6.479}. Je nach Reifegrad des Informationssystems sind die beobachteten Fälle sehr unterschiedlich weit. Teilweise ist der Aufwand insgesamt noch grösser, weil das Informationssystem sich noch im Aufbau befindet {4.040}. Teilweise bringt das Informationssystem so weit eine Entlastung, dass «damit das Marktwachstum […] abgefedert werden kann» {3.059}. Teilweise ist ein Produktivitätsfortschritt der menschlichen Arbeit um mehrere Faktoren den ausgereiften IT-Systemen gutzuschreiben {7.052}. Teilweise wird der Produktivitätsfortschritt als unbedingte Notwendigkeit betrachtet, weil man bedingt durch die Altersverteilung bei den Mitarbeitenden mit wegfallenden Fachkräften rechnen muss, die voraussichtlich auf dem Arbeitsmarkt «gar nicht mehr wirklich zu ersetzen sind» {7.061}.

Erweiterung der Datenbasis

Eine Gemeinsamkeit zahlreicher Einsatzgebiete ist ein erweiterter Blick, so dass «…man auf Grund von […] indirekt erhobenen Daten eine Information bekommt oder daraus interpretiert» {6.214}. Denn «man kann ja vieles messen heute und man sieht was man bewirkt» {7.110}.

Mit dieser Erweiterung des Blicks kommt es (in gewissem Sinn paradoxerweise) dazu, dass «die zunehmende Digitalisierung […] auch zu einer Änderung in der Sichtbarkeit [der Arbeit der Fachkräfte führt]» {2.036}, weil sich die Fachkräfte weniger im Feld befinden.

Sicherheitsgewinne

Dies bringt bei bestimmten Konstellationen ein Zugewinn an Sicherheit für die Fachleute, was je nach dem einen wichtigen oder sogar einen sehr entscheidenden Vorteil darstellen kann {6.058}.

Schnellere Prozesse

Hinreichend für die Rechtfertigung von Mensch-Maschine-Systemen ist in mehreren Beispielen auch die erzielte Zeitersparnis, um zu Ergebnissen zu kommen {2.022; 5.012; 7.012}. Möglichst rasch erzielte Resultate können Erfolgsquoten grundsätzlich erhöhen, was eine sehr hohe Bedeutung haben kann. Oder es kann die Kundenzufriedenheit steigern, oder auch zur Effizienz insgesamt beitragen: «Jetzt geht es in einem Schritt, der Kunde ist happy» {3.063}.

Gezieltere Allokation von Arbeit auf Mitarbeitende

Aus Sicht der Organisation können die Mensch-Maschine-Systeme dazu führen, dass die Arbeitsbelastung gleichmässiger auf die einzelnen Mitarbeitenden aufgeteilt wird {3.033}. Das ist gerade dort, wo der «Betrieb […] sehr straff organisiert» {6.533} ist, als Vorteil genannt, «die zeitliche Planung der Aufgaben würde auch eine genauere Personalplanung ermöglichen» {2.002}.
Diese Vorteile müssen nicht nur zu Gunsten der Effizienz ausgegeben werden, sondern können auch zur Mitarbeiterbindung eingesetzt werden. «Die Mitarbeitenden haben damit ein breites Arbeitsgebiet, und für die Mitarbeitenden bedeutet das ein Job Enrichment» {3.065}. Oder Mitarbeitende mit unterschiedlichen Kompetenzen können gezielter ihren Fähigkeiten entsprechend eingesetzt werden. «Organisatorisch sieht das so aus, dass die einfachen [Fälle] durch die weniger gut ausgebildeten Mitarbeitenden veranlagt werden» {4.023}.

HERAUSFORDERUNGEN FÜR DIE ORGANISATION

Die Mensch-Maschine-Systeme bringen jedoch nicht nur Vorteile, sondern auch bestimmte neue Herausforderungen, die in der Organisation zu bewältigen sind. So entstehen beispielsweise neue fachliche Abhängigkeiten und damit auch Fragen darüber, wer wem zu unterstellen ist, oder welche Funktionen mit welchen Löhnen fair entschädigt sind. «[IT-Spezialisten] bekommen oftmals höhere Löhne und das führt dann punktuell zu einer Diskussion über Gerechtigkeit und Ungerechtigkeit» {2.034}.
Vereinzelt wird bestätigt, dass durch Automatisierungen für den Menschen Lernmöglichkeiten weniger werden. «Da die ganz einfachen Fälle nun automatisiert werden, bleiben weniger davon für die Lernenden übrig als Übungsfeld» {3.072}.
Ein Zuviel an Effizienz kann in gewissen Fällen auch menschlichen Kontakt und damit verbundene Beziehungsqualitäten vermindern. So sind beispielsweise automatisierte Prozesse an sich «auch Aufhänger für weitere Kundengespräche» {3.078}.
Als grosse technische Herausforderung für Fortschritte wird oft die «fehlende bzw. mangelnde Integration» {3.107} verschiedener Informationssysteme genannt. «Wir scheitern immer und immer wieder an der Vernetzung der Tools oder der Programme gegenseitig» {3.108}.

REGIONALE UND GLOBALE RAHMENBEDINGUNGEN

Bei den Rahmenbedingungen werden in mehreren Interviews die Verfügungsrechte erwähnt, etwa wie Algorithmen im Eigentum des Unternehmens bleiben können {5.059}.
Was oben als Vorteil der Informationssysteme beschrieben wurde, verschiebt sich teilweise zu den notwendigen Bedingungen zur Erfüllung der gestellten Aufgaben, etwa wenn Sicherheitsvorschriften es überhaupt nicht mehr erlauben, Mitarbeitende ins Feld und damit in gefährliche Situationen zu bringen {6.056}. Häufig sind fehlende Rechtsgrundlagen aber eher ein Hemmnis zur Ausschöpfung des vollen Potenzials von IT-Systemen, insbesondere Datenschutzbestimmungen limitieren den Austausch und damit die Möglichkeiten {2.015}. «Technisch wäre da noch vieles machbar, was aber datenschutztechnisch nicht zulässig ist» {4.010}.

FAZIT UND AUSBLICK

Insgesamt zeigt sich, dass die Auswirkungen von KI-Anwendungen auf die beruflichen Tätigkeiten vielfältig zu sein scheinen. Es sollen hier nochmals die kritischen Auswirkungen erwähnt werden.
So wird die Anforderung einer gesteigerten Wahrnehmungsfähigkeit bei KI-Anwendungen von den Mitarbeitenden erwähnt. Hier stellt sich die Frage, inwiefern diese Erhöhung bzw. Sensibilisierung der Wahrnehmungsfähigkeit erlernt werden kann. In der Hochschulbildung wird das Thema KI als bedeutsam und aktuell eingestuft, aber es kann dennoch sein, dass es den Weg in Nicht-Informatikstudiengängen nicht findet, weil die Verantwortlichen das Thema zu wenig kennen. Mittlerweile könnte ein Mangel an KI-Fachkräften ebenfalls dazu kommen. Erwähnt wurde ebenfalls, dass man den KI-Anwendungen im Fehlerfall keine Schuld geben kann, der Mensch ist schlussendlich verantwortlich, aber dennoch bleiben Unsicherheiten wie mithilfe von Qualitätssicherungsmassnahmen die notwendige Fehlerfreiheit erreicht werden kann und wie man im Fehlerfall damit umgeht. Mit der Einführung von KI-Anwendungen werden frühere fehlerhafte Arbeiten von Mitarbeitenden sichtbar. Das ist unangenehm für die Mitarbeitenden, und für die Unternehmen stellt sich die Frage, wie sie damit nachträglich umgehen wollen bzw. auch müssen, sowohl auf personeller als auch fachlicher Ebene. Das sind drei ausgewählte kritische Auswirkungen von KI-Anwendungen, die einerseits den Forschungsbedarf aufzeigen, in diesem Fall die Erhöhung bzw. Sensibilisierung der Wahrnehmungsfähigkeit im Rahmen von KI-

Anwendungen und andererseits den Handlungsbedarf im Sinne einer schnellen Integration von grundsätzlich neuen Themen, in diesem Fall KI, in die Curricula von Nicht-Informatikstudiengängen.

In der untenstehenden Tabelle 3 werden die Auswirkungen von Anwendungen mit Künstlicher Intelligenz, die in diesem Forschungsprojekt herausgefunden wurden, zusammenfassend dargestellt und dienen auch als übersichtliche, zusammenfassende Beantwortung der zentralen Fragestellung in diesem Forschungsprojekt «*Wie wirkt sich die Einführung von neuen Informationssystemen, insbesondere Anwendungen der Künstlichen Intelligenz, auf das Anforderungsprofil und das Learning-on-the-Job von Arbeitenden aus, und welche Auswirkung hat dies auf die Arbeitsproduktivität und auf die arbeitende Person selbst?*»

Tabelle 3: Asuwirkungen von Künstlicher Intelligenz (eigene Darstellung)

Auswirkungen auf Anforderungsprofil
Komplexere Aufgaben, keine Routinetätigkeiten müssen erledigt werden
Beziehungsarbeit kann (teilweise) wegfallen durch Kundenarbeit (Datenerfassung)
Training und Pflege der KI, Ergebniskontrolle
Auswirkungen auf Learning-on-the-job
Einfach Tätigkeiten für Lernende sind nicht mehr möglich/verfügbar
Fehlende Ausbildungsinhalte müssen im Job erlernt werden
Fehlender Einsatz von KI führt zur Kündigung (Enttäuschung)
Auswirkungen auf Produktivität der Arbeit
Verarbeitung von grösseren Datenmengen
Zeitersparnis
Fehlerquote wird reduziert, Sicherheit wird erhöht
Auswirkungen auf Mitarbeitende
Anforderungen/Erwartungen an Fachkompetenz und Technikaffinitität steigen
Abnehmende Kommunikation mit der Kundschaft
Leistungen der Arbeitenden werden transparent

LITERATUR

an Tang, Tam, R., Cadrin-Chênevert, A., Guest, W., Chong, J., Barfett, J. et al. (2018). Canadian Association of Radiologists White Paper on Artificial Intelligence in Radiology. *Canadian Association of Radiologists journal = Journal l'Association canadienne des radiologistes*, 69 (2), 120-135.

Arntz, M. (2020). Das Ende der Arbeit? Zwischen Potenzial und Rentabilität. *Maschine & Mensch* (16), 126-135.

Balser, M. (2022, 13. Februar). KI soll die Bahn pünktlicher machen. *Süddeutsche Zeitung.* https://www.sueddeutsche.de/wirtschaft/bahn-ki-verspaetung-1.5527651 (11.08.2022)

Barton, T. & Müller, C. (Hrsg.) (2021). *Künstliche Intelligenz in der Anwendung. Rechtliche Aspekte, Anwendungspotenziale und Einsatzszenarien* (Angewandte Wirtschaftsinformatik). Wiesbaden: Springer Fachmedien.

Bizer, J. (2019). Bestandsaufnahme und Perspektiven der Digitalisierung im Steuerrechtsverhältnis aus Sicht der Verwaltung. In Deutsche Steuerjuristische Gesellschaft (Hrsg.), *Digitalisierung im Steuerrecht* (S. 135-144). Köln: Verlag Dr. Otto Schmidt.

Busch, S. (2019). Versicherungsunternehmen im kognitiven Zeitalter. In M. Reich & C. Zerres (Hrsg.), *Handbuch Versicherungsmarketing* (2. Aufl., S. 243-260). Berlin, Heidelberg: Springer.

Castellino, R. A. (2005). Computer aided detection (CAD): an overview. *Cancer Imaging, 5* (1), 17-19.

Deckert, R. & Meyer, E. (2020). *Digitalisierung und Künstliche Intelligenz. Kooperation von Menschen und Maschinen aktiv gestalten (essentials).* Wiesbaden: Springer Fachmedien. https://link.springer.com/content/pdf/10.1007/978-3-658-31795-9.pdf

Deutsche Bahn (Hrsg.). (o.D.). *Künstliche Intelligenz bei der DB.* https://www.deutschebahn.com/de/kuenstlicheintelligenz-6898594

European Commission. (2020). *The journey begins – 2021 is the European Year of Rail! Press release*. Brussels. https://ec.europa.eu/commission/presscorner/detail/en/ip_20_2528 (28.07.2021)

Fichter, A. (2017). *Künstliche Intelligenz im Kundendienst. Braucht jedes Unternehmen einen Chatbot?,* Swisscom. https://www.swisscom.ch/de/business/enterprise/themen/digital-business/des-2017-008-servicesmit-kuenstlicher-intelligenz.html (03.08.2021)

Frost, M., Guhlemann, K., Cordes, A., Zittlau, K. & Hasselmann, O. (2020). Produktive, sichere und gesunde Arbeitsgestaltung mit digitalen Technologien und Künstlicher Intelligenz – Hintergrundwissen und Gestaltungsempfehlungen. *Zeitschrift für Arbeitswissenschaft*, 74 (2), 76-88. https://link.springer.com/content/pdf/10.1007/s41449-020-00200-3.pdf

Giering, O. (2022). Künstliche Intelligenz und Arbeit: Betrachtungen zwischen Prognose und betrieblicher Realität. *Zeitschrift für Arbeitswissenschaft, 76* (1), 50-64. https://link.springer.com/content/pdf/10.1007/s41449-021-00289-0.pdf

Gödeke, J., Muensterer, O. & Rohleder, S. (2020). Künstliche Intelligenz in der Kinderchirurgie: Gegenwart und Zukunft. *Der Chirurg; Zeitschrift für alle Gebiete der operativen Medizin,* 91 (3), 222-228. https://link.springer.com/content/pdf/10.1007/s00104-019-01051-3.pdf

Gruhn, V. (2018). Versicherungen: Von Natur aus für Künstliche Intelligenz geeignet. *Wirtschaftsinformatik & Management*, 10 (4), 104-111.

Hahn, L. & Zwiesler, J. (2018). Wie können Versicherer ihre Daten intelligent nutzen?, Institut für Finanz- und Aktuarwissenschaften. https://www.ifa-ulm.de/fileadmin/user_upload/download/sonstiges/2018_ifa_Hahn-Zwiesler_Wie-koennen-Versicherer-ihre-Daten-intelligent-nutzen.pdf (02.08.2021)

Hämmerli, M. (2019). Thurgauer Steueramt überlegt, mit Algorithmen nach Steuerhinterziehern zu suchen. St. Galler Tagblatt. https://www.tagblatt.ch/ostschweiz/thurgauer-steueramt-ueberlegt-mit-algorithmen-nach-steuerhinterziehern-zu-suchen-ld.1100813 (11.08.2022)

Haubold, J. (2020). Künstliche Intelligenz in der Radiologie: Was ist in den nächsten Jahren zu erwarten? *Der Radiologe*, 60 (1), 64-69.

Jungmann, F., Kuhn, S. & Kämpgen, B. (2018). Grundlagen und Einsatzmöglichkeiten von Natural Language Processing (NLP) in der Radiologie. *Der Radiologe*, 58 (8), 764-768.

Krupinski, E. A. (2003). The Future of Image Perception in Radiology. *Academic Radiology*, 10 (1), 1-3.

Kleinz, T. (2018). Das automatische Finanzamt. *Algorithmenethik*. https://algorithmenethik.de/2018/03/27/das-automatische-finanzamt/ (27.07.2021)

Mangei, T. (2019). Entwicklungstendenzen und Herausforderungen in der Versicherungswirtschaft. In M. Reich & C. Zerres (Hrsg.), *Handbuch Versicherungsmarketing* (2. Aufl., S. 139-151). Berlin, Heidelberg: Springer.

Maris, S. (2022) Wie Künstliche Intelligenz einzelne Aspekte und Werkzeuge kreativer Prozesse verändern könnte. In D.-K. Mah & C. Torner (Hrsg.), *Künstliche Intelligenz mit offenen Lernangeboten an Hochschulen lehren. Erfahrungen und Erkenntnisse aus dem Fellowship-Programm des KI-Campus* (S. 50-56).

Miotto, R., Li, L., Kidd, B. A. & Dudley, J. T. (2016). Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Scientific reports, 6,* 26094.

Miskolczi, A. & Thingna, Z. (2022). Die neue Gretchenfrage der Anwälte: Künstliche Intelligenz in der Rechtspraxis. In C. Schieblon (Hrsg.), *Digitalisierung und Innovation in Kanzleien* (S. 113-128). Wiesbaden: Springer Gabler.

Moshammer, T. (2020) Smart Bogie 4.0 – Fahrwerksdiagnose im Zeitalter des Internets der Dinge (IoT) und Artificial Intelligence (AI). In *IRSA 2019: Tagungsband, Proceedings: 2nd International Railway Symposium Aachen, Aachen, Germany, 26-28 November 2019* (S. 487-502). Aachen: RWTH. https://publications.rwth-aachen.de/record/775840/files/775840.pdf

Noser Engineering AG. (o.J.). abilio, die App für‘s Unterwegs. https://www.noser.com/showcase/abilio/ (29.07.2021)

Nufer, M. & Ackermann, R. (2021). *Digitale Steuerverwaltung (Obwalden). Präsentation - Schweizerische Vereinigung diplomierter Steuerexperten*. https://www.svds.ch/files/2187/nufer-digitalisierung.pdf (11.08.2022)

OECD (Hrsg.) (2021). *Künstliche Intelligenz in der Gesellschaft.* Paris: OECD Publishing. https://www.oecd-ilibrary.org/sites/4a748a0a-de/index.html?itemId=/content/component/4a748a0a-de (24.07.2021)

Rädiker, S. & Kuckartz, U. (2019). *Analyse qualitativer Daten mit MAXQDA. Text, Audio und Video* (Lehrbuch). Wiesbaden: Springer VS. https://link.springer.com/content/pdf/10.1007/978-3-658-22095-2.pdf

Rainsberger, L. (2021). Vertriebstechnologie: Ein Ozean an Möglichkeiten. In L. Rainsberger (Hrsg.), *Digitale Transformation im Vertrieb* (Edition Sales Excellence, S. 207-300). Wiesbaden: Springer Fachmedien.

Reich, M. & Braasch, T. (2019). Die Revolution der Prozessautomatisierung bei Versicherungsunternehmen: Robotic Process Automation (RPA). In M. Reich & C. Zerres (Hrsg.), *Handbuch Versicherungsmarketing* (2. Aufl., S. 291-306). Berlin, Heidelberg: Springer.

Schacker, M. & Fuchs, A. (2018). Chatbots im Kundenservice: Ein Verfahren zur Kosten-Nutzen-Analyse. *Wirtschaftsinformatik & Management*, 10 (6), 8-17. https://link.springer.com/content/pdf/10.1007/s35764-018-0114-x.pdf

Schmidt, J.-P. (2020) Chancen und Herausforderungen für die Versicherungswirtschaft durch Künstliche Intelligenz. In H. Müller-Peters, J.-P. Schmidt & M. Völler (Hrsg.), *Revolutionieren Big Data und KI die Versicherungswirtschaft? 24.Kölner Versicherungssymposium am 14. November 2019 (15-17)*. Köln. Institut für Versicherungswesen.

Siemens AG. (2016). Daten treiben Züge an. Interview mit Gerhard Kress, Leiter Siemens Mobility, Data Services Center. *Zukunft Deutschland,* 5. https://www.inpactmedia.com/wirtschaft/zukunft-deutschland/daten-treiben-zuege (29.07.2021)

Simmler, M., Brunner, S. & Schedler, K. (2021). *Smart Criminal Justice. – Eine empirische Studie zum Einsatz von Algorithmen in der Schweizer Polizeiarbeit und Strafrechtspflege* (Studienbericht der Universität St. Gallen): Helbing Lichtenhahn Verlag.

Staatskanzlei Kanton Zürich. (2021). *Einsatz Künstlicher Intelligenz in der Verwaltung: rechtliche und ethische Fragen.* Schlussbericht vom 28. Februar 2021 zum Vorprojekt IP6.4, Zürich. https://www.zh.ch/content/dam/zhweb/bilder-dokumente/themen/politik-staat/kanton/digitale-verwaltung-und-e-government/projekte_digitale_transformation/ki_einsatz_in_der_verwaltung_2021.pdf (11.08.2022)

Stephanie. (2021). Die besten Chatbot Use Cases aus verschiedenen Branchen, Omlin. https://onlim.com/die-besten-chatbot-use-cases/ (03.08.2021)

Stucki, T., D'Onofrio, S. & Portmann, E. (2020). *Chatbots gestalten mit Praxisbeispielen der Schweizerischen Post. HMD Best Paper Award 2018* (Essentials Series). Wiesbaden: Springer Fachmedien. https://link.springer.com/content/pdf/10.1007/978-3-658-28586-9.pdf

Systematische Rechtssammlung SRL - Kanton Luzern. (2011). Zentralschweizer Fachhochschul-Vereinbarung - Nr. 520. https://srl.lu.ch/app/de/texts_of_law/520/versions/1454 (12.08.2022)

Wittpahl, V. (2019). Künstliche Intelligenz. Berlin, Heidelberg: Springer.

Zabel, T. (2020) Altes im neuen Gewand?! – KI und Data Analytics im Versicherungssektor. In H. Müller-Peters, J.-P. Schmidt & M. Völler (Hrsg.), *Revolutionieren Big Data und KI die Versicherungswirtschaft? 24.Kölner Versicherungssymposium am 14. November 2019 (46-53).* Köln. Institut für Versicherungswesen.

Zeier Röschmann, A. & Erny, M. (2019). Transformation zum digitalen Versicherungsbroker – das Beispiel Optimatis. In A. Uhl & S. Loretan (Hrsg.), *Digitalisierung in der Praxis* (S. 171-181). Wiesbaden: Springer Fachmedien.

KONTAKT

CHRISTOPH HAUSER ist Ökonom und Dozent für Regional- und Institutionenökonomie an der Hochschule Luzern – Wirtschaft. Dort leitet er das Kompetenzzentrum Management & Law und ist regelmässig in Projekten zur Digitalisierung und Innovationspolitik tätig. Seine E-Mail-Adresse ist: christoph.hauser@hslu.ch und sein Personenprofil ist hier: https://www.hslu.ch/de-ch/hochschule-luzern/ueber-uns/personensuche/profile/?pid=222

UTE KLOTZ ist Dozentin für Informations- und Innovationsmanagement. Sie ist Fokusgruppenleitende für „Technologien für die digitalisierte Arbeitswelt der Zukunft“ beim interdisziplinären Themencluster „Digitale Transformation der Arbeitswelt“. Ihre Forschungsinteressen liegen beim Thema „Zukunft der Arbeit“ und „Technologien im Alltag“. Ihre E-Mail-Adresse ist: ute.klotz@hslu.ch und ihr Personenprofil ist hier: https://www.hslu.ch/de-ch/hochschule-luzern/ueber-uns/personensuche/profile/?pid=228

AI-based Product Quality Controlling in an Anodizing Process

Joshua Prim
Cognitive Information Systems, KITE - Kompetenzzentrum für Informationstechnologie
Technische Hochschule Mittelhessen
Wiesenstr. 14
35390 Gießen
joshua.prim@mnd.thm.de

Henrik Pöschl
Seidel GmbH & Co. KG
Rosenstraße 8
35037 Marburg
henrik.poeschl@seidel.de

Michael Guckert
Cognitive Information Systems, KITE - Kompetenzzentrum für Informationstechnologie
Technische Hochschule Mittelhessen
Wiesenstr. 14
35390 Gießen
michael.guckert@mnd.thm.de

Keywords
Artificial Intelligence, Machine Learning, Quality Assurance, Automation Process Monitoring

ABSTRACT
Product quality is a crucial factor of customer satisfaction and thus directly influences the competitiveness of a company. In manufacturing companies the quality of production processes obviously has significant impact on product quality. Therefore, establishing automated quality control offers considerable leverage for improving processes without necessarily increasing work efforts and costs. In this paper an artificial intelligence based pattern recognition method for increasing the output of an anodising process for aluminium parts is discussed. In the use case presented here, customers have high aesthetic requirements regarding the products which are used in an expensive market segment with only limited fault tolerance. Preparation of the product parts before going through the anodising process is a manual, tedious, and error-prone task that nevertheless requires highest precision. Small deviations can lead to quality problems causing rejections and enforcing repetitions in production. We discuss the application of visual image processing with an artificial intelligence algorithm integrated into the information system of the company to monitor the process and prevent human errors. Results show that our approach reaches high accuracy and can potentially improve delivery reliability with respect to time and quantity by reducing cost-intensive manufacturing errors.

1. PROBLEM DESCRIPTION
Efficiency of processes and quality of products play a decisive role in running economically successful industrial production. Therefore, product quality assurance (QA) is becoming an increasingly important element of the value chain. Seidel GmbH & Co. KG in Marburg is a leading producer of packaging material for cosmetic brands and components for aluminium products. Aluminium foliage is formed coldly in progressive a press and then receives surface treatment in an anodising process.

Before entering the anodising process aluminium parts are placed on racks. Each rack has a capacity of up to 300 parts, depending on the product and the process configuration. Racks are put on carriers which are then moved automatically into the anodising plant. The electrochemical anodising process covers the raw aluminium parts with a layer of aluminium oxide. Colours and varying product appearances can be created.

Furthermore, the aluminium part receives a calloused, scratch-resistant high-quality surface. This complicated process requires product carriers to be loaded strictly according to defined patterns. Due to their size and shape, different aluminium parts require a careful setup of this process.

Depending on the item in production in a given batch, different plug patterns are used to load the product carriers. A plug pattern defines how racks loaded with products have to be placed on the carrier. A product carrier can hold 28 racks (14 on each side of the product carrier), resulting in several million possible combinations for the arrangement of racks on the product carrier. However, only a few plug patterns are feasible in the anodising process and have proven to lead to good results. Fig. 1 shows an example of such a product carrier.

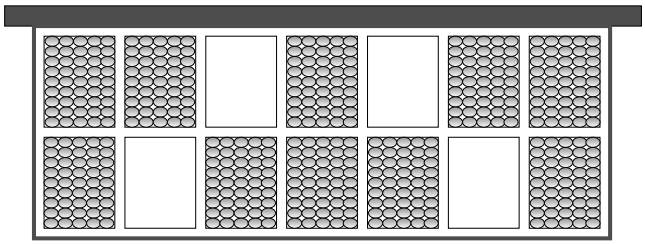


Figure 1: Schematic representation of a loaded product carrier. The round balls symbolise products on a rack which, due to their arrangement in the slots represented by the rectangles, result in the plug pattern on the rack.

Loading product carriers is a manual task in which time pressure and the complexity of the task lead to errors. A typical error is misplacing racks so that the loaded product carrier does not match the pattern required for that particular product. Such a pattern mismatch in the feed of the anodising plant leads to low output quality, so the complete batch has to be rejected. In this case, this batch has to be

reproduced incurring additional costs and finally shipping delays.

Therefore, the use of artificial intelligence methods to monitor the loading process of the anodising plant was analysed. Image recognition with a Deep Learning algorithm was implemented to interfere and stop the process in case of loading errors. An independent functional unit (in the following called QA-Engine) was developed, which compares patterns detected in images of the product carrier to patterns that are defined in the master data of the enterprise resource planning system (ERP system). In this paper, we describe the proof of concept for the use of the QA-Engine. We first discuss related work and then present our methodology. The results achieved are then presented. The paper concludes with a look at future work in general and the specific use case presented here.

2. RELATED WORK

In recent years, machine learning methods have been used in many different environments and their fields of application have already been extensively illuminated. Quality control, especially process control, has also been investigated to a certain extent.

One area that has been considered is recognition of patterns such as shifts, trends and cycles in quality control charts with the goal to minimise deviations by identifying the cause and adjusting the process accordingly. Guh and Shiue [Guh and Shiue, 2005] showed that decision trees can be used to identify patterns in control charts. Wang et al. [Wang et al., 2018] also managed to use decision trees to identify anomalies. Gauri and Chakraborty [Gauri and Chakraborty, 2007] successfully trained a neural network to identify features in control charts.

Other studies aimed at determining whether machine learning can be used to classify processes with or without control. Smith [Smith, 1994] used a neural network for X bar R charts to determine whether a process is out of control. For detecting significant shifts in mean values, the researchers were able to achieve equally good results with neural networks as with regular control limits. For small shifts, however, neural networks exceeded conventional control limits. Shao and Chiu [Shao and Chiu, 1999] trained a neural network to identify different assignable causes in an attempt to integrate statistical process control with feedback control for a set of parameters.

Pacella and Semeraro [Pacella and Semeraro, 2007] point out the problem that many quality characteristics correlate and use a neural network to monitor the quality of autocorrelated process data. Low et al. [Low et al., 2003] also consider autocorrelated data, but focus specifically on detecting variations of variance.

Machine learning has also been used to identify anomalies in surface textures ([Weimer et al., 2016]; [Wang et al., 2018]). Methods of image recognition to identify anomalies on textile textures were also investigated ([Ngan et al., 2011]; [Sajid, 2012]).

Plastic injection moulding was also investigated, using various parameters from production as input ([Ribeiro, 2005]; [Tellaeche and Arana, 2013]). Zhao et al. [Zhao et al., 2017] describe an automatic image recognition approach for quality assurance for cold rolling processes. Villalba-Diez et al. [Villalba-Diez et al., 2019] have applied machine learning in the printing industry. The researchers show how a neural network can be combined with a high-resolution optical quality control camera to increase product quality and reduce costs in the printing industry. Ferguson et al. [Ferguson et al., 2018] on the other hand, demonstrate the use of neural networks to identify casting defects in X-ray images.

The fast-food chain Domino's Pizza is also using machine learning to monitor the quality of its pizzas. In 2019, the company introduced a scanning technology in its kitchens in Australia that uses machine learning to analyse images of pizzas. According to Domino's, it has succeeded in increasing the quality of the pizzas monitored in this way by fifteen per cent ([Dominos, 2019]). The US company eBay Inc. also relies on machine learning for quality assurance. eBay uses a neural network to classify whether a UX component meets the desired quality criteria or not (see [Sharan et al., 2018]).

3. METHODOLOGY

For error detection, we use a convolutional neural network (CNN) architecture. In this section we will at first provide some theoretical background about CNNs in 3.1 and then briefly describe our our database structure in 3.2.

3.1 Convolutional neural networks

CNNs ([LeCun et al., 1989]) represent a special variant of artificial neural networks (ANN). Due to their structure, they are particularly well suited for image recognition and are preferably used for the classification of images and videos ([Schwaiger and Steinwendner, 2019]). The CNN model was inspired by the mechanisms of the visual cortex of the brain. A significant difference to conventional ANNs is that CNNs apply filters and create feature maps to detect patterns and structures in images. These can be contours, colours or textures, which are then combined into more complex structures.

Architecturally, the structure of CNNs shows specific differences compared to conventional ANNs. The input layer usually takes three-dimensional input in the form of the spatial extension of the image (width * height) and has a depth representing the colour channels (usually three for the RGB colour channels). This is followed by convolution layers and pooling layers, explained below. These two types of layers can be repeated with tailored parameterisation. Typically, a classification layer is used as output layer, represented by a fully linked layer for generating the scores. A high-level general CNN architecture is shown in Fig. 2.

- Input: The CNN input is usually the 3-channel colour image or 1-channel grey image matrices, containing the intensity values at each position.
- Conv: The core of a CNN is the convolutional layer. This layer performs the mathematical operation called convolution. Convolution is a special kind of linear operation. CNNs are thus ordinary neural networks that use convolution instead of general matrix multiplica-

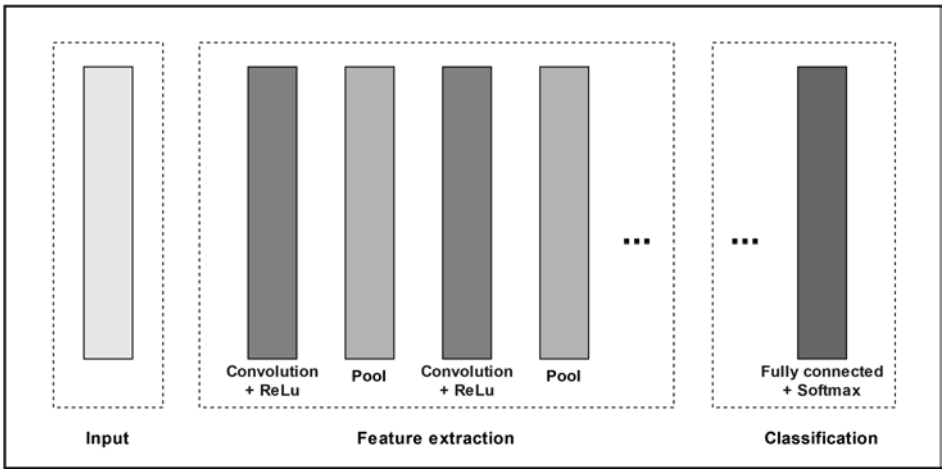


Figure 2: General CNN architecture

tion in at least one of their layers ([Goodfellow et al., 2016]).
During convolution, the size of the filter (kernel size) (e.g. 3 x 3) is defined first. Then the filter scans the pixel matrix of the input like a window with a constant step size. The filters move from left to right across the input matrix and jump to the next lower row after each pass. The so-called padding determines how the filter should behave when it hits the edge of the matrix. The filter has a fixed weight for each point in its viewport, and it calculates a result matrix from the pixel values in the current viewport and these weights. The result of the convolution is called a feature map in the terminology of CNNs. The size of this resulting matrix depends on the kernel size of the filter, the padding and especially on the step size. A non-linear activation function is used for the feature map. For modern CNNs, the default activation function is the rectified linear activation function (ReLu).

- Pooling: The pooling layer leads to a reduced spatial dimension by using the pooling function according to Zhou and Chellappa [Zhou and Chellappa, 1988]. It aims to reduce the amount of network parameters and the calculation costs. The pooling layer is often placed between two successive convolution layers. Different aggregation functions can be used for pooling. The most common aggregation functions are max pooling and average pooling ([Lin et al., 2013]).
- Dense: The Fully Connected Layer or Dense Layer is a standard neural network structure in which all neurons are connected to all inputs and all outputs. Moreover, the last fully connected layer produces the output of the entire net. Each value of the k-dimensional output represents the probability of the corresponding label using the softmax function.

Combined, these layers provide a complete CNN into which input can now be fed for network decision-making.

3.2 Dataset

Images taken in the anodising plant serve as the data basis for the QA-Engine. For this reason, a system for image acquisition was integrated into the anodising process, which records the pattern set by the employees. The system uses an RPI3-CM01 camera to take images of the product carrier with the racks at the start of the anodising process. Such an image is shown in Fig. 3.

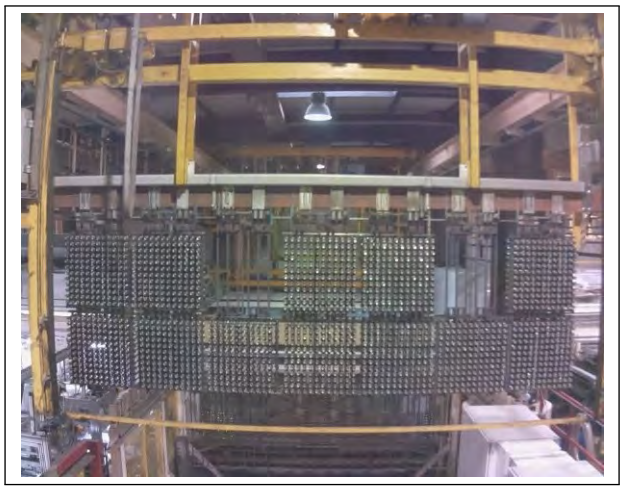


Figure 3: Example of an image taken in the anodising plant

After the product carrier has been completely loaded with the racks by the machine operators, it is automatically moved through the anodising system. While the product carrier is moving through the anodising plant to the preprocessing, an image is taken automatically triggered by an ultrasonic sensor of the type HC-SR04. Thus, images are always taken at the same point and all have identical distance from the camera. Additional information of the current order is required to evaluate the pattern. For each new order, information is captured by the ERP system and stored together with the information which plug pattern is used for the order. The information describing the orders is delivered via an API from the ERP system. In preparation for our experiments more than 20000 images were reviewed with low-quality images eliminated. The remaining rest was labelled with a plug pattern resulting in a data set of 1000 labelled

images (elox_seidel_1k). When creating the data set, care was taken to ensure that the distribution of images per plug pattern was balanced. This data set is used for the training of the CNN and for all experiments carried out. The elox_seidel_1k dataset was then further processed to be suitable for the training of the CNN. The QA-Engine responsible for data preparation reads the data set and processes the data.

In these and further steps, an X-data set and a Y-data set are created, representing input and target values. The X-data set consists of an array of the form (1000, 256, 320, 3). The first value represents the number of image instances. The second and third values are the dimensions of the image. The fourth value represents the three colour channels of the image (RGB). The Y-data set corresponds to the form (1000, 10), whereby the first value also represents the number of image instances. The second value specifies the plug pattern class in an one-hot-encoding. Finally, the image instances are divided into training, validation and test data set. The data records can now be used to train and validate the CNN.

4. QA-ENGINE

In this section, the architecture of the QA-Engine is described, and the CNN implementation is discussed.

4.1 The QA-Engine in the anodising plant

The architecture of the QA-Engine was developed so that it can be combined with the existing components of Seidel GmbH & Co. KG and can easily be integrated into the system landscape of Seidel. See Fig. 4 for an overview of that landscape.

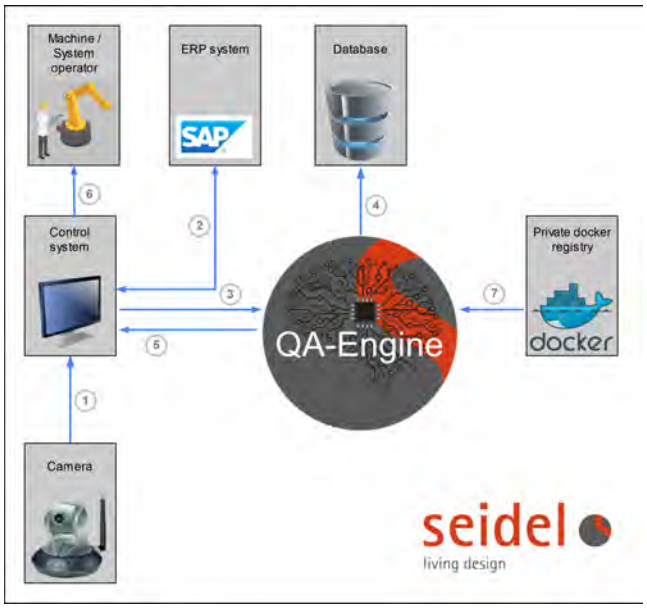


Figure 4: Integration of the QA-Engine into the Seidel environment

The architecture is designed to support the quality assurance process in the anodising plant in the best possible way. Initially an image of the product carrier with the racks is taken in the anodising plant. This image is then forwarded to a control system in the anodising process [1]. The plant control system asks the ERP system for the order currently in the anodising plant [2]. The recorded image, together with the current order from the ERP system, is sent by the plant control system to the QA-Engine [3]. The QA-Engine analyses the data and returns an assessment of the plug pattern to the plant control system [5]. It is also planned that the QA-Engine will store generated assessments in a database for potential further evaluation in the future. [4]. The results database created by the QA-Engine is forwarded by the system control if required. If the QA-Engine detects an error in the plug pattern, the result is forwarded to the system operator [6]. It is subjected to a human who can, if necessary, intervene in the anodising process. In terms of a fully automated production, the result of the QA-Engine can also be forwarded directly from the system control to the machine in the anodising plant. As a result, the machine interrupts the process and the plug pattern can be corrected. The fully implemented QA-Engine was provided as a docker image in a private Seidel registry [7].

Communication with the QA-Engine is done via a representational state transfer application programming interface. This enables interaction with the system using hypertext transfer protocol requests. The system control of the anodising plant sends a request to the QA-Engine and receives a response from the QA-Engine. Various endpoints have been implemented for communication with the QA-Engine API. Among other things, the images can be sent to the engine and hyperparameters can be requested and adjusted.

4.2 Model Implementation

The main component of the QA-Engine for quality assurance is a CNN. Based on the pictures taken, the CNN recognises the plug pattern depicted in the anodising plant. This way classification of the images and thus of the plug patterns is to be carried out. One class represents one plug pattern in the anodising process. The CNN was implemented in the programming language Python (version 3.6) (cf. [van Rossum and Drake Jr, 1995]). TensorFlow (Version 1.9) ([Abadi et al., 2016]) was used for this purpose. TensorFlow is an open-source program library for machine learning. TensorFlow combines the computational algebra of compilation optimisation techniques and thus facilitates the calculation of many mathematical expressions. Time-consuming calculations can thus be processed much faster ([Zaccone, 2016]). Keras ([Chollet, François, 2015]), an interface for TensorFlow, was also used for training and validation of the model.

Different architectures were implemented and assessed to determine the network architecture most suitable for the problem. Based on the results of the comparison, the architecture shown in Fig. 5 was selected. As shown in Fig. 5, the network consists of nine layers with weights; the first three are folded and the remaining six are for regulation, as well as fully connected. The output of the final fully connected layer is fed into a 10-way softmax, which creates a distribution over the ten class labels.

The first convolution layer filters the $256x320x3$ input image with 32 filters with a $5x5$ window and a stride of one pixel.

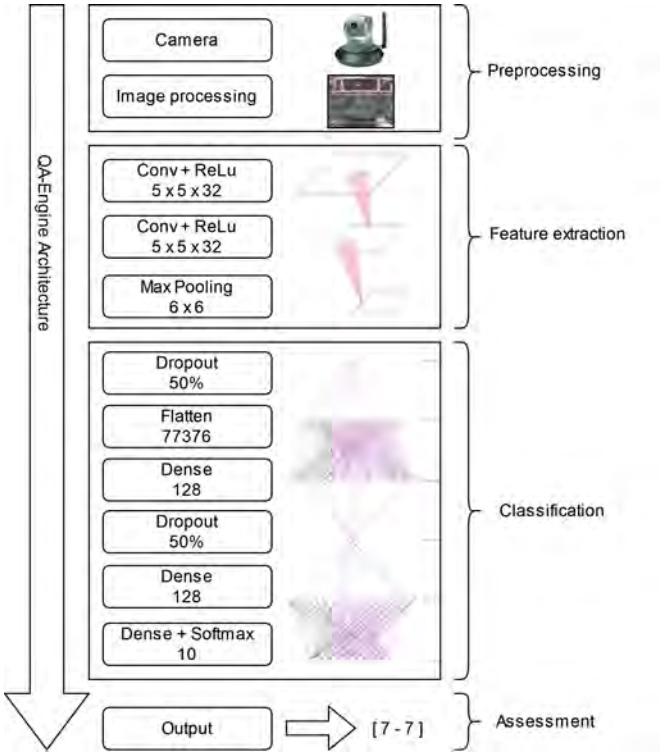


Figure 5: Visualisation of a schematic representation of the architecture of the QA-Engine.

The second convolutional layer takes the output of the first convolutional layer as input and also applies 32 filters with a $5x5$ window. The two layers are followed by a pooling layer with a window size of $6x6$. Padding is not used in either layer.

There is a flatten layer between the convolutional and the fully connected layer. Flattening converts the two-dimensional matrix of features into a vector for processing by the subsequent fully connected layers. The fully connected layers consist of 128 neurons. To prevent the CNN from overfitting, dropout was used. The softmax function underlying the final dense layer transforms the transferred values. The output of the softmax function corresponds to a categorical probability distribution and thus indicates a probability of belonging to a plug pattern.

4.3 Further experiments

Furthermore, the influence of contrast, brightness and sharpness on the CNN used was investigated. On the one hand, this makes it possible to determine whether an adjustment of contrast, brightness and sharpness during the preprocessing of the images possibly leads to better results. On the other hand, the robustness of the network for the factors will be investigated.

For this experiment, the preprocessing was adapted and the network was then trained and tested, using the previously described dataset. Contrast, brightness, and sharpness was manipulated for each of the 1000 images (800 training, 100 and 100 test). Subsequently, the CNN was trained. Nine training and test runs were performed for each factor, each with different value assignments. In addition, the vulnerability of the CNN of the QA-Engine was tested under poor conditions. For example, dust or other impurities may collect on the camera lens when used in the anodising plant. Also, lamps in the production hall may fail temporarily. These and other situations lead to different conditions when the image is captured. Therefore, it was investigated how well the mesh performs with changes in contrast, brightness and sharpness. Therefore, 100 images were randomly taken for this purpose and deliberately manipulated. They were then given to the CNN of the QA-Engine for assessment.

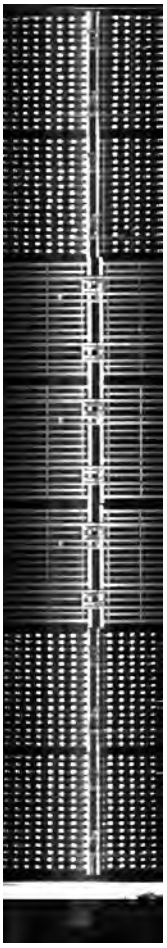


Figure 6: Example image taken with the line scan camera

The possibility of generating patterns dynamically at runtime was also investigated to react more flexibly to the introduction of new patterns. For this purpose, the recorded images were divided into individual quadrants, which were checked separately for a loaded or unloaded slot. A modified test setup using a line scan camera was used for this purpose. The line scan camera is triggered by a incremental encoder when the product carrier is moved into the anodising system. An example of a recorded image is shown in Fig. 6. Since this is an experimental setup, only the central area of the container was recorded. 14 quadrants were defined on the image, which were checked for an existing rack. At runtime, the plug pattern can be dynamically generated depending

on whether a rack with aluminium parts needs to be on the quadrant or not. A modified version of the CNN described above was trained to recognise and distinguish between existing and non-existing racks. The network has an accuracy of 98,99%, but the static definition of the quadrants proves to be error-prone in the case of shifted images. Since a fine adjustment of the encoder can solve this problem, this approach will be further investigated in the future.

5. RESULTS

The CNN we created was trained using the previously described data set. Our model was trained on a NVIDIA GeForce GTX 1080Ti 11 GB. 10-fold cross validation was used. The network has an accuracy of 98.09%. The results show that our model can recognise plug patterns and differentiate between plug patterns.

During the experiments, it was found that varying contrast, brightness, and sharpness in preprocessing did not positively affect the predictive ability of the CNN. Tendentially, an extreme adjustment of these variables led to worse results. Further investigations showed that the CNN of the QA-Engine is not very resistant to changes in brightness and contrast. Among other things, the predictive ability deteriorates significantly when brightness is lowered. Changes in image sharpness, on the other hand, are harmless.

6. CONCLUSION AND FUTURE WORK

In this work, we have described the prototypical development of a system for quality assurance. The system was developed for use in the anodising process of the company Seidel GmbH & Co. KG to monitor this process automatically. Thereby, images of product carriers containing racks with production goods arranged in a plug pattern are automatically recorded. These serve as input for the system. The current order from the ERP system is also transferred to the QA-Engine. The image is recognised and classified using a previously trained artificial neural network. For this purpose, the image with the product carrier is assigned to a pattern. The plug pattern classified by CNN is compared with the intended plug pattern from the ERP system. In case of an incorrect plug pattern, the system gives feedback. Errors occurring in the anodising can be detected early thus reducing quality assurance costs.

Currently a fully usable prototype is available which can be integrated into the infrastructure with container virtualisation. Further tests are required and camera technology will be consolidated.

The prototype created in this work has a potential high long term impact to significantly improve the anodising process. Additionally, the prototype is considered to have the potential to be transferred to other use cases, e.g. high precision determination of the loss of products. Individual aluminium parts tend to detach from the racks at odd times and remain in the anodising plant. The QA-Engine can be extended to count the aluminium parts on the individual racks. In this case, images would have to be taken at start and end of the anodising process which could then be used to compare counts giving the exact loss of products. However, this use case could not yet be implemented as camera technology with sufficient precision was not available. Later versions of the QA-Engine can implement a higher degree of automation by independently interrupting processes and initiating reloading of the carriers.

In conclusion, with implementing the QA-Engine for quality assurance Seidel GmbH & Co. KG has made a step towards fully automated and monitored production processes.

REFERENCES

[Abadi et al., 2016] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *https://arxiv.org/pdf/1603.04467*.

[Chollet, François, 2015] Chollet, François (2015). Keras. https://keras.io/, last updated 25.05.2022.

[Dominos, 2019] Dominos (2019). Dom pizza checker. https://dompizzachecker.dominos.com.au/, last updated 21.05.2022.

[Ferguson et al., 2018] Ferguson, M. K., Ronay, A., Lee, Y.-T. T., and Law, K. H. (2018). Detection and segmentation of manufacturing defects with convolutional neural networks and transfer learning. *Smart and sustainable manufacturing systems*, 2.

[Gauri and Chakraborty, 2007] Gauri, S. K. and Chakraborty, S. (2007). A study on the various features for effective control chart pattern recognition. *The International Journal of Advanced Manufacturing Technology*, 34(3):385–398.

[Goodfellow et al., 2016] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. Adaptive computation and machine learning. MIT Press, Cambridge, Massachusetts and London, England.

[Guh and Shiue, 2005] Guh, R.-S. and Shiue, Y.-R. (2005). On-line identification of control chart patterns using self-organizing approaches. *International Journal of Production Research*, 43(6):1225–1254.

[LeCun et al., 1989] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551.

[Lin et al., 2013] Lin, M., Chen, Q., and Yan, S. (2013). Network in network. *arXiv preprint arXiv:1312.4400*.

[Low et al., 2003] Low, C., Hsu, C.-M., and Yu, F.-J. (2003). Analysis of variations in a multi-variate process using neural networks. *The International Journal of Advanced Manufacturing Technology*, 22(11):911–921.

[Ngan et al., 2011] Ngan, H. Y., Pang, G. K., and Yung, N. H. (2011). Automated fabric defect detection—a review. *Image and Vision Computing*, 29(7):442–458.

[Pacella and Semeraro, 2007] Pacella, M. and Semeraro, Q. (2007). Using recurrent neural networks to detect changes in autocorrelated processes for quality monitoring. *Computers & Industrial Engineering*, 52(4):502–520.

[Ribeiro, 2005] Ribeiro, B. (2005). Support vector machines for quality monitoring in a plastic injection molding process. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 35(3):401–410.

[Sajid, 2012] Sajid, T. (2012). Fabric defect detection in textile images using gabor filter. *IOSR Journal of Electrical and Electronics Engineering*, 3(2):33–38.

[Schwaiger and Steinwendner, 2019] Schwaiger, R. and Steinwendner, J. (2019). *Neuronale Netze programmieren mit Python*. Rheinwerk Computing. Rheinwerk Computing, 1. auflage edition.

[Shao and Chiu, 1999] Shao, Y. E. and Chiu, C.-C. (1999). Developing identification techniques with the integrated use of spc/epc and neural networks. *Quality and Reliability Engineering International*, 15(4):287–294.

[Sharan et al., 2018] Sharan, Y., Wang, H., and Rath, S. (2018). Gui testing powered by deep learning.

[Smith, 1994] Smith, A. E. (1994). X-bar and r control chart interpretation using neural computing. *The International Journal of Production Research*, pages 309–320.

[Tellaeche and Arana, 2013] Tellaeche, A. and Arana, R. (2013). Machine learning algorithms for quality control in plastic molding industry. *2013 IEEE 18th Conference on Emerging Technologies & Factory Automation (ETFA)*, pages 1–4.

[van Rossum and Drake Jr, 1995] van Rossum, G. and Drake Jr, F. L. (1995). *Python tutorial*, volume 620. Centrum voor Wiskunde en Informatica Amsterdam.

[Villalba-Diez et al., 2019] Villalba-Diez, J., Schmidt, D., Gevers, R., Ordieres-Meré, J., Buchwitz, M., and Wellbrock, W. (2019). Deep learning for industrial computer vision quality control in the printing industry 4.0. *Sensors (Basel, Switzerland)*, 19(18).

[Wang et al., 2018] Wang, T., Chen, Y., Qiao, M., and 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.

[Weimer et al., 2016] Weimer, D., Scholz-Reiter, B., and Shpitalni, M. (2016). Design of deep convolutional neural network architectures for automated feature extraction in industrial inspection. *CIRP Annals - Manufacturing Technology*, 65(1):417–420.

[Zaccone, 2016] Zaccone, G. (2016). *Getting started with TensorFlow: Get up and running with the latest numerical computing library by Google and dive deeper into your data!* Community experience distilled. Packt Publishing Ltd, Birmingham, UK.

[Zhao et al., 2017] Zhao, Y. J., Yan, Y. H., and Song, K. C. (2017). Vision-based automatic detection of steel surface defects in the cold rolling process: considering the influence of industrial liquids and surface textures. *The International Journal of Advanced Manufacturing Technology*, 90(5):1665–1678.

[Zhou and Chellappa, 1988] Zhou, Y. and Chellappa, R. (1988). Computation of optical flow using a neural network. In *IEEE International Conference on Neural Networks 1993*, pages 71–78 vol.2, Piscataway. IEEE.

Acknowledgments

This research was partially supported by Seidel GmbH & Co. KG and by Hessen Agentur LOEWE 3 (project IPDU - HA project no. 593/18-16)

Contact

For more information contact:

Joshua Prim
Cognitive Information Systems, KITE - Kompetenzzentrum für Informationstechnologie
Technische Hochschule Mittelhessen
Wiesenstr. 14
35390 Gießen
joshua.prim@mnd.thm.de

SERVERLOSE DATENVERARBEITUNG IN DER CLOUD

Einschränkungen und Implikationen für eine IoT-Anwendung am Beispiel von AWS Lambda

Simon Rapp
HS Pforzheim
Tiefenbronnerstr. 65,
75175 Pforzheim
rappsimo@hs-pforzheim.de

Frank Morelli
HS Pforzheim
Tiefenbronnerstr. 65,
75175 Pforzheim
frank.morelli@hs-pforzheim.de

KEYWORDS

Cloud Computing, Serverlose Datenverarbeitung, Amazon Web Services (AWS), AWS Lambda Serviceverhalten und Einschränkungen, Internet of Things (IoT).

ABSTRACT

Ein populäres Alleinstellungsmerkmal der Cloud ist das Angebot der serverlosen Datenverarbeitung. Bei diesem bezahlen die Kunden nur für jene Ressourcen, die sie tatsächlich nutzen. Zudem erfolgt die Ressourcenallokation durch den Cloud-Anbieter dynamisch. Aus der Sicht des Kunden ist ein serverloser Cloud-Service somit automatisch hoch verfügbar und das Risiko einer Unter- oder Überprovisionierung wird minimiert.

Die Kehrseite dieses Ausführungsmodells ist jedoch, dass die Kunden nahezu alle Verwaltungsaufwendungen hinsichtlich der physischen und virtuellen Infrastruktur dem Cloud-Anbieter überlassen. Dies hat zur Folge, dass Kunden auf das Verhalten der von ihnen genutzten Services nur eine eingeschränkte Kontrolle ausüben können, was sich in konkreten Nachteilen äußern kann.

Der vorliegende Beitrag befasst sich mit den Gründen für diese Einschränkungen und zu welchen erkennbaren Nachteilen diese führen können. Im Mittelpunkt ist dabei der serverlose Cloud Service AWS Lambda, von dem das Serviceverhalten im Rahmen eines IoT-Use Cases genauer untersucht wird.

EINLEITUNG

Serverlose Datenverarbeitung, auch kurz als „serverlos“ bezeichnet, ist ein für die Cloud einzigartiges Ausführungsmodell. Hierbei führt der Cloud-Anbieter den Anwendungscode seiner Kunden durch dynamische Zuweisung von Rechenressourcen aus. Im Rahmen dieses Ansatzes sind nach wie vor Server, wie beispielsweise physische Server, virtuelle Maschinen oder Container, für die Ausführung von Code erforderlich. Im Vergleich zu serverbasierter Datenverarbeitung wird dabei jedoch eine weitere Abstraktionsschicht, oberhalb von der Cloud-Infrastruktur, hinzugefügt. Anwendungsentwickler müssen somit die zugrunde liegende Infrastruktur, die für die Ausführung von Code erforderlich ist, nicht selbst konfigurieren oder im laufenden Betrieb verwalten. Diese Aktivitäten werden durch den Cloud-Anbieter gemanagt, der in Abhängigkeit vom Bedarf die erforderliche Infrastruktur automatisiert bereitstellt und skaliert (Bahga und Madisetti 2019).

Ein weiteres Merkmal der serverlosen Datenverarbeitung ist das Abrechnungsmodell. So wird jede Ausführung separat berechnet, wobei die Dauer und die Menge der verwendeten Ressourcen den Preis determinieren. Dieser Ansatz ermöglicht eine Optimierung der Kosten: Bei der serverbasierten Datenverarbeitung müssen die zugrunde liegenden Ressourcen fix bereitgestellt werden, was die Gefahr einer Überprovisionierung begünstigt und damit zu Leerkosten bzw. ungenutzten Serverkapazitäten führen kann (Artasanchez 2021).

Das serverlose Konzept ist deshalb besonders für jene Anwendungstypen attraktiv, die stark schwankende oder zum Teil auch unvorhersehbare Datenströme verarbeiten sollen, wie zum Beispiel eine IoT-Anwendung. Hier ist es möglich alle zentralen Vorteile der serverlosen Implementierung zu nutzen, d.h. hohe Robustheit und Flexibilität des Systems bei begrenztem Verwaltungsaufwand und zugleich geringeren Kosten (Laszewski et al. 2018).

Mit der zunehmenden Popularität und Adaption der serverlosen Datenverarbeitung in der Cloud (Dremel und Herterich 2018), ist das Konzept allerdings auch kritisch zu würdigen. Ein wesentlicher Grund hierfür besteht darin, dass die Kunden eines serverlosen Cloud-Services die Kontrolle über die zugrunde liegende Infrastruktur nahezu vollständig dem Cloud-Anbieter überlassen. Im Einzelnen ergeben sich folgende Fragen aufgrund der eingeschränkten Kontrolle:

- Welche technischen Einschränkungen stehen den Vorteilen gegenüber?
- Wie lassen sich diese erkennen?
- Welche Auswirkungen haben sie auf das Verhalten einer Applikation?

Der vorliegende Beitrag beschäftigt sich mit diesen Fragestellungen am Beispiel des serverlosen Cloud-

Services „AWS Lambda“ von Amazon Web Services (AWS).

Zuerst werden hierzu die grundlegenden Servicefunktionen sowie bereits bekannte Einschränkungen von AWS Lambda charakterisiert. Danach erfolgt die Betrachtung eines industriellen Use Cases, am Beispiel einer IoT-Anwendung. In diesem Kontext wurde im Vorfeld eine vertiefende Analyse des Serviceverhaltens von AWS Lambda durchgeführt, deren Ergebnisse dieser Beitrag vorstellt und diskutiert. Abschließend wird die Thematik zusammengefasst und mit einem Ausblick gewürdigt.

AWS LAMBDA

AWS Lambda ist im Angebot von AWS ein häufig verwendeter Service für serverlose Implementierungen. Bei Lambda erfolgt die Strukturierung des Codes im Rahmen von Funktionen, die durch ein Ereignis eines anderen AWS-Service ausgelöst werden. Zugehörige Ereignisse sind zum Beispiel ein HTTP-Request an ein API Gateway, ein neu eingetragener Datensatz in einer Datenbank, eine neue hochgeladene Datei in einem Cloud-Speicher, eine neue Nachricht in einem Messaging-Queue, ein Überwachungsalarm oder ein terminiertes Ereignis (Gupta 2018). Jedes Mal, wenn eine Funktion ausgelöst wird, initialisiert AWS automatisch die Ressourcen einer Ausführungsumgebung und führt die Funktion innerhalb dieser aus (Amazon Web Services Inc. o.D. a).

Jede Funktion ist dabei zustandslos und hat keine Beziehung zur zugrunde liegenden Infrastruktur. Auf diese Weise kann Lambda schnell so viele Kopien einer Funktion starten, wie für die Skalierung erforderlich ist, um die Anzahl der eingehenden Ereignisse abzuarbeiten (Amazon Web Services Inc. o.D. b).

Bei der Erstellung einer Lambda-Funktion können Kunden verschiedene Konfigurationseinstellungen festlegen, wie zum Beispiel die Größe des Arbeitsspeichers, die Laufzeit oder die maximale Ausführungszeit. Lambda verwendet diese Informationen, um die Ausführungsumgebung einzurichten (Amazon Web Services Inc. o.D. c).

Grundlegendes Serviceverhalten

Das grundlegende Serviceverhalten von AWS Lambda beschreibt der typische Lebenszyklus einer Lambda-Ausführungsumgebung. Er besteht aus drei Phasen, die Abbildung 1 skizziert (Amazon Web Services Inc. o.D. c):

In Phase 1, der Init-Phase, erstellt Lambda gemäß der Kundenkonfiguration eine Ausführungsumgebung, lädt den Code für die Funktion herunter, initialisiert die Laufzeit und führt dann den Initialisierungscode der Funktion (d.h. der Code außerhalb des Haupthandlers bzw. der eigentlichen Hauptfunktion) aus.

Die Init-Phase, auch „Kaltstart-Phase“ genannt, wird bei der Standardbereitstellung von Lambda durch das „erste“ Ereignis gestartet, das eine Funktion erhält. Wenn der Kunde jedoch gegen einen Aufpreis eine sogenannte „Provisioned Concurrency“ hinzubucht, erfolgt der Start dieser Phase direkt im Anschluss an die Erstellung bzw. nach dem Deplyoment der Funktion in AWS.

In der zweiten Phase „Invoke“ ruft Lambda auf Basis des erhaltenen Ereignisses die Hauptfunktion auf und führt diese aus. Nachdem die Ausführung vollständig beendet wurde, lässt sie sich durch ein weiteres Ereignis erneut starten.

Phase 3, die Shutdown-Phase, wird eingeleitet, wenn die Funktion für eine gewisse Zeit keine Aufrufe erhält. Lambda fährt dann die Laufzeit herunter und entfernt die Ausführungsumgebung. Die Verarbeitung eines neuen Ereignisses beginnt bei der Standardbereitstellung somit wieder mit Phase 1.

Bekannte Einschränkungen des Serviceverhaltens

Die bekannten Einschränkungen des Serviceverhaltens von Lambda werden durch die dynamische Bereitstellung des Services verursacht. Die Auswirkungen daraus zeigen sich in der Ausführungsdauer der Funktion, wobei zwei Parameter sich in diesem Zusammenhang als relevant erweisen (Beswick 2021):

Der erste Parameter beinhaltet die Ressourcenallokation, welche durch den Ressourcenbedarf der Kunden von Lambda bestimmt wird. AWS weist Kunden immer die Anzahl an Ressourcen zu, die sie zum aktuellen Zeitpunkt benötigen. Ändert sich der Bedarf, wird neu allokiert. Folglich weist AWS permanent neue Ressourcen zu bzw. entzieht diese bei fehlendem Bedarf.

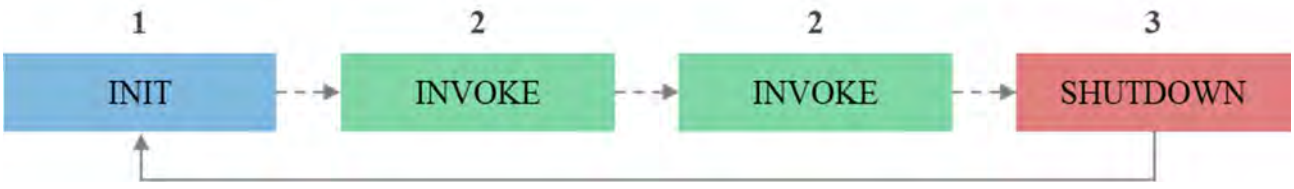


Abbildung 1: Lebenszyklus einer Lambda-Ausführungsumgebung (Quelle: Eigene Darstellung in Anlehnung an Amazon Web Services Inc. o.D. c)

Dies zeigt sich beispielsweise darin, dass die Ausführungsumgebung einer Lambda-Funktion, sobald initiiert, nicht dauerhaft aktiv bleibt. Zwar bleibt eine Funktion nach der Verarbeitung eines Ereignisses für kurze Zeit in einem sogenannten „warmen“ Zustand. In dieser Situation lässt sich beim Eintritt eines neuen zugehörigen Ereignisses Phase 1 überspringen und die Verarbeitung in der „Invoke“-Phase unmittelbar in der Ausführungsumgebung durchführen. Ansonsten erfolgt nach einer bestimmen Zeitdauer im Rahmen von Phase 3 durch AWS eine Löschung der Ausführungsumgebung, damit die zugrunde liegenden Ressourcen wieder frei für andere Kunden sind. Kundenseitig kann im Rahmen der Standardbereitstellung auf die Dauer dieses Zeitraums kein Einfluss genommen werden.

Erfolgen die Aufrufe der Funktion lediglich sporadisch ist davon auszugehen, dass die Verarbeitung bei einem hinzukommenden Ereignis erneut mit Phase 1 beginnt. Diese Kaltstart-Phase benötigt zusätzlich Zeit, die aufgrund der Initialisierungsaktivitäten oftmals signifikant länger dauert als die eigentliche Verarbeitung in Phase 2.

Dies gilt ebenso für parallele Funktionsaufrufe, da hierbei jedes Ereignis eine separate Ausführungsumgebung benötigt. Entsprechend erfahren sämtliche parallele Verarbeitungsvorgänge, für die keine „warme“ Umgebung bereitsteht, ebenfalls einen Kaltstart.

Der zweite Parameter im Rahmen der bekannten Einschränkungen des Serviceverhaltens von Lambda ist die Ressourcenverfügbarkeit, welche in der Verantwortung von AWS liegt.
AWS ist bestrebt diese auf möglichst hohem Niveau zu halten und permanent zu optimieren. Aktuell wird für Lambda z. B. ein Service Level Agreement (SLA) mit einer Verfügbarkeitsgarantie von 99,95% den Kunden zugesprochen (Amazon Web Services Inc. o.D. d).

Um dies zu gewährleisten, wird der Service und damit auch die Ausführungsumgebungen der Kunden über mehrere Rechenzentren hinweg verteilt und verwaltet. Dabei werden verschiedene Sicherheitsmechanismen aktiv. Sie sorgen beispielsweise in Abhängigkeit vom erzeugten Datenvolumen durch die Kunden und der daraus resultierenden Auslastung der einzelnen Rechenzentren dafür, dass ein Lastausgleich für eine Funktion vorgenommen werden kann. Das bedeutet, dass bei einer Überlastung oder bei einer ungleichen Lastverteilung zwischen den Rechenzentren Ressourcen umverteilt und neu initialisiert werden.

Deshalb ist es möglich, dass eine Funktion innerhalb eines kurzen Zeitraums zweimal aufgerufen wird und beide Ausführungen dennoch aufgrund dieser Lastausgleichsaktivitäten einen Kaltstart erfahren.

USE CASE IOT-ANWENDUNG

Der folgende Use Case handelt von einem Industrieunternehmen, das im Rahmen der AWS Cloud eine IoT-Anwendung entwickelt, die Maschinendaten aus der Produktion verknüpft und analysiert. Auf dieser Basis sollen zu ergreifende Maßnahmen definiert und automatische Workflows, zum Teil in Echtzeit, initiiert werden.

In der Umsetzung ist AWS Lambda ein wichtiger Bestandteil des Dateneingabe-Moduls der Anwendung. Dieses veranschaulicht Abbildung 2 anhand eines Architekturdiagramms. Hierbei werden die Maschinendaten, nach ihrer Übertragung in die Cloud, zunächst durch einen SQS-Queue zwischengepuffert.

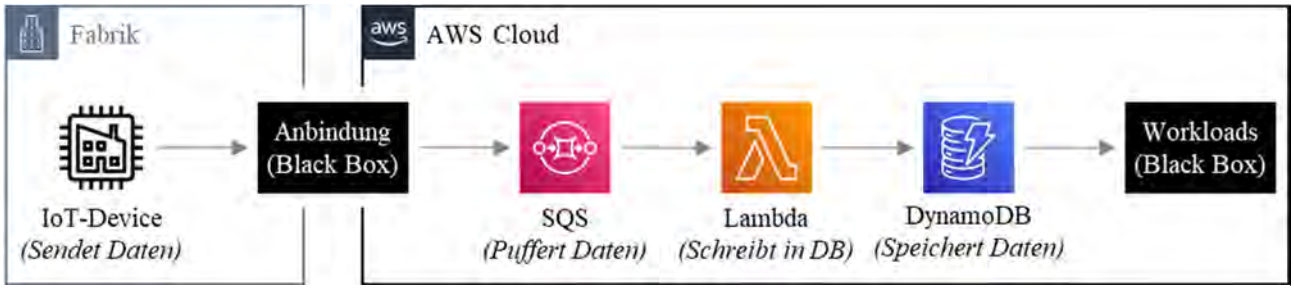


Abbildung 2: Use Case Architektur (Quelle: Eigene Darstellung)

Sobald neue Datensätze vorliegen, sendet der Queue ein Ereignis an eine Lambda-Funktion, welche daraufhin die Daten aus dem Queue entnimmt und in eine DynamoDB Datenbank schreibt. Dieser Datenhub fungiert als zentraler Ausgangspunkt für weitere Verarbeitungsschritte.

Besonderheiten und Methodik

Im vorliegenden Use Case gibt es kaum Abweichungen bei der Eingaberate der Maschinendaten. Weiterhin verarbeitet Lambda die SQS-Ereignisse in Batches, wobei dies im Sekundenrhythmus erfolgt oder wenn 10 Stück in weniger als einer Sekunde aufgelaufen sind. Bei kleineren Lastspitzen muss die Funktion folglich nicht sofort skalieren.

Entsprechend ist aufgrund der geschilderten Ausganssituation davon auszugehen, dass es bei diesem Setup nur dann zu einem verzögernden Kaltstart von Lambda kommen kann, wenn AWS einen Lastausgleich zwischen seinen Rechenzentren vornimmt.

Im Rahmen einer vertiefenden Analyse wird dieser angenommene Sachverhalt überprüft. Ein weiteres Ziel in diesem Kontext besteht darin, einen Überblick über die Häufigkeit solcher erkennbaren Lastausgleichsaktivitäten innerhalb eines bestimmten Zeitraums zu erhalten.

Die Vorgehensweise orientiert sich an der Auswertung von CloudWatch Logs, die jede Lambda-Funktion standardmäßig für alle Ausführungen im Betrieb generiert. Anhand der Logs lässt sich zum Beispiel ablesen wie lange die Ausführung gedauert hat oder ob es zu einem Kaltstart gekommen ist und wie viel Zeit dieser in Anspruch genommen hat.

Die Auswertung der Logs einschließlich Visualisierung und Filterung erfolgt im Rahmen der CloudWatch Konsole und der Abfragesprache von CloudWatch Logs Insights.

Ergebnisse der AWS CloudWatch Log Analyse

In einem ersten Schritt der Log-Analyse wurde ein Betrachtungszeitraum von einem vollen Tag gewählt, der um 00:00:00 Uhr beginnt und um 23:59:59 Uhr endet. Die IoT-Anwendung ist dabei durchgehend in Betrieb. Abbildung 3 veranschaulicht dazu den zeitlichen Verlauf der Funktionsaufrufe und der jeweiligen Ausführungsdauer für die Verarbeitung der Batches.

Ausführungsdauer anstatt wie im Durchschnitt wenige Millisekunden, mehr als 750 Millisekunden dauert.

Der Grund hierfür liegt nicht in der Anwendung, da

- der Umfang der zu verarbeitenden Datenpakete nur geringfügig variiert,
- die Anzahl an verwendeten Ausführungsumgebungen für die Batchverarbeitung nicht verändert wurde, denn sie lag konstant bei eins, und
- grundsätzlich keine Fehler aufgetreten sind, welche zu einer Neuinitialisierung der Ausführungsumgebung geführt hätten.

Wie die Auswertung der CloudWatch Logs zeigt, wurde aber jede der zu erkennenden Abweichungen durch die Neuinitialisierung der Lambda-Ausführungsumgebung, also einen Kaltstart, verursacht. Die in Tabelle 1 dargestellten Ergebnisse verdeutlichen dies.

Den Ergebnissen der Log-Analyse zufolge wurden alle Ausführungen, die länger als 143,27 Millisekunden dauerten, durch einen Kaltstart verzögert. Für alle 11 erheblichen Abweichungen von der durchschnittlichen Ausführungsdauer, die in Abbildung 3 zu erkennen sind, trifft das zu.

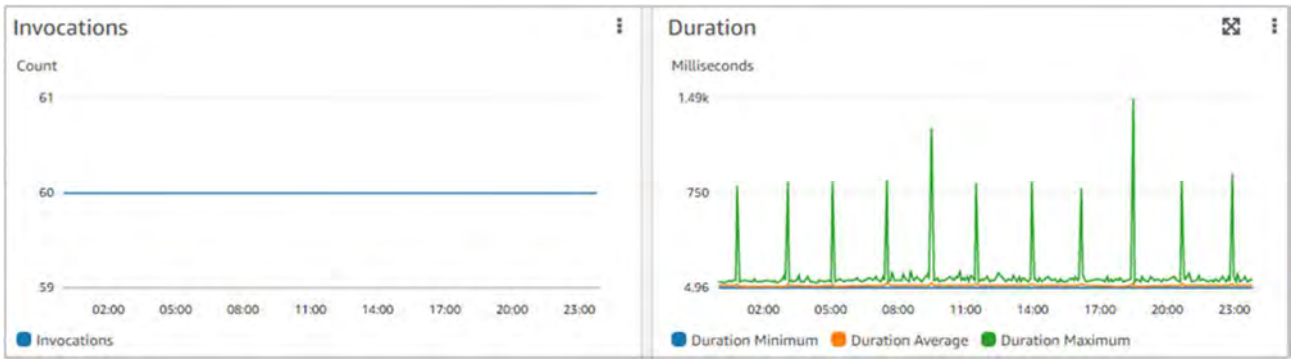


Abbildung 3: Zeitlicher Verlauf der Funktionsaufrufe und der jeweiligen Ausführungsdauer im Zeitraum 00:00:00 Uhr bis 23:59:59 Uhr am 23.03.2022 (Quelle: Eigene Darstellung)

Wie dargestellt, sind die Aufrufe (Invocations) der Lambda-Funktion über den gesamten Zeitraum konstant. Dennoch kommt es bei der Ausführungsdauer (Duration) aber wiederholt zu Abweichungen, wobei die Ursache hierfür liegt nicht bei dem Verhalten oder einem Fehler der Anwendung. Sie muss demnach bei einem internen Prozess des Lambda-Service bzw. bei AWS liegen.

Tabelle 1: Auswertung CloudWatch Logs der Lambda-Funktion im Zeitraum 00:00:00 Uhr bis 23:59:59 Uhr am 23.03.2022 (Quelle: Eigene Darstellung)

	Ausführungen mit Kaltstart-Phase	**Kaltstart-Phase**	**Ausführungen ohne Kaltstart-Phase**
Minimale Dauer (in MS)	787,92	449,27	4,96
Durchschnittliche Dauer (in MS)	932,51	590,35	21,77
Maximale Dauer (in MS)	1.494,95	1.143,81	143,27
Anzahl	11	11	17.269

Es erscheint aber unwahrscheinlich, dass es sich um Auswirkungen handelt, die aufgrund von intern durchgeführten Lastausgleichsaktivitäten herrühren. So sind anhand der Datenlage regelmäßige Zeitabstände zwischen den einzelnen Abweichungen erkennbar, die darauf schließen lassen, dass AWS den Austausch der Ausführungsumgebungen geplant bzw. terminiert ansteuert.

Eine mögliche Ursache hierfür könnte in der Absicherung der Service-Verfügbarkeit liegen, wenn davon ausgegangen wird, dass das Setup einer Ausführungsumgebung mit der Zeit an Stabilität verliert. Die genaue Ursache für die regelmäßigen Neuinitialisierungen lässt sich jedoch nicht aus der Auswertung ableiten.

Durch dieses Verhalten muss der Kunde Performanceschwankungen tolerieren, über die er keine Kontrolle hat. Es handelt sich somit um eine Einschränkung der Servicequalität.

Um die beobachtete Regelmäßigkeit zu untermauern, wurde der Betrachtungszeitraum anschließend auf drei Tage erweitert, von 00:00:00 Uhr am ersten Tag bis um 23:59:59 Uhr am dritten Tag. Abbildung 4. veranschaulicht hierzu wieder den zeitlichen Verlauf der Funktionsaufrufe und der Ausführungsdauer für die Verarbeitung der Batches.

Das gleiche Resultat ergibt die Analyse der Log-Files, wonach jede Ausführung, die länger als 784,68 Millisekunden, oder ohne diesen Ausreißer, länger als 148,84 Millisekunden dauerte durch einen Kaltstart verzögert wurde. Bei den meisten Abweichungen, die in Abbildung 4 zu erkennen sind, tritt der beschriebene Sachverhalt somit auf. Tabelle 2 zeigt hierzu die Ergebnisse der Auswertung.

In Summe kommt es im betrachteten Zeitraum zu 39 Kaltstartes, obwohl auch hier der Umfang der zu verarbeitenden Datenpakete nahezu immer gleich war, die Anzahl an verwendeten Ausführungsumgebungen konstant bei 1 lag und es auch zu keinem Ausfall gekommen ist.

Diskussion und Implikationen

Die Ergebnisse der Log-Analyse zeigen, dass die Durchgängigkeit einer Lambda-Ausführungsumgebung auch bei permanent aktiver Umgebung nicht gewährleistet ist. Im dauerhaften Betrieb wird sie immer wieder ausgetauscht und neu initialisiert, was einen Kaltstart zur Folge hat, der die Ausführung um bis zu 1,37 Sekunden verzögert.

Es fällt auf, dass der Austausch der Ausführungsumgebungen in einer gewissen Regelmäßigkeit stattfindet. Dieser Sachverhalt wurde im

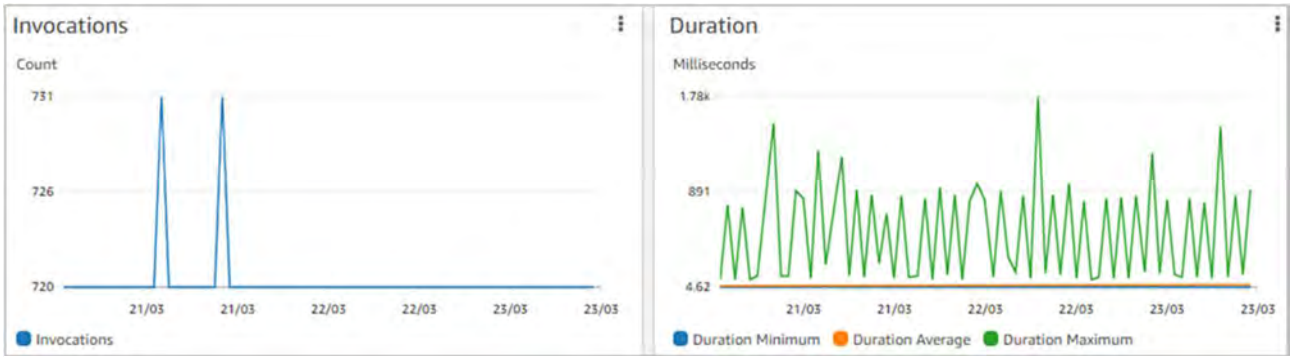


Abbildung 4: Zeitlicher Verlauf der Funktionsaufrufe und der jeweiligen Ausführungsdauer im Zeitraum 21.03.2022 um 00:00:00 Uhr bis 23.03.2022 um 23:59:59 Uhr (Quelle: Eigene Darstellung)

Wie dargestellt, setzt sich die zuvor beobachtete Regelmäßigkeit in den Abweichungen der Ausführungsdauer (Duration), trotz überwiegend konstanter Aufrufe, fort.

Rahmen der ersten Log-Analyse mit einem Betrachtungszeitraum von einem Tag beobachtet und durch eine erweiterte Analyse mit einem Beobachtungszeitraum von drei Tagen bestätigt. Aus

Tabelle 2: Auswertung CloudWatch Logs der Lambda-Funktion im Zeitraum 21.03.2022 um 00:00:00 Uhr bis 23.03.2022 um 23:59:59 Uhr (Quelle: Eigene Darstellung)

	Ausführungen mit Kaltstart-Phase	**Kaltstart-Phase**	**Ausführungen ohne Kaltstart-Phase**
Minimale Dauer (in MS)	685,27	381,36	4,61
Durchschnittliche Dauer (in MS)	930,82	579,03	22,1
Maximale Dauer (in MS)	1.780,11	1.371,38	(784,68) 148,84
Anzahl	39	39	51.823

diesem Grund erscheint es wenig plausibel, dass ein interner Lastausgleich zwischen den Rechenzentren von AWS dafür verantwortlich ist. Anzunehmen ist vielmehr, dass der Austauschprozess terminiert ist und nach einer gewissen Einsatzdauer einer Lambda-Ausführungsumgebung initiiert wird.

Im behandelten Use Case und dem dabei betrachtetet Ausführungszeitraum werden die Umgebungen in einem durchschnittlichen Rhythmus von 1,85 Stunden ausgetauscht. Der relative Anteil der Ausführungen, die dabei durch einen Kaltstart verzögert werden, ist zwar gering – bei 51.862 Ausführungen beträgt dieser 0.0752% –, allerdings ist zu beachten, dass in größeren Anwendungsszenarien oftmals mehrere Lambda-Funktionen voneinander abhängen bzw. miteinander verkettet sind, wobei jede Funktion vom dargestellten Sachverhalt betroffen sein kann.

Wenn der zeitliche Rhythmus der Neuinitialisierungen für jede Funktion versetzt erfolgt, steigt als Konsequenz mit jeder Funktion, welche ein Datensatz durchläuft, die Wahrscheinlichkeit, dass die Verarbeitung durch einen Kaltstart verzögert wird. Anhand der Use Case Daten veranschaulicht Tabelle 3 diese Auswirkungen im Detail. Als Grundlage der Berechnung dient die folgende Formel:

$$P = 1 - (1 - 0{,}000752)^n$$

Ferner gibt die in Abbildung 5 skizzierte Architektur, welche sich am Setup des Use Case orientiert, ein praktisches Beispiel für einen Workload, bei dem mehrere Funktionen miteinander verkettet sin: Die erste Lambda-Funktion, welche die IoT-Datensätze aus dem SQS-Queue entnimmt, ruft über eine zweite Funktion weitere Informationen aus einer weiteren Datenbank ab, um die IoT-Daten mit statischen Attributen anzureichern. Anschließend wird durch die DynamoDB eine dritte Funktion getriggert, die anhand der aktualisierten Datensätze eine Reihe von Kennzahlen in einem Dashboard aktualisiert oder einen weiteren Prozess anstößt.

Alle drei Funktionen weisen demzufolge Abhängigkeiten voneinander auf, wobei die Ausführungsrate von jeder Funktion jeweils durch die Eingaberate der IoT-Daten bestimmt wird. Sie sind somit dauerhaft in Verwendung. Die Wahrscheinlichkeit, dass bei einer dieser drei Funktionen der Verarbeitungsprozess der IoT-Daten verzögert wird, liegt demnach bei 0,2254%.

Tabelle 3: Wahrscheinlichkeit für Kaltstart in Abhängigkeit der Anzahl an verketteten Funktionen (Quelle: Eigene Darstellung)

Verkettete Funktionen (n)	Wahrscheinlichkeit (P) für prozessverzögernden Kaltstart
1	0,0752%
2	0,1503%
3	0,2254%
4	0,3005%
5	0,3754%
6	0,4504%

Bei einem Volumen von 1 Millionen Batchverarbeitungsvorgänge, die von jeder Funktion zu leisten sind, wären folglich 2.254 Verarbeitungsvorgänge von einem verzögernden Kaltstart betroffen.

Die einzige Möglichkeit, um den Lambda-Service zu verwenden und dieses Problem zu umgehen, besteht im Hinzubuchen einer Provisioned Concurrency für jede Lambda-Konfiguration. Es ist allerdings zu beachten, dass in der Folge die Nutzungskosten von AWS Lambda erheblich steigen können.

Dies kann zum Beispiel anhand einer Rechnung mit dem Online frei verfügbaren AWS Pricing Calculator demonstriert werden (Amazon Web Services Inc. o.D. d). Dabei ist eine mögliche Konfiguration für Lambda eine

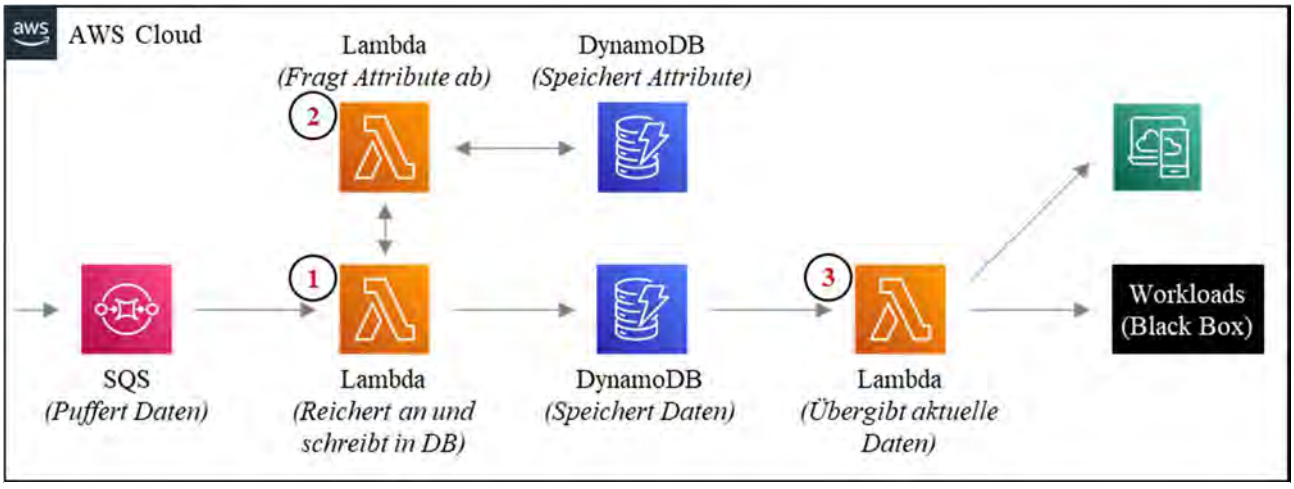


Abbildung 5: Erweiterte Architektur von Use Case (Quelle: Eigene Darstellung)

x86 Architektur mit einem Arbeitsspeicher von 1024 MB. Die bereitstellende Region, in welcher die Funktion gehostet wird, ist für dieses Beispiel Irland. Die Provisioned Concurrency wird auf 1 gesetzt und soll für einen Tag aktiviert sein – das bedeutet, dass eine Ausführungsumgebung nach Erstellung der Funktion für 24 Stunde, sofern sie nicht gerade ein Ereignis verarbeitet, permanent in einem warmen Zustand gehalten wird.

Die Kosten dafür betragen 0,40 USD. Hinzu kommen die Kosten für die eigentliche Datenverarbeitung in dieser Zeit. Bei einer Anzahl von 1 Millionen Anfragen an die warme Umgebung, belaufen sich diese auf 0,25 USD. Dabei wird impliziert, dass ein Verarbeitungsvorgang im Durchschnitt 5 Millisekunden dauert. Somit beträgt der Preis für eine Funktion mit einer Provisioned Concurrency von 1 pro Tag in Summe 0,65 USD.

Wenn keine Provisioned Concurrency für die verwendete Umgebung aktiviert ist, liegen die Kosten für die Datenverarbeitung etwas höher bei 0,28 USD. In diesem Fall handelt es sich aber um den einzigen Kostenfaktor. Die Differenz zu einer Lambda-Konfiguration mit Provisioned Concurrency von 1 beträgt demnach pro Tag 0,37 USD.

Auf ein Jahr gerechnet, summiert sich der Kostenunterschied in der Bereitstellung für diese eine Funktion und Ausführungsumgebung auf 135,05 USD. Die Kosten für die Funktion ohne Provisioned Concurrency würden dabei 102,2 USD betragen, während für die Funktion mit einer Provisioned Concurrency von 1 insgesamt 237,25 USD anfallen.

Für jeden vergleichbaren Use Case stellt sich daher die Frage, ob sich zur Lösung der herausgearbeiteten Kaltstart-Problematik eine Provisioned Concurrency als sinnvoll und ob die hierfür zusätzlich anfallenden Kosten tragbar sind. Dies hängt von den Anforderungen an den jeweiligen Use Case bzw. den individuellen Rahmenbedingungen ab.

Bei der Entscheidungsfindung ist davon auszugehen, dass sich die drei nachfolgenden Fragen als relevant erweisen:

1. Gefährden regelmäßige Kaltstarts einer Lambda-Funktion die Effizienz der Anwendung?
2. Erweist sich eine serverlose Umsetzung mit AWS Lambda als gerechtfertigt, auch wenn das kostenpflichtige Hinzubuchen einer Provisioned Concurrency erforderlich ist, um die Performance-Anforderungen an die Applikation erfüllen zu können?
3. Ist eine alternative Bereitstellung, trotz höherer Verwaltungsaufwände, im Vergleich zu AWS Lambda mit aktivierter Provisioned Concurrency kostengünstiger? Im Falle einer Kostengleichheit gilt es zu bedenken, dass kundenseitig mehr Einflussnahme über die zugrunde liegende Infrastruktur ausgeübt werden kann.

FAZIT UND AUSBLICK

Serverlose Cloud-Services sind dafür bekannt, dass sie hoch verfügbar sind und ein attraktives Kostenmodell bieten. Dieses Verhalten gilt es jedoch kritisch zu würdigen.

Am Beispiel des serverlosen Cloud-Services AWS Lambda zeigt der vorliegende Beitrag auf, dass Kunden aufgrund der eingeschränkten Kontrolle über die zugrunde liegende Infrastruktur auch technische Einschränkungen hinnehmen müssen. Diese äußern sich im Performanceverhalten des Service.

So werden die zugrunde liegenden Ressourcen einer Lambda-Funktion durch AWS dynamisch optimiert. Den Kunden werden somit immer nur so viele Kapazitäten bereitgestellt, wie aktuell benötigt werden. Wenn die Anwendungslast steigt, ist zwar durch die SLAs abgesichert, dass AWS den Bedarf deckt, jedoch kann es zu Kaltstartzeiten kommen. Dies liegt daran, dass neu zugeteilte Ressourcen vor der Nutzung immer zuerst hinsichtlich der Kundenkonfiguration initialisiert werden müssen.

Des Weiteren zeigt eine vertiefenden Use Case-Analyse, dass AWS dauerhaft aktive Laufzeitumgebungen einer Lambda-Funktion regelmäßig austauscht. Dies wirkt wie ein interne Lastausgleich und führt ebenfalls zu Kaltstarts und damit einhergehenden Verzögerungen.

Die von AWS Lambda angebotene Konfigurationsmöglichkeit „Provisioned Concurrency“, mit der es möglich ist, die Kaltstart-Phase für eine definierte Anzahl an Ausführungsumgebungen zu überspringen, wird ebenfalls betrachtet. Es zeigt sich allerdings, dass im Vergleich zu den Preisen, welche typischerweise für die eigentliche Datenverarbeitung anfallen, mit erheblichen Mehrkosten zu rechnen ist. Dies gilt insbesondere dann, wenn diese Lösung dauerhaft adaptiert wird. Es ist daher in Abhängigkeit vom Anwendungsfall und dem dahinterstehenden Business Case zu entscheiden, ob die Einrichtung einer Provisioned Concurrency sinnvoll und wirtschaftlich tragbar ist.

Entsprechend kann es sich als sinnvoll erweisen, Alternativen zu AWS Lambda hinsichtlich Kosten und Leistung zu evaluieren. Wenn Kaltstartzeiten zwingend zu vermeiden sind, kann eine serverbasierte Lösung im Vergleich zum serverlosen Lambda-Service mit Provisioned Concurrency kostengünstiger sein und gleichzeitig mehr Kontrolle über die zugrunde liegende Infrastruktur bieten.

Weitere Studien zu diesem Thema können hier ansetzen. Unternehmen dürften in diesem Kontext vor allem an der

Frage interessiert sein, ab welchem Datenvolumen eine serverbasierte Lösung gegenüber AWS Lambda, ein besseres Kosten und Leistungsverhältnis bietet.

Für eine ganzheitlichen Überblick erweist es sich ebenfalls von Interesse, wie sich die Kaltstartthematik bei anderen führenden Cloud-Anbieter, z.B. Microsoft Azure oder Google Cloud, verhält und welche Optionen diese anbieten.

LITERATUR

Amazon Web Services Inc. o.D. a. *"AWS Lambda Developer Guide. Stichwort: What is AWS Lambda?"* https://docs.aws.amazon.com/lambda/latest/dg/welcome.html (besucht am 17.09.2022).

Amazon Web Services Inc. o.D. b. *"AWS Lambda Developer Guide. Stichwort: AWS Lambda foundations."* https://docs.aws.amazon.com/lambda/latest/dg/lambda-foundation.html (besucht am 17.09.2022).

Amazon Web Services Inc. o.D. c. *"AWS Lambda Developer Guide. Stichwort: AWS Lambda execution environment."* https://docs.aws.amazon.com/lambda/latest/dg/lambda-runtime-environment.html (besucht am 17.09.2022).

Amazon Web Services Inc. o.D. d. *"AWS Service Level Agreement (SLA)."* https://aws.amazon.com/de/legal/service-level-agreements/ (besucht am 17.09.2022).

Amazon Web Services Inc. o.D. e. *"AWS Pricing Calculator. Die Kosten für Ihre Architektur-Lösung schätzen."* https://calculator.aws/#/ (beucht am 17.09.2022).

Artasanchez, A. 2021. *"AWS for Solutions Architects. Design your cloud infrastructure by implementing DevOps, containers, and Amazon Web Services."* Packt Publishing, Birmingham.

Bahga, A und V. Madisetti. 2019. *"Cloud Computing Solutions Architect – A Hands-On Approach. A Competency-based Textbook for Universities and a Guide for AWS Cloud Certification and Beyond."*

Beswick, J. 2021. *"Operating Lambda. Performance Optimization – Part 1."* https://aws.amazon.com/de/blogs/compute/operating-lambda-performance-optimization-part-1/ (besucht am 17.09.2022).

Dremel, C. und M. Herterich. 2018. *„Digitale Cloud-Plattformen als Enabler zur analytischen Nutzung von operativen Produktdaten im Maschinen- und Anlagenbau." In Cloud Computing – Die Infrastruktur der Digitalisierung 2018, Reinheimer, S. (Hrsg.).* Wiesbaden, 73–88.

Gupta, M. 2018. *„Serverless Architectures with AWS – Discover how you can migrate from traditional deployments to serverless architectures with AWS."* Packt Publishing, Birmingham.

Laszewski, T.; K. Arora; E. Farr und P. Zonooz. 2018. *„Cloud native architectures. Design high-availability and cost-effective applications for the cloud."* 1. Aufl. Packt Publishing, Birmingham.

Application of Artificial Intelligence to improve Customer Understanding: Transformer based Topic Modeling in practice

Nils Blessing
Pforzheim University
Tiefenbronner Straße 65
75175 Pforzheim
blessing@hs-pforzheim.de

Prof. Dr. Frank Morelli
Pforzheim University
Tiefenbronner Straße 65
75175 Pforzheim
frank.morelli@hs-pforzheim.de

KEYWORDS

Artificial Intelligence, Unsupervised Learning, Natural Language Processing, Cloud Computing

1 ABSTRACT

Digitalization offers useful data, especially unstructured data, to increase customer understanding. However, without targeted and efficient methods this poses a great challenge to cope with. In this article, the development of a prototype alongside the CRISP-DM model is considered in order to outline a suitably holistic solution.

In order to create the prototypical application, textual data is cleaned and transformed into a unified language by using machine translation. A transformer model then is employed to generate numerical representations of texts, while preserving semantic relationships. Finally, algorithms for dimensionality reduction, creating topics and corresponding representations as well as the specification of individual descriptive words are applied.

Within the presented prototype, a sentence is considered as the smallest unit of information. By splitting customer feedback into sentences, several clusters or topics can be assigned to one document. The use case shows, that fine-grained analysis is possible by using short texts with comparatively few sentences. However, if the character of the data changes, this procedure has to be adapted under given conditions according to the CRISP-DM approach. While a major challenge emerges in evaluating the quality of clustering results, considerable potential in the use of GPU resources with respect to processing time and cost can be presented as key findings.

2 INTRODUCTION

With the continuing or even further accelerating pace of digitalization, the amount of data is growing exponentially and is following more and more the big data nature. An estimation by the International Data Corporation (IDC) indicates that the amount of data generated worldwide will increase approximately tenfold over the years from 2016 to 2025 (Potnis, 2018, p. 2). Concerning the acquisition of product quality data, with the digital age there are advanced possibilities besides the classic and targeted customer studies. Domain-specific, purely digital data from alternative channels such as digital services, mobile apps or social media platforms is now also gathered and considered a valuable asset across industries.

With a special focus on the automotive industry, customer data can nowadays already be collected via vehicles themselves. With the advances and possibilities of the digitalization, increasingly connected products and services continue to emerge. Today's vehicles are no longer just a means of transportation, but rather a digitally connected lifestyle product. This brings new communication channels and opportunities, creating advanced interactions between the enterprise, product and customer. Isson (2018) described the overwhelming volume of unstructured data as follows:

> "It comes from various sources, such as social media, [...] product reviews, market research, customer care conversations, voice of customer, consumer feedback, and employee narrative through performance review. The challenge is to derive reliable and relevant insights from the ever-increasing stream of data and maximize the business value buried in the unstructured data." (p. 34)

According to Potnis (2018), "it is important that organizations have the means to bring order to their data management, as unstructured datasets continue to grow within siloed storage infrastructure" (p. 3). As the name already implies here, a structure must first be established for unstructured

data in order to make it usable for further analyses and business processes. Given the large and constantly increasing volume as well as the lack of structure for the data, this ties up a great deal of capacities as well as resources. At this point, "without appropriate tools to manage data across silos, organizations can expect an increase in management overheads while also missing out on valuable data insights" (Potnis, 2018, p. 3).

Since the early detection of topics and issues within customer feedback makes a significant contribution to the product development, a suitable and efficient solution approach therefore offers thoroughly high potential.

3 CRISP-DM FRAMEWORK

As a foundation for the practical approach within this work, it is relied on the CRISP-DM framework, which "stands for CRoss Industry Standard Process for Data Mining. More popularly known by the acronym itself, CRISP-DM is a tried, tested, and robust industry standard process model followed for data mining and analytics projects" (Sarkar et al., 2018, p. 45). The framework covers six phases and links them into an overall process. "The sequence of the phases is not rigid. Moving back and forth between different phases is always required. The outcome of each phase determines which phase, or particular task of a phase, has to be performed next" (Chapman et al., 2000, p. 10). The underlying database has an essential role as a central element, and the process is not only theoretically but also practically oriented around it. The original descriptions for the respective phases are summarized and provided by means of a brief summary in the following.

3.1 Business Understanding

Capture and understand the requirements as well as objectives of a project looking at the requirements from a business perspective. The insights gained are then used to derive a data mining or data problem from it and define corresponding goals as well as a roadmap to achieve them. The phase combines, on the one hand, the business view and, on the other hand, the technical view of a given problem.

3.2 Data Understanding

This phase starts with an initial examination of the data, often in an exploratory way, to become familiar with and understand it. Furthermore, it is about gaining initial insights, uncovering obstacles due to, for example, quality issues, and deriving hypotheses regarding hidden information within the dataset.

3.3 Data Preparation

The data preparation phase deals with the preparation of the initial raw data so that it is available in the appropriate quality for the subsequent steps and can be processed further. Typical steps are the selection of relevant data so that a lean dataset is created. Furthermore, the data is transformed and cleaned according to the requirements. It should be mentioned that these steps can or even must be added to and repeated as the requirements are not always fully known or continue to arise and adapt in the course of the process.

3.4 Modeling

Modeling is the selection of techniques and methods, such as algorithms, to solve the identified problem. Often it is not just a matter of finding and using a single technique or algorithm, but rather the right orchestration of different ones, which then complement each other. Furthermore, in this phase, adaptations in the previous steps can become obvious, which have to be done by a further iteration loop.

3.5 Evaluation

To validate the achievement of the business objectives, this phase involves the evaluation of the model(s) generated. This may include the evaluation of quantitative metrics as well as the evaluation of qualitative feedback from relevant stakeholders, which serve to assess the quality. Also in this phase it is possible that previous steps or even the initial phase, i.e. the business understanding, are repeated and new learnings are generated.

3.6 Deployment

In this final phase, the deployment of the solution for its actual application in business processes is being considered. Depending on the requirements, the deployment can range from the creation of simple reports to the integration of the data computing process into a holistic information architecture system to serve a broad range of users and applications within the organization.

4 TEXT REPRESENTATION

To enable a machine to work with texts or natural language, these must first be converted into a suitable and computer-readable format. There are several methods or approaches for the numerical representation of text. In this chapter, the two concepts of local representations as well as distributed representations of texts are highlighted to familiarize the reader with these concepts. Both of these are used in the later implementation.

4.1 Local Representations

Local representations are a very straightforward and widely used concept for converting texts into a numerical format. However, Liu et al. (2020) pointed out a shortcoming in the use of simple, computer-readable representation methods like the one-hot vector; whereas the dimension is equal to the vocabulary size and 1 is assigned to the word's corresponding position and 0 to others. With such a proceeding "it is apparent that one-hot vectors hardly contain any semantic information about words except simply distinguishing them from each other" (p. 4). Even though there are already more advanced methods and approaches according to the state-of-the-art, local representations still have their raison d'être.

To enhance such simple representation models, the importance of a given term in each document (relative document frequency or df) can be calculated as an aspect to a given representation. This is achieved, for example, by the term frequency (tf), which is then multiplied by the inverse document frequency (idf). Such a concept is better known under the term $tf-idf$. Its function is thus be used to calculate a factor for weighting the specificity of a term based on its absolute frequency within a collection of documents (called corpus) and the document frequency of a given word or term.

First, the term frequency (tf) must be calculated, which is the sum of the occurrences of a term in a given document in relation to the total number of words in this document. The logic for this is shown under Equation (1). To calculate the inverse document frequency as a weighting factor, its equation is presented right below under (2).

$$tf(t,d) = \frac{n_{t,d}}{\sum_k n_{t,d}} \tag{1}$$

$$idf(t) = \log_{10}(\frac{N_D}{df_t}) \tag{2}$$

Here, the index t refers to a given term, N is the absolute number of all documents D within the corpus. The denominator df_t refers to the number of documents that actually contain a certain term. As a result, very frequent words (e.g. stop-words), which are present in almost every document, get no weight, while less common but still frequent words would receive a higher weight.

4.2 Distributed Representations

By the explanation of DeepAI (2021), distributed representations can be seen as methods which "describe the same data features across multiple scalable and interdependent layers. Each layer defines the information with the same level of accuracy, but adjusted for the level of scale. These layers are learned concurrently but in a non-linear fashion" (para. 1). This is complemented by Liu et al. (2020), who stated that "each feature is represented by a pattern of activation distributed over multiple elements, and each computing element is involved in representing multiple entities" (p. 5).

A very fundamental concept of distributed representations are word embeddings. One of the best-known models here is the so-called "Word2Vec" model, developed by Mikolov et al. (2013). Such a model learns from a collection of documents through shallow Neural Networks (only one or two hidden layers), which is composed by either general documents like Wikipedia articles or a collection of individual, domain-specific documents. "While learning such semantically rich relationships, Word2Vec ensures that the learned word representations are low dimensional (vectors of dimensions 50–500), instead of several thousands, as with previously discussed local representations in this chapter) and dense (most values in these vectors are non-zero)" (Vajjala et al., 2020, p. 94).

More generally, once a word embedding model has been trained, it can be used to retrieve a corresponding numerical vector representation for each word it contains. To obtain a representation for a text, i.e. more than just one word, in practice often all word embeddings of the words contained in this text are summed up and then their mean vector is calculated. However, this can have disadvantages: On the one hand, the word embeddings have no context to the given text. On the other hand, the sharpness of detail is blurred by averaging all embeddings.

To address the limitations of word embeddings, so-called pre-trained Language Models (PLM) have been on the rise since about 2018 and represent the state-of-the-art in Natural Language Processing to these days. As Vajjala et al. (2020) described, "there are several variants to these model architectures like Convolutional Neural Networks (CNN) or Long Short Term Memory (LSTM), and new models are being proposed every day by NLP researchers. [...] It is a constantly evolving area in NLP research; the state-of-the-art keeps changing every few months" (p. 146). When one refers to pre-trained models, it is referred to models that were previously trained on data. Instead of having machines train on data from scratch to perform NLP tasks, one starts with pre-trained language models that have already been trained on lots and lots of data to perform language modeling to good levels of performance.

5 TRANSFORMER MODELS

According to Vajjala et al. (2020), transformer models (also referred to as transformers) are the latest entry in the league of Deep Learning models to tackle Natural Language Processing tasks:

> "Transformers have achieved state-of-the-art in almost all major NLP tasks in the past few years. They model the textual context but not in a sequential manner. Given a word in the input, the model prefers to look at all the words around it (known as self-attention) and represent each word with respect to its context. For example, the word "bank" can have different meanings depending on the context in which it appears. If the context talks about finance, then "bank" probably denotes a financial institution. On the other hand, if the context mentions a river, then it probably indicates a bank of the river. Transformers can model such context and hence have been used heavily in NLP tasks due to this higher representation capacity as compared to other deep networks." (p. 25)

A transformer model essentially contains two components, namely an encoder and a decoder. A representative illustration of the associated architecture is shown in Figure 1.

TensorFlow (2021a) emphasizes the core idea

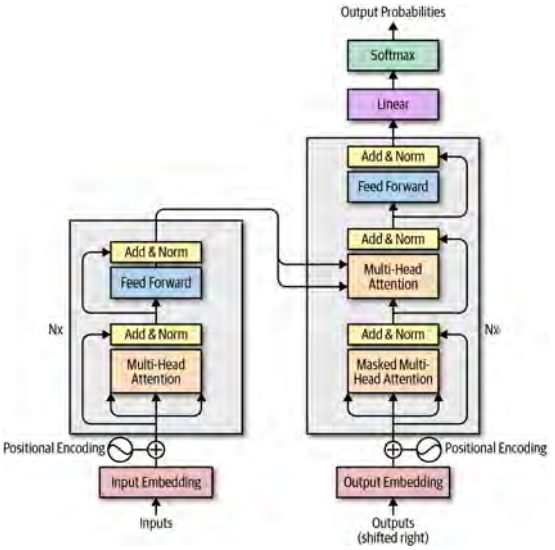


Figure 1: Transformer architecture with its encoder (left) and decoder (right) (Vaswani et al., 2017, p. 3)

behind transformer models, which is "self-attention - the ability to attend to different positions of the input sequence to compute a representation of that sequence. Transformers achieve this by the creating stacks of self-attention layers, which is also referred as multi-head attention" (para. 1). An example for a sequence could be a text that is transformed from a source language first into the numeric space and afterwards from the numeric space into a target language or sequence. This is also one of the main origins of the transformer models, namely the machine translation. Such a model was first introduced in the paper "Attention is All You Need" where Vaswani et al. (2017) show how it is possible to create powerful Neural Networks for sequential modeling that do not require complex recurrent or convolutional architectures but instead only rely on attention mechanisms. Transformer architectures today "power some of the most impressive practical examples of generative modeling, such as Google's BERT and GPT-2 for language tasks and MuseNet for music generation" (Foster, 2019, para. 3). Since the encoder part and its multi-head attention mechanism of the transformer architecture is mainly used to encode textual data in the context of this paper, it will be focused in the following.

5.1 Multi-Head Attention

With multi-head attention, one of the essential steps of the transformer comes into play, which is often referred to as the attention mechanism. The

task is to establish the associations between the individual words and thus to establish the contextual information. A more in-depth view of the multi-head attention mechanism can be seen in Figure 2 below.

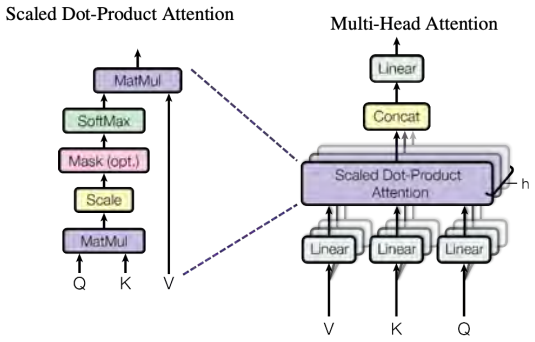


Figure 2: Scaled Dot-Product Attention (left). Multi-Head Attention consists of several attention layers running in parallel (right) (Vaswani et al., 2017, p. 4)

The authors Vaswani et al. (2017) described the attention mechanism as follows:

> "Mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key." (p. 3)

The previously presented figure shows the query (Q), keys (K) and values (V) by their initial letters, furthermore the stacked attention layers (h) are marked accordingly. In practice, the query, key, and value are each fed into a linear layer and are subjected to a matrix multiplication by a dot product, resulting in a score matrix. Subsequently, this score matrix is used to determine the focus or attention per word. Logically, a high weight in the matrix represents a higher attention. An additional step is the scaling of the weights, whereby these are converted to a probability, i.e. values between zero and one. Finally, the value vector can be multiplied by the attention weights to obtain an output vector. To turn this process into a multi-head attention, Phi (2020) summarizes the necessary steps as follows:

> "One need to split the query, key, and value into (N) vectors before applying self-attention. The split vectors then go through the self-attention process individually. Each self-attention process is called a head. Each head produces an output vector that is concatenated into a single vector before going through the final linear layer. In theory, each head would learn something different therefore giving the encoder model more representation power." (para. 7)

By combining all the aforementioned operations, a distributed representation is generated by the encoder, which incorporates attentional information from its input. The resulting dense vectors can now be used for a variety of applications.

6 DIMENSIONALITY REDUCTION

Dimensionality reduction can be described as "getting rid of uninformative information while retaining the crucial bits. There are many ways to define uninformative" (Zheng and Casari, 2018, ch. 6). The necessity is explained by Géron (2019) with the following rationale:

> "Many Machine Learning problems involve thousands or even millions of features for each training instance. Not only do all these features make training extremely slow, but they can also make it much harder to find a good solution. This problem is often referred to as the curse of dimensionality." (p. 213)

More generally, there are two basic concepts when considering the area of dimension reduction. Patel (2019) synthesized these major two concepts as follows:

> "The first is known as linear projection, which involves linearly projecting data from a high-dimensional space to a low-dimensional space. This includes techniques such as principal component analysis (PCA), singular value decomposition (SVD), and random projection. The second concept is known as manifold learning, which is also referred to as nonlinear dimensionality reduction. This involves techniques such as isomap, [...], multidimensional scaling (MDS), locally linear embedding (LLE), t-distributed stochastic neighbor embedding (t-SNE), [...] and independent component analysis." (ch. 3)

Linear projection and manifold learning have been widely used for quite some time, but due to the ever growing amount of data, they may reach their limits. McInnes et al. (2018) argue that it is "desirable to have an algorithm that is both scalable to massive data and able to cope with the diversity of data available." This is mainly because "dimension reduction techniques are being applied in a broadening range of fields and on ever-increasing sizes of datasets" (p. 2).

6.1 Uniform Manifold Approximation and Projection

The Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) algrorithm was proposed by McInnes and Healy (2018). It promises some significant advantages over previously known algorithms in terms of its scalability with very large and high-dimensional real-world datasets. Based on the research of McInnes et al. (2018), this results in an algorithm which is "a practical scalable algorithm that applies to real world data, [...] competitive with well-known algorithms such as t-SNE for visualization quality, and arguably preserves more of the global structure with superior run time performance" (p. 1). As the mathematical foundations of the UMAP research would exceed the scope of this paper, it is referred to a summary given by the authors McInnes et al. (2018):

> "In overview, UMAP uses local manifold approximations and patches together their local fuzzy simplicial set representations. This constructs a topological representation of the high dimensional data. Given a low dimensional representation of the data, a similar process can be used to construct an equivalent topological representation. UMAP then optimizes the layout of the data representation in the low dimensional space, minimizing the cross-entropy between the two topological representations. The construction of fuzzy topological representations can be broken down into the two problems: approximating a manifold on which the data is assumed to lie; and constructing a fuzzy simplicial set representation of the approximated manifold." (p. 2)

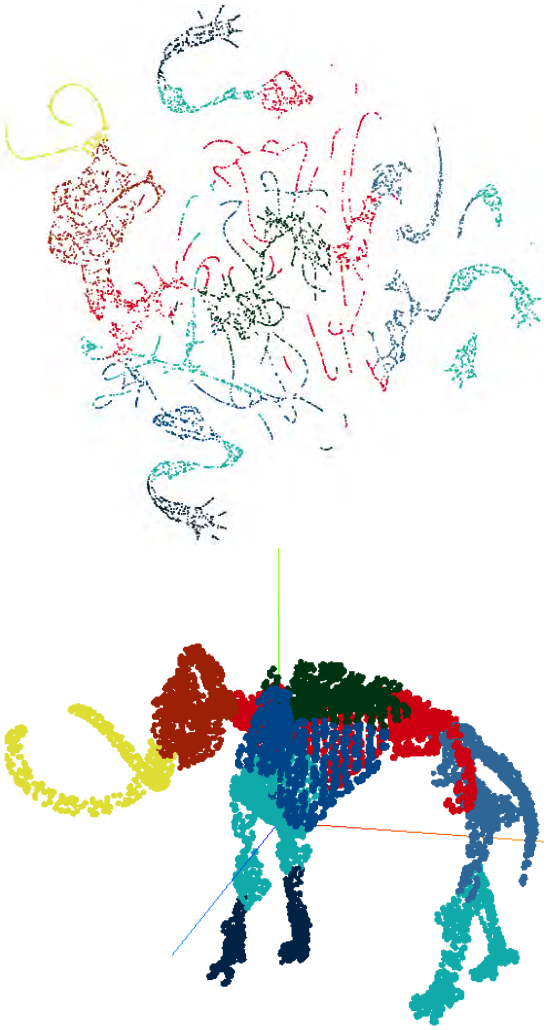


Figure 3: Sample projection results (top) by using UMAP on a three-dimensional dataset (bottom) to non-linearly project it in a lower, two-dimensional space (Coenen and Pearce, 2019)

To envision the capabilities of the algorithm, Figure 3 shows a simplified visual example. Here, sample data points (skeleton of a mammoth) are transferred from an initially three-dimensional space into the two-dimensional space by applying UMAP. What is particularly impressive here is that individual subsets of the original data retain their structure and appear to be flattened out. This can be clearly seen by the color-coding, with data points from contiguous regions appearing mostly as a composite even after they have been reduced.

7 CLUSTERING

Clustering is seeing a variety of use cases and opportunities within the industry to apply the technology in holistic systems. It can be categorized in the field of Machine Learning (ML) as an unsupervised learning technique. Aggarwal (2014)

describes the clustering problem as follows:

> "Clustering has been widely studied in the Data Mining and Machine Learning literature because of its numerous applications to summarization, learning, segmentation, and target marketing. In the absence of specific labeled information, clustering is considered a concise model of the data which can be interpreted in the sense of either a summary or a generative model. The basic problem of clustering may be stated as follows: Given a set of data points, partition them into a set of groups, which are as similar as possible." (ch. 1)

For example, in credit card fraud detection, clustering can group fraudulent transactions together, separating them from normal transactions. On the other hand, if only a few labels for the observations in our dataset are available, one could use clustering to group the observations first (without using labels). Then, the labels of the few-labeled observations could be transferred to the rest of the observations within the same group. This is a form of transfer learning, a rapidly growing field in Machine Learning (Patel, 2019, ch. 5).

The portfolio of algorithms in the area of clustering is vast and offers a wide range of choice. The selection of the appropriate algorithm here usually depends on the nature of the data to be clustered. Since each algorithm has its advantages as well as disadvantages for certain applications, this is always subjected to an initial condition.

7.1 Hierarchical Density Based Clustering of Applications with Noise

The Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm is an extension to the DBSCAN clustering algorithm and was introduced by Campello et al. (2013). At its core, it remains a clustering approach that works based on dense areas within the data but completes it by adding a hierarchical structure, which makes it possible to combine or further differentiate related sub-clusters based on its parameterization. An essentially advantage of such a density based method is above all that the clusters can also be present in arbitrary shapes, sizes and numbers as well that there is a certain robustness against impure data. This robustness is achieved not at least by the possibility to classify corresponding data points as outliers or noise. Another interesting point here is that the number of clusters does not have to be known in advance, but can rather be controlled indirectly via the parameterization.

8 TOPIC MODELING

Topic modeling, also known as topic detection, is the use of "a suite of probabilistic Machine Learning algorithms to discover and annotate large archives of documents with thematic information." Corresponding algorithms "are statistical methods that analyze the words of the original texts to discover the themes that run through them, how those themes are connected to each other, and how they change over time" (Blei, 2012, p. 77). Two common assumptions underlie all topic modeling methods: Each document consists of a mix of topics, and each topic consists of a collection of words or terms, and the topics are "hidden" or "latent" constructs in between documents and words. The goal of topic modeling is to discover the latent variables (i.e. topics) that shape the meaning or semantics in a document collection (Delen, 2020, ch.7). Topic modeling operationalizes the intuition of bringing out some words, which contribute to an understanding of a given corpus. It tries to identify the "key" words present in a text corpus without prior knowledge about it, unlike the rule-based text mining approaches that use regular expressions or dictionary-based keyword searching techniques (Vajjala et al., 2020, p. 253).

While methods such as Latent Semantic Indexing (LSI) and probabilistic topic models, such as Latent Dirichlet Allocation (LDA) are based on local representations, Angelov (2020) presents in his paper "Top2Vec: Distributed Representations of Topics" how distributed representations can be used to perform topic modeling. The author states that to achieve optimal results with traditional topic modeling methods they often require the number of topics to be known as well as further preprocessing steps such as custom stop-word lists, stemming, and lemmatization are required. "Additionally these methods rely on local representation of documents, which ignore the ordering, and semantics of words. Distributed representations of documents and words have gained popularity due to their ability to capture semantics of words and documents" (Angelov, 2020, p. 1).

Since the concept of Angelov (2020) turned out to be difficult when working with distributed representations by transformer models, Grootendorst (2022) proposes a paper "BERTopic: Neural topic modeling with a class-based TF-IDF procedure" which introduces a topic representation by using class-based term frequency - inverse document frequency (c-TF-IDF) and Maximal Marginal Relevance (MMR). Following the author, its approach "generates coherent topics and remains competitive across a variety of benchmarks involving classical models and those that follow the more recent clustering approach of topic modeling" (Grootendorst, 2022, p. 1).

9 IMPLEMENTATION

To put together the presented concept and make it usable usable in practice, it now has to be implemented productively. To this end, the requirements have already been determined in advance and the data reviewed. The goal is to process an existing database with a large number of customers feedback through the process in order to assign a corresponding topic to each data point as a result. The nature of the underlying data in this case is short texts, i.e. texts consisting of a small number of sentences or words. In addition, these are available in various languages and are to be converted into a uniform language, in this case English. Preprocessing is generally not required, but is carried out in this example, since the texts contain very specific (non-natural) patterns, which are removed in the process. Such patterns can be removed by simply using regular expressions.

9.1 Data Preparation

Data preparation is used to select and prepare relevant data for the subsequent modeling phase. In the given application, the unstructured data, i.e. the customer feedback, is initially selected from the overall database. For the actual preparation of the data, a process that is illustrated in Figure 4 was defined. The sub-process steps can be seen here as modules with sub-functions of the preprocessing. This modular design allows individual components to be adapted or even completely exchanged later during operational use in a well-organized manner and has significant advantages in terms of maintainability. After preparation, the database is enriched with the newly generated information, which is indicated by a dashed line.

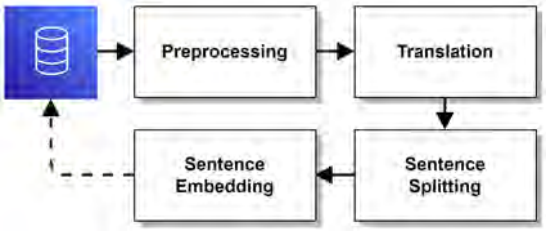


Figure 4: Data preparation process including corresponding sub-steps

In order to create a uniform language basis for the customer feedback, all texts are to be converted into English language if this is not already the case. To also ensure that the translation is efficient, an identification of the language of the texts to be processed is carried out in advance of the actual translations of the texts. In general, this step could be omitted if a multilingual transformer model is used during the embedding process. However, this would also lead to topic representations with very similar words, which in turn come from different languages, making interpretability difficult in some cases. Another peculiarity of the structure is that the texts are each split into their sentences. The reason for this is that in the underlying example, the smallest unit of information is assumed to be at the sentence level.

9.2 Modeling

As a starting point for the modeling phase, the previously translated texts and their contextualized embeddings that were generated are used. The included steps of the modeling phase are first the reduction of the embedding vectors, which are available in a high dimensional manner. These are reduced by UMAP so that further algorithms can be applied to them in a target-oriented fashion. Subsequently, the clustering of semantically similar texts follows as well as the derivation of relevant words of each cluster, which finally define and describe the clusters as a topic. The sub-steps of the modeling process are shown in Figure 5.

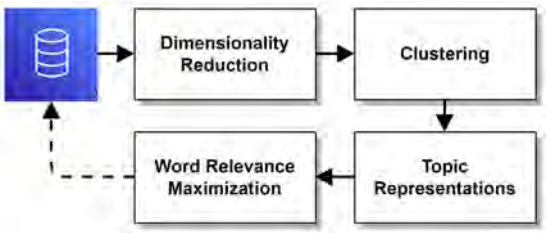


Figure 5: Modeling process with corresponding sub-steps

Again, the dashed line denotes a feedback of the enriched data to the overall database. The setup basically represents the proposed concept of Grootendorst (2022).

9.3 Deployment

The data processing and corresponding algorithms described during the previous chapter were implemented and operationalized in the course of this work. A simplified overview of the deployed solution with its respective layers and the corresponding systems or services is seen in Figure 6.

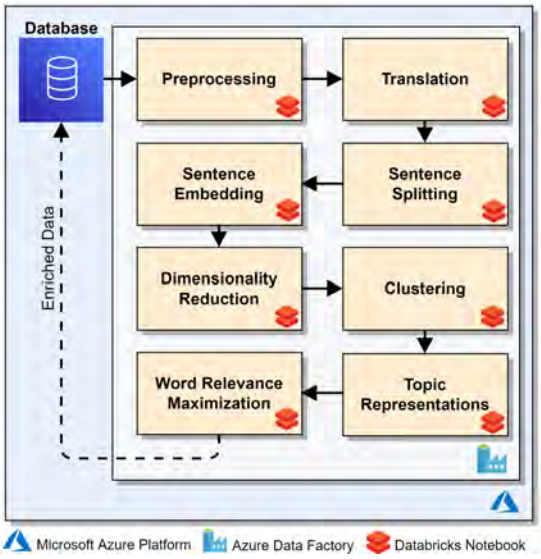


Figure 6: Operationalized process within the Microsoft Azure cloud platform

As a central basis, the solution is built on the Microsoft Azure platform, where the database is also located or connected to. Individual customer feedback texts contained therein are subsequently selected and fed to the processing steps. The respective steps and the necessary source code are provided each as a Databricks Notebook service. This modular design allows for targeted maintenance and customization, with individual notebooks being edited and reviewed at a time. The Azure Data Factory service is used to orchestrate the notebooks and assign corresponding computing instances for processing. Depending on the computational effort, the instances here can also be adjusted either manually or automatically. After each step, newly generated information is transferred to the next step. After all steps have been completed, the original database is enriched by the additional data generated and is thus available for further applications and data consumers.

10 EVALUATION

After the structure of the implementation was presented in the previous chapter, the results will now be further investigated. This includes a general reflection on the observed results and practical benefits using a real-world example. With the second part, insights and thoughts on the evaluation of quality will be shared. With the evaluation of scalability and performance, the key differences as well as potentials compared to an existing solution are addressed. Furthermore, results are presented from an experimental evaluation in which parallel processing was used with the goal of proving a particular hypothesis as well as the introduction of GPU acceleration.

10.1 Evaluation of the Observed Results

With the presented configuration of the system, a sentence is considered as the smallest unit of information. This assumption is mainly based on the properties and the explorative investigation of the underlying data. The assumption is that each sentence may address one topic at a time, but not necessarily. Such an approach may differ depending on the application and should therefore be considered individually. To get an impression of the result, Figure 7 shows an exemplary text as well as the assignment of the topics for each sentence from the model. It should be noted that the third sentence (gray coloring) is a data point that was not assigned to a valid cluster or was identified as noise. For the output of the topics or labels shown underneath, the top 3 words were determined in each case by the associated topic representation.

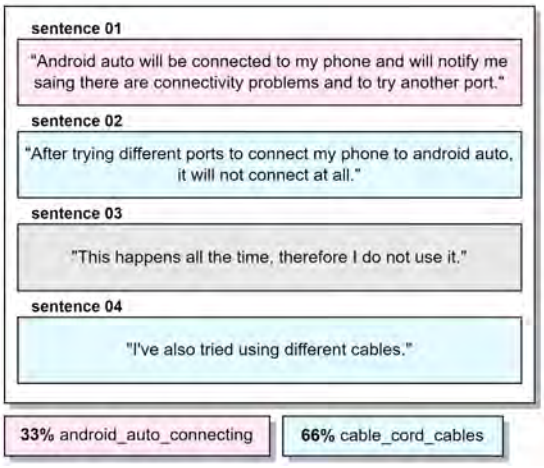


Figure 7: Sample text split into its individual sentences as components as well as the automatically generated and assigned topics and their proportions

By splitting the customer feedback into the sentences contained in it, several clusters or topics can therefore be assigned to one customer feedback under certain circumstances. In the presented use case and the underlying data, this is a valid method, since it concerns short texts with comparatively few sentences and additionally a fine-grained analysis is to be made possible. However, if the character of the data changes, this procedure should be adapted under given conditions according to the CRISP-DM logic. This is especially true if the length of the texts would increase significantly and consequently the number of sentences contained in them would increase as well.

10.2 Evaluation of Quality

Evaluating the quality of an unsupervised and exploratory approach is challenging. While there are a number of known extrinsic measures to quantify the quality of a clustering, there are also limitations and concerns. At the start of this work, extensive test data (about 20,000 records) were collected. The data thus obtained came from various departments such as research and development as well as quality management, which had previously labeled the customer feedback manually.

Using established evaluation criteria, comparative measurements of human-generated and machine-generated clusters were to be carried out. However, it turned out that metrics varied very inconsistently and thus did not contribute to a clear statement or insight about the overall quality. It was particularly interesting to observe that individual clusterings (results from individual runs) were significantly worse in a quantitative assessment of quality than in a qualitative evaluation by individual experts during various interviews. Here, experts or end-users who consume the enriched data in the form of a dashboard and use it for their daily work to derive insights from it were interviewed. A major assumption of this observation is that the diversity of backgrounds of individual experts in labeling the data leads to different assessments, not at least due to the subjective perception of the individual. This conjecture suggests that the quantitative assessment of a clustering by unsupervised methods may well be more complex than initially thought. For this reason, the decision has been taken to discontinue further investigation of an evaluation method based on metrics, as this would exceed the scope of this paper. Nevertheless, it is important that it is clearly advocated and recommend for continuous monitoring of quality through, i.e., expert interviews and the resulting evaluation in qualitative terms. Gibbons et al. (2008) emphasizes the importance of such a methodology:

> "This is especially important when the user base gradually expands. The problem is not primarily technical, finding the right kind of methods or establishing the correct procedures. Rather, it is one of the learning to handle complexity, developing procedures which leave room for experimental planning and preserving opportunities for feedback in order to allow intervention in time to change the course of events, if that is necessary. The number of participants in knowledge production is increasing as the number of "centers" addressing a particular problem. Further, the whole process is sensitive to the changes that inevitably occur in social, economic and technical environments." (p. 66)

It is furthermore important since the approach can deliver changing results due to stochastic components and changes in the database. The chosen approach should also provide this characteristic in order to be able to react to new topics or clusters within the data from a technical perspective. With a continuous feedback mechanism by the end-users, it is ensured that changes in the data do not lead to impairments or that early interventions can be taken.

10.3 Evaluation of Scalability and Performance

For a detailed assessment of scalability and performance, the process was not considered as a whole, but was divided into its essential components. Thus, the processing time in minutes was measured from the beginning to the end of each step. For the individual analysis, the process was divided into the steps preprocessing, language detection, translation, sentence splitting, embedding and clustering. It is worth mentioning that the clustering step encompasses all further steps up to the creation of the topics. In the initial setup, a single-node computing instance with the specifications seen in Table 1 was used for this purpose.

Table 1: Initial single-node CPU computing instance configuration (pricing information taken from Microsoft (2022))

Instance Name	vCPU (cores)	RAM (GByte)	Cost (€/hour)
D64a_v4	64	256	3.4999

To validate the scalability as well as the performance of the approach, the entire dataset was processed in the first run. With a randomized 50% sample of the data, a second measurement was subsequently performed to achieve comparability; results obtained are shown in Table 2. Since the database contains texts in different languages and lengths, the 50% sample was drawn in a randomized manner.

Table 2: Overview of process steps and corresponding durations measured on the single-node CPU (rounded values)

Process Step	Full Sample (min.)	50% Sample (min.)	Change (%)
Preprocessing	1	0.5	-50
Language Detect.	74	35.4	-48
Translation	88.3	43.9	-50
Sentence Split.	11.9	5.7	-48
Embedding	16	8	-50
Clustering	6	3	-50
Total	197.2	97	-49

As mentioned in chapter 9.1 Data Preparation, texts that are already available in English are not forwarded to the translation process but only retained. Depending on the order of the data, this would mean that a corresponding sample could possibly lead to a biased result in the time measurement, since the texts may not even need to be translated or differ in length. The almost linear progression of data quantity in relation to the required processing time can be described with the time complexity notation $O(n)$. It is assumed that the processing time could probably be further reduced, for example, with the use of a more powerful computing instance. Steps like language detection, translation, sentence splitting and embedding offer a starting point for verifying parallelized processing. This is a valid hypothesis since the data can still be considered independently at this stage.

For the comparison with the predecessor solution and its approach using local representations only, the processing time of old versus new was benchmarked. It is referred to the same dataset for each approach and to an estimated processing time of 36-48 hours[1] of the predecessor approach. In comparison with the values shown in Table 2, time savings by a rounded value of 91-93% are determined. This value sounds very high at first and raises questions, but it can be justified. With the previous system, mainly local representations of the individual texts were implemented. With the use of such concepts like one-hot vector or Bag of Words model, the size of a representation is directly proportional to size of the vocabulary, and most real-world corpora have large vocabularies. This results in a sparse representation where most of the entries in the vectors are zeroes, making it computationally inefficient to store and compute with (Vajjala et al., 2020, p. 86). Transformer models, on the other hand, provide dense representation vectors with a fixed size and already contribute a large proportion of the performance improvements. In addition, the new approach uses other state-of-the-art algorithms that are characterized by their efficiency and scalability. These arguments serve as substantial justification and reinforce the validity of the values presented.

10.4 Evaluation of Parallelization

Following the evaluation of scalability and performance in the previous chapter, some assumptions arose from the consideration of the two disproportionately large shares of processing times for the identification of the language and the translation compared to the other steps. These led to formulate a hypothesis for parallelization, which will be validated through this chapter. Since the two sub-process mentioned are small and independent data processing steps, it seems obvious that further optimization in terms of time and cost-effectiveness could be possible by using a parallelized mode of operation. In fact, this applies for the two steps described, but has not been possible so far due to the use of a single-node computing instance. While initially all processing steps were executed on a single-node cluster with very high performance specifications, individual steps that require less computing power may well run in parallel on less powerful cluster instances with adapted specifications. The mentioned specifications of the single-node cluster are nevertheless justified, steps like the reduction of the dimensionality make it necessary here for example that all data is loaded into the memory and processed at

[1] Based on an estimations taken from expert interviews.

the same time.

For the experimental setup, an instance suitable for parallelization was used for this purpose in order to perform another measurement of the processing times for the stated steps. At this point, one instance serves as a driver and other instances as workers, which are assigned the operations in order to utilize their threads optimally. An overview of the computing cluster-instance setup is seen in Table 3. It should be noted that the amount of worker instances is flexibly defined, so that they can be automatically scaled between a minimum of two and a maximum of five instances if needed. Furthermore, the costs per hour shown refer to one instance unit of each type described.

Table 3: Parallel CPU computing instance configuration (pricing information taken from Microsoft (2022))

Instance Name	vCPU (cores)	RAM (GByte)	Amount	Unit cost (€/hour)
D8s_v3 (driver)	8	32	1	0.4565
DS3_v2 (worker)	4	14	2-5	0.2587

With lower cluster-instance specifications such as virtual CPU cores or RAM, the costs for the corresponding instance type drop compared to the originally used single-node cluster setup. However, it is much more interesting to see the major impact on the processing time required for the two process steps. A significant advantage of parallelized processing comes from the optimized distribution of iteration chunks on multiple threads. Although less computing power is available, the large number of iterations is distributed and therefore executed faster. As shown in Table 4, this can result in even more efficient use through more cost-effective resources.

Table 4: Comparison of single-node CPU and driver-worker CPU computing instance setup (rounded values)

Process Step	Single-Node Cluster (min.)	Parallel Cluster (min.)	Change (%)
Lang. Det.	74	6.5	-91.2
Translation	88.3	7.6	-91.4
Total	162.3	14.1	-91.3

Although there was the option of automatically scaling up to three additional instances, this did not occur or was not necessary. The fact that only two instances were utilized is not obvious, but still worth mentioning. With a reduction in processing times for the two presented steps of more than 90%, it is clear how important and urgent it is to further optimize the resources involved. Moreover, the generally lower costs for the adapted instances or the selected setup have another positive effect. The hypothesis of parallelization has thus been confirmed, and at the same time, an example of tangible improvements with the experiment performed was provided.

10.5 Evaluation of GPU Acceleration

To investigate a further increase in efficiency, the use of Graphics Processing Units (GPU) was considered. While neural networks and the associated tensor operations make GPU acceleration highly utilizable, this is only supported to a limited extent by some well-known algorithms. With the open-source GPU ecosystem of Rapids Development Team (2018), several algorithms are developed and offered as GPU implementations. This evaluation focuses in particular on the process steps ”Embedding” and ”Clustering”. While the GPU hardware can be used directly for creating vector space embeddings by transformer models, the RAPIDS ecosystem is utilized for the clustering phase and the associated algorithms. Two types of GPU computing instances are used here, their specifications are shown in Table 5.

Table 5: GPU accelerated computing instance configuration (pricing information taken from Microsoft (2022))

Instance Name	vGPU (cores)	RAM (GByte)	Amount	Unit cost (€/hour)
NC12	2	112	2	2.2188
NC6_v3	1	112	2	3.6359

Measured in terms of cost per hour and compared to the original single-node CPU used, differences can be seen here. Although the NC6_v3 instance is slightly higher in cost, the advantages in terms of processing time become clear when looking at Table 6. Significant reductions in processing time of 80% and 82.5% can be seen.

This observation becomes more interesting when looking at the total computing costs in Table 7 compared to the originally used single-node CPU computing instance. Here, too, savings of 74.6% and 61.9% are possible using the given data set as well as GPU accelerated processes. Considering both process steps, on average about 71%

Table 6: Comparison of single-node and GPU computing instance setup bases on processing time (rounded values)

Process Step	Single-Node Cluster (min.)	GPU Cluster (min.)	Change (%)
Embedding	16	3.2[a]	-80
Clustering	6	1.1[b]	-82.5
Total	22	4.3	-80.5

(a) measured on NC_12 (b) measured on NC6s_v3

of the costs can be saved. The urgency of selecting and specifying hardware appropriately becomes clear once more.

Table 7: Comparison of single-node and GPU computing instance setup based on total computing costs (rounded values)

Process Step	Single-Node Cluster (€)	GPU Cluster (€)	Change (%)
Embedding	0.93	0.23[a]	-74.6
Clustering	0.35	0.13[b]	-61.9
Total	1.28	0.36	-71.2

(a) measured on NC_12 (b) measured on NC6s_v3

11 CONCLUSION

With the progress of digitalization, the volume and speed of data flowing into business processes continue to grow. Given the large and constantly increasing volume as well as the lack of structure for the data, this ties up a great deal of capacity as well as resources. "Without appropriate tools to manage data across silos, organizations can expect an increase in management overheads while also missing out on valuable data insights" (Potnis, 2018, p. 3). The key here are efficient methods for processing customer feedback or texts in general, so that valuable insights and conclusions can be derived from it. Based on this, enriched data and information sources were created, which enable stakeholders to find and use relevant information in an uncomplicated and task-oriented manner. As part of this, the data may be used to build further applications such as interactive dashboards or systems and thus consume the data product for individual purposes. All of this contributes to making business processes more efficient and to sustainably increasing customer satisfaction as well as product quality.

In this paper, a holistic approach to optimize the customer understanding with the aid of Artificial Intelligence as well as Cloud Computing is presented. Based on theoretical work as well as the research of a variety of authors, synergies were leveraged and concepts put into business-oriented practice. First, it was shown how data in textual form is cleaned and transformed into a unified language through the use of machine translation. A state-of-the-art transformer model was then used to generate numerical representations of the texts, while semantic relationships were preserved within these. This was followed by dimensionality reduction using a new and efficient algorithm and the associated clustering using a density-based algorithm. For a humaninterpretable labeling of the clusters, further algorithms for the generation of a topic representation as well as the specification of individual descriptive words were performed.

Since the approach was developed and implemented on an existing enterprise platform, the solution can be immediately utilized in production. Here, the Microsoft Azure cloud platform proved to be the medium of choice. Based on this, the source code can be managed modularly by the Databricks Notebook service and orchestrated for various scenarios of use by the Azure Data Factory service. Thus, scaling and adaptation of deployed computing resources is ensured for further use cases or the expansion of the application. This is particularly useful when additional stakeholders want to leverage the solution with additional data.

A major challenge in this work has been the assessment of quality regarding the clustering results. Although a wide range of quality metrics is available for the evaluation of a clustering, this cannot always be used as intended. At the beginning, about 20,000 test data records from various departments were collected in order to compare the cluster results and to assess their quality using various known metrics. Strongly fluctuating and varying results leave reason to assume that the subjective perception of individuals leads to different interpretations. Apart from the quantitative results, qualitative assessments were determined in the course of expert workshops. Most interestingly, these assessments by experts were significantly better and there was a clear message that the results were valid. Once again, the statement of Gibbons et al. (2008) should be emphasized here:

> "The problem here is not primarily technical, finding the right kind of methods or establishing the correct procedures. Rather, it is one of the learning to handle complexity, developing proce-

> dures which leave room for experimental planning and preserving opportunities for feedback in order to allow intervention in time to change the course of events, if that is necessary." (p. 66)

Following on from these findings, it is clearly recommend to continuously tracking the quality through the involvement of experts and their feedback in order to track the quality of such an unsupervised approach. Feedback iterations can ensure that the quality is guaranteed in the long term and it does not develop unintentionally in the context of changes within the underlying data.

With the evaluation of scalability and performance, it was possible to show that using state-of-the-art technologies and algorithms, a quite significant increase in efficiency can be brought about in comparison to a predecessor system. During this work, it was referred to a comparatively small dataset with about 215,000 records. However, with the knowledge gained about scalability, it is assumed that the performance behaves similarly with extended data. Further positive effects have been shown during the evaluation of parallelization. At this point, it also became clear that special care should be taken when selecting computing instances with respect to specific processing steps and corresponding requirements. Furthermore, this was underlined by the investigation of acceleration and cost reduction through the use of GPU architecture. With the appropriate selection and adaption to a given use case, significant savings can be achieved.

References

Aggarwal, C. C. (2014). An introduction to cluster analysis. In C. C. Aggarwal & C. K. Reddy (Eds.), *Data clustering*. CRC Press.

Angelov, D. (2020). Top2vec: Distributed representations of topics. https://doi.org/10.48550/ARXIV.2008.09470

Blei, D. M. (2012). Probabilistic topic models. *Communications of ACM*, (4), 77–84.

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In D. Hutchison, T. Kanade, & J. Kittler (Eds.), *Advances in knowledge discovery and data mining* (pp. 160–172). Springer.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *Crisp-dm 1.0: Step-by-step data mining guide*.

Coenen, A., & Pearce, A. (2019). Understanding umap (Google Pair, Ed.). Retrieved December 13, 2021, from https://pair-code.github.io/understanding-umap/

DeepAI. (2021). Distributed representations. Retrieved December 1, 2021, from https://deepai.org/machine-learning-glossary-and-terms/distributed-representation

Delen, D. (2020). *Predictive analytics: Data mining, machine learning and data science for practitioners: Data mining, machine learning and data science for practitioners* (2nd ed.). Pearson FT Press.

Foster, D. (2019). *Generative deep learning: Teaching machines to paint, write, compose, and play* (1st ed.). O'Reilly.

Géron, A. (2019). *Hands-on machine learning with scikit-learn, keras, and tensorflow: Concepts, tools, and techniques to build intelligent systems* (2nd ed.). O'Reilly.

Gibbons, M., Limoges, C., Nowotny, H., Schwartzmann, S., Scott, P., & Trow, M. (2008). *The new production of knowledge: The dynamics of science and research in contemporary societies*. SAGE Publications.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. https://doi.org/10.48550/ARXIV.2203.05794

Isson, J.-P. (2018). *Unstructured data analytics: How to improve customer acquisition, customer retention, and fraud detection and prevention*. Wiley. https://doi.org/10.1002/9781119378846

Liu, Z., Lin, Y., & Sun, M. (2020). *Representation learning for natural language processing* (1st ed.). Springer.

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. https://doi.org/10.48550/ARXIV.1802.03426

Microsoft. (2022). Windows virtual machines pricing. Retrieved June 9, 2022, from https://azure.microsoft.com/en-

us / pricing / details / virtual - machines / windows/#pricing

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. https://doi.org/10.48550/ARXIV.1301.3781

Patel, A. A. (2019). *Hands-on unsupervised learning using python: How to build applied machine learning solutions from unlabeled data* (1st ed.). O'Reilly.

Phi, M. (2020). Illustrated guide to transformers: Step by step explanation. Retrieved June 17, 2022, from https://towardsdatascience.com/illustrated-guide-to-transformers-step-by-step-explanation-f74876522bc0

Potnis, A. (2018). Illuminating insight for unstructured data at scale. Retrieved December 4, 2021, from https://www.ibm.com/downloads/cas/Z2ZBAY6R

Rapids Development Team. (2018). Rapids: Collection of libraries for end to end gpu data science. Retrieved July 26, 2022, from https://rapids.ai

Sarkar, D., Bali, R., & Sharma, T. (2018). *Practical machine learning with python: A problem-solver's guide to building real-world intelligent systems*. Apress.

TensorFlow. (2021a). Transformer model for language understanding. Retrieved November 18, 2021, from https://www.tensorflow.org/text/tutorials/transformer

Vajjala, S., Majumder, B., Gupta, A., & Surana, H. (2020). *Practical natural language processing: A comprehensive guide to building real-world nlp systems* (1st ed.). O'Reilly.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. https://doi.org/10.48550/ARXIV.1706.03762

Zheng, A., & Casari, A. (2018). *Feature engineering for machine learning: Principles and techniques for data scientists* (1st ed.). O'Reilly.

Vergleich von Process Mining Methoden bei praxisnahen, verteilten Prozessen im Unternehmensumfeld

Simon Kopold

Hochschule Pforzheim

Master Information Systems
Tiefenbronner Straße 65
75175 Pforzheim
simon_kopold@web.de

Ramona Becker

SEEBURGER AG

BIS Application Development
Edisonstraße 1
75015 Bretten
E-Mail: r.becker@seeburger.de

Frank Morelli

Hochschule Pforzheim

Master Information Systems
Tiefenbronner Straße 65
75175 Pforzheim
frank.morelli@hs-pforzheim.de

Schlüsselwörter
Process Mining, Multi-Event-Log, objekt-zentriertes Process Mining, Business Integration, Digitalisierung, Prozessoptimierung

Problemstellung
Process Mining gilt als aufstrebende Methode um digitalisierte Prozesse zu erfassen und zu optimieren. Die SEEBURGER AG bietet für ihre Kunden eine Vielzahl an technischen Integrationsmöglichkeiten in Form der hauseigenen Business Integration Suite (BIS) und versteht sich in diesem Sinne als Treiber der Digitalisierung. Im Rahmen einer Masterarbeit wird untersucht, welche Mehrwerte durch Process Mining bei der SEEBURGER AG erzeugt werden können. Besonders bei übergreifenden, komplexen Prozessen, welche über mehrere Systeme verteilt sind, greift der klassische Process Mining Ansatz zu kurz. Mit der Multi-Event-Log (MEL) Funktion von Celonis sollen auch mehrere verteilte Prozesse mittels Process Mining in der Unternehmenspraxis erfasst und analysiert werden können. Im Rahmen der Masterarbeit wird klassisches Process Mining sowie MEL in verschiedenen Einsatzgebieten der SEEBURGER AG angewendet. Durch einen Ausblick auf objekt-zentriertes Process Mining wird eine weitere Methodik aus der Wissenschaft beleuchtet.

Klassisches Process Mining
Um (klassisches) Process Mining anwenden zu können benötigt man ein Event-Log mit Fall-ID, Aktivitäten und Zeitstempel als Basisdaten (Peters und Nauroth 2019, S. 15). Bei einer Bestellung lässt sich beispielsweise die Bestellnummer als Fall-ID, „Bestellungseingang“ oder „Bestellung absenden“ als Aktivitäten und der Zeitpunkt, an dem eine Aktivität begonnen wird, als Zeitstempel interpretieren (vgl. linke Teiltabelle in Abbildung 1). Mit diesen Kenngrößen kann Process Mining grundsätzlich angewendet werden und der tatsächliche Prozessverlauf sowie zeitliche Zusammenhänge und Engpässe erkannt werden. Durch zusätzliche Daten (z.B. Kundenummer, Priorität, …), ist es möglich, die Analyse anzureichern (Peters und Nauroth 2019, S. 15).

Process Mining bei praxisnahen Prozessen
Die dargestellte Datenstruktur sieht genau eine Fall-ID je Fall vor, welche eindeutig über den Gesamtprozess definiert ist und beschreibt damit „klassisches“ Process Mining. Betrachtet man einen ganzheitlichen Einkaufsprozess von der Bestellung der Ware, über die Lieferung mit anschließenden Rechnungserstellung, muss über alle Prozessschritte hinweg ein eindeutiger Identifizierer vorhanden sein, um klassisches Process Mining anwenden zu können. Im vorliegenden Beispiel ist die Bestellnummer eine potenzielle Fall-ID des Gesamtprozess, welche auch bei der Lieferung und der Rechnung angegeben sein muss. In der Praxis lässt sich dies nicht immer umsetzen, da die Daten für Process Mining aus verschiedenen Systemen mit verschiedenen Identifizierern (Fall-ID) stammen. Bezogen auf den obigen Bestellprozess ist es möglich, dass im Lieferschein anstelle der ursprünglichen Bestellnummer die Rechnungsnummer angegeben wird. Dies ist vor allem für Teillieferungen und mehreren (Teil-) Rechnungen typisch. Im vorliegenden Fall wird angenommen, dass ein Lieferant je (Teil-)Lieferung eine Rechnung für die tatsächlich gelieferte Ware ausstellt. Für dieses einfache Beispiel ist es denkbar aus den unterschiedlichen Identifizierern (Bestell-, Lieferschein- und Rechnungsnummer) eine zentrale Fall-ID zu generieren (van der Aalst 2020). Eine solche Modellierung scheitert jedoch, wenn obiger Beispielprozess nicht mehr linear ist: Dieser Sachverhalt tritt beispielsweise ein, wenn Lieferungen fehlerhaft sind, Bestellungen storniert oder Rechnungen für mehrere Lieferungen gesammelt ausgestellt werden. Solche Einzelfälle lassen sich nicht durch Modellierung im Vorfeld abgebilden und damit nicht mit „klassischem“ Process Mining erfassen. Mögliche Fehlerquellen bestehen darin, dass ein Event (das tatsächliche Auftreten einer Aktivität) fälschlicherweise mit mehreren Fällen in Verbindung gebracht und somit mehrmals im Event-Log aufgeführt wird (Konvergenz). Weiterhin kann eine Aktivität fälschlicherweise mehrmals innerhalb eines Falls auftreten (dabei handelt es sich um eine sogenannte „Divergenz“) (van der Aalst 2020). Als Folge repräsentiert das Event-Log nicht mehr die Realität.

Celonis Multi-Event-Log

Celonis (Celonis 2021) bietet mit dem Multi-Event-Log (MEL) eine Funktion an, welche es Unternehmen erlaubt mehrere „klassische“ Event-Logs ohne gemeinsame Fall-ID zu verbinden, sodass ein Gesamtprozess aus mehreren Event-Logs dargestellt werden kann (Celonis 2020). Aus den Untersuchungen der Masterthesis geht hervor, dass hierbei nach wie vor „klassische“ Process Mining Algorithmen zum Einsatz kommen, da im Rahmen der Modellierung die Betrachtungsperspektive auf einer vordefinierten Identifikationskenngröße liegen muss. Konvergenz und Divergenz können hierdurch weiterhin auftreten.

MEL erlaubt das Definieren von mehreren Event-Logs mit eigner Fall-ID, Zeitstempel und Aktivität sowie Referenzen zu anderen Event-Logs. Abbildung 1 stellt einen ganzheitlichen Bestellprozess aus Sicht eines Lieferanten von Bestellungs- bis zum Zahlungseingang dar. Anstelle eines zentralen Event-Logs mit einer zentralen Fall-ID sind hier drei einzelne Event-Logs mit einzelnen Fall-IDs aufgeführt. Die Event-Logs sind über Referenzen miteinander in Beziehung gesetzt. Die linke Tabelle beschäftigt sich ausschließlich mit den internen Prozessen bis zum Versand einer Bestellung. Über die Bestellnummer, welche hier auch als Fall-ID der linken Tabelle fungiert, sind die zugehörigen Rechnungen verknüpft: Bestellung *1002* wird hier auf zwei getrennte Rechnungen (*2002* und *2003*) aufgeteilt. Grund hierfür ist, dass die Bestellung in zwei Teillieferungen unterteilt und je Lieferung eine Rechnung erstellt wird. Der Zahlungseingang seitens des Kunden bezieht sich auf eine konkrete Rechnung basierend auf dem Lieferschein und damit nicht direkt auf die ursprüngliche Bestellnummer. Durch MEL ist es trotzdem möglich den Prozessablauf von Bestellungseingang über die Lieferung mit anschließender Zahlung nach Rechnungseingang durch Process Mining zu analysieren, ohne vorab eine zentrale Fall-ID zu generieren. Hierfür wird vom Anwender gefordert, dies explizit im Celonis-Modell zu definieren. Eine automatische Erkennung der Zusammenhänge ist nicht möglich. Der menschliche Modellierer muss als Vorgabe eine Gesamtperspektive für die Prozessabläufe wählen. Je nach Ausführung können dadurch ungenaue Prozessmodelle entstehen.

Abbildung 1: Bestellprozess bis Zahlungseingang als MEL-Konzept; eigene Darstellung

Objekt-zentriertes Process Mining

Die händische Modellierung und damit verbundene Wahl der Perspektive bei MEL soll bei objekt-zentriertem Process Mining ebenso wie Konvergenz und Divergenz entfallen (van der Aalst und Berti 2020). Wil van der Aalst benennt objekt-zentriertes Process Mining als eines der wichtigsten Zukunftsfelder in diesem Bereich (van der Aalst 2021).

Objekt-zentriertes Process Mining benötigt nach van der Aalst und Berti ein objekt-zentriertes Event-Log, welches anschließend mittels eigener objekt-zentrierter Process Mining Algorithmen analysiert und daraus ein objekt-zentriertes Petri-Netz erzeugt wird (van der Aalst und Berti 2020, S. 7 ff.). Der Vorteil dieses Verfahrens besteht darin, dass ganzheitliche Prozessmodelle erzeugt werden können, die Konvergenz und Divergenz vermeiden, obwohl keine durchgängige Fall-ID vorhanden ist (van der Aalst und Berti 2020, S. 35). Die Umsetzung dieses Verfahren ist aktuell mittels einer zusätzlichen Erweiterung für das Python Process Mining Modul PM4PY möglich (Javert899) und wird bei der SEEBURGER AG in einer weiteren Thesis untersucht.

Literaturverzeichnis

Celonis (2020): Celonis setzt neue Maßstäbe für das Execution Management. Online verfügbar unter https://www.celonis.com/de/press/celonis-raises-the-bar-for-execution-management-as-only-solution-to-enable-optimization-of-multiple-interconnected-processes, zuletzt geprüft am 30.11.2021.

Celonis (2021): Celonis | Unternehmen. Online verfügbar unter https://www.celonis.com/de/company/, zuletzt geprüft am 19.11.2021.

Javert899: pm4py-mdl. Online verfügbar unter https://github.com/Javert899/pm4py-mdl, zuletzt geprüft am 25.01.2022.

Peters, Ralf; Nauroth, Markus (2019): Process-Mining. Geschäftsprozesse: smart, schnell und einfach. Wiesbaden: Springer Fachmedien Wiesbaden (essentials).

van der Aalst, Wil (2020): Object-Centric Process Mining: Dealing With Real-Life Processes. Online verfügbar unter https://blog.rwth-aachen.de/pads/2020/10/09/object-centric-process-mining-dealing-with-real-life-processes/, zuletzt geprüft am 30.11.2021.

van der Aalst, Wil (2021): The 2021 Celonis Ecosystem Summit. Celonis. München, 21.09.2021. Online verfügbar unter https://www.celonis.com/ecosystem-summit/, zuletzt geprüft am 21.10.2021.

van der Aalst, Wil M. P.; Berti, Alessandro (2020): Discovering Object-Centric Petri Nets. Online verfügbar unter http://arxiv.org/pdf/2010.02047v1.

Konzeption und Integration eines Chatbots in Moodle

Emilia Kunowsky

Hochschule für Technik und Wirtschaft Berlin

Fachbereich 4
Treskowallee 8
10318 Berlin
emilia.kunowsky@web.de

Prof. Dr. Verena Majuntke

Hochschule für Technik und Wirtschaft Berlin

Fachbereich 4
Treskowallee 8
10318 Berlin
verena.majuntke@htw-berlin.de

Prof. Dr. Birte Malzahn

Hochschule für Technik und Wirtschaft Berlin

Fachbereich 4
Treskowallee 8
10318 Berlin
birte.malzahn@htw-berlin.de

Kategorie

Bachelorarbeit

Schlüsselwörter

Chatbot, Chatbot Builder, Moodle, Umfrage, Integration

Zusammenfassung

Chatbots haben sich als hilfreiches Tool bei der Interaktion mit Nutzer*innen bewährt und werden heutzutage in zahlreichen Bereichen eingesetzt, wie zum Beispiel als interaktive FAQs (vgl. Celina 2021). Vorteile von FAQ-Chatbots sind unter anderem eine ständige Erreichbarkeit sowie eine gleichbleibende Freundlichkeit beim Beantworten von sich wiederholenden Fragen (vgl. Celina 2021).
Das Ziel der Bachelorarbeit war es, einen Chatbot für das Modul „Grundlagen der Programmierung“ aus dem Studiengang Wirtschaftsinformatik an der HTW Berlin zu entwerfen und diesen in den zugehörigen Kurs der E-Learning-Plattform Moodle zu integrieren. Das Modul „Grundlagen der Programmierung“ ist für das erste Semester vorgesehen und wird daher von einer Vielzahl an Studierenden belegt, die neu an der HTW Berlin sind. Der Chatbot sollte daher als zentrale Anlaufstelle für Fragen dienen und alle vorhersehbaren Fragen der Studierenden beantworten, um die Anzahl der an Lehrkräfte gesendeten E-Mails zu reduzieren. Neben der Entlastung der Lehrkräfte bedeutet dies gleichzeitig auch eine große Zeitersparnis für Studierende. Um die aus Sicht der Studierenden wichtigsten Themengebiete für den Chatbot zu ermitteln, wurde eine Umfrage durchgeführt. Diese ergab, dass Studierende sich Informationen zu allen organisatorischen, strukturellen und inhaltlichen Aspekten des Moduls wünschen und des Weiteren zusätzliche Übungs- und Lernmöglichkeiten in einem Chatbot als hilfreich erachten würden.
Basierend auf den Umfrageergebnisse wurden Anforderungen definiert, die als Kriterien zur Auswahl des geeignetsten Chatbot Builders dienten. Anhand einer Bewertungstabelle mit diesen sieben Kriterien – *verfügbare Gestaltungselemente, Individualisierbarkeit des User Interface, Funktionsumfang des Dashboards, Intuitivität der Bedienung, Übersichtlichkeit sowie Kosten und Einbindbarkeit in Moodle* – wurden die drei Chatbot Builder *Landbot*, *Tidio* und *BotPenguin* gegeneinander abgewogen (vgl. Kaiser et al. 2019, Khan / Das 2018). Der Chatbot Builder *Tidio* gewann aufgrund der intuitiven Bedienbarkeit, des kostenlosen Basis-Accounts und des Elements „Kartennachricht“, welches das Erstellen einer Themenübersicht in einem Chatbot vereinfacht (vgl. Tidio).
Für den Chatbot Builder *Tidio* wurde darauffolgend das Konzept erstellt. Mit Hilfe des Elements „Kartennachricht“ wurde eine baumartigen Informationsstruktur des Chatbots erarbeitet. Die User*innen gelangen zuerst in eine Themenübersicht mit folgenden Themengebieten: *Allgemeine Infos, Bewertung, Kommunikation, Übungsmöglichkeiten und Erstie-Infos*. Mithilfe von Buttons können User*innen in eines der Themen hineinnavigieren, um sich konkrete Informationen ausgeben zu lassen.
Die geplante Vorgehensweise zur Realisierung des Konzeptes umfasste drei Phasen: die Erstellung des Chatbots, die Einbindung in Moodle und das Testen des Chatbots auf fehlerfreie Funktion durch die Entwicklerin. Das Einbinden in Moodle erfolgte mithilfe eines von *Tidio* zur Verfügung gestellten JavaScript Codes, der in den Moodle-Kurs eingefügt wurde.
Abschließend wurden Weiterentwicklungsmöglichkeiten des Chatbots erarbeitet, wobei unter anderem die Optimierung und Erweiterung des Chatbots auf Basis von Analysen und weiteren Umfragen betrachtet wurde, um eine optimale Unterstützung der Studierenden zu ermöglichen. Des Weiteren wurde die durch ein Feature in *Tidio* ermöglichte Verknüpfung des Chatbots mit weiteren zukünftigen Chatbots untersucht (vgl. Tidio).
Die Verknüpfung von Chatbots könnte es ermöglichen, ein Netzwerk spezialisierter Chatbots zu erstellen, die miteinander interagieren und User*innen somit immer an den für ihre Frage passenden Chatbot weiterleiten könnten. Beispielsweise könnte ein zentraler Chatbot auf der HTW-Website erstellt werden, der die User*innen je

nach ihren Bedürfnissen an spezialisierte Chatbots weiterleitet. So könnte ein Chatbot pro angebotenem Studiengang oder auch für organisatorische Themen wie den Bewerbungsprozess an der HTW Berlin erstellt werden.

Literatur

Celina (2021): Warum Chatbots auch 2021 immer wichtiger werden. assono GmbH. Kiel. Online verfügbar unter https://www.assono.de/blog/die-entwicklung-der-chatbots-der-erfolg-in-zahlen. Letzter Zugriff: 15.12.2022.

Khan, Rashid; Das, Anik (2018): Build Better Chatbots. A Complete Guide to Getting Started with Chatbots. Apress.

Kaiser, Markus; Buttkereit, Aline-Florence; Hagenauer, Johanna (2019): „Journalistische Praxis: Chatbots. Automatisierte Kommunikation im Journalismus und in der Public Relation“. Springer VS.

Tidio (o.J.): Automate sales with a powerful chatbot builder. Online verfügbar unter: https://www.tidio.com/chatbots/. Letzter Zugriff: 15.12.2022.

Einsatz künstlicher Intelligenz in der Medizin – Chancen und Herausforderungen

Dong Tuyen Nguyen

Technische Hochschule Mittelhessen

Fachbereich MND
Wilhelm-Leuschner-Str. 13
61169 Friedberg
E-Mail:
dong.tuyen.nguyen@mnd.thm.de

Prof. Dr. Harald Ritz

Technische Hochschule Mittelhessen

Fachbereich MNI
Wiesenstraße 14
35390 Gießen
E-Mail: harald.ritz@mni.thm.de

Prof. Dr. Frank Kammer

Technische Hochschule Mittelhessen

Fachbereich MNI
Wiesenstraße 14
35390 Gießen
E-Mail:
frank.kammer@mni.thm.de

Kategorie

Masterarbeit

Schlüsselwörter

Künstliche Intelligenz, Machine Learning, Big Data, Digitalisierung, Gesundheitswesen, Prognosen, Python, Entscheidungsbäume, Regressionsanalyse, Random Forest, Neuronales Netz, Data Augmentation

Zusammenfassung

Das Interesse und die Fortschritte bei medizinischen KI-Anwendungen sind in den letzten Jahren aufgrund der enorm gestiegenen Rechenleistung moderner Computer sowie der Zunahme digitaler Daten und ihrer Komplexität sprunghaft angestiegen. Der Einsatz kognitiver Technologien ermöglicht es Organisationen im Gesundheitswesen, auf riesige Mengen von Patientendaten zuzugreifen und diese zu verarbeiten.

KI-Technologien werden bereits in zahlreichen Bereichen des Gesundheitswesens eingesetzt, von Systemen, die bei Therapieentscheidungen helfen, über die Krebsdiagnose bis hin zur Forschung und Entwicklung von Medikamenten. Um das Potenzial der KI vollständig nutzen zu können, müssen jedoch auch die potenziellen Auswirkungen und Hindernisse verstanden werden.

Anhand von Analysen wurde die Effizienz des Einsatzes von KI in einigen medizinischen Anwendungsbereichen untersucht. Mit Hilfe der Regressionsanalyse und dem Random Forest konnte der Erfolg einer Warzenbehandlung bei Patienten mit hoher Wahrscheinlichkeit vorhergesagt werden. Allerdings ist es schwierig, aus der kleinen Stichprobe eine repräsentative Verteilung der Daten zu extrahieren, so dass es nicht möglich ist, zu beurteilen, ob es sich um ein aussagekräftiges Modell handelt.

Der Entscheidungsbaum wurde verwendet, um die Entscheidungswege zu visualisieren, die die KI zur Diagnose von Herz-Kreislauf-Erkrankungen durchläuft. Die Ergebnisse zeigen, dass der Bluthochdruck mit etwa 70% das wichtigste Merkmal ist, gefolgt von dem Alter und den hohen Cholesterinspiegel.

Ein neuronales Netz wurde für die Bildgebung erstellt, um Hirntumore in Röntgen-, CT- und MRT-Bildern zu erkennen. Durch Data Augmentation wurde die Anzahl der eindeutigen Trainingsdaten erhöht, wodurch die Genauigkeit des Modells trotz der bereits erreichten 97% auf 98% gesteigert werden konnte.

Damit die entwickelten Modelle genaue Ergebnisse liefern, müssen die Daten vorverarbeitet und auf mögliche diskriminierende Verzerrungen und Datensatzverschiebungen überprüft werden. Nur wenn eine hohe Datenqualität vorliegt, kann das Modell genaue Ergebnisse liefern. Falsche Vorhersagen können dazu führen, dass die falschen Medikamente und Behandlungen vorgeschlagen werden, was sich lebensbedrohlich auf die Gesundheit der Patienten auswirken könnte.

Die rechtlichen Rahmenbedingungen u.a. bezüglich der Sicherheit der gesammelten Daten müssen gewährleistet sein, da sonst auf die Gesundheitsdaten der Patienten zugegriffen werden kann und sie z.B. bei der Suche nach einem Arbeitsplatz oder einer Versicherung aufgrund ihrer Krankengeschichte diskriminiert werden könnten.

Eine Maschine kann die persönliche Beziehung und das Einfühlungsvermögen, das ein Arzt seinen Patienten entgegenbringt, nicht ersetzen. Aber sie kann anhand der Krankengeschichte, der persönlichen Vorlieben und der kulturellen Identität die Verhaltensmuster der Patienten untersuchen und daraus Rückschlüsse auf ihr künftiges Verhalten oder ihre Beweggründe ziehen, was zu einer besseren Arzt-Patienten-Beziehung führen kann.

Literatur

Becker, A.; Marcon, M.; Stocker-Ghafoor, S. et al.: Deep Learning in Mammography: Diagnostic Accuracy of a Multipurpose Image Analysis Software in the Detection of Breast Cancer, Invest Radiol. 2017 Jul; 52(7):434-440. PMID: 28212138

Kaul, V.; Enslin, S. & Gross, S. A.: History of artificial intelligence in medicine, American Society for Gastrointenstinal Endoscopy Volume 92, No.4, 807-812, New York, 2020

Kulikowski, C. A.: Beginnings of Artificial Intelligence in Medicine (AIM): Computational Artifice Assisting Scientific Inquiry and Clinical Art – with Reflections on Present AIM Challenges, AI Magazine Volume 6 Number 3, 122-134, 1985

Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H. & Ferrante, E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proc. Natl Acad. Sci. USA 117, 12592–12594, 2020, URL: https://www.pnas.org/doi/abs/10.1073/pnas.1919012117

Algorithmischer Handel

Nutzen von Machine Learning und künstlicher Intelligenz im Wertpapierhandel

Benedikt Ortwein

Technische Hochschule Mittelhessen

Fachbereich MND
Wilhelm-Leuschner-Straße 13
61169 Friedberg
benedikt.ortwein@mnd.thm.de

Prof. Dr. Harald Ritz

Technische Hochschule Mittelhessen

Fachbereich MNI
Wiesenstraße 14
35390 Gießen
harald.ritz@mni.thm.de

Prof. Dr. Oliver Hein

Technische Hochschule Mittelhessen

Fachbereich MND
Wilhelm-Leuschner-Straße 13
61169 Friedberg
oliver.hein@mnd.thm.de

Kategorie

Bachelorarbeit

Schlüsselwörter

Handel, Finanzmarkt, Machine Learning, Künstliche Intelligenz, Hochfrequenzhandel, Algorithmen

Zusammenfassung

Der Kursverlauf von Wertpapieren wird durch viele verschiedene Faktoren beeinflusst, der Mensch als Faktor emotionsgetriebener Entscheidungen spielt dabei eine wichtige Rolle. Mithilfe von künstlicher Intelligenz und automatisiertem Handel kann versucht werden, diesen Faktor beim Handeln zu minimieren und den Kauf und Verkauf von Wertpapieren zu automatisieren. Automatisierter Handel wird auch als algorithmischer Handel oder Algorithmic Trading bezeichnet. Dabei können der Automatisierungsgrad und die Strategie des eigenen Handelns in Form des Algorithmus definiert und verändert werden.

Durch die fortschreitende Digitalisierung der Finanzmärkte ist ein Umbruch zu erkennen, der zu einer erhöhten Liquidität und einer Beschleunigung von Finanztransaktionen führt. Ein wesentlicher Treiber ist der automatisierte Handel, welcher in den letzten Jahren immer stärker an Bedeutung gewonnen hat und bereits im Jahr 2012 schätzungsweise 85% aller Transaktionen auf dem Finanzmarkt ausmacht.

Dabei gibt es für die verwendeten Algorithmen keine Mindestkomplexität, sie können trivialen Wenn-Dann-Bedingungen entsprechen. Beispielsweise als Stopp Loss Order, welche bei Erreichen oder Unterschreiten eines Kurswerts eine Order ausführt. Für die Algorithmen können aber auch Ansätze aus den Bereichen des maschinellen Lernens, der Optimierung und der künstlichen Intelligenz genutzt werden, um beispielsweise Vorhersagen über die Entwicklung von Aktienkursen zu treffen.

Die automatisierte Art des Börsenhandels wird jedoch auch kritisch betrachtet, da sie zu sogenannten Flash Crashs beiträgt oder diese sogar auslösen kann. Auch bieten sich durch das automatisierte Handeln neue Möglichkeiten der Marktmanipulation, weshalb die Bundesanstalt für Finanzdienstleistungen seit dem 3. Januar 2018 eine Anzeigepflicht bei algorithmischem Handel eingeführt hat.

Im Rahmen der Bachelorarbeit soll die Frage beantwortet werden, ob und inwieweit künstliche Intelligenz in Verbindung mit dem automatisierten Handel für den einzelnen Marktteilnehmer zum Erfolg auf dem Finanzmarkt beitragen kann. Dafür werden zunächst die Grundlagen des Finanzmarktes und der psychologischen Verhaltensweisen von Menschen beim Handel mit Finanzprodukten herausgearbeitet.

Anschließend werden die künstliche Intelligenz und ihre Anwendungsmöglichkeiten im Bereich des automatisierten Handels näher betrachtet. Darauf folgen die Betrachtung und Evaluation verschiedener Modelle, basierend auf bestehenden Studien (siehe Tabelle 1).

Modell / Jahr	KI-Systeme	Zielsetzung
Modell 1 Mokhtari et al. 2021	ML Verfahren, Neuronales Netz mit LSTM	Voraussage von Aktienkursen, basierend auf technischer Analyse mit Machine Learning und Stimmungsanalyse mit der Klassifizierung von Stimmungen aus Twitter.
Modell 2 Mohan et al. 2019	Rekurrentes neuronales Netzwerk mit LSTM	Vorhersage von Aktienkursen als hybrides Modell aus technischer Analyse und Stimmungsanalyse.
Modell 3 Gehring et al. 1999	Zweistufiges neuronales Netz	Entscheidungsunterstützungssystem zur Maximierung der Rendite eines sog. Evoked Set mithilfe einer Risikonutzen Funktion.
Modell 4 Vaidya et al. 2015	Neuronales Netz	Entscheidungsunterstützungssystem unter Zuhilfenahme verschiedener Indikatoren und grafischer Aufbereitung der Ergebnisse für Endanwender.

Tabelle 1 - Übersicht der Modelle

Abschließend werden auch die Auswirkungen im allgemeinen Börsenumfeld betrachtet. Ergänzend dazu wird auf regulatorische Aspekte und Risiken des Handels mit Algorithmen eingegangen.

Innerhalb dieser Arbeit kann gezeigt werden, dass es potenziell möglich ist, Überrenditen am vorliegenden Aktienmarkt zu erzielen. Zum aktuellen Zeitpunkt gibt es jedoch kein ausgereiftes Modell, welches einem unerfahrenen und sonst börsentechnisch uninformierten Marktteilnehmer langfristige Überrenditen garantieren könnte. Börsenteilnehmer können jedoch bereits heute Modelle der künstlichen Intelligenz in ihre eigenen Analysen mit einbeziehen und so rationalere Entscheidungen unter dem Einfluss weniger Emotionen treffen.

Literatur

Buxmann, P. & Schmidt, H. (2021). Künstliche Intelligenz: Mit Algorithmen zum wirtschaftlichen Erfolg, Springer, Berlin Heidelberg.

Ferreira, F., Gandomi, A. & Cardoso, R. (2021). Artificial Intelligence Applied to Stock Market Trading: A Review. (9). IEEE Access.

Gehring, H. et al. (1999). Ein Entscheidungsunterstützungssystem zur Aktienanlage auf der Basis eines genetisch lernenden neuronalen Netzwerks. Hagen : FernUniversität in Hagen, 1999. URL: https://docplayer.org/14836917-Ein-entscheidungs-unterstuetzungssystem-zur-aktienanlage-auf-der-basis-eines-genetisch-lernenden-neuronalen-netzwerks.html

Mohan, S., Mullapudi, S., Sammeta, S., Vijayvergia, P. & Anastasiu, D. (2019). Stock Price Prediction Using News Sentiment Analysis. 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService), S. 205-208.

Mokhtari, S., Yen, K. K. & Liu, J. (2021). Effectiveness of Artificial Intelligence in Stock Market Prediction Based on Machine Learning. International Journal of Computer Applications 2021, Bd. 183, 7.

Vaidya, A. M., Waghela, N. H. & Yewale, S. S. (2015). Decision Support System for the Stock Market using Data Analytics and Artificial Intelligence. International Journal of Computer Applications. 08. Mai 2015, S. 21-28.

Konzeption und Entwicklung einer Datenpipeline zur automatisierten Validierung und Verarbeitung von Bankdaten am Beispiel der Mittelstand.ai

Stani Lennart Schlegel

Technische Hochschule Mittelhessen

Fachbereich MNI
Wiesenstr. 14
35390 Gießen
stani.lennart.schlegel@mni.thm.de

Prof. Dr. Harald Ritz

Technische Hochschule Mittelhessen

Fachbereich MNI
Wiesenstr. 14
35390 Gießen
harald.ritz@mni.thm.de

Dr. Michel Becker

Mittelstand.ai GmbH & Co. KG

Data Science
Schiffenberger Weg 110
35394 Gießen
michel.becker@mittelstand.ai

Kategorie

Bachelorarbeit

Schlüsselwörter

Data Engineering, Data Warehousing, Data Science, Apache Spark, Data Governance

Zusammenfassung

Mit dem Voranschreiten der Digitalisierung im Bankensektor und der stetig wachsenden Menge an Daten, die durch Banken erfasst und gespeichert werden können, ergeben sich viele verschiedene Möglichkeiten der Verwertung dieser unterschiedlichen Arten von Daten.

Ein Beispiel der Verarbeitung dieser Daten ist die Vertriebssteuerung anhand von Data-Science-Analysen und mittels Machine-Learning-Modellen. Die Aussagekraft und Qualität der Analysen und Machine-Learning-Modellen hängt stark von der Qualität der Daten ab, die als Grundlage für die Analysen und Modelle dienen.

Eine Herausforderung in diesem Prozess der Datenverarbeitung besteht darin, die Datenqualität automatisiert sicherzustellen. Stammen Daten aus externen Quellsystemen, so sind diese Systeme nicht direkt kontrollierbar und können durch verschiedene Arten von Fehlern die Qualität der bereitgestellten Daten negativ beeinflussen.

Diese Thesis beschäftigt sich damit, eine Datenpipeline anhand von fest definierten Datenqualitätsrichtlinien zu konzipieren und zu implementieren. Mit der Implementierung soll eine automatisierte Validierung der Qualität und Konsistenz von Bankdaten anhand von einem konkreten Anwendungsfall umgesetzt werden.

Die Datenpipeline basiert auf dem Apache Spark Framework und wird auf einem Dataproc Cluster in der Google Cloud ausgeführt. Das Cluster wird nicht dauerhaft betrieben, sondern dynamisch gestartet, sobald neue Daten zur Verarbeitung bereitstehen. Die Speicherung der validierten Daten findet in einem BigQuery Data Warehouse statt.

Diese Arbeit geht auf einige theoretische Grundlagen hinter den Konzepten von Data Warehousing und ETL-Prozessen, sowie den Systemen Apache Spark und BigQuery, ein. Es wird anhand eines Prototyps einer cloudbasierten Datenpipeline gezeigt, wie diese theoretischen Aspekte in der Praxis umgesetzt werden können.

In der Arbeit wird untersucht, ob durch den Einsatz der prototypischen Implementierung die Datenqualitätsanforderungen der Eindeutigkeit, Vollständigkeit, Konsistenz und Gültigkeit in einer sequenziellen bzw. parallelen Verarbeitung der Daten erfüllt werden können. Außerdem wird das Erfüllen von weiteren nicht funktionalen Anforderungen überprüft.

Die Evaluation stellt heraus, dass sämtliche Datenqualitätsanforderungen durch die Implementierung der Datenpipeline bei einer sequenziellen Abarbeitung der extrahierten Daten sichergestellt werden können. Im Falle der parallelen Verarbeitung wird ein Szenario gefunden, in dem die Implementierung die funktionalen Anforderungen nur teilweise erfüllen kann.

Es stellt sich weiterhin heraus, dass dynamisch erstellte Dataproc Cluster nicht für zeitkritische ETL-Prozesse geeignet sind, da die Prozesse des dynamischen Startens und Beendens des Clusters per Dataproc API lange Wartezeiten nach sich ziehen. Daher eignen sich für diese Art von Anforderung dauerhaft ausgeführte Dataproc Cluster.

Literatur

Cai, Li; Zhu, Yangyong: The Challenges of Data Quality and Data Quality Assessment in the Big Data Era, Data Science Journal, 2015

Gluchowski, Peter: Data Governance: Grundlagen, Konzepte und Anwendungen, Heidelberg: dpunkt.verlag, 2020

Provost, Foster; Fawcett, Tom: Data Science for Business, Sebastopol: O`Reilly Media Verlag, 2013

Salloum, Salman; Dautov, Ruslan; Chen, Xiaojun; Xiaogang Peng, Patrick; Zhexue Huang, Joshua: Big data analytics on Apache Spark, 2016, URL: https://link.springer.com/content/pdf/10.1007/s41060-016-0027-9.pdf (besucht am 05.04.2022)