Dissertationsvorhaben von Nicolas Weeger

Thema: AI Engineering Blueprints for practicable Machine Learning systems

Hochschule für angewandte Wissenschaften Ansbach, Fakultät Technik

Erstbetreuer: Prof. Dr. Christian Uhl

Zweitbetreuer: Prof. Dr. Stefan Geißelsöder

angestrebter Titel: Dr. rer. nat.

1 Forschungsthema

Künstliche Intelligenz (KI) verändert zahlreiche Industrien und Anwendungsbereiche. Für Unternehmen ist es von entscheidender Bedeutung, KI-Techniken einzusetzen, um geschäftlichen Erfolg zu erzielen [1, 2]. Der Einsatz dieser Techniken kann insbesondere für kleine und mittlere Unternehmen (KMU) vorteilhaft sein, da er ihnen die Möglichkeit bietet, ihre Kompetenzen in spezifischen Bereichen, wie beispielsweise der Kundenerfahrung, der Produktionsüberwachung und den Entscheidungsprozessen, zu optimieren [3]. Die Entwicklung von Machine-Learning-Modellen (ML-Modellen) im eigenen Unternehmen, die innerhalb oder als Bestandteil eines Produkts zum Einsatz kommen und als KI-Systeme bezeichnet werden, kann in einem organisatorischen Kontext eine Reihe von Herausforderungen mit sich bringen. [4, 5, 6, 7]. Dazu gehört das Verständnis für die Feinheiten der KI, einschließlich ihrer funktionalen Anforderungen und Einsatzszenarien. Die Integration zusätzlicher Prozesse, wie beispielsweise die Datengenerierung und -vorverarbeitung oder das Modelltraining und -einsatz in den traditionellen Softwareentwicklungsprozess, kann insbesondere für KMU potenziell zu organisatorischen Problemen führen.[6]. Der ML-Modellentwicklungszyklus umfasst neben DevOps zusätzliche Praktiken für die Daten und Modelle. Hierzu zählen MLOps- und DataOps-Techniken, welche eine Kultur, Praktiken und Werkzeuge für den Umgang mit Daten und Modellen umfassen. Darüber hinaus ist eine Abstimmung der Systemarchitektur für KI-Systeme auf die Anforderungen des zugrunde liegenden Modells erforderlich. Die Integration von Trainings- und Inferenzumgebungen sowie die Speicherung und Versionierung von Daten und verschiedenen Artefakten ist eine essenzielle Komponente, um die Funktionalität des KI-Systems zu gewährleisten. Die effektive Implementierung von KI-Systemen erfordert demnach eine sorgfältig durchdachte Architektur, die auf die spezifischen Anforderungen der jeweiligen KI-Anwendung zugeschnitten ist. In dieser Dissertation wird die Entwicklung von Blueprints untersucht, die auf die Anforderungen der verschiedenen Arten von KI und deren Entwicklungsstufen abgestimmt sind. Die Blueprints verbinden die Prinzipien von AI-Engineering, DevOps, MLOps und DataOps um die Herausforderungen bei der Entwicklung von KI-Systemen zu bewältigen. KMUs können die Blueprints anwenden, indem sie Referenzarchitekturen und geeignete Automatisierungsansätze für verschiedene Arten von KI umsetzen und implementieren.

2 Stand der Forschung

2.1 AI Engineering

Das Fachgebiet AI Engineering stellt eine Weiterentwicklung des Bereichs Software Engineering dar. Es ist als ein aufstrebendes Feld zu betrachten, das sich aufgrund des raschen Wachstums im Bereich der maschinellen Lernens (ML) herausgebildet hat. Gemäß Gartners „AI Hype Cycle for 2024“(https://www.gartner.com/en/articles/hype-cycle-for-artifical-intelligence) befindet sich die Disziplin des AI Engineerings gegenwärtig auf dem Höhepunkt der Erwartungen. Laut Gartner bildet AI Engineering die Grundlage für die unternehmensweite Bereitstellung von KI und GenAI in großem Umfang. Den meisten Unternehmen mangelt es jedoch an den erforderlichen Daten-, Analytik- und Software-Grundlagen, um einzelne KI-Projekte produktiv zu nutzen und ein Portfolio von KI-Lösungen zu betreiben.
In [8] wurden mehr als ein Dutzend Projekte untersucht und es wurde festgestellt, dass die Herausforderungen von AI Engineering zu Problemen im produktiven Betrieb von KI-Projekten führen. Die vorliegende Studie gelangt zu dem Schluss, dass ein signifikanter Anteil der Unternehmen, die sich mit der Entwicklung von Modellen für maschinelles Lernen befassen, bei dem Versuch, diese in die Produktion zu überführen, mit Herausforderungen konfrontiert sind. Autoren präsentieren eine Forschungsagenda und einen Überblick über die Fragestellungen, die in dieser Richtung adressiert werden müssen.
Wie [9] darlegt, wurde der Bereich des Software Engineering bereits umfassend wissenschaftlich erörtert, während das Themengebiet des AI Engineerings weit weniger intensiv untersucht wurde. Es existiert eine begrenzte Anzahl an Publikationen, die konkrete Erfahrungen mit der Anwendung von AI Engineering-Prinzipien aufzeigen. Die Autoren wählten zehn AI Engineering-Praktiken aus mehreren Kategorien aus der Literatur aus und wendeten sie auf eine Beispielimplementierung an, um die Praktiken und ihre Systemarchitektur zu bewerten. Darüber hinaus haben Gespräche und Fragebögen mit KMUs gezeigt, dass der Wunsch besteht KI in ihre Systeme zu implementieren. Der Erfolg ist jedoch davon abhängig, ob die zuvor genannten Herausforderungen bewältigt werden können. Die Anwendung von AI Engineering-Praktiken kann für Unternehmen einen entscheidenden Vorteil darstellen, indem sie die Entwicklung, den Einsatz und den Betrieb von Modellen für maschinelles Lernen optimiert.

2.2 MLOps

Die Idee hinter MLOps besteht in der Bereitstellung von Techniken und Werkzeugen für den Einsatz und den Betrieb von KI-Systemen [10]. Das Ziel besteht darin, eine Strategie zur Lösung realer Problemen beim Einsatz von ML-Modellen zu entwickeln. In mehreren Studien werden verschiedene Literaturquellen in diesem Bereich untersucht und Pipelines, Taxonomien, Werkzeuge, Methoden und Herausforderungen in diesem Bereich angeboten [11, 12, 13]. In der Publikation [14] wird eine systematische Mapping-Studie für MLOps-Architekturen durchgeführt eine Anzahl von 35 Architekturkomponenten aufgezeigt. Es werden verschiedene Architekturvarianten für unterschiedliche Anwendungsfälle beschrieben und gängige Werkzeuge für diese Architekturkomponenten bereitgestellt.

2.3 Weitere relevante Forschungsarbeiten

In [15] wurde eine Referenzarchitektur für spezifische Anwendungsfälle in der Prozessindustrie vorgestellt, die sich mit Edge Devices befasst. Die Autoren demonstrierten die Architektur durch die Implementierung einer Fallstudie für einen realen Anwendungsfall und bewiesen die Funktionalität mit dieser Anwendung. In [16] wurde eine Referenzarchitektur zur Erleichterung der Nutzung von Big-Data-Techniken im Edge-Computing-ML entwickelt. Die Studie präsentiert eine Analyse Architektur der Modellentwicklung und deren Anwendung für diesen Anwendungsfall. Eine weitere Studie [17] stellt eine Vision für „disziplinierte, wiederholbare und transparente modellgetriebene Entwicklung und Machine Learning Operations (MLOps) von intelligenten Unternehmensapplikationen“ vor. Die Autoren präsentieren ein dreistufiges Metamodell für die modellbasierte Entwicklung von KI/ML-Blueprints auf Basis intelligenter Anwendungsarchitekturen.
In mehreren Studien werden Entwurfsmuster für KI-basierte Systeme mit Blick auf Software und Architektur diskutiert [18, 19, 20, 21]. Es wird ein Überblick über Entwurfsmuster gegeben, die für KI-Anwendungsfälle angepasst sind. Zudem werden die Anwendung und die daraus resultierenden Vorteile bei der Entwicklung von Modellen für maschinelles Lernen aufgezeigt.

2.4 Zusammenfassung des Standes der Forschung

Zusammenfassend lässt sich sagen, dass die Literatur Einblicke in die Bedeutung, mögliche Architekturen und Prinzipien für AI Engineering und MLOps-Praktiken ermöglicht. Die Anwendung dieser Erkenntnisse ist derzeit jedoch auf einige wenige Referenzarchitekturen in spezifischen Bereichen wie Big Data oder Edge Devices beschränkt. Andere Studien fokussieren sich auf die Definition von Architekturen und Mustern und demonstrieren ihre Anwendbarkeit in Fallstudien. Die in dieser Dissertation entwickelten Blueprints basieren auf den Prinzipien des AI Engineering, welche die Grundlage für deren Entwicklung bilden. Im Rahmen der Entwicklung von KI-Systemen finden MLOps-Pipelines und -Tools sowie bestehende Referenzarchitekturen und Frameworks Anwendung. Ihr Einsatz dient der Unterstützung des Entwicklungsprozesses, der damit rationalisiert, standardisiert und beschleunigt wird. Software- und Architekturentwurfsmuster finden bei der Beschreibung der Entwicklung Anwendung, um die nicht-funktionalen Anforderungen (NFRs) für die verschiedenen Entwürfe zu erfüllen. Der Einsatz in Feldprojekten erlaubt eine flexible, hochautomatisierte Anwendung sowie einen ressourcenschonenden Betrieb für unterschiedliche Anforderungen in KMUs.

3 Ziele und wissenschaftlicher Beitrag der Promotion

Die Entwicklung von KI-Systemen ist zunehmend komplexer und unterstreicht damit die wachsende Bedeutung von AI-Engineering und MLOps-Techniken. Kleine und mittlere Unternehmen sehen sich mit beträchtlichen Herausforderungen konfrontiert, wenn es darum geht, künstliche Intelligenz (KI) in ihre Produkte oder Prozesse zu integrieren. Unternehmen mangelt es vielfach an den erforderlichen Ressourcen und dem entsprechenden Fachwissen, um KI-Systeme zu entwickeln, einzusetzen und zu betreiben, die auf ihre spezifischen Probleme zugeschnitten sind. In Anbetracht der mangelnden Forschung hinsichtlich der Implementierung von AI Engineering-Praktiken, insbesondere in KMU, fokussiert sich diese Dissertation auf die Entwicklung von Blueprints für den Umgang von Modellen für maschinelles Lernen (ML) unter Einsatz von AI-Engineering und MLOps-Praktiken. Diese Blueprints dienen als Fundament für die Entwicklung von KI-Systemen in KMU. Die Bereitstellung von

Referenzarchitekturen und adäquater Automatisierungsansätze für diverse Arten von ML ermöglicht die Entwicklung und den Betrieb von KI-Systemen. Die Effizienz der Blueprints wird durch ihre Anwendung auf einer Reihe von Feldprojekten evaluiert. Aus diesen resultieren weitere Anforderungen, und somit zusätzliche Entwicklungsschleifen zur Verallgemeinerung der Blueprints. Die Evaluation der Vorteile, die sich aus dem Einsatz von Blueprints für Organisationen ergeben, erfolgt durch die Beobachtung des Prozesses der Entwicklung von ML-Modellen sowie durch Interviews mit den Entwicklern.

4 Methodik

Im Rahmen der Untersuchung des Nutzens der Entwürfe für Unternehmen erfolgt die Validierung mittels der Methode des Design Science Research (DSR) [22, 23]. Die Herausforderungen und Anforderungen von KMU werden mithilfe von Interviews und einer umfassenden Literaturrecherche identifiziert. Zudem werden Verbesserungspotenziale und -möglichkeiten ermittelt. Basierend auf diesen Erkenntnissen besteht die Möglichkeit, Geschäftsanforderungen in Zusammenarbeit mit den relevanten Interessengruppen festzulegen, wobei sich diese an deren spezifischen Bedürfnissen orientieren sollten. Die Integration der fachlichen Anforderungen mit den Anforderungen der verschiedenen KI-Typen, wie Algorithmen, Datenspeicherung, Rechenkapazitäten und NFRs, ermöglicht die Entwicklung eines umfassenden Rahmenwerks. Im Anschluss besteht die Möglichkeit, die betreffenden Elemente einer iterativen Prüfung und Validierung zu unterziehen. Schließlich können die Artefakte in den Projekten der Beteiligten als Feldtest eingesetzt werden. Der Prozess wird in der Folge so lange iteriert, bis die Anforderungen finalisiert sind und die Artefakte die Anforderungen nachweislich erfüllen. Der Prozess wird in mehreren Projekten wiederholt, um die Ergebnisse zu verallgemeinern und sie für KMU so anwendbar wie möglich zu machen.

5 Arbeits- und Zeitplan

In der ersten Phase der Dissertation wurde die Definition des Themas und der Ziele durchgeführt. Die interdisziplinäre Ausrichtung des Themas, mit den Hauptbestandteilen in Softwarearchitektur und KI, verlangte eine umfassende Literaturrecherche in den Bereichen AI-Engineering, MLOps und Softwarearchitektur für Machine Learning Anwendungen. Diese wurde in den ersten sechs Monaten der Dissertation durchgeführt.
In der zweiten Phase wurde die Notwendigkeit und die Ziele der Forschung in Richtung AI-Engineering Blueprints für KMU untersucht. Die Ergebnisse dieser Phase wurden in einem Paper zusammengefasst, welches als peer reviewed Konferenzbeitrag im Rahmen des Software Architecture for Machine Learning (SAML) Workshop der International Conference on Software Architecture (ICSA) 2025 in Odense, Dänemark im April 2025 veröffentlicht [**?**].
In der dritten Phase wird die Entwicklung von Blueprints für verschiedene Anforderungen beim Training von Time Series Analysis Modellen durchgeführt. Diese Blueprints werden auf die spezifischen Anforderungen von Forschungsgruppen und KMU abgestimmt und sollen eine Grundlage für die Entwicklung von KI-Systemen in diesem Bereich bieten. Die Entwicklung der Blueprints wird in enger Zusammenarbeit mit den beteiligten Forschungsgruppen und Unternehmen erfolgen, um sicherzustellen, dass sie den praktischen Anforderungen entsprechen. Die Blueprints beinhalten Empfehlungen, Vor- und Nachteile und Validierungen verschiedener Architekturen, Prozesse und Werkzeuge,

die für die Entwicklung von KI-Systemen erforderlich sind. Darunter beispielsweise Datenhaltung, Daten- und Modellversionierung oder Experimenttracking. Diese Phase wird voraussichtlich bis Ende 2025 andauern.
In der vierten Phase werden Blueprints für die Verwendung von generativer KI auf Unternehmensdaten entwickelt. Dabei soll die Entwicklung von RAG Systemen und Agentic RAG Systemen, basierend auf Unternehmensdaten und -prozessen, im Fokus stehen. Diese Blueprints sollen eine Grundlage für die Entwicklung von KI-Systemen bieten, bei denen generative KI-Techniken genutzt werden. Sie konzentrieren sich auf Entwicklung, Architektur und Bereitstellung sowie den Betrieb von produktiven Use-Cases in KMUs. Die Validierung der Blueprints erfolgt durch die Anwendung in Feldprojekten, um sicherzustellen, dass sie den praktischen Anforderungen entsprechen. Diese Phase wird voraussichtlich parallel zur dritten Phase beginnen und bis Ende 2026 andauern.
Die fünfte und finale Phase der Dissertation wird die Zusammenfassung der Ergebnisse und die Erstellung der Dissertation umfassen. In dieser Phase werden die Ergebnisse der vorherigen Phasen zusammengefasst und in einem kohärenten Dokument präsentiert. Die Dissertation wird die entwickelten Blueprints, ihre Anwendung in Feldprojekten und die daraus gewonnenen Erkenntnisse detailliert beschreiben. Diese Phase wird voraussichtlich bis Oktober 2027 andauern.

6 Zuordnung zum Promotionskolleg REDIG

In dieser Dissertation werden die Prinzipien des AI-Engineering, MLOps und Softwarearchitektur kombiniert, um Blueprints für die Entwicklung von KI-Systemen zu erstellen. Diese Blueprints zielen darauf ab, KMU bei der effektiven Integration von KI-Techniken in ihre Produkte und Prozesse zu unterstützen. Darüber hinaus wird die Dissertation durch die Anwendung der entwickelten Blueprints in Feldprojekten validiert. Dies trägt zur Digitalisierung, Automatisierung und somit Effizienzsteigerung in KMU bei und fördert die Entwicklung von KI-Systemen, die auf die spezifischen Bedürfnisse und Anforderungen von KMU zugeschnitten sind. Die Arbeit leistet somit einen wertvollen Beitrag zur Forschung im Bereich AI-Engineering und bietet gleichzeitig praktische Lösungen für KMU.

Literatur

[1] I. M. Enholm, E. Papagiannidis, P. Mikalef, and J. Krogstie, “Artificial Intelligence and Business Value: A Literature Review,” *Information Systems Frontiers*, vol. 24, no. 5, pp. 1709–1734, 2022.

[2] S. M. C. Loureiro, J. Guerreiro, and I. Tussyadiah, “Artificial intelligence in business: State of the art and future research agenda,” *Journal of Business Research*, vol. 129, pp. 911–926, 2021.

[3] K. Bhalerao, “A study of barriers and benefits of artificial intelligence adoption in small and medium enterprise,” *Academy of Marketing Studies Journal*, vol. 26, no. 1, 2022.

[4] L. Fischer, L. Ehrlinger, V. Geist, R. Ramler, F. Sobiezky, W. Zellinger, D. Brunner, M. Kumar, and B. Moser, “AI System Engineering—Key Challenges and Lessons Learned,” *Machine Learning and Knowledge Extraction*, vol. 3, no. 1, pp. 56–83, 2020.

[5] L. E. Lwakatare, I. Crnkovic, and J. Bosch, “DevOps for AI – Challenges in Development of AI-enabled Applications,” in *2020 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, (Split, Croatia), pp. 1–6, IEEE, 2020.

[6] M. Schönberger, “Artificial Intelligence for Small and Medium-sized Enterprises: Identifying Key Applications and Challenges,” *Journal of Business Management*, vol. 21, pp. 89–112, 2023.

[7] E. D. S. Nascimento, I. Ahmed, E. Oliveira, M. P. Palheta, I. Steinmacher, and T. Conte, “Understanding Development Process of Machine Learning Systems: Challenges and Solutions,” in *2019 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, (Porto de Galinhas, Recife, Brazil), pp. 1–6, IEEE, Sept. 2019.

[8] J. Bosch, H. H. Olsson, and I. Crnkovic, “Engineering AI Systems: A Research Agenda,” in *Advances in Systems Analysis, Software Engineering, and High Performance Computing* (A. K. Luhach and A. Elçi, eds.), pp. 1–19, IGI Global, 2021.

[9] M. Grote and J. Bogner, “A Case Study on AI Engineering Practices: Developing an Autonomous Stock Trading System,” in *2023 IEEE/ACM 2nd International Conference on AI Engineering – Software Engineering for AI (CAIN)*, pp. 145–157, 2023.

[10] G. Symeonidis, E. Nerantzis, A. Kazakis, and G. A. Papakostas, “MLOps - Definitions, Tools and Challenges,” in *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, (Las Vegas, NV, USA), pp. 0453–0460, IEEE, 2022.

[11] M. Testi, M. Ballabio, E. Frontoni, G. Iannello, S. Moccia, P. Soda, and G. Vessio, “MLOps: A Taxonomy and a Methodology,” *IEEE Access*, vol. 10, pp. 63606–63618, 2022.

[12] D. Kreuzberger, N. Kühl, and S. Hirschl, “Machine Learning Operations (MLOps): Overview, Definition, and Architecture,” *IEEE Access*, vol. 11, pp. 31866–31879, 2023.

[13] M. Steidl, M. Felderer, and R. Ramler, “The pipeline for the continuous development of artificial intelligence models—Current state of research and practice,” *Journal of Systems and Software*, vol. 199, 2023.

[14] F. A. Najafabadi, J. Bogner, I. Gerostathopoulos, and P. Lago, “An Analysis of MLOps Architectures: A Systematic Mapping Study,” in *European Conference on Software Architecture*, vol. 14889, pp. 69–85, Springer Nature Switzerland, 2024.

[15] R. Wostmann, P. Schlunder, F. Temme, R. Klinkenberg, J. Kimberger, A. Spichtinger, M. Goldhacker, and J. Deuse, “Conception of a Reference Architecture for Machine Learning in the Process Industry,” in *2020 IEEE International Conference on Big Data (Big Data)*, (Atlanta, GA, USA), pp. 1726–1735, IEEE, 2020.

[16] P. Pääkkönen and D. Pakkala, “Extending reference architecture of big data systems towards machine learning in edge computing environments,” *Journal of Big Data*, vol. 7, no. 1, pp. 1–29, 2020.

[17] W.-J. Van Den Heuvel and D. A. Tamburri, “Model-Driven ML-Ops for Intelligent Enterprise Applications: Vision, Approaches and Challenges,” in *Business Modeling and Software Design* (B. Shishkov, ed.), vol. 391, pp. 169–181, Cham: Springer International Publishing, 2020.

[18] L. Heiland, M. Hauser, and J. Bogner, “Design Patterns for AI-based Systems: A Multivocal Literature Review and Pattern Repository,” in *2023 IEEE/ACM 2nd International Conference on AI Engineering – Software Engineering for AI (CAIN)*, pp. 184–196, 2023.

[19] R. Sharma and K. Davuluri, “Design patterns for Machine Learning Applications,” in *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, (Erode, India), pp. 818–821, IEEE, 2019.

[20] R. Cabral, M. Kalinowski, M. T. Baldassarre, H. Villamizar, T. Escovedo, and H. Lopes, “Investigating the Impact of SOLID Design Principles on Machine Learning Code Understanding,” in *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering - Software Engineering for AI*, (Lisbon Portugal), pp. 7–17, ACM, 2024.

[21] M. Take, S. Alpers, C. Becker, C. Schreiber, and A. Oberweis, “Software Design Patterns for AI-Systems,” *EMISA*, pp. 30–35, 2021.

[22] Hevner, March, Park, and Ram, “Design Science in Information Systems Research,” *MIS Quarterly*, vol. 28, no. 1, p. 75, 2004.

[23] M. Ivarsson and T. Gorschek, “A method for evaluating rigor and industrial relevance of technology evaluations,” *Empirical Software Engineering*, vol. 16, no. 3, pp. 365–395, 2011.