Towards an AI-driven business development framework: A multi-case study

Received: 1 November 2020 | Revised: 26 May 2021 | Accepted: 22 June 2021

DOI: 10.1002/smr.2432

SPECIAL ISSUE - EMPIRICAL PAPER

Journal of Software: Evolution and Process WILEY

Towards an AI-driven business development framework: A multi-case study

Meenu Mary John[1] | Helena Holmström Olsson[1] | Jan Bosch[2]

[1]Department of Computer Science and Media Technology, Malmö University, Malmö, Sweden

[2]Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg, Sweden

Correspondence
Meenu Mary John, Department of Computer Science and Media Technology, Malmö University, Malmö, Sweden.
Email: meenu-mary.john@mau.se

Funding information
Software Center

Abstract

Artificial intelligence (AI) and the use of machine learning (ML) and deep learning (DL) technologies are becoming increasingly popular in companies. These technologies enable companies to leverage big quantities of data to improve system performance and accelerate business development. However, despite the appeal of ML/DL, there is a lack of systematic and structured methods and processes to help data scientists and other company roles and functions to develop, deploy and evolve models. In this paper, based on multi-case study research in six companies, we explore practices and challenges practitioners experience in developing ML/DL models as part of large software-intensive embedded systems. Based on our empirical findings, we derive a conceptual framework in which we identify three high-level activities that companies perform in parallel with the development, deployment and evolution of models. Within this framework, we outline activities, iterations and triggers that optimize model design as well as roles and company functions. In this way, we provide practitioners with a blueprint for effectively integrating ML/DL model development into the business to achieve better results than other (algorithmic) approaches. In addition, we show how this framework helps companies solve the challenges we have identified and discuss checkpoints for terminating the business case.

KEYWORDS
AI-driven business development framework, artificial intelligence, challenges, deep learning, iterations and triggers, machine learning

1 | INTRODUCTION

Digitalization implies a transition from a hardware and product-based business to one that relies primarily on software, data and artificial intelligence (AI) to improve products, offer purely software-based products provide new digital and data-driven services to customers.[1] In addition, digital technologies are enabling companies to significantly accelerate value creation, replace transactional business models with more service-oriented business models, move to a continuous customer relationship characterized by entirely new ways of retrieving, responding to and redefining customer and market needs.[2] Digital and data-driven services are being used in a variety of areas such as preventive maintenance of vehicles, mobility services with a focus on subscription and "pay-as-you-go" business models, automation and as a key component in autonomous driving.[2] However, as companies advance from a hardware and product-based business to one that is increasingly focused on digital technologies

such as software, data and AI, they will need to evolve their current systems and complement them with new technologies. From this perspective, AI/machine learning (ML)/deep learning (DL) technologies offer excellent opportunities as they enable innovation and new ways of meeting customer needs.[2] Moreover, these technologies enable companies, especially in the embedded systems domain to adapt to new ways of working characterized by continuous integration (CI) practices and the deployment of not only of software features but also of data and ML/DL components.[2]

Companies across all domains are using AI/ML/DL[3,4] technologies to improve, scale and optimize their business. Adoption of these technologies in companies improves business scalability, productivity, customer intimacy and operational efficiency. ML/DL technologies are effective due to the vast amounts of data available and provide opportunities for data-driven decision making, business insights, pattern recognition and predictions, but few people have the skills required to develop these solutions.[5] Due to the shortage of data scientists, software developers and other roles and functions within a company are often asked to take responsibility for developing, deploying and evolving models. Although software developers recognize the value of their implementation, they find it difficult to apply ML/DL technologies because they lack expertise in data science.[6] Furthermore, non-experts with little or no data science background run the risk of developing and deploying invalid ML/DL models for their daily needs if they view accuracy as the only evaluation metric of a model performance.[5] Regardless of the level of expertise and/or access to data scientists, companies across the domains struggle to create high-performance models due to a lack of established and systematic design methods and processes. Software companies need to evolve their development practices and processes to build intelligence into their products as the ML/DL technology matures. In conventional software development, there are various approaches and procedures that assist developers; however, this is not the case with ML/DL models. Although previous research has identified the development, deployment and evolution of ML/DL models and their integration into larger systems as a challenge,[7,8] there is little, if any, support for the development, deployment and continued evolution of ML/DL models in a systematic and structured manner. In previous literature studies,[9-14] existing processes and structures representing ML/DL development tend to focus on the data-intensive context or on a mixture of data and model requirements contexts and have little to no focus on business case generation, selection and validation.

In our previous research, and based on a multi-case study in three embedded systems companies "Developing ML/DL Models: A Design Framework" presented at the International Conference on Software and Systems Process (ICSSP) 2020,[15] we identified the typical phases that data scientists go through when designing ML/DL models. The seven typical phases are (i) Business case Specification, (ii) Data Exploration, (iii) Feature Engineering, (iv) Experimentation, (v) Development, (vi) Deployment and (vii) Operational. For each phase, we identified the approaches as well as the major challenges experienced by the data scientists involved in our study. Finally, we outlined the iterations between the different development phases and the events that trigger these iterations to optimize the design process.

This journal is an extension of the paper "Developing ML/DL Models: A Design Framework" (ICSSP) 2020.[15] In this extension, we provide the following contributions in addition to those already presented in the conference paper. First, we add three new case companies, in addition to the original three, to provide additional empirical results and further generalization of our findings on the development, deployment and evolution of ML/DL models. Second, we present a conceptual framework in which we describe not only the activities involved in the development of ML/DL models but also the roles and company functions involved as well as the iterations that occur and the activities they trigger to optimize the process. In this framework, the phases originally identified in the ICSSP are integrated into three high-level activities (i.e., focusing on business, data and model) that companies perform in parallel to develop, experiment with and optimize the ML/DL models they develop. With this framework, we provide a blueprint for how to effectively integrate AI/ML/DL into the business of companies in a systematic way. The framework details how ML/DL business cases are generated, selected and validated by clients and practitioners; how the datasets needed for business cases are collected and explored to find interesting insights; and how key features are coined as well as develop, deploy and evolve models. Finally, we show how our framework can help solve the key challenges identified during our empirical study and explore various checkpoints for immediately terminating business cases that do not offer significant cost reductions, time savings or other benefits to clients. The framework represents the continuous delivery of ML/DL systems to accelerate AI-driven business with companies and provides an agile way of working instead of a sequential way of working to better manage development, deployment and evolution. By AI-driven, we refer to the inclusion of ML/DL models into software-intensive systems with the intention of generating better results than other (algorithmic) approaches. The ML/DL model achieves better results because it learns from data. Businesses benefit from AI technology by enabling better ways of working, automating complex processes, enriching customer experience, increasing productivity, improving operational efficiency and freeing up time for innovation. This in turn helps companies improve their products and ultimately their business. This can only happen if they are able to successfully integrate ML/DL development into the larger system development and business context.

The rest of the paper is organized as follows: In Section 2, we review the contemporary literature on how digital technologies are changing embedded systems and how AI/ML/DL can contribute to this transformation and its applications. In Section 3, we present the research methodology and the six case companies involved in the study. In Section 4, we report our empirical findings from the multi-case study research in the six case companies and the challenges they experienced. In Section 5, we derive a framework based on our empirical findings to help companies effectively integrate the development of ML/DL into their business and the larger system of which the ML/DL component is a part. In Section 6, we review the related work and in Section 7, we discuss threats to validity. In Section 8, we conclude the paper and provide an outlook for future research.

2 | BACKGROUND

In this section, we first review the contemporary literature on how digital technologies are changing current business practices in the embedded systems domain, especially through the adoption of DevOps, DataOps and MLOps practices. Second, we discuss previous studies on AI/ML/DL technologies and their applications in various companies.

2.1 | Digitalization: New technologies and ways of working

Most companies have undergone significant changes in recent years as a result of increasing digitalization.[16] According to Gartner,[17] digitalization generates new revenue and value by using digital technologies to transform a business model. With sophisticated mechanisms to support massive data collection, processing and execution as well as novel ways of connecting and communicating, digital technologies are shaping the way embedded systems companies operate and experience the emergence of new and faster business opportunities than ever before. Introducing functionality in software rather than hardware is enabling companies to improve the customer experience by upgrading and refining products and extending product life.

In recent years, and as reported in previous studies, CI and continuous deployment (CD) practices are leading to shorter development times and faster placement of release candidates into production environments. Recently, new ways of working have emerged that are becoming increasingly important for companies seeking to take advantage of these new technologies. Practices such as DevOps,[18-20] DataOps[21] and MLOps[22,23] are being adopted with the intention of driving both product automation and quality in terms of development, data and ML operations. Penners and Dyck[24] propose DevOps as a collaboration of teams working in development and IT operations within a software-intensive company to deliver faster software changes.[18] Key benefits of adopting DevOps in companies[25] include faster delivery of software changes, higher operational productivity, better quality and improved organizational culture and attitudes. On the other hand, companies face some difficulties in adopting DevOps. These include inadequacies in infrastructure automation, high skill and knowledge requirements, project and resource constraints and monitoring challenges. According to Munappy et al,[21] DataOps can be defined as "an approach that accelerates the delivery of high-quality results by automation and orchestration of data life cycle stages. DataOps chooses the best practices, processes, tools and technologies from agile software engineering (SE) and DevOps to regulate analytics development, optimizing code verification, building and delivering new analytics, and thereby fostering a culture of collaboration and continuous improvement." Challenges to adopting DataOps include lack of pipeline robustness, data silos, organizational and restructuring. MLOps adopts and applies DevOps principles to ML models instead of software and merges the development cycles followed by data scientists and ML engineers with those of operational teams to ensure consistent delivery of high-performance ML models.[22] One of the biggest challenges is training for AI operations as most data scientists are not trained computer scientists by education and most data-intensive companies have little to no idea on how to manage their data.[26]

Digitalization is indeed much more than DevOps, DataOps and MLOps practices, but they enable embedded systems companies to adopt shorter and continuous cycles of software, data and ML technologies. These are called continuous development and deployment practices (for software, data and ML models). This can also benefit the mechanic and electronic parts of the system (for instance, by enabling software updates for an existing system providing means to effectively use the data collected by the system) as these are considered critical. The case companies we worked with in this study also consider these as core concepts and competences to accelerate digitalization and therefore see them as important mechanisms to succeed in a digital world. The companies see it as critical to integrate DevOps, DataOps and MLOps practices into their workflows as data and software components add value to their business by opening up new opportunities.

2.2 | AI/ML/DL and its applications

Due to advances in ML, DL, Big Data and cloud computing, AI has gained popularity. According to Mitchell,[27] ML can be defined as "a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E." Jordan and Mitchell[3] describe methods, recent advances and research opportunities in ML. According to Bengio,[28] "Deep learning algorithms seek to exploit the unknown structure in the input distribution in order to discover good representations, often at multiple levels, with higher-level learned features defined in terms of lower-level features." Goodfellow et al[4] present an overview of DL concepts and techniques. ML analyzes and recognizes data patterns for value generation. In contrast, DL uses multiple neural layers to learn from data. DL differs from ML in that it learns to represent data automatically using multiple abstraction layers.[29] Unlike DL, ML requires a variety of feature engineering tasks to manually build this representation. ML/DL differs from conventional SE in that its behavior is heavily dependent on external data. The main difference between ML/DL and non-ML/DL systems is that in ML/DL systems, data take the place of code and data patterns are recognized using an algorithm rather than hard coding.

ML/DL has emerged as the preferred method in many fields such as robotics, speech and image recognition, natural language processing (NLP)[30] and computer vision.[31,32] Significantly, ML/DL technologies have been widely used to improve customer satisfaction in both large and small companies as well as online companies. Thus, ML/DL technologies have been used by companies such as Google,[33] Apple,[34] Microsoft[35] and Facebook[36] in their services such as Google Translator, Google Street View, Siri, Bing search, Cortana virtual assistant and DeepFace. In addition, ML/DL technologies are used in a variety of companies.[37] ML/DL applications in retail include recommendation engines, market segmentation and inventory planning. In contrast, ML/DL is used in demand forecasting, condition monitoring and process optimization. Some of the use cases in healthcare are real-time alerts and diagnostics, disease and risk detection while in financial services, use cases include evaluation of credit scoring, risk analytics and regulation, and customer segmentation. ML/DL technologies are used in the energy and utilities sector where use cases are power use analytics, carbon emissions, trading, and customer-specific pricing. The use of ML/DL technologies is widespread and advanced in the online sector, for instance, King and Peltarion. Moreover, ML/DL is becoming increasingly important for software-intensive embedded systems, for example, image recognition and prediction services.[38] The development, deployment and evolution of a complex ML-based business system continues to be a major challenge for software-intensive embedded systems.[8,39]

3 | RESEARCH METHOD

The research presented in this paper is part of a larger research initiative in which seventeen embedded systems companies and five Swedish universities are collaborating to improve digitalization capabilities within these companies*. In this context, we have conducted longitudinal multi-case study research with several of the companies on topics related to the development, deployment and evolution of AI/ML/DL technologies. For the purpose of this paper, we worked closely with five of these seventeen companies as well as one company that is external to this collaboration which was deemed highly relevant because it provides a platform for AI deployment and has several embedded systems companies as clients.

In our previous research, we conducted multiple-case study research in three embedded systems companies (Case companies A, B and C) to explore and identify the key phases and challenges faced by data scientists in developing, deploying and maintaining models.[15] As an extension of this previous research and to advance the empirical insights as well as the conceptualization and generalization of the results, we added three additional case companies. Two of the three additional case companies (Case companies D and E) are embedded systems companies and one of them (Case company F) is a non-embedded systems company that is not part of the research initiative. We included Case company F in our study because it offers an end-to-end ML/DL deployment platform for embedded systems companies. Following the guidelines of Runeson and Höst,[40] we conducted a multiple-case study to investigate the activities involved in the development, deployment and evolution of ML/DL models. In particular, we were interested in the challenges practitioners face and how these can be mitigated by better integrating the development of ML/DL models into the overall business context of a company. The case study method is an empirical research approach that relies on an in-depth analysis of a contemporary phenomenon that cannot be studied independently in its real-world context.[41] Case studies are conducted in SE to understand, clarify or demonstrate new technique capabilities, methods, tools, processes or business structures. Multiple-case studies allow for an even deeper understanding of each case in its whole by comparing both their similarities and differences.[42] During our study, we worked closely with practitioners from six different software-intensive embedded systems companies to understand the activities they undertake, the challenges they face and the ways in which the development of ML/DL models is becoming increasingly important to the companies and the businesses in which they operate. By using the multiple-case study approach, we were able to identify common activities that different roles in embedded systems companies perform in parallel and simultaneously in the development, deployment and evolution of ML/DL models. Based on this understanding, we derived a framework to incorporate ML/DL into the larger business context of a company.

3.1 | Case companies

In this section, we present the six case companies that participated in our previous and current study. For each company, we describe the different ML/DL use cases we studied. All the reported use cases use real-time datasets to build ML/DL models in real-world environments. Table 1 provides a detailed description of each case company as well as the practitioners involved in the study and their roles within the company. In this table, Case companies A, B, and C are part of the previous study, whereas companies D*, E* and F* are part of the recent study. *I* stands for interview participants and *W* for workshop participants in the study.

*Software Center - https://www.software-center.se/

TABLE 1 Description of case companies and practitioners involved in interviews and workshops

Case company	Description	Practitioners	
		ID	Roles
A	A company providing software, services, and infrastructure in communication technology	I1	Senior data scientist
		I2	Senior data scientist
		I3	Instigator
		I4	Data scientist
		W1	Data scientist
B	A company manufacturing and marketing vehicles	I5	Technology specialist AI/ML
		W2	Director
		W3	Manager
		W4	Software system architect
		W5	Research engineer
C	A company offering packaging and processing solutions for food products	I6	Data science manager
		I7	Solution architect
		I8	Data scientist analyst
		I9	AI application specialist
		W6	Head of data science
		W7	Director
		W8	Manager
		W9	Technology specialist
D*	A company providing mobility solution for vehicles	I10	Vice president
		I11	ML/AI engineer
		I12	Data scientist
E*	A company manufacturing pumps and electronics for pump control	I13	Project head
		I14	Senior data scientist
F*	A company providing platform for AI development	I15	Senior data scientist
		I16	Head of research
		I17	Chief AI officer
		I18	Senior data scientist

Case company A—Telecommunications: A telecommunications company with a multinational network of partners, for instance, vendors, operators and customers. In Case company A, we have studied four use cases in different business areas.

A1. Log analysis: With the advent of ML/DL technologies, log data can be used to gain valuable insights. Logs generated by various products are retrained by reusing all existing infrastructures, models and parameters to generate profits for the company. The dataset consists of gigabytes of log files generated by different products within the company.

A2. Paging cell phone: To access a specific cell phone, there is a need to page one or more base stations. The company analyzes the mobility patterns of user equipment (UE) within the network to specifically page the mobile phone while it is idle using an existing paging algorithm. The dataset contains all events that occurred for a specific node in the company within a week (about a thousand gigabytes of data).

A3. Detection of garbled speech frame: Speech frames are encoded when voice data are encoded over audio. Certain flags in a given speech frame contain information about the previous and subsequent frames. The company predicts whether the speech frame is garbled or not by analyzing this information. The ML should be super fast and accurate to be inserted into the system. The dataset consists of speech frames collected from the company.

A4. Prediction of hardware faults: The company predicts hardware faults in order to minimize the number of devices sent in for repair by customers. The information gathered from the crash log reports of different customers helps in building the ML model. Two aspects are considered when screening the hardware: (a) If the hardware is faultless, it is returned to the customer and (b) If the hardware is faulty, it is sent to the repair centre. The dataset consists of crash log reports collected from the company.

Case company B—Automotive I—*Autonomous driving vehicles*: A company that manufactures trucks, buses, construction equipment and also supplies marine systems. The vehicle needs faster DL algorithms when driving on high-speed roads. The company also needs to ensure that the failure rate for these safety-critical products is minimal. The company develops DL models in collaboration with external partners. The dataset consists of millions of manually created test cases based on accident statistics or on replicating typical traffic situations which are then mutated or combined in various ways. It also contains thousands of hours of data recorded during real-life driving.

Case company C—Packaging—*Defect detection*: A company that provides processing and packaging solutions for food and beverages. The company uses DL models to detect defects in packaging, for instance, dents and wrinkles, at each client site during processing. The global model in the cloud is trained with the knowledge gained from local training at each client site. To more effectively detect client-specific defects, the learnings from the cloud are then fed back to the client sites for inference using transfer learning. The dataset consists of packages with different patterns, types and colours.

Case company D—Automotive II—*Object detection*: The company acts as an innovation centre providing mobility solutions for passenger cars. The company focuses on redefining automotive technology by introducing smart solutions. The company uses DL models to detect forgotten items of taxi passengers. These DL models are used by the company to detect items, for instance, keys and wallets in the taxi. The use case has proved to be worthwhile as collecting and storing these forgotten items is quite expensive. The dataset consists of a large number of images with forgotten items in the taxi.

Case company E—Pump supplier—*Water supply rate for pumps*: The company manufactures various types of pumps and the electronics to control them. The company has an extensive network in many countries and distributes its products in these countries with the help of local distributors. The company designs and optimizes pump solutions to meet the needs of its customers. When there is an excess of water, the company uses ML models that provide information on how much water is flowing into the pumps. The dataset contains both mechanical and operational data from pumps.

Case company F—AI development platform provider—*Audio analysis*: A company that provides a platform for the development and deployment of AI models. The teams in their client companies can easily operationalize AI and grow their business using this platform. The main goal of company is to make AI accessible to everyone so they can focus on the business case and not spend time on repetitive coding tasks. The company is developing DL models to distinguish audio signals for industrial predictive maintenance. The DL models are used in client site to detect defects in machinery at an early stage. The dataset contains both good and bad audio signals from healthy and unhealthy machines.

3.2 | Data collection

The research reported in this paper is based on a semi-structured interview study conducted between September 2019 and May 2020 to collect qualitative data. With an increasing understanding and knowledge of the development, deployment and evolution of AI/ML/DL technologies, the first author developed the interview protocol. Based on the feedback and recommendations of the other two authors, who have extensive experience with AI/ML/DL technologies, the first author added some additional questions, merged similar questions and removed some irrelevant questions. The interview protocol consisted of two parts. Part I focused on the role and experience of the interviewee and Part II focused on current development practices and activities related to the development of ML/DL models and explored the challenges that practitioners experience in developing ML/DL models. We provide the interview protocol covering Part I and Part II in the appendix (Appendix A1). Based on the objective of the study, we planned the study by identifying key contacts in six case companies that are part of a larger research initiative based on their experience in dealing with AI/ML/DL projects. We used a snowballing technique to identify suitable practitioners in each company with experience in developing, deploying and evolving ML/DL models for the study after receiving suggestions from these key contacts. Once we had identified suitable practitioners, we sent a personalized invitation to the interviewees (key contacts + suitable practitioners) and arranged a suitable time slot. As a result, 18 practitioners from 6 large companies representing different business areas and domains participated in our interview study. The practitioners who participated in our study have a wide range of experience from 2 to over 10 years. All interviews lasted 1 hour and were conducted via video conferencing. With the consent of the interviewees, all interviews were recorded and transcribed for analysis. An opportunity for follow-up questions was arranged at the end of each interview. In addition to the interviews and as part of the overall research initiative, we continuously met with case company practitioners at various workshops, meetings and events organized by both case companies and university researchers to collect secondary qualitative data. We organized a total of five workshops where we visited the same case companies to present our preliminary findings from the study, gather their feedback and better understand their ways of working. These workshops provided an opportunity to discuss, share insights and experiences, and give us different perspectives from a group of practitioners interested in ML/DL models. Furthermore, the workshops provided excellent opportunities to explore many of the challenges associated with developing ML/DL models. In addition to the interviews and workshops, we also held several meetings and events with practitioners of the case companies to improve the quality of our research study. At these meetings, we obtained additional information about their current working practices, overall

system development processes, visions and goals. Furthermore, the AI-driven business development framework shown in Figure 2 was presented for validation during these meetings and events.

3.3 | Data analysis

During the analysis, each interview transcription was carefully read by the first author and summarized and discussed with the other two authors. Notes were also taken during the interviews to summarize the key content. In addition, any possible question or misunderstanding was discussed with the interviewee to avoid misinterpretation. We took notes when we presented our preliminary findings to practitioners in the follow-up workshops and when we validated the AI-driven business development framework in Figure 2 at the meetings and events. Our notes as well as the feedback we received from the company practitioners were continuously incorporated into the analysis process. In analyzing the interviews, workshops, meetings and events, we applied elements of open coding to identify key concepts and then grouped them into conceptual categories that captured the development activities undertaken by the companies, the roles and functions involved, and the many challenges they experienced along the way in developing, deploying and evolving ML/DL models.[43] Triangulation increases the accuracy and validity of our study.[44] Triangulation provided multiple perspectives on the topic under study[45] because there were three of us authors conducting the interviews and reading the interview transcripts, notes and summaries. The first iteration aimed to identify additional empirical knowledge to generalize our previous study findings. The second iteration was conducted to identify the high-level activities that companies undertake in developing, deploying and evolving ML/DL models. The third iteration was conducted to identify the challenges in designing models. In the fourth iteration, we identify iterations and triggers between phases to optimize ML/DL development and the roles and company functions involved based on the study. The final iteration was conducted to evaluate various checkpoints for immediate termination of less valuable business cases. The results derived from the analysis were sent to the practitioners involved in the study for validation.

4 | EMPIRICAL FINDINGS

We report empirical findings from each of the six case companies involved in our study. Companies adopting ML/DL technologies are interested in knowing whether or not they are useful/suitable for their business context, whether they are cost-effective and reliable, how much it will cost to set up the necessary infrastructures, what are the valuable data collection procedures, and the development, deployment and evolution of models. These questions are currently not widely explored in either the companies or academia leading to difficulties in determining the appropriate business case for introducing such technologies into the companies.[46] As a structure for presenting our empirical findings, we divide them into three viewpoints: Business, Data and Models. The introduction of ML/DL technology in companies is pointless if it does not bring benefits to customers. High-quality data are needed to train models that can identify hidden valuable patterns in the dataset. The value of a particular ML/DL business case can only be realized by operationalizing the ML/DL models. We believe that this categorization helps to provide an overview of the different elements that companies need to work on and advance towards becoming an AI-driven business. Below, we describe the business opportunities for AI at each of the case companies, their organizational structure and how they deal with data and ML/DL models.

4.1 | Business, data and model opportunities in case companies

4.1.1 | Case company A: Telecommunications

Business: Case company A provides information and communication technology products, software, services and infrastructure to service providers. The company uses AI in all services and products to improve automation, advance existing products, create new business opportunities, increase revenue and improve customer experience. In Case company A, the typical use cases for ML/DL are ticket routing of trouble reports and classification of faults in logs. Company A have experimental prototyping or half-applied research teams in addition to product development teams. They focus on proof-of-concept or pilot projects for the product development teams. In Company A, research and development (R&D) is organized according to the DevOps principle where developers and operations (e.g., release) work in close collaboration and cross-functional feature teams.

Case company A implements ML/DL to a product if it adds value to customers, reduces human effort and saves time compared with implementing traditional approaches. We observe that data scientists spend a day or two running simple ML/DL models over a structured dataset to check whether the business case criteria can be met or not when profitable business cases are proposed. They set up discussions with product owners to determine whether the accuracy achieved is worthwhile or whether they should invest time in improving accuracy. On the other hand,

certain business cases can only be solved with ML/DL to generate added value, even if they involve high costs. For instance, building an infrastructure to collect and examine radio data for the implementation of an analysis tool involves high costs.

> I3: *"It is very hard to nail down exactly the business case when it comes to ML/DL because typically there are a lot of costs involved than in ordinary projects."*

Business cases are validated by clients based on end-goal requirements before they are deployed. The requirements usually include accuracy, prediction time, available computing resources, understandability, interoperability with other environments, reproducibility and model execution environment. For instance, accuracy was given more importance because misclassifying the correct frame is much worse than not classifying the wrong frame in garbled speech frame detection. In another scenario, ML analyzes customer trouble reports to determine the team to which the report should be routed. It will take a while for a person to check the report when it is generated. The prediction time is not important in this case.

Data: Data scientists with the help of developers generate data that can be used for ML/DL business cases in Case company A. Data scientists find it very difficult to collect data once they sell a product to a customer who owns the data. In such cases, they either have to ask the customers directly or try other methods to get valuable and representative data. For instance, data scientists spend 3 months analyzing events in the dataset to reverse engineer the baseline in the case of paging cell phone use case as access to the node configuration is restricted and the data is anonymized. Data scientists spend time with domain experts to understand the data and the system. They use visualization techniques for data exploration. They either use domain knowledge experts to label the data or use unsupervised or clustering techniques for labeling. One of the data scientists in Case company A suggests that it would be a good idea for domain experts or data scientists with domain knowledge to physically sit down with the team to gain an understanding of the data. This seems to be a big problem in the company as it is complex and has many dependencies. Once data scientists have access to the data, they must therefore manually search for and contact domain experts. Data scientists continuously create and test hypotheses about the data to better understand the data and the system. It also provides insights into nonfunctional requirements such as computational budgets and model expectations.

> I2: *"We always start with the data, look at the data and just explore the data as much as physically possible. Just to understand it and to get some intuition around what is going on and what is behind all this data…. It is highly unstructured in the way that we do not know where the data will take us, so we try building and experimenting with the hypotheses that we have around data."*

Most data scientists compose features with the help of domain experts using two approaches: (a) Start with a high-dimensional feature set and scale it to a low-dimensional set by removing irrelevant features that do not affect model performance and (b) scale up from a low-dimensional to a high-dimensional feature set by adding relevant features. Data scientists often add features to the feature set that are not explicitly part of the dataset from a domain knowledge perspective.

Model: Data scientists follow certain guidelines when implementing ML/DL models. They use classical ML techniques for smaller datasets, DL for large labeled datasets and clustering techniques when clusters become visible during data exploration. Some data scientists use random forest for small dimensional problems and pairwise correlation between features for relatively small dimensional problems. Typically, data scientists take a small sample from a large dataset and identify the class of the problem, for instance, image or text. They select two to three approaches ranging from simple and classical to more advanced approaches such as DL that have a proven track record with the identified class of problem.

> I1: *"There is no scientific precise method for decision making. It is completely based on from one problem to another problem. That is not well-defined. I don't think anyone can say, "yes, this is like a formula that you can use"…So the key is just to have a scientific and objective approach, I think that is probably the ticket rather than just having the formula."*

We find that data scientists are interested in automating experimentation tasks with tools like H2O.ai, AutoML and Auto-WEKA. They have different attitudes when working with ML/DL models. Some data scientists do not care how the model works as long as it works, whereas others use accessible domain knowledge for simple models, which contributes to an easily explained model. To optimize the model, data scientists undertake hyperparameter tuning that can be based on previous literature or on any findings or experiments that data scientists wish to make. According to *I3*, it is difficult to significantly improve performance by experimenting with a different model after a particular model has been finalized. In this situation, experimenting with the already finalized model is a good choice.

With the introduction of ML/DL in the company, data scientists and ML engineers are focusing more on the development and deployment of ML/DL models. Data scientists place requirements on ML/DL models before they start deployment to ensure code review, unit testing of all components, deployment infrastructure ready for use, proper model training review, checking whether a model created by one data scientist is understandable to another data scientist or not, maintenance, robustness and stability testing. Practitioners focus on three aspects to get the model

into deployment and integration. First, they prepare the code for easy deployment in the docker container; secondly, a plan for integration with existing internal systems in the company; and thirdly, the use of reusable functions that exist in API services to package the model so that it can provide intended services. According to *I1*, the transition from prototype to production-ready model is quite slow, despite huge investments in AI and data scientists in the company. The data scientists put the model under the strict supervision of A/B testing. They also use a model execution environment with built-in support for A/B testing and canary selection as a deep part of the product. The deployed services provide a training interface, an inference interface and an evaluation interface to train data; retrain with new data; and compare new models with old models to decide whether to roll back or not. Data scientists often use a strategy of adapting the old model with new features so that a newer version of the model is available in the same infrastructure. Then, they compare the old and the new model to choose the best one. In contrast to this approach, data scientists perform continuous retraining by adding features to the existing model. Data scientists perform continuous monitoring and logging activities to ensure that the model is performing as expected.

> I3: *"Maybe what is a bit specific putting ML/DL models in production is that you want some way to monitor that the model keeps performing the way you need it to."*

4.1.2 | Case company B: Automotive I

Business: Case company B is engaged in the design, production and supply of automotive and the use of ML/DL technologies for predictive maintenance, image and speech recognition. The company considers automation as the key aspect of AI adoption, which in turn increases customer satisfaction. At Case company B, AI is mainly used for perception. Because autonomous driving comes with many privacy and security concerns, the vehicles will only hit the road after numerous V&V (Verification & Validation) activities. The company relies on external collaboration to develop product-level components for autonomous driving vehicles for public roads.

In Case company B, the software development organization practices an agile way of working. Practitioners work together in a team, across teams and in the company network. During the initial outbreak of AI techniques, the company established an ML team. We find that a single person is responsible for different roles in the company. For instance, *I5* acts as product owner for two different teams and as project manager for other projects. The company relies on conducting advanced engineering or research projects for autonomous driving. The data scientists in company use DL for perception and a mix of ML/DL for other projects. Case company B confirms the need for established processes at a later stage compared with the existing development model and processes.

> I5: *"I think that question is very much dependent on who gets it and what their context is…Since the company is big, we have many different departments, contexts and application areas. But for us who work with autonomous driving, it is a very immature area that is research-oriented where everyone is frantically trying to build products."*

Data: Compared with other case companies, Case company B needs to collect extensive data for V&V as these activities are necessary for safety-critical products. Consequently, the whole design of the autonomous driving vehicle together with the key performance indicators (KPIs) and methodology is based on V&V. For instance, V&V is important because there is a high probability that people will be killed or injured while driving on the road. According to *I5*, data collection is not difficult if enough resources are provided. On the other hand, practitioners have different opinions when it comes to perception use case. DL models need a large dataset for better performance compared with ML models. For instance, datasets of safety-critical products are collected by manually creating, mutating and simulating test cases, and recording thousands of hours of data while driving on the roads to prove that failure rates are low. The company relies on a third party for data annotation. For instance, the Case company B builds and drives vehicles, whereas a third party provides the infrastructure to access the data.

> I5: *"Need enormous amount of data for perception systems, so it is hard to collect all that data…We need to prove that the failure rates are low enough."*

Model: According to Case company B, there are specific ways of working with ML/DL and non-ML/DL systems. Data scientists focus on a state-of-the-art approach to find the current best algorithms for research-oriented projects and on known and explored algorithms for a deeper understanding of small-scale pilot products. For instance, data scientists search for the state-of-the-art and select the best available algorithm because object classification is a key issue for all autonomous driving vehicles. They also conduct a feasibility study before selecting a particular algorithm. In Case company B, high AI and DL knowledge is required for perception. Knowledge includes understanding how to combine networks and how to be confident with the performance they provide, and challenges when integrating individual DL models into a well-functioning ensemble model. For established products such as vehicle perception, data scientists make realistic architectural decisions in terms of computational speed, memory speed and real-time execution. For instance, it is important to choose an algorithm that matches the performance of the hardware

available during the lifetime of the specific product development. For safety-critical products, the practitioners apply different method levels to the ML/DL component to ensure proper system behavior. For instance, they apply ISO 26262 and Safety Of The Intended Functionality Standard (SOTIF). For data scientists, the ML/DL project is challenging when the product itself changes frequently. For instance, changes to algorithms would lead to higher costs as autonomous driving requires extensive validation activities for safety certification. After completion of the required V&V, the DL models are installed in the vehicles for perception.

> I5: *"It costs a lot of money to change the algorithm in such a large project because a lot of validation is going on in the case of safety-critical products."*

4.1.3 | Case company C: Packaging

Business: The Case company C manufactures high-end packaging, processing and filling machines for various food products. The company uses AI to improve automation, to increase productivity and reduce costs. The company ensures customer satisfaction, brand presence and protection of food product. Typical use cases in Case company C include detection of defects in finished or semi-finished packages, quality control of seals and detection of misplaced food. Detecting defects by mounting a camera on the production line improves value creation. Triggering a warning during defect detection saves a lot of time and costs. The company also tries to apply for patents on its valuable and novel inventions to improve its market position and prevent price competition among competitors. Compared with most case companies, Case company C is more advanced in the implementation of ML/DL. It is developing suitable architectures and frameworks for the implementation of ML/DL business cases with the help of solution architects in the company. The company is focusing on the digitalization of food production in close cooperation with external partners. A data science team in the case company usually consists of a data science manager, senior data scientists, junior data scientists and solution architects. According to the data scientists at Case company C, the company will benefit from having both production and research-oriented teams operating simultaneously. The production-oriented teams rely on the traditional cloud approach, whereas the research-oriented teams experiment with new concepts, for instance, edge (re)training.

> I6: *"Some teams in company perform standard tasks while others try trending concepts."*

Data: Case company C already has mechanisms for cloud (re)training and edge inference. It intends to move from centralized cloud approach to a fully decentralized edge approach. The company has a global model for cloud (re)training and a local model for edge (re)training. The dataset required for the global model is collected either through direct sampling or manual dataset generation while the local dataset is unique to each client site. The company occasionally employs interns to label the datasets. To improve the performance of the global model in the cloud, mislabeled data are updated in the global dataset after edge inference. As the company moves towards decentralized approaches, appropriate decisions need to be made regarding the amount of data to be transferred to the cloud, how to handle labeling, feature selection, quality issues, model compression, Type 1 and Type 2 errors.

Model: We note that the company is taking initiatives to schedule frequent meetings and discussions between ML/DL and non-ML/DL practitioners to bridge the communication gap between them. For instance, architects in the company are provided with end-to-end knowledge of the entire ML/DL business case. The company searches for published work to find state-of-the-art learning. Before settling on a particular algorithm, *I8* experiments with available algorithms that are suitable for the business case. They also attend conferences and workshops to increase their knowledge of the latest advances in the field. The global model that has been (re)trained in the cloud is used at the edge when it outperforms the local model. The performance of the model is evaluated using metrics such as average precision, accuracy, false positive and false negative. To reduce power and resource requirements at the edge, the global model is compressed before being deployed at the edge and decompressed later without sacrificing accuracy. The company tries to use techniques like transfer learning and federated learning to optimize the performance of the model. The company ensures model management by sending logs to the cloud when errors (Type 1 or Type 2) occur.

> I8: *"The simplest part is deploying AI on the edge, while most challenging part is evaluating global model retraining."*

4.1.4 | Case company D: Automotive II

Business: An innovation company dedicated to providing intelligent mobility solutions for the automotive industry. Based on AI, the company seeks to strengthen its innovation capabilities and add value to products. Most business cases arise from client requirements or as part of internal research initiatives in the company. In the case of research projects, the user experience (UX) designer or the innovation management in the company determine whether the business case is valuable to any of the business owners/clients. If so, the company transforms the business case into

a real project for these clients. Once the business case is confirmed, the company sets up data science teams. A typical data science team in the company consists of product owner, data scientists and solution architects. When a business case is delayed or fails due to insufficient data, data scientists investigate what more can be done for the business case in the future. Unfortunately, they cannot apply their expertise in the current situation and are only involved in preparing future projects. The company is expanding the number of data scientists to apply ML/DL technologies to more valuable use cases. If the proposed business case does not add value to the business owners/clients from the start, it should be discontinued as soon as possible. It is quoted below:

> I10: *"From the first idea, if there is no business case, you stop"*

Data: To label data, the company uses the services of MTurk. Confidence is always low because MTurk involves people from all over the world in the labeling process. The company integrates data version control (DVC) into the workflow. If the annotations are initially bad, they can be refined and gradually updated and this data can be used to train models. In Case company D, DVC and global information tracker (GIT) work together. During scheduled meetings, data scientists encourage business owners with less ML/DL knowledge to complete a series of questionnaires. These questionnaires can bridge the communication gap between data scientists and business owners. For instance, the questionnaires include questions such as: What do you want to answer? What kind of data do you need to do this? Do you know any domain experts who can tell you what kind of data you need to do this? Most business owners believe that data scientists understand their business and can formulate their questions and answer their concerns, but this is not the case for the vast majority of data scientists. Often, data scientists try to clear up misconceptions about AI and tell business owners that they can automate the decision-making process when valuable relationships can be extracted from their data. According to *I10* and *I11*, exploratory data analysis is project dependent.

> I11: *"The specifics of how you do it, what you are going to look at it, depends a bit on the project. But I think the overall reasoning that you have to do exploratory data analysis is very much there."*

Model: The company is still in the early stages of developing ML/DL models with little integration with existing systems. They also rely on state-of-the-art to identify trending algorithms in their domains. We note that there is a slight change in the conventional wisdom of the small model versus the large model in the company. For instance, when it comes to text data, more blog posts and articles seem to argue for using a very large model and training it for a small period of time. Some of the projects are put on hold for deployment in the company. According to the data scientists in the company, very few have working code that can be experimented with, whereas many algorithms are available in publications. So there is a high probability that publications without working code will be rejected if the business case needs to be implemented in a short time.

> I11:*"Since we do have some time pressure on how to do what we want, we have taken the kind of publications that had working code and we have looked at it."*

4.1.5 | Case company E: Pump supplier

Business: Case company E is both a pump manufacturer and a pump retailer. The company uses AI, digital services and cloud networks to promote real-time monitoring and fault prediction of pumps. A particular use case for the company is condition monitoring of pumps and other critical equipment to determine the need for maintenance and repairs. Traditionally, product teams develop, maintain and operate services in Case company E. The company is trying to digitalize the services it offers to increase efficiency, and this idea is initiated by a business developer. For instance, a business developer proposes the concept of a wastewater network in the company. This will allow wastewater utilities to monitor the status of pumps when there is an influx of excess water. Case company E discusses different business cases in its internal ML network meetings. A typical data science team in this company consists of data scientists, data pipeline owner, domain experts and business owner. Other typical roles include back-end or front-end software developers and subject matter expert groups. One of the data science teams in the company consists of 10 practitioners and students, whereas the other team consists of 12 data scientists and 8–10 data engineers. The company has cross-functional teams that enable them to reach out to different teams to develop ML/DL models. For instance, the data pipeline team works with the marketing team as both teams have data scientists and data engineers who understand data and its purpose.

> I13: *"And then we support the teams with some subject matter expert groups, some of them are for software development like DevOps, some of them are from my team - data pipeline and platform. They support on how to set up data architecture or tools to use to set up data pipelines and how to make good data quality measures"*

Data: The company sets up devices to collect data for different business cases. They often start with a small data set as it is very costly to start a project with a large data set. After data collection, the company uses a quality assessment method to have a structured dialogue about the data quality dimensions of the dataset. According to *I13*, data scientists in other companies ignore the data dimensions when applying software development methods. Data scientists collaborate with domain experts in data exploration. They often start with small data sets that they explore in the alpha phase of the project. According to *I13*, the main difference between DevOps and DataOps is that more emphasis is placed on data architecture than on technical architecture. Data scientists confirm that the training-serving skew leads to poor performance of models in production. For instance, the typical format, alignment and timing of the dataset varies depending on training set.

> I13: *"So we have to rethink all the way how you can develop your in-house solutions because that is what I see when we use the software development methods, we disregard the data dimension and then we see all kind of problem coming up all over the place."*

Model: To introduce digitalization of services in Case company E, it is using cutting-edge AI, ML and cloud computing technologies. The company relies on a literature study to identify appropriate statistical methods for the business case and relies on business owners to get a reasonable idea of KPIs. The case company has tools to test and deploy models, easily switch between models and deal with data lineage. According to Case company E, the old model needs to be decommissioned when a new model is created and deployed in production. The use of agile development, DevOps, data science and MLOps makes it easier for Case company E to develop, deploy and evolve models.

4.1.6 | Case company F: AI development platform provider

Business: By providing a platform for building and deploying systems, the company aims to simplify AI development so that its benefits are accessible to non-experts. It supports client companies in their transition from the analog to the digital world by helping them adopt AI. Case company F investigates relevant potential problems before finalizing the business case. It provides AI platform lifecycle management to increase productivity and helps client companies solve their problems by providing a DL development platform. In addition to the main data science team working with the business owner, there are research teams working on new techniques, for instance, federated learning. When the business case is finalized, data science teams are formed within the company and junior and experienced data scientists are brought together to work on projects. The data scientist working in a particular business unit should become familiar with the field and eventually become a domain expert. Because the number of data scientists is low, the company believes that it is important to teach data science to software engineers in order to add value.

> I17: *"Defining exactly what should be the problem before going into the project and work for few months is much more important than you think."*

Data: The company plans to hold early meetings with business owners to gain a fuller understanding of the business case. These meetings aim to bridge the gap between the lack of knowledge and understanding of how ML/DL works among business owners and how the domain works among data scientists. These activities can also help avoid a scenario where the problem seems to have sufficient data and the prediction seems reasonable, but the difficulty is that there is no business case to justify the model development. For instance, building a fraud detection system by collecting a huge amount of data and spending hours on it is pointless if fraud only occurs once a year. Data scientists enlist the support of domain experts for data collection. In the early stages of data exploration, they conduct a quality test with a sample dataset to assess which data can be used for the business case and which cannot. According to data scientists in Case company E, the main difference between ML and DL is that traditional ML requires a significant amount of work in feature engineering. If relevant features in ML are removed, there is a high risk of information loss. Because data scientists are costly, their efforts can be used for other valuable tasks when working with DL models. By using DL instead of ML, we can avoid errors that occur when people are biased in feature engineering. On the other hand, feature engineering in ML is useful in two ways: (a) Reducing the dataset and finding good data points from the point of view of noise to remove impure data and (b) Simple algorithm cannot handle a large number of features.

> I17: *"DL Models are so complex that they can learn if you have enough data"*

Model: Data scientists rely on a literature study to find the state-of-the-art for developing DL models. To achieve model optimization, data scientists (a) Tune the hyperparameters, (b) Introduce more variances in the dataset, (c) Fine-tune the annotation, (d) Regularize the weights and (e) Understand the part of the data where the model underperforms. The robustness of the algorithm can be improved by artificially introducing more variances into the dataset. For instance, add variation to the input images by cropping or shifting them to create previously unseen images. Rethink the annotation as there is a risk that the model will underperform. In such cases, refine the annotation and perhaps update the data used to train the models. Weight regularization reduces the weights of the network to a small value if they do not contribute to the performance of the model. If a model developed on the basis of previous work fails, data scientists can check the discrepancies between the two datasets as a first step.

If possible, they try to validate the business case with the dataset described in previous work. If the problem still persists, they experiment with new algorithms. Data scientists assume that business owners have sufficient resources for integration and are equipped with interfaces to input data and obtain results. In contrast, in most cases, the data science team has to set up the integration for business owners for the first time. This may be because there is a lack of resources in the company or a lack of a diverse engineering organization or because they are busy with other tasks.

> I16: *"Even though we have a lot of experience having those meetings and even though we know many of the problems, we tend to fail. Even if we set up a problem then also we miss things that we have not communicated well on expectation"*

> I17: *"But I think we have evolved a lot and I think we are getting to a point that we need to solidify on a set of best practices but we have not yet."*

4.2 | Challenges

During our study, we identified several challenges faced by company practitioners in their daily practice of developing, deploying and evolving ML/DL models. Below is a summary of the main challenges faced by data scientists that are found in all companies. The challenges are grouped into three categories that relate to the development of ML/DL models: (a) Pre-Deployment, (b) Deployment and (c) Non-Technical challenges. We detail on each challenge and present them in Figure 1. The frequency of these challenges in each company can be found in Table 2.

4.2.1 | Pre-Deployment

1. **Representative and valuable dataset**: The majority of data scientists involved in our study confirm that DL models require a larger dataset than ML models. In contrast, one of the case companies disagrees and reports that there are several misconceptions about DL models such as they need a large amount of data. Most of the time, the data collected is noisy, has few labels or is completely unlabeled. According to data scientists, the dataset collected for ML/DL needs to be a valuable and representative sample. For instance, a dataset with data points for antibiotic-containing and non-containing prescriptions is needed to classify antibiotic-containing prescriptions. In another scenario, the dataset should include data from sites where the process is running normally and sites where it is not to determine whether an industrial process is running normally or not.

> I10: *"If I want to do machine learning, I will also need negative samples. Most datasets are skewed toward abnormality, this is usually not a deal-breaker, it is important to keep in mind early on."*

2. **Improper feature selection**: Most data scientists consider feature selection as an important step and point out that a feature set with insignificant features has implications for model performance. Without a thorough understanding of each feature, data scientists find it difficult to add relevant features to the feature set. For instance, if two features contain the same information then it is difficult to select the significant feature.

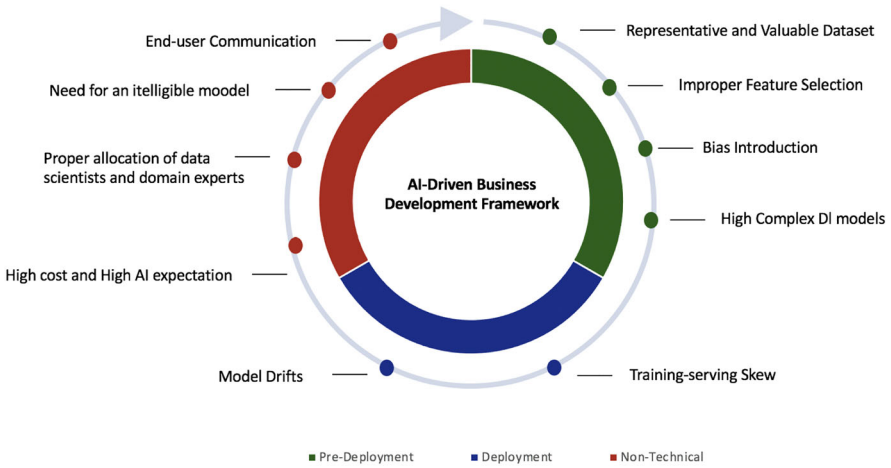

FIGURE 1 Challenges in ML/DL model development

TABLE 2 Frequency of challenges in case companies

Categories	Challenges	Case Company					
		A	*B*	*C*	*D*	*E*	*F*
Predeployment	Representative and valuable dataset	H	M	L	H	H	L
	Improper feature selection	M	L	L	M	M	L
	Bias introduction	M	L	L	L	L	L
	High complex DL models	L	H	M	H	L	L
Deployment	Training-serving skew	H	H	H	L	L	H
	Model drifts	H	H	H	L	L	H
Nontechnical	High cost and AI expectation	H	M	M	H	H	L
	Proper allocation of data scientists and domain experts	M	L	L	H	H	L
	Need for an intelligible model	H	M	M	H	M	H
	End-user communication	H	H	H	M	H	L

I4: *"Adding a new feature is costly in the way we work. So, we want to have an idea of what this feature can bring us before we do the implementation."*

3. **Bias introduction**: We find that data scientists introduce bias based on their experience in selecting algorithms for both ML/DL models and in selecting features for ML. For instance, most data scientists prefer näive Bayesian models, logistic regression, tree models, random forest and support vector machines as preferred models to achieve baseline performance. We note a conflict in the study when one of the data scientists points out that the random forest is computationally fast to train and gives results within seconds, whereas another data scientist states that the random forest is flexible but less interpretable.

I1: *"They bring their own experiences and then introduce own biases into the whole working."*

4. **High complex DL models**: Despite the popularity of DL models, most data scientists prefer ML models for training because they are less complex. Compared with ML, they believe that deep knowledge is required to implement DL models. For instance, deep DL knowledge is required to understand neural network limitations, network merging, and activation function in safety-critical use cases such as perception.

I15: *"We need to have a lot of AI and DL knowledge to understand the limitations of the networks, when they may fail a lot …. We are only half done after we have trained the networks."*

4.2.2 | Deployment

1. **Training-serving skew**: Data scientists consider training-serving skew as the discrepancy between training and serving performance. For instance, the system is trained on events to ensure it performs well on safety-critical products, but it is less suited to real-world situations while serving.

I5: *"It is very hard to capture real data for these long-tailored events because some of them will happen so rarely."*

2. **Model drifts**: Data scientists see model drifts as a potential threat to model performance. Therefore, it is necessary to properly identify model drifts and determine when they need to be retrained or updated. For instance, models are retrained daily, weekly, monthly or from year-to-year, depending on the domain and when the input data changes.

4.2.3 | Non-Technical

1. **High cost and high AI expectations:** According to the study, the implementation of ML/DL models requires significant infrastructure costs compared with traditional software development. According to Case company A, experts with little or no experience in data science have high AI expectations

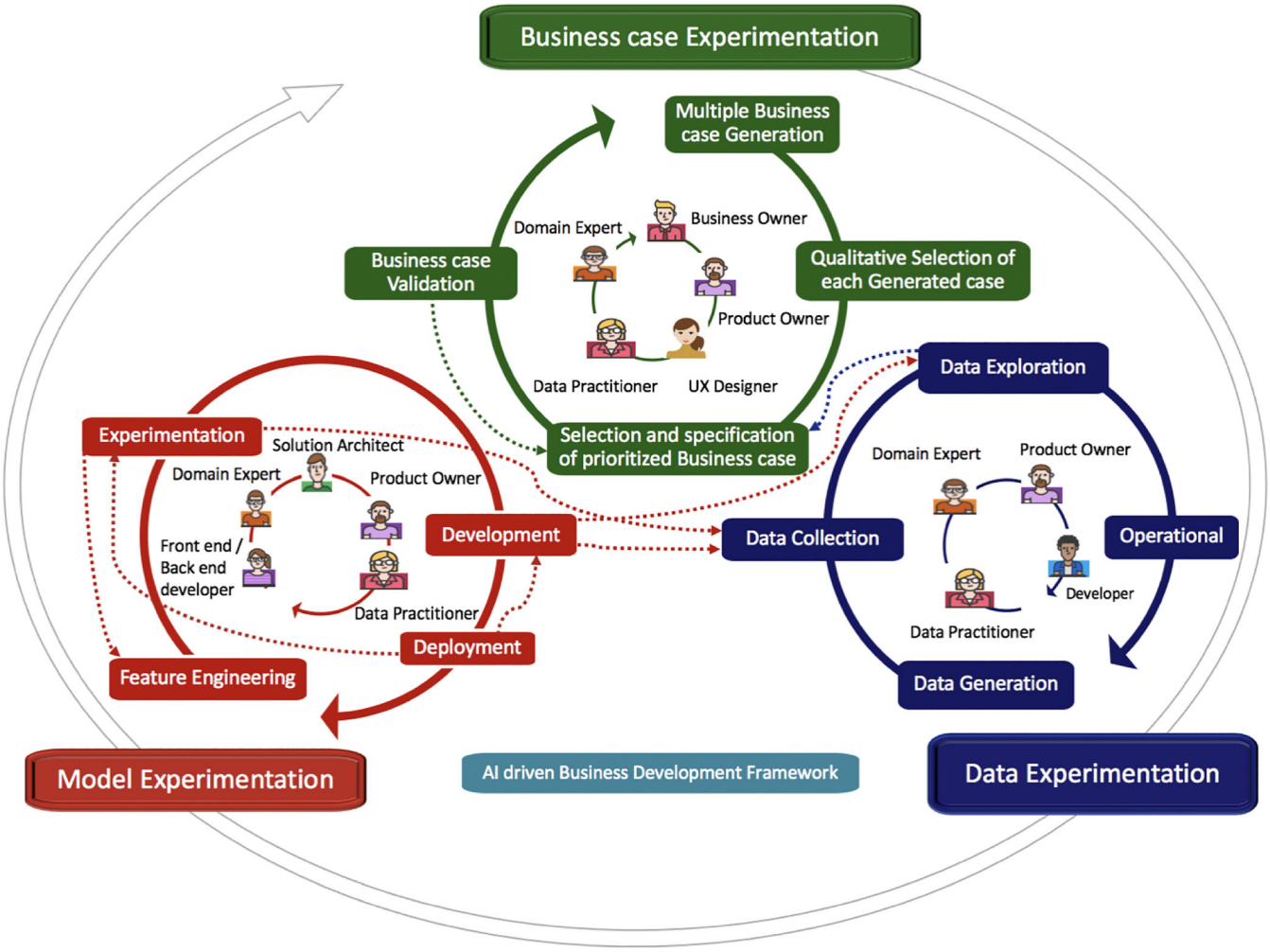

FIGURE 2 AI-driven business development framework

making it possible to gain early insights into what can be achieved in a given business case. For instance, when product owners hear the term "AI," they have extremely high expectations. Unfortunately, data scientists are usually pressured at a high-level to solve business cases with ML/DL. The majority of data scientists agree that it is difficult to formulate a business case for business owners who have a lot of data but no knowledge of AI.

> I4: *"We are assigned the business case, but we never get any information about the quantitative specifications to be optimized…. To make good business value, we need input from stakeholders, and sometimes we need to play an educational role because they are not aware of the potential risks involved in the design of ML/DL systems."*

2. **Proper allocation of data scientists and domain experts:** Most practitioners involved in our study confirm that data scientists and domain experts need to be carefully allocated to the project. Because data scientists are scarce resources, they need to be allocated to projects with high business value. For instance, we find that the number of data scientists in the case companies is significantly lower than the number of other technical staff. On the other hand, it is more difficult for data scientists to consult domain experts in a larger company because of their small numbers. We note that this could be because the companies have assigned domain experts to several ML/DL projects at the same time. We also note that one of the case companies relies on interns, particularly students, to label data. Although one of the companies uses a labeling tool to label the data, reviewing these labels is a tedious task. It is also difficult to find optimal models due to the lack of suitable data scientists to tune the hyperparameters.

> I1: *"Even then as a data scientist, you need a domain expert to label it."*

3. **Need for an intelligible model**: Most data scientists stress the importance of understanding the model developed by other data scientists. In the worst case, they spend a significant amount of time trying to understand the model, which slows down the overall speed of the project.

I4: *"It is very hard for another data scientist to understand the model in the case of some models that we have taken over from a data scientist."*

4. **End-user communication**: All data scientists report that there is a need to encourage communication with end-users as the model is difficult for them to understand. For instance, the data scientists have answered all the questions asked by customers because they do not have much confidence in the AI in the case of the analysis tool developed to analyse radio data. In other cases, the data scientists work with testers who have no experience in developing ML/DL models. The data scientists must spend time explaining the models to them because these testers have difficulty understanding the models.

I4: *"We need to educate them because they are very interested in understanding why we have done some specific predictions."*

I3: *"If you take an ordinary software component no one questions it as it is conceptually quite understandable; but with ML/DL models, there are many more questions."*

In Table 2, we classify the frequency of challenges that occur in six case companies into three ranges. These are high (H), medium (M) and low (L). As can be seen from the table, the majority of the companies face challenges in deployment. This could explain why the transition of ML/DL models from prototype to production quality is less pronounced in the case companies. Companies with very few deployment activities consider that the occurrence of these challenges is very low. Most companies believe that a dedicated team should be set up to monitor deployment issues. Communication with end users is a challenge for all companies when they have to interact with clients, testers, developers, architects, and others who have little or no ML/DL knowledge. It should be noted that companies working specifically on DL algorithms have fewer problems with selecting the wrong features and introducing bias because DL models learn patterns directly from the data. In some companies, the availability of experienced data scientists is still a nightmare.

5 | DISCUSSION

In this section, we discuss empirical findings and present a framework inductively derived from our empirical results. Based on the empirical findings, the framework provides a better understanding of how the adoption of ML/DL technologies brings profit to companies and how they can be integrated to deliver high business value, the different activities undertaken by practitioners, their way of working with business cases, data and models, and iterations and triggers to optimize model design, and the roles and company functions involved. The framework provides an end-to-end conceptual framework to ensure continuous CI/CD of ML/DL models for software-intensive components in embedded systems. In this context, we present three parallel and concurrent high-level activities that take place in companies as part of the development, deployment and evolution of ML/DL models, that is, business case experimentation, data experimentation and model experimentation. These three high-level activities are important because companies use ML/DL models only when they add value to the customer business. We also discuss various decision checkpoints that are useful for early rejection of low-value business cases. Figure 2 shows an overview of the framework and different iterations and triggers that can optimize the framework. Finally, we show how our framework helps solve the key challenges we identified in Section 4.2. Below, we describe the framework in detail.

5.1 | AI-driven business development framework

Our study shows that companies are struggling with the introduction of ML/DL components, model development and related practices into the business context of the company. Both in our interviews and during the workshops, this was expressed by several company practitioners when they reported challenges in applying new technologies, adopting new ways of working, and difficulties in introducing and training different company roles in relation to ML/DL model development. Challenges were also identified outside the direct context of developing ML/DL models as these ultimately need to be integrated and incorporated into a larger system and the overall context of a company. The end-to-end ML/DL process from business case generation to deployment as outlined in the framework can serve as a blueprint for those working to develop, deploy and evolve ML/DL models. AI companies will only benefit when their products are deployed in production. In such a context, this framework can accelerate the entire process and shorten the time between prototype and production-ready implementation of models. Moreover, the framework can be used as a guide for beginners with little or no background in data science and others who want to learn more about ML/DL models. Based on our findings, we note that there are three activities that all six case companies undertake in parallel and simultaneously when developing ML/DL models, that is, business case experimentation, data experimentation and model experimentation. By capturing these high-level activities and detailing them with the roles and iterations taking place, our framework provides a blueprint for the lifecycle management of ML/DL model development that mirrors the continuous practices of DevOps, DataOps and MLOps. The high-level activities outlined in the framework are explained in more detail below.

5.1.1 | Business case experimentation

Business case experimentation refers to the generation and validation of business cases suitable for ML/DL. The generation of multiple business cases are based either on the needs of business owners or on research projects. Often business cases are generated as a result of meetings or ideation processes or as part of brainstorming sessions within the company. For the qualitative selection of each business case generated, companies set up frequent meetings with the business owners to better understand the business cases and explore possible legal implications. A data treatment agreement can be created to ensure the proper use of data as agreed with two parties under the general data protection regulation (GDPR), especially for critical projects. Practitioners try to solve the business case in a non-ML/DL approach to see if it makes sense initially. Before selecting and specifying a prioritized business case, data scientists attempt a proof-of-concept based on random algorithms or a literature review to justify the actual model development. Data scientists plan early discussions about the evaluation metrics with the business owners and determine the metric that needs to be optimized for the business case. Once the business case has been validated by the business owners, it should be moved into production to make realistic data-based decisions.

5.1.2 | Data experimentation

Data experimentation refers to the generation, collection and exploration of data and monitoring of model performance based on inference data. Once the business case is specified, the next objective is to generate a dataset suitable for ML/DL models. It is shocking that the whole world praises the great availability of data and talks about the need for huge data storage mechanisms, but the amount of useful data is limited in practise. When collecting data, it is advisable to formulate a set of assumptions, questionnaires, checklists and facts to be checked about the data collected at an early stage to guide the approach and later conduct an investigation. In the initial stage of data collection, a quality test should be conducted on a sample dataset for evaluation. To gain a clear understanding of the data during the exploration, data scientists continuously build and test hypotheses about the data and provide insight into non-functional requirements. Various visualization techniques are used to understand the data distributions. It is recommended to ensure that the same data pre-processing techniques are used before training and after placing the model into operation. In the operational phase, the performance of the model should be monitored. In Figure 2, operational is part of the data experimentation as monitoring model performance is closely related to input and output data.

5.1.3 | Model experimentation

Model experimentation refers to the selection of appropriate features, experimentation with algorithms, finalization of an appropriate algorithm and placing the model into production. Once suitable data for ML/DL models have been generated and explored, relevant features are selected for modeling. The feature set is used as input and experimented with several algorithms to find a specific algorithm or set of algorithms. The finalized algorithm is tuned with hyperparameters to optimize the performance of the model and needs to be placed into production in parallel with existing processes and systems in software-intensive companies to achieve benefits. The deployed model should be under strict supervision of A/B testing. Based on model and data drifts, the model needs to be redeployed/retrained.

5.2 | Roles and company functions

Our empirical results show that different profiles of practitioners are involved in the development, deployment and evolution of ML/DL models. Companies set up data science teams as soon as the business case is finalized. A typical large data science team consists of product owner, data scientists, domain experts, business owner, back-end or front-end software developers. All roles in the company may be involved in different ML/DL projects simultaneously and in parallel. The product owner in the company keeps track of best practices and all responsibilities assigned to the teams and organizes the work and hosts team meetings. In companies, teams working on ML/DL projects are often divided into two groups: (a) The main data science team that works mainly on projects with business owner and (b) The research team that tries out the latest AI techniques for real-world business cases. All ML/DL projects require developers to generate data suitable for specific business cases. However, it is still difficult for data scientists to collaborate with developers who do not have a data science background. Often, data scientists ask to work with domain experts who are tasked with data collection and data scientists to help them. They help either by contributing to a tool that can be used for data collection or by discussing different ways to help them collect the data they need. If there is a lack of domain experts, data scientists can use clustering techniques or semi-supervised/unsupervised techniques for labeling.

Most companies hire data scientists by setting high standards to save the time required for their training. When a junior data scientist works with an experienced data scientist, it increases productivity and allows both parties to learn from each other and share skills. The company

expects data scientists working in a particular business unit to quickly understand the business context and become a domain expert as the number of data scientists and domain experts in companies is limited. Therefore, it is beneficial to teach software engineers how to use data science in a good, solid, repeatable and predictable way. To achieve this, software engineers need to be familiarized with different ways of working, processes or methods that they can use when dealing with ML/DL models. On the other hand, engineering teams in companies try to improve tooling and build platform support for the project. Solution architects plan to integrate ML/DL models with the rest of the software-intensive embedded systems. Some companies even use interns to (a) Bridge the communication gap between data scientists involved in developing ML/DL model development and solution architects involved in deployment, (b) Perform labeling tasks, (c) Be involved in data collection and (d) Engage in state-of-the-art learning. Data scientists retrain or redeploy a new model when performance degrades and add it to the list of available models for A/B experimentation. Automation of business-driven development can reduce bias among data scientists when performing feature engineering for ML models and in selecting preferred algorithms.

5.3 | Decision checkpoints for business case termination

Based on the analysis of our empirical data, we identify various check points for business case rejection during ML/DL model development. Below, we elaborate on the different checkpoints. The business case generated based on the needs of the business owners is quickly terminated if it does not make sense or have value in the initial stages of business case experimentation. Data scientists may initially treat the business case as a non-ML/DL case and apply basic thinking skills to check whether it creates value for the business owners. On the other hand, business cases that have generated as a result of internal meetings or brainstorming processes within the company are terminated if innovation management cannot find business value for that particular case. In addition, a business case may have to be aborted if the innovation management present in the company considers the business case to be valuable but cannot find a business owner who sees value in the case. Business cases are cancelled or delayed if no dataset is available as some companies strictly limit the availability of dataset. On the other hand, a business case is also abandoned when it has sufficient data and seems to make predictions, but the proof of concept does not justify model development. Expensive and difficult data generation and data collection methods also lead to a stage of dissatisfaction, slowness and even termination of business cases. The business case is terminated when the implementation of the ML/DL solution fails because it is not cost-effective or has not been approved by the business owners during the validation of the business case.

5.4 | Iterations and triggers

Based on empirical findings, we identified multiple iterations among the high-level activities of the AI-driven business development framework as shown in Figure 2. Although following the framework helps in establishing AI-driven business development, using iterations can help in building high-performing ML/DL models by ensuring better predictions, efficient inferences and high business value. Below, we outline high-level activities, events that trigger these iterations, solutions to optimize the iterations and indicate whether evaluation metrics, data, features or the model itself change when these iterations are triggered. Figure 3 illustrates the iterations.

1) *Business case validation to Selection and specification of prioritized business case*:
   The iteration proceeds from business case validation to selection and specification of business case. In exceptional cases, the business owners propose new/additional evaluation metrics during the session dedicated to the validation of the already agreed and defined metrics of the business case. These new/additional metrics are never communicated to the data scientists in the initial phase of the project. This leads to a high cost and time loss. Therefore, this iteration is triggered to specify the already existing business case with new metrics.
2) *Experimentation to Feature engineering*:
   This iteration leads from experimentation to feature engineering. After experimenting with a simple base model, data scientists examine the results and try to understand the relevance of the features and the correlation between them. This iteration is useful when it is cost-effective to obtain the features. Therefore, this iteration is triggered to get a proper understanding of the system and feature selection. This triggering can provide further insight in complicated data science problems.
3) *Deployment to Experimentation*:
   Iteration extends from the deployment to the experimentation phase. Once the model is in production, it makes real data predictions. We find that data scientists investigate the state-of-the-art learning to find better learning algorithms to replace a well-performing deployed model. Replacement takes place when the domain or requirements change over time, for instance, to optimize prediction accuracy, hardware cost, explainability, latency and prediction time. This iteration is only triggered when data scientists confirm that the new replacement algorithms perform even better than the old deployed model and increase business value. In such cases, the model needs to be trained with a new learning algorithm on existing data and features and finally proceed with the redeployment of the new model.

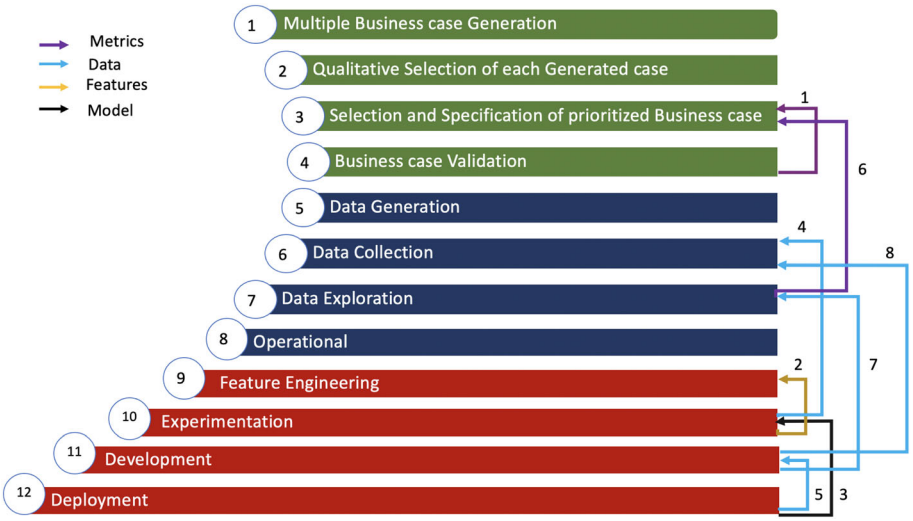

FIGURE 3 Iterations to optimize design model

4) *Experimentation to Data collection*:
Iteration proceeds from experimentation to data collection. Data scientists validate the discrepancies between two datasets if the algorithm underperforms when experimenting with algorithms published in previous works. If discrepancies exist, the data collection process/collected data can be verified to ensure data quality. This iteration is triggered when it is believed that verifying the dataset and correcting if necessary will improve the performance of the model than when experimenting with a new algorithm.
5) *Deployment to Development*:
Iteration is from deployment to development phase. We find that deploying the model to production is not a one-time activity in most software-intensive embedded systems companies. Rather, it is a continuous process. The model makes accurate predictions when the data used for prediction and training have a similar data distribution. To mitigate data drifts or keep the model up to date, this iteration is triggered. The simplest solution is to retrain the model with new data and validate the model to ensure that the model still produces accurate results. The learning algorithm and hyperparameter space remain the same as the trained model when retraining with new data.
6) *Data exploration to Selection and specification of prioritized business case*: The iteration goes from data
exploration to business case specification. Until data exploration, it is usually difficult to define the right business goal for the project. This is because the expectations of the business owners during the discussions do not match what can be achieved with their data. This is due to the knowledge gap between people who have a data science background and those who do not. As a result, this iteration to fine-tune the business case is triggered or even leads to rejection of the business case after data exploration.
7) *Development to Data exploration*:
The iteration goes from the development to the data exploration phase. This iteration is triggered when it is suspected that there are opportunities to improve the performance of the model by re-exploring and understanding the part of the data where the model underperforms. If interesting aspects are found after exploring data sets, extract the features and add them to the existing feature set, build and train the model from scratch and proceed with deployment.
8) *Development to Data collection*:
The iteration goes from development to data collection. This iteration is triggered by an attempt to fine-tune the experimented algorithm that produces good results, thinking that introducing more variances into the dataset or refining the annotations will optimize the performance of the model. In this iteration, the model is built and trained using the updated dataset and features. Once the model is finalized, it is deployed and later put into operation for monitoring and logging.

5.5 | Correlation between framework and challenges

As shown in the above sections, there are several challenges that practitioners face in the development, deployment and evolution of ML/DL models. Based on the generalizations of the practices and experiences of practitioners in the case companies, we have developed a framework to

improve their ML/DL development and deployment practices. In this way, we also aim to help mitigate the challenges they face. Below, we describe how the framework helps to address the identified challenges.

Data scientists can work with developers or help implement the necessary tools in companies to generate data suitable for ML/DL business cases. Otherwise, they can ensure that the datasets are available with clients before they start the business case. The following iteration from development to data collection in Section 5.4 provides the opportunity to revise the dataset to introduce variance or refine the labeling to make the dataset more valuable and representative. If obtaining features is cost-effective, data scientists can experiment with the selected feature set and consider a larger feature set with better understanding based on the results. As data scientists work simultaneously and in parallel with data science teams, they can gain knowledge by participating in different ML/DL projects. They can also collaborate with domain experts to understand the business case domain and label the dataset. Scheduling stand-up meetings, sprint meetings, participating in demos, internal workshops and so forth can ensure that all team members are aware of each step in the development, deployment and evolution of models for a particular business case in the companies. This can ensure that the models are understandable and avoid risks when data scientists are assigned to other projects in the middle of working on a particular project. Working with different practitioners on different projects can reduce misunderstandings about ML/DL models among data scientists. For instance, DL models are complex compared with ML models. Adopting DL models can reduce the significant time spent on feature engineering. Ensuring that training and test data go through the same process of data collection, cleaning and have the same formats would avoid training-serving bias. Also, retraining models with more data in an appropriate format can reduce training-serving skew to some extent. Making sure that corner cases are also collected when collecting datasets can reduce this skew. When model drifts occur, use the experimentation stage and experiment with a better algorithm if the drift is due to a change in technology or domain. A proof-of-concept can reduce the time required for model development and give a realistic idea of what can be achieved with the business case. Based on the proof-of-concept, data scientists and domain experts can be assigned to projects with high value. Because business owners are also involved in experimenting the business case, communication with business owners can be improved by asking them to fill in questionnaires, participate in frequent meetings and train them.

6 | RELATED WORK

ML/DL systems differ from traditional software in the following ways: (a) Because ML/DL models learn from data, collecting, processing and updating data takes time, (b) In addition to SE skills, teams need deep ML/DL knowledge to develop high-performing models and (c) Unlike traditional software development, deploying models requires additional effort. Furthermore, the mechanisms for monitoring and logging are unique to ML/DL models. According to Yang et al,[5] the development of ML models involves experts, intermediate users and amateurs. Amateurs are non-experts who are not familiar with the ML technology. Very little research has been done on the work of intermediate users and amateurs. According to Hill et al,[10] experienced programmers had difficulty using ML. Patel et al[6] examined the implementation of ML systems by software engineers. Several tools are available to support people working with ML/DL technologies. Although ML/DL offers powerful tools, the knowledge required to apply these tools to specific scenarios remains a challenge. Smaller non-profit companies have extensive local and domain expertise but lack the skills to implement ML/DL in their context.[47]

Wang et al[49] mentioned the current working practices of data scientists and the impact of automated AI (AutoAI) on these practices. Muller et al[48] suggested approaches to data processing used by data scientists. In Hirt et al,[50] a holistic process model is presented that explains the activities, initiation, error estimation and deployment of a supervised ML classification model. The ML workflow[35] has nine stages, including data-oriented and model-oriented stages. A comprehensive overview of the ML process for developing intelligent systems is described.[10] Data management, model learning, model verification and model deployment are the activities associated with the four stages of the ML workflow.[51] These are similar to data science workflows such as cross-industry standard process (CRISP-DM)[14] and knowledge discovery databases (KDD).[13] CRISP-DM enables companies to apply data mining in real-world scenarios regardless of the technology. KDD is used to discover knowledge from data using data mining tasks. Both workflows have a data-centric focus with multiple feedback loops between each stage.

People working on the development, deployment and evolution of ML/DL models face several challenges. Lwakatare et al[7] attempts to identify and classify numerous SE challenges encountered in the development and deployment of ML components in software-intensive systems. Experts encounter problems in developing systems that use ML algorithms as described.[10,52] Yang et al[5] explored how non-experts develop ML model based on their experience, knowledge, blind spots and also uncovered their unique potentials and pitfalls. ML specific risks should be considered during system design using the SE technical debt framework.[8] Some of the technical debts are boundary erosion, hidden feedback loops, data dependencies, configuration debts, changes in the outside world and so forth. Studies have been conducted to test software[53] and ML models. Challenges in the intersection of SE and ML were described.[54] However, studies looking at the combination of SE and ML have not been extensively researched.[55] Similar to software, the implementation of a production-ready ML system needs to be tested[56] and can be divided into development, production and company-specific challenges. Unlike ML and traditional software development, DL technology requires dedicated infrastructure support[57] as scaling DL models increases performance to a higher degree.[57]

7 | THREATS TO VALIDITY

We focus on mitigating validity threats by focusing on construct validity, reliability and external validity.[58] To improve the construct validity of our research study, the authors and the practitioners involved are well versed in the development, deployment and evolution of ML/DL models. We used various techniques (interviews, workshops, meetings and events) and sources (senior data scientists, technology specialists, AI application specialists and so on) to collect empirical data. Our study included six case companies, and the results were reviewed by practitioners who were both directly and indirectly involved in our research. To ensure reliability, we validated the findings with practitioners in the companies through workshops, meetings and events. In terms of external validity, the main contribution of the study can be applied to similar companies with software-intensive embedded systems that are interested in developing, deploying and evolving ML/DL models. The empirical findings reported are based on interviews as well as insights and observations from workshops, meetings and events where practitioners shared their subjective perceptions and experiences.

8 | CONCLUSION

Although ML/DL technologies are becoming increasingly popular, companies face several challenges in developing ML/DL models. To address these challenges and help embedded systems companies advance their ML/DL model development and deployment process, this research study focuses on designing an end-to-end process for developing, deploying and successfully evolving ML/DL models. In this study, we identify the high-level activities that companies undertake in parallel and simultaneously to develop, deploy and evolve models. In addition, we describe in detail the activities, iterations and triggers that optimize model design as well as the roles and company functions. We also show how this study helps companies solves challenges we identify and discuss different decision checkpoints for immediately terminating less valuable business cases. In future research, we plan to focus on the continuous delivery of MLOps as well as how the adoption of MLOps affects existing practices in companies and the challenges they face.

ACKNOWLEDGMENTS

We would like to thank the practitioners from all six companies involved in this study for providing their experiences and examples. This work is funded by the Software Center.

DATA AVAILABILITY STATEMENT

Data available on request from the authors. The interview data is not shared as it may contain sensitive company information.

ORCID

Meenu Mary John https://orcid.org/0000-0003-3972-2265
Helena Holmström Olsson https://orcid.org/0000-0002-7700-1816

REFERENCES

1. Bosch J. Digital transformation: a holistic perspective for business leaders. ISBN 978-91-519-2465 –6; 2019.
2. Olsson HH, Bosch J. Going digital: disruption and transformation in software-intensive embedded systems ecosystems. *J Softw Evol Process*. 2020; 32(6):e2249.
3. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. *Science*. 2015;349(6245):255-260.
4. Goodfellow I, Bengio Y, Courville A, Bengio Y. *Deep Learning* (Vol. 1, No. 2). Cambridge: MIT press; 2016.
5. Yang Q, Suh J, Chen NC, Ramos G. Grounding interactive machine learning tool design in how non-experts actually build models. In: Proceedings of the 2018 Designing Interactive Systems Conference; 2018:573-584.
6. Patel K, Fogarty J, Landay JA, Harrison BL. Examining difficulties software developers encounter in the adoption of statistical machine learning. In: AAAI; 2008:1563-1566.
7. Lwakatare LE, Raj A, Bosch J, Olsson HH, Crnkovic I. A taxonomy of software engineering challenges for machine learning systems: an empirical investigation. *International Conference on Agile Software Development*. Cham: Springer; 2019:227-243.
8. Sculley D, Holt G, Golovin D, et al. Hidden technical debt in machine learning systems. *Adv Neural Inform Process Syst*. 2015;28:2503-2511.
9. Salvaris M, Dean D, Tok WH. Microsoft AI platform. *Deep learning with azure*. Berkeley, CA: A press; 2018:79-98.
10. Hill C, Bellamy R, Erickson T, Burnett M. Trials and tribulations of developers of intelligent systems: a field study. In: 2016 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC). IEEE; 2016:162-170.
11. Machine learning workflow. https://cloud.google.com/ai-platform/docs/ml-solutions-overview, accessed: 2021-03-16.
12. Patel K, Fogarty J, Landay JA, Harrison B. Investigating statistical machine learning as a tool for software development. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; 2008:667-676.
13. Fayyad U, Piatetsky-Shapiro G, Smyth P. The KDD process for extracting useful knowledge from volumes of data. *Commun ACM*. 1996;39(11):27-34.

14. Wirth R, Hipp J. *CRISP-DM: Towards A Standard Process Model for Data Mining. In Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining* (Vol. 1). London. UK: Springer-Verlag; 2000.
15. John MM, Olsson HH, Bosch J. Developing ML/DL models: a design framework. In: Proceedings of the International Conference on Software and System Processes; 2020:1-10.
16. Bosch J, Olsson HH, Crnkovic I. It takes three to tango: requirement, outcome/data, and AI driven development. In: SiBW; 2018:177-192.
17. https://www.gartner.com/it-glossary/digitalization, accessed 2021-03-16.
18. Erich F, Amrit C, Daneva M. A mapping study on cooperation between information system development and operations. In: International Conference on Product-Focused Software Process Improvement. Springer, Cham; 2014:277-280.
19. Lwakatare LE, Karvonen T, Sauvola T, et al. Towards DevOps in the embedded systems domain: why is it so hard? In: 2016 49th Hawaii International Conference on System Sciences (HICSS). IEEE; 2016:5437-5446.
20. Stahl D, Martensson T, Bosch J. Continuous practices and devops: beyond the buzz, what does it all mean? In: 2017 43rd Euromicro Conference on Software Engineering and Advanced Applications (SEAA). IEEE; 2017:440-448.
21. Munappy AR, Mattos DI, Bosch J, Olsson HH, Dakkak A. From ad-hoc data analytics to DataOps. In: Proceedings of the International Conference on Software and System Processes; 2020:165-174.
22. Alla S, Adari SK. What is MLOps? *Beginning MLOps with MLFlow*. Berkeley, CA: A press; 2021:79-124.
23. Mäkinen S, Skogström H, Laaksonen E, Mikkonen T. Who needs MLOps: what data scientists seek to accomplish and how can MLOps help? arXiv preprint arXiv:2103.08942; 2021.
24. Penners R, Dyck A. Release engineering vs. DevOps—an approach to define both terms. *Full-scale Softw Eng*. 2015:49-54.
25. Lwakatare LE, Kilamo T, Karvonen T, et al. DevOps in practice: a multiple case study of five companies. *Inform Softw Technol*. 2019;114:217-230.
26. Tamburri DA. Sustainable MLOps: trends and challenges. In: 2020 22nd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC). IEEE; 2020:17-23.
27. Mitchell TM. Does machine learning really work? *AI Mag*. 1997;18(3):11-11.
28. Bengio Y. Deep learning of representations for unsupervised and transfer learning. In: Proceedings of ICML workshop on unsupervised and transfer learning, JMLR Workshop and Conference Proceedings; 2012:17-36.
29. Bengio Y, Bengio S. Modeling high-dimensional discrete data with multi-layer neural networks. *Adv Neural Inform Process Syst*. 2000;12:400-406.
30. Schmidhuber J. Deep learning in neural networks: an overview. *Neural Netw*. 2015;61:85-117.
31. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Adv Neural Inform Process Syst*. 2012;25:1097-1105.
32. Radford A, Metz L, Chintala S. arXiv preprint arXiv:1511.06434; 2015.
33. Jones N. Computer science: the learning machines. *Nat News*. 2014;505(7482):146.
34. Efrati A. How "deep learning" works at apple, beyond; 2017.
35. Amershi S, Begel A, Bird C, et al. Software engineering for machine learning: a case study. In: 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP). IEEE; 2019:291-300.
36. Hazelwood K, Bird S, Brooks D, et al. Applied machine learning at facebook: a datacenter infrastructure perspective. In: 2018 IEEE International Symposium on High Performance Computer Architecture (HPCA). IEEE; 2018:620-629.
37. https://towardsdatascience.com/demystified-ai-machine-learning-deep-learning-c5259d38678e. accessed 2021-03-16.
38. https://acerta.ai/blog/artificial-intelligence-industry-4-0-5-manufacturing-applications-for-ai/. accessed 2021-03-16.
39. Dahlmeier D. On the challenges of translating NLP research into commercial products. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers); 2017:92-96.
40. Runeson P, Höst M. Guidelines for conducting and reporting case study research in software engineering. *Empir Softw Eng*. 2009;14(2):131-164.
41. Walsham G. Interpretive case studies in IS research: nature and method. *European J Inform Syst*. 1995;4(2):74-81.
42. Yin RK. *Case Study Research and Applications: Design and Methods*: Sage publications; 2017.
43. Holton JA. The coding process and its challenges. *Sage Handbook Grounded Theory*. 2007;3:265-289.
44. Wilson V. Research methods: triangulation. *Evid Based Libr Inform Pract*. 2014;9(1):74-75.
45. Stake RE. *The Art of Case Study Research*: Sage; 1995.
46. Rana R, Staron M, Hansson J, Nilsson M, Meding W. A framework for adoption of machine learning in industry for software defect prediction. In: 2014 9th International Conference on Software Engineering and Applications (ICSOFT-EA). IEEE; 2014:383-392.
47. Bopp C, Harmon E, Voida A. Disempowered by data: nonprofits, social enterprises, and the consequences of data-driven work. In: Proceedings of the 2017 CHI conference on human factors in computing systems; 2017:3608-3619.
48. Muller M, Lange I, Wang D, et al. How data science workers work with data: discovery, capture, curation, design, creation. In: Proceedings of the 2019 CHI conference on human factors in computing systems; 2019:1-15.
49. Wang D, Weisz JD, Muller M, et al. Human-AI collaboration in data science: exploring data scientists' perceptions of automated AI. In: Proceedings of the ACM on Human-Computer Interaction, 3(CSCW); 2019:1-24.
50. Hirt R, Koehl NJ, Satzger G. An end-to-end process model for supervised machine learning classification: from problem to deployment in information systems. In: Designing the Digital Transformation: DESRIST 2017 Research in Progress Proceedings of the 12th International Conference on Design Science Research in Information Systems and Technology. Karlsruhe, Germany: Karlsruher Institut für Technologie (KIT); 2017:55-63.
51. Ashmore R, Calinescu R, Paterson C. Assuring the machine learning lifecycle: desiderata, methods, and challenges. arXiv preprint arXiv:1905.04223; 2019.
52. Sculley D, Holt G, Golovin D, et al. Machine learning: the high interest credit card of technical debt; 2014.
53. Kanewala U, Bieman JM. Testing scientific software: a systematic literature review. *Inform Softw Technol*. 2014;56(10):1219-1232.
54. Arpteg A, Brinne B, Crnkovic-Friis L, Bosch J. Software engineering challenges of deep learning. In: 2018 44th Euromicro Conference on Software Engineering and Advanced Applications (SEAA). IEEE; 2018:50-59.
55. Breck E, Cai S, Nielsen E, Salib M, Sculley D. What's your ML test score? A rubric for ML production systems; 2016.
56. Murphy C, Kaiser GE, Arias M. An approach to software testing of machine learning applications; 2007.

57. Dean J, Corrado GS, Monga R, et al. *Large scale distributed deep networks*; 2012.
58. Easterbrook S, Singer J, Storey MA, Damian D. Selecting empirical methods for software engineering research. *Guide to advanced empirical software engineering*. London: Springer; 2008:285-311.

How to cite this article: John MM, Olsson HH, Bosch J. Towards an AI-driven business development framework: A multi-case study. *J Softw Evol Proc*. 2023;35(6):e2432. doi:10.1002/smr.2432

APPENDIX A: Interview protocol

Below, we provide the interview guide that served as the basis for our discussions with company practitioners during the interviews. For each question, we provided interviewees the opportunity to add their input and insights as well as information they found was important but was not covered in the question. The questions worked as a foundation for discussing and reflecting on the challenges as well as opportunities that the company practitioners experienced. It should be noted that the questions were used as a basis for the discussions and not as a protocol to control the conversation.

Introduction questions

1. Background and contextual information of the interviewee
   - Role in the industry
   - Previous work
   - Current work
   - Overall experience
   - Assigned responsibilities
   - Current state of use of ML/DL model in the company
   - Use cases

Main questions

1. How do you select/ensure/choose or define a business case? What are the typical activities involved in selecting a business case?
2. In what ways are the various stakeholders involved (who are they, what roles do they play, what is the interaction, etc.)? Are there challenges involved in interacting with the stakeholders? If so, what are typical challenges?
3. How/when do you consult domain experts to understand the data? Have you encountered a situation where the lack of domain experts impacts data exploration? If so, can you provide an example?
4. How do you ensure that sufficient data is available and accessible to implement the business case? Who will provide you with the data set?
5. When dealing with noisy, unlabeled, partially labeled or data privacy concern, how did you explore the data? With examples?
6. Who is responsible for labeling in the company? Data scientists, domain experts or interns?
7. How do you handle feature engineering? Would you scale up or scale down features to the feature set?
8. Have you come across a situation where features that are indirectly dependent on the dataset, when added to the feature set increase model performance?
9. How do you deal with -
   - If you find two features with the same information? How do you decide which one to choose? How do you go about in such a situation? Are there any challenges associated with it?
   - If a feature is outside the training range and the model behaves unpredictably?
10. Have you experienced the following situations? If so, what are the typical challenges that occur?
    - The model does not provide the accuracy you were looking for?
    - The model performs poorly on a given data set, i.e., it performs well 80% of the cases but poorly for 20% of the cases?
    - The model suffers from robustness problems?
11. How do you find the current best available algorithms (state-of-the-art) for a given business case?
12. Do you use automated tools for experimentation? If so, what are they?
13. Do you have a preference between ML and DL models? If yes, why? And when?

14. How would you describe the competence level and access within your company regarding the development of ML/DL models? How do you welcome new people and what support is available for them to acquire the skills to develop ML/DL models? What are the key skills people need (in your opinion) for effective ML/DL model development?
15. We believe that the choice of the final model depends on the end goals. Which end goals do you place more importance on? What factors guide your decision to choose a model?
16. How do you handle hyperparameter tuning? Rough estimate or based on literature review or previous experience? Are there cases where tuning the hyperparameters increases the model performance to a greater extent?
17. What are the requirements placed on the ML/DL models before they are deployed?
18. Have you encountered the following situations? If so, what were the implications of these?
    - Problems integrating ML/DL components into software-intensive systems.
    - Internal deployments due to the strong disconnect between the customer network and the organization
19. When did you decide to retrain or redeploy a particular model? What was the reason for this decision?
20. What do you do when you encounter instances of -
    - Training-serving skew?
    - Model drifts?

Finally, is there anything you would like to add in relation to the above questions that you think is important for the development, deployment and evolution of ML/DL models?