Machine Learning Application Development Guidelines Using CRISP-DM and Scrum Concept

Muhammad Tamiramin Hayat Suhendar
School of Electrical & Informatics Engineering
Institut Teknologi Bandung
Bandung, Indonesia
13519129@std.stei.itb.ac.id

Yani Widyani
School of Electrical & Informatics Engineering
Institut Teknologi Bandung
Bandung, Indonesia
yani@informatika.org

2023 IEEE International Conference on Data and Software Engineering (ICoDSE) | 979-8-3503-8138-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICODSE59534.2023.10291438

***Abstract*— In the field of data mining, machine learning (ML) has been utilized in the search for solutions to various problems. One widely used model process for ML application development is the Cross Industry Standard Process for Data Mining (CRISP-DM). On the other hand, Scrum has emerged as the most popular agile method for software development in recent years. In this research, we proposed an ML application development guideline for data mining by incorporating relevant Scrum concepts into CRISP-DM. The process involves analyzing CRISP-DM and the development situation through interviews with experienced ML software developers. Furthermore, an analysis of the implementation of Scrum concepts in CRISP-DM is conducted. The proposed guideline is represented in Essence and tested through a case study, qualitative evaluation, and evidence map. The evidence map is used to analysis the importance of proposed guideline components is examined. The results indicate that the proposed guideline can be utilized to assist in the development of ML software.**

Keywords—machine learning; CRISP-DM; scrum; software development; guideline; Essence

I. INTRODUCTION

Artificial intelligence is a field of study that seeks to enable machines to perform tasks requiring human intelligence [1]. Machine learning (ML) is part of the field of artificial intelligence. ML focuses on enhancing the performance and knowledge of machine learning models as they evolve. Enhancements in performance and knowledge typically involve data analysis [2]. Data mining is the practice of identifying significant correlations, patterns, and trends through the utilization of statistical techniques, machine learning (ML), and data visualization on vast amounts of data [3]. When conducting the data mining process, a widely utilized model process called the Cross Industry Standard Process for Data Mining (CRISP-DM) is commonly employed [2]. The CRISP-DM process model encompasses the general development process of machine learning within data mining.

Various software development process models have been developed since the 1960s. In 1970, the sequential waterfall SPDM emerged. By the early 2000 agile emerged which iterative and adaptable to user needs [4]. The most popular framework from agile methodologies is Scrum. During each sprint, the product backlog is broken down into smaller tasks, known as the sprint backlog. Scrum ceremonies and artifacts further assist developers in dividing tasks into smaller steps and monitoring the progress of development.

Situational method engineering (SME) is one of the disciplines in software engineering that focuses on designing, tailoring, adapting methods, techniques, and tools for system development [5]. In SME, methods are designed with consideration given to the software development situation. Subsequently, method components are selected from the available options that best fit the software development situation [5]. The Object Management Group (OMG) has developed Essence as a standard for method representation. Essence provides software engineers with a shared perspective on methods [6]. Considering the incorporation of certain Scrum concepts in software development and the widespread adoption of the CRISP-DM model process in data mining, there is an opportunity to create a more detail guide that can assist developers in building machine learning software.

II. LITERATURE REVIEW

A. Machine Learning

Machine Learning is derived from various disciplines, including artificial intelligence, probability and statistics, the complex theory of computation, philosophy, psychology, and neurobiology. The "learning" in machine learning refers to a computer program's ability to learn from experience [7].

When it comes to data extraction and data modeling, the techniques employed are commonly referred to as machine learning. ML algorithms are used to enhance the performance of ML models over time. The development process of data mining has been standardized and outlined by the Cross Industry Standard Process for Data Mining [2], which can be observed in Figure 1.

1. Business Understanding is vital to understand the problem to be solved.
2. Data Understanding, the data serves as the available raw material from which the solution will be constructed. It is crucial to understand the strengths and limitations of the data.
3. Data Preparation often proceeds along with data understanding, where the data is manipulated and transformed into formats that yield better results.

4. Modelling, the output of this stage is some sort of model or pattern capturing regularities in the data.
5. Evaluation is to assess the data mining results rigorously and to gain confidence that they are valid and reliable before moving to deployment.
6. Deployment, the result of data mining is put into real use and involves implementing a predictive model in some information system or business process.

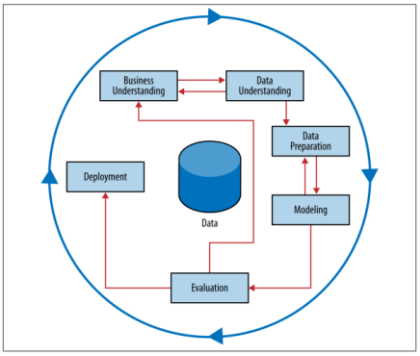

Figure 1 CRISP-DM

B. Scrum

Scrum is a lightweight framework that helps people, teams, and organizations generate value through adaptive solutions to complex problems [8]. Scrum is proposed incompletely, it only defines the essential component to implement Scrum theory. Scrum takes an iterative and incremental approach to optimizing risk prediction and control. Scrum consists of three primary components, namely Scrum team, Scrum ceremony, and Scrum artifacts [8].

The Scrum team consists of a Scrum Master, a Product Owner, and developers. Each member of the Scrum team must possess the necessary skills to deliver value during each sprint. Typically, a Scrum team comprises fewer than ten people. With fewer members, the team can communicate more effectively and be more productive. In Scrum ceremonies, the sprint becomes the container for all ceremonies. The ceremonies of Scrum are divided into four, there is 1) sprint planning, 2) daily scrum, 3) sprint review, and 4) sprint retrospective.

Scrum artifacts represent work or value. They are designed to maximize information transparency so that all members can view and adapt them. There are three artifacts, product backlog, sprint backlog, and product increment [8].

C. Situational Method Engineering

Method Engineering (ME) is a discipline that encompasses the design, construction, and adaptation of methods, tools, and techniques for system development. Situational Method Engineering (SME) refers to the application of ME in specific contexts. Both ME and SME emphasize the formulation of methods for system development. SME constructs methods using method chunks stored in a repository or method base [5].

One technique for modelling the situation is the approach proposed by [9]. Their approach involves assigning a value to each factor of the development situation in the context of specific method engineering. Modelling the situation encompasses four primary facets: organizational facet, human facet, development strategy facet, and application domain facet.

D. Essence

Essence is a software development standard developed by the Object Management Group (OMG). It provides a universal language for defining methods and best practices in software engineering, with a specific focus on software engineering methods [6]. Essence also encompasses its own architectural method. The architecture consists of methods, practice, the Kernel, and Essence language is a DSL (Domain Specific Language). The kernel is described using the kernel language, which focuses on three key areas: alphas, activity space, and competencies [6].

III. Proposed Guidelines

A. Improving CRISP-DM

The Cross Industry Standard Process for Data Mining (CRISP-DM) divides the stages of software development into six key phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. CRISP-DM elucidates these stages by offering an overview of the objectives for each phase, as outlined in Section I.D.

In facilitating the decomposition of goals into smaller steps and managing those steps, Scrum offers an opportunity to complement this process. Scrum elucidates task management and tracks tasks through ceremonies and artifacts. It captures goals in the product backlog, which are further decomposed into a sprint backlog, consisting of tasks for a single sprint. By incorporating the Scrum concept into CRISP-DM, it can provide developers with guidance in constructing machine learning software, thereby enabling more detailed work steps, and aiding in task management to enhance developer productivity.

B. ML Application Development Situation

The analysis of the Machine Learning software's development situation is elaborated upon, grounded in real-world developmental conditions. Following this, a method design framework is employed to conduct a structured situational analysis, accounting for the nuances of the development conditions. The interviews were conducted virtually, utilizing one-on-one video conferencing sessions. A total of five seasoned developers with prior experience in ML software development participated in the interviews, providing invaluable insights into the practical developmental conditions. Among these developers, three were engaged in research-oriented roles, while the remaining two held full-time positions as machine learning engineers within the information technology sector.

Based on the interview results, the team consists of one to eight members, divided into three roles: developers, data scientists, and machine learning engineers. Researcher interviewees have one member in the team, so it must handle

multiple roles. To track progress, the team holds meetings with stakeholders once a week or as necessary. Machine learning developed by six milestone: requirement gathering, data exploration, feature engineering, training, evaluation, and deployment. These milestones align with the six phases of CRISP-DM. The training and evaluation phase is carried out iteratively.

The results of the interviews are then classified into situational factors, which are further divided into four facets, as explained in Section II.4, as identified by [9]. The first facet is the organizational facet. Within the organizational facet, the nature of limited resources is defined, including informational resources and human resources. This limitation arises because obtaining the necessary data for model training is not always feasible, and in cases where the team consists of only one member, that member must handle multiple roles.

In the human facet, relevant characteristics include resistance to conflict, expert roles, and user involvement. Within the application domain facet, machine learning software exhibits high complexity, normal repetitiveness, and falls under the category of intra-organizational applications. As for the development strategy facet, relevant characteristics of machine learning development encompass the source system, development strategy, delivery strategy, and project tracing. The development strategy is defined as iterative and phase-wise, while project tracing is considered weak.

C. Implementing Scrum Concepts in CRISP-DM Stages

The analysis of applying the Scrum concept involves a thorough examination of each stage in the CRISP-DM process model, incorporating Scrum ceremonies and artifacts that can enhance and complement these stages.

1. Business Understanding, it has a goal to understand the problem to be solved. The Scrum concept that can be applied is the product backlog artifact. The product backlog assists stakeholders in documenting the issues they aim to address through a list of software requirements.
2. Data Understanding, it is to comprehensively grasp the strengths and limitations of the data. Scrum concepts can be applied by initiating the decomposition of the product backlog into sprint backlogs. The data understanding stage involves alternating arrow directions with the business understanding stage. Hence, during the data understanding stage, iterations can be conducted until the results are accepted by stakeholders. Scrum iterations can be facilitated through the sprint concept, leading to the production of product increments that stakeholders can directly utilize. However, in the case of data understanding iterations, the results cannot be directly used by stakeholders. Consequently, the full implementation of the sprint concept becomes unfeasible.
3. Data Preparation, the objective of data preparation is to manipulate the structure of the data to achieve improved outcomes. Complementary Scrum concepts include the artifacts within the sprint backlog. Data preparation involves alternating arrow directions with the CRISP-DM modeling stages, allowing for integration with the modeling stages. This provides an opportunity to combine data preparation with the modeling stages.
4. Modeling is the stage where existing machine learning techniques and algorithms are applied. The Scrum concept that can be complemented is the sprint backlog, achieved through decomposing the product backlog. When combined with data preparation, it can be further enhanced by utilizing the Scrum concept of sprints. Like data understanding, the full implementation of product increments resulting from sprints is not feasible.
5. Evaluation is to assess the reliability and validity of the outcomes obtained from the modeling stage. This stage presents an opportunity for integration with both the modeling and data preparation stages. Once the machine learning (ML) model is generated, it undergoes evaluation, and the ML model with the highest performance is selected. The chosen ML model is then reported to the stakeholders. This process can be facilitated by conducting a sprint review within the Scrum framework. If stakeholders do not agree, the stages can be iterated by recombining different data features and algorithms.
6. Deployment, it is to integrate ML models into the software. The Scrum concept that can be complemented is sprint planning, as it becomes crucial to consider the software architecture that will be combined with the ML model at this stage. The outcomes of the deployment are subsequently reported to stakeholders. This process can incorporate the sprint review ceremony in Scrum, along with the handover of the ML software to stakeholders. However, if the software architecture is extensive and the software possesses additional functionalities, the sprint planning can be transformed into a sprint.

D. Process Modeling using Essence

The proposed process guideline is depicted in the form of an Essence. The activities within the proposed guideline are aligned with the Essence activity space. The outcomes of these activities are mapped to alphas. Additionally, the roles specified in the proposed guideline are correlated with the competencies Essence element.

There are four elements within the activity space of the Essence that are not applicable to the proposed guideline. These four elements are: 1) use and 2) operate the system, 3) prepare to do the work, and 4) stop the work. Use and operate the system occurs after deployment, hence they fall outside the scope of the development process. Since the team roles are established at the beginning of the development process, prepare to do the work becomes irrelevant. Similarly, stop the work is irrelevant because, during the maintenance phase after deployment, it can revert to the business understanding stage.

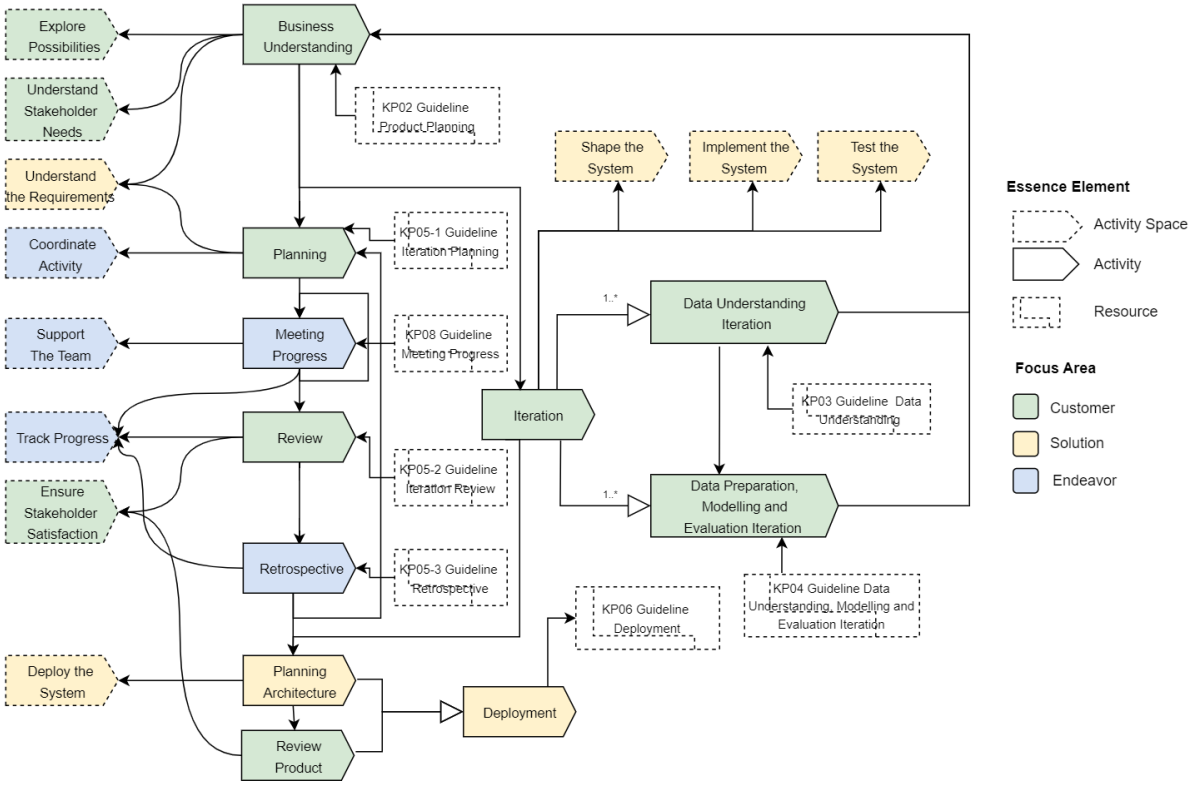

Figure 2 Proposes Activity

E. The Process Model

The proposed guideline consists of four primary activities that are iterative. These stages, in sequential order, include 1) business understanding, 2) data understanding iteration, 3) data preparation, modeling, and evaluation iteration, and 4) deployment. The flow of activities can be observed in Figure 2.

In addition to the business understanding activities, there are other activities that encompass smaller activities within them. These activities are adapted from Scrum ceremonies to fit the CRISP-DM process model. Each activity is accompanied by a resource that serves as a guide for executing the respective activities. Each activity also is linked back to the business understanding activity through an arrow, ensuring that the results of the iterations can effectively address the problem at hand.

Iteration planning and planning architecture incorporate sprint planning from Scrum to break down tasks during the development process. Meeting Progress is adapted from the daily scrum, which may not be conducted daily due to the nature of the development situation, aiming to enhance progress monitoring. Review product and iteration review are derived from the sprint review, which involves collectively assessing the outcomes during iterations. Retrospective originates from a sprint retrospective, serving to further enhance productivity.

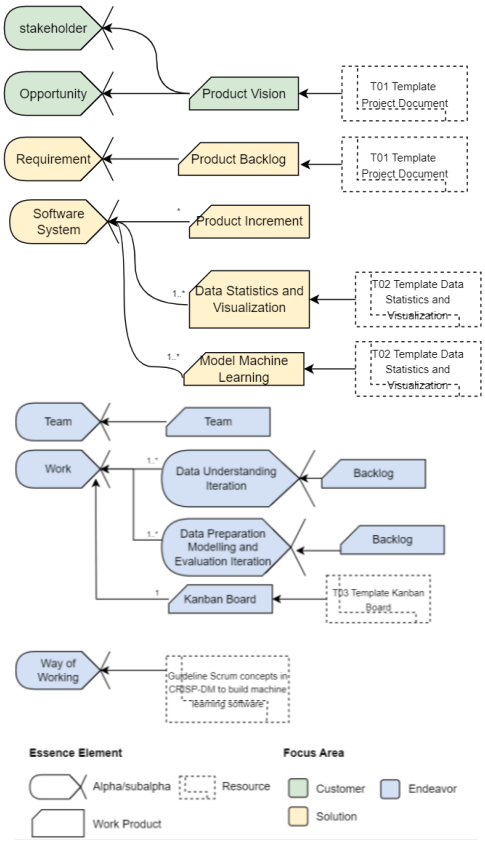

Figure 3 Proposed Guideline Alphas

Within the alphas, there exist work products such as the product vision and product backlog, which are outcomes of the business understanding activities. These work products are from agile and Scrum methodologies to assist stakeholders in comprehending the problem by documenting it in a shared document.

The work product, namely the product increment, is generated through the deployment activities. The statistical data and visualization outcomes are derived from the data understanding iteration activities. ML models are created through data preparation, modeling, and evaluation iteration activities. Some of these work products are accompanied by resources to aid in the development process.

The roles in the proposed guideline are divided into five roles: 1) product owner, 2) scrum master, 3) developer, 4) machine learning engineer, and 5) data scientist. The product owner serves as the representative of stakeholders. Apart from the product owner, all roles possess management skills. Scrum masters possess additional competencies in leadership and analysis. Data scientists are equipped with analysis competence, while machine learning engineers possess competencies in analysis and testing. Developers have development competence.

IV. Evaluation

Testing is conducted to evaluate the established proposed guideline. The series of tests includes case studies of ML software development with the same situation as section III.B. Additionally, qualitative analysis of case study implementations and argumentative assessments using evidence maps are performed. The case study is conducted by the researcher who assumes concurrent roles as the product owner, scrum master, data scientist, developer, and machine learning engineer. The case study involves building software capable of predicting whether employees will remain with the company for more than two years or choose to leave.

A. Qualitative Analysis of Case Study

Qualitative analysis of the case studies was performed to ensure that the produced guidelines can effectively guide the implementation of case study. The analysis was conducted sequentially, aligning with the activities outlined in the proposed guideline. The discussion of work products is tailored to the corresponding activities that generate them.

1. The first activity is business understanding. Stakeholders are provided with assistance in crafting product visions and defining problems, which helps them articulate the purpose of the product being developed more clearly. The product backlog aids developers in decomposing problems.
2. The second is data understanding iteration. During this iteration, the data scientist did not encounter any issues.
3. The subsequent iterations involve data preparation, modeling, and evaluation. These iterations are executed twice because, at the end of the first iteration, the machine learning engineer had not completed generating the ML model. However, since it is part of an iterative process, the unfinished task does not pose a problem and can be revisited in subsequent iterations to complete it.
4. In the two preceding activities, there were additional smaller activities, namely Iteration Planning, Meeting Progress, Iteration Review, and Retrospectives. These activities assist every role involved in managing their tasks effectively. Meeting Progress are conducted once a day to accommodate the time constraints faced by developers. With these adjustments, every involved role comprehended the implications of a two-day workload and encountered no problems.
5. In the final stage, deployment, no issues were encountered. The documentation of each activity helps stakeholders to track the progress of the development process.

B. Examination of Guideline Completeness

The guideline's completeness will be expressed through an evidence map. An Evidence Map is systematic search of a broad field to identify gaps between goals and arguments in user friendly format [10]. Arguments are evaluated based on supporting evidence. The evidence map for the proposed guideline consists of three arguments, which are elements of Essence (alphas, activity spaces, competencies). Each argument is associated with one or more goals. Goals are considered achieved when there is supporting evidence that justifies the arguments. The Evidence Map is depicted using the Goal State Notation (GSN), where squares represent goals, parallelograms represent arguments, and circles represent evidence. The Evidence Map for the proposed guideline can be viewed in Figure 4.

The evidence work product comprises four pieces of evidence, namely product documents, data statistics and visualization notebooks, kanban boards, and the ML Software development team. In the case study, the product document is divided to document the product vision, product backlog, and logs for each activity in the proposed guideline. The product document serves as a reference throughout the development process. Data statistics and visualization notebooks are utilized for data analysis and constructing machine learning models. Kanban boards are employed during development to facilitate task management.

Evidence from activities is categorized into six types of evidence that correspond to the four main stages in the proposed guideline. The deployment activity has two pieces of evidence for planning architecture and product review activities. Additionally, there is evidence for smaller activities carried out during iterations. These pieces of evidence contribute to the data understanding iterations and data preparation, modeling and evaluation iterations, while modeling and evaluation play crucial roles in problem-solving during development. The

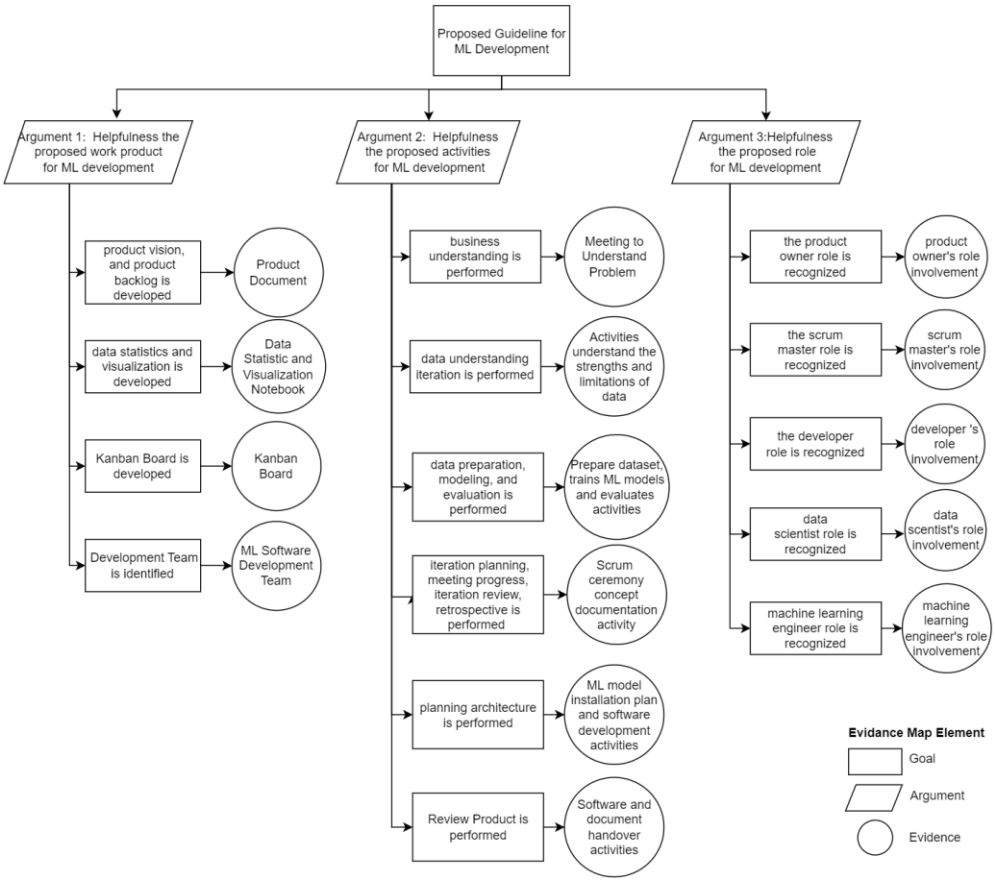

Figure 4 Evidence Map

evidence obtained from planning architecture is plan for installation of ML models and software development. Evidence of competencies stems from the five roles involved in ML software development. The findings indicate that analysis is the most crucial competency, which data scientists and machine learning engineers must possess to effectively address problems.

V. CONCLUSION

The development of the method commenced with an analysis of CRISP-DM, revealing that Scrum could enhance CRISP-DM by providing detailed task breakdown and progress tracking. The development situation was characterized by interviewing experienced ML developers, highlighting aspects such as organizational, human, application domain, and development strategy.

The proposed guideline was evaluated through case study analysis and evidence maps to determine the relevance and importance of its components in development. The findings indicated that the proposed guideline can indeed facilitate the development of ML software. For future research, interviews with various roles, incorporating data variations from additional case studies, and involving external parties in the evaluation process could be conducted to minimize bias.

VI. REFERENCES

[1] S. J. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 4th Edition, New Jersey: Pearson Education, 2021.

[2] F. Provost and T. Fawcett, Data Science for Business, United States of America: O'Reilly Media, 2013.

[3] Y.-P. Sun, Introduction to Data Mining and its Applications, Tamil Nadu: Springer, 2006.

[4] I. Sommerville, Software Engineering, 10th Edition, Pearson, 2016.

[5] B. Henderson-Sellers and J. Ralyte, "Situational Method Engineering: State-of-the-Art Review," *JOURNAL OF UNIVERSAL COMPUTER SCIENCE,* p. 56, 2010.

[6] O. M. G. OMG, Essence – Kernel and Language for Software Engineering Methods, 2018.

[7] T. M. Mitchell, Machine Learning, New York: McGraw-Hill Science/Engineering/Math, 1997.

[8] K. Schwaber and J. Sutherland, The Scrum Guide, Creative Commons, 2020.

[9] E. Komyshova, R. Deneckere and B. Claudepierre, "Contextualization of Method Components," *2010 4th International Conference on Research Challenges in Information Science - Proceedings, RCIS 2010,* pp. 235 - 246, 2010.

[10] I. H. S. S. R. e. a. Miake-Lye, "What is an evidence map? A systematic review of published evidence maps and their definitions, methods, and products," *Systematic Reviews,* 2016.