Journal of Theoretical and Applied Information Technology
15th December 2025. Vol.103. No.23
© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

EVALUATION METRICS OF MACHINE LEARNING OPERATIONS (MLOPS)

NUR FARAH AFIFAH AHMAD SUKRI[1], WAN MOHD AMIR FAZAMIN WAN HAMZAH[2], MOHD KAMIR YUSOF[3], ISMAHAFEZI ISMAIL[4], HARMY MOHAMED YUSOFF[5], AZLIZA YACOB[6]

[1,2,3,4]Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Terengganu, Malaysia

[2]Artificial Intelligence Research Centre for Islam and Sustainability (AIRIS), Universiti Sultan Zainal Abidin, Terengganu, Malaysia

[5]Faculty of Medicine, Universiti Sultan Zainal Abidin, Terengganu, Malaysia

[6]University College TATI, Kemaman, Terengganu, Malaysia.

E-mail: [1]sl4722@putra.unisza.edu.my, [2]amirfazamin@unisza.edu.my

ABSTRACT

Machine Learning Operations (MLOps) has emerged to address challenges associated with deploying, integrating, monitoring, and scaling machine learning (ML) models in production environments. However, to effectively evaluate how well MLOps improves ML models, it is important to have clear and standardised evaluation metrics for such measurement. This study introduced thematic analysis to compile and map evaluation metrics from previous research, select the most relevant criteria and then define a comprehensive set of metrics to assess MLOps implementations. The key findings presented in this article focus on eight main themes: data management, automation and pipelines, model performance, resource and time efficiency, deployment and scalability, usability and collaboration, monitoring and observability, as well as compliance and security.

Keywords: *Machine Learning (ML), Machine Learning Operations (MLOps), Machine Learning Evaluation Metrics, MLOps evaluation, Thematic Analysis.*

1. INTRODUCTION

This study employed thematic analysis to develop evaluation metrics for MLOps. To understand the progression of this work, it is crucial to first understand key concepts such as Machine Learning Operations (MLOps), Continuous Integration (CI), Continuous Delivery (CD) and thematic analysis. Defining these terms provides the essential context for a shared understanding and establish the foundation necessary to accurately interpret the methods, processes and results presented by this study.

1.1. Definitions

MLOps is a set of practices, tools and processes of machine learning models deployment, monitoring and management in real-world environments. MLOps is based on DevOps ideas to solve challenges related to machine learning workflows, such as data changes, model retraining and performance monitoring [44].

Continuous Integration is a practice where code changes are frequently combined into a shared repository. Each change is automatically tested and integrated to ensure the software remains stable. In MLOps, CI also includes automated testing and validation of machine learning pipelines and models to ensure they work reliably before deployment [45].

Continuous Delivery (CD) builds on Continuous Integration (CI) by automating the deployment of validated code or models to operational environments. It ensures that updated models and pipelines can be reliably delivered with minimal human intervention [46]. In machine learning systems, CD manages not only the code but also model files and data. It ensures that new versions can be safely deployed with options to roll back if needed and to scale deployments as required [47].

Thematic Analysis is a qualitative research method used to identify, analyse and report patterns or themes within data. [41] highlight that Thematic Analysis is practically not only about summarising

data but also involves interpreting and providing a meaningful understanding of the underlying ideas or concepts of the data. Methods implemented in TA includes six phases: familiarizing with data, generating initial codes, searching for themes, reviewing themes, defining and naming themes and producing the final report.

1.2. Research Background

Machine learning (ML) is a part of artificial intelligence (AI) that trains computers and machines to learn from data and imitate human learning processes, further improving performance and efficiency. ML is a powerful tool that can analyse large datasets to find hidden patterns that traditional statistical methods may struggle to identify [16,17]. ML can be categorised into two primary types: supervised and unsupervised learning, each with a different learning ability and purpose. Supervised learning involves algorithms learning from labelled data, where input-output pairs are used to guide the learning process. This method is effective for classification and regression tasks, especially when applied to large labelled datasets with clearly defined target variables [2]. In contrast, unsupervised learning trains models on unlabelled data to discover hidden patterns or structures within the data. Common applications of unsupervised learning include clustering and dimensionality reduction [2]. ML has been applied to many real-world applications, such as cybersecurity, healthcare and intelligent transportation systems [3]. These applications showcase ML's ability to solve complex challenges across domains, enhancing efficiency, accuracy and decision-making.

However, the ML model faces several challenges in real-world deployment, which drive the need for Machine Learning Operations (MLOps). Some challenges include inconsistent, incomplete or biased data that can lead to unreliable models [18]. Deploying ML models becomes even more challenging from the difficulties in scaling, integration with older systems and maintenance of reproducibility [19]. Additionally, complex ML model requires continuous training and tuning to maintain their accuracy and relevancy. To address these challenges, MLOps enforce robust data pipelines and governance to ensure data accuracy and relevance. They also provide automated pipelines, unified platforms, continuous monitoring, scalable infrastructure and built-in security, ensuring ML models are reliable, scalable and secure in production.

MLOps is theoretically grounded in frameworks that adapt and extend DevOps, an approach that combines development (Dev) and operations (Ops) to make software production effortless and more reliable. There are two primary DevOps principles: Continuous Integration (CI) and Continuous Delivery (CD). CI is where developer teams integrate their code changes into a shared repository multiple times a day. This is to check on the viability of the new code, ensuring that it aligns with existing functionality and does not introduce bugs or conflicts [20]. Meanwhile, CD involves designing, building, testing and releasing software in short cycles. This approach builds upon CI by automating the release process, ensuring that code changes are not only built and tested but also deployed to production environments. This automation ensures that software updates are consistently delivered in a timely and dependable manner [20].

While DevOps focuses mainly on automating and improving the software development and deployment lifecycle, MLOps extends these practices to handle unique complexities of machine learning. These includes continuous training, model monitoring and testing and versioning of data and models [21].

A crucial aspect of MLOps is creating automated pipelines for tasks like testing and deploying ML software. These pipelines structure the hidden and repetitive processes between teams, thereby increasing the efficiency of the entire machine learning software lifecycle [34]. By emphasising continuous monitoring, updating and retraining of models, MLOps ensure that models maintain their performance and adapt to changing data conditions, which is essential for maintaining their reliability and effectiveness.

To assess the effectiveness of MLOps mosdels compared to static models, a thorough evaluation is essential. Using thematic analysis methodology, this study systematically extracts evaluation criteria from previous research, maps them to related codes and organises these codes into themes for developing a consolidated list of mapped criteria. This process results in the definition of a standardised set of criteria for evaluating MLOps.

2. LITERATURE REVIEW

This section reviews related works on MLOps applications and the evaluation methods used by each work. As MLOps practices continue to

evolve, a standardised evaluation framework and metrics are needed to assess the effectiveness, reliability and scalability of MLOps systems [29]. Evaluating MLOps is not only done by assessing model effectiveness but also covers every stage from data collection to deploying the model in a real-world environment. This includes technical components, such as automated pipelines, management and deployment processes. Additionally, it involves organisational aspects like collaboration, governance and implementation of best practices to ensure smooth ML operations across an organisation. However, as highlighted by [29] and [30], the academic community is still developing thorough and standardised approaches for MLOps evaluation. Much of the current knowledge comes from peer-reviewed research and real-world sources like industry reports, blog posts and industry practitioners.

While MLOps have made things easier, questions remain about how to make these systems as strong and reliable as possible. Hence, a study by [38] was conducted to provide an overview of how trustworthy MLOps systems are. It explained technical methods to make MLOps systems more robust, reviewed current research on making ML systems reliable in real-world use, and looked at the tools and software available for building these systems. The study proposed a "Continuous Monitoring Approaches" to monitor the development performance, including completion time, CPU and GPU usage, memory usage, disk input/output (IO) operations and network traffic.

Several studies have established theoretical frameworks that describe MLOps as a practice that is derived from DevOps principles combined with machine learning lifecycle management. According to [49], MLOps emphasises automation, collaboration between data scientists and operations teams, model management, continuous monitoring and governance to ensure reliable ML deployment and maintenance. These foundational principles include CI/CD automation, reproducibility, workflow orchestration and versioning of data, models and code, which differentiate MLOps from traditional DevOps by addressing the challenge of data-driven machine learning models.

[47] provide a comprehensive framework that integrates these principles with thorough designs, functions and workflows, highlighting the importance of automating not only deployment but also model training, testing and process management. Their work discusses challenges related to data quality, model drift and the need for continuous evaluation, highlighting the repetitive and collaborative nature of MLOps.

Furthermore, [50] proposes an MLOps maturity model that outlines progression stages from Ad-hoc to Automated MLOps and additional higher stages that represent continuous improvement. Their model defines maturity across five key dimensions: Data, Model, Deployment, Operations and Infrastructure and Orchestration. This framework guides organisations to advance from initial automation efforts to fully integrated pipelines that support continuous monitoring and feedback loops that is very essential in real-world applications of machine learning. Nevertheless, empirical research on implementation challenges, scalability constraints and governance issues in practical deployments remains limited [34].

Complementing academic findings, extensive industry analyses such as those by Neptune.ai provides an extensive overview and critical review of various MLOps tools and platforms. It discusses their strengths and weaknesses, key features and important factors to consider when evaluating these solutions, such as cloud alignment, integration capabilities, and user support. While the article focuses on reviewing current tools, it effectively critiques previous methodologies by highlighting gaps and practical challenges in evaluating MLOps platforms comprehensively [39]. Similarly, The Xebia blog presents a practical framework for evaluating MLOps performance by focusing on people, processes and technology. It highlights the importance of well-defined, cross-functional ML teams and regular user feedback to improve maturity. The blog states that MLOps metrics encompass evaluations for both ML teams and MLOps teams. According to [40], ML teams are a subset of MLOps teams, so assessing ML team performance is essential before evaluating the broader MLOps team. The evaluation of ML teams employs four key metrics derived from DORA (DevOps Research and Assessment): deployment frequency, lead time for changes, change failure rate and time to restoration. For MLOps teams, metrics include platform adoption, user satisfaction, coverage of capabilities and cost per user. These authoritative sources enrich the literature by bridging the gap between theoretical frameworks and operational realities.

Thematic analysis is a widely used qualitative research method that systematically identifies, analyses and reports patterns or themes within data [41]. Thematic analysis is a flexible and valuable research method that can provide a deep and detailed understanding of complex qualitative data [52]. Unlike other qualitative methods that aim to build theories or grounded theory, thematic analysis focuses on finding repeated themes that capture the main idea of the data [22].

Braun and Clarke [41] are recognised as the originators of the widely adopted thematic analysis method. They first introduced their approach in a paper titled "Using Thematic Analysis in Psychology", which has since become the primary reference for this qualitative analysis technique across numerous academic studies. The process in thematic analysis involves a six-step approach: familiarization with data, generating initial codes, searching for themes, reviewing themes, defining and naming themes and reporting [41].

According to [42], thematic analysis originated in psychology as a systematic and rigorous method for examining qualitative data. Its early applications focused on identifying and interpreting meaningful patterns within textual data, making it a valuable tool for exploring human behaviour and experiences. Researchers highlights the importance of reflexivity and thorough coding practice to ensure trustworthiness and validity when interpreting themes [52].

Given its structured approach and ability to extract large qualitative data into clear themes, thematic analysis is particularly suitable for synthesising complex topics like MLOps evaluation criteria. Its iterative process ensures that researchers can refine themes to accurately reflect the data while maintaining transparency and replicability.

According to [31], an assessment is necessary to determine the deployment capability of MLOps systems, considering factors such as maintainability, scalability and deployment cost. In this study, the deployment in the context of production was defined, while relevant dimensions for deployment along the ML operations were introduced. The dimensions included were data integration, data preparation, dimension for ML model and dimension for deployment.

Urias and Rossi [32] evaluated three frameworks of MLOps services in their study, namely Machine Learning as a Reusable Microservice (MLRM), Minerva and Machine Learning in Microservices Architecture (MLMA). The study conducted a qualitative analysis that considers the capacity for sharing resources, the scope of use by users and the use of the cloud environment. According to [32], it is essential to evaluate based on qualities to ensure that the built system meets the stakeholders' requirements. Table 3 presents a result obtained by [32]. A green square represents that the framework fully meets the criteria, a yellow triangle represents partial adherence, whereas a red circle indicates that the framework does not meet the criteria.

In the view of [33], automation is a fundamental aspect of MLOps, especially in enabling the continuous delivery of ML models. In the study, a method of checking how well MLOps systems support automation was created by systematically analysing their Architectural Design Decisions (ADDs). Ordinal regression analysis was then employed to validate the effectiveness of these metrics in predicting real-world assessments. This validation proves that their metrics and approach are valuable contributions to assessing automation in MLOps systems.

There are no established methods or frameworks in existing research and industry to properly evaluate MLOps platforms and tools [34]. Some studies evaluated MLOps platforms based on criteria, such as secure ML production support, pipeline setup complexity and management capabilities. Meanwhile, other industries highlighted platform capabilities in automation, secure pipeline creation, responsible AI support, collaboration and architecture [34].

Zhou et al. [35] developed an ML platform with DevOps capabilities by integrating CI/CD tools with Kubeflow. Their work aimed at providing practical insights and guidance for building efficient ML pipeline platforms in real-world settings. To evaluate the overall platform performance, the resource and time consumption across different pipeline stages were analysed, identifying potential bottlenecks, including GPU utilisation. This approach demonstrated the feasibility of combining established CI/CD tools with Kubeflow to create operational ML pipelines that support continuous model training and deployment. Their study serves

as a valuable reference for practitioners designing scalable and efficient MLOps platforms.

The study by [36] focused on evaluating MLOps tools specifically within the Kubernetes ecosystem, emphasising their integration and compatibility with Kubernetes. The evaluation was conducted in two phases. The first phase provided a broad market overview, identifying and defining the capabilities of currently available MLOps tools. The second phase involved an in-depth, practical assessment of three platforms that are native to Kubernetes: Kubeflow, Pachyderm and Polyaxon. These platforms were evaluated based on functionality, usability, community vitality and performance to understand their real-world applicability and effectiveness.

A comprehensive and integrated MLOps pipeline can enhance the development and operational processes of ML, making them accessible and effective even for organisations with limited technical expertise [37]. This approach presents a complete framework for building an MLOps pipeline from early analysis to implementation and assessment in an actual environment. For evaluation, the study structured the MLOps pipeline into different components to evaluate and select suitable tools for each component. After identifying these key components, an assessment was conducted to determine whether the existing DevOps framework already supports the evaluation of those components. The study then combined a range of widely used ML platforms and tools for the evaluation. Platforms used were: Azure ML, Databricks, MLflow, Kubeflow, Iguazio and Weights & Biases. As for tools, there were numerous tools introduced for each component. Examples of these tools included Labelbox, Scale, Snorkel Flow, Doccano, DVC, Pachyderm, Delta Lake and LakeFS, among others.

Numerous methodologies and tools have been developed to evaluate MLOps systems with each having its own strengths and limitations. Leading platforms such as Amazon SageMaker, Microsoft Azure ML, Google Vertex AI, MLflow and Kubeflow offer comprehensive solutions including experiment tracking, pipeline automation, model deployment and monitoring. These platforms are excellent for workflow orchestration and supporting scalable deployments with a focus on enhancing automation and ensuring reproducibility throughout the machine learning lifecycle.

Strengths of these methodologies include the integration with cloud environments enabling flexible multi-cloud or hybrid deployments, real-time monitoring to detect model drift and data quality issues and robust governance features that address enterprise security and compliance requirements. Additionally, open-source tools like MLflow offer flexible options that support customization, enabling users to tailor workflows to their specific needs.

However, these methodologies also have significant limitations. There is no standard evaluation metric used by each tool. This has led to inconsistent assessment standards across organisations. Moreover, there are not many tools that offer direct support for embedding AI practices. Furthermore, effectively managing the interactions among data, models and code throughout their lifecycles remains a persistent challenge that current frameworks have yet to fully address.

The comparative analysis across these methodologies underscores a need for standardised evaluation criteria. This study addresses these gaps by proposing standardised evaluation metrics derived from previous works through thematic analysis, which systematically extracts and synthesises key themes from prior research.

3. METHODOLOGY

This section outlines the step-by-step process for developing metrics to evaluate MLOps in real-world environments. The thematic analysis approach used in this study follows the detailed step-by-step process described by [48], providing a thorough framework for coding and theme development. A study by [48] presented a six-step thematic analysis technique to guide researchers through a systematic process, including the selection of keywords and quotations, coding, theme development, interpretation and model development.

Figure 1 illustrates the overview of the research process conducted in this study. All the processes involved are explained in the following subsection.

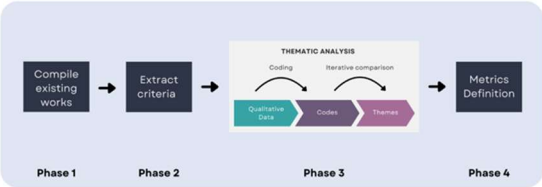

Figure 1: Overview of the Research Process

Phase 1: Compile Existing Works

This first phase is where a review is conducted, with results and outputs of previous related works being compiled. Papers were collected from four different databases: IEEEXplore, ScienceDirect, SpringerLink and ACM Digital Library, focusing on articles published between the years 2020 and 2024. Keywords used in searching for the best articles are: MLOps evaluation, MLOps assessment, MLOps overview, MLOps evaluation metrics, MLOps framework, MLOps application and MLOPs overview. The inclusion and exclusion criteria were then defined to decide whether or not a particular study should be included.

Inclusion criteria

1. Studies on MLOps evaluation
2. Studies written in English
3. Studies published in the last five years (2020–2024)

Exclusion criteria

1. Studies that did not provide clear methodologies or results related to MLOps evaluation
2. Irrelevant titles, abstracts and keywords
3. Review papers

Table 1: List of Existing Works Reviewed

Paper ID	Reference	Paper ID	Reference
P1	[31]	**P6**	[36]
P2	[32]	**P7**	[37]
P3	[33]	**P8**	[38]
P4	[34]	**P9**	[39]
P5	[35]	**P10**	[40]

Phase 2: Extract Criteria Used by Previous Studies

This is where all the criteria used in previous studies are recorded for evaluating MLOps systems. While some studies refer to these as "components" and others as "dimensions," the terminology was standardised in this study using the term "criteria". The keywords were then identified from the criteria to standardise the definition of the criteria. Table 2 lists the extracted criteria from each study and the identified keywords of each criterion.

Table 2: Extracted Criteria and Keywords From Previous Works

Paper ID	Extracted criteria	Keywords
P1	Data integration environment consistency	Data consistency
	Data integration volume of data	Volume of data
	Data preparation performance	Data performance
	Data preparation robustness	Data robustness
	Data preparation explainability	Data explainability
	Data preparation impact	Data impact
	Data preparation computing time	Data computing time
	Modelling performance	Model performance
	Modelling robustness	Model robustness
	Modelling safety	Safety
	Modelling explainability	Model explainability
	Model run time	Run time
	Model storage size	Storage size
	Modelling effort	Effort
	Modelling transfer learning	Transfer learning
	Deployment time restrictions	Time restriction
	Deployment value for user	Value for user
	Deployment cost	Cost
	Deployment complexity	Complexity
	Deployment maturity	Maturity
	Deployment pipeline automation	Pipeline automation
	Deployment scalability	Scalability
	Deployment cyber security	Secure
P2	Problem generalisation	Problem generalisation
	Different programming languages	Languages
	Resource sharing	Resource
	User coverage	User
	Quantitative results	Result
	Suitable for cloud utilisation	Cloud utilisation
P3	Build and deployment scripts	Deployment

	CI/CD pipeline	Pipeline
	Machine learning orchestrator	Orchestrator
	Pipeline triggers on commit	Pipeline
	Pipeline triggers on schedule	Pipeline
	Availability of new training data	Data training
	Model performance degradation	Model performance
	Changes to the data distribution	Changes of data
	Data pipeline	Data pipeline
	Etl pipeline	Pipeline
	Data processing component	Data component
P4	Secure ML production support	Secure
	Pipeline setup complexity	Pipeline complexity
	Artefact management support	Management support
	Platform capabilities in automation	Platform automation
	Secure pipeline creation	Secure
	Responsible AI support	Responsible
	Collaboration	Collaboration
	Architecture	Architecture
P5	Time consumption	Time consumption
	Resource consumption	Resource
P6	Functionality	Functionality
	Usability	Usability
	Vitality	Vitality
	Performance	Performance
P7	Data labeling	Data label
	Data versioning & management	Data version
	Exploratory data analysis	Data analysis
	Feature stores	Features stores
	Code repository	Repository
	CI/CD	CI/CD
	Model training pipeline	Pipeline
	Hyperparameter tuning	Hyperparameter tuning
	Experiment tracking & metadata store	Pipeline
	Model testing & validation	Model performance
	Model registry	Registry
	Model deployment & serving	Deployment
	Monitoring & observability	Monitoring
P8	Continuous validation	Continuous
	Continuous versioning	Continuous
	Continuous monitoring	Continuous
	Continuous update	Continuous
	Data cleaning	Data cleaning
	Data anomaly detection	Data anomaly
	Distribution shift	Distribution shift
	Data augmentation	Data augmentation
	Hyperparameter tuning	Hyperparameter tuning
	Concept drift	Concept drift
	Generalisability	Generalisability
	Label noise	Noise
P9	Cloud and technology strategy	Cloud
	Data sources	Data source
	Data engineering platform	Data engineering
	Code repositories	Repository
	CI/CD pipelines	CI/CD pipelines
	Monitoring systems	Monitoring
	Commercial details	Commercial
	Knowledge and skills in the organisation	Organisation skills
	Key use cases and/or user journeys	User journeys
	User support arrangements	User support
	Active user community	User community
P10	Deployment frequency	Deployment frequency
	Lead time for changes	Lead time
	Change failure rate	Failure rate
	Time to restoration	Restoration time
	Platform adoption	Platform adoption
	User satisfaction	User satisfaction
	Coverage of capabilities	Coverage capabilities
	Cost per user	User cost

Phase 3: Thematic Analysis

After compiling all the criteria and keywords, the steps involved in Thematic Analysis (TA) were undertaken. There were three stages involved in TA. Figure 2 illustrates the thematic analysis methodology employed in this phase.

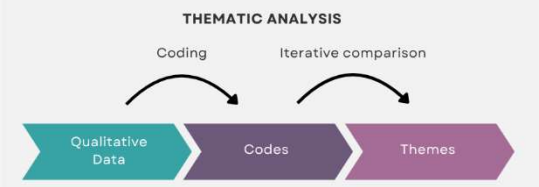

Figure 2: Thematic Analysis Methodology

Step 1: Data collection and keyword selection

Data was collected from relevant sources and carefully reviewed to extract significant keywords that consistently appeared and were contextually relevant to evaluating MLOps. The selection focused on terms capturing important aspects of MLOps practices. For example, keywords like "data consistency," "volume of data" and "data explainability" were identified as recurrent and meaningful within the MLOps evaluation context, indicating key topics of interest.

Step 2: Coding

In this stage, each keyword was assigned a code representing the underlying concept of that data. This involved analysing the semantic meaning and operational relevance of keywords to group related terms under a collective label. For instance, the keywords "data consistency" and "volume of data" were coded as "Data Drift" because both relate to changes and variations in datasets affecting MLOps. Similarly, "data explainability" was coded under "Complexity" as it reflects interpretability challenges.

Step 3: Theme Development

This is where the generated codes were organized into broader themes by identifying conceptual connections and ensuring comprehensive coverage of the data. Codes addressing related dimensions of MLOps were grouped hierarchically under overarching themes. For example, the codes "Data Drift," "Data Quality" and "Data Robustness" were grouped into the "Data Management" theme to show related parts of handling data in ML pipelines. This approach clarifies the relationships among concepts and enhances analytical depth.

Figures 3–10 illustrate the thematic analysis for each theme derived from Table 2 above. Based on the keywords derived from the criteria data compiled from previous works, each keyword was then labelled to form a suitable code. The codes were then grouped according to their definition to generate themes.

As shown in Figure 3, the keywords “data consistency,” “volume of data,” “data training,” “changes of data,” “data pipeline,” “data component” and “distribution shift” were coded under the label “Data Drift.” This grouping reflects how these terms relate to variations and shifts in data over time that can affect model performance, highlighting the dynamic nature of data in operational MLOps systems.

For “Data Quality,” keywords such as “data explainability,” “data impact,” “data performance,” “data label,” “data version,” “data analysis,” “data cleaning,” “data augmentation” and “data source” were categorised together as all these codes emphasise processes and attributes that ensure the quality and usability of data, which are essential for model training and evaluation.

Finally, “data robustness,” “data anomaly,” “noise” and “data engineering” were grouped as “Data Robustness.” This category focuses on the resilience of data handling and the system’s ability to cope with data irregularities and distortion.

Together, Data Drift, Data Quality and Data Robustness codes compose the broader Data Management theme, encompassing key dimensions of managing data challenges in MLOps.

Figure 4 presents the Automation & Pipeline theme. Keywords defining continuous integration and delivery processes were coded as CI/CD Pipeline. “Pipeline automation,” “orchestrator,” “pipeline complexity” and “pipeline” were grouped as Pipeline Automation because they collectively describe the mechanisms and challenges involved in automating pipelines in ML workflow processes. The term “orchestrator” was coded with “pipeline automation” as it specifically refers to tools enabling automatic management of pipeline steps within the workflows. These two codes were combined under the Automation & Pipeline theme to capture both the procedural and systemic automation aspects of MLOps workflows.

Figure 5 shows the Model Performance theme. Keywords such as “model explainability,” “complexity,” “languages” and “concept drift” were coded as Complexity because these terms highlight model interpretability and structural challenges that impact overall reliability. Hyperparameter tuning extracted from [37] and [38] was coded as Hyperparameter Tuning. Terms like “model performance,” “model robustness” and “result”

Journal of Theoretical and Applied Information Technology
15th December 2025. Vol.103. No.23
© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

formed the Model Performance code, capturing the overall effectiveness and stability of the model. Meanwhile, "transfer learning," "problem generalisation" and "generalisability" were categorised as Model Transfer Learning to represent the ability of models to adapt learned knowledge to different tasks or domains. This coding reflects diverse factors influencing model behavior and adaptability.

Figure 6 illustrates thematic analysis for the Deployment & Scalability theme that highlights infrastructure and scaling concerns. Codes that define this theme include deployment and scalability. "Architecture," "deployment" and "deployment frequency" were defined as Deployment to capture implementation and operational aspects of model delivery, while "scalability," "cloud utilisation" and "cloud technology" were defined as Scalability, reflecting the capacity to efficiently expand resources and handle increased workload.

Figure 7 outlines the Resource & Time Efficiency theme. Terms related to computational and storage resources such as "storage size," "resource," "management support," "features," "model registry" and "repository" were coded as Resource Utilisation. Keywords that represent time-related aspects, such as "computing time", "run time", "effort", "time restriction," "time consumption" and "time restoration" were coded as Time Efficiency.

The Usability & Collaboration theme, depicted in Figure 8 was developed from keywords emphasising human and team factors. The term "collaboration", "organisation skills" and "platform adoption" was coded as Collaboration to highlight cooperative activities within MLOps teams. User-centric keywords such as "value for user," "user," "usability," "user journeys," "user support," "user satisfaction" and "functionality" were grouped as User Satisfaction, reflecting user experience and acceptance considerations. "Vitality," "commercial", "coverage capabilities" and "user cost" was coded as Vitality to capture the financial aspects of the system.

Figure 9 presents the Monitoring & Observability theme. Keywords "monitoring" was coded as Monitoring to represent ongoing monitoring of MLOps system health and performance. This thematic distinction enhances understanding of both pre-deployment validation and post-deployment oversight.

Lastly, Figure 10 details the Compliance & Security theme. Keywords such as "safety," "secure," "security" and "responsible" were combined into the Security code, reflecting concerns related to ethical standards, regulatory adherence and protection measures within MLOps.

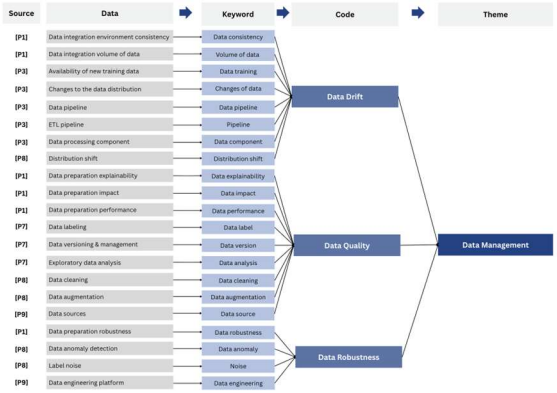

Figure 3: Thematic Analysis for the Data Management Theme

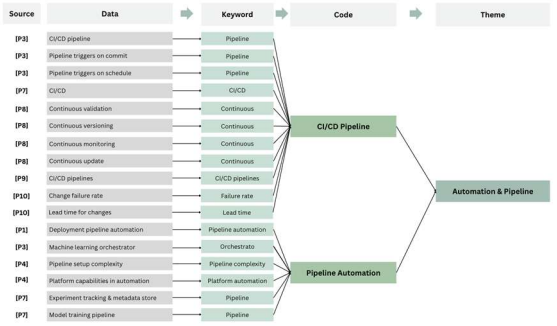

Figure 4: Thematic Analysis for the Automation & Pipeline Theme

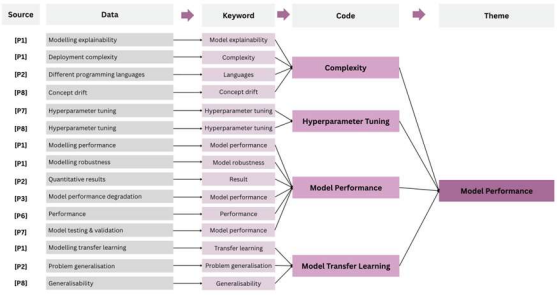

Figure 5: Thematic Analysis for the Model Performance Theme

Journal of Theoretical and Applied Information Technology
15th December 2025. Vol.103. No.23
© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

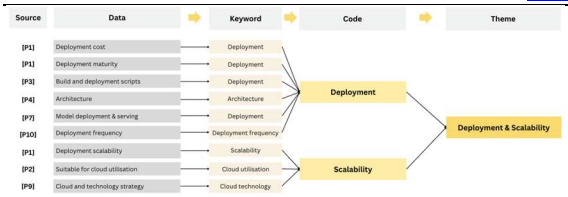

Figure 6: Thematic Analysis for the Deployment & Scalability Theme

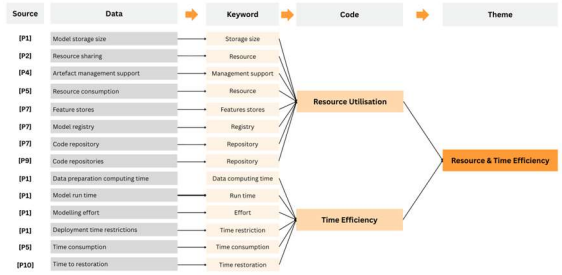

Figure 7: Thematic Analysis for the Resource & Time Efficiency Theme

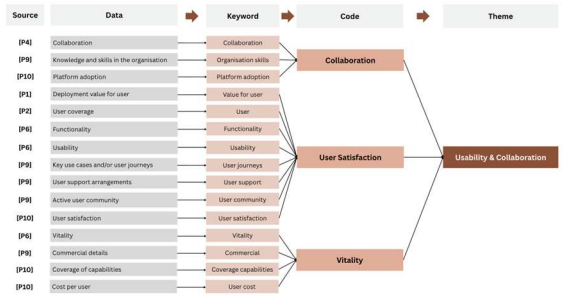

Figure 8: Thematic Analysis for the Usability & Collaboration Theme

Figure 9: Thematic Analysis for the Monitoring & Observability Theme

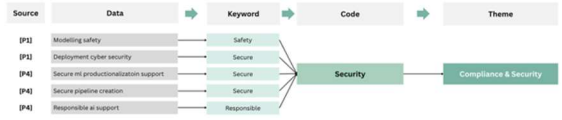

Figure 10: Thematic Analysis for the Compliance & Security Theme

After defining codes and themes for each criterion, it is concluded that resource utilisation is the most frequently mentioned criterion, appearing in 60% of the reviewed papers. CI/CD pipeline, deployment, model performance and user satisfaction followed closely, each cited in 50% of the previous studies. Although hyperparameter tuning, monitoring and security were mentioned in only 20% of the papers, they were included in this study due to their critical roles. Hyperparameter tuning optimises model accuracy and efficiency, monitoring ensures ongoing system health and early issue detection and security protects data and models from breaches, ensuring trustworthiness and compliance. This result is illustrated as Figure 11 below.

The list of MLOps evaluation criteria, along with their corresponding reference papers is summarised and presented in Table 3 below.

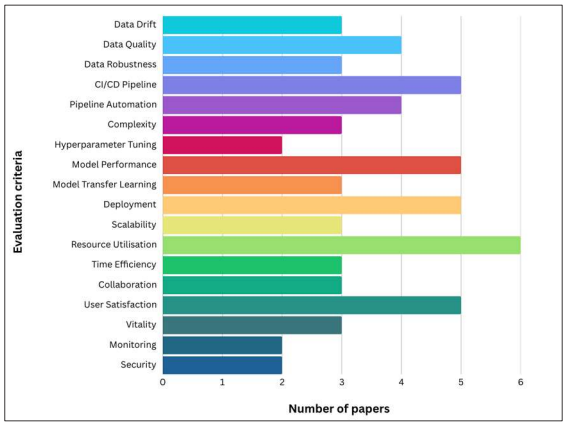

Figure 11: Frequency of MLOps Evaluation Criteria Mentioned in Reviewed Studies

Table 3: Evaluation Criteria with Source Papers

Category	Criteria	Source Papers
Data Management	Data drift	P1, P3, P8
	Data quality	P1, P7, P8, P9
	Data robustness	P1, P8, P9
Automation & Pipeline	CI/CD pipeline	P3, P7, P8, P9, P10
	Pipeline automation	P1, P3, P4, P7
Model Performance	Complexity	P1, P2, P8
	Hyperparameter tuning	P7, P8
	Model performance	P1, P2, P3, P6, P7
	Model transfer learning	P1, P2, P8
Deployment & Scalability	Deployment	P1, P3, P4, P7, P10
	Scalability	P1, P2, P9
Resource & Time Efficiency	Resource Utilisation	P1, P2, P4, P5, P7, P9
	Time efficiency	P1, P5, P10
Usability & Collaboration	Collaboration	P4, P9, P10
	User satisfaction	P1, P2, P6, P9, P10
	Vitality	P6, P9, P10

Monitoring & Observability	Monitoring	P7, P9
Compliance & Security	Security	P1, P4

Phase 4: Metrics Definition

After mapping the criteria, this study proceeded to define specific evaluation metrics for each one. This phase is important because good metrics enable a reliable assessment of MLOps systems across different platforms and use cases. Good evaluation metrics translate qualitative criteria into measurable indicators, allowing fair comparison and evaluation of MLOps implementations.

Defining metrics involves identifying measurable indicators aligned with each criterion, specifying how they are calculated and setting thresholds or benchmarks.

By establishing clear metrics, organisations can comprehensively evaluate the effectiveness, robustness and maturity of their MLOps systems, ensuring that ML products deliver continuous value in production environments. The list of evaluation metrics is tabulated and explained in the results section.

4. RESULTS

This section presents the evaluation metrics defined in this study. Table 4 lists all the evaluation metrics defined to assess various aspects of MLOps systems. In Data Management, data drift detection measures the ability to detect changes in data distribution over time. This is important since data drift can affect model training and reliability. The Population Stability Index (PSI) is used to measure the change in distribution between two samples and is widely used for credit scoring and risk modelling. Data quality is very important in ML as it affects how well models perform, how trustworthy they are and how reliable their predictions will be. If the data is poor or incorrect, the models will also be unreliable. According to a survey made by [43], dimensions involved to evaluate data quality are completeness, self-consistency, timeliness, confidentiality, accuracy, standardisation, unbiasedness and ease of use. Lastly, the Interquartile Range (IQR) method is used for detecting outlier rates to assess the data robustness in this Data Management category. These Data Management metrics contribute to evaluating MLOps by ensuring the foundational data remains reliable, consistent and robust, which directly impacts overall system performance and stability.

The Automation & Pipeline category in MLOps measures how much of the machine learning workflow is automated and how reliable that automation is, focusing mainly on two metrics: automation coverage percentage and failure rate of CI/CD pipelines. This category defines the stability, robustness and maturity from the perspective of automated ML workflow quality and process maturity. High automation coverage with low failure rates reflects a mature, well-managed pipeline infrastructure which is critical for scaling ML operations efficiently.

For Model Performance, standard classification metrics, such as accuracy, precision, recall and F1-score, are used to evaluate model performance. Model complexity is assessed by the model size, typically the number of trainable parameters. Larger models with more parameters typically have higher complexity as they can capture more intricate and varied relationships. The success rate of hyperparameter tuning measures how often hyperparameter tuning leads to improved model performance, indicating how effectively models are optimised. Transfer learning is evaluated by the accuracy gain compared to training models from scratch, especially valuable in data-scarce situations. These metrics evaluate core aspects of model performance and adaptability, providing insights into both basic predictive accuracy and the effectiveness of optimisation processes in MLOps implementations.

In the context of Deployment & Scalability, metrics such as deployment success rate measure the reliability and effectiveness of deploying models into production, while throughput scaling efficiency evaluates how well the system maintains or improves processing capacity as workload increases. Together, these metrics provide insights into the agility and resilience of the deployment process under varying operational demands, highlighting how well models perform and scale in real-world environments.

The Resource & Time Efficiency category evaluates the consumption of computational resources, including CPU and GPU memory usage and time efficiency metrics, such as the number of tasks completed within a given time frame. These measurements capture resource utilisation and operational speed, providing vital insights into how efficiently models train and generate predictions that are critical for optimising performance and managing costs in production environments.

The Usability & Collaboration category assesses user satisfaction through tools like the System Usability Scale (SUS), a widely used, simple

questionnaire designed to measure how usable a system is from the user's perspective. Collaboration can be evaluated using qualitive feedback that based on (Satisfaction, Performance, Activity, Communication, Efficiency) SPACE framework. This dimension checks if the team shares information clearly, helps each other, and solves problems together. Additionally, vitality is measured financially via the cost per user metric, which is calculated by dividing the total operational costs (including infrastructure, labor, and maintenance) by the number of users served. This metric provides insight into the system's efficiency and scalability from a financial standpoint. Together, this category emphasises the critical human and team dynamics influencing successful MLOps adoption and sustained collaboration.

The Continuous Monitoring and Observability category in MLOps evaluates the ongoing health and performance of machine learning models and their underlying data using specific metrics and processes. Dashboard coverage refers to how comprehensively the monitoring dashboard of the system provides visibility into system components. This ensures the models remain reliable and effective after deployment by catching issues early and supporting timely intervention.

Finally, the Compliance & Security category encompasses audit trails, data privacy, access controls and model versioning to ensure compliance with regulatory requirements, apart from maintaining secure and traceable model management.

Together, all these metrics collectively create a comprehensive set of evaluation measures that encompass the technical, operational and organisational dimensions of MLOps systems. Each theme contributes uniquely. Data Management secures input integrity, Automation & Pipeline reflects process maturity, Model Performance captures predictive quality, Deployment & Scalability assesses operational stability, Resource & Time Efficiency ensures practical viability, Usability & Collaboration addresses human factors, Monitoring & Observability ensure ongoing reliability and Compliance & Security uphold governance and trust.

This integrated approach allows for a thorough and balanced assessment by capturing not only the performance and efficiency of the underlying technology but also the effectiveness of workflows, team collaboration and compliance with organisational standards. By addressing these multiple aspects, the evaluation metrics provide a holistic understanding of how well an MLOps system functions in real-world production environments.

Table 4: MLOps Evaluation Metrics

Category	Criteria	Evaluation Metric
Data Management	Data drift	Population Stability Index (PSI)
	Data quality	Completeness, self-consistency, timeliness, confidentiality, accuracy, standardisation, unbiasedness and ease of use
	Data robustness	Outlier detection rates using IQR method
Automation & Pipeline	CI/CD pipeline	Pipeline change failure rate (CFR)
	Pipeline automation	Automation coverage percentage
Model Performance	Complexity	Model size (number of parameters)
	Hyperparameter tuning	Hyperparameter tuning success rate
	Model performance	Accuracy, F1-score, precision & recall
	Model transfer learning	Transfer learning accuracy gain
Deployment & Scalability	Deployment	Deployment success rate
	Scalability	Throughput scaling efficiency
Resource & Time Efficiency	Resource Utilisation	CPU & GPU memory usage
	Time efficiency	Number of tasks completed over time taken
Usability & Collaboration	Collaboration	Collaboration Quality (Qualitative Feedback) evaluated using SPACE framework.
	User satisfaction	System Usability Scale (SUS)
	Vitality	Cost per user
Monitoring & Observability	Monitoring	Dashboard Coverage

Compliance & Security	Security	Audit trails, data privacy & access control

5. DISCUSSION

This study proposes a comprehensive metric for evaluating MLOps by systematically reviewing existing literature, extracting and standardising evaluation criteria, conducting thematic analysis and defining precise metrics aligned to those criteria. In the initial phase, a broad set of existing studies related to MLOps evaluation was compiled, as summarised in Table 1. This literature review provided a foundational understanding of diverse approaches and criteria used by researchers and practitioners to assess MLOps performance and maturity. Next, all evaluation criteria from these studies were extracted, and keywords were identified for each criterion.

Thematic Analysis allowed the organisation of diverse criteria in this study into meaningful codes, which were subsequently grouped into eight primary themes: Data Management, Automation & Pipeline, Model Performance, Deployment & Scalability, Resource & Time Efficiency, Usability & Collaboration, Monitoring & Observability, and Compliance & Security. These themes reflect the nature of MLOps in practice.

This was followed by the definition of clear and measurable metrics that connect the ideas behind the criteria to practical assessments. For example, the Interquartile Range (IQR) helps measure data robustness, automation coverage shows how mature the workflow is and the Population Stability Index (PSI) for data drift. This study covers technical, operational, organisational and human aspects reflecting the complex reality of MLOps in production. Important metrics like deployment success rate, resource use, and user satisfaction (measured by the System Usability Scale) give useful insights for everyone involved. This approach improves how reliably and fairly MLOps systems can be evaluated across different tools and use cases.section.

These evaluation metrics have direct implications for the design, implementation and management of MLOps systems. For instance, recognising data drift through PSI encourages implementation of automated data monitoring and retraining strategies, improving model reliability over time. Automation coverage metrics guide teams in identifying gaps in their workflow automation and encourage them to invest in CI/CD tools that can reduce errors and speed up model delivery. Model performance metrics help engineering teams to track how well models are optimised and guide decisions on model selection, ensuring the deployed models are reliable and easy to understand.

Deployment and scalability metrics like deployment success rate and throughput scaling efficiency help operations teams identify bottlenecks and improve architecture scalability to meet workload demands. Resource and time efficiency metrics help manage costs and plan system capacity by measuring computational usage and processing times. Usability and collaboration evaluations highlight the importance of user-centered design and effective teamwork, which are crucial for sustainable MLOps adoption.

Continuous monitoring metrics help detect model performance degradation and trigger timely interventions, reducing downtime and performance drift in production environments. Compliance and security metrics ensure MLOps systems adhere to regulatory requirements and ethical standards, critical for trust and oversight in sensitive domains.

Together, these metrics provide a balanced and data-driven way to evaluate MLOps systems, encompassing technical quality, operational stability, user experience and governance. Implementers can use these insights to benchmark maturity, prioritize improvements and align MLOps practices with organisational goals, ultimately driving more reliable, scalable and accountable MLOps in real-world settings.

Despite its strengths, thematic analysis has several limitations that should be considered in this study. A key limitation is its interpretative nature, where identifying themes depends heavily on the researcher's perspective. This subjectivity can lead to different findings if other researchers analyse the same data, which may affect the results. Additionally, the lack of a standardised procedure in thematic analysis can cause inconsistencies in how themes are developed and reported. Moreover, thematic analysis can be time-consuming, especially with large or complex datasets, and may have difficulty capturing complex relationships within the data fully.

Regarding the generalisability of this study, the findings are based on a literature review of selected MLOps evaluation criteria and may be influenced by the scope of the included studies. Factors specific to certain industries or organisations may limit how well the results apply to other situations.

Future research could address these limitations by using complementary methods, such as quantitative analyses or mixed-methods approaches, to increase objectivity and depth. Longitudinal studies examining MLOps metric application in real-world deployments would provide empirical validation and insights into practical challenges. Expanding the dataset to cover a wider range of industries and organisational contexts would improve the generalisability of findings. Finally, investigating how different evaluation themes work together and affect MLOps success would improve understanding and help create better practices.

6. CONCLUSION

This study makes a significant contribution to advancing the understanding of MLOps evaluation metrics through the development and application of a comprehensive framework. It thoroughly assesses MLOps systems across essential domains such as automation, deployment, resource efficiency, monitoring and compliance. By synthesising different evaluation methods from previous studies and turning ideas into measurable indicators, these evaluation metrics can help organisations to clearly assess and improve their MLOps maturity, performance and reliability. This comprehensive approach aids in making better decisions, continuous improvement and sustains the success of machine learning operations in real-world production environments. The study significantly advances understanding of MLOps evaluation by providing a structured, multidimensional evaluation framework that can be adapted across different contexts. Future work should focus on expanding these metrics to cover emerging MLOps scenarios and provide precise formulas for each criterion. Ultimately, this work lays a robust foundation for delivering reliable, scalable, and compliant machine learning solutions in production environments, bridging the gap between academic research and practical deployment.

7. ACKNOWLEDGEMENT

This study was funded by the Ministry of Higher Education Malaysia (MOHE) through the Fundamental Research Grant Scheme (FRGS) with project reference code: FRGS/1/2023/ICT06/UNISZA/02/1. A special thanks to the Centre for Research Excellence & Incubation Management (CREIM) of Universiti Sultan Zainal Abidin (UniSZA).

8. CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] Badillo, S., Banfai, B., Birzele, F., Davydov, I. I., Hutchinson, L., Kam-Thong, T., Siebourg-Polster, J., Steiert, B., & Zhang, J. D. (2020). An Introduction to Machine Learning. Clinical pharmacology and therapeutics, 107(4), 871–885. https://doi.org/10.1002/cpt.1796

[2] Shakya, L. M. (2023). Supervised vs. unsupervised machine learning: A comparative analysis. In Proceedings of the GCSSD Conference 2023 (pp. 441–451). ConferenceWorld. http://proceeding.conferenceworld.in/GCSSD-2K23/62.pdf

[3] Jhaveri, R. H., Revathi, A., Ramana, K., Raut, R., & Dhanaraj, R. K. (2022). A review on machine learning strategies for real-world engineering applications. Computational Intelligence and Neuroscience, 2022, Article 1833507. https://doi.org/10.1155/2022/1833507

[4] Kao, Y., Wu, Y.-J., Hsu, C.-C., Lin, H.-J., Wang, J.-J., Tian, Y.-F., Weng, S.-F., & Huang, C.-C. (2022). Short- and long-term recurrence of early-stage invasive ductal carcinoma in middle-aged and old women with different treatments. Scientific Reports, 12(1), 4422. https://doi.org/10.1038/s41598-022-08328-4

[5] Koume, M., Seguin, L., Mancini, J., Bendiane, M. K., Bouhnik, A. D., & Urena, R. (2025). Predicting Fear of Breast Cancer Recurrence in women five years after diagnosis using Machine Learning and healthcare reimbursement data from the French nationwide VICAN survey. International journal of medical informatics, 193, 105705. https://doi.org/10.1016/j.ijmedinf.2024.105705

[6] C. Bosetti, et al., Cancer mortality in Europe, 2005-2009, and an overview of trends since 1980, Ann. Oncol. 24 (2013) 2657–2671.

[7] R. Siegel, K. Miller, N. Wagle, A. Jemal, Cancer statistics, 2023, CA Cancer J. Clin. 73 (1) (2023) 17–48, https://doi .org /10 .3322 /caac .21763.

[8] Ginsburg, O., Yip, C. H., Brooks, A., Cabanes, A., Caleffi, M., Dunstan Yataco, J. A., Gyawali, B., McCormack, V., McLaughlin de Anderson,

M., Mehrotra, R., Mohar, A., Murillo, R., Pace, L. E., Paskett, E. D., Romanoff, A., Rositch, A. F., Scheel, J. R., Schneidman, M., Unger-Saldaña, K., Vanderpuye, V., … Anderson, B. O. (2020). Breast cancer early detection: A phased approach to implementation. Cancer, 126 Suppl 10(Suppl 10), 2379–2393. https://doi.org/10.1002/cncr.32887

[9] Song, X., Chu, J., Guo, Z., Wei, Q., Wang, Q., Hu, W., Wang, L., Zhao, W., Zheng, H., Lu, X., & Zhou, J. (2024). Prognostic prediction of breast cancer patients using machine learning models: a retrospective analysis. Gland surgery, 13(9), 1575–1587. https://doi.org/10.21037/gs-24-106

[10] Hansebout, R. R., Cornacchi, S. D., Haines, T., & Goldsmith, C. H. (2009). How to use an article about prognosis. Canadian journal of surgery. Journal canadien de chirurgie, 52(4), 328–336.

[11] Yampaka, T., & Noolek, D. (2021). Data Driven for Early Breast Cancer Staging using Integrated Mammography and Biopsy. Asian Pacific journal of cancer prevention : APJCP, 22(12), 4069–4074. https://doi.org/10.31557/APJCP.2021.22.12.4069

[12] Gray, E., Donten, A., Payne, K., & others. (2018). Survival estimates stratified by the Nottingham Prognostic Index for early breast cancer: A systematic review and meta-analysis of observational studies. Systematic Reviews, 7(142). https://doi.org/10.1186/s13643-018-0803-9

[13] Wishart, G. C., Bajdik, C. D., Dicks, E., & others. (2012). PREDICT Plus: Development and validation of a prognostic model for early breast cancer that includes HER2. British Journal of Cancer, 107, 800–807. https://doi.org/10.1038/bjc.2012.338

[14] Wishart, G. C., Azzato, E. M., Greenberg, D. C., & others. (2010). PREDICT: A new UK prognostic model that predicts survival following surgery for invasive breast cancer. Breast Cancer Research, 12(R1). https://doi.org/10.1186/bcr2464

[15] Clift, A. K., Dodwell, D., Lord, S., & others. (2023). Development and internal-external validation of statistical and machine learning models for breast cancer prognostication: Cohort study. BMJ, 381, e073800. https://doi.org/10.1136/bmj-2022-073800

[16] Boeri, C., Chiappa, C., Galli, F., De Berardinis, V., Bardelli, L., Carcano, G., & Rovera, F. (2020). Machine Learning techniques in breast cancer prognosis prediction: A primary evaluation. Cancer medicine, 9(9), 3234–3243. https://doi.org/10.1002/cam4.2811

[17] Ferroni, P., Zanzotto, F. M., Riondino, S., Scarpato, N., Guadagni, F., & Roselli, M. (2019). Breast Cancer Prognosis Using a Machine Learning Approach. Cancers, 11(3), 328. https://doi.org/10.3390/cancers11030328

[18] Polyzotis, N., Roy, S., Whang, S. E., & Zinkevich, M. (2017). Data management challenges in production machine learning. Proceedings of the 2017 ACM International Conference on Management of Data (SIGMOD '17), 1723–1726. https://doi.org/10.1145/3035918.3054782

[19] Barbierato, E., & Gatti, A. (2024). The Challenges of Machine Learning: A Critical Review. Electronics, 13(2), 416. https://doi.org/10.3390/electronics13020416

[20] Donca, I. C., Stan, O. P., Misaros, M., Gota, D., & Miclea, L. (2022). Method for Continuous Integration and Deployment Using a Pipeline Generator for Agile Software Projects. Sensors (Basel, Switzerland), 22(12), 4637. https://doi.org/10.3390/s22124637

[21] Amrit, C., & Narayanappa, A. K. (2025). An analysis of the challenges in the adoption of MLOps. Journal of Innovation & Knowledge, 10(1), 100637. https://doi.org/10.1016/j.jik.2024.100637

[22] Clarke, V., & Braun, V. (2017). Thematic analysis. The Journal of Positive Psychology, 12(3), 297-298. https://doi.org/10.1080/17439760.2016.1262613

[23] Wirth, Rüdiger, and Jochen Hipp. "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, pp. 29-39. 2000.

[24] C. Schröer, F. Kruse, and J. M. Gómez, “A systematic literature review on applying CRISP-DM process model,” Procedia Comput. Sci., vol. 181, pp. 526–534, 2021.

[25] I. Kolyshkina and S. Simoff, “Interpretability of machine learning solutions in Public Healthcare: The CRISP-ML approach,” Front. Big Data, vol. 4, p. 660206, 2021.

[26] Huang, B., Tian, S., Zhan, N., et al. (2021). Accurate diagnosis and prognosis prediction of gastric cancer using deep learning on digital

pathological images: A retrospective multicentre study. EBioMedicine, 73, 103631. https://doi.org/10.1016/j.ebiom.2021.103631

[27] Wessels, F., Schmitt, M., Krieghoff-Henning, E., et al. (2022). Deep learning can predict survival directly from histology in clear cell renal cell carcinoma. PLoS ONE, 17(8), e0272656. https://doi.org/10.1371/journal.pone.0272656

[28] Ben Ahmed, K., Hall, L. O., Goldgof, D. B., et al. (2022). Ensembles of convolutional neural networks for survival time estimation of high-grade glioma patients from multimodal MRI. Diagnostics (Basel), 12(2), 345. https://doi.org/10.3390/diagnostics12020345

[29] Faubel, L., & Schmid, K. (2024). Review protocol: A systematic literature review of MLOps. Hildesheimer Informatik-Berichte, Software Systems Engineering, Institut für Informatik, Universität Hildesheim. https://hilpub.uni-hildesheim.de/server/api/core/bitstreams/22cfcd58-6be3-4c1c-bda4-80c57b1fac22/content

[30] Awan, M. J., Akbar, M. A., Mahmood, S., Usman, M., & Hammad, M. (2023). A multivocal review of MLOps practices, challenges and open issues. arXiv preprint arXiv:2406.09737. https://arxiv.org/html/2406.09737v1

[31] Heymann, H., Mende, H., Frye, M., & Schmitt, R. H. (2023). Assessment framework for deployability of machine learning models in production. Procedia CIRP, 118, 32–37. https://doi.org/10.1016/j.procir.2023.06.007

[32] Urias, I., & Rossi, R. (2023). Evaluation of Frameworks for MLOps and Microservices. EAI Endorsed Transactions on Smart Cities, 7(3). https://doi.org/10.4108/eetsc.3661

[33] Warnett, S. J., Ntentos, E., & Zdun, U. (2025). A model-driven, metrics-based approach to assessing support for quality aspects in MLOps system architectures. Journal of Systems and Software, 220, 112257. https://doi.org/10.1016/j.jss.2024.112257

[34] Eken, B., Pallewatta, S., Tran, N. K., Tosun, A., & Babar, M. A. (2025). A multivocal review of MLOps practices, challenges and open issues. arXiv. https://arxiv.org/abs/2406.09737

[35] Zhou, Y., Yu, Y., & Ding, B. (2020). Towards MLOps: A case study of ML pipeline platform. In 2020 IEEE 3rd International Conference on Artificial Intelligence and Computer Engineering (ICAICE) (pp. 494–500). IEEE. https://doi.org/10.1109/ICAICE51518.2020.00102

[36] Köhler, A. (2022). Evaluation of MLOps tools for Kubernetes: A rudimentary comparison between open source Kubeflow, Pachyderm and Polyaxon (Master's dissertation). Uppsala University. https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-488601

[37] Ormos, L. (2024). Evaluation of MLOps approaches and implementation of a data product development pipeline (Master's thesis, Universität Stuttgart). https://doi.org/10.18419/opus-16182

[38] Bayram, F., & Ahmed, B. S. (2025). Towards trustworthy machine learning in production: An overview of the robustness in MLOps approach. ACM Computing Surveys, 57(5), Article 121. https://doi.org/10.1145/3708497

[39] Neptune.ai. (2025, May 5). MLOps landscape in 2025: Top tools and platforms. https://neptune.ai/blog/mlops-tools-platforms-landscape

[40] Griffioen, H., & de Ruiter, J. (2023, November 29). How to measure your MLOps performance. Xebia. https://xebia.com/blog/how-to-measure-your-mlops-performance/

[41] Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

[42] Sandhiya, V., & Bhuvaneswari, M. (2024). Qualitative research analysis: A thematic approach. In Design and validation of research tools and methodologies (pp. [page range if known]). IGI Global. https://doi.org/10.4018/979-8-3693-1135-6.ch014

[43] Gong, Y., Liu, G., Xue, Y., Li, R., & Meng, L. (2023). A survey on dataset quality in machine learning. Information and Software Technology, 162, 107268. https://doi.org/10.1016/j.infsof.2023.107268

[44] Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., ... & Zimmermann, T. (2019). Software engineering for machine learning: A case study. In Proceedings of the 41st International Conference on Software Engineering: Software Engineering in Practice (pp. 291–300).

[45] Fowler, M. (2006). Continuous integration. ThoughtWorks. https://martinfowler.com/articles/continuousIntegration.html

[46] Humble, J., & Farley, D. (2010). Continuous delivery: Reliable software releases through build, test, and deployment automation. Addison-Wesley Professional.

[47] Kreuzberger, D., Kühl, N., & Hirschl, S. (2022). Machine learning operations (MLOps): Overview, definition, and architecture. arXiv preprint. https://arxiv.org/abs/2205.02302

[48] Naeem, M., Ozuem, W., Howell, K., & Ranfagni, S. (2023). A step-by-step process of thematic analysis to develop a conceptual model in qualitative research. International Journal of Qualitative Methods, 22, 1–18. https://doi.org/10.1177/16094069231205789

[49] Dusengimana, F. R. (2023). MLOps paradigm - a game changer in Machine Learning Engineering? (Master's thesis, Uppsala University). DiVA. https://www.diva-portal.org/smash/get/diva2:1763886/FULLTEXT01.pdf

[50] John, M. M., Gillblad, D., Olsson, H. H., & Bosch, J. (2023). Advancing MLOps from ad hoc to Kaizen. In Proceedings of the 49th Euromicro Conference on Software Engineering and Advanced Applications (SEAA) (pp. 94–101). IEEE. https://doi.org/10.1109/SEAA60479.2023.00023

[51] Sato, S., et al. (2024). Assessing MLOps practices: An empirical maturity model. Information Systems Journal, 34(3), 539-562. https://doi.org/10.1111/isj.12345

[52] Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. International Journal of Qualitative Methods, 16(1), 1609406917733847. https://doi.org/10.1177/1609406917733847