

Cross-Cloud Mlops Benchmarking For AI Workloads In BFSI Enterprises

Satishkumar Rajendran

University of Central Missouri, Warrensburg, USA.

Abstract

This paper presents a systematic literature review of the existing literature on cross-cloud MLOps benchmarking of artificial intelligence workloads in banking, financial services, and insurance (BFSI) companies. The paper provides an overview of the current literature on distributed ML pipeline infrastructures, cloud-based MLOps systems, and benchmarking methods that are used to evaluate enterprise AI pipelines. The literature search covered Scopus, IEEE Xplore, ACM Digital Library, and Google Scholar, however, the primary focus was on the publications from 2015 onward. The review employs a structured methodological approach to establish the dimensions of performance, architectural designs, and operational issues that are commonly documented in cross-cloud machine learning deployment. The review suggests that the distributed computing architectures and containerized orchestration systems play a critical role in improving scalability and model-training efficiency and cross-cloud implementations are further complicated by data synchronization, network overhead, and governance limitations. Some of the most significant evaluation metrics found in the literature reviewed include training latency, inference throughput, resource consumption and infrastructure cost. Based on the synthesis, a conceptual benchmarking framework of BFSI-oriented cross-cloud MLOps systems is proposed, incorporating performance, architectural, and governance factors. The study also identifies future research directions, including standardized benchmarking guidelines, intelligent workload orchestration, and AI infrastructure design that is compliance-conscious.

Keywords: Cross-cloud MLOps; Distributed machine learning; Financial AI infrastructure; Cloud benchmarking; Enterprise AI systems.

1. Introduction

Artificial intelligence (AI) has rapidly transformed the operational landscape of banking, financial services, and insurance (BFSI) institutions and altered the mode of operation of banks, financial services providers, and insurance companies (BFSI). The latest machine learning algorithms are actively used in credit scoring and fraud detection, algorithmic trading, customer analytics, risk prediction and regulatory compliance monitoring. Such As AI workloads have increased in scale and complexity, their infrastructural and computational demands have also grown in relation to the infrastructural and computational demands. Banks are forced to handle very large volumes of transactional information and to facilitate real-time insights and huge volumes of regulatory compliance to privacy, transparency, and operational resilience rules [1]. These factors have led to the adoption of cloud computing systems that can scale machine learning pipelines.

Cloud computing has evolved from a basic on-demand resource model into a complex ecosystem, but a complex ecosystem, which provides managed machine-learning services, distributed training systems, model monitoring services and in-built data engineering pipelines. The large cloud providers have developed end-to-end machine-learning platforms which support the entire AI development lifecycle which comprises data preparation, model training, deployment, monitoring and governance. These capabilities are typically part of the MLOps paradigm, therefore, enabling models to be created, deployed, monitored, and continually updated in a predictable way [2]. MLOps frameworks have emerged as critical components of enterprise AI strategies due to the issues of model reproducibility, model lifecycle, model monitoring, and model scale.

MLOps has become important in the BFSI because financial decision systems are mission critical. The machine learning models that are utilized in financial institutions should not only be highly predictive, but they should also be stable, traceable and meet regulatory requirements. Such

algorithmic decision-making systems are also becoming explainable and auditable by financial regulators and that has led to the emergence of adequate lifecycle management infrastructures capable of explaining the provenance of data and model versions and model performance indicators in deployment landscapes [3]. Cloud-based MLOps pipelines have therefore played an important role in the financial institution scaling and compliant AI deployment.

Despite the benefits of cloud-native machine learning infrastructure, the problem of determining the most appropriate deployment architectures for enterprise AI workloads still exists. The BFSI organisations increasingly use multi-cloud or cross-cloud environments to avoid vendor lock-in, enhance operational resiliency and address the regulatory requirements of the area. The resulting opportunity to distribute machine learning pipelines across cloud providers can aid in resilience and data availability and enable organizations to leverage platform-specific capabilities [4]. Nonetheless, these architectures pose other issues regarding interoperability, performance benchmarking, cost optimization and operation governance.

Cross-cloud MLOps benchmarking has thus become an important field of study where it seeks to quantify the performance, reliability and efficiency of machine learning tasks in heterogeneous clouds. Typical benchmarking points include the training latency, inference throughput, cost effectiveness, scalability, model deployment latency, and resource usage across cloud environments. They are significant performance indicators in organisations that are interested in optimisation of AI pipelines in production systems where business performance is directly related to the computational costs as well as service latencies [5].

In the context of the BFSI industry, benchmarking research assumes greater importance since financial tasks have peculiarities in terms of which many other machine learning applications cannot be compared. financial data are generally large-scale, highly structured, and frequently updated through transactional streams through streams of transactions. Such models as fraud detection, in turn, should be capable of processing high-velocity data at very low latency and with very high accuracy to prevent both financial losses and regulatory fines. Likewise, the risk-assessment models require consistent performance over extended time horizons and model governance needs to be of high quality. Such applications require both batch and real-time machine learning workloads to be highly optimized by MLOps infrastructures [6].

Recent developments in distributed ML pipeline models have also enhanced the applicability of benchmarking in cloud settings. Contemporary AI pipelines often use massive deep learning implementations, which are trained on a cluster of distributed GPUs. These models are resource intensive in terms of the quantity of computation they need and complex orchestration systems that are able to coordinate the distributed training processes. Cloud platforms have provided dedicated hardware accelerators and container-orchestration systems to meet scaling requirements to fulfil these needs. However, the performance of these systems may vary greatly across the providers, and the cross-cloud testing is required to identify the optimal strategies of deployment [7].

In addition to the computational capability, other aspects of operations such as cost-effectiveness and compliance with regulations are central towards cloud platform selection within the BFSI organizations. Pricing models in cloud services can differ greatly based on how much compute power, storage and network bandwidth is used and specialized hardware accelerators. Consequently, financial institutions that apply AI pipelines on a large scale need to not only focus on the computational performance but also the financial consequences of various implementation plans in the long term. Besides this, data-governance regulations such as the General Data Protection Regulation (GDPR) and financial supervision frameworks have extremely rigid policies on data residency, security, and auditability that could have an impact on the considerations of cloud infrastructure [8].

Although the available literature on the subject of cloud computing performance, distributed ML pipeline structures, as well as enterprise MLOps frameworks, has ample coverage, comparatively scarce attention has been given to systematic cross-cloud MLOps environment benchmarking in the context of BFSI environments. The majority of available studies focus on the overall and/or individual machine learning workloads without consideration of the nature of operation of financial AI pipelines. In addition, variations in benchmarking procedures, evaluation metrics and experimental designs have also been considered to be a challenge in comparison of the results obtained in the studies. Absence of cohesive benchmarking models has thus hampered the ability

of financial institutions to come up with well-informed judgments regarding the cross-cloud AI deployment plans [9].

The cross-cloud MLOps benchmarking also presupposes a systemic analysis to integrate the existing body of knowledge, demonstrate the tendencies of the approaches and also identify underexplored research opportunities in the specified field. It is possible to utilize this review to comprehend the performance properties of distributed AI pipelines, the problems of the operational multi-cloud machine learning infrastructure, and the technological innovations that should be offered to maintain the next-generation financial AI infrastructure. The literature review enables identifying the key gaps in the methodology of benchmarking, present the tendencies in the approach to the implementation of AI in the business, and the possibilities of how in the future the activities of the cloud-based machine learning will be more reliable and efficient.

In this respect, the present review discusses the active literature on the cross-cloud MLOps benchmarking of AI workloads in BFSI firms. This review has three objectives. First, the current literature on cloud-based machine learning processes and distributed AI infrastructure is critically evaluated to trace the current benchmarking practices and performance metrics. Second, the methodological frameworks of the existing benchmarking studies are also analyzed in order to determine their suitability to financial AI workloads. Third, emerging technological challenges and research directions are identified to help in building scalable, reliable, and cost-efficient cross-cloud MLOps architecture to BFSI enterprises. The review is supposed to contribute to the overall knowledge of technological and methodological background of effective benchmarking of enterprise AI systems in the financial environment. It is important to note that this work does not empirically benchmark cloud platforms. Instead, it adopts a methodological literature review process to synthesize the existing benchmarking methods and performance dimensions that have been reported in the literature. It aims to organize the fragmented research environment into a rational analytical system and propose a conceptual benchmarking framework tailored to satisfy BFSI-oriented cross-cloud MLOps systems.

2. Literature Review

This section reviews the literature across three key dimensions: (i) models of MLOps lifecycle and enterprise AI deployment, (ii) benchmarking models of cloud-based machine learning systems, and (iii) cross-cloud architecture with special consideration to BFSI settings. The increasing popularity of cloud computing to deploy enterprise-scale artificial intelligence systems has given rise to a major research on how machine learning algorithms work. AI-based analytics systems in the BFSI industry operate under strict compliance, reliability, and performance requirements. This suggests that successful implementation of machine learning pipelines requires formal operational frameworks to run the lifecycle of a model, scale to large, and monitor continuously. The requirements have led to various studies on the problems of Machine Learning Operations (MLOps), distributed cloud computing systems and benchmarking techniques to determine the performance of cloud infrastructure.

MLOps emerged at the intersection of machine learning engineering and DevOps. Earlier machine-learning research focused largely on algorithm development and predictive accuracy and predictive accuracy. There were other complexities that were added to enterprise deployment environments such as model reproducibility, pipeline orchestration, and lifecycle governance. This led to the development of MLOps systems that have facilitated the use of machine learning models to support continuous integration and continuous deployment (CI/CD). These systems involve automated model training, model versioning, monitoring systems and infrastructure management tools to support AI implementation in production services at scale [10].

The financial institutions have embraced MLOps due to the need to have transparency in the operations and governance of the model. Credit scoring, fraud detection, and risk management models must remain reliable under dynamic data conditions while maintaining consistent performance and satisfying regulatory audit requirements. MLOps frameworks thus include data lineage tracking mechanisms, model explainability assessment mechanisms and automated retraining pipelines to ensure model accuracy in the long term [11]. The implementation of these functions into cloud systems has helped organizations to use large-scale AI systems without compromising on their operations.

In recent studies, there has been a focus on the necessity of having standardized MLOps architectures that can be used to support heterogeneous computing environments. The standard components in the workflow of a multi-stage pipeline are data ingestion modules, feature engineering, distributed model training systems, containerized deployment, and monitoring services. These distributed pipelines are commonly managed through container orchestration solutions (e.g., Kubernetes) to allow scaling resource allocation to cloud infrastructures [12].

Containerized MLOps pipelines have been especially useful in BFSI infrastructure to support flexible model deployment in a variety of computing environments. Very often financial institutions host hybrid infrastructures composed of an on-premises computing systems and a public cloud platform. Containerization allows machine learning models to be implemented in these environments with uniformity and minimize operational variations and streamline pipeline management [13].

The cloud computing infrastructures are gaining popularity as enterprise AI infrastructure owing to its scalability and flexibility. Major cloud vendors offer dedicated machine-learning services to support model development and scale. The training systems that are distributed, pipeline coordination systems and monitoring systems that are mostly used therein assist in the whole machine learning lifecycle.

Comparative analysis of cloud-native machine learning platforms has shown that there can be a large performance difference depending on the workload nature and infrastructure configuration. Training latency, inference latency, throughput, and scalability depend strongly on distributed-training architecture and network capability. Benchmarking study has thus been conducted to determine the performance variations between the vendors of clouds [14].

In the cloud environment, Apache Spark and TensorFlow are widely used frameworks for large-scale and distributed machine-learning workloads to offer parallel machine learning workloads. The frameworks can be used to distribute computing to groups of compute nodes and may be useful in reducing training time on large datasets significantly. Such architectures have also been embraced by many financial institutions to process a lot of transactional data, and to train more complex predictive models with the ability to either identify fraudulent behaviour, or predict financial risk [15].

Nevertheless, the advantages of performance that come with distributed cloud computing infrastructures are thoroughly reliant on both the efficiency of the networks and the strategy of resource allocation. Poor task scheduling or network overload can dramatically lower distributed training performance. Consequently, benchmarking studies have highlighted the necessity to consider not only the computational performance, but also system-level aspects including communication latency and scalability of a cluster [16].

The research in cloud benchmarking has been developed considerably during the last decade. Initial benchmarking research mainly aimed at determining the computational performance measures of processor speed, memory throughput and storage latency of the cloud providers. Nevertheless, the current AI workloads demand more sophisticated benchmarking models that can assess distributed ML pipelines.

The recent benchmarking techniques have been found to generally consider multiple performance dimensions at the same time, resource utilization efficiency, and operational cost. These metrics are a holistic measure of the performance of the cloud platforms to the machine learning workloads. Also, industry-specific benchmarks have been created to recreate the scenario of real-life enterprise AI pipelines so that researchers could test the performance of the systems in real-life settings [17].

Benchmarking systems like MLPerf have been proposed to standardize and compare the performance of machine learning systems with both hardware and cloud systems. They provide standardized datasets, training targets, and evaluation metrics, training goals and evaluation metrics which are intended to assist the repeatability of the benchmarking experiments. The frameworks have helped to increase the comparability of the studies and systematically study the machine learning infrastructure performance [18].

The financial AI workloads have benchmarking criteria not founded strictly on computational performance. Transactional data streams are often time sensitive and high velocity and any system able to support real time inference pipelines is needed. As a result, benchmarking research

in the BFSI environment often assesses the latency-sensitive workloads and streaming data processing services [19].

Multi-cloud and cross-cloud computing have become increasingly important strategies within the enterprise settings. Organizations deploy workloads across different cloud providers to decrease the reliance on a single vendor, enhance the service reliability, and lower costs of the infrastructure. Cross-cloud architecture in BFSI-based structures is also capable of aiding regulatory compliance by allowing data storage and processing to be done in particular geographic locations.

Studies on multi-cloud design suggest that distributed cloud architectures can enhance system resilience and scalability. The allocation of workloads among providers can reduce the risk of service failures due to provider interruptions or infrastructure crashes within the organization. Nonetheless, the cross-cloud setups present operational issues related to interoperability, data transfer costs as well as orchestrating pipelines across heterogeneous infrastructure settings [20]. Machine learning pipelines scheduled on different cloud providers need advanced orchestration systems that are able to coordinate distributed computing resources. These pipelines have been managed by workflow orchestration systems like Kubeflow and Apache Airflow. Such systems allow companies to specify complicated workflows with distributed training tasks, automated evaluation tasks, and deployment operations across more than one infrastructure environment [21].

In spite of such developments, cross-cloud machine learning infrastructures remain difficult to manage in practice because of differences in cloud service interfaces and security frameworks as well as resource management models. Benchmarking studies are thus vital in the determination of best architectural solutions in deploying machine learning pipelines on heterogeneous cloud environments. Table 1 summarizes representative studies included in the review and highlights their research focus, infrastructure context, and benchmarking dimensions.

Table 1. Summary of Selected Studies Included in the Review

Reference	Research Focus	Infrastructure	Benchmark Metrics	Key Findings
[10]	MLOps lifecycle management	Enterprise cloud	Deployment reliability	Demonstrated need for lifecycle governance
[11]	Financial AI model governance	BFSI datasets	Model transparency	Highlighted importance of explainability
[12]	Container orchestration	Kubernetes clusters	Resource scaling	Improved pipeline scalability
[13]	Hybrid cloud ML deployment	Hybrid enterprise systems	Deployment portability	Containerization improved portability
[14]	Cloud performance comparison	Public cloud providers	Compute performance	Identified significant provider variation
[15]	Distributed ML training	Apache Spark clusters	Training time	Distributed frameworks reduce training latency
[16]	Distributed system optimization	Cluster computing	Network communication	Network overhead impacts training efficiency
[17]	ML infrastructure benchmarking	Multi-cloud environments	Latency, throughput	Multi-dimensional benchmarking required
[18]	MLPerf benchmarking framework	AI hardware systems	Standardized metrics	Improved benchmarking reproducibility

[19]	Financial AI streaming systems	BFSI infrastructure	Real-time latency	Streaming analytics critical for fraud detection
------	--------------------------------	---------------------	-------------------	--

Table 2 consolidates commonly reported benchmarking metrics identified across the reviewed literature and their relevance to BFSI applications.

Table 2. Benchmarking Metrics Identified in Reviewed Literature

Metric	Description	Importance for BFSI
Training Latency	Time required for model training	Impacts model retraining cycles
Inference Throughput	Predictions per second	Critical for real-time fraud detection
Resource Utilization	CPU/GPU efficiency	Determines infrastructure cost
Deployment Latency	Time to deploy updated models	Important for model governance
Operational Cost	Cloud infrastructure expenditure	Influences enterprise adoption

Several emerging themes can be identified in the reviewed literature on cross-cloud benchmarking in BFSI settings. First, increasing attention has been directed toward automated pipeline-orchestration systems capable of managing distributed ML processes across heterogeneous environments. These structures are designed to automate the retraining, evaluation and deployment to streamline the operations. Second, cost optimization is a research subject of interest as AI-related workloads are huge and prohibitively expensive to use computationally. As cloud pricing schemes become increasingly complex to calculate, network storage and usage, benchmarking research is increasingly considering infrastructure design that optimizes performance with respect to both the computational cost and the operational cost [22]. Third, adherence to the rules has been a major design consideration of enterprise AI infrastructure. The financial institutions require auditable and transparent machine learning systems, thus leading to the use of explainability tools and governance systems in MLOps pipelines [23].

Despite these trends, substantial gaps remain in the systematic analysis of cross-cloud MLOps infrastructures with BFSI workloads. Much of the work published is not specific to regulatory, latency, and reliability requirements of financial systems. Furthermore, benchmarking methodology differs and this makes it hard to directly compare the studies. This gap limits the relevance of existing literature to inform the evidence-based decisions made by institutions on cross-cloud AI implementation strategies.

In this respect, additional studies need to be done to create standard benchmarking models, which will be able to measure distributed MLOps pipelines in heterogeneous clouds. Such frameworks will be in a position to assist the financial institutions that have AI systems in making better decisions regarding their infrastructure. Despite a large body of literature available on cloud computing, distributed ML pipelines, and MLOps systems, the literature remains fragmented in terms of benchmarking techniques, evaluation metrics, and domain aspects. Precisely, the requirements related to the BFSI, e.g. the regulatory compliance, auditability, the decision systems suggesting the latency sensitivity of the decision system, etc. are not included in the current benchmark models systematically. This is the reason why it should be systematically synthesized and conceptually coherent so that it would help to cross-cloud MLOps benchmarking in financial AI settings.

3. Methodology

This paper used a systematic literature review approach to examine the literature on cross-cloud MLOps benchmarking of AI workloads in banking, financial services, and insurance (BFSI) enterprises. The purpose of the review was to provide a summary of the existing evidence on the dimensions of benchmarking, architectural patterns, orchestration mechanisms, and governance requirements of cloud-based machine learning processes in a financial environment.

The literature search was conducted with the assistance of the major academic databases, namely Scopus, IEEE Xplore, ACM Digital Library, and Google Scholar. The search focused primarily on studies published from 2015 onward, while selected earlier foundational studies were retained where necessary. Search queries included a mix of the following keywords: MLOps, machine learning operations, cloud benchmarking, cross-cloud AI, multi-cloud infrastructure, distributed machine learning, BFSI, financial AI, fraud detection and credit risk.

Peer-reviewed journal articles and conference papers in the review were at least one of the following areas: cloud-based machine learning systems, MLOps frameworks, benchmarking methods, distributed AI infrastructure, or BFSI relevant AI deployment environments. The research papers were considered in cases where they discussed the performance analysis, system design, orchestration, lifecycle management, governance, or compliance challenges in enterprise AI activities. Studies focused solely on algorithm development, without discussion of infrastructure, deployment, benchmarking, or operational management, were excluded. Non-peer-reviewed materials, editorials, white papers, and articles lacking methodological or technical relevance were excluded.

Study selection was conducted in four steps, which included identification, screening, eligibility assessment, and final inclusion. The initial course of action was to locate records by searching the databases. After filtering out duplicates, abstracts and titles were filtered. The remaining articles were subsequently reviewed by conducting a full-text review in line with the pre-determined inclusion and exclusion criteria. Figure 1 summarizes the process of studies selection in PRISMA style, giving the identification, screening, eligibility assessment and eventually the inclusion process. The final set of studies included after the inclusion and exclusion criteria were applied formed the basis of the thematic synthesis.

Data extraction was conducted through the structured review matrix. In each of the included studies, where feasible, the following aspects were recorded: the publication information, infrastructure context, type of workload, orchestration or deployment architecture, benchmarking metrics, cloud environment, and governance or compliance applicability of BFSI. Thematic synthesis was also used to analyse the material extracted to generate recurrence dimensions of benchmarking, architectural factors and domain specific operational needs.

Synthesis was organized into three analysis dimensions, performance dimensions, architectural and orchestration dimensions and BFSI-specific governance dimensions. The conceptual benchmarking framework proposed in this paper is based on this framework. The review process was developed as shown in Figure 1, to traverse the general database search to the limited set of final study retrieval. This fact is notable as the theoretical framework proposed in the next sections of the paper will be built upon a selective and thematically ordered literature, rather than a non-systematic selection of narratives.

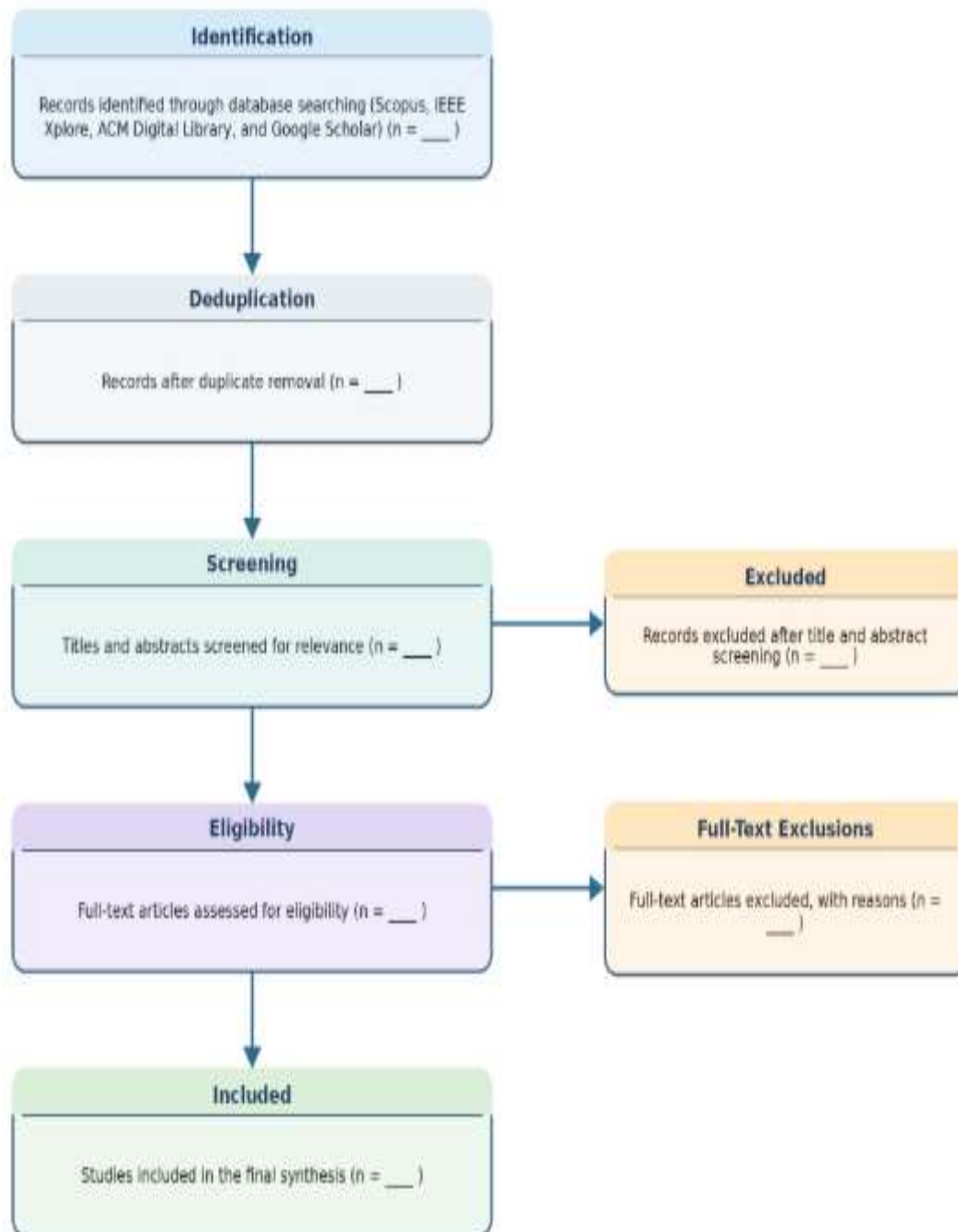


Figure 1. PRISMA-style flow diagram of study identification, screening, eligibility assessment, and final inclusion

4. Conceptual Benchmarking Framework for Cross-Cloud MLOps in BFSI

Based on the literature review synthesis, it is postulated that a conceptual benchmarking framework of cross-cloud MLOps systems in the BFSI environment should be proposed. The framework will attempt to integrate the different dimensions of benchmarking, architectural concerns, and governance requirements that have been developed in the past literature into a single analytical framework. The primary contribution of the study, this theoretical framework, should be applied in the design of benchmarking of BFSI-oriented cross-cloud MLOps environments.

The proposed model consists of three large dimensions. The former dimension also includes performance-related metrics that are generally reported in benchmarking research including training latency, inference latency, throughput, scalability, resource consumption, and cost of infrastructure. These measures provide a quantitative basis of evaluating the efficiency of distributed ML pipelines in clouds.

The second dimension relies on system-level and architectural issues, including orchestration mechanisms, distributed storage system strategies based on containerization, and overhead of data transfer. All these factors influence the behaviour of cross-cloud MLOps pipelines and determine their scalability and reliability. These dimensions are divided into a BFSI-oriented benchmarking framework in Table 3 and graphically illustrated in Figure 2 as the relationship between performance, architecture and orchestration, and governance and compliance to the representative financial AI workloads. Together, the table and the figure suggest that benchmarking criteria should not be considered as such independent technical measures but rather in the context of workload.

Table 3. BFSI-Oriented Cross-Cloud MLOps Benchmarking Framework

Dimension	Sub-dimension	Example metrics	BFSI relevance
Performance	Training performance, inference performance, scalability, resource efficiency, cost efficiency	Training latency, inference latency, throughput, scalability, CPU/GPU utilization, memory utilization, infrastructure cost	Supports rapid model retraining, low-latency fraud detection, efficient large-scale deployment, and cost-aware operation of financial AI systems
Architecture and orchestration	Containerization, workflow orchestration, interoperability, distributed storage, data transfer overhead, system resilience	Deployment latency, failover readiness, storage access time, data transfer time, portability across cloud platforms, orchestration efficiency	Enables reliable coordination of cross-cloud pipelines, improves operational resilience, and supports multi-region or multi-provider deployment requirements
Governance and compliance	Auditability, explainability, data lineage, model versioning, data residency, security and access control	Audit trail availability, explainability support, model traceability, version control coverage, policy compliance, access logging	Essential for regulatory compliance, model risk management, supervisory review, and transparent financial decision-making
Representative BFSI workloads	Fraud detection, credit risk modelling, AML monitoring, real-time financial analytics	Latency sensitivity, retraining frequency, governance burden, workload criticality, reliability requirements	Ensures that benchmarking reflects real financial AI use cases instead of generic machine learning workloads

The third dimension considers BFSI-specific conditions of operations and governance present in the literature. They include model auditability, explainability, data governance, regulatory compliance, and system resiliency. Financial AI systems rely on the lifecycle management, and transparency, and, therefore, these aspects are essential to benchmark an enterprise MLOps infrastructure.

Representative BFSI workloads, such as fraud detection, credit risk modelling, and real-time financial analytics are also in scope, and are extremely strict in terms of the latency, reliability, and compliance demands on AI systems.

The proposed framework provides a systematic framework on how the benchmarking studies should be structured and may be applied in future research to provide standardized evaluation mechanisms in cross-cloud machine learning systems when applied in a financial environment. The framework is premised on three levels of evaluation, namely, the review of the literature reviewed: performance dimensions, architectural and orchestration dimensions, and dimensions specific to the BFSI governance. This framework will make it possible to benchmark cross-cloud MLOps systems not only based on technical efficiency but also on compliance-based and operationally sensitive requirements of financial AI systems.

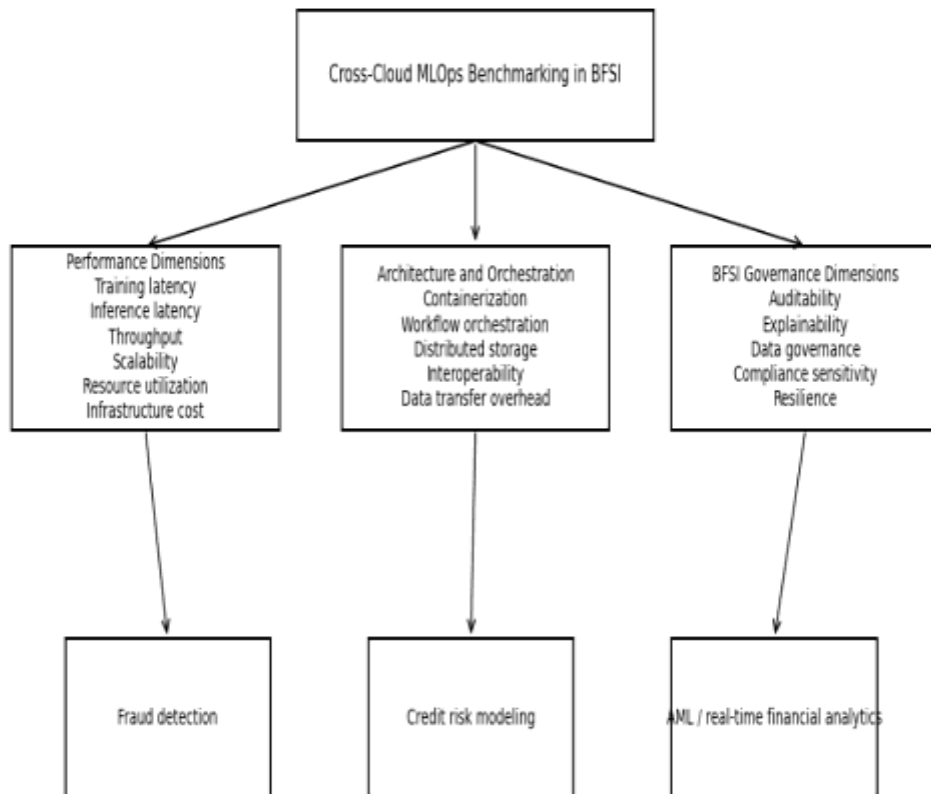


Figure 2. Conceptual benchmarking framework for cross-cloud MLOps systems in BFSI enterprises

5. Discussion

The literature review shows that cross-cloud MLOps systems in the context of BFSI need to be viewed as a more holistic entity than mere technical performance. Some of the typical metrics in traditional benchmarking are training latency, inference latency, scalability, resource utilization, and cost, but auditability, explainability, compliance, and operational resilience are also needs in BFSI deployments. This is a big difference because the infrastructure architecture that performs well in a typical enterprise setting may not be suitable to the financial requirements when it fails to satisfy the governance and reliability requirements. The literature shows benchmarking in BFSI as a performance issue and risk-management issue.

One of the most frequently-used dimensions of cloud-based machine learning benchmarking studies is distributed training performance. The machine learning models applied in BFSI problems like fraud detection, credit score and predicting portfolio risks are typically trained on large volumes of data of millions of transactional records. The models are computationally expensive to be trained on single-node computing systems. This has in turn led to extensive usage of distributed computing systems to facilitate parallel training of sets of compute nodes.

Reductions in model training time are often reported with distributed training architectures through empirical studies. Distributed training systems like Apache Spark and TensorFlow

provide frameworks, which the training activities can be distributed among a very large number of compute nodes and, consequently, run model optimization algorithms simultaneously. However, the efficiency of such disseminated training performance greatly relies on the behaviour of the communication among nodes. Poor data synchronization mechanisms or high network latency can significantly reduce the expected performance benefits of distributed architectures [34].

The cross-cloud benchmarking studies also show that the various infrastructure providers have a large disparity in the training latency due to the disparity in the hardware acceleration capabilities, networking infrastructure, and the resource scheduling algorithms. Cloud services that provide specialized hardware accelerators such as GPUs or TPUs tend to have significant performance benefits in the training of deep learning models. These variations indicate the significance of cross-cloud benchmarking of machine learning workloads in order to develop enterprise AI deployment models.

Besides the configuration of the hardware, the distributed training performance is also affected by the data partitioning mechanisms and the workload balancing mechanisms. Lack of uniform distribution of data among cluster nodes may result in poor resource utilization and increased training time. Examples of experiments involving benchmarking often compare load balancing algorithms and cluster scheduling strategies to determine the best configuration to run large-scale machine learning workloads [35].

Although training performance is a valuable evaluation measure, inference latency can be more crucial to functional AI applications in BFSI. Systems such as fraud detection platforms require high volumes of transactional data to be processed within a real-time environment such that suspicious fraudulent activities are detected before fraudulent transactions are completed. On the same note, the algorithmic trading systems depend on very low latency predictions to achieve trading strategies.

The studied papers suggest that the inference latency is highly dependent on the architecture of deployment pipelines. Containerized microservice-based architectures have gained popularity in the deployment of machine learning inference services since they allow the deployment of predictive models in a scalable and modular fashion. These architectures enable the distribution of inference workloads to more than one compute node to allow prediction services to scale horizontally [36].

Nonetheless, deployments across clouds create new latency issues that are a result of network communication overheads across cloud environments. Data transfer latency between cloud providers may significantly increase in the total response time of inference, especially when inference pipelines use distributed data sources. Therefore, a number of studies suggest the implementation of latency-sensitive inference pipelines to single cloud environments and training and batch processing tasks to different clouds.

Another possible solution that has been suggested to solve the inference latency problem in financial AI systems is the use of edge computing architectures. By deploying lightweight inference models closer to data sources, edge computing systems can decrease the network latency and enhance the real-time decision-making process. Edge computing architectures remain nascent in the financial infrastructures, but benchmarking studies propose that hybrid cloud-edge architectures could be highly beneficial in improving the performance of AI workloads with a high sensitivity to latency [37].

Another major dimension of MLOps benchmarking concerns resource efficiency, particularly the effective use of computing resources during machine learning processes. Cloud infrastructure is one of the biggest operational challenges of enterprises that operate large-scale AI pipelines. When trained on distributed GPU clusters, training deep learning models can be highly costly especially with inefficient resource allocation strategies.

To determine cost-efficient deployment strategies, benchmarking studies often examine the connection between the cost of infrastructure and the computational performance. Findings of numerous works suggest that distributed cloud architectures can decrease the total training time but can cause higher operational costs in cases when resources of clusters are not used efficiently. Unutilized compute resources, inefficient task scheduling and unnecessary inter-node data transfer can also increase infrastructure costs by a large margin [38].

The strategies of optimization of costs are usually associated with the dynamic provisioning mechanisms that can be used to scale the computing resources according to the workload demand. In cloud platforms, auto-scaling systems allow the cloud platform to add more compute nodes at peak training times and remove more nodes at times of reduced workload. These mechanisms come in especially handy with financial institutions, whereby machine learning loads might change considerably based on the situation in the market or the volume of transactions taking place.

The relevance of comparing the cost of data transfer between cloud vendors is also noted in the case of cross-cloud benchmarking research. Implementing multi-cloud architectures may require the movement of large datasets between cloud computing systems when performing training or evaluation of a model. Network data transfer fees can thus constitute a huge proportion of total operations cost. Some of the studies suggest reducing cross-cloud data transfer via local data storage policies or distributed replication of data [39].

Along with computational performance and cost, MLOps benchmarking research also assesses more and more operational governance mechanisms within machine learning pipelines. BFSI businesses are highly regulated with transparency, auditability, and reliability of the algorithmic decision-making systems. Consequently, the machine learning infrastructure should accommodate powerful model governance protocols in a position to address regulatory compliance demands.

Structured lifecycle management models have thus emerged to be an important part of enterprise MLOps architectures. These frameworks have the ability to track different versions of models, monitor model performance, identify model drift and set up automatic retraining processes in the case of performance deterioration. The monitoring systems constantly compare the accuracy of prediction and the distribution patterns of data to detect any changes in the underlying data properties that can influence the model performance [40].

Most of the benchmarking studies have shown that machine learning can be significantly improved with regards to its operational reliability with the help of automated monitoring and retraining. Constant monitoring may further allow organizations to realize the impending performance drop and, by so doing, corrective actions can be taken before the financial or operation blow may be felt greatly. The models can also be permitted to be trained to the changing trend of the financial data with least manual correction by automated retraining pipelines.

The adhered evidence (Table 4) confirms that the most common performance metrics continue to be at the core of assessing cross-cloud MLOps infrastructures yet the meaning of them will also heavily rely on the application scenario. The advantage of reduced duration of training in BFSI is not only due to its capability to improve computational efficiency but also due to the fact that the possibility to retrain more often in case of a change in the market or fraud patterns. Likewise, low inference latency is neither a technical nor operationally significant benefit, in time-constrained applications like fraud detection and real-time decision support.

Table 4. Synthesized Performance Characteristics from Literature

Metric	Typical Observation	Implication for BFSI AI Systems
Training latency	Reduced significantly with distributed clusters	Enables faster model retraining cycles
Inference latency	Highly sensitive to network overhead	Critical for real-time fraud detection
Resource utilization	Often inefficient in poorly balanced clusters	Directly affects infrastructure cost
Scalability	Improved through container orchestration	Supports high-volume financial workloads
Data transfer overhead	Significant in cross-cloud architectures	May increase operational costs

Table 5. Comparative Evaluation of Cloud Infrastructure Factors Affecting AI Pipelines

Infrastructure Factor	Observed Impact	Operational Consideration
GPU acceleration	Large reduction in training time	Higher infrastructure cost
Container orchestration	Improved deployment scalability	Increased architectural complexity
Distributed storage	Faster data access for training	Requires data consistency management
Multi-cloud redundancy	Increased system resilience	Higher network communication overhead
Edge inference deployment	Reduced latency for predictions	Limited model complexity support

Combined, Tables 4 and 5 show that there is a predictable trade-off across the literature that has been reviewed: performance benefits are seldom unconditioned by architectural and cost-related impacts. Example: Distributed clusters and hardware accelerators can significantly decrease training time, but may come with increased cost of infrastructure, more orchestration, or a new communication overhead. In the case of BFSI organizations, this implies that speed should not be the sole factor when making an infrastructure selection. Instead, the best cross-cloud architecture is the one that ensures a balance between responsiveness, resilience, governance needs, and the operational cost in the long term.

However, cross-cloud architectures also bring significant complexity of operation which should be handled cautiously. Issues of operation that are of concern to a distributed cloud environment include key challenges in data synchronization, overhead in the network communication and optimization of the infrastructure cost. They ought therefore to have good orchestration designs and smart structure of workload scheduling to effectively handle cross cloud machine learning pipelines.

The discussion also shows the growing importance of standardized benchmarking infrastructure that may measure the performance of machine learning infrastructure within and across heterogeneous computing environments. Without standardized benchmarking approaches, infrastructure studies cannot be compared in terms of performance and, thus, enterprises cannot make well-informed decisions based on AI deployment strategies.

Overall, the results show that successful implementation of cross-cloud MLOps systems must be premised on the convergence of distributed computing systems, automated pipeline orchestration systems and complete integration of monitoring systems that are able to facilitate regulatory governance requirements in financial environments.

6. Limitations

The research has a few shortcomings since it depends on the literature published before and lack of empirical benchmarking and experimental validation of cloud platforms. Differences in the scope, datasets, methods of evaluation, and reporting standards of previous research can decrease the direct comparability of results between the literature. Also, not all studies reviewed apply to BFSI because some of them are on enterprise machine learning infrastructure in general. However, the review offers a systematic overview of the evidence at hand and suggests a conceptual framework that can underpin future empirical and domain-specific benchmarking research.

7. Future Directions

Among the most evident gaps in the existing literature, one can note the lack of standardized benchmarks that would be used to specifically target machine learning operations across clouds. Even though benchmark suites like MLPerf have enhanced reproducibility, they mainly test the performance of hardware and model training. Consequently, they fail to completely represent the operational complexity of enterprise MLOps pipelines, especially in cases where pipelines are spread among several cloud providers. The way forward in future work then is to benchmark frameworks which assess the entire lifecycle of machine learning systems in cross-cloud environments and not just isolated computational tasks in isolation.

This then needs to be followed by future research work to create holistic benchmarking models that will be in a position to evaluate the whole lifecycle performance of machine learning systems in distributed clouds. This should encompass model training performance metrics, inference latency metrics, system scalability metrics, infrastructure cost efficiency metrics, and capabilities related to governance (model versioning and monitoring). It would be possible to make comparisons between cloud platforms more credible and make evidence-based decisions in relation to infrastructure with the help of standardized benchmarking protocols to BFSI enterprises [41].

In addition to technical benchmarking predictors, future designs would have workload-focused assessment tasks that would emulate real financial AI pipelines. Such workloads might contain fraud detection models, credit risk prediction pipelines, and financial transaction monitoring pipelines. More realistic performance insights would be obtained by benchmarking experiments using such datasets than with generic machine learning tasks.

Cross-cloud machine learning systems require advanced coordination mechanisms to manage distributed computing resources effectively. The existing orchestration systems like Kubernetes and Kubeflow offer highly effective mechanisms to administer containerized machine learning pipelines, but these are usually based on fixed policies of resource allocation, which might not offer optimal performance in heterogeneous cloud systems.

This leaves room to conduct a future study on the integration of intelligent workload scheduling algorithms that can be used to dynamically deploy machine learning tasks on various cloud providers depending on real-time performance indicators. It is possible to use reinforcement learning and optimization algorithms to find the optimal resource allocation strategies to use when distributing AI workloads. Such adaptive schedulers might be able to constantly assess the performance of infrastructure and automatically scale workloads across cloud platforms to enhance performance and lower operations costs [42].

These types of smart orchestration systems would be especially useful to BFSI businesses with large-scale AI pipelines that have very fluctuating workload demand. As an illustration, financial transaction volumes can be very high during peak periods or seasonal occasions of the economy. Such conditions may enhance the efficiency of machine learning systems through adaptive orchestration systems that have the capability to dynamically optimize infrastructure resources. Financial institutions operate under stringent regulatory requirements, which demand transparency within the algorithmic decision-making systems. As a result, the incorporation of explainable artificial intelligence (XAI) methods into enterprise MLOps pipelines is a valuable research field. Although machine learning models have a number of interpretability methods that have been worked out, it is not easy to incorporate the methods into the large-scale distributed AI infrastructure.

In the future, MLOps architectures will have automated explainability systems that are able to produce interpretable model outputs during training and inference. Such mechanisms have potential to allow financial institutions to be regulatory compliant and, at the same time, implement advanced machine learning models. Also, benchmarking systems will contain measures of the computational cost of explainability tools since interpretability methods often introduce additional computational overheads to machine learning systems [43].

Additional studies are necessary to come up with a governance model capable of managing the model lifecycle transparency in distributed infrastructures. These structures can contain automated audit logging that captures the data provenance, model configuration parameters and training processes of each deployed model version. Such audit records would allow the regulatory bodies to check whether there is adherence to financial governance rules.

Data privacy is a major concern for financial institutions that have implemented machine learning systems. Financial data are usually very sensitive information, such as personal financial transactions, credit records and identity checking records. The regulatory frameworks therefore place stringent restrictions on the way in which such data can be stored, processed as well as be transmitted within infrastructure environments.

Federated learning has become one of the promising methods of facilitating collaborative machine learning within distributed data sources and maintaining data privacy. Federated learning designs have machine learning models that are trained on decentralized data without the need to

aggregate data centrally. Rather, the updates of models are exchanged among the involved nodes with the data underlying the model being localized within the original environment.

Future studies ought to be conducted regarding the incorporation of federated learning methods into cross-cloud MLOps systems in applications of BFSI. These architectures may allow financial institutions to develop models of machine learning through collaborating without providing sensitive datasets to each other. Besides, benchmarking research needs to assess the performance implication of federated learning structures, especially in terms of communication overhead and the rate of model convergence [44].

Differential privacy and secure multiparty computation are privacy-preserving machine learning methods that can be important in future financial AI infrastructures. Such methods allow machine learning models to be trained with encrypted or anonymized data and have statistical value. Such approaches to enterprises as an MLOps pipeline are a significant research problem demanding additional study.

The information processing requirements of the large-scale AI workloads have increased the concern over the ecological footprint of machine learning infrastructures. The process of training deep learning models in a distributed set of GPUs may consume a significant portion of energy and play a role in the carbon footprint of the data centre. The machine learning infrastructure will be a consideration as financial institutions transition more to AI-driven analytics systems.

In the next round of benchmarking studies, attention should therefore be given to energy efficiency indicators in assessing cloud-based machine learning systems. Such metrics may quantify the energy usage of the process of model training and model inference on various hardware architectures and cloud platforms. The implementation of AI can also be more sustainable with the help of studies on benchmarking since they can offer information about energy-saving infrastructure designs [45].

Special hardware accelerators can also be developed to make machine learning workloads more energy-efficient. New classes of processor architectures that are explicitly tailored to computation in machine learning can radically decrease the energy used to train large models. Much of the future research focus on evaluating the performance and sustainability implications of such hardware architectures is of interest.

Financial markets produce vast amounts of real-time data, such as exchange records, market prices and economic variables. Conventional machine learning pipelines have typically been based on batch-based processing systems that process historical data as opposed to data streams that change over time. But applications like fraud detection, algorithmic trading, or dynamic risk management are becoming more and more dependent on real-time analytics applications.

The next generation of cross-cloud MLOps should thus include streaming data processing frameworks that can be used to run real-time machine learning pipelines. Apache Kafka and distributed event-processing systems are the technologies that could be used to allow ongoing ingestion of financial data streams and subsequent analysis. With this kind of technologies using machine learning to organize structures, financial institutions could have the opportunity to utilize dynamic AI systems, which could react to the rapidly changing market conditions [46]. Similarly, these streaming machine learning pipelines are to be benchmarked in terms of performance measures like latency, throughput, fault tolerance, and scalability.

8. Conclusion

The study presented a literature review of the subject of benchmarking AI workloads to cross-cloud MLOps in the BFSI business. The overview has summarized the literature available on distributed machine learning infrastructure, enterprise MLOps systems, and benchmarking dimensions that can be utilized in the financial AI implementation environment.

Its findings indicate that cross-cloud machine learning systems are useful in terms of scalability, resilience and flexibility of deployment, but introduce new challenges of complexity, orchestration, overhead in data transfer, cost control and governance requirements. The reviewed literature identifies the latency, throughput, resource utilization, and cost as the key dimensions of benchmarking and the BFSI environments include the extra attention to auditability, explainability, compliance, and operational transparency.

This study has come up with a conceptual benchmarking model founded on the synthesis and integrates the performance, architectural and BFSI-specific governance dimensions. The

framework will provide a systematic basis on the future benchmarking research, and could lead to the development of more standardized and domain specific assessment frameworks to enterprise AI systems in financial environments.

References

- [1] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255–260.
- [2] Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems*, 28, 2503–2511.
- [3] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.
- [4] Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., & Zaharia, M. (2010). A view of cloud computing. *Communications of the ACM*, 53(4), 50–58.
- [5] Li, A., Yang, X., Kandula, S., & Zhang, M. (2010). CloudCmp: Comparing public cloud providers. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement* (pp. 1–14).
- [6] Kou, G., Peng, Y., & Wang, G. (2014). Evaluation of clustering algorithms for financial risk analysis using MCDM methods. *Information Sciences*, 275, 1–12.
- [7] Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. *Communications of the ACM*, 51(1), 107–113.
- [8] Voigt, P., & von dem Bussche, A. (2017). *The EU General Data Protection Regulation (GDPR): A practical guide*. Springer.
- [9] Zaharia, M., Chen, A., Davidson, A., Ghodsi, A., Hong, S. A., Konwinski, A., Murching, S., Nykodym, T., Ogilvie, P., Parkhe, M., Xie, F., & Zumar, C. (2018). Accelerating the machine learning lifecycle with MLflow. *IEEE Data Engineering Bulletin*, 41(4), 39–45.
- [10] Kreuzberger, D., Kühl, N., & Hirschl, S. (2023). Machine learning operations (MLOps): Overview, definition, and architecture. *IEEE Access*, 11, 31866–31879.
- [11] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115.
- [12] Burns, B., Grant, B., Oppenheimer, D., Brewer, E., & Wilkes, J. (2016). Borg, Omega, and Kubernetes. *Communications of the ACM*, 59(5), 50–57.
- [13] Pahl, C. (2015). Containerization and the PaaS cloud. *IEEE Cloud Computing*, 2(3), 24–31.
- [14] Leitner, P., & Cito, J. (2016). Patterns in the chaos: A study of performance variation and predictability in public clouds. *ACM Transactions on Internet Technology*, 16(3), Article 15.
- [15] Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2010). Spark: Cluster computing with working sets. In *Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing*.
- [16] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., ... Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation* (pp. 265–283).
- [17] Huppler, K. (2009). The art of building a good benchmark. *ACM SIGMETRICS Performance Evaluation Review*, 37(4), 18–30.
- [18] Mattson, P., Cheng, C., Damos, G., Coleman, C., Micikevicius, P., Patterson, D., Tang, H., Wei, G.-Y., Bailis, P., Bittorf, V., Brooks, D., Chen, D., Dyzma, M., Gardner, A., Hubara, I., Kanter, D., Kumar, N., Liao, J., Narayanan, D., ... Zaharia, M. (2020). MLPerf: An industry standard benchmark suite for machine learning performance. *IEEE Micro*, 40(2), 8–16.
- [19] Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, 50(3), 559–569.

- [20] Li, Z., O'Brien, L., Zhang, H., & Cai, R. (2012). On a catalogue of metrics for evaluating commercial cloud services. In *Proceedings of the 13th International Conference on Grid Computing* (pp. 164–173).
- [21] Matskin, M., Tahmasebi, S., Layegh, A., Payberah, A. H., Thomas, A., Nikolov, N., & Roman, D. (2021). A survey of big data pipeline orchestration tools from the perspective of the DataCloud project. In *Data Analytics and Management in Data Intensive Domains 2021*.
- [22] Llorido-Botran, T., Miguel-Alonso, J., & Lozano, J. A. (2014). A review of auto-scaling techniques for elastic applications in cloud environments. *Journal of Grid Computing*, 12(4), 559–592.
- [23] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229).
- [24] Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. In *2015 IEEE Symposium Series on Computational Intelligence* (pp. 159–166).
- [25] Islam, S., Keung, J., Lee, K., & Liu, A. (2012). Empirical prediction models for adaptive resource provisioning in the cloud. *Future Generation Computer Systems*, 28(1), 155–162.
- [26] Jouppi, N. P., Young, C., Patil, N., Patterson, D., Agrawal, G., Bajwa, R., Bates, S., Bhatia, S., Boden, N., Borchers, A., Boyle, R., Cantin, P.-L., Chao, C., Clark, C., Coriell, J., Daley, M., Dau, M., Dean, J., Gelb, B., ... Yoon, D. (2017). In-datacenter performance analysis of a Tensor Processing Unit. In *Proceedings of the 44th Annual International Symposium on Computer Architecture* (pp. 1–12).
- [27] Buyya, R., Vecchiola, C., & Selvi, S. T. (2013). *Mastering cloud computing: Foundations and applications programming*. Morgan Kaufmann.
- [28] Breck, E., Cai, S., Nielsen, E., Salib, M., & Sculley, D. (2017). The ML test score: A rubric for ML production readiness and technical debt reduction. In *2017 IEEE International Conference on Big Data* (pp. 1123–1132).
- [29] Calheiros, R. N., Ranjan, R., Beloglazov, A., De Rose, C. A. F., & Buyya, R. (2011). CloudSim: A toolkit for modelling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. *Software: Practice and Experience*, 41(1), 23–50.
- [30] Ben-Nun, T., & Hoefler, T. (2019). Demystifying parallel and distributed deep learning: An in-depth concurrency analysis. *ACM Computing Surveys*, 52(4), Article 65.
- [31] Li, M., Andersen, D. G., Park, J. W., Smola, A. J., Ahmed, A., Josifovski, V., Long, J., Shekita, E. J., & Su, B.-Y. (2014). Scaling distributed machine learning with the parameter server. In *Proceedings of the 11th USENIX Symposium on Operating Systems Design and Implementation* (pp. 583–598).
- [32] Newman, S. (2015). *Building microservices: Designing fine-grained systems*. O'Reilly Media.
- [33] Satyanarayanan, M. (2017). The emergence of edge computing. *Computer*, 50(1), 30–39.
- [34] Gholami, A., Kim, S., Yao, Z., Mahoney, M. W., & Keutzer, K. (2021). AI and machine learning for cloud computing. *IEEE Micro*, 41(6), 33–40.
- [35] Peng, Y., Bao, Y., Chen, Y., Wu, C., & Guo, C. (2018). Optimus: An efficient dynamic resource scheduler for deep learning clusters. In *Proceedings of the Thirteenth EuroSys Conference* (pp. 1–14).
- [36] Dragoni, N., Giallorenzo, S., Lluch-Lafuente, A., Mazzara, M., Montesi, F., Mustafin, R., & Safina, L. (2017). Microservices: Yesterday, today, and tomorrow. In M. Mazzara & B. Meyer (Eds.), *Present and ulterior software engineering* (pp. 195–216). Springer.
- [37] Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. *IEEE Internet of Things Journal*, 3(5), 637–646.
- [38] Walker, E. (2009). The real cost of a CPU hour. *Computer*, 42(4), 35–41.
- [39] Corodescu, A.-A., Nikolov, N., Khan, A. Q., Soylu, A., Matskin, M., Payberah, A. H., & Roman, D. (2021). Big data workflows: Locality-aware orchestration using software containers. *Sensors*, 21(24), 8212.
- [40] Schlegel, M., & Sattler, K.-U. (2022). Management of machine learning lifecycle artifacts: A survey. *SIGMOD Record*, 51(4), 18–29.
- [41] Folkerts, E., Alexandrov, A., Sachs, K., Iosup, A., Markl, V., & Tosun, C. (2013). Benchmarking in the cloud: What it should, can, and cannot be. In *Selected Topics in Performance Evaluation and Benchmarking (Lecture Notes in Computer Science, Vol. 7755, pp. 173–188)*. Springer.

- [42] Mao, H., Netravali, R., & Alizadeh, M. (2016). Resource management with deep reinforcement learning. In Proceedings of the 15th ACM Workshop on Hot Topics in Networks (pp. 50–56).
- [43] Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing, and interpreting deep learning models. *IEEE Signal Processing Magazine*, 34(6), 76–86.
- [44] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., El Rouayheb, S., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., ... Zhao, S. (2021). Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2), 1–210.
- [45] Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645–3650).
- [46] Kreps, J., Narkhede, N., & Rao, J. (2011). Kafka: A distributed messaging system for log processing. In Proceedings of the NetDB Workshop (pp. 1–7).