Kelvin Open Science Publishers
Connect with Research Community

Research Article *Volume 2 | Issue 1*

KOS Journal of AIML, Data Science, and Robotics
https://kelvinpublishers.com/journals/aiml-data-science-robotics.php

Streaming Telemetry Analytics for AI-Driven Service Performance Prediction in Large-Scale Microservices Architectures

Manisha Konda[1*] and Kamalakar Reddy Singi[2]

[1]Senior Analyst, Analytics, Starcom (Publicis Groupe), USA
[2]Senior Software Engineer, Valparaiso university, USA

[*]**Corresponding author:** Manisha Konda, Senior Analyst, Analytics, Starcom (Publicis Groupe), USA

Received: May 21, 2026; **Accepted:** May 25, 2026' **Published:** May 27, 2026

Citation: Manisha Konda, et al. (2026) Streaming Telemetry Analytics for AI-Driven Service Performance Prediction in Large-Scale Microservices Architectures. *KOS J AIML, Data Sci, Robot.* 2(1): 1-8.

1. Abstract

Distributed systems today, such as cloud infrastructures, microservice architectures, and hybrid cyber-physical networks, are increasingly difficult to monitor with respect to fault detection and diagnostics because of their large size, complexity, and dynamics. Telemetry data, a record of detailed operational metrics and event logs, is important for understanding how a system behaves and for proactively maintaining it. The paper describes an AI-based telemetry system capable of predicting service performance in a microservices system. The architecture leverages streaming telemetry, such as system metrics, logs, and traces, from cloud-native setups. The data are cleaned and normalized, and their features are extracted using Principal Component Analysis (PCA) to enhance data quality and reduce dimensionality. To learn temporal patterns of dependencies and identify anomalies in service behavior, an LSTM-based deep learning model is employed. A privacy protection layer is a safety measure that ensures the secure handling of sensitive information. The presented model achieves 94% accuracy with excellent predictive performance, low detection time, and few false positives. The comparative analysis reveals that the framework is superior to existing solutions and can be used to reliably monitor microservice performance in real time.

2. Keywords:

Microservices Architecture, Service Performance Prediction, Telemetry Data, Artificial Intelligence, Machine Learning, Privacy Preservation, Real-Time Monitoring.

3. Introduction

Over the last ten years, distributed software systems have been more popular due to the growing need for Internet of Things applications and advancements in cloud infrastructure [1,2]. As a new architectural style that encourages the use of fine-grained services and collaboration across cloud nodes, microservice architecture (MSA) has arisen in response to this need and in reaction to advancements in service-oriented architecture (SOA) [3,4]. Software applications that use SOA or monolithic style often encounter issues with availability, fault tolerance, scalability, and maintainability [5,6]. The architecture of microservices is gaining interest in both academia and business, and it is often contrasted with monolithic design [7-9]. Many of the findings of these research publications contradict each other in terms of performance of various designs. Network devices are producing an increasing amount of data, including control, statistics, and user data [10,11].

Telemetry data is essential for maintaining the integrity of spacecraft systems and guaranteeing the success of spacecraft missions. Therefore, to guarantee the safe functioning of spacecraft subsystems, it is essential to promptly identify and notify of any unexpected occurrences connected to their functionality [12]. There have been a number of approaches to anomaly identification in spacecraft telemetry data monitoring that have been developed in the last few years [13,14]. After many decades, the commercial and military telemetry groups reaped the rewards of their investment in these standards.

The microservices architecture is radically different in approach compared to traditional monolithic applications, as it involves a system of distributed services, each of them encapsulating a specific business function or capability [15,16]. It has become very popular because of its flexibility, scalability, and adaptability regarding the handling of complex software demands across various industries. Recent developments in artificial intelligence (AI) offer bright solutions to such problems. Reinforcement learning (RL), predictive analytics (PA), and evolutionary algorithms (EA) are some of the AI methods that promise more advanced resource optimization methods [17-19]. The benefit of RL is that systems can learn optimal policies for allocating resources through interactions with their environment and feedback on the outcomes of their actions. Artificial intelligence (AI) is a tool that can potentially offer solutions to these challenges by automating and improving the creation of microservice architectures, especially in new software development [20,21]. Machine learning (ML), and natural language processing (NLP) are used as AI techniques to help with critical design activities, including requirement analysis, service identification, and architectural decision making [22,23]. There has been a convergence of telemetry data with data mining, ML, and DL over the past ten years. This partnership aims to implement new automated techniques (on-board) and semi-automated techniques (at the ground station). Some of the space systems' health monitoring processes can be automated using these methods [24,25]. Many DM, ML, and DL techniques have been proposed in recent years to address telemetry anomaly detection and health monitoring in aerospace systems.

The key contributions of this research are as follows:

- Design of a real-time AI-driven telemetry analytics framework for service performance prediction in large-scale microservices architectures using streaming metrics, logs, and traces.
- Integration of a Kubernetes-native observability pipeline, leveraging container orchestration environments (Kubernetes clusters with Prometheus and Open Telemetry) for real-world cloud deployment compatibility.
- Development of a hybrid deep learning model (PCA + LSTM) enhanced with a privacy-preserving layer, enabling accurate, scalable, and secure performance prediction with reduced dimensionality and temporal dependency learning.

3.1. Motivation and novelty of the study

Large-scale microservice architectures are being used more and more in modern cloud-native applications to provide scalability, flexibility, and rapid deployment, but they are distributed and highly dynamic and therefore difficult to ensure that service performance is consistent. Conventional methods of monitoring with fixed limits and reactive alarms sometimes fail to notice small-scale anomalies or anticipate the existence of performance decreases in real-time especially when workloads fluctuate, and there exist complicate inter-service relationships. In the meantime, microservice-based systems continuously produce vast amounts of telemetry data, such as CPU utilization, memory usage, network throughput, and latency, and manual analysis is not feasible, and traditional methods are not powerful enough to discover nonlinear temporal relationships. Recent changes in artificial intelligence and deep learning promise to do this prediction of performance in advance based on incoming data, but many current systems do not scale to deploy at large scale or respond to performance in real time.

The novelty and validity of this study are the sophistication and changeability of the microservices architecture and the necessity to address the issue with an artificial intelligence-based telemetry framework. The proposed algorithm is a combination of live stream telemetry logs and an AI-based controlled layered model to predict the service performance accurately and detect anomalies by detecting temporal dependencies in system behavior. It is a state-of-the-art with regards to privacy, unlike alternative methods; it merges end-to-end pre-processing, PCA-based feature reduction, and a privacy-preserving layer. The design enables the high-volume telemetry data to be analyzed appropriately, quick and scalable and with low overhead on monitoring. As the existing approaches usually prove to be unwieldy, fail to capture the time-related features, or cause excessive expenses, the proposed framework provides a critical and proactive answer to the problem of performance prediction by making system stability, downtime reduction, and the efficient management of the cloud-native microservices environment.

3.2. Organization of the paper

The paper is structured in the subsequent way: Section II provides a review of related literature on AI-based telemetry analytics frameworks in Microservices Performance. Section III explains the methodology, i.e. data pre-processing, model development and evaluation. In Section IV, the results of the experiments and comparison of the models are given. Section V wraps up with fundamental conclusions and recommendations about the future research on enhancing the efficiency of the commercial computing systems.

4. Literature Review

The literature shows the enhancement of real-time performance, fault detection, scalability, and efficiency through the use of AI-driven telemetry, predictive analytics, and microservices architectures, exposing gaps in predictive performance modelling and edge deployment.

S. O. Awodele, et al. introduce RCA Sage, an AI-augmented inference engine for autonomous root cause analysis in microservices. LSTM/NLP anomaly detection, and an Autonomous Inference Engine integrating GNNs with Neural Granger Causality. RCA Sage minimizes the MTTR by more than 90 percent, offers clear causal explanations with XAI, and addresses DevOps with JIT defect predictions [26].

M. K. Gaddam proposes a scalable observability architecture specifically designed to support AI-driven microservices, that is, the natural opacities and dynamics of AI workloads. The framework proposed provides an opportunity to perform

better visibility over the whole AI pipeline, including data ingestion and model inference by combining fine-grained telemetry with trace correlation, anomaly detection, and model-centric metrics. The methodology shows enhanced root cause analysis, decreased system downtime, and heightened openness of model performance behaviour. The architecture is validated through a real-world deployment on a Kubernetes-based AI platform, showing minimal performance overhead while maintaining high fidelity in observability signals [27].

G. Dkmak, et al. introduces an innovative unsupervised method of microservices anomaly detection the Night Watch algorithm. The methodology eliminates significant limitations of existing approaches through temporal considerations and multi-source information. Depending on the size of the training set, their findings show that the Night's Watch algorithm greatly enhances recall (39% improvement) and precision (92% improvement). These results show that the algorithm is capable of improving real-time anomaly detection in microservice setup [28].

B. Barua and M. S. Kaiser, create a system for real-time travel reservations using a microservices architecture and prescriptive analytics. By decoupling components, the system achieves modularity, scalability, and fault tolerance. ML models optimize demand forecasting, dynamic pricing, and performance, improving response time, throughput, transaction success, and prediction accuracy. This allows making data-based decisions in time and increasing operational efficiency, which is a blueprint to other multifaceted, data-driven applications [29].

D. K. Pentyala, suggests an AI-based system of fault identification in cloud-optimized data engineering systems, using the techniques of machine learning models to process telemetry, logs and performance indicators in real-time. The framework covers system failures predicting anomalies and optimizing resources utilization minimizing down time and enhancing reliability. Quantitative evaluation reveals it is much better, e.g., it reduces the fault detection latency by 78% and resource efficiency by 65% compared to traditional monitoring plans, which demonstrates the extent to which AI-enabled predictive analytics can be used to monitor cloud services [30].

G. P. Menaud, suggest a hybrid ML-based system that entails the combination of past trace statistics and instant telemetry data to predict and optimize performance. The architecture employs supervised learning to make prediction and reinforcement learning to optimize adaptively. Their analysis and suggested model reveal that with different workloads, ML performs much better in terms of flexibility and efficiency of micro services and decreases the latency and enhances the throughput more than 30 times in simulated deployments [31].

Table 1 provides the summary of the methods, key results, advantages, constraints, and prospective studies of the recent research and reveals the gaps in predictive analytics and scalable monitoring with the help of microservices.

Table 1: Research Gaps in Ai-Driven Telemetry and Microservices Performance Prediction.

Author (Year)	Methodology	Key Findings & Advantages	Limitations	Future Work
S. O. Awodele, et al. (2026)	RCASage: AI-augmented inference engine; three-stage pipeline (telemetry ingestion, LSTM/NLP anomaly detection, Autonomous Inference Engine with GNN + Neural Granger Causality)	Reduces MTTR by >90%; identifies true root causes; integrates Explainable AI (XAI); bridges DevOps via JIT defect prediction	Focused primarily on root cause analysis rather than direct service performance prediction; requires extensive telemetry data	Extend to predictive performance optimization for microservices; integrate edge computing telemetry
M. K. Gaddam (2025)	Scalable observability architecture for AI-driven microservices; telemetry, trace correlation, anomaly detection, model-centric metrics; deployed on Kubernetes	Improved visibility across AI pipeline; minimal overhead; reduces downtime; enhances model transparency	Lacks explicit predictive performance modeling; limited evaluation metrics beyond observability	Extend framework to predictive analytics for latency/failure forecasting; incorporate AI-based performance prediction models
G. Dkmak, et al. (2025)	Night's Watch algorithm: unsupervised anomaly detection in microservices; multi-source data + temporal features	Precision up to 92%, recall up to 39%; reduces false positives; enhances real-time anomaly detection	Focused only on anomaly detection, not root cause analysis or predictive performance; recall is moderate	Integrate predictive models for service degradation; enhance recall and scalability for large microservices systems
B. Barua & M. S. Kaiser (2024)	Microservices architecture with predictive analytics; ML models for demand forecasting, dynamic pricing, and system performance optimization	Improves response time, throughput, transaction success; scalable and modular; supports real-time analytics; enhances customer satisfaction	Case study limited to travel reservation systems; does not address cross-service dependencies or telemetry-driven anomaly detection	Extend to AI-driven microservices monitoring and real-time failure prediction; test in other domains like healthcare and industrial systems
D. K. Pentyala (2024)	AI-driven fault identification in cloud-optimized data engineering systems; ML	Reduces fault detection latency by 78%; increases	Focused on fault detection, not predictive service	Integrate with real-time performance prediction in microservices; extend

	models for telemetry, logs, and performance metrics	resource efficiency by 65%; improves reliability	performance; limited to cloud infrastructure, not microservices	to multi-layered distributed architectures
G. P. Menaud (2023)	Hybrid ML framework: historical trace analysis + real-time telemetry; supervised learning for prediction, reinforcement learning for adaptive optimization	Reduces latency and increases throughput by >30%; improves adaptability under varying workloads	Simulated deployment; lacks real-world important microservices validation	Apply in real-world large-scale microservices; integrate edge computing telemetry; include multi-service failure prediction

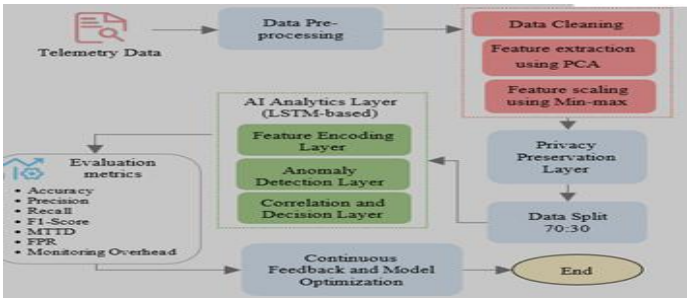
5. Methodology

The proposed framework is to be used to predict the performance of AI-driven services in the microservices architecture of large scale as illustrated in Fig. 1. The system is deployed to a Kubernetes-based cloud-native environment, with microservices being deployed in the form of containerized workloads. Every service is an independent pod, which is orchestrated by Kubernetes, which supports dynamic scaling, load balancing, and self-healing in the fluctuating workload scenarios. The framework begins with gathering of telemetry data where operational metrics such as CPU utilization, memory utilization, network throughput and service latency are continuously emitted by distributed microservices. Prometheus is used to gather telemetry data in terms of metrics scraping and Open Telemetry is used to provide distributed tracing and logging, which keeps the service interactions in the cluster visible end-to-end. The resulting telemetry streams are then work upon at the data pre-processing stage that involves cleaning of data, feature extraction by application of Principal Component Analysis (PCA) and feature scaling through application of Min-Max normalization in order to achieve a homogenous representation of the features. A privacy-preserving layer is also built to anonymize sensitive identifiers within the telemetry streams and then proceed to further analysis. The generated data are then subjected to the AI analytics layer, where the temporal feature encoding and anomaly detection are achieved through the assistance of DL techniques to identify the abnormal behavior of the system and predict service degradation. Lastly, a feedback and model optimization system should be continuous to allow the system to respond to the changing telemetry patterns.

5.1. Dataset description

The data available in this research is streaming telemetry data that is gathered in cloud-native microservices setups. The data contains the system metrics like CPU usage, memory usage, network throughput and service latency, application logs and distributed traces. In Kubernetes-based systems, the monitoring systems can be used to obtain these telemetry signals including Open Telemetry and Prometheus.

Figure 1: AI-Based Telemetry Analytics Framework for Microservices Performance.



5.2. Data preprocessing

The telemetry data has been collected and initially pre-processed to guarantee the quality and consistency of the data. This phase involves a number of processing steps data.

- **Data cleaning:** Telemetry records which are invalid, duplicate, or inconsistent are eliminated, in order to permit consistent observations.
- **Noise filtering:** The use of smoothing methods to minimize spike anomalies, transient peaks and monitoring artifacts in such measurements as CPU usage, latency, and network throughput.
- **Removal of corrupted records:** Discovery and deletion of telemetry records that lack timestamps or do not contain full metric values or have corrupted logs to ensure data integrity.
- **Consistency verification:** Checking of the telemetry times and range of metrics to keep the value in reasonable working ranges.
- **Data standardization:** Organizing the cleaned telemetry data into a standard form to extract feature easily and then analyze the data using ML.

Pre-processing is important to make sure that abnormal spikes and monitoring artifacts do not adversely affect the learning of models.

5.3. Feature extraction

Use of feature extraction helps to minimize redundancy in the attributes of telemetry but maintain the most informative attributes of the data set. Correlated metrics, like the use of CPU, consumption of memory, network throughput, and service latency, are found in the telemetry streams in large-scale microservices environments. These associations are capable of raising the complexity of computing and lowering the effectiveness of ML models. Hence, PCA is used as a dimensionality reduction method to condense the initial telemetry feature space into a smaller collection of independent components. PCA determines directions of maximum variance in data and retains significant components, which practically eliminates redundancy and preserves the required patterns for predicting service performance using AI.

5.4. Feature scaling using min-max

Normalization of range of telemetry attributes in a microservices dataset is done using feature scaling. All features then are scaled with min-max normalization to range of 0-1. The change guarantees equal input of features and the learning efficiency of the model. The Min–Max normalization formula is defined in Equation (1):

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad (1)$$

Here, X is original feature value, X_{min} and X_{max} are minimum and maximum values of feature, and X' represents the normalized value after scaling.

5.5. Privacy preservation layer

To safeguard sensitive operational data, the framework incorporates a privacy-preserving layer before AI processing. Telemetry data can include identifiable data, such as service instance identifiers or infrastructure addresses. Consequently, hash-based pseudonymization is considered to be an anonymization approach that conceals sensitive identifiers. This layer provides secure data processing whilst losing the analytical value of telemetry signals.

After pre-processing, the data is separated into prediction model training/test sets to evaluate prediction model. The telemetry data used in this study is in 70:30 ratio, 30 percent of the data is allocated to testing and evaluation, and 70 percent to training.

5.6. AI analytics layer

The AI analytics layer is the central element of the suggested framework to predict service performance degradation in large-scale microservice architectures. This layer has three sublayers:

5.6.1. Feature encoding layer: The feature encoding layer transforms the telemetry data which has been processed into structured sequential inputs that are accepted by LSTM network. Distributed microservice telemetry, such as CPU usage, memory usage, service latency, and network traffic, are arranged in time-ordered sequences reflecting the behavior of a system over time [32]. This sequential form helps the LSTM to capture patterns and time dependencies to be ready to achieve effective learning and prediction of performance.

5.6.2. Anomaly detection layer: An LSTM network is used as the anomaly detection layer to learn the past system behavior and detect the abnormal deviation in the telemetry patterns that are signs of performance degradation. LSTMs are time-series models that are also suitable because they are able to capture long-term dependencies by using internal memory structures to model them. The functions of LSTM such as forget gate, input gate, candidate cell state, cell state update, output gate, and hidden state are described collectively in Equations (2) to (7), and make the model effective in terms of capturing the temporal relationships and detecting anomalies:

$$forget\ gate = f_t = \sigma\left(W_f.[h_{t-1}, x_t] + b_f\right) \quad (2)$$

$$Input\ gate = \ i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \quad (3)$$

$$Cell\ Candidate = \tilde{C}_t = \tanh(W_c.[h_{t-1}, x_t] + b_c) \quad (4)$$

$$Cell\ State\ update = \ C_t = f_t.C_{t-1} + i_t.\tilde{C}_t \quad (5)$$

$$Output\ gate = \ o_t - \sigma(W_o.[h_{t-1}, b_o] + b_o) \quad (6)$$

$$hidden\ state = h_t = o_t.tanh(C_t) \quad (7)$$

The LSTM network through these gated operations captures time relationship in telemetry signals and detects the abnormal deviation in service performance metrics.

5.6.3. Correlation and decision layer: The final prediction on system performance conditions is produced by the correlation and decision layer. This layer interprets the representations learned by the LSTM and checks the correlation between various telemetry metrics, in order to decide whether the system is functioning normally or there is a performance degradation in the system. This layer compares the trends of telemetry patterns that are observed between various monitoring attributes to generate final anomaly warnings and service performance forecasts. These forecasts can be used to monitor proactively through the detection of possible failures of the services that might cause a major effect on the operations of the systems.

5.7. Performance assessment metrics

The standard ML metrics are used to evaluate the performance of the proposed framework. In order to understand these metrics, one should define the key terms:

- **TP (True Positives):** This represents the number of service performance degradations that are predicted as an anomaly, which are correct.
- **FP (False Positives):** The value of the number of normal system behaviours that are falsely predicted to be anomalous.
- **FN (False Negatives):** These are service performance degradations that are predicted to be normal.
- **TN (True Negatives):** The count of behaviours of the normal system that were rightly forecasted as normal.

The performance measures using these are:

5.7.1. Accuracy: The ratio of properly identified measurements is the statistic used to assess the system's efficiency; it is given in Equation (8):

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (8)$$

5.7.2. Precision: The accurately anticipated positive instances are measured by this metric. Equation (9) measures it as the ratio of the properly categorized values (TP) to the total forecasted values (TP + FP):

$$Precision = \frac{\text{TP}}{\text{TP+FP}} \quad (9)$$

5.7.3. Recall: Sensitivity is a measure that takes into account both the number of items in the dataset (TP + FN) and the proportion of properly categorized values (TP/FN). Therefore, Equation (10) shows the formula:

$$Recall = \frac{TP}{TP+FN} \quad (10)$$

5.7.4. F1 score: This statistic, which goes by many names than just "F Measure", is weighted average of two metrics: recall and precision. Equation (11) yields a minimum of 0, and a maximum of 1, indicating the classifier's optimal performance:

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (11)$$

MTTD (Mean Time to Detection): The average time taken by the system to detect anomalies or performance degradations, with lower values indicating faster detection. Mathematically, represented in Equation (12):

$$\text{MTTD} = \frac{Total\ Detection\ Time\ for\ All\ Events}{Number\ of\ Detected\ Events} \quad (12)$$

False Positive Rate (FPR): The proportion of normal system

behaviours incorrectly flagged as anomalies. A more dependable monitoring system is indicated by a lower FPR. Formulated in Equation (13).

$$FPR = \frac{FP}{FP+TN} \qquad (13)$$

Monitoring Overhead: The ratio of extra system resources used by the monitoring framework, which is efficient in a large-scale microservices setting with the formula presented in Equation (14).

$$\text{Overhead } (\% = \frac{\text{Monitoring Resource Usage}}{Total\ System\ Resource} \times 100 \qquad (14)$$

This measure guarantees the suggested framework's continued effectiveness in extensive microservices environments.

5.8. Continuous feedback and model optimization

The last step of the methodology involves continuous feedback to improve the performance of the model after some time. The events of the real system are compared with the predictions of the AI model with a perspective of identifying the errors in prediction. The model can be retrained according to the feedback with new telemetry information gathered hence giving it the ability to adapt to varying system behaviours. It is a constant learning process that will result in the framework being highly predictive in a dynamic microservices environment.

6. Result Analysis and Discussion

These findings show that the PCA-based dimensionality reduction and LSTM-based temporal modelling are effective to enhance feature representation and sequential dependency learning in a microservice telemetry environment of Kubernetes. All experiments have been carried out in Python on Scikit-Learn and an edge computer system using an ARM-based processor and 16 GB RAM that have been optimized with low-latency inference. Since Table II shows that the balance between predictive accuracy and operational efficiency is very high, this implies the applicability of the proposed framework to real-time microservices monitoring with low detection latency and controlled overhead.

Table 2: Predictive performance of the proposed model in microservices-based service performance forecasting.

Metric	Proposed AI Framework
Accuracy	0.94
Precision	0.93
Recall	0.90
F1-Score	0.91
MTTD (min)	5.8
False Positive Rate	0.07
Monitoring Overhead	21%

The confusion matrix in **Figure 2** is well balanced and there is a high proportion of correct classifications in the main diagonal, which demonstrates a high separability between the normal and degraded service states. The relatively small values of false positives (72) and false negatives (49) confirm the strength of the model to reduce the error of misclassification, which is important to ensure the reliability of cloud-native monitoring systems.

Figure 2: Confusion Matrix of the Proposed Model.

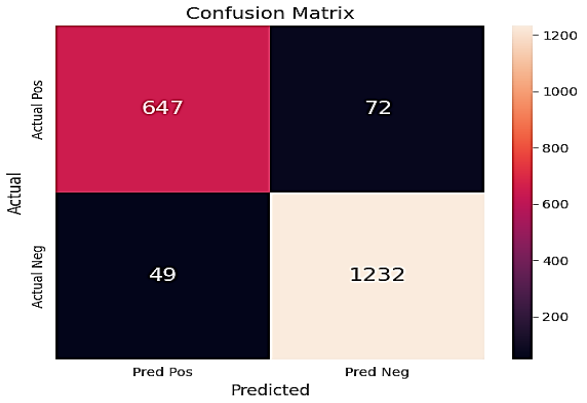


Figure 3: Performance of the Proposed AI Framework.

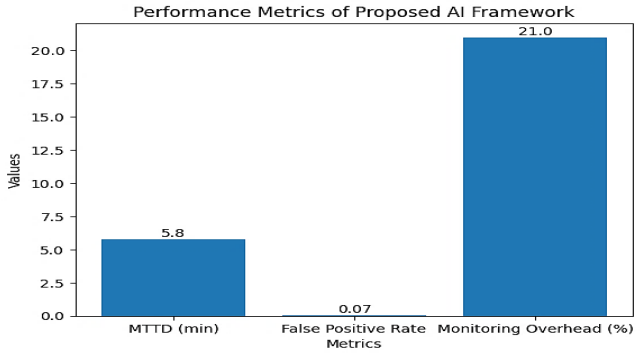


Figure 3 shows the efficiency of the proposed framework in operations with a low Mean Time to Detect (5.8 minutes) and makes it possible to respond practically to service degradation in dynamic microservices environments. The fact that the false positive rate (0.07) is low guarantees a high level of alertness whereas the monitoring overhead (21%), is a practical trade-off between computational cost and predictive performance in edge-enabled applications.

A. Comparative Analysis

The comparative outcomes indicate that the proposed system is the most effective in terms of all evaluation measures compared to the existing baseline strategies. According to Table III, the proposed model outperforms classic machine learning baselines, such as Random Forest (RF), Gradient Boosting (GB), and Isolation Forest. RF and GB have moderate precision and recall, but cannot discover temporal dependencies in telemetry data streams. Isolation Forest, which is an unsupervised approach, has lower recall because it has a low ability to distinguish complex performance degradation patterns. However, the suggested LSTM-based model demonstrates better and balanced performance, being able to capture the temporal relationships and increase the accuracy of service performance prediction in the microservices environments significantly.

Table 3: Comparative analysis of service performance across microservices architectures using telemetry data.

Ref.	Accuracy	Precision	Recall	F1-Score
RF [33]	-	85	78	81
GB [34]	-	89	86	87.5
Isolation forest [35]	90.19	78.42	60.35	68.59
Proposed	94.0	93.0	90.0	91.0

The proposed AI-based telemetry model demonstrates obvious superiority to the existing solutions, providing more

robust predictive ability, reliability, and detection. Compared to the methods that have been previously used, the proposed model demonstrates a reduced or partial level of achievement on the key metrics, but it is more balanced and robust, which basically reduces the number of errors in addition to the possibility to detect the issue and use resources more efficiently. This highlights its relevance to predicting the performance of microservices accurately and on-demand within a real-time environment.

B. Discussion

The experimental findings confirm the usefulness of the suggested AI based telemetry framework to enhance predictive accuracy and real-time responsiveness of microservices environments, and the results have demonstrated classification accuracy, precision, recall and F1-score of 90 to 94 percent. Moreover, Confusion Matrix verification shows that the system is extremely strong, and a significant percentage of correct classifications and a small percentage of misclassifications are obtained, which means that it is possible to discriminate between normal and degraded service states successfully. The fact that the mean time to detect anomalies is low (5.8 minutes) indicates that this framework is capable of near real-time anomaly detection that is essential to dynamic cloud-based services. The low false positive rate (0.07) also provides assurances of reliability with minimal unnecessary alerts, and the 21% monitoring overhead provides a reasonable balance between system performance and resource utilization. The performance of the new model outperforms all previous studies in terms of each of the key metrics, which demonstrates the value of combining existing pre-processing methods (e.g., PCA and normalisation), privacy preservation and LSTM-based temporal learning methods for accurate, timely service performance prediction.

The proposed framework has major benefits and implications for the real-world implementation of very large-scale microservices architectures, including providing the ability to monitor systems proactively, detect performance degradation quickly, and ensure improved operational reliability by using intelligent analysis of telemetry data. Implementation of privacy-preserving mechanisms will also make it ideal for use in sensitive cloud environments where data security is of utmost importance. Despite these benefits, several limitations must also be addressed. First, because LSTM models tend to use computationally intensive computations, they may run into challenges with regard to scalability and performance on extremely large or high-frequency telemetry streams. Secondly, although acceptable, monitor overheads may still affect systems that are constrained with respect to available resources; for example, edge systems. Thirdly, the ability of the model to perform is dependent upon having high-quality and diverse training data, which could limit its ability to generalize across differing kinds of infrastructure or workloads that have not already been experienced. In the future research, work may involve lightweight architecture exploring, adaptive learning techniques and cross-domain validation to improve scaling, efficiency and robustness of such frameworks.

7. Conclusion and Future Scope

Telemetry data is very heterogeneous, and it is produced in very large amounts, thus, it is hard to analyze without sophisticated analytical models. To conclude, the paper presents a telemetry model based on AI that predicts service performance in microservice large-scale architecture. The framework utilizes streaming telemetry data and other powerful pre-processing algorithms like cleaning, normalization, and PCA-based feature extraction. The LSTM-based model has the ability to capture the temporal dependencies and identify anomalies in the system behaviour. Moreover, a privacy saving layer will provide safe data processing in a manner that it does not interfere with the analysis. The presented framework has a high predictive accuracy, low false positives and can be detected at a high speed with a satisfactory monitoring overhead. Its superiority in comparison to other current methods has been confirmed, and it is therefore appropriate in monitoring and management of microservices in real time in the cloud native environments. The suggested AI-powered telemetry model has a major implication in the real-life cloud-native system, especially on Kubernetes-based production systems like financial services, e-commerce systems, healthcare systems, and IoT systems. The system allows performance degradation to be caught proactively, reducing downtime, improving service-level agreement (SLA) compliance, and the user experience in applications where latency is a concern. Moreover, the privacy-sensitive telemetry processing is integrated in the framework, which means that it can be applied to industries that work with sensitive information related to operations or users, and it will not violate the data security principles, at the same time being highly accurate in terms of analysis. Future work will focus on enhancing scalability through lightweight deep learning architectures such as transformers, enabling adaptive real-time learning, and improving edge–cloud collaboration for distributed intelligence in microservice ecosystems.

8. References

1. M Parikh AA, Soni SM Shah, AR Jha. (2026) Big Data Workload Profiling for Energy-Aware Cloud Resource Management.
2. VK Sharma. (2025) Cloud Computing & IoT: 5G Focused IoT with Cloud Solutions. Int. J. Artif. Intell. Big Data, Cloud Comput. Data Sci. 6(3): 21-25.
3. MRR Deva. (2025) DevOps and Continuous Delivery Adoption: Trends, Challenges, and Best Practices in Modern Software Development Life Cycle. Int. J. Adv. Res. Comput. Sci. 16(4): 118-124.
4. SR Sirikonda. (2026) Reducing SRE Toil via Safe Autonomous Remediation in Cloud-Native Systems. Am. J. Technol. 5(3): 30-49.
5. G Maddali. (2025) An Efficient Bio-Inspired Optimization Framework for Scalable Task Scheduling in Cloud Computing Environments. Int. J. Curr. Eng. Technol. 15(3): 229-238.
6. A Parupalli, H Kali. (2023) An In-Depth Review of Cost Optimization Tactics in Multi-Cloud Frameworks. Int. J. Adv. Res. Sci. Commun. Technol. 3(5): 1043-1052.
7. AK Padhy, C Medicherla, B Vulugundam, et al. (2025) Latency-Optimized Microservices Orchestration for Real-Time E-Commerce in Multi-Cloud Environments. In: 2025 International Conference on Computer and Applications (ICCA). 1-6.
8. TP Patel. (2026) Adaptive Token Routing for Heterogeneous LLM Inference in Edge-Cloud Continuum. In: SoutheastCon. 1-7.
9. MRC Mukkolakkal. (2025) InfraLLM: A Generic Large Language Model Framework for Production-

Grade Microservice Auto-Scaling in Cloud Infrastructure. Int. J. Sci. Res. Mod. Technol. 4(11): 113-123.
10. HP Cyril. S Kumara. (2026) DevSecOps-Driven Security Integration in the Software Development Lifecycle Using CI/CD Pipelines. In: 2026 IEEE 5th International Conference on AI in Cybersecurity (ICAIC), IEEE. 1-6.
11. R SINHA, R Gulati, A Muttayane, et al. (2022) Network latency time measurement using DNS and web server messages. US11245606B1.
12. SK Chintagunta. (2023) Survey of Containerization, Orchestration, and CI/CD Integration on DevOps in Modern Software Development. Int. J. Curr. Eng. Technol. 13(6): 610-618.
13. SK Davuluri, V Challagulla, V Mudapaka, et al. (2025) AI-Driven DevOps in Telecommunications: Bridging Predictive Analytics with Continuous Delivery for Network Agility. In: 2025 IEEE International Conference and Expo on Real Time Communications at IIT (RTC), IEEE. 1-4.
14. V Ramamoorthi. (2024) AI-Enhanced Performance Optimization for Microservice-Based Systems. J. Adv. Comput. Syst. 4: 17.
15. KK Mohammed. (2025) The Future is Cloud: Modernizing Big Data for the Cloud Era. Int. J. Sci. Res. Eng. Trends. 11(5): 1-5.
16. JB Mehta. (2026) Predictive Quality Engineering in Distributed Data Platforms Using Machine Learning. In: 2026 IEEE International Systems Conference (SysCon). 1-6.
17. JE Kofi. (2025) Data-Driven Cloud Workload Optimization Using Machine Learning Modeling for Proactive Resource Management. Int. J. Artif. Intell. Data Sci. Mach. Learn. 6(4): 27-37.
18. VK Bollu. (2026) Threat Landscape in Artificial Intelligence Systems: Taxonomy, Attack Vectors and Security Implications. World J. Adv. Res. Rev. 29(1): 285-294.
19. V Methuku, S Kamatala, P Naayini, et al. (2022) From Ethical Principles to Technical Safeguards: A Unified Framework for Safe and Human-Centered Artificial Intelligence. Am. Int. J. Comput. Sci. Technol. 4(5): 26-34.
20. S Kilaru. (2013) Automated ETL Intelligence: Metadata-Orchestrated Framework with Rule-Based Heuristics for Monitoring and Reporting. Int. J. Inf. Electron. Eng. 3(6): 14.
21. RN Rajendran, DK Rai, SK Anumula, et al. (2025) Zero Trust Security Model Implementation in Microservices Architectures Using Identity Federation. In: 2025 2nd International Conference on Recent Trends in Electrical, Electronics and Computing Technologies (ICRTEECT). 1-6.
22. S Singamsetty. (2025) An Intelligent Framework for Secure and Fair Cloud Resource Distribution. In: 2025 7th International Conference on Innovative Data Communication Technologies and Application (ICIDCA), IEEE. 686-690.
23. B Krishnan, S Perla, S Maddela, et al. (2025) Adaptive Multi-Cloud Infrastructure for CRM Analytics: Real-Time ML and Data Sync with LLMs. In: 2025 IEEE 3rd Global Conference on Wireless Computing and Networking (GCWCN), IEEE. 1-8.
24. MA Obied, FFM. Ghaleb, AE Hassanien, et al. (2023) Deep Clustering-Based Anomaly Detection and Health Monitoring for Satellite Telemetry. Big Data Cogn. Comput. 7(1): 39.
25. A Katangoori. (2026) The Role of Big Data in Advancing Artificial Intelligence: Methods and Case Studies. Int. J. Artif. Intell. Mach. Learn. 6(1): 37-54.
26. ASO, et al. (2026) Ai-Driven Root Cause Analysis Framework for Distributed Microservices Architectures. Multidiscip. J. Eng. Technol. Sci. 3(1): 8-17.
27. MK Gaddam. (2025) Architecting Observability for AI-Driven Microservices at Scale. In: 2025 3rd International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI), IEEE. 1830-1838.
28. G Dkmak, B Can, O Sevinc, et al. (2025) AI-Driven Anomaly Detection in Cloud-Native Microservices: The Night's Watch Algorithm. Appl. Sci. 15(23): 12762.
29. B Barua, MS Kaiser. (2024) Microservices-Based Framework for Predictive Analytics and Real-time Performance Enhancement in Travel Reservation Systems. arXiv. 1-10.
30. DK Pentyala. (2024) Artificial Intelligence for Fault Detection in Cloud-Optimized Data Engineering Systems. Int. J. Soc. Trends. 2(4): 147-160.
31. GP Menaud. (2023) Machine Learning-Based Performance Prediction and Optimization in Microservices-Oriented Artificial Intelligence Systems. Int. J. Mach. Intell. 1(5): 1-6.
32. SA Pushkala. (2026) Financial Fraud Identification Using Graph Neural Network And LSTM With Autoencoder-Based Data Refinement. J. Int. Cris. Risk Commun. Res. 9(1): 198-213.
33. T Ali, R Iqbal, NM Ansari, et al. (2025) Ai-Powered Anomaly Detection in Software Logs: A Machine Learning Approach for Proactive Fault Diagnosis and Self-Healing Systems. Spectr. Eng. Sci. 3(3): 302-322.
34. R Malaiyalan. (2026) AI/ML-Driven Microservices Architecture for Scalable Cloud Computing Applications. World J. Multidiscip. Stud. 3(3): 28-35.
35. H Ge, Xin Ji, Fang Peng, et al. (2024) SRdetector: Sequence Reconstruction Method for Microservice Anomaly Detection. Electronics.14(1): 65.