Universal Library of Engineering Technology Research Article

ISSN: 3064-996X

Open Access | PP: 43-52

DOI: https://doi.org/10.70315/uloap.ulete.2022.006

A Deep-Review based on Predictive Machine Learning Models in Cloud Frameworks for the Performance Management

Sunil Jacob Enokkaren[1], Avinash Attipalli[2], Varun Bitkuri[3], Raghuvaran Kendyala[4], Jagan Kurma[5], Jaya Vardhani Mamidala[6]

[1]ADP, Solution Architect.

[2]University of Bridgeport, Department of Computer Science.

[3]Stratford University ,Software Engineer.

[4]University of Illinois at Springfield, Department of Computer Science.

[5]Christian Brothers University, Computer Information Systems.

[6]University of Central Missouri, Department of Computer Science.

Abstract

Cloud computing (CC) has increasingly become a critical part of modern-day digital infrastructures that offer dynamic and flexible resources to suit various applications. However, it is all complicated by the necessity to maintain consistency in the ongoing cloud environments. Predictive Performance Management (PPM) aims to identify problematic performance issues early and correct them before they compromise the stability of the systems, as well as the user experience (UX). In the following paper, the author offers a comprehensive review of such approaches to PPM on Cruise Control as Machine Learning (ML) and Artificial Intelligence (AI). It explores the way traditional reactive types of monitoring have been replaced by intelligent predictive frameworks which use real-time information as well as automated decision making. Key parts that are studied are acquisition of performance data, analysis of metrics, models of forecasting and adaptive control mechanisms. The research classifies ML techniques according to their use in workload prediction, anomaly detection, and resource optimization (RO), and then goes on to describe their roles in unsupervised learning (UL), semi-supervised learning (SSL), reinforcement learning (RL), and supervised learning (SL). A discussion is held regarding the predictive task effectiveness of commonly used algorithms, such as decision trees (DT), ensemble approaches (EA), regression models (RM), support vector machines (SVM), and deep learning networks (DLN). The report also highlights the most significant obstacles to using AI/ML for CC performance management and suggests avenues for further study to develop intelligent, predictive approaches that can make CC infrastructures more robust and capable of self-optimization (SO).

Keywords: *Machine Learning, AI, Predictive Performance Management, Cloud Systems.*

INTRODUCTION

Cloud computing has revolutionised the way computing resources are accessible, managed, and provided in today's interconnected world. Many different kinds of applications, from simple web hosting to complicated enterprise systems, favour this infrastructure because of its on-demand service, scalability, and budget-friendly pricing [1]. The growing complexity and variability of cloud environments pose a significant challenge to service providers and users in achieving optimal performance [2]. Workloads are effectively mobile, resource requirements too can vary unpredictably and infrastructure components frequently disperse within many geographic regions and service planes (IaaS, PaaS, SaaS).

Predictive performance management overcomes this challenge by identification of most likely performance problems before they arise therefore allowing proactive management measures to be taken [3]. It entails: tracking system metrics, predicting resource utilisation, anticipating bottlenecks and compliance with service-level agreements (SLAs) [4]. Proper prediction influences the reliability, resources and expenses applied in operation, and user satisfaction positively.

Predictive performance management in cloud systems is now impossible without AIs and ML [5]. These methods make predictions about future performance by sifting through mountains of data, both historical and real-time, in search of patterns and outliers. Due to their unique adaptability and learning ability, ML models are effectively used in the dynamic cloud environment [6]. The methodologies currently deployed and studied in this context include DL, RL, SL, and unsupervised learning, only to name a few. To facilitate the predictive performance management, this survey will group some of the AI and ML techniques, explore their strengths and weaknesses and analyze how they can be applied in practice [7]. The synthesis of the current advances

helps the paper to give a structured description of the field and defines major areas that need to be unveiled in future researches and advancements.

Structure of the Paper

The structure of this document is as follows In Section II, they covered the groundwork for managing cloud performance predictions. Part III Visit explore AI and machine learning in predictive management. IV AI and Machine Learning Techniques of Performance Prediction. The fifth section (V) examines associated literature and case studies, and the last section (VI) comes up with future research directions.

FUNDAMENTALS OF CLOUD PERFORMANCE MANAGEMENT

The application of AI and ML to cloud performance management is therefore predictive Performance Management (PPM). Project Portfolio Management (PPM) allows for efficient and timely resource management in ever-changing cloud settings by predicting future problems, including bottlenecks or failures, in contrast to reactive solutions [8]. This approach enhances decision-making for resource provisioning, fault tolerance, and overall system reliability. Preventing equipment failure and keeping it running smoothly are two of the most important goals of maintenance in industrial operations. Data analytics and ML backup cloud-enabled maintenance systems that facilitate predictive approaches to assisting businesses to increase efficiency and asset uptime at low costs of operation.

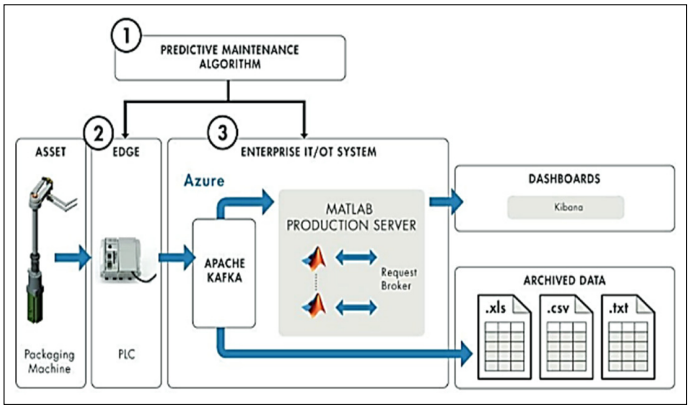

Fig 1. Working of Predictive Maintenance System

Predictive maintenance represents a groundbreaking shift in this paradigm, leveraging cloud technology to analyze data and utilize ML algorithms for predicting equipment failures, thereby streamlining maintenance activities and enhancing overall operational efficiency (as shown in Figure 1). Cloud-based predictive maintenance has the potential not only to save companies substantial sums in maintenance costs but also to substantially enhance equipment reliability. Working of the predictive Maintenance system.

Evolution of Performance Management Systems

The legacy performance management systems were usually centred upon static and backwards-looking forms of assessment with emphasis placed upon financially driven and lagging measures [9]. These systems were based so much on periodic manual reporting, top-down goal-setting, and subjective appraisals. Performance reviews were commonly conducted annually or quarterly, which limited the organization's ability to respond to dynamic market conditions [10]. Key performance indicators (KPIs) in traditional models often lacked alignment with real-time operational data, thereby impeding responsiveness and agility.

In contrast, modern performance management frameworks emphasize continuous feedback, agile goal-setting, and the integration of real-time data. These systems are driven by dashboards, automated metrics, and analytics tools that support forward-looking assessments [11]. KPIs have become more granular, cross-functional, and aligned with strategic priorities, including innovation, customer engagement, and sustainability. Organizations now track a combination of financial and non-financial indicators, including employee engagement, customer satisfaction, and operational efficiency

Key Components in AI-Based Predictive Maintenance

AI-based predictive maintenance systems in cloud environments comprise several interdependent components that collectively enable accurate forecasting, timely maintenance decisions, and optimized resource utilization. These components span the data collection, processing, modelling, and deployment layers, each playing a crucial role in the end-to-end predictive maintenance pipeline [12]. An AI-based Pd.M. consists of six primary components, as illustrated in Figure 2: data pre-processing, AI algorithms, decision-making modules, communication and integration, user interface and reporting.

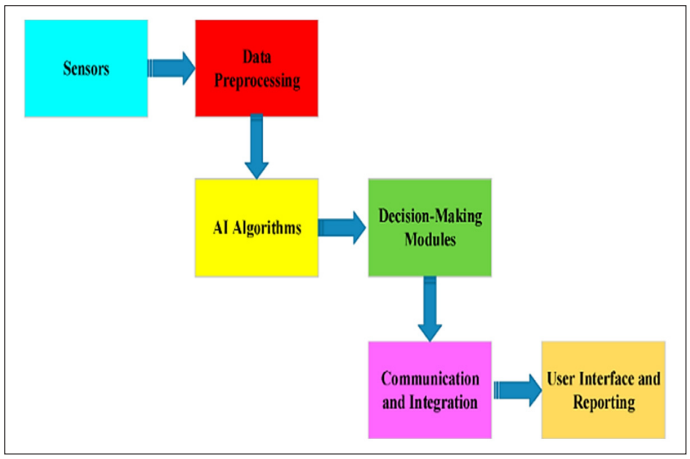

Fig 2. Key Components of An AI-Based Pd.M. System.

The following are the primary parts of predictive maintenance:

Sensors

A Pd.M. system's sensors are its primary data collectors. Various factors, including temperature, pressure, vibration,

and others, are constantly monitored by these specialised devices that are strategically installed on equipment and machinery.

Data Pre-Processing

Noise and irregularities are common in sensor raw data. In order to get data ready for analysis, the first step is data pre-processing. Data normalisation, cleansing, and addressing missing data are all part of it. Precise modelling relies on high-quality data.

AI Algorithms

Computer programs that use artificial intelligence, such as DL and ML. In order to determine which data aspects are most indicative of possible problems, the algorithms sift through the information. They are able to anticipate equipment failures, outliers, and RUL by learning from past data.

Decision-Making Modules

The modules responsible for making decisions examine the predictions and insights generated by the AI systems. Identifying the need for maintenance is the responsibility of these modules. They are capable of making maintenance job recommendations (both preventative and corrective), creating maintenance schedules, and, when required, alerting maintenance crews.

Communication and Integration

The successful translation of the system's findings into action depends on communication and integration. Maintenance workers and upper management are among the many parties involved in this component's interactions.

User Interface and Reporting

Powerful reporting and user interfaces are necessary to make these insights available to decision makers and maintenance workers. The data visualisation, dashboard, and reporting features offered by the products facilitate users' comprehension of intricate data patterns, allowing them to make well-informed decisions. Data visualisation tools and dashboards are a great way for decision-makers and maintenance teams to understand data insights and get forecast information.

Performance Metrics in Cloud Environments

Performance indicators are very important for figuring out how efficient and effective different computing models are in the cloud, especially for Internet of Things (IoT) apps. The four cornerstones of any functional framework are The Internet of Things, Cloud, Fog, and Edge Computing. Each layer brings new problems and effects to measuring performance [13]. IoT device availability for offloaded activity management is critical in a mist computing model, for instance. At every level, from the edge to the fog, availability and response time are crucial; yet, the relative weight of these two metrics could change based on the design.

Computing Model

Fig. 1 depicts a generic representation of the cloud-to-things landscape. Academics and businesses have proposed a four-tiered architecture for IoT applications in the real world. These levels are the IoT itself, edge computing, fog computing, and cloud computing [14]. The suggested model illustrates the interaction between the landscapes and the orchestration process, which can be optimally optimised with a MAPE-K loop. The MAPE-K loop can be applied independently to each layer in this general model.

Taxonomy

The computer model or layer that controls the Internet of Things application greatly affects the performance measures. The availability of Internet of Things devices to carry out offloaded tasks is guaranteed in a cloud model or layer, but it's a major concern in a mist model. When it comes to cloud, edge, and fog computing, availability and response time are two of the most important factors. The weight that is given to each performance statistic, however, varies. In addition, as future computing paradigms like fog and edge computing follow cloud computing in some ways, they might have comparable performance measurements.

AI AND MACHINE LEARNING TECHNIQUES IN PREDICTIVE PERFORMANCE MANAGEMENT

ML has completely changed the game when it comes to predictive maintenance. It does this by sifting through massive datasets in search of irregularities that could be signs of impending equipment failure. Using a cloud environment, it analyses large-scale operational data, i.e., logs, metrics, and telemetry, to identify any slight decline or potential issue. Such algorithms are able to learn any previous failures to be able to tell what may go wrong then correct it before it actually happens. To further enhance the quality of forecasts, the ML algorithms also have the opportunity to adapt to the changes within the cloud. Predictive maintenance cloud infrastructure through ML is an excellent idea as cloud environments are a mixture of various environments and changing all the time [15]. A cloud system consists of various interconnected hardware, software, and services [8]. DL and RL are two ML methods that can capture such complex relations and identify the issues within a system magnificently. Moreover, ML enhances resource optimising to achieve a lot with minimal disruptions [16].

Types of Machine Learning Paradigms

In ML, RL, SL, unsupervised learning, and semi-supervised learning are the four primary categories of models. In supervised learning, models are taught to make predictions by using labelled data. When it comes to classification and regression, this strategy is absolutely crucial. Clustering and dimensionality reduction are examples of unsupervised learning techniques that use unlabelled data to discover hidden patterns and groups. Through the use of a combination

of labelled and unlabelled data, semi-supervised learning is able to increase the accuracy of learning [17]. RL lets agents learn the best way to deal with an environment by making mistakes and trying again. It is commonly used in control and decision-making systems. Figure 3 shows how these ML models relate to each other and where they can be used:

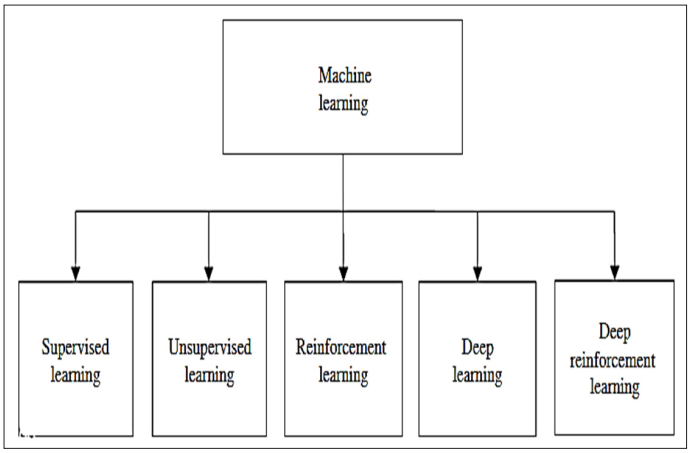

Fig 3. Machine Learning Paradigms

The types of ML models that follow are listed below.

Supervised Learning

The process of SL involves training the machine with the use of labelled data. Accompanying the tagged data is a supervisory role. The model is taught by using both inputs and outputs. Following that, the model would make predictions for more data points. All sorts of things can be classified, including speech recognition, handwriting recognition, documents, biometric identity, and a plethora of others. The functioning of a supervised learning model can be shown in Figure 4.

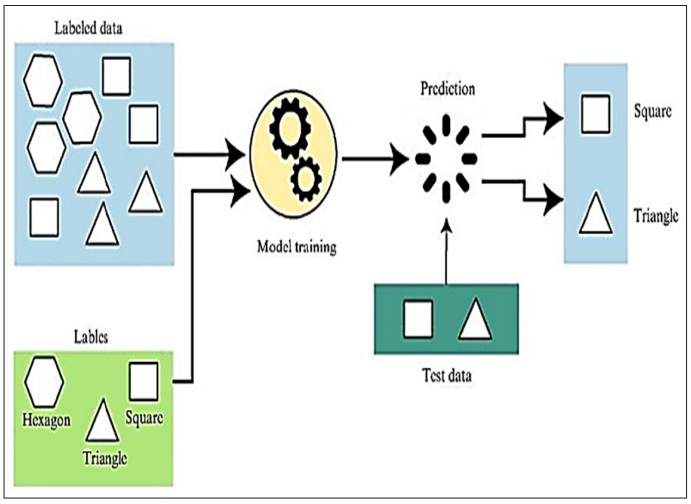

Fig 4. Working Model of Supervised Learning

In cloud systems, SL is a popular ML technique for problems involving predicted performance. It requires knowing the outcome variable beforehand and training models using labelled datasets. Common performance measures that are predicted using algorithms like Random Forests, Support Vector Machines (SVM), and linear regression include system throughput, CPU utilisation, and reaction time.

Unsupervised Learning

Machines engage in unsupervised learning when they are taught to process input that does not contain labels. There is no requirement for a supervisor in unsupervised learning. When left to their own devices, models can learn and uncover hidden patterns and data. These algorithms are able to uncover previously unseen data structures or patterns without knowing the results in advance. K-Means and DBSCAN are two examples of clustering algorithms that can help with workload classification and anomaly detection by grouping comparable performance profiles.

Reinforcement Learning

ML learning techniques like RL allow agents to learn how to interact with their surroundings by doing actions and then analysing the results. As a means of learning its surroundings and producing an outcome, the agent relies on trial and error. A reinforcement learning system primarily consists of two parts: the environment and agents. See Figure 5 for a model of reinforcement learning in action.

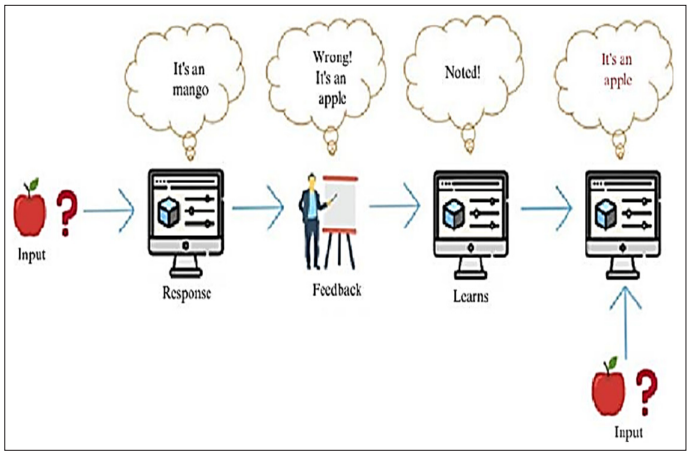

Fig 5. Working Model of Reinforcement Learning

A strong paradigm for adaptive and dynamic decision-making in cloud systems is RL. By interacting with their environment and learning from the consequences (rewards or penalties), agents in RL learn to make optimal decisions.

Deep learning

One area where DL models are finding more and more use is in the analysis of time-series and multivariate performance data. RNNs, LSTMs, and CNNs are among the most used DL models in this context. These models can capture complex temporal dependencies and nonlinear relationships within large datasets, making them highly effective for predicting trends, identifying long-term patterns, and understanding the dynamics of systems [18].

Techniques Used in Prediction Performance-Based AI/ML

ML techniques can be one of the tools used when predicting the success of any employees in their work. These methods allow systems to learn based on the trends of data and make

predictions without explicitly programmed information as the methods are better than standard analytics. As an example, fed past data performance, supervised learning models, such as random forests and neural networks could be used to predict what is next, thus having a more advanced sense of the factors that affect the output on an individual and team level. With the help of the correlations and trends in data, the ML strategies were useful in identifying the performance of the employees in their jobs. In this regard, the following ML techniques find an application in this region:

Linear Regression

Predicting a continuous result, like employee performance ratings, is possible with the use of the SL technique known as linear regression. In order to determine the target components, it uses the input features to create a linear relationship with them. Using training data from prior scenario, linear regression can give scenario-specific predictions based on feature values.

Random Forests

One of the forms of an ensemble learning is called RF, and it constructs and enhances the projections of a set of decision trees. The analysis of a great number of features can lead RF to develop a comprehensive model for predicting the performance of employees with a minimum chance of overfitting. The predictive capability and practicality of the model are both enhanced by the ensemble method.

Artificial Neural Networks

ANNs are ML models modelled on brain neural structure, with input-processed neurons routed to outputs interconnected and generating outputs. Here, the ANNs are sufficiently trained to give an output that can correspond to an input and its training is such that it can be highly accurate and so it is trained with the Levenberg-Marquardt with back propagation algorithm.

Decision Tree

A DT is a supervised ML algorithm which has a tree-like representation to show decisions and possible outcomes given some attributes [19]. It begins with a root node, followed by internal nodes that test attributes, and concludes with leaf nodes representing the final outcomes. Each branch reflects a decision path derived from the dataset.

Support Vector Machine

Support Vector Machines (SVMs) are supervised methods that use Vapnik's theory of learning to classify data, perform regression, and identify outliers. A kernel function is employed to transform the input data into a high-dimensional feature space. From there, it builds an ideal hyperplane to divide classes. Strong generalisation is achieved by SVM without relying on domain-specific information because this hyperplane is defined only by critical data points, called support vectors.

K-Nearest Neighbours

KNN is an easy-to-understand ML technique that uses the consensus of nearby objects to determine their classification. The object is being placed in the most popular class among its immediate neighbours. Typically, K is a tiny positive number [20]. The class of the object's nearest neighbour is used if k is equal to 1. In binary (two-class) classification problems, using an odd number of integers as the value of k allows to remove of tied votes. Choosing the k-parameter is the most critical step in this procedure.

APPLICATIONS OF AI/ML IN CLOUD PERFORMANCE MANAGEMENT

ML and AI have become significant components of the process of making computer systems more efficient, reliable, and flexible. This allows them to approach performance in a proactive way, reduce downtime, help use resources to the best advantage and ensure that Service Level Agreements (SLAs) are complied with; they can learn from past data, recognize patterns and also be able to make predictions. In this section, the author discusses the most significant applications of AI and ML in cloud performance management.

Workload Prediction

Predicting workload in the cloud environment is a key aspect in effective resource when it comes to efficient planning of resources in cloud environment since it can make systems proactively manage the workload using the computational resources available: through workload prediction. The cloud platforms are capable of predicting the user traffic to web applications reliably with the historical usage patterns using ML models specifically time-series based models including LSTM, GRU, XGBoost, and SARIMA [21]. These proactive forethoughts allow proactive changes to be made in resource distribution, cutting down on the chances of over-provisioning at low usage times and under-provisioning at high utilization times. The consequence is that workload prediction improves system elasticity, cost efficiency and the process of user experience is smoother.

Resource Utilization Forecasting

The future mapping of the use of CPU, memory, disc, and network may be perfectly copied or anticipated with the help of AI and ML models. Being able to tune and schedule resources before job requirement helps cloud environments. Regression models, neural networks (ANN, RNN) also examine past and real-time data to make a conjecture of the required resource amount [22]. Such models assist in monitoring the crucial metrics, such as CPU load, memory consumption, and I/O operations per second so that the performance bottlenecks could be avoided, the risk of SLA violation could be reduced, and the overall responsiveness and operational efficiency of the system could be enhanced.

Auto-Scaling and Load Balancing

AI-based auto-scaling allows dynamic scaling of virtual

machines (VMs) or containers with regard to real-time and projected workload patterns to guarantee efficient use of resources. ML also improves the process of load balancing because, it is a smart way of distributing all the incoming traffic system based on the load in the system, and the system's prediction concerning the traffic demand [23]. Some of the common ways used in optimization of such processes are reinforcement learning and decision trees. Techniques such as implementations onto platforms such as Kubernetes where the use of custom auto scalers can be conducted or deployments on AWS Auto Scaling with predictive policies can reduce latency, reduce operational cost overheads and retain high levels of system reliability and performance.

SLA Violation Prediction

The assumption of proactive ML models is the ability to predict the probability of violations of SLA before they happen through well-processed system telemetry and usage representations, making a possibility to intervene early before performance degradation affects the outcome. With major characteristics including latency and response time, as well as success rate (of request), models such as RF, LR, and SVM are capable of predicting future attempts of breach successfully. This predictive ability helps keep the customers contented, reduces the financial fine and also enables SLA-aware orchestration in the cloud settings.

Energy-Aware and Cost-Effective Scheduling

Long-term in cloud computing Energy-efficient and cost-efficient schedule is significant since it addresses environmental concerns and the increase in energy prices. Smart schedulers can determine how much energy a job will consume using methods of AI/ML such as reinforcement learning, genetic algorithms, and multi-objective optimisation and how to execute them in the most efficient way, with the least possible resource consumption [24]. By leveraging historical and real-time data, these models dynamically schedule workloads, such as deferring non-critical tasks to off-peak periods when renewable energy is available. This approach reduces carbon footprints, enhances energy efficiency, supports green SLAs, and promotes eco-friendly, cost-effective cloud operations.

LITERATURE REVIEW

Existing work on AI and ML in cloud systems has primarily focused on predictive performance management, there is an emphasis on integrating DL models for improved prediction accuracy and developing frameworks for predictive maintenance to enhance operational efficiency.

Panicucci et al. (2020) describe an innovative approach to estimating the remaining useful life (RUL), which permits predictive maintenance of industrial equipment by utilising partial knowledge of its degradation function and the parameters that impact it. In addition, the previously mentioned idea is integrated into the design and prototype implementation of a fully functional, end-to-end cloud architecture that supports predictive maintenance of industrial equipment. With the architecture being plug-and-play. This is done by a number of apps integrated together such as scheduling, data visualisation, predictive analytics, and data repositories on the edge and in the cloud. The proposed method has been implemented in a real-life situation where it is applied in the robotic arm maintenance process. The results obtained in this era of the fourth industrial revolution reveal that the proposed approach is efficient and effective towards facilitating predictive analytics [25].

Andreazi et al. (2020) the MoHRiPA allocation module, which the architects of private cloud architectures created to achieve peak performance while allocating virtual machines. Its architecture and specifications of components shall be the focus of this presentation. They will revolve around the resource allocation module, which possesses hybrid nature, and is capable of hosting diverse virtualisation tools. They introduce the findings based on a multi-factor planning that enables us to better evaluate the specifications and qualities of the allocation module of MoHRiPA. Bring into the light MoHRiPA allocation module which was intended to provide high-quality performance at the time of allocating virtual machines in Architecture Management of Hybrid Resources in Private Cloud Architektur [26].

Aslanpour, Gill and Toosi (2020b) thorough literature study, provide a taxonomy of real-world metrics for assessing the performance of cloud, fog, and edge computing, survey the literature to find common metrics and their uses, and (3) point out areas that need more investigation. The results of this comprehensive benchmark study might considerably improve the chances of researchers' and developers' success in achieving their objectives in real-world production environments by providing them with useful metrics and standards for effective performance evaluation.. An integral aspect of cloud computing, optimisation is especially important with the rise of fog and edge computing [11].

Fila, Khaili and Mestari (2020) suggest a fresh method of prediction that showcases the Prognosis as a Service, built around the multitenancy principle and the Cloud Computing model. In response to a client's request, this method improves service quality while providing an effective prognostic solution. Performance metrics for the prediction system, including accuracy, precision, mean squared error, and Quality of Service, define the solution's efficacy (Qos). The fundamental pillar of predictive maintenance is known as Health Management and Prognostics (PHM) [27].

Fargo et al. (2019) demonstrate a system for autonomously managing power consumption and performance. The suggested method begins by gathering information about the environment in relation to power usage and the deployment of security technologies. Then, it uses the system's behaviour to determine which security technologies to deploy and how

to provision virtual resources to lower power consumption without sacrificing performance. In high-performance computing (HPC), cloud computing (CC), server, embedded, and other systems, this method allows for the efficient use of a variety of secure resources [28].

Jodayree, Abaza and Tan (2019) introduces Cicada, an end-to-end system that uses rule-based task balancing algorithms to make predictions. The approach will achieve lower computational demand and faster workload balancing by simulating cloud services using Cloud Sim, a cloud service simulator. The results will demonstrate the efficacy of a predictive workload balancing approach, which can reduce computer power consumption while increasing speed [29].

A comparative analysis of the background study, including its author(s), approaches, key findings, limitations, and future work, is provided in Table I:

Table I. Summary of Recent Studies on for Predictive Management in AI and Machine Learning

Reference	Study Focus	Approach	Key Findings	Challenges	Future Directions
Panicucci et al. (2020)	Predictive maintenance using RUL estimation	Prototype tested on robotic arm; end-to-end plug-and-play cloud architecture merged with partial knowledge-based RUL prediction	To back up Industry 4.0 predictive analytics effectively and efficiently	Real-time data integration and model adaptability	Extend RUL methodology to other industrial use cases; enhance real-time performance and scalability
Andreazi et al. (2020)	Virtual machine (VM) allocation in hybrid private clouds	MoHRiPA architecture with a hybrid resource allocation module supporting different virtualization tools; factorial analysis conducted	High-performance VM allocation; flexibility in supporting multiple virtualization tools	Managing complexity of hybrid environments and resource scheduling	Improve modularity and integration with other cloud orchestration tools
Aslanpour, et.al (2020b)	The cloud, fog, and edge: performance metrics and benchmarks	Classification of practical measures; review of relevant research; determination of typical assessment variables	Comprehensive classification of performance metrics; benchmark for realistic system evaluation	Lack of unified standards; inconsistent metric usage across domains	Develop standardized benchmarking suites for hybrid cloud-fog-edge environments
Fila, et.al. (2020)	Prognosis as a Service (PaaS) for predictive maintenance using cloud and multitenancy	Created a multitenant cloud-based prognostic framework to provide PHM as a service provider	To prove improved precision and QoS, performance metrics including accuracy and mean squared error (MSE) were employed..	Handling variability in tenant requirements and maintaining consistent QoS across clients	Enhance scalability, tenant isolation, and real-time performance; integrate advanced ML/AI models for better prognostic accuracy
Fargo et al. (2019)	Power and security in computer systems through autonomous management	Environment-aware provisioning of resources and deployment of security tools based on behaviour	Efficient utilization of secure resources with reduced power consumption	Dynamic adaptation of resources to changing workloads and security needs	Apply to more heterogeneous environments like IoT and real-time systems
Jodayree, et.al. (2019)	Optimising cloud workloads using predictive models	Rule-based workload-balancing algorithm using Cicada system and simulated via CloudSim	Achieved faster workload balancing with reduced computational overhead	Limited accuracy and adaptability in dynamic workloads	Enhance prediction accuracy; real-world validation with diverse workloads

CONCLUSION AND FUTURE SCOPE

This survey reviewed the current landscape of AI and ML techniques for Predictive Performance Management in cloud systems. By shifting from reactive to proactive strategies, these intelligent approaches enable cloud providers to anticipate performance degradation, optimize resource allocation, and ensure service-level objectives, categorized ML methods into various learning paradigms and discussing their application in key areas such as workload forecasting, anomaly detection, and system adaptation. Scalability, data heterogeneity, model interpretability, and real-time processing are some of the obstacles that current systems encounter, despite their promise. The ongoing integration of cloud computing with cutting-edge AI, such as hybrid models, DL, and reinforcement learning, offers great promise for creating autonomous, robust cloud systems.

Research on cloud-based PPM should go towards creating AI models that are both flexible and scalable so that they can make good use of heterogeneous, multi-cloud setups. One important step is to use edge AI and federated learning to make decentralised performance prediction possible while protecting privacy and not letting data leak.

REFERENCES

1. S. Gupta and S. Prakash, "QoS and load balancing in cloud computing-an access for performance enhancement using agent-based software," Int. J. Innov. Technol. Explor. Eng., vol. 8, no. 11, pp. 641–644, 2019.
2. W. Venters and E. A. Whitley, "A critical review of cloud computing: Researching desires and realities," J. Inf. Technol., vol. 27, no. 3, pp. 179–197, 2012, doi: 10.1057/jit.2012.17.
3. P. Ganesh, D. Evangelingeetha, and T. V. S. Kumar, "Prediction of Cloud Application Performance using SMTQA Tool," SSRG Int. J. Comput. Sci. Eng., vol. 3, no. 11, pp. 27–33, 2016.
4. B. Kang, D. Kim, and S.-H. Kang, "Periodic Performance Prediction for Real-time Business Process Monitoring.," Ind. Manag. Data Syst., vol. 112, 2011, doi: 10.1108/02635571211193617.
5. S. Ray, "A Quick Review of Machine Learning Algorithms," in 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), IEEE, Feb. 2019, pp. 35–39. doi: 10.1109/COMITCon.2019.8862451.
6. A. Qayyum et al., "Securing Machine Learning in the Cloud: A Systematic Review of Cloud Machine Learning Security," Front. Big Data, vol. 3, Nov. 2020, doi: 10.3389/fdata 2020.587139.
7. Y. Hu, H. Wang, and W. Ma, "Intelligent cloud workflow management and scheduling method for big data applications," J. Cloud Comput., vol. 9, no. 1, p. 39, Dec. 2020, doi: 10.1186/s13677-020-00177-8.
8. N. Fareghzadeh, M. A. Seyyedi, and M. Mohsenzadeh, "Dynamic performance isolation management for cloud computing services," J. Supercomput., vol. 74, no. 1, pp. 417–455, Jan. 2018, doi: 10.1007/s11227-017-2135-2.
9. M. Ficco, M. Rak, S. Venticinque, L. Tasquier, and G. Aversano, "Cloud Evaluation: Benchmarking and Monitoring," in Quantitative Assessments of Distributed Systems, Wiley, 2015, pp. 175–199. doi: 10.1002/9781119131151.ch7.
10. D. J. Schleicher, H. M. Baumann, D. W. Sullivan, and J. Yim, "Evaluating the effectiveness of performance management: A 30-year integrative conceptual review," J. Appl. Psychol., vol. 104, no. 7, pp. 851–887, 2019, doi: 10.1037/apl0000368.
11. M. S. Aslanpour, S. S. Gill, and A. N. Toosi, "Performance evaluation metrics for cloud, fog and edge computing: A review, taxonomy, benchmarks and standards for future research," Internet of Things, vol. 12, Dec. 2020, doi: 10.1016/j.iot.2020.100273.
12. O. Motaghare, A. S. Pillai, and K. I. Ramachandran, "Predictive Maintenance Architecture," in 2018 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), IEEE, Dec. 2018, pp. 1–4. doi: 10.1109/ICCIC.2018.8782406.
13. H. Gangwar, "Cloud computing usage and its effect on organizational performance," Hum. Syst. Manag., vol. 36, no. 1, pp. 13–26, Mar. 2017, doi: 10.3233/HSM-171625.
14. K. Hwang, X. Bai, Y. Shi, M. Li, W. G. Chen, and Y. Wu, "Cloud Performance Modelling with Benchmark Evaluation of Elastic Scaling Strategies," IEEE Trans. Parallel Distrib. Syst., vol. 27, no. 1, pp. 130–143, 2016, doi: 10.1109/TPDS.2015.2398438.
15. J. Sahlin and J. Angelis, "Performance management systems: reviewing the rise of dynamics and digitalization," Cogent Bus. Manag., vol. 6, no. 1, Jan. 2019, doi: 10.1080/23311975.2019.1642293.
16. M. Balaji, C. A. Kumar, and G. S. V. R. K. Rao, "Predictive Cloud resource management framework for enterprise workloads," J. King Saud Univ. - Comput. Inf. Sci., vol. 30, no. 3, pp. 404–415, Jul. 2018, doi: 10.1016/j.jksuci.2016.10.005.
17. J. Alzubi, A. Nayyar, and A. Kumar, "Machine Learning from Theory to Algorithms: An Overview," J. Phys. Conf. Ser., vol. 1142, p. 012012, Nov. 2018, doi: 10.1088/1742-6596/1142/1/012012.
18. M. F. Mushtaq, U. Akram, I. Khan, S. Naqeeb, A. Shahzad, and A. Ullah, "Cloud Computing Environment and Security Challenges: A Review," Int. J. Adv. Comput. Sci.

Appl., vol. 8, no. 10, pp. 183–195, 2017, doi: 10.14569/IJACSA.2017.081025.

19. I. D. Mienye, Y. Sun, and Z. Wang, "Prediction performance of improved decision tree-based algorithms: a review," Procedia Manuf., vol. 35, pp. 698–703, 2019, doi: 10.1016/j.promfg.2019.06.011.
20. A. S. Hashim, W. A. Awadh, and A. K. Hamoud, "Student Performance Prediction Model based on Supervised Machine Learning Algorithms," IOP Conf. Ser. Mater. Sci. Eng., vol. 928, no. 3, 2020, doi: 10.1088/1757-899X/928/3/032019.
21. J. Kumar and A. K. Singh, "Workload prediction in cloud using artificial neural network and adaptive differential evolution," Futur. Gener. Comput. Syst., vol. 81, pp. 41–52, Apr. 2018, doi: 10.1016/j.future.2017.10.047.
22. T. Mehmood, S. Latif, and S. Malik, "Prediction of Cloud Computing Resource Utilization," in 2018 15th International Conference on Smart Cities: Improving Quality of Life Using ICT & IoT (HONET-ICT), IEEE, Oct. 2018, pp. 38–42. doi: 10.1109/HONET.2018.8551339.
23. A. Kushwaha, P. Pathak, and S. Gupta, "Review of optimize load balancing algorithms in cloud," Int. J. Distrib. Cloud Comput., vol. 4, no. 2, pp. 1–9, 2016.
24. S. S. S. Neeli, "Serverless Databases: A Cost-Effective and Scalable Solution," Int. J. Innov. Res. Eng. Multidiscip. Phys. Sci., vol. 7, no. 6, p. 7, 2019.
25. S. Panicucci et al., "A Cloud-to-Edge Approach to Support Predictive Analytics in Robotics Industry," Electronics, vol. 9, no. 3, p. 492, Mar. 2020, doi: 10.3390/electronics9030492.
26. G. T. Andreazi, J. C. Estrella, S. M. Bruschi, A. M. A. Ferreira, and W. da Silva Martins, "Performance Evaluation in an Architecture which Instances Hybrid Resources in Private Cloud," in 2020 IEEE Cloud Summit, IEEE, Oct. 2020, pp. 108–113. doi: 10.1109/IEEECloudSummit48914.2020.00023.
27. R. Fila, M. El Khaili, and M. Mestari, "Cloud Computing for Industrial Predictive Maintenance Based on Prognostics and Health Management," Procedia Comput. Sci., vol. 177, pp. 631–638, 2020, doi: 10.1016/j.procs.2020.10.090.
28. F. Fargo, O. Franza, C. Tunc, and S. Hariri, "Autonomic Resource Management for Power, Performance, and Security in Cloud Environment," in 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA), IEEE, Nov. 2019, pp. 1–4. doi: 10.1109/AICCSA47632.2019.9035213.
29. M. Jodayree, M. Abaza, and Q. Tan, "A Predictive Workload Balancing Algorithm in Cloud Services," Procedia Comput. Sci., vol. 159, pp. 902–912, 2019, doi: 10.1016/j.procs.2019.09.250.
30. Polu, A. R., Vattikonda, N., Buddula, D. V. K. R., Narra, B., Patchipulusu, H., & Gupta, A. (2021). Integrating AI-Based Sentiment Analysis With Social Media Data For Enhanced Marketing Insights. Available at SSRN 5266555.
31. Katari, A., & Kalla, D. (2021). Cost Optimization in Cloud-Based Financial Data Lakes: Techniques and Case Studies. ESP Journal of Engineering & Technology Advancements (ESP-JETA), 1(1), 150-157.
32. Polu, A. R., Vattikonda, N., Gupta, A., Patchipulusu, H., Buddula, D. V. K. R., & Narra, B. (2021). Enhancing Marketing Analytics in Online Retailing through Machine Learning Classification Techniques. Available at SSRN 5297803.
33. Polu, A. R., Buddula, D. V. K. R., Narra, B., Gupta, A., Vattikonda, N., & Patchipulusu, H. (2021). Evolution of AI in Software Development and Cybersecurity: Unifying Automation, Innovation, and Protection in the Digital Age. Available at SSRN 5266517.
34. Gupta, K., Varun, G. A. D., Polu, S. D. E., & Sachs, G. Enhancing Marketing Analytics in Online Retailing through Machine Learning Classification Techniques.
35. Kakani, A. B., Nandiraju, S. K. K., Chundru, S. K., Vangala, S. R., Polam, R. M., & Kamarthapu, B. (2021). Big Data and Predictive Analytics for Customer Retention: Exploring the Role of Machine Learning in E-Commerce. International Journal of Emerging Trends in Computer Science and Information Technology, 2(2), 26-34.
36. Polam, R. M., Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., Chundru, S. K., & Vangala, S. R. (2021). Data Security in Cloud Computing: Encryption, Zero Trust, and Homomorphic Encryption. International Journal of Emerging Trends in Computer Science and Information Technology, 2(3), 70-80.
37. Polam, R. M., Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., Chundru, S. K., & Vangala, S. R. (2021). Big Text Data Analysis for Sentiment Classification in Product Reviews Using Advanced Large Language Models. International Journal of AI, BigData, Computational and Management Studies, 2(2), 55-65.
38. Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., Chundru, S. K., Vangala, S. R., & Polam, R. M. (2021). Advanced Machine Learning Models for Detecting and Classifying Financial Fraud in Big Data-Driven. International Journal of Artificial Intelligence, Data Science, and Machine Learning, 2(3), 39-46.
39. Vangala, S. R., Polam, R. M., Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., & Chundru, S. K. (2021). Smart Healthcare: Machine Learning-Based Classification of Epileptic Seizure Disease Using EEG Signal Analysis. International Journal of Emerging Research in Engineering and Technology, 2(3), 61-70.

40. Gangineni, V. N., Pabbineedi, S., Penmetsa, M., Bhumireddy, J. R., Chalasani, R., & Tyagadurgam, M. S. V. (2021). Strengthening Cybersecurity Governance: The Impact of Firewalls on Risk Management. International Journal of AI, BigData, Computational and Management Studies, 2(4), 60-68.

41. Pabbineedi, S., Penmetsa, M., Bhumireddy, J. R., Chalasani, R., Tyagadurgam, M. S. V., & Gangineni, V. N. (2021). An Advanced Machine Learning Models Design for Fraud Identification in Healthcare Insurance. International Journal of Artificial Intelligence, Data Science, and Machine Learning, 2(1), 26-34.

42. Tyagadurgam, M. S. V., Gangineni, V. N., Pabbineedi, S., Penmetsa, M., Bhumireddy, J. R., & Chalasani, R. (2021). Enhancing IoT (Internet of Things) Security Through Intelligent Intrusion Detection Using ML Models. International Journal of Emerging Research in Engineering and Technology, 2(1), 27-36.

43. Penmetsa, M., Bhumireddy, J. R., Chalasani, R., Tyagadurgam, M. S. V., Gangineni, V. N., & Pabbineedi, S. (2021). Next-Generation Cybersecurity: The Role of AI and Quantum Computing in Threat Detection. International Journal of Emerging Trends in Computer Science and Information Technology, 2(4), 54-61.

Citation: Sunil Jacob Enokkaren, Avinash Attipalli, et al., "A Deep-Review based on Predictive Machine Learning Models in Cloud Frameworks for the Performance Management", Universal Library of Engineering Technology, 2022; 43-52. DOI: https://doi.org/10.70315/uloap.ulete.2022.006.