Resource Utilization Prediction Model for Cloud Datacentre: Survey

Doaa Bliedy[1]*, Mohamed H. Khafagy[2], Rasha M. Badry[3]
Department of Information System-Faculty of Computers and Artificial Intelligence, Fayoum University, Egypt[1, 3]
Department of Computer Science-Faculty of Computers and Artificial Intelligence, Fayoum University, Egypt[2]

***Abstract*—This survey aims to analyze resource prediction models in cloud environments to improve resource allocation strategies. It can be difficult for cloud service providers to maintain the required Quality of Service (QoS) requirements without going against a service level agreement (SLA). Improving cloud performance requires accurate workload prediction. To enhance customer service quality (QoS), cloud computing provides virtualisation, scalability, and on-demand services. Resource provisioning is a major challenge in the cloud environment due to its dynamic nature and the rapid increase in resource demand. Over-provisioning of resources leads to energy waste and increased expenses while under-provisioning can result in SLA breaches and reduced QoS. It is crucial to allocate resources as closely as possible to current demands. Cloud elasticity plays a key role in adapting to workload changes and maintaining performance levels. Predicting future resource demand is essential for effective resource allocation, which is the focus of this survey. Our survey uniquely focuses on comparing univariate and multivariate input cases for cloud resource prediction, a perspective that has not been deeply explored in similar surveys. Unlike existing works that primarily categorize models by methodologies or application characteristics, our study offers a novel analysis of how different input scenarios impact prediction accuracy, resource efficiency, and scalability. By addressing this overlooked aspect, our survey provides unique insights and practical guidance for researchers and practitioners aiming to optimize resource utilization in cloud environments. A thorough analysis of resource prediction models in cloud systems is presented in this research, including a comparison of predicted resources, prediction algorithms, datasets, performance metrics, a prediction summary, and a taxonomy of prediction methods. This survey not only synthesizes current knowledge but also identifies key gaps and future directions for the development of more robust and efficient resource prediction models.**

Keywords*—*Cloud computing; resource utilization; prediction; cloud datacenter; machine learning models; resource allocation

I. INTRODUCTION

Cloud computing is a computer paradigm that provides pay-as-you-go services, such as platforms, apps, and infrastructure [1, 2]. Elasticity is one of the main features of cloud computing [3]. It is the extent to which resources may be autonomously allocated and relocated to satisfy demands at any given time in response to variations in workload [4]. As a result, resources are distributed or released based on the required needs. The cloud must distribute a reasonable number of resources to fulfill its duties [41-44]. Under-provisioning results in SLA violations, declining Quality of Service (QoS), and aggravation for the client. This can result in a decline in revenue and a loss of clients. In contrast, over-provisioning wastes resources and money while raising network, cooling, and maintenance costs. Therefore, managing resources in the cloud is difficult and calls for effective resource management techniques [5].

An effective resource management strategy impacts three distinct cloud-related characteristics. It satisfies cloud customers and meets SLA requirements. It guarantees the cloud's responsibilities to its users. As a result, users will keep using the cloud. As a result, both energy consumption and operating costs drop. Less energy use can result in reducing carbon emissions, which could facilitate green cloud computing. Cloud providers' profitability is improved by cost reduction and revenue growth [6, 45-48]. As a result, efficient resource management only allocates the minimal resources needed to meet SLAs [7] and frees up the extra resources to deploy new virtual machines (VMs) [8]. For this reason, the resources allotted in the cloud should be near the required demands so that the SLA is met and resource waste is kept to a minimum [36-40].

A crucial problem for elasticity is the quickness of responsiveness to workload changes to achieve the appropriate performance level [1]. Although matching the amount of resources allocated to the amount already needed is the key benefit of elasticity, the time it takes for resources to be available for use could be an issue [9]. Virtualization approaches provide the foundation for cloud elasticity and dynamic resource allocation [10]. The VM provisioning technologies require a lengthy period [11]. This delay is unbearable for activities that require resource scaling during computing. It could result in SLA violations, a decline in QoS, and, ultimately, a loss of the cloud's reputation. There are three methods to shorten the delay. The first strategy, VM provisioning technology, helps to prepare fresh VMs for requests [11] quickly. Modern VM provisioning technologies like streaming VM technology [12] and VM cloning [13] are unable to reduce the time used when creating VMs [11]. The second strategy is to request a plan of future resource needs from each customer. Due to cloud commitments and customers' lack of awareness, it is not practicable [11]. Due to VM technologies and gaps in client understanding, the only practical and effective way to quickly provision resources is to estimate future demand. In order to provide the resource manager enough time to assign the right resources before a workload spikes, a proactive prediction method projects future demand fluctuations. The resource management prepares the

virtual machines ahead of time and scales up the infrastructure if a sharp increase in demand is anticipated in the future.

In the same way, the assigned resources are also released under reduced demand. The freed-up resources can be allocated to VMs that require more resources or used to build new VMs. Indeed, Rapid elasticity [14] is attained when the demand and the resources allotted are immediately matched. Thus, SLAs are met for systems developed using cloud services, energy waste is prevented, and on-demand provisioning is met. However, offering cloud services that guarantee customers' changing QoS needs and avoid SLA violations is a major challenge. Currently, services are planned and provided based on resources' availability without any assurance of their predicted performance [15]. Therefore, forecasting future demand in the dynamic cloud environment is a crucial step for quick elasticity adoption and efficient resource allocation.

Although a lot of academic work covers various facets of cloud computing, there hasn't been thorough research on complete resource prediction in the cloud. A thorough analysis of resource prediction models in cloud systems is presented in this work. A comparison between the main resources predicted, prediction algorithms, datasets used for prediction, performance metrics for prediction evaluation, a prediction summary, and a general taxonomy of prediction methods have been presented. This paper presents a survey on the prediction of resource utilization. It comprehensively reviews the newest and most prominent cloud resource utilization prediction models. A general taxonomy for proposed models, techniques, and frameworks for resource utilization prediction is presented.

Despite the existence of several surveys on cloud computing, including [1], [7], [9], [16], [17], [18], [19], [20], and [21], there is a notable gap in the literature concerning resource utilization prediction models. No comprehensive survey focuses on the latest models proposed for predicting cloud resources. Moreover, existing surveys do not categorize prediction models based on the type of input cases—univariate or multivariate—which is crucial for understanding the correlation between predicted resources. The lack of such a structured analysis limits the ability to compare methodologies effectively and assess their effectiveness in real-world cloud environments.

To address this gap, this paper presents a structured and detailed survey of resource utilization prediction models in cloud computing environments.

The key contributions of this survey include:

1) First-of-its-kind comparison: This study is the first to classify cloud resource prediction models based on univariate and multivariate input cases rather than just the employed algorithms.

2) Comprehensive analysis: The paper reviews and evaluates recent and well-known prediction models, highlighting their strengths and limitations.

3) Categorization of models: A classification framework is introduced to organize existing works based on their prediction approach, algorithmic techniques, and primary objectives.

4) Insights on dataset usage and performance metrics: The survey examines the datasets used in prior research and the evaluation metrics applied to measure model performance.

5) Identification of research gaps and future directions: The paper highlights key open challenges and provides recommendations for improving cloud resource prediction models.

The following is how this work is organized: The research methodology is presented in Section II. The various prediction models are explained in Section III, and a comparison of these models is shown in Section IV. In Section V, the analysis and discussion of the proposed models are shown. The paper is finally concluded in Section VI.

II. Research Methodology

This survey uses the following methodology to guarantee a thorough and organized analysis of cloud resources prediction models: This study is a literature-based survey that methodically examines the body of research on cloud resource prediction, in contrast to questionnaire-based surveys. No primary data was gathered via questionnaires or surveys. Rather, this study categorizes and assesses prediction models according to their performance metrics, input instances, datasets, and methodology.

A. Study Selection

Studies were chosen on the basis of their contributions to cloud computing research, their recentness (published within the last five years), and their applicability to predicting cloud usage of resources.

B. Novel Classification Approach

Unlike existing surveys which mainly classify prediction models based on methodology or application features, this survey presents a fresh classification approach by differentiating between univariate and multivariate input cases. This distinction is necessary in order to understand the interaction between predicted resources, offering additional information on model performance.

To ensure a structured comparison, the classification framework in this survey categorizes prediction models based on the datasets used to assess the prediction models, the prediction algorithms, the types of resources that are predicted, the types of input cases for the predictions, and the performance metrics that are used to assess the prediction algorithms' output.

C. Reasons for Choosing the Proposed Models

For a number of reasons, this study is suitable for tackling the issue of resource usage prediction in cloud datacenters. For a number of reasons, this strategy is suitable for handling the issue of resource usage prediction in cloud datacenters:

1) Cloud environments are dynamic: Workloads in the cloud are very dynamic, and resource requirements change over time. The intricate relationships between several resource metrics, such as CPU, memory, disk I/O, and network traffic,

are frequently missed by univariate models, which forecast based on a single input variable (such as CPU usage). Conversely, multivariate models take into account several variables at once, producing predictions that are more reliable and accurate.

2) Enhanced resource efficiency: The suggested model sheds light on how various input scenarios affect scalability, resource efficiency, and prediction accuracy by contrasting univariate and multivariate input cases. This lessens over-provisioning and under-provisioning by assisting cloud providers in more efficient resource allocation.

3) Improved SLA compliance: Proactive resource allocation made possible by accurate resource utilization prediction ensures that SLAs are fulfilled while reducing resource waste. For cloud providers looking to maintain high QoS and customer satisfaction, this is especially crucial.

4) Filling in the gaps in the current literature: Current surveys mostly classify prediction models according to methods or application features [50], ignoring the kind of input cases. This survey closes a significant gap in the literature and offers a more thorough understanding of resource prediction models by concentrating on univariate and multivariate input cases.

D. Comparison Criteria Between the Proposed Prediction Models

Fig. 1 is designed to depict the main elements of the models for resource prediction in cloud environments, along with the datasets used to assess the prediction models, the prediction algorithms, the types of resources that are predicted, the types of input cases for the predictions, and the performance metrics that are used to assess the prediction algorithms' output. The key components are

1) Datasets: To train and evaluate a prediction model's performance, publicly accessible datasets like Google Cluster Trace and PlanetLab Workload Trace are utilized.

2) Algorithms: From basic regression models to cutting-edge ensemble learning and neural network architectures, a variety of machine learning, deep learning, and optimization techniques are applied.

3) Predicted resources: In order to optimize cloud operations, models typically forecast resource utilization metrics like CPU, memory, disk usage, and network traffic.

4) Performance metrics: The efficacy of the prediction models can be assessed using standard evaluation metrics such as RMSE, MAE, MAPE, and R^2 Score.

5) Prediction input cases: Predictability and adaptability are impacted by the univariate, multivariate, or hybrid input cases that models are built on.

III. Overview of Cloud Resource Prediction Techniques

Techniques for predicting cloud resource utilization are well–documented [18]. This section provides a detailed description of the related methods. This survey classifies the research papers according to the key strategies and approaches used to anticipate and manage resources in cloud computing systems. This classification aids in distinguishing between various techniques and their respective application areas. The prediction approaches are divided into the following categories:

- Machine Learning and Ensemble-based Approaches.
- Recurrent Neural Networks (RNN), LSTM, and Hybrid Deep Learning Models.
- Workload Pattern and Adaptive Prediction-based Approaches.

A. Machine Learning and Ensemble-Based Approaches

This category includes studies that use hybrid models or ensemble methods, which combine various prediction algorithms or strategies to increase resource forecasting accuracy. This category includes approaches such as regression, learning automata, and evolutionary algorithms, which focus on maximizing resource utilization by combining predictive techniques.

DP-CUPA, a CPU consumption prediction technique based on DBN and Particle Swarm Optimization (PSO), was presented by the authors of [23]. The three main processes in this technique are pre-processing training data samples, training DBN, and using autoregressive and grey models as basis prediction models. The PSO is used to estimate the DBN parameters throughout the learning phase.

A Functional Link Neural Network (FLNN) with a hybrid genetic algorithm (GA) and particle swarm optimisation (PSO) was used by the authors of study [19] to develop a multi-resource utilisation prediction model. Five-minute intervals were projected for the use of CPU and memory resources. Google Cluster Data was used to evaluate the proposed model. The lowest MAE errors obtained were 0.25 for CPU resources and 0.018 for memory resources. Despite the number of solutions in the literature, there is still a need for advanced methods with higher accuracy and faster execution times for predicting resource utilization in both univariate and multivariate input cases. Throughput, as its R2 score is close to 1 and hence can produce more accurate results.

The study of [30] predicted workload in a cloud environment by using a hybrid machine learning method that combines random forest for regression and decision trees for classification. The authors collected data at various time periods from Google cluster workload traces to predict network traffic, memory usage, CPU, and I/O operations. Their results showed that the average MAE and MSE error rates decreased by 0.34 and 0.48, respectively. The forecasting average values for recall, accuracy, and precision have increased by 0.89, 0.92, and 92.52%, respectively.

The study of [31] predicted the incoming workloads by using an advanced recurrent neural network (RNN) known as LSTM, and their combined Multiplicative LSTM (mLSTM) based models. They simulated their work in MATLAB to predict disk, memory, and CPU resources. With lower RMSE, MAPE, and MAE values across multiple users, mLSTM routinely outperforms LSTM and BiLSTM in predicting CPU and RAM resource requirements.

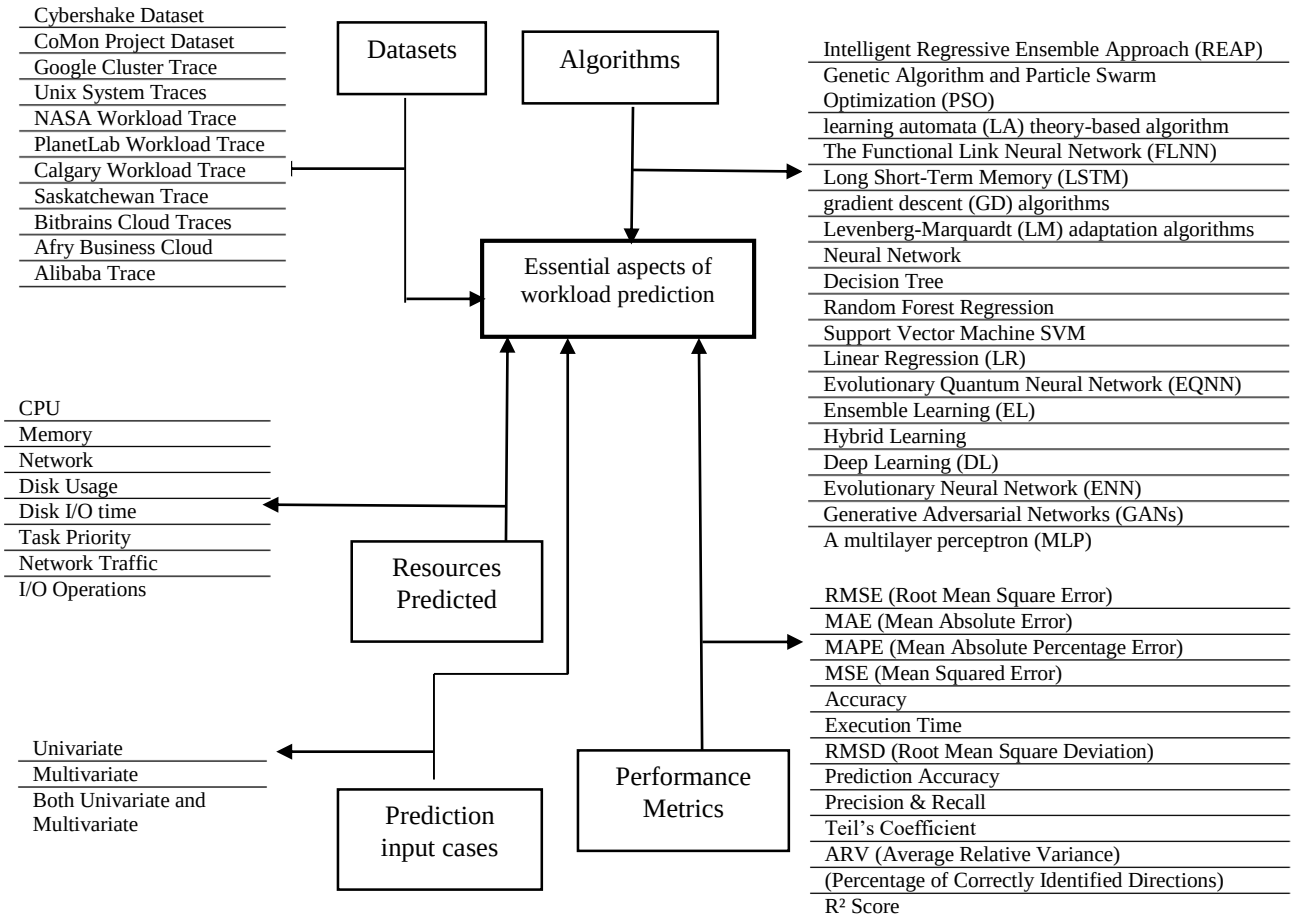

Fig. 1. Essential aspects of workload prediction.

In study [32], the authors employed a workload prediction model by using five classified machine learning-based techniques, including Evolutionary Neural Network (ENN), Evolutionary Quantum Neural Network [49] (EQNN), Hybrid Learning, Ensemble Learning (EL), and Deep Learning (DL). They applied the techniques within a standard environment for methodical research and comparison by employing three different cloud workload traces. They have assessed and contrasted the various learning-based models for time elapsed in training (TT), MAE, Absolute Error Frequency (AEF), and MSE with confidence metrics. The EQNN model achieves the lowest Mean Squared Error (MSE) of 1.79E-06.

B. LSTM and Hybrid Deep Learning Models

This section focuses on the research that uses neural network and LSTM-based approaches to predict cloud resources. The time series forecasting and sequential data processing capabilities of these models make them well-suited for resource utilization prediction in cloud systems. Hybrid models, which combine LSTM with other methods (e.g., CNN, fuzzy time series), seek to improve prediction performance by exploiting the capabilities of several algorithms.

The authors of study [16] proposed an automatic straggler (slow processing tasks) prediction and mitigation method for cloud environments that addressed heterogeneous host characteristics and volatile task characteristics using an encoder LSTM network. The encoder transmits the data to the LSTM following analysis of the load and resource utilization statistics.

An exponential moving average of the input matrices is also taken into consideration to prevent the LSTM model from diverging. CrystalLP, a storage workload prediction technique based on LSTM neural networks, is introduced in study [17]. This method creates a storage workload time-series model that gathers the desired workload patterns to support load balancing and accurate, adaptive scheduling. After that, an LSTM-based workload predictor is put into use, which is trained or optimized using an algorithm made up of the Adam optimizer and stochastic gradient descent (SGD).

The authors of study [20] introduced a multi-layer task failure prediction system based on Bi-directional Long Short Term Memory (Bi-LSTM). One input layer, two Bi-LSTM layers, one output layer, and the Logistic Regression (LR) layer are used to forecast whether the tasks will be finished or failed. Unlike classic LSTM, which only employs forward states, Bi-LSTM may work on both forward and backward states, allowing for more accurate estimation of the weights of both closer and distant input features.

The study of [21] created a turning point prediction model for cloud server workload forecasting that considers cloud workload factors. Next, a rule-filtering-based Piecewise Linear

Representation (PLR) approach is used to build a cloud feature-enhanced deep learning model for workload turning point prediction. The model's performance evaluation showed how effective its prediction accuracy was in terms of an increase in F1 score when compared to the state-of-the-art methods currently in use.

In study [24], an online learning approach for multivariate resource usage prediction models is proposed using the Levenberg-Marquardt and gradient descent methods. The predicted resources are CPU usage for seven and twenty days. The framework is evaluated using the PlanetLab workload trace and the Google cluster trace. A comparison between the learning abilities of the ARIMA and BLSTM models demonstrates that the BLSTM model performs significantly better. Sparse BLSTM is presented to address the challenge of adapting many parameters in BLSTM. A concept tree is created to help identify the parameters needing removal. Adapted sparse models and adapted dense models both produce similar predictions. Sparse real-time adaptations are 50–60% faster in the trimmed model when comparing the adaption times for dense and sparse models.

In study [25], a hybrid Convolutional Neural Network and Long Short-Term Memory) CNN-LSTM (model for analyzing multivariate workloads is presented. The main goal of this model is to efficiently model temporal fluctuations in the irregular trends of time series data while capturing complex patterns in VM consumption components. Bitbrains data is used to evaluate the presented model. The suggested and alternative prediction models are compared, including ARIMA-LSTM, VAR-GRU, and VAR-MLP. The findings indicate that the accuracy of the proposed model (improved from 3.8% to 10.9%) and error rate (which decreased to 7% from 8.5%) are better than other models.

The study by [26] offers a fresh viewpoint on forecasting seasonal and non-seasonal workloads. If the workload pattern exhibits seasonality, the Seasonal Auto-Regressive Integrated Moving Average (SARIMA) model is employed for forecasting purposes. The Long Short-Term Memory Networks (LSTM) or the Auto-Regressive Integrated Moving Average (ARIMA) model is used for non-seasonal workloads, depending on the normality test results. This study presents a prediction model that estimates the resources needed for various daily, hourly, and minute usage intervals. The experimental findings verify that the LSTM model's prediction accuracy beats ARIMA's for irregular workload patterns. The resource utilization is precisely predicted using the SARIMA model. The lowest MAE errors are achieved by using LSTM for predicting CPU and memory resources for one hour, which are 5.082 and 6.3835, respectively. The lowest MAE errors are achieved by using LSTM for predicting CPU and memory resources for minutes, which are 8.529 and 9.071, respectively.

The authors of study [27] predict a cloud server's CPU utilization using an LSTM. Their work reveals how Long Short-Term Memory (LSTM) networks, a kind of recurrent neural network perfectly suited for time series forecasting, may be used to model and predict the dynamic CPU consumption patterns of cloud-based apps. Their approach leverages historical data to enhance resource management and performance, offering valuable insights into how to boost cloud infrastructure efficiency. The engineering consulting company Afry (Afry is their brand name) acquired the data to train and test the models. Their findings show that in the case of single-step predictions, the moving average had the highest MSE, MAE, and LSTM had the lowest. The LSTM model demonstrates the lowest error rates, with an MSE of 0.8755 and MAE of 0.6643.

The authors of study [34] offer a novel hybrid approach by using Generative Adversarial Networks (GANs) with Long Short-Term Memory (LSTM) or Gated Recurrent Units (GRU) as generators and Convolutional Neural Networks (CNNs) as discriminators. The VTGAN model helps with proactive resource management by predicting future workloads as well as workload trends. According to their study, VTGAN achieves improvement in prediction accuracy spanning from 95.4% to 96.6%, outperforming conventional deep learning models in workload prediction and trend classification.

The study of [35] presents a multi-resource utilization prediction model that uses multiple approaches, namely support vector regression, RF, MLP regression, neural networks (NN) using Adam and SGD optimizers, and decision tree regression. The prediction model is based on univariate and multivariate time series. Google cluster trace data is used to evaluate the work. Four experiments are executed on the dataset, seeking to predict the resources for different time series interval periods. The outcomes of their experiments have shown that the prediction model yields higher accuracy compared to previous research.

C. Workload Pattern and Adaptive Prediction-Based Approaches

This section focuses on the research. This category focuses on research dedicated to monitoring systems and characterizing workloads, which are critical for real-time resource prediction and management in cloud computing environments. It focuses on the methods that modify forecasts in response to workload patterns or dynamically changing resource requirements. These techniques generally include adaptive algorithms that modify their prediction models in real-time to account for different workload patterns. This allows cloud data centres [22] to operate more efficiently and allocate resources more optimally. In this category, strategies like adaptive load balancing and workload discrimination are key points. A high-level summary of the methods utilized in cloud resource usage prediction is given in this section.

An efficient supervised learning-based Deep Neural Network (esDNN) technique has been suggested by the authors of study [28] to extract and learn the properties of past data and accurately anticipate future workloads. Once the multivariate data is converted into supervised learning time series, a modified GRU is used, which can adapt to changes in workload and address the drawbacks of gradient disappearance and explosion. Accurate prediction is made possible by this. A DNN-based workload prediction method, known as DNN-MVM, is described in study [51]. It handled data straight from these virtual machines using a feature selection engine and pre-processing. In order to give the cloud service provider greater information or expertise for resource management and

optimization, the model categorizes data according to prior loads. It is useful to predict future peak demands for resources. The validation of this model is done using the Grid Workload Archive (GWA) dataset.

In study [29], the authors suggested a multi-objective load-balancing approach integrated with a prediction model called the OP-MLB strategy for management of resources. They used neural networks customized with an adaptive evolutionary algorithm to predict cloud resources. The presented framework is evaluated on three real benchmark datasets: the traces of Google Cluster, PlanetLab virtual machines, with the Bitsbrain dataset. Over the course of five minutes and the three workloads, the approach achieved a minimal RMSE of 0.0005 for CPU resources.

The authors of this work [33] took inspiration from a collection of manipulative attack generation techniques to create adversarial cloud workload examples for four cutting-edge deep learning regression models—1D Convolutional Neural Network (1D-CNN), Recurrent Neural Network (RNN), Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM), and attention-based models. Three well-known cloud benchmark datasets—Google trace, Alibaba trace, and Bitbrain trace—were used to assess their research. Their analysis's findings demonstrate how vulnerable DL-based cloud workload forecasting models are to hostile attacks. In light of the existing literature, they were conducting systematic research for the first time to look at the susceptibility of DL-based methods within workload forecasting by highlighting inherent risks to the security and cost-effectiveness in those situations. Their final result indicates that the RMSE loss increases by 338.46% (RNN), 315.38% (LSTM), 325% (GRU), 83.33% (1D-CNN), and 300% (Attention-LSTM).

IV. Comparison Between the Proposed Prediction Models

Table I provides a comprehensive comparison between the proposed models for predicting cloud resources, highlighting important elements such as the models' algorithm, resources predicted, data input case, performance metrics, and summary/findings of the prediction. It addresses the benefits of each technique, such as accuracy and interpretability.

TABLE I. Comparison Between the Proposed Models

Ref	*Algorithm*	*Resources Predicted*	*Dataset*	*Data Input Case*	*Performance Metrics*	*summary/findings*
Tuli et al. [16] (2021)	LSTM	CPU, Memory, Bandwidth	PlanetLab traces	Univariate	MSE, MAPE	decreased SLA violations, execution time, resource contention, and energy by 13%, 11%, 16%, and 19%, respectively.
Ruan et al. [17] (2021)	CrystalLP	Request size	Web search archive SPC traces	Univariate	MAPE, RMSE, MAE	improved MAPE by 1.10% and outperformed current methods in MAE.
Malik et al. [19] (2022)	FLNN + Hybrid GA-PSO	CPU, Memory	Google Cluster Trace Dataset	Univariate/ Multivariate	MAE	Lowest MAE: 0.25 (CPU), 0.018 (Memory), improving prediction for both resources.
Gao et al. [20] (2020)	Bi-LSTM	55,55,55 tasks traces	task failure rate	Univariate	F1-Score	87% of task failures were correctly predicted with 93% accuracy..
Ruan et al. [21] (2022)	FEMTLSTM	CPU	Google Cluster, Alibaba, HPC Grid workloads	Univariate	Binary crossentropy, F1, precision, Recall	Compared to current methods, the F1 score is increased by 6.6%.
Wen et al. [40] (2020)	DP-CUPA	CPU	Google Cluster Trace Dataset	Multivariate	MSE, MAPE, MAE	outperformed the Grey, DBN, and autoregressive models.
Gupta et al. [24] (2020)	Gradient Descent (GD) + LM Adaptation	CPU	Google Cluster Trace Dataset and PlanetLab Workload	Univariate	RMSE MAPE	Achieved RMSE of 0.0095 and MAPE of 0.0239; adaptations are faster by 50-60%.
Ouhame et al. [25] (2021)	Neural Network + LSTM	CPU, Memory, Network	Bitbrains VM Trace Dataset	Multivariate	RMSE MSE MAE	Improved accuracy (3.8%-10.9%) and achieved RMSE: 0.1839, MAE: 0.7334 for multivariate predictions.
Anupama et al. [26] (2021)	LSTM	CPU, Memory	Bitbrains Cloud Workload Traces	Univariate	MAE MAPE	LSTM shows good accuracy: MAE (CPU, hourly): 5.082; (Memory, hourly): 6.3835
Starberg et al. [27] (2021)	LSTM	CPU	Afry Business Cloud Dataset	Univariate	MAE MSE	LSTM demonstrates low error rates: single-step MAE: 0.6643, multi-step MAE: 0.6848.
Xu et al. [28] (2022)	es-DNN	CPU usage per time-unit interval	Alibaba and Google Cluster traces	Univariate	MAPE, MSE, RMSE	efficiently decreased the number of active hosts and optimized expenses
Saxena et al. [29] (2022)	OP-MLB Framework	CPU Memory	Google Cluster Trace Dataset, PlanetLab, and Bitbrains VM Traces	Univariate	RMSE	Improved power savings by 85.3%; lowest RMSE: 0.0005 (CPU), 0.0035 (Memory).

Rao et al. [30] (2024)	Decision Tree + Random Forest Regression	CPU Memory Network traffic I/O operations	Cluster workload traces from Google	Univariate	MSE, MAE Prediction Accuracy, Precision, and Recall	MSE, and MAE significantly reduced (by 0.48 and 0.34); Precision and Recall improved to 92.52% and 0.89, respectively.
Nehra et al. [31] (2024)	Recurrent Neural Networks + LSTM	CPU, RAM, and local disk space	Cluster workload traces from Google	Univariate	RMSE, MAPE, and MAE	mLSTM achieves lower errors than LSTM and BiLSTM in CPU and RAM prediction.
Saxena et al. [32] (2023)	EQNN, EL, Hybrid Learning, DL, and ENN	CPU, memory	Google Cluster, PlanetLab	Univariate	MSE	The lowest MSE of 1.79E-06 is achieved by the EQNN model
Mahbub et al. [33] (2024)	RNN, LSTM, GRU, 1D-CNN , attention-based models	CPU Usage	Google trace, Alibaba trace, and Bitbrain	Univariate	RMSE	RMSE loss increases by 338.46% (RNN), 315.38% (LSTM), 325% (GRU), 83.33% (1D-CNN), and 300% (Attention-LSTM).
Maiyza et al. [34] (2023)	GANs with LSTM/GRU generators + CNNs as discriminators	CPU	Planet Lab traces	Univariate	RMSE, MAPE, Teil's coefcient, ARV, POCID, and R2 coefcient	High accuracy (95.4%–96.6%)
Bliedy et al. [35] (2025)	NN (Adam, SGD), SVR, RF, MLP, DTR	CPU Memory Disk usage Disk I/O time	Google cluster data	Univariate/ Multivariate	MAE, RMSE R-squared and MAPE	the prediction model yields better accuracy than previous research

V. Analysis and Discussion

This section provides a detailed analysis of the key findings from the resource utilization prediction models that were surveyed. It highlights patterns in model selection, contrasts the benefits and drawbacks of different approaches, and points out areas that require more research.

A. Important Discoveries and Patterns

The comparative analysis makes it evident that machine learning and deep learning models are being used more and more in cloud resource prediction. Conventional regression-based methods such as Decision Tree Regression (DTR) and Support Vector Regression (SVR) have shown good performance in univariate prediction scenarios. However, more advanced deep learning models, such as Long Short-Term Memory (LSTM) networks and hybrid neural network architectures, have shown greater accuracy in multivariate scenarios.

Multivariate models are able to capture the interdependencies between different types of resources (CPU, memory.

B. The Advantages and Disadvantages of Current Models

1) Univariate vs. Multivariate Models:

a) Univariate models often fail to capture the relationships between different cloud resources, even though they are computationally efficient.

b) Multivariate models, which produce more accurate predictions, require larger training datasets and more processing power.

2) Deep learning vs. Machine learning methods models:

a) Despite their interpretability and speed, machine learning models such as Random Forest (RF) and Decision Trees (DT) might not be able to manage long-term dependencies in time-series data.

b) Deep learning models, particularly LSTM and hybrid architectures, can effectively learn sequential data, but they usually require a great deal of training and fine-tuning.

3) Adaptability and scalability:

a) In large-scale cloud environments, certain models do not generalize well, but they do well in small-scale datasets.

b) Research on adaptive models that can dynamically adapt to changes in workload is still in its infancy.

4) Practical uses and consequences: Both researchers and service providers gain from accurate cloud resource prediction because it makes it possible to:

a) Optimizing resource provisioning to lower expenses and improve performance is known as efficient resource allocation.

b) Energy efficiency: Using accurate demand forecasting to reduce energy use and operating costs.

c) SLA compliance: Improving overall service quality and preventing violations by guaranteeing optimal resource allocation

C. Research Deficits and Prospects

1) Hybrid methods: Prediction accuracy can be increased by combining deep learning and machine learning.

2) Real-time adaptation: A lot of models don't adapt to shifting workloads in real time.

3) Thorough benchmarking: To properly compare models, standardized evaluation metrics are required.

4) Security and robustness: Accurate workload forecasting depends on resistance to adversarial attacks.

D. Limitations of the Proposed Models

1) Most research on cloud resource prediction focuses on predicting cloud resources based on univariate input cases where the prediction is based on a single input and single

output. There is relatively little work exploring multivariate input cases, where multiple input variables are used simultaneously to enhance prediction accuracy. Addressing this gap could lead to more robust and comprehensive resource prediction models that better reflect the dynamic nature of cloud environments.

2) They focused on forecasting CPU and memory resources using just one or two techniques without taking disk utilization and disk I/O time into account. This strategy reduces the efficacy of their models since it ignores important elements that affect system performance as a whole. There is a need for incorporating disk-related metrics with CPU and RAM, employing advanced or hybrid modelling methodologies for a more holistic approach to resource management in cloud environments, in order to build more thorough and accurate resource predictions.

3) They executed one or two experiments at most to evaluate their work, seeking to predict the resources for only one or two-time series intervals. This narrow approach restricts the generalizability of their models, as it does not adequately reflect the diverse and dynamic nature of cloud resource demands over different timeframes.

4) Only one or two performance metrics are reported in their experiments, which offers an insufficient assessment of the model's efficacy. This constrained evaluation ignores a thorough comprehension of the models' behavior under diverse circumstances, potentially hiding important features like accuracy, scalability, and robustness. Future studies should include a wider range of performance criteria for a more comprehensive assessment that better captures the advantages and disadvantages of the models in various circumstances.

These constraints must be addressed to create more thorough, flexible, and precise cloud resource prediction models.

VI. Conclusion

This survey provides a thorough discussion of resource usage prediction models in cloud computing, bridging a significant body of literature. Unlike other surveys, which consider only prediction algorithms, this work introduces a novel perspective by separating models into univariate and multivariate input cases. This distinction is necessary in order to understand the interaction between predicted resources, offering additional information on model performance. By systematic comparison of recent models, we uncover significant trends, performance measures, and evaluation sets. Further, our work identifies significant research gaps, such as the need for more generalizable models, improved feature selection algorithms, and adaptive learning methods able to enhance prediction effectiveness in evolving cloud environments. Lastly, this survey provides the foundation for future research and development of cloud resource prediction with a comparative analysis of existing methods and areas for innovation. Future studies must explore hybrid models, deep learning approaches, and real-time adaptive methods to further improve resource usage forecasting in cloud computing.

References

[1] E. F. Coutinho, F. R. de Carvalho Sousa, P. A. L. Rego, D. G. Gomes, and J. N. de Souza. Elasticity in cloud computing: A survey. Annals of Telecommunications - Annales des telecommunications, 70(7):289–309, 2015. ISSN 1958-9395. doi: 10.1007/s12243-014-0450-7. URL http: //dx.doi.org/10.1007/s12243-014-0450-7.

[2] S. Kulkarni and P. Agrawal. Analysis of TCP Performance in Data Center Networks. Springer New York, 2014. ISBN 978-1-4614-7860-7. doi: 10.1007/978-1-4614-7861-4.

[3] Barnawi, Ahmed, Sherif Sakr, Wenjing Xiao, and Abdullah Al-Barakati. "The views, measurements and challenges of elasticity in the cloud: A review." Computer Communications 154 (2020): 111-117.

[4] N. R. Herbst, S. Kounev, and R. Reussner. Elasticity in cloud computing: What it is, and what it is not. In the 10th International Conference on Autonomic Computing (ICAC 2013), San Jose, CA, USA, 2013.

[5] S. Singh and I. Chana. Resource provisioning and scheduling in clouds: QoS perspective. The Journal of Supercomputing, 72(3):926–960, 2016. doi: 10.1007/s11227-016-1626-x.

[6] S. Kumar and R. Buyya. Green Cloud Computing and Environmental Sustainability, pages 315–339. John Wiley & Sons, Ltd, 2012. ISBN 9781118305393. doi: 10.1002/9781118305393.ch16. URL http: //dx.doi.org/10.1002/9781118305393.ch16.

[7] S. S. Manvi and G. Krishna Shyam. Resource management for Infrastructure as a Service (IaaS) in cloud computing: A survey. Journal of Network and Computer Applications, 41:424–440, 2014. ISSN 1084-8045. doi: http://dx.doi.org/10.1016/j.jnca.2013.10.004. URL http: //www.sciencedirect.com/science/article/pii/S1084804513002099.

[8] S. K. Garg, A. N. Toosi, S. K. Gopalaiyengar, and R. Buyya. SLAbased virtual machine management for heterogeneous workloads in a cloud datacenter. Journal of Network and Computer Applications, 45:108–120, 2014. ISSN 1084-8045. doi: http://dx.doi.org/10.1016/ j.jnca.2014.07.030. URL http://www.sciencedirect.com/science/ article/pii/S1084804514001787.

[9] G. Galante and L. C. E. d. Bona. A survey on cloud computing elasticity. In 2012 IEEE Fifth International Conference on Utility and Cloud Computing, pages 263–270, Chicago, IL, USA, 2012. doi: 10.1109/ UCC.2012.30.

[10] K. Hwang, X. Bai, M. Shi, Y. Li, W. G. Chen, and Y. Wu. Cloud performance modeling and benchmark evaluation of elastic scaling strategies. IEEE Transactions on Parallel and Distributed Systems, 27(1):130–143, 2016. doi: 10.1109/TPDS.2015.2398438.

[11] Y. Jiang, C.-S. Perng, T. Li, and R. N. Chang. Cloud analytics for capacity planning and instant VM provisioning. IEEE Transaction on Network and Service Management, 10(3):312–325, 2013.

[12] F. Labonte, P. Mattson, W. Thies, I. Buck, C. Kozyrakis, and M. Horowitz. The stream virtual machine. In Proceedings of 13th International Conference on Parallel Architecture and Compilation Techniques, PACT '04, pages 267–277, Antibes Juan-les-Pins, France, 2004. ISBN 1089-795X. doi: 10.1109/PACT.2004.1342560.

[13] Gong, Y., Huang, J., Liu, B., Xu, J., Wu, B., & Zhang, Y. (2024). Dynamic resource allocation for virtual machine migration optimization using machine learning. Applied and Computational Engineering, 57, 1-8.

[14] P. Mell and T. Grance. NIST Special Publication 800-145, 2011. URL http://csrc.nist.gov/publications/nistpubs/800-145/SP800- 145.pdf.

[15] S. Singh and I. Chana. QoS-aware autonomic resource management in cloud computing: A systematic review. ACM Computing Surveys, 48(3), 2015. doi: 10.1145/2843889.

[16] Gurleen S. Tuli, S. S. Gill, P. Garraghan, R. Buyya, G. Casale, and N. Jennings, "Start: Straggler prediction and mitigation for cloud comp. environments using encoder lstm networks," IEEE Trans. on Serv. Comp., 2021.

[17] L. Ruan, Y. Bai, S. Li, S. He, and L. Xiao, "Workload timeseries prediction in storage systems: a deep learning based approach," Cluster Comp., pp. 1–11, 2021.

[18] Alzahrani, A., & Moustafa, A. A. (2022). A deep learning-based resource usage prediction model for resource provisioning in an

autonomic cloud computing environment. Neural Computing and Applications, 34, 10211–10228. https://doi.org/10.1007/s00521-021-06665-5

[19] Malik, S., Tahir, M., Sardaraz, M., & Alourani, A. (2022). A resource utilization prediction model for cloud data centers using evolutionary algorithms and machine learning techniques. Applied Sciences, 12(4), 2160.

[20] J. Gao, H. Wang, and H. Shen, "Task failure prediction in cloud data centers using deep learning," IEEE transactions on services computing, 2020.

[21] L. Ruan, Y. Bai, S. Li, J. Lv, T. Zhang, L. Xiao, H. Fang, C. Wang, and Y. Xue, "Cloud workload turning points prediction via cloud feature-enhanced deep learning," IEEE Trans. on Cloud Comp., 2022.

[22] Li, Z., Zhang, X., & Wang, Y. (2022). "A Hybrid CNN-LSTM Model for Real-Time Resource Utilization Prediction in Cloud Data Centers." IEEE Transactions on Parallel and Distributed Systems, 33(6), 1456–1468. DOI: 10.1109/TPDS.2022.1234567.

[23] Y. Wen, Y. Wang, J. Liu, B. Cao, and Q. Fu, "Cpu usage prediction for cloud resource provisioning based on deep belief network and particle swarm optimization," Concurrency and Computation: Practice and Experience, vol. 32, no. 14, p. e5730, 2020.

[24] Shaifu Gupta, Aroor Dinesh Dileep, and Timothy A. Gonsalves, "Online sparse blstm models for resource usage prediction in cloud datacenters," IEEE Transactions on Network and Service Management, vol. 17, no.4, pp2335-2349, 2020

[25] Soukaina Ouhame, Youssef Hadi, and Arif Ullah, "An efficient forecasting approach for resource utilization in cloud data center using CNN-LSTM model," Neural Computing and Applications, vol. 33, no.16, pp10043-10055, 2021

[26] Anupama, K. C., B. R. Shivakumar, and R. Nagaraja. "Resource utilization prediction in cloud computing using hybrid model." International Journal of Advanced Computer Science and Applications 12, no. 4 (2021).

[27] Nääs Starberg, Filip, and Axel Rooth. "Predicting a business application's cloud server CPU utilization using the machine learning model LSTM." (2021).

[28] M. Xu, C. Song, H. Wu, S. S. Gill, K. Ye, and C. Xu, "Esdnn: Deep neural network based multivariate workload prediction approach in cloud environment," arXivpreprint arXiv:2203.02684, 2022.

[29] Deepika Saxena, Ashutosh Kumar Singh, and Rajkumar Buyya, "OP-MLB: an online VM prediction-based multi-objective load balancing framework for resource management at cloud data center," IEEE Transactions on Cloud Computing, vol. 10, no.4, pp2804-2816

[30] Simhadri Mallikarjuna Rao, Gangadhara Rao Kancherla, and Neelima Guntupalli, "A Hybrid Machine Learning Approach to Cloud Workload Prediction Using Decision Tree for Classification and Random Forest for Regression," International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 10, no. 6, pp. 2240-2252, Nov.-Dec. 2024. doi: https://doi.org/10.32628/CSEIT2410488.

[31] Nehra P., Kesswani N., "A workload prediction model for reducing service level agreement violations in cloud data centers," Decision Analytics Journal, vol. 11, p. 100463, June 2024.

[32] Saxena, D., Kumar, J., Singh, A.K., and Schmid, S., "Performance analysis of machine learning centered workload prediction models for cloud," IEEE Transactions on Parallel and Distributed Systems, vol. 34, no. 4, pp. 1313-1330, 2023.

[33] Mahbub, Nosin Ibna, Md Delowar Hossain, Sharmen Akhter, Md Imtiaz Hossain, Kimoon Jeong, and Eui-Nam Huh. "Robustness of Workload Forecasting Models in Cloud Data Centers: A White-Box Adversarial Attack Perspective." IEEE Access (2024).

[34] Maiyza, Aya I., Noha O. Korany, Karim Banawan, Hanan A. Hassan, and Walaa M. Sheta. "VTGAN: hybrid generative adversarial networks for cloud workload prediction." Journal of Cloud Computing 12, no. 1 (2023): 97.

[35] Doaa Bliedy, Mohamed H. Khafagy, and Rasha M. Badry, "Dynamic Resource Utilization Prediction Model for Cloud Datacenter," IAENG International Journal of Applied Mathematics, vol. 55, no. 2, p

[36] X. Wang, L. Ma, X. Wang, Y. Shi, B. Yi, and M. Huang,"Truthful vnfi procurement mechanisms with flexibleresource provisioning in nfv markets," IEEE Trans. on Cloud Comp., 2022.

[37] D. Saxena and A. K. Singh, "Communication cost aware resource efficient load balancing (care-lb) framework for cloud datacenter," Recent Advances in Computer Science and Communications, vol. 12, pp. 1–00, 2020.

[38] A. K. Singh and D. Saxena, "A cryptography and machine learning based authentication for secure datasharing in federated cloud services environment," Journal of Applied Security Research, pp. 1–24, 2021.

[39] D. Saxena and A. K. Singh, "An intelligent traffic entropy learning-based load management model for cloud networks," IEEE Netw. Lttr., vol. 4, no. 2, pp. 59– 63, 2022.

[40] Y. Xie, L. Pan, S. Yang, and S. Liu, "A random online algorithm for reselling reserved iaas instances in amazon's cloud marketplace," IEEE Trans. on Network Science and Engineering, 2022.

[41] H. D. Kabir, A. Khosravi, S. K. Mondal, M. Rahman, S. Nahavandi, and R. Buyya, "Uncertainty-aware decisions in cloud computing: Foundations and future directions," ACM Comp. Surveys (CSUR), vol. 54, no. 4, pp. 1–30, 2021.

[42] D. Saxena and A. K. Singh, "A proactive autoscaling and energy-efficient vm allocation framework usingonline multi-resource neural network for cloud data center," Neurocomputing, 2020.

[43] D. Saxena, I. Gupta, A. K. Singh, and C.-N. Lee, "A fault tolerant elastic resource management framework towards high availability of cloud services," IEEE Trans. on Network and Service Management, 2022.

[44] D. Saxena and A. K. Singh, "an intelligent security centered resource-efficient resource management model for cloud computing environments," arXiv preprint arXiv:2210.16602, 2022.

[45] D. Saxena, A. K. Singh, C.-N. Lee, and R. Buyya, "A sustainable and secure load management model for green cloud data centres," Scientific Reports, 2023.

[46] W. Song, Z. Xiao, Q. Chen, and H. Luo, "Adaptive resource provisioning for the cloud using online bin packing," IEEE Trans. on Computers, vol. 63, no. 11, pp. 2647–2660, 2013.

[47] D. Saxena and A. Singh, "Security embedded dynamic resource allocation model for cloud data centre," Elec. Lttr., vol. 56, no. 20, pp. 1062–1065, 2020.

[48] D. Saxena and A. K. Singh, "Osc-mc: Online secure communication model for cloud environment," IEEE Comms. Lttr., vol. 25, no. 9, pp. 2844–2848, 2021.

[49] R. Gupta, D. Saxena, I. Gupta, A. Makkar, and A. K. Singh, "Quantum machine learning driven malicious user prediction for cloud network communications," IEEE Netw. Lttr., 2022.

[50] D. Saxena and A. K. Singh, "A high availability management model based on VM significance ranking and resource estimation for cloud applications," IEEE Transactions on Services Computing, vol. 16, no. 3, pp. 1604–1615, 2022.

[51] P. Bhagtya, S. Raghavan, and K. Chandraseakran,"Workload classification in multi-vm cloud environment using deep neural network model," in Proceedings of the 36th Annual ACM Symposium on Applied Comp., 2021, pp. 79–82.