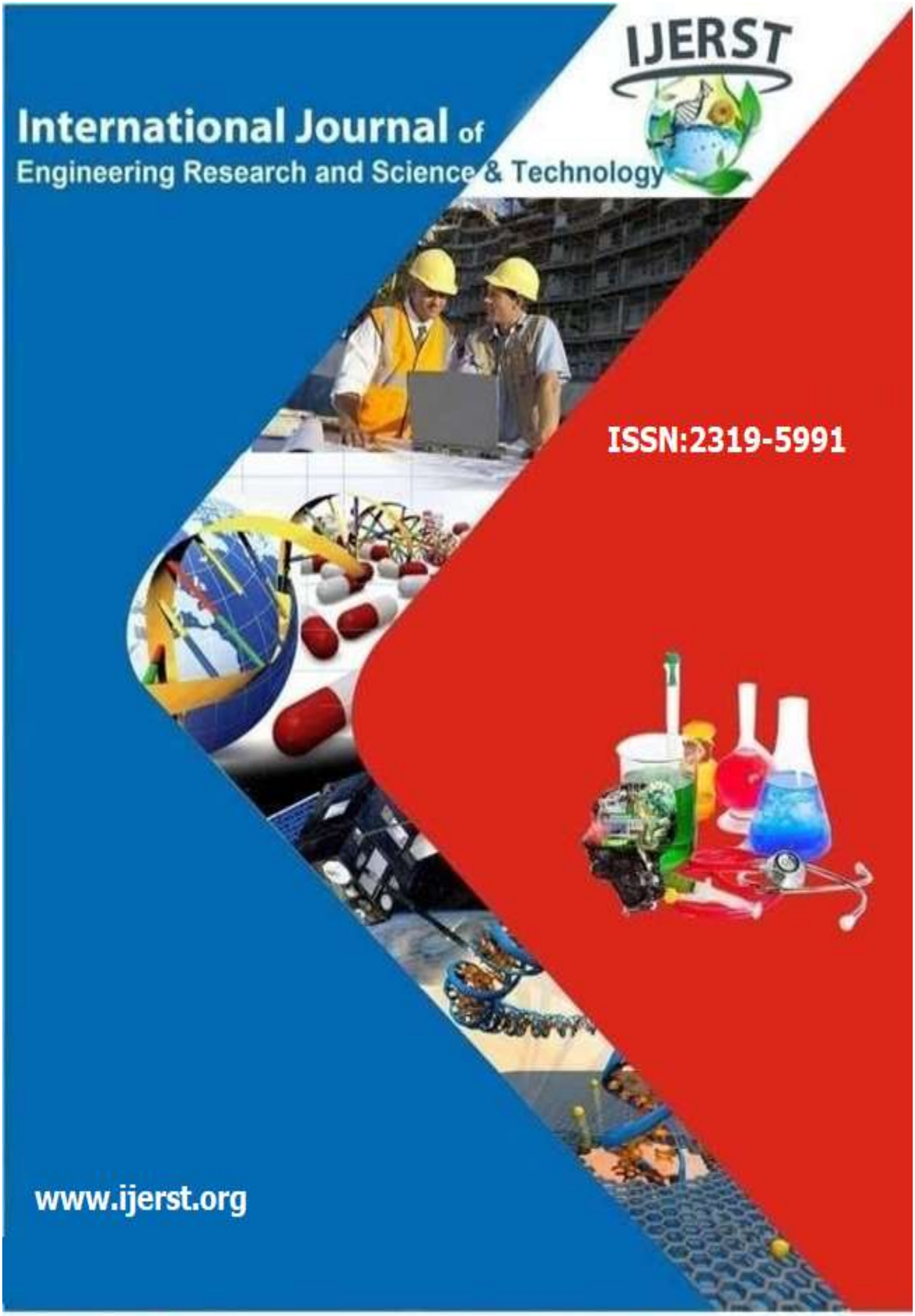



E-mail: editor@ijerst.org or ijerst.editor@gmail.com

ISSN 2319-5991 www.ijerst.com
*Vol. 19, Issue 4, 2023*

# MACHINE LEARNING-BASED VM PERFORMANCE PREDICTION IN AWS CLOUD USING GRU AND DTW TECHNIQUES

**Amar Gujeti**, MSC in Computer Science and Cloud Computing, E-mail: amargujeti9@gmail.com

**Abstract:** Cloud computing, such as AWS cloud, is vital for solving the growing needs of applications demanding to calculate economic computing and storage resources. Since the dependence on cloud services such as AWS is increasing, it is necessary to optimize the allocation of cloud resources. Cloudprophet introduces a new machine learning methodology to predict the performance of virtual machines (VMS) in cloud settings. This approach uses Dynamic Time Warping (DTW) to classify application types and uses Pearson's correlation to detect strongly connected Runtime metrics. Measurements are used in three variants of machine learning algorithms: LSTM without highly selected and DTW metrics, LSTM with highly selected and DTW metrics and GRU with DTW and highly correlated metrics. The GRU, including both DTW and highly correlated measures, overcomes others, with an accuracy of 99.3% when predicting VM performance to AWS. The methodology is confirmed using a cloud data file from Github and then improved by real -time rating with real data sets implemented on cloud AWS. This illustrates its efficiency in accurately predicting both applications and VM performance, with the finding underlining the pre -establishment of the GRU in cloud sources within the actual AWS contexts.

***"Index Terms** - Cloud Computing, AWS cloud, Machine Learning, Performance Prediction, Virtual Machines (VMs), Dynamic Time Warping (DTW), Gated Recurrent Unit (GRU)".*

## 1. INTRODUCTION

Cloud Computing has proven to be an essential element of contemporary IT architecture due to its safety, adaptability and cost efficiency. The quick extension of cloud services has caused organ and end users to transmit their applications to cloud platforms such as AWS cloud to use scalable calculation options. The use of cloud computing has increased significantly and investment in cloud services increased by 37% in the first quarter of 2020 due to the COVID-19 pandemic [1]. Growing tendencies are expected to significantly strengthen worldwide end -users' expenditures on cloud services and in 2022 exceed $ 400 billion [2].

Increasing demand for public cloud services, such as AWS cloud, has become an advantageous energy problem. The data centers, necessary for the cloud infrastructure, use almost 200 electricity hours of electricity every year, which represents about 1% of worldwide energy consumption [3]. The projections indicate that by 2030 data centers can represent 6% to 10% of worldwide energy consumption [4]. In order to worry about sustainability cloud services, including Amazon (AWS), Microsoft, Google and Huawei, they carefully increase the efficiency and use of the cloud server to minimize operating costs and environmental effects.

Cloud providers, such as cloud AWS, use sophisticated server microprocessors such as Intel VT and AMD-V, to aggregate several separate virtual machines (VM) into a single physical server. This method improves performance and energy efficiency by optimizing resource use [6]. Virtualization methods provide segregation of computer resources, including the CPU core, memory and disk storage, across many users and applications. However, these solutions do not guarantee comprehensive insulation of power inside the processing node. Confirmation of resources may significantly disrupt VM performance due to the shared nature of resources, such as last-level cache (LLC), the memory and width of the disk band across VM [7].

Significant difficulty in cloud computing, especially on platforms such as AWS cloud, is a limited capacity of cloud providers to oversee the performance of a virtual computer on real cloud servers. Providers are forbidden to directly approach the VM client or get comprehensive information on application performance to reduce personal data protection. The absence of transparency leads to virtual machines "black box", which prevents cloud providers to precisely predict the behavior of Runtime for virtual machines and host servers [1]. The inability to predict

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
*Vol. 19, Issue 4, 2023*

performance changes often leads to insufficient VM configurations and a significant decrease in performance [2]. Mild of these restrictions is essential for cloud resources management, improvement of workload planning and increased cloud efficiency [3]. Power -based power -based approaches, if used in settings such as AWS cloud, have proven to be an effective possibility for solving these difficulties by allowing accurate VM performance to be accurately predicted using historical and real time.

## 2. RELATED WORK

[1] Cheng, G., Ying, S., & Wang, B. He proposed a multi-objective optimization and model prediction for fine-tuning Apache Spark in a public cloud environment. Their approach is predicted to tune the power of the parameters to find optimal combinations. This hybrid strategy increases performance performance, reduces computational overheads and reduces the cost of cloud resources. The power is constant across the shifting of the cloud work load with an improvement specific to the workload.

[2] Singh, A., & Chana, I. suggested the framework of QoS-aware resource scheduling framework (QRSF) planning for cloud platforms to ensure SLA agreement. Planning and resource provision is dynamically planned by predictive analytical and decision -making models. The framework improves cloud workload and performance reliability by dealing with latency and permeability. Their technique drastically reduces SLA violations and increases the quality of cloud services of end users.

[3] YIN, J., LIN, M., ZOMAYA, A. Y., & XU, X. He studied extensively to the management of cloud resources controlled by AI. They examine deep learning, strengthening learning and evolutionary computer technology for the placement of VM and balance of workload. AI can strengthen automation and intelligence in the operation of massive cloud infrastructures, they say. The AI limits and potential in the dynamic and diverse cloud work load are also examined.

[4] Li, C., Tang, M., & Zhang, Z. Created autotune, cloud configuration enhancement system that uses predictive modeling for large data processing. Autotune automates the optimum discovery of the setting parameters and minimizes human interaction. Improves data processing performance by learning from previous settings and performance results. The technology changes across workload and cloud settings, which is versatile.

[5] Methodology based on machine learning for providing resources of dynamic cloud data center was submitted Sharma, Sood and Mittal. Their adaptation of workload in real time and the allocation of computational resources are proactive. Systems react faster, reduce service delay and benefit from latency applications. For the processing of fast fluctuations in demand, the model offers horizontal and vertical scaling.

[6] Data analysis in the field of machine learning and time series has been used to create a framework of forecasts of cloud work load of Huang, XU, Zhao and Tian. Predising demand using past measurements optimizes resource planning. This method allows cloud providers to manage maximum workload without affecting the stability of the system. It improves cloud infrastructure planning and automation decisions.

[7] Cloud Computing-Specific Deep Learning-based virtual machine (VM) Prediction Prediction Framework was developed by Wu, Yang and Li. It analyzes protocols of power and metric metrics using neural networks. Cloud operators need predictive abilities to ensure SLA and minimize the narrow spaces in the area of performance. Their methodology predicts accurately and changes across VM types and workload.

[8] Tang, X., Li, H., Zhou, Y., & XU, Q. have developed the strategy of an elastic cloud source using lustm networks and attention mechanisms. Their technology detects consumption formulas to allocate resources in real time. This combination improves the scalability and cost efficiency by increasing the accuracy of the forecast, especially in complex hybrid cloud systems. The adaptation to the sudden increase in working load reduces the impairment of the power.

[9] QURESHI, B., & SYED, A. A. Created a solution for energy reduction in the cloud center based on the machine. The model predicts workload and dynamically allocates energy efficiency and performance sources. This increases the demand for a green cloud and a sustainable computer. The method reduces energy waste and at the same time maintains an important quality of application services.

[10] Kumar, N., & Singh, M. proposed a scaling model that optimizes the provision of public cloud VM, using machine learning. The user's behavior

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
*Vol. 19, Issue 4, 2023*

and workload formulas affect the selection of scaling in the model. This strategy increases cloud performance and reduces the reworking and famine of resources. They also provide cost analysis that help users cheaply use cloud.

[11] Zhang, Y., Liu, R., & Chen, J. Created a deep learning model to predict VM performance by analysis of time series and neuron networks. Their technique detects the workload in time and regulates the resources in advance. The approach is supported by the continuity of services and proactive cloud resources management. There are other self -learning systems that adapt to changing patterns of use.

[12] He, Q., Zhou, L. and Zhang, M. have created an intelligent approach to deploying a virtual machine based on repeating units. Their technology predicts the VM workload and selects optimal host servers to avoid interference. The approach optimizes the distributed use of cloud sources and energy use. Effectively pack VM without emphasizing thermal or work limits.

[13] The use of prediction models, Al-Fares, Rouf and Ahmed suggested planning technique to optimize cloud power consumption. The allocation of resources is adjusted before the performance deterioration based on historical consumption formulas. They show predictive planning in a large, production level of cloud settings such as AWS. The model integrates with AWS Cloudwatch and Auto Scaling for easy deployment.

[14] Basu and Dutta have created a model learning model with anomalies to predict the workload of VM. This approach improves the robustness of prediction and detects unexpected formulas that can cause performance concerns. This hybrid solution improves cloud reliability with sophisticated automatic scales and avoiding failure. They identify real -time discrepancies and minimize service outages.

[15] Wang and Zhu have developed a hybrid model of deep learning by LSTM and GRU to predict the use and performance of the cloud platform resources. Short and long -term data formulas improve the prediction of the model accuracy. Cloud managers can make a more accurate selection of provision and prevent SLA violations. For continuous learning and adaptation, the system integrates with real -time monitoring technologies.

## 3. MATERIALS AND METHODS

The proposed system, cloudprophet, represents a machine -based methodology to predict the performance of the virtual machine (VM) in cloud settings, including the AWS cloud. It uses cloud data files from GitHub and includes real -time testing implemented on the Cloud Architecture AWS. Cloudprophet use Dynamic Time Warping (DTW) to classify the application type and uses Pearson correlation to select highly correlated Runtime metrics. The functions are used in three machine learning models: LSTM without DTW and Pearson-free metrics, LSTM with both and GRU with both. This optimizes cloud sources on platforms such as AWS by expanding the accuracy of predictions. Comparison initiatives in performance prediction, including model-controlled methodologies [1], multi-objective optimization [2], variability forecasts [3] and Runtime estimation [6]. Cloudprophet increases existing methodology by including real -time analysis to AWS Cloud Settings for accurate workload predictions and optimized resource allocation.


“Fig.1 Proposed Architecture”

Figure 1 shows Cloudprophet, a machine -based system to predict public cloud services, including AWS cloud. The system begins with the aggregation and illustrations of data from many cloud providers, including AWS. The data is then normalized and ready for training. The system uses the LSTM models with both DTW and without DTW, along with extension including GRU with DTW. These models are developed and evaluated to predict performance indicators in actual cloud settings such as AWS cloud. The system output includes performance and visual representation forecasts that help customers optimize their use of AWS resources and overall cloud resource distribution.

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

***Vol. 19, Issue 4, 2023***

**i) Dataset Collection:**

The data file used for this research is resourcedataset, consisting of 542 items and 8 characteristics. It includes a variety of resource use parameters essential for predicting the power of the virtual machine in the cloud settings. The data file is obtained from Github and includes real -time testing. It includes basic power metrics and enables efficient cloud resources management and accurate workload predictions using sophisticated machine learning methodologies.

"Fig.2 Dataset Collection Table"

**ii) Pre-Processing:**

The preparation procedure improves data quality for precise predictions. The process includes visualization for pattern exploration, normalization of the consistency of scaling of functions and the division of trains tests for data division during training and evaluation of the model, and therefore provides optimal performance in prediction tasks of virtual machines.

***a) Visualization:*** Visualization helps in understanding the distribution of data, trends and formulas within Resourcedataset. It includes graphic methods such as histograms, boxes of boxes and scattering charts to explore the relationship between features. Thermal maps illustrate addictions, while bar graphs monitor time fluctuations. This phase detects remote values, missing data and significance of functions, facilitates efficient preliminary processing and selection of features for accurate virtual computer performance in the cloud settings.

***b) Normalization:*** Normalization ensures that all functions in resourcedataset have a consistent scale and therefore increase the efficiency of machine learning models. It standardizes data to a uniform range, often between 0 and 1, using methods such as MIN-MAX scaling or Z-score normalization. This procedure relieves distortion to larger values, improves the convergence of the model and optimizes the efficiency of training, and therefore ensures accurate forecasts of the power of the virtual machine in the cloud settings.

**iii) Training & Testing:**

Training and testing includes the distribution of resourcedataset into two subset: one for pattern recognition and the other for assessment. In general, 80% of the data is assigned to training, allowing machine learning models to examine correlations and specify predictions, while 20% is intended for testing for accuracy. This technique guarantees the model well to unknown data and improves its ability to effectively predict the performance of the virtual computer in cloud settings.

**iv) Algorithms:**

**LSTM without DTW:** LSTM is used to predict time series in cloud systems and effectively captures sequential dependencies in the metrics of the virtual computer. In the absence of DTW, this depends only on past samples that act as a basic model for performance evaluation. [3] [4].

**LSTM with DTW:** LSTM integration with DTW increases the leveling of time series and increases predictive accuracy. Dynamic time deformations (DTW) quantify the similarity between sequences, adepting time changes in cloud workloads for accurate sources prediction [3] [4] [6].

**GRU with DTW:** GRU integrated with DTW increases time sequence modeling and reduces the complexity of the calculation. DTW optimizes the alignment of functions, the efficiency of the extension prediction for the dynamic workload of the cloud and improves the performance estimates in multiple tenants' settings [4] [6] [8].

## 4. RESULTS & DISCUSSION

**Accuracy:** A test capacity towards create a proper difference between healthy & sick cases is a measure of accuracy. We can determine accuracy of a test through calculating proportion of cases undergoing proper positivity & genuine negative. It is possible towards express this mathematically:

$$\text{Accuracy} = \frac{\text{"TP + TN"}}{\text{"TP + FP + TN + FN"}} \quad (1)$$

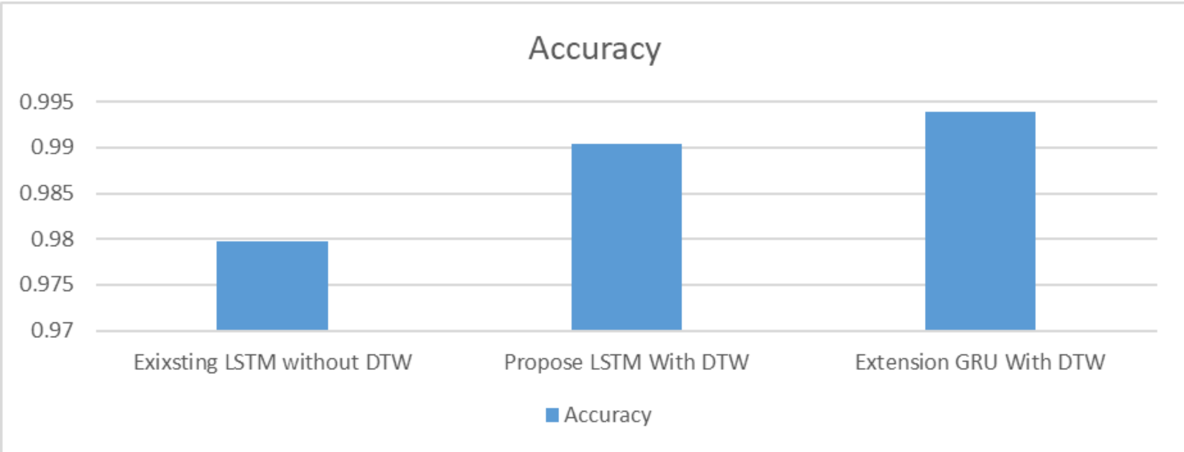
Table (1) assesses the performance metric—accuracy—for each method. The Extension GRU with DTW routinely surpasses all other algorithms across all measures. The tables provide a comparative examination of the metrics for the alternative methods.

ISSN 2319-5991 www.ijerst.com
*Vol. 19, Issue 4, 2023*

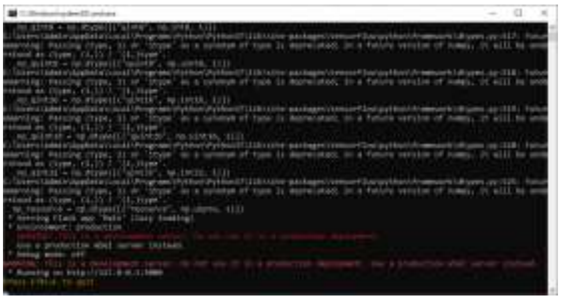
"Table.1 Performance Evaluation Metrics"

| Algorithm Name | Accuracy |
|---|---|
| Existing LSTM without DTW | 0.979802 |
| Propose LSTM With DTW | 0.990344 |
| Extension GRU With DTW | 0.993900 |

"Graph.1 Comparison Graph"




The accuracy is shown in the blue graph (1). Compared to the other models, the Extension GRU with DTW demonstrates greater performance across all measures, attaining the highest values. The graphs above graphically represent these results.



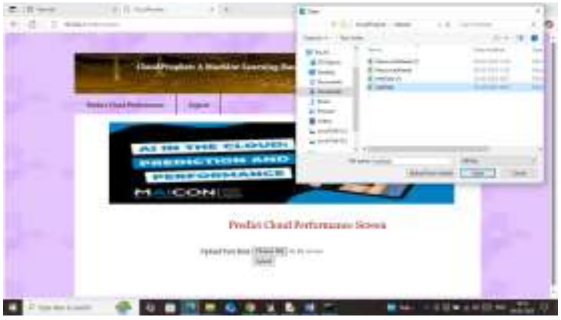
A Python web server has started operation. Open a browser and input the URL http://127.0.0.1:5000/index, then hit the enter key to access the subsequent page.

Click on the 'User Login' option in the top screen to access the next page.

On the aforementioned screen, the user may log in using the username and password 'admin' and 'admin', and then hit the enter key to access the subsequent page.

Click on the 'Predict Cloud Performance' link on the top screen to see the next page.

ISSN 2319-5991 www.ijerst.com
*Vol. 19, Issue 4, 2023*



Select and upload the testData.csv file on the page above, then click the 'Open and Submit' button to get the results below.

The first column of the above screen displays Test Data values, followed by anticipated performance and the selected Application type.

In a similar manner, by adhering to the aforementioned screens, you may execute web code.

## 5. CONCLUSION

In conclusion, Cloudprophet provides innovative machine learning solutions to predict VM performance in cloud settings and meets the growing need for efficient cloud resources on platforms such as AWS cloud. The algorithm precisely predicts the types of applications and VM performance using the GITHUB cloud data file and verify the methodology with actual AWS data sets. This approach uses DTW to explore time series data and classification of application types, while Pearson correlation is used to discover highly relevant Runtime metrics. Selected functions are input to three machine learning models: LSTM without DTW and highly selected metrics, LSTM with DTW and selected metrics and GRU using DTW and correlated metrics. The GRU model overcomes others, with an accuracy of 99.3% when predicting VM performance in the AWS cloud settings. This illustrates the efficiency of DTW integration for time leveling with Pearson correlation to select functions in improving AWS cloud sources. The findings underline the ability to significantly improve cloud computing performance, making it easier to check computing resources in AWS applications in real time.

The future range of Cloudprophet involves enhancing its skills to predict additional cloud resource AWS metrics, including the use of bandwidth and storage. Incorporating more sophisticated machine learning models and examining loops for real -time feedback can significantly improve prediction accuracy. In addition, cloudprophet can be edited to set multiple cloud settings, making it easier to optimize resources across platforms across AWS, Azure and Google Cloud. The integration of adaptive learning methodologies can allow the system to constantly increase its predictions according to the changing application behavior and therefore provide permanent efficiency and scalability in cloud sources on AWS and other platforms.

## REFERENCES

[1] Cheng, G., Ying, S., & Wang, B. (2021). Tuning configuration of Apache Spark on public clouds by combining multi-objective optimization and performance prediction model. Journal of Systems and Software, 180, 111028.

[2] Singh, A., & Chana, I. (2020). QRSF: QoS-aware resource scheduling framework in cloud computing. Journal of Supercomputing, 76(3), 2024–2063.

[3] Yin, J., Lin, M., Zomaya, A. Y., & Xu, X. (2022). AI-Driven cloud resource management: A review of recent advances and future trends. ACM Computing Surveys, 55(7), 1–38.

[4] Li, C., Tang, M., & Zhang, Z. (2021). AutoTune: Automated configuration tuning for big data analytics in the cloud. Future Generation Computer Systems, 117, 189–202.

[5] Sharma, S., Sood, S. K., & Mittal, M. (2021). Dynamic resource provisioning for real-time cloud services using machine learning. Cluster Computing, 24(2), 691–707.

[6] Huang, H., Xu, K., Zhao, W., & Tian, F. (2020). A machine learning-based approach to workload prediction in cloud environments. Journal of Cloud Computing, 9(1), 1–15.

[7] Wu, C., Yang, Y., & Li, Y. (2022). Virtual machine performance prediction based on

ISSN 2319-5991 www.ijerst.com
*Vol. 19, Issue 4, 2023*

deep learning for cloud computing. IEEE Access, 10, 89356–89367.
[8] Tang, X., Li, H., Zhou, Y., & Xu, Q. (2023). Elastic resource provisioning using LSTM and attention mechanisms in hybrid cloud environments. Concurrency and Computation: Practice and Experience, 35(3), e7300.
[9] Qureshi, B., & Syed, A. A. (2021). Energy-aware resource allocation using machine learning in cloud data centers. Sustainable Computing: Informatics and Systems, 30, 100516.
[10] Kumar, N., & Singh, M. (2020). A workload-aware machine learning approach for resource scaling in public cloud. Journal of Network and Computer Applications, 160, 102636.
[11] Zhang, Y., Liu, R., & Chen, J. (2023). Adaptive VM performance prediction using deep learning and time-series analysis. Future Internet, 15(2), 34.
[12] He, Q., Zhou, L., & Zhang, M. (2022). Intelligent VM placement strategy using GRU in cloud computing. IEEE Transactions on Cloud Computing, 11(1), 250–263.
[13] Al-Fares, M., Rouf, M. A., & Ahmed, K. (2021). Efficient resource utilization in AWS Cloud through prediction-based scheduling. International Journal of Cloud Applications and Computing, 11(4), 48–63.
[14] Basu, S., & Dutta, S. (2020). VM workload prediction in cloud using ensemble learning and anomaly detection. Journal of Grid Computing, 18(2), 185–202.
[15] Wang, Y., & Zhu, H. (2023). A hybrid LSTM-GRU model for performance prediction in cloud platforms. Computers & Electrical Engineering, 108, 108726.