Prediction of Resource Utilisation in Cloud Computing Using Machine Learning

Ruksar Shaikh, Cristina Hava Muntean and Shaguna Gupta[a]
School of Computing, National College of Ireland, Dublin, Ireland

Keywords: Cloud Computing, Machine Learning, Deep Learning, Resource Utilization, CPU Utilization, Network Transmission Throughput.

Abstract: In today's modern computing infrastructure, cloud computing has emerged as a pivotal paradigm that offers scalability and flexibility to satisfy the demands of a wide variety of specific applications. Maintaining optimal performance and cost-effectiveness inside cloud settings continues to be a significant problem and one of the most important challenges is efficient resource utilisation. A resource utilization prediction system is required to aid the resource allocator in providing optimal resource allocation. Accurate prediction is difficult in such a dynamic resource utilisation. The applications of machine learning techniques are the primary emphasis of this research project which aims to predict resource utilisation in cloud computing systems. The dataset GWA-T-12 Bitbrains have provided the data of timestamp, cpu usage, network transmitted throughput and Microsoft Azure traces has provided the data of cpu usage of a cloud server. To predict VM workloads based on CPU utilization, machine learning models such as Linear Regression, Decision Tree Regression, Gradient Boosting Regression, and Support Vector Regression are used. In addition to these, deep learning models such as Long Short-Term Memory and Bi-directional Long Short-Term Memory have also been evaluated in our approach. Bi-directional Long Short Term Memory approach is considered more effective as compared to other models in terms of CPU Utilisation and Network Transmitted Throughput as its R2 score is close to 1 and hence can produce more accurate results.

1 INTRODUCTION

Cloud service providers often adopt a pay-as-you-go pricing model, which can result in cost savings and increased flexibility for cloud users. The vast variety of improvements in cloud computing technology has resulted in a considerable growth in cloud users and the development of cloud-based applications to access various cloud computing services. Several scientific applications use cloud computing services, resulting in varying utilisation of cloud resources. As a result, efficient resource management is required to handle the shifting demand of users. Efficient resource management in a cloud computing environment can help to optimise resource utilisation, save costs, and improve performance. Resource utilisation prediction is used to accomplish efficient resource management (Malik et al., 2022). Predicting the consumption of cloud resources such as CPU, memory, and network throughput is critical for effective resource management (Kaur et al., 2019). CPU utilisation is one of the most essential metrics for measuring the performance of host machines. It is also a prominent indicator for researchers to evaluate when attempting to anticipate the performance of hosts in the future. The central processing unit (CPU) is typically the resource that is subject to the highest amount of demand in virtualized settings. As a result, it is a significant contributor to resource shortages on cloud host devices (Mason et al., 2018). Machine learning algorithms have gained a lot of attention and are becoming commonplace in cloud computing applications in recent years. Inspired by the structure of the brain, the Neural Network is one of the most versatile and successful machine learning techniques available. Because neural networks approximate functions, they can be used to solve a wide range of issues, from robotics to regression (Duggan et al., 2017). Efficient resource utilisation is still a major difficulty in modern cloud computing settings, affecting cloud systems' cost-effectiveness and performance. It is difficult to forecast resource utilisation in such dynamic contexts, even with cloud models' inherent scalability and flex-

[a] https://orcid.org/0000-0002-9361-3097

Shaikh, R., Muntean, C. and Gupta, S.
Prediction of Resource Utilisation in Cloud Computing Using Machine Learning.
DOI: 10.5220/0012742200003711
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 14th International Conference on Cloud Computing and Services Science (CLOSER 2024)*, pages 103-114
ISBN: 978-989-758-701-6; ISSN: 2184-5042

ibility. In this paper, we predict virtual machine CPU utilization using ML and DL predictive models. This research aims to investigate the accuracy of predictive models for predicting CPU utilization and network throughput transmission and comparing them.

This paper is discussed as follows: Section 2 presents the overview of the existing works related to the prediction of resource utilization using ML and Dl models. Section 3 presents the method of research, implementation steps of the predictive models. Section 4 presents evaluated results and section 5 presents conclusion and future of the research.

2 LITERATURE REVIEW

Resource usage prediction is becoming more and more popular due to recent advancements in the field of resource management (Amiri and Mohammad-Khanli, 2017). The various prediction techniques based on deep learning and machine learning methodologies are compiled in this section.

2.1 Resource Utilisation Using Machine Learning Techniques

(Borkowski et al., 2016) introduces Cloud resource provisioning through the use of machine learning-based models to predict resource utilisation at the task and resource levels. Evaluations demonstrate significant gains in accuracy, with 20% reduction in prediction errors and up to 89% improvements. (Conforto et al., 2017) presents a unique machine learning-based resource utilisation prediction system for IaaS clouds that dynamically estimates resource requirements. It offers major improvements in IaaS infrastructure management and optimisation by combining historical data with real-time monitoring to optimise resource allocation, increase cost efficiency, and improve overall IaaS performance.

(Mehmood et al., 2018) emphasises how crucial it is to allocate resources precisely on cloud platforms to prevent waste or deterioration in service. It suggests utilising machine learning approaches to build precise predictive models for an ensemble-based workload prediction system. In large-scale production, (Morariu et al., 2020) investigates how machine learning might improve scheduling and resource allocation. Making use of previous data to develop prediction models, it tackles the intricacies of industrial operations. Learning from past trends, these models—which include supervised and unsupervised machine learning algorithms—optimize scheduling choices. (Daid et al., 2021) investigates data centre scheduling, with a focus on optimising CPU utilisation and using machine learning (ML) to fulfil service level agreement (SLA) needs. The paper focuses into issues with CPU efficiency and SLA fulfilment and suggests a hybrid machine learning strategy that combines regression and clustering models for scheduling.

(Manam et al., 2023) suggests a unique method for cloud computing that optimises resource scheduling and lowers costs by using the Random Forest algorithm. The approach builds decision trees for classification and regression and is well-known for ensemble learning. A novel approach to predicting CPU utilization in virtualized environments is presented by (Estrada et al., 2023). It increases forecast accuracy by clustering related virtual machines according to resource utilisation trends using a streamlined VM clustering technique. (Khurana et al., 2023) focuses on improving Gradient Boosting models to predict CPU utilisation in cloud environments. This probably entails a lot of parameter optimisation, such as feature engineering, cross-validation methods, and hyperparameter fine-tuning.

2.2 Resource Utilisation Using Deep Learning Techniques

(Wang et al., 2016) introduces a proactive VM deployment approach in cloud computing, using CPU utilization predictions via the ARIMA-BP neural network. By foreseeing performance issues, it revolutionizes deployment strategies, ensuring service quality and server efficiency. (Duggan et al., 2017) investigates the use of recurrent neural networks (RNNs) to predict CPU utilization in cloud computing. By analyzing historical CPU and network data, the study employs RNNs to capture temporal dependencies and forecast usage patterns. (Nääs Starberg and Rooth, 2021) focuses on managing CPU fluctuations in cloud computing by introducing an LSTM model. It forecasts CPU usage up to 30 minutes ahead, aiding in dynamic capacity scaling. Through performance evaluations against RNNs and state-of-the-art models, its accuracy in predicting future utilization is assessed.

(Shivakumar et al., 2021) proposes a hybrid model for cloud resource utilization forecasting, combining SARIMA for seasonal workloads and LSTM/ARIMA for non-seasonal patterns. It highlights LSTM's accuracy in irregular patterns, SARIMA's effectiveness in forecasting future usage, and its significance in helping providers avoid resource over or under-provisioning.

In Table 1 Review of works related to Resource Utilization Prediction Techniques.

Table 1: Summarized related works of resource utilisation in cloud computing.

Author	Title	Dataset	Tool	Technique	Result
(Duggan et al., 2017)	Predicting host CPU utilization in cloud computing using recurrent neural networks	No application/ Dataset of CoMon project	PlanetLab	Recurrent Neural Network	Prediction accuracy is improved.
(Daid et al., 2021)	An effective scheduling in data centres for efficient CPU usage and service level agreement fulfilment using machine learning	Randomly generated data	Matlab	Linear Regression	Prediction accuracy is improved.
(Manam et al., 2023)	A Machine Learning Approach to Resource Management in Cloud Computing Environments	Materna dataset Trace 3	Google Colaboratory platform	Random Forest algorithm	Prediction accuracy is improved.
(Mehmood et al., 2018)	Prediction Of Cloud Computing Resource Utilization	Google cluster usage trace data	Cloud system	Ensemble based workload prediction mechanism	Prediction accuracy is improved.
(Shivakumar et al., 2021)	Resource Utilization Prediction in Cloud Computing using Hybrid Model	Bitbrains dataset	Experiment was conducted using fastStorage, real trace data of Bitbrains data center	SARIMA, LSTM, ARIMA	Prediction accuracy is improved.
(Conforto et al., 2017)	Adaptive Resource Utilization Prediction System for Infrastructure as a Service Cloud	Bitbrains dataset	fastStorage of Bitbrains data center	ARIMA and Autoregressive Neural Network (AR-NN)	Prediction accuracy is improved.
(Wang et al., 2016)	Research on the Prediction Model of CPU Utilization Based on ARIMA-BP Neural Network	IBM Server	Xen System	ARIMA-BP neural network	Prediction can be improved.
(Nääs Starberg and Rooth, 2021)	Predicting a business application's cloud server CPU utilization using the machine learning model LSTM	Afry dataset	Python	LSTM	Prediction accuracy is improved.

The Table 1 showcases a variety of approaches leveraging different datasets, tools, and machine learning algorithms such as recurrent neural networks, linear regression, random forests, and LSTM among others. Several studies demonstrate improved prediction accuracy when forecasting resource utilization in cloud environments. However, a compelling trend surfaces from the reviewed literature: the utilization of LSTM-based models consistently demonstrates enhanced predictive capabilities across various datasets. The Bidirectional LSTM, with its ability to capture long-term dependencies and process sequential data bidirectionally, presents itself as a robust choice for modeling the complex temporal patterns inherent in cloud resource usage.

The choice of BiLSTM model stems from its capacity to effectively capture both past and future context, which is particularly relevant in resource utilization forecasting where historical trends and future behavior significantly impact predictions. The utilization of this model offers the potential to enhance accuracy, thereby aiding in proactive resource allocation and optimization in cloud environments.

3 RESEARCH METHODOLOGY

The research methodology followed in this research consists of the following steps:

- **Research Understanding:** With a focus on optimising resource utilisation, enhancing performance, and cost reduction, the study aims to predict the accuracy of VM CPU Utilisation and Network Transmission Throughput using ML and DL prediction models in cloud computing environments.
- **Data Collection:** Both qualitative and quantitative information on virtual machine CPU utilisation and network transmission throughput was taken from open-source repositories (BitsBrain dataset from gwa-t-12-bitbrains, Microsoft Azure traces from GitHub).
- **Data Pre-processing:** For model readiness, feature engineering and data cleaning are performed. Addressing missing values and getting datasets prepared for the training of ML and DL prediction models.
- **Predictive Models Creation:** Using the selected datasets, train the predictive ML and DL model. BiLSTM is chosen for its capacity to capture complicated temporal correlations, aiming to predict VM CPU Utilisation and Network Transmission Throughput accurately.
- **Evaluation:** Metrics include Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error, and R-squared (R2) value to assess prediction accuracy.
- **Performance Criteria:** Aimed at achieving efficient resource utilization, enhancing performance, and reducing operational costs.
- **Experimentation and Feedback:** Experimental scenarios are meticulously designed with controlled variables to rigorously test the predictive models.

3.1 Dataset Description

This research utilises two key datasets, the Bitbrains dataset obtained from gwa-t-12-bitbrains and the Microsoft Azure Traces 2017 dataset sourced from GitHub. Both CPU utilisation and network transmission throughput are particularly predicted by the BitBrains dataset, while the Microsoft Azure Traces dataset focuses on CPU utilisation in the context of time series. Timestamp, CPU utilisation, and network transmission throughput data are the three main parameters in the BitBrains dataset. Similarly, CPU utilisation and network transmission throughput patterns can be predicted using data from the Microsoft Azure Traces 2017 dataset. Key resource variables, such as CPU utilisation, are accessible through the datasets and are essential for the predictive models used in this research.

3.2 Resource Provisioning Framework

In cloud environments, the availability of computing resources such as network capabilities, storage capacities, and CPU power forms the cornerstone of service provision. Predictive ML and DL models play a crucial role by forecasting CPU usage and Network Transmission Throughput, significantly impacting these resources. These models enable cloud service providers to anticipate resource requirements more accurately, thus optimizing the allocation of network, storage, and CPU resources to align with projected demand. By harnessing the insights from these models, cloud environments achieve enhanced resource utilization and allocation efficiency. When considering resource allocation strategies, the influence of predictive modeling insights is profound in both reservation-based and on-demand scenarios. For reservations, predictive models inform allocation strategies by accurately predicting resource needs over time. This approach ensures resources are reserved efficiently, minimizing wastage while guaranteeing sustained usage in alignment with anticipated demands. Simultaneously, for on-demand scenarios, predictive models drive real-time resource optimization by dynamically adjusting CPU and network resources based on immediate requirements.

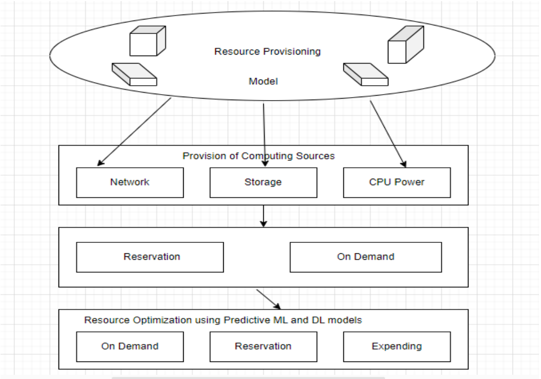


Figure 1: Resource Provisioning Framework.

Optimization strategies, predictive ML, and DL models play a pivotal role in various aspects of resource management. Real-time on-demand resource optimization leverages these models to swiftly adapt CPU and network resources to immediate needs. This agile adjustment significantly improves resource utilization efficiency, thereby enhancing system performance during fluctuating demands. Additionally, reservation-based optimization harnesses predictive models to minimize unnecessary resource reservations, ensuring system stability while maintaining efficiency.

This paper proposes a resource provisioning model within cloud computing that integrates various predictive Machine Learning (ML) and Deep Learning (DL) models, including Linear Regression, Decision Tree Regression, Support Vector Regression, Gradient Boosting Regression, LSTM, and Bi-LSTM. This model acts as a comprehensive framework, guiding the allocation and management of critical resources like network bandwidth, storage capacities, and CPU power based on insights derived from these predictive models in figure 1.

3.2.1 Linear Regression (LiR)

Linear Regression is a conventional regression model that seeks to establish a linear correlation between independent variables (features) and a dependent variable (resource utilization). Linear Regression is a method used to predict CPU utilization and Network-transmitted throughput in cloud computing. It aims to identify direct linear relationships between different factors that affect resource usage and the actual utilization levels.

3.2.2 Decision Tree Regression (DTR)

Decision Tree Regression allows resource utilisation to be predicted by building a tree-like structure based on data features. Decision Tree Regression would generate decision rules based on characteristics like CPU usage and network transmission throughput in order to forecast resource utilisation levels in the context of cloud resource prediction.

3.2.3 Gradient Boosting Regression (GBR)

Gradient Boosting Regression creates an ensemble of decision trees and uses error minimization to improve predictions iteratively. This model integrates several techniques to improve forecasts in cloud computing resource prediction by comprehending intricate interactions between various elements influencing resource usage.

3.2.4 Support Vector Regression (SVR)

Support Vector Regression is a method that determines the most accurate hyperplane to reflect the connection between input data and resource utilisation. Within the realm of cloud computing, it establishes a multidimensional threshold to predict CPU utilisation and Network-transmitted throughput by considering several aspects.

3.2.5 Long Short Term Memory (LSTM)

LSTM, a form of recurrent neural network (RNN), excels at capturing dependencies in sequence data. When predicting resource utilisation in cloud computing, LSTM would focus on temporal patterns in CPU usage and network-transmitted throughput, recognising minor changes and trends that emerge over time.

3.2.6 Bi-Direction Long Short Term Memory (BiLSTM)

BiLSTM enhances LSTM by performing data processing in both the forward and backward directions, enabling the simultaneous capture of both past and future context. BiLSTM is used in cloud computing resource prediction to analyse temporal dependencies in both directions, allowing for a more comprehensive understanding of CPU utilisation and Network-transmitted traffic patterns. The comparison between traditional regression models like Linear Regression, Decision Tree Regression, Gradient Boosting Regression, and Support Vector Regression with deep learning models like LSTM and BiLSTM mirrors the evaluation of simpler, rule-based approaches against complex, memory-enhanced models in their ability to predict resource utilization patterns in cloud computing, particularly focusing on CPU utilization and Network-transmitted throughput.

The proposed approach is grounded in utilizing the BiLSTM (Bidirectional Long Short-Term Memory) model to improve the accuracy of predictions for both VM CPU Utilization and Network Transmission Throughput. Firstly, datasets from Bitsbrain (for CPU Utilization and Network Transmission Throughput) and Microsoft Azure Traces 2017 (for CPU Utilization) are collected and meticulously preprocessed to ensure completeness and relevance in the context of the study. After that, the predictive models are implemented and rigorously trained using these datasets. The primary focus lies in assessing key predictive metrics such as Mean Absolute Error(MAE), Mean Squared Error(MSE), Root Mean Squared Error(RMSE), and R2 score to evaluate the model's accuracy in predicting VM resource utilization.

3.3 Framework Implementation

This research project employs Python programming language within Google Colab to predict resource utilization. Figure 2 shows the implementation steps.

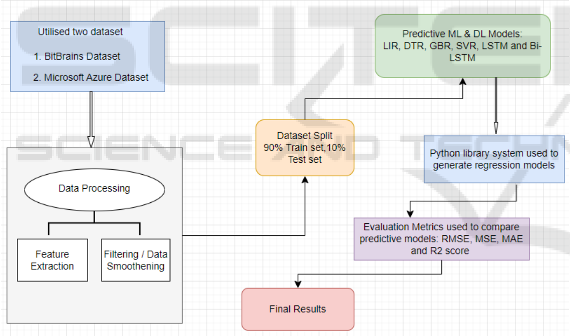

Utilised two dataset
1. BitBrains Dataset
2. Microsoft Azure Dataset
Data Processing
Feature Extraction
Filtering / Data Smoothening
Dataset Split 90% Train set,10% Test set
Predictive ML & DL Models: LIR, DTR, GBR, SVR, LSTM and Bi-LSTM
Python library system used to generate regression models
Evaluation Metrics used to compare predictive models: RMSE, MSE, MAE and R2 score
Final Results

Figure 2: Roadmap of Implementation.

This study utilizes two datasets to predict resource utilization: the Bitbrains dataset, which includes CPU usage, network transmission throughput, and timestamp data. Selected the minimum, maximum, and average CPU utilization, as well as the timestamp, from the Microsoft Azure Traces dataset. Both datasets are implemented individually. The Bitbrains dataset is utilized for predicting CPU utilization and network transmission throughput, whereas the Microsoft Azure Traces dataset is specifically employed for predicting CPU utilisation.

Prior to initiating the modeling process, meticulous checks are conducted to identify any missing values and address them appropriately, in order to guarantee the integrity and comprehensiveness of the datasets. The data normalization process in both the Bitbrains and Microsoft Azure Traces datasets involves the use of feature scaling algorithms, specifically MinMaxScaler. The essential preprocessing stage normalizes the attributes, guaranteeing consistency and optimal efficiency throughout the training of the machine learning model.

The implementation phase commences with a thorough Exploratory Data Analysis (EDA) focused on gaining a comprehensive understanding of the datasets. This comprehensive analysis involves closely examining important aspects such as the shape, size, data types, mean values, column names, counts, standard deviations, and the range between the minimum and maximum values of the dataset. These statistical insights offer a comprehensive perspective on the datasets, which is crucial for subsequent modelling.

The machine learning method begins with identifying the target column, referred to as 'y', which will be predicted by the models. In order to streamline the process of training and evaluating the model, the dataset is split into four distinct subsets: X-train, X-test, y-train, and y-test. The data is split into two sets using a 90-10 ratio, with 90% given for training and 10% for testing. It is easy to re-train the machine learning and deep learning models if the new dataset contains the predicted parameters i.e. timestamp, CPU usage, and network transmitted throughput.

This enables the efficient execution of many machine learning and deep learning techniques such as Linear Regression, Decision Tree Regression, Gradient Boosting Regression, Support Vector Regression, Long Short Term Memory, and Bi-directional Long Short Term Memory.

These algorithms are executed using a robust library system, which ensures accurate results. In addition, this method effectively adapts the algorithms by using the features provided by the library system.

Evaluation metrics such as Mean Square Error, Mean Absolute Error, R Square Score, and Root Mean Square Error are calculated to predict the accuracy of each regression model. It is also noted that if the number of layers in a model i.e. BiLSTM is increased, the accuracy will increase. Specifically, it is seen that when we increased the number of layers by 1 then accuracy showed an improvement of 4% approximately.

Real-time testing is conducted to validate the models' effectiveness in practical scenarios using test data, ensuring their viability and accuracy in a live cloud environment. Continuous monitoring and optimization of these models remain pivotal, allowing for adjustments based on evolving cloud infrastructure dynamics and patterns within the datasets. Ultimately, this implementation aims to provide a robust predictive system facilitating efficient resource allocation, improved performance, and cost reduction within cloud computing infrastructures.

4 EVALUATION

In this section, the effectiveness of conventional machine learning algorithms as described in literature is assessed against the proposed approach. Employing the Scikit-Learn library, the experiments are conducted on the Google Colaboratory platform, serving as the environment for training and testing. Four distinct machine learning algorithms are evaluated: Linear Regression, Decision Tree Regression, Gradient Boosting Regression, Support Vector Regression, and two deep learning models such as LSTM and BiLSTM.

4.1 Performance Metrics

4.1.1 Root Mean Squared Error (RMSE)

RMSE is a measure of the differences between predicted values and observed values. It represents the square root of the average of the squared differences between the predicted and actual values. The formula for RMSE is:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1}$$

Here, y_i represents the actual value, y_i^2 represents the predicted value, and n is the number of samples. (Sourced from (Chugh, 2020))

4.1.2 R-squared (R2) Score

R2 score represents the proportion of the variance in the dependent variable that is predictable from the independent variables. It is calculated as the ratio of the explained variation to the total variation. The formula for R2 score is:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2}$$

Here, y_i represents the actual value, y_i^2 represents the predicted value, and n is the number of samples. (Sourced from (Chugh, 2020))

4.1.3 Mean Squared Error (MSE)

MSE measures the average of the squares of errors or deviations. It's calculated by taking the average of the squared differences between predicted and actual values. The formula for MSE is:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3}$$

Here, y_i represents the actual value, y_i^2 represents the predicted value, and n is the number of samples. (Sourced from (Chugh, 2020))

4.1.4 Mean Absolute Error (MAE)

MAE is the average of the absolute differences between predicted and actual values. It measures the average magnitude of errors without considering their direction. The formula for MAE is:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4}$$

Here, y_i represents the actual value, y_i^2 represents the predicted value, and n is the number of samples. (Sourced from (Chugh, 2020))

4.2 Evaluation of Resource Utilisation for Machine Learning and Deep Learning Models

In this research project, we have conducted implementation using two public available datasets i.e BitBrains and Microsoft Azure Traces 2017. We have evaluated the CPU Utilisation and Network Transmitted Throughput using BitBrains dataset, alongwith CPU Utilisation evaluated for Microsoft Azure dataset.

4.2.1 Evaluation of CPU Utilisation

The evaluated results of CPU Utilisation using BitBrains dataset are presented in details in Table 2 and Figure 4. Also for MicroSoft Azure dataset, the evaluated results of CPU Utilization are presented in details in Table 3 and Figure 5. The MSE, MAE, RMSE and R2 metrics of the ML and DL algorithms compared in this paper are shown in Table 2 and Table 3. The results shows that for RMSE, Decision Tree Regression and Gradient Boosting Regression algorithms had higher error values when compared to BiLSTM and Linear Regression model which performed better than the compared models shown in Table 2 and Table 3. The evaluated results of the CPU utilization for the prediction and actual values of the machine learning models are presented in Figure 4 and Figure 5. Hence, BiLSTM and Linear Regression performed better than the compared approaches followed by LSTM.

Table 2: For BitBrains dataset - Comparison of machine learning and deep learning algorithms for CPU Utilization prediction.

Model	MSE	MAE	RMSE	R2 score
LiR	0.0027	0.0215	0.0521	0.7794
DTR	0.0099	0.0402	0.0995	0.1951
GBR	0.0055	0.0294	0.0747	0.5465
SVR	0.0030	0.0389	0.0556	0.7488
LSTM	0.0026	0.0233	0.0515	0.7843
BiLSTM	0.0024	0.0224	0.0490	0.8042

Table 3: For Microsoft Azure dataset - Comparison of machine learning and deep learning algorithms for CPU Utilization prediction.

Model	MSE	MAE	RMSE	R2 score
LiR	0.0002	0.0131	0.0169	0.9833
DTR	0.0023	0.0390	0.0480	0.8661
GBR	0.0010	0.0255	0.0321	0.9399
SVR	0.0015	0.0337	0.0388	0.9127
LSTM	0.0009	0.0239	0.0304	0.9462
BiLSTM	0.0004	0.0169	0.0214	0.9732

Figure 4 and figure 5 illustrates the predictions for CPU utilization using a range of machine learning and deep learning techniques, including Linear Regression(LIR), Gradient Boosting Regression(GBR), Decision Tree Regression(DTR), Support Vector Regression(SVR), Long Short Term Memory(LSTM) and Bi-directional Long Short Term Memory (BiLSTM). These predictive models enable accurate forcasting of the CPU utilization, providing valuable insights into the resource demands and usage patterns within the cloud environment. Figure 6 presents the predictions for the network transmission throughput, utilizing machine learning and deep learning models, including LIR, GBR, DTR, SVR, LSTM, and BiLSTM. These predictions offer valuable insights into the anticipated network throughput trends and patterns within the cloud environments, aiding in the proactive management and optimization of network resources.

4.2.2 Evaluation of Network Transmission Throughput

The evaluated results for network transmission throughput are presented in Table 4 and Figure 6. From the results, it can be seen that BiLSTM has very close values when compared to the actual value. BiLSTM and Linear Regression have achieved higher network transmission throughput prediction accuracy than the compared models with 0.9 and 0.92 for R2 metrics respectively and lower error rates for RMSE and MAE.

Table 4: For BitBrains dataset - Comparison of machine learning and deep learning algorithms for Network Transmission Throughput prediction.

Model	MSE	MAE	RMSE	R2 score
LiR	0.0012	0.0114	0.0359	0.9256
DTR	0.0043	0.0473	0.0656	0.752
GBR	0.00183	0.0259	0.0428	0.894
SVR	0.0037	0.0495	0.0610	0.786
LSTM	0.0026	0.0232	0.0513	0.848
BiLSTM	0.00172	0.0203	0.0415	0.901

4.2.3 Performance Comparison of ML and DL Models for Predicting CPU Utilization

In comparing models for CPU Utilization predictions using BitBrains and Microsoft Azure datasets, the BiLSTM and Linear Regression model consistently stood out as the most accurate in Figure 3. Compared to LSTM, SVR, GBR, and DTR models, BiLSTM and Linear Regression consistently demonstrated superior performance across both datasets. BiLSTM has advantages over Linear Regression as its strength in capturing complex temporal dependencies allowed for more precise predictions of CPU Utilization dynamics. Also, it works better where data might exhibit non-linear patterns. While other models showed promise to varying degrees, none matched the robustness of BiLSTM in handling the intricacies within these datasets. This underlines the pivotal role of model architecture in effectively predicting CPU Utilization across diverse datasets. The new information of this research paper is that the BiLSTM shows consistently better performance for both the chosen datasets as compared to other models.

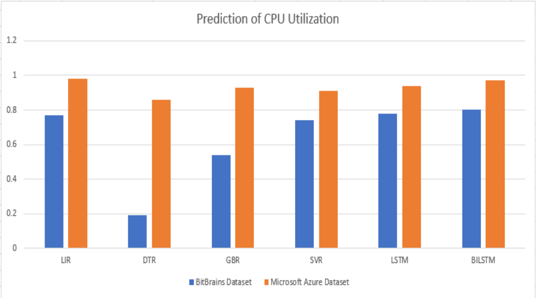
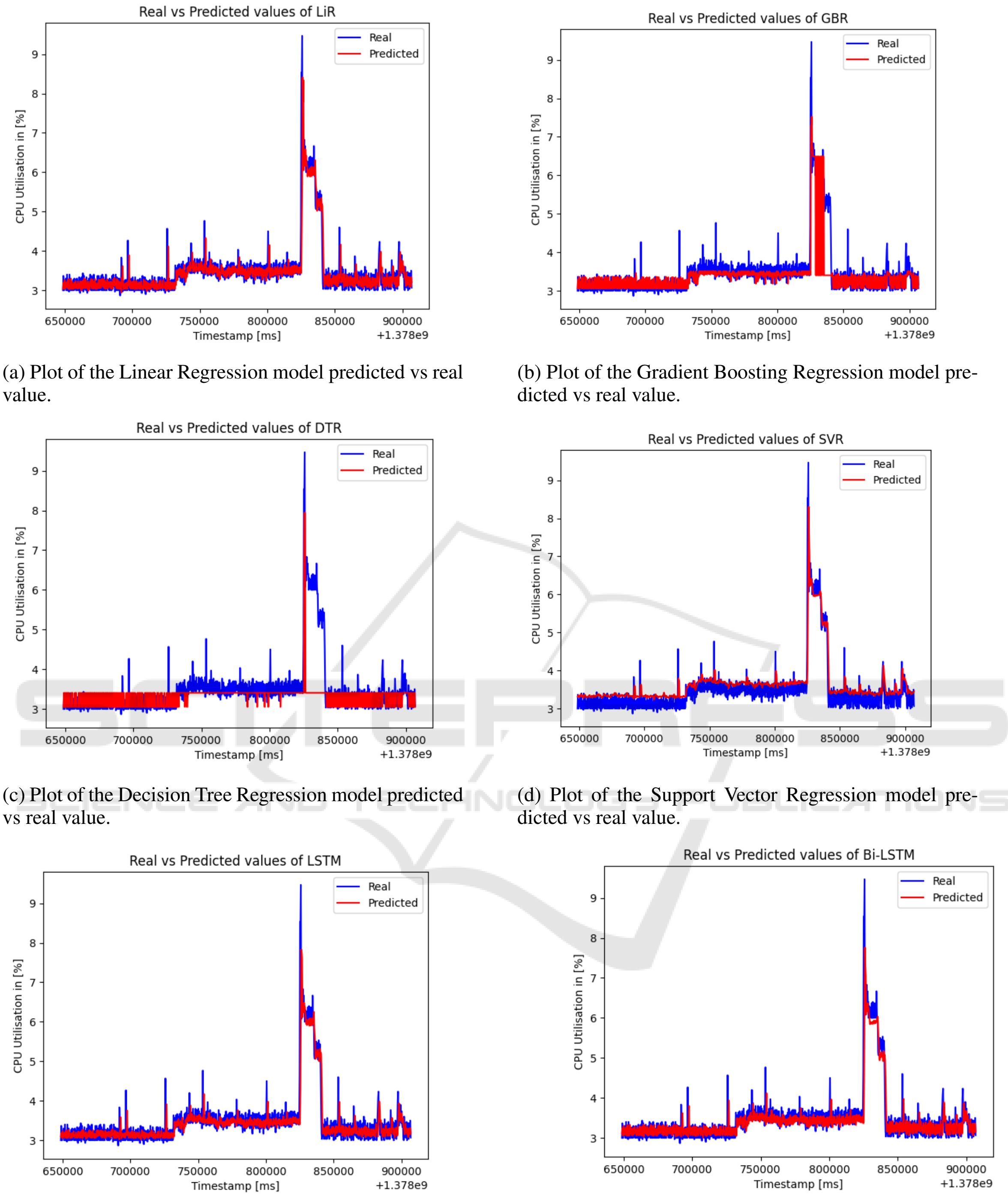


Figure 3: Comparison of Prediction of CPU Utilisation.



(a) Plot of the Linear Regression model predicted vs real value.

(b) Plot of the Gradient Boosting Regression model predicted vs real value.

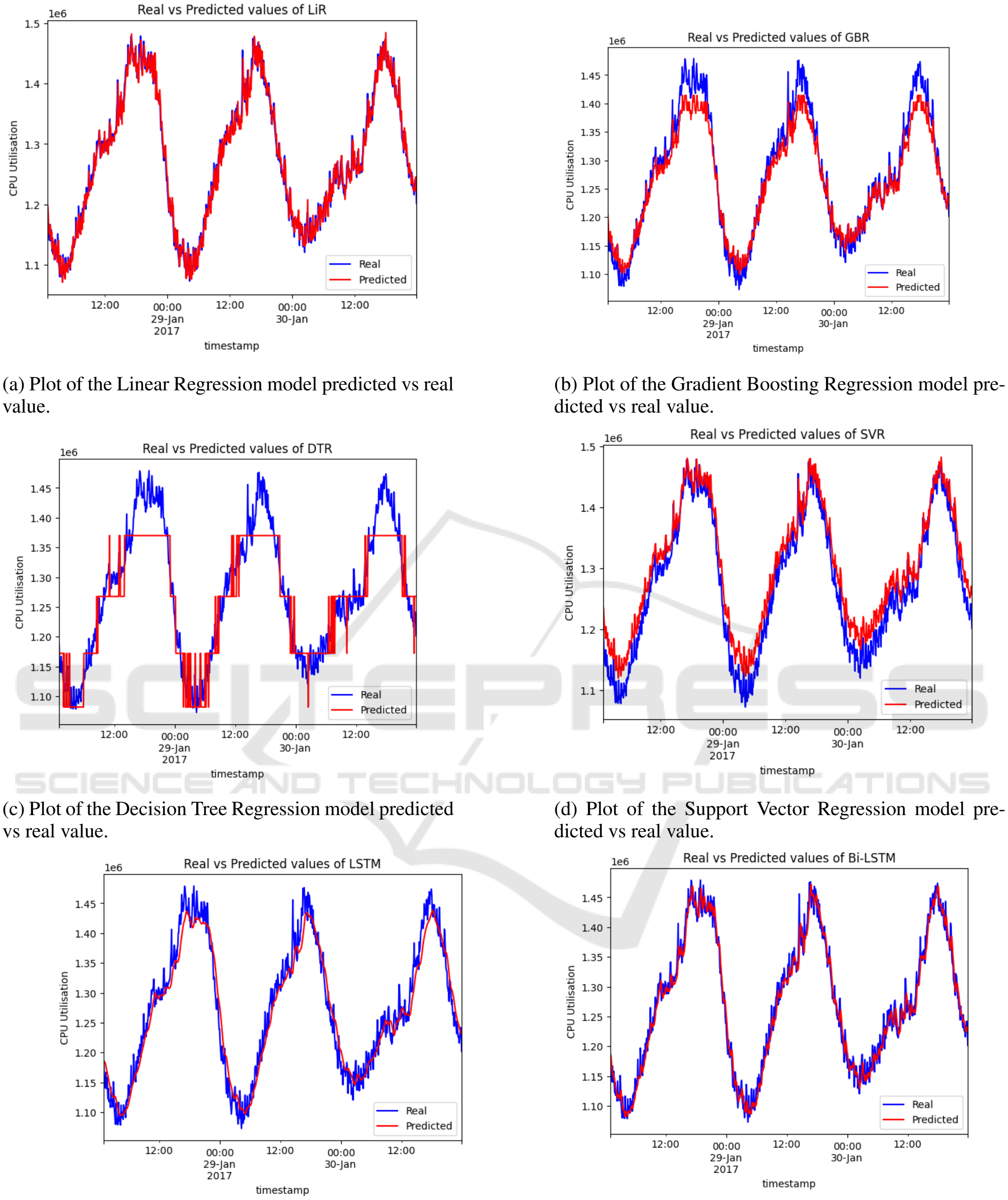
(c) Plot of the Decision Tree Regression model predicted vs real value.

(d) Plot of the Support Vector Regression model predicted vs real value.

(e) Plot of the Long Short Term Memory model predicted vs real value.

(f) Plot of the Bi-directional Long Short Term Memory model predicted vs real value.

Figure 4: Prediction of CPU Utilisation using BitBrains dataset.



(a) Plot of the Linear Regression model predicted vs real value.

(b) Plot of the Gradient Boosting Regression model predicted vs real value.

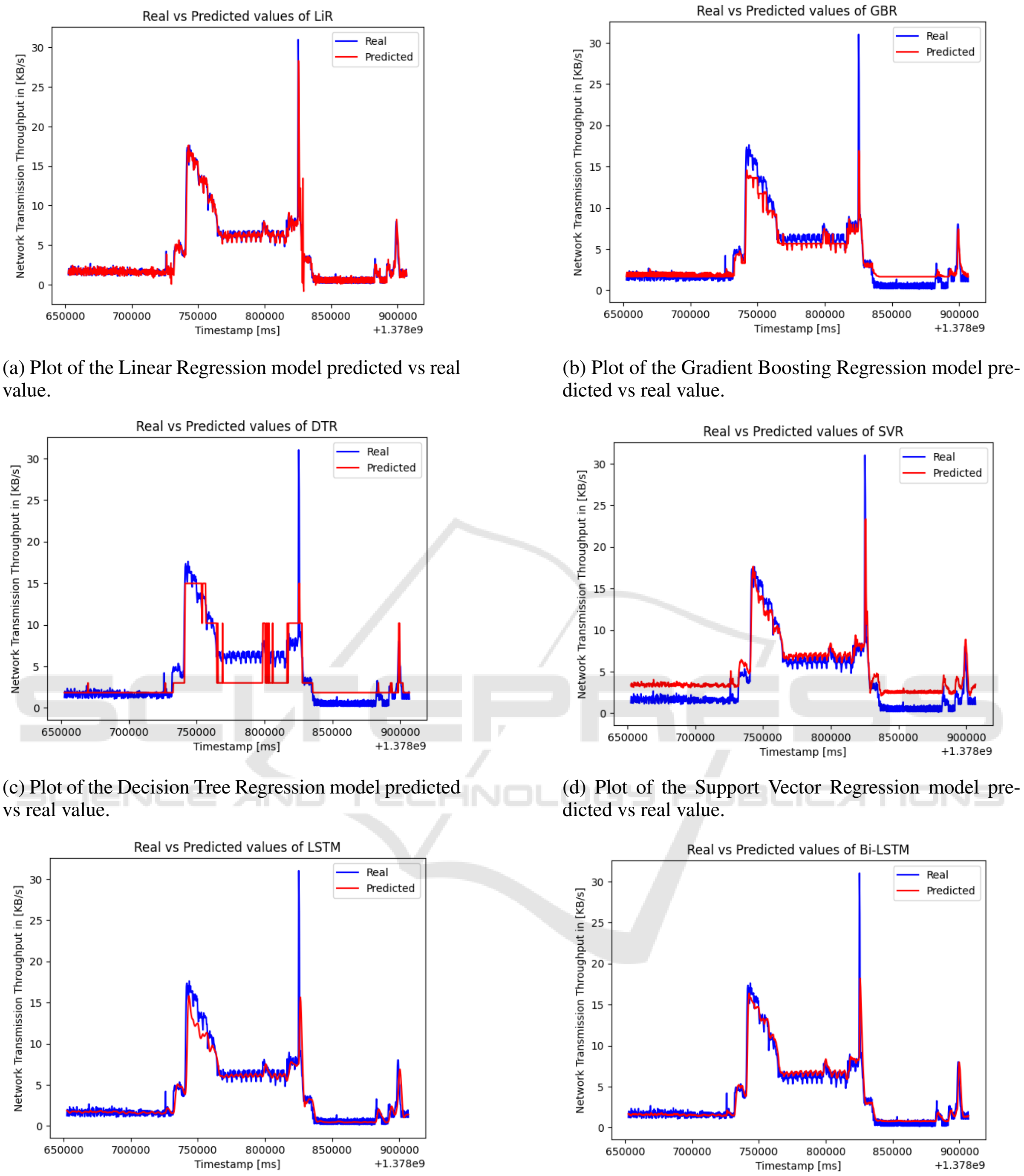
(c) Plot of the Decision Tree Regression model predicted vs real value.

(d) Plot of the Support Vector Regression model predicted vs real value.

(e) Plot of the Long Short Term Memory model predicted vs real value.

(f) Plot of the Bi-directional Long Short Term Memory model predicted vs real value.

Figure 5: Prediction of CPU Utilisation using Microsoft Azure dataset.



(a) Plot of the Linear Regression model predicted vs real value.

(b) Plot of the Gradient Boosting Regression model predicted vs real value.

(c) Plot of the Decision Tree Regression model predicted vs real value.

(d) Plot of the Support Vector Regression model predicted vs real value.

(e) Plot of the Long Short Term Memory model predicted vs real value.

(f) Plot of the Bi-directional Long Short Term Memory model predicted vs real value.

Figure 6: Prediction of Network Transmission Throughput using BitBrains dataset.

5 CONCLUSION

This study conducted a comprehensive exploration into predicting resource utilization within cloud computing frameworks through a diverse range of machine learning and deep learning models. Python programming within Google Colab was utilized alongside BitBrains and Microsoft Azure datasets, encompassing critical metrics such as CPU usage, network transmission throughput, and timestamps. The findings strongly emphasized the efficacy of the Bi-directional Long Short-Term Memory (BiLSTM) model, surpassing other machine learning algorithms in accuracy and performance. The achieved R-square values and Root Mean Square Error (RMSE) metrics highlight the BiLSTM model's exceptional predictive abilities in anticipating resource utilization, offering pivotal insights for optimizing cloud computing efficiency.

Based on this research, there are a number of interesting directions for further study. Prediction accuracy might be increased even more by investigating ensemble learning strategies to integrate different models. A more thorough grasp of resource usage patterns may be obtained by extending the dataset's reach outside BitBrains and Microsoft Azure. Further research into other real-time data aspects may improve prediction accuracy; nevertheless, improving the models' interpretability is still a crucial step towards gaining more profound understanding.

REFERENCES

Amiri, M. and Mohammad-Khanli, L. (2017). Survey on prediction models of applications for resources provisioning in cloud. *Journal of Network and Computer Applications*, 82:93–113.

Borkowski, M., Schulte, S., and Hochreiner, C. (2016). Predicting cloud resource utilization. In *2016 IEEE/ACM 9th International Conference on Utility and Cloud Computing (UCC)*, pages 37–42.

Chugh, A. (2020). Mae, mse, rmse, coefficient of determination, adjusted r-squared: Which metric is better. *Medium*. http://tiny.cc/137ivz.

Conforto, S., Zia Ullah, Q., Hassan, S., and Khan, G. M. (2017). Adaptive resource utilization prediction system for infrastructure as a service cloud. *Computational Intelligence and Neuroscience*, 2017:4873459.

Daid, R., Kumar, Y., Hu, Y.-C., and Chen, W.-L. (2021). An effective scheduling in data centres for efficient cpu usage and service level agreement fulfilment using machine learning. *Connection Science*, 33(4):954–974.

Duggan, M., Mason, K., Duggan, J., Howley, E., and Barrett, E. (2017). Predicting host cpu utilization in cloud computing using recurrent neural networks. In *2017 12th International Conference for Internet Technology and Secured Transactions (ICITST)*, pages 67–72.

Estrada, R., Valeriano, I., and Aizaga, X. (2023). Cpu usage prediction model: A simplified vm clustering approach. In *Conference on Complex, Intelligent, and Software Intensive Systems*, pages 210–221. Springer.

Kaur, G., Bala, A., and Chana, I. (2019). An intelligent regressive ensemble approach for predicting resource usage in cloud computing. *Journal of Parallel and Distributed Computing*, 123:1–12.

Khurana, S., Sharma, G., and Sharma, B. (2023). A fine tune hyper parameter gradient boosting model for cpu utilization prediction in cloud.

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

Manam, S., Moessner, K., and Asuquo, P. (2023). A machine learning approach to resource management in cloud computing environments. In *2023 IEEE AFRICON*, pages 1–6.

Mason, K., Duggan, M., Barrett, E., Duggan, J., and Howley, E. (2018). Predicting host cpu utilization in the cloud using evolutionary neural networks. *Future Generation Computer Systems*, 86:162–173.

Mehmood, T., Latif, S., and Malik, S. (2018). Prediction of cloud computing resource utilization. In *2018 15th International Conference on Smart Cities: Improving Quality of Life Using ICT & IoT (HONET-ICT)*, pages 38–42.

Morariu, C., Morariu, O., Răileanu, S., and Borangiu, T. (2020). Machine learning for predictive scheduling and resource allocation in large scale manufacturing systems. *Computers in Industry*, 120:103244.

Nääs Starberg, F. and Rooth, A. (2021). Predicting a business application's cloud server cpu utilization using the machine learning model lstm.

Shivakumar, B. R., Anupama, K. C., and Ramaiah, N. (2021). Resource utilization prediction in cloud computing using hybrid model. *International Journal of Advanced Computer Science and Applications*, 12:2021.

Wang, J., Yan, Y., and Guo, J. (2016). Research on the prediction model of cpu utilization based on arima-bp neural network. In *MATEC Web of Conferences*, volume 65, page 03009. EDP Sciences.