Neural Computing and Applications (2021) 33:10043–10055
https://doi.org/10.1007/s00521-021-05770-9

ORIGINAL ARTICLE

An efficient forecasting approach for resource utilization in cloud data center using CNN-LSTM model

Soukaina Ouhame[1] · Youssef Hadi[1] · Arif Ullah[2]

Received: 29 November 2019 / Accepted: 21 January 2021 / Published online: 10 March 2021

Abstract
Cloud computing provides different kind of services for users and provides with the help of internet. The Infrastructure as a service is a service model that provides virtual computing resources such as, networking, hardware, and storage services as needed for users. However, cloud-hosting initialization takes several minutes delay in the hardware resource allocation process. To resolve this issue, we need to predict the future amount of computing. In this paper, we propose a convolutional neural network and long short-term memory model for predicting multivariate workload which are the central processing unit, memory, and network usage. Firstly, the input data are analyzed by the vector auto regression method which filters the linear interdependencies among the multivariate data. Then, the residual data are computed and entered into the convolutional neural network layer which extracts complex features of each of the virtual machine usage components after the long short-term memory neural network, which is suitable for modeling temporal information of irregular trends in time series components in the proposed hybrid model latest activation function scaled polynomial constant unit used. The proposed model is compared with other predictive models. Based on the result, the proposed model shows the accuracy rate enhanced by approximately 3.8% to 10.9% and the error percentage rate also reduces by approximately 7% to 8.5% as compared to the other different models. It means the proposed technique improves central processing unit, memory, disk, and network usage in the network taking less amount of time due to the good predication approach compare to other models. In future research, we implement the proposed technique for VM energy section as well data predication system in cloud data center.

Keywords Multivariate prediction · CNN · LSTM · Cloud computing · Infrastructure · Vector auto regressive

1 Introduction

Cloud computing becoming the most prominent technology due to it services because it provides the services with the help of internet and work on the base of pay and gain rule these services consist of hardware and software. These applications are used in various fields of life and improve the quality of our life. Cloud computing is emerging technology due to its main property which is known as virtualization process. Different physical machines run as virtual machine. Virtual machine plays important role in cloud computing. Cloud computing becomes popular among the communities of business and researcher due to its virtualization property [1, 2]. Infrastructure as a service provides flexible and fast IT resources on demand the majority of cloud providers offer scalable services that automatically provide computer resources (such as CPU, memory, and storage). Virtualization is the process in which physical instance of single application or resource share among multiple organizations or users. This technique is done by assigning physical resources as logical form. Its play an important role in cloud computing like

✉ Soukaina Ouhame
Soukaina.ouhame@uit.ac.ma

Youssef Hadi
hadi@uit.ac.ma

Arif Ullah
arifullahms88@gamil.com

[1] Department of Computer Science, Faculty of Science, Ibn Tofail University, Kénitra, Morocco

[2] Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400 Johor, Malaysia

sharing data as application but actually with the help of virtualization user share the infrastructure services [3, 4]. Furthermore, the main elements of virtualization are known as virtual machine. A single physical resource can appear as multiple resources this process can be achieved with the help of a virtual machine.It contains matched environment for a physical computer system its run an operating system and applications. It can be implemented through with the help of software, framework, and hardware [5, 6]. However, the scaling time to initialize a new virtual machine mainly introduces a delay of several minutes. To reduce scaling time, it is important to fix the exact amount of resources in advane. Consequently, predicting VM utilization is the key solution to solve this problem [7, 8]. Resource and work load prediction are important parameter of cloud management system or system platform system. Predication process improves accuracy rate, and it directly affects the security, QoS, economical and management process which improve the performance of cloud computing. Normally load and application predication are used to describe the future behavior of resource and application on the specific aspect of the collected information base [9, 10]. One of the main issues in cloud computing are resource allocation systems in cloud computing not astonishingly, previous work has established number of solution to provide cloud resources in a competent manner. However, in order realize holistic resource provisioning models a predication of the future resource consumption of upcoming generation of resource are necessary [11, 12]. Time series is a sequence of observation which is indexed in time order depends on problems situations. It defines as a continues state process of observation where specify discrete time point are mentioned. One of the main areas of research is forecasting where future data are points are forecasted using past data points nowadays widely used in different areas of computer science [13, 14]. Cloud hosting initialization prediction is significant for improving resource allocation and utilization in cloud computing. Due to the higher variance than that in a resource system, accurate prediction remains a challenge in the cloud system. Therefore, in this paper, we are going to present modified model for improvement in resource allocation system predication in cloud data center [15]. The rest of this paper is organized as: The necessary background for the resource predication is discussed in Sect. 2. In Sect. 3: present the proposed technique for cloud resource consummation predication system with CNN-LSTM model. We also introduce several different mechanisms in this section. In Sect. 4: present the CNN model.In Sect. 5: we present the test, evaluation for the proposed models to prove the proposed model effectiveness. Section 6: present the conclusion and define some of the future works direction.For the comfort of readers, we provided a list of the most frequently used acronyms in the paper are mention in Table 1.

2 Related work

There exit several predicting resource utilization techniques in literature based on single and multiple machine learning techniques some of them are given below.

The author [16] present novel technique for cloud resource prediction which consists of functional link neural network (FLNN) and genetic algorithm (GA) used to train learning model for forecast of effectiveness. The proposed technique also consists of several mechanisms which are combined together to enable which process different resource type in the system. The performances of system good as compared to traditional system.

For data center accurately estimate resource utilization is important therefore author [17] extant approach trains a classifier based on statistical features of historical resources usage to decide the appropriate prediction model. This model use for given resource utilization system for specific time interval. The proposed system improves the delivers 6% to 8% improved in resource utilization estimation accuracy compared to baseline methods.

The author [18] intend host over/under loading detection algorithm along with new VM replacement algorithm. The author presents robust linear regression prediction model for energy aware system for VM in cloud data center. The

Table 1 List of acronyms

Acronyms	Meaning
CC	Cloud computing
VM	Virtual machine
LB	Load balancing
CNN	Convolutional neural network
LSTM	Long short-term memory
DAG	Directed acyclic graph
AIC	Akanke information criterion
BIC	Bayesian Information Criterion
HQC	Hannan-Quinn Criterion
ADF	Augmented dickey–fuller
MAE	Mean Absolute Error
RMSE	Mean Squared Error
MSE	Mean Squared Error
QoS	Quality of service
FLNN	Functional link neural network
CPU	Central processing unit

methods amend the prediction and squint toward over-prediction by adding the error to the prediction.

According to author [19] linear regression method for prediction policy for futuristic resource utilization in cloud computing. The proposed techniques improve service level agreement (SLA) due to that QoS improve in cloud data center.

The author [20] proposed a novel model for cloud auto-scaling system which is known as multivariate fuzzy LSTM (MF-LSTM). The proposed model consists of pre-processing phase in which fuzz fraction technique used for reducing flotation of monitoring data and long short-term memory (LSTM) neural network is employed to predict the resource consumption with multivariate time series data at the same time. The proposed techniques improve the per-former of the system.

According to author [21] introduces server less application analytics framework (SAAF), for (FaaS) function runtime that allows profiling workload performance and resource utilization predications. For that resign multiple regressions is used for estimate work load across hetero-geneous CPU's with different memory size and for alter-nate (FaaS) platform used.

The author [22] present novel scheduling algorithm by using directed acyclic graph (DAG) based on the prediction of tasks computation time algorithm (PTCT) for the esti-mate of makespan or network time improvement with the help of load balancing technique and reducing the expected time to compute (ETC). The proposed technique result compare with standard technique in terms of scheduling performance.

The author [23] proposed a hybrid method EEMD-RT-ARIMA, for utilization prediction of resource in cloud computing. The proposed method consists of autoregres-sive integrated moving average (ARIMA) and empirical mode decomposition (EEMD), for run test (RT), and (EEMD) used for decompose the no stationary host uti-lization sequence into relatively stable intrinsic mode function (IMF) components and a residual component to improve prediction accuracy.

In the cloud computing over and under loaded resource are share able using different technique but when it time to use unused resource and they have no access and control for their distribution so therefore the author [24], proposed multi sharing resources in hybrid cloud (MSRHC) model. The proposed model design with the help of carry forward of unused resources (CFUR) and access control for con-tributed user (ACCU) the working citer was that it shares multiple resource to multiple used and the control of all these by unused resource. Therefore, it used 97% resource utilization of cloud computing. Table 2 presents summary of related work.

Resource provisioning system involves dynamic allo-cation technique by scaling resource up and down depending on the current and future demand of user and provider. Recently, different researchers present prediction based resource allocation system for cloud computing which are mentions in Table 1 as summary. While pre-diction accuracy rate for recourse consumption is very attractive issues and challenges. The problems related with multiple processing metric (CPU, and memory usages, disk capacity and I/O, network throughput and so forth) are given less time from researcher when production working going on. The working resource likes (between CPU and memory usage, and between disk I/O and memory I/O) and their forecasting and working production is very different due to their relationship and reaction. One of the main goal of prediction of resource technique is to predict the resource consummation in an advance in cloud computing. For that resign different researcher implement and forecast models for cloud computing. Resource provisioning approaches are future divided in to group which are pre-dictive and reactive types in reactive approach measure a system based on state like the utilization of (VM, servicer, CPU) based on current task request and make decision based on that. While predictive approach aims to predict the feature of system based on behavior. For illustration based on the number of request or data packages to be used the necessary amount of resource of future prediction time is calculated. Cloud resources are released based on the predicted resource requirements and tasks are distributed among the resource. These techniques improve the per-formance the system and network life time. Designing predication based on resource allocation technique for cloud computing become attractive topic for scientists research now a days. But due to different barriers these technique face some issue which are accuracy, efficiency and network life time. For improvement of these issues, different machine learning techniques are used. So in this paper, we are going to propose a hybrid approach.

3 Proposed approach

After the study of the related work, it seem that most of the research working resource utilization taking single resource of host or more but need more accuracy efficiency in this area of study. Therefore, in our research, we choose CNN-LSTM for resource utilization prediction in cloud computing. CNN is used to remove noise and to take into account the correlation between multivariate variables, and LSTM models temporal information and maps time series into separable spaces to generate predictions. We propose a CNN-LSTM neural network combining CNN and LSTM to predict VM utilization. VM utilization is multivariate time

Table 2 Summary of related work

Predication technique	Platform	Metric	Pre processing	Prediction section	References
(FLNN) /GA	Google Trace Dataset/ cloudsim	Accuracy, Memory Prediction	Yes	Multi Section	[25]
Adaptive Prediction Models	Ali baba Data Set/Cloudsim	Data Center Resource Utilization	No	Multi Section	[26]
SLA Model	Cloudsim/Google Data set	Energy	No	One Section	[27]
Linear Regression/ UPLRegA	Cloud Analyst	Throughput, Requests, Time	No	Two section	[28]
(LSTM (MF-LSTM)	Google Trace Data/ Cloudsim	Accuracy, Effectiveness Prediction	No	Multi Section	[29]
Framework (SAAF)	Google Trace Data/ Cloudsim	Accurate, Performance Predictions	No	One Section	[30]
Directed Acyclic Graph (DAG)/ (PTCT)	Math lab 2013 B/ Google Trace Dataset/ Math lab 2013	Tasks Reduces Makespan Maximizes Resource Utilization	No	Multi Section	[31]
ARIMA	Google Trace data/ cloudsim	QoS, Short-term host Utilization Prediction	No	One Section	[31]
MSRHC model	Google Trace data/ cloudsim	Unused Resources	No	Two Section	[32]

series that is recorded over time, including spatial information among variables and irregular patterns of temporal information. Therefore, we propose a CNN-LSTM model for predicting resource usage metrics, which are CPU, Memory, and Network usage. Firstly, the input data are analyzed by the VAR regression method to filter the linear interdependencies among the multivariate data. Then, the residual data are computed and entered to the CNN layer, which extract complex features of each of the VM usage components, and after to the LSTM, which models temporal information of irregular trends in time series components and generates the predictions. Figure 1 presents the working section of proposed model.

The main contributions of this paper are as.

a. Propose a CNN and LSTM model for multivariate resource forecasting in cloud data center for cloud computing.
b. Approximation and comparison of the proposed model with different models.
c. To conduct experiments and evaluate the proposed method result using real world workload traces of GWAT-12 Bit brains service provider.

3.1 Proposed approach method

Our multivariate time series data are composed of two portions, the linear and the nonlinear portion. Thus, we can express as follows and Fig. 1 shows the schema of the proposed model. The spatial features are extracted by the CNN model then entered as inputs to the LSTM model, which is appropriate for modeling temporal information and generates the final predictions. First, we introduce some background concepts of the two models. After that, we describe the proposed model for multivariate workload prediction in the cloud [33]. Figure 2 presents the proposed approach structure of this study.

$$x_t = L_t + N_t + \varepsilon$$

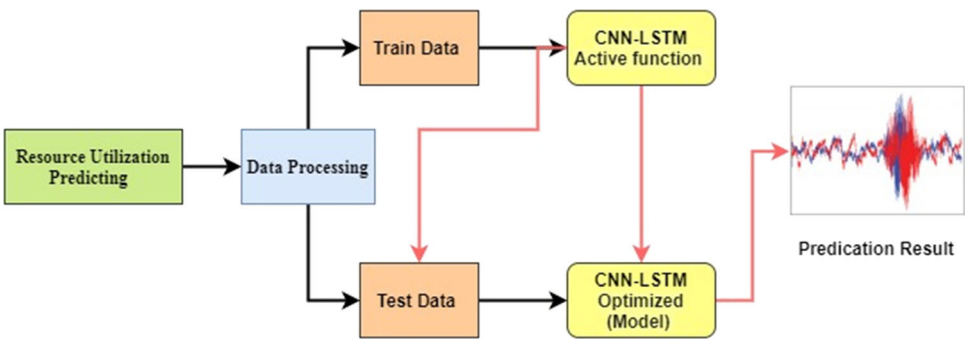


Fig. 1 Working steps of proposed model

L_t represents the linearity of data at time t, while N_t signifies nonlinearity. The $\mathcal{E}$ value is the error term. Firstly, the multivariate time series x_t are analyzed by the VAR model, which captures the linear trends. The residuals of the model (the nonlinear part N_t) contain spatial and temporal information [34, 35].

$$N_t = S + T$$

The spatial features are extracted by the CNN model then entered as inputs to the LSTM model, which is appropriate for modeling temporal information and generates the final predictions [36].

3.2 Vector autoregressive model

VAR models introduced by Sims [37] are a univariate model extension for predicting multivariate time series. The structure is that each variable is a linear function of past lags of itself and past lags of the other variables.

$$\begin{aligned} y_1(t) &= a_1 + w_{11}y_1(t-1) \\ &\quad + w_{12}y_2(t-1) + w_{13}y_3(t-1) \\ &\quad + w_{14}y_4(t-1) + e_1(t-1) \\ y_2(t) &= a_2 + w_{21}y_1(t-1) \\ &\quad + w_{22}y_2(t-1) + w_{23}y_3(t-1) \\ &\quad + w_{24}y_4(t-1) + e_2(t-1) \\ y_3(t) &= a_3 + w_{31}y_1(t-1) \\ &\quad + w_{32}y_2(t-1) + w_{33}y_3(t-1) \\ &\quad + w_{34}y_4(t-1) + e_3(t-1) \\ y_4(t) &= a_4 + w_{41}y_1(t-1) \\ &\quad + w_{42}y_2(t-1) + w_{43}y_3(t-1) \end{aligned} \tag{3}$$

where $y_1(t)$,$y_2(t)$, $y_3(t)$ and $y_4(t)$ are the CPU, memory, disk and network usage at moment t, $y_1(t-1)$, $y_2(t-1)$, $y_3(t-1)$ and $y_4(t-1)$ are the CPU, memory, disk and network usage at moment $t-1$ (here the lag value is 1). a_1, a_2, a_3 and a_4 are the constant terms, etc. are the coefficients, and e_1,e_2, e_3 and e_4 are the error terms. Before we can estimate a vicariate VAR model for the two series, we must specify the order p [38, 39].

3.3 VAR order selection

The most common approach for model order selection involves selecting a model order that minimizes one or more information criteria evaluated over a range of model orders; akanke information criterion (AIC), bayesian information criterion. (BIC) or hannan-quinn criterion (HQC). In this paper, we resolve to use the AIC metric to estimate parameters.

$$AIC = -2ln(\hat{L}) + 2k \tag{4}$$

The $ln(\hat{L})$ notation is the value of the likelihood function, and k is the degree of freedom, that is, the number of parameters used. A model that has a small AIC value is generally considered a better model. The residual values are computed and entered to the subsequent CNN-LSTM model. As the VAR model has identified the linear trend, the residual is assumed to comprise the nonlinear features.

$$x_t - L_t = N_t$$

4 CNN-LSTM model

4.1 Convolutional neural network

The convolutional neural network, or CNN for short, is a specialized type of neural network model designed for working with two-dimensional image data, although they

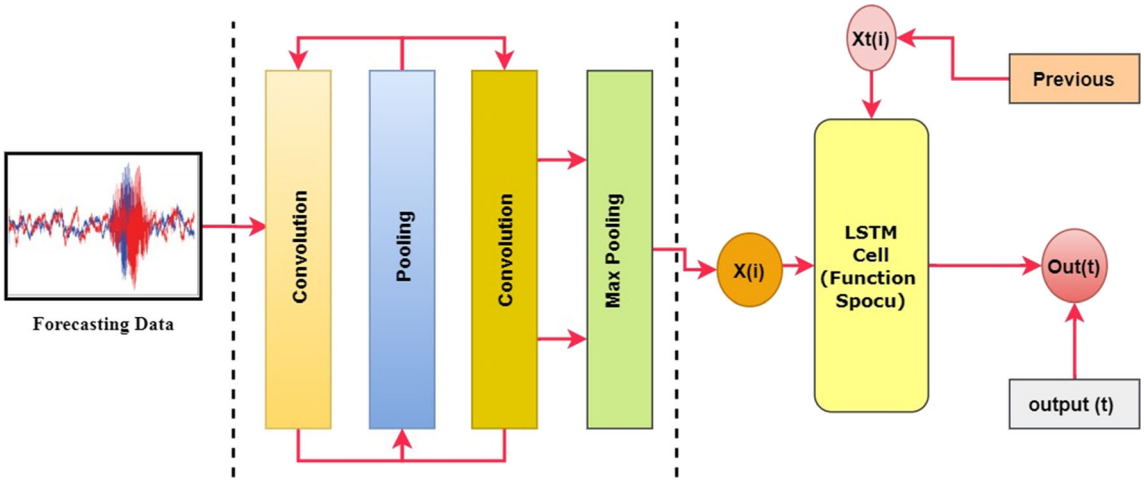


Fig. 2 Proposed model structure

can be used with one-dimensional or with three-dimensional data. In the time series forecasting problem, A 1D CNN is capable of reading across sequence input and automatically learning the salient features. A one-dimensional CNN is a CNN model having a convolutional hidden layer that operates over a 1D sequence. This is followed by a second convolutional layer in some cases, such as very long input sequences. Equation (6) is the result of the vector y_{ij}^{1} output from the first convolutional layer [40, 41]. Figure 3 shows the structure of CNN.

$$y_{ij}^{1} = \sigma\left(b_j^{1} + \sum_{m=1}^{M} w_{m,j}^{1} x_{i+m-1,j}^{0}\right)$$

where x_{ij}^{1} is the input vector, b_j^{1} represents the bias for the jth feature map, w is the weight of the kernel, m is the index value of the filter, and σ is the activation function like ReLU. Then the convolutional layer is followed by the pooling layer whose job it is to distill the output of the convolutional layer to the most salient elements. The pooling layer reduces the space size of the representation to reduce the number of parameters and network computation costs. The max-pooling used for resource usage prediction uses the maximum value from each neuron cluster in the previous layer. This also has the effect of adjusting over fitting. Equation (7) represents the operation of the max-pooling layer. T is the stride that decides how far to move the area of input data, and R is the pooling size of less than the size of the input y. Figure 4 shows the working rule of max-pooling layer.

$$p_{ij}^{1} = \max y_{r\in R}^{1} i \times T + r.j$$

The convolutional and pooling layers are followed by the LSTM layer that interprets the features extracted by the convolutional part of the model. A flatten layer is used between the convolutional layers and the LSTM layer to reduce the feature maps to a single one-dimensional vector [42, 43].

4.2 Long short-term memory neural networks

LSTM, which is a lower layer of CNN-LSTM, stores time information about important characteristics of power demand extracted through CNN. LSTM provides a solution by preserving long-term memory by consolidating memory units that can update the previous hidden state. This function makes it easy to understand temporal relationships on a long-term sequence. The output values from the previous CNN layer are passed to the gate units. The LSTM network is well suited for predicting power demand by addressing explosive and vanishing gradient problems. The LSTM cell comprises four interactive neural networks, each representing the forget gate, input gate, input candidate gate, and the output gate. The forget gate outputs a vector whose element values are between zero and one. It serves as a forgetter that is multiplied to the cell state C_{t-1} from the former time step to drop values that are not needed and keep those that are necessary for the prediction [44].

$$f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f)$$

The σ function, also denoted with the same symbol in Fig. 5, is the logistic function, often called the spocu. It is the activation function that enables nonlinear capabilities for the model.

$$\sigma(X) = \frac{1}{1 - e^{-x}}$$

In the next phase, the input gate and the input candidate gate operate together to render the new cell state C_t, which will be passed on to the next time step as the renewed cell state. The input gate uses the spocu as activation function and the input candidate utilizes the hyperbolic tangent,

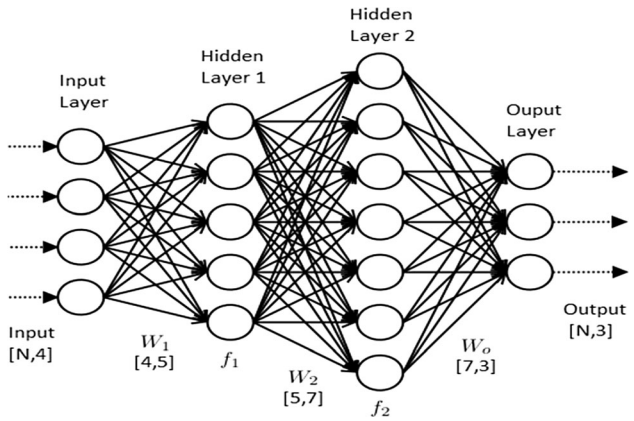


Fig. 3 CNN structure

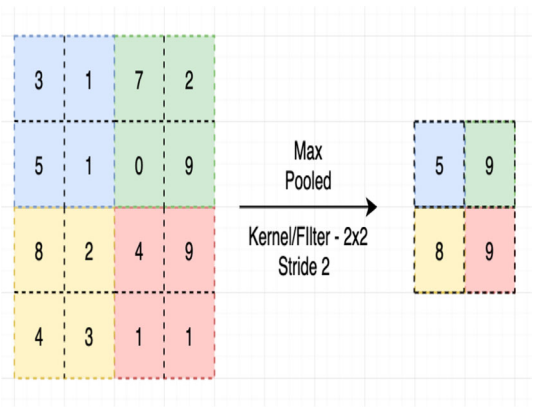


Fig. 4 Pooling layer structure

each outputting i_t and C'_t. The i_t selects, which feature in C' should be reflected in to the new cell state C_t.

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

$$C_t = \tanh(W_c, [h_{t-1,} x_t] + b_c)$$

The tanh function in Fig. 6 is the hyperbolic tangent. Unlike the spocu, which renders value between zero and one, the hyperbolic tangent outputs value between − 1 and1 [45].

$$\tanh(X) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

Finally, the output gate decides what values are to be selected, combining o_t with the tanh applied state C_t as output h_t. The new cell state is a combination of the forget gate applied former cell state C_{t-1} and the new tanh applied state C_t .

$$o_t = \sigma(W_o, [h_{t-1}, x_t] + b_0)$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot C'_t$$

$$h_t = o_t \cdot \tanh(C_t)$$

The cell state C_t and h_t output will be passed to the next time step, and will go through a same process [46]. Figure 7 presents the Pseudo-code of algorithm 1.

5 The proposed CNN-LSTM model

The inputs of the algorithm are the CPU, memory, network and disk usage time series, which are formed using the historical data of the workload. The stationary of each time series is checked by using the augmented dickey–fuller (ADF) test; if they are not stationary, we differentiate them.

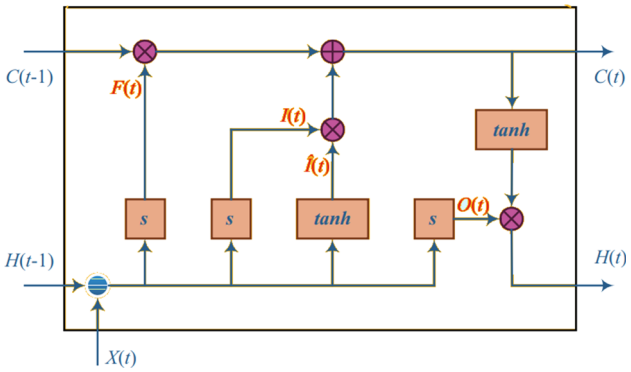
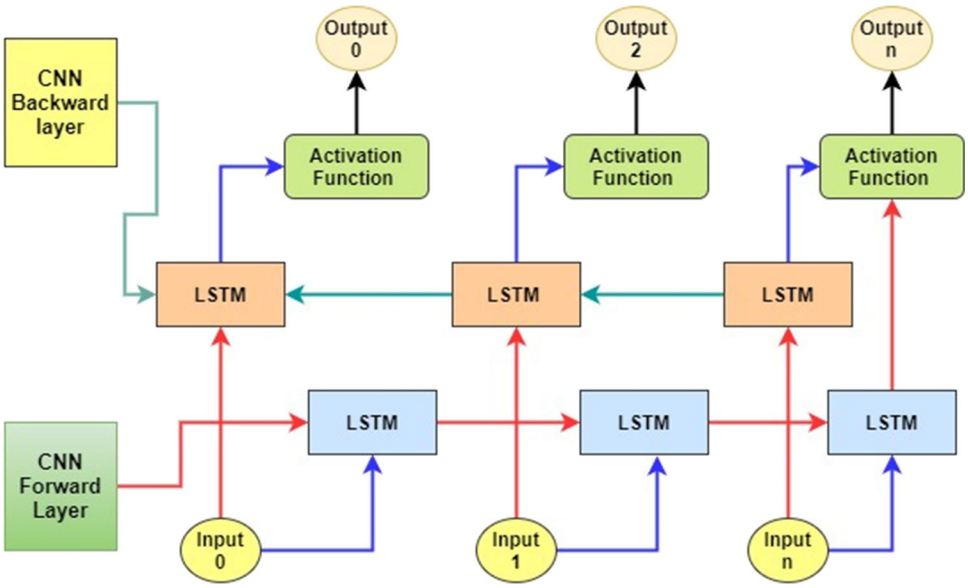


Fig. 6 Structure memory cell

In the following step, we select the lag order with the least value for model fitting and then we compute the residual values. Figure 8 presents the Pseudo-code of algorithm 2..

The CNN-LSTM prediction algorithm works in four main phases: data preprocessing, fixing model parameters, model fitting and estimation, and model prediction. As cited before the residual values calculated by the algorithm1 are entered to the CNN-LSTM model. We use four time steps; every sample is split into a pair of subsequences. The CNN model can interpret every subsequence and therefore the LSTM can piece along the interpretations from the subsequences. As such, we will split every sample into two sub sequences of two times per subsequence. The CNN is defined to expect two time steps per subsequence with four options. The whole CNN model is then wrapped in time distributed wrapper layers so it is often applied to every subsequence within the sample. The results are then interpreted by the LSTM layer, that uses fifty neurons or blocks, and eventually the dense layer outputs the

Fig. 5 Activation function



Algorithm 1: **VAR model fitting algorithm**	
Input:	Y_1: CPU usage time series
Y_2:	Memory usage time series
Y_3:	Network usage time series
Y_4:	Disk usage time series
P_{max}:	The max lags order
Output:	Residual: residual values of the two series
1.	**If** Y_1, Y_2, Y_3 and Y_4 are not stationary
2	Y_1=Difference (Y_1)
3.	Y_2=Difference (Y_2)
4.	Y_3=Difference (Y_3)
5.	Y_4=Difference (Y_4)
6.	**End If**
7.	Y= concatenate (Y_1, Y_2, Y_3, Y_4)
8.	Train, Test = divide (Y, 0.7)
9.	Select_order (maxlags = P_{max})
10.	P_lag = order lag with least AIC value
11.	Mfit = fit VAR (P_lag)
12.	**for all** data **in** Test **do**
13.	Residual = Test - predict (Train, Mfit)
14.	**Return** Residual

Fig. 7 Pseudocode of algorithm 1

prediction. The scaled polynomial constant unit (SPOCU) activation function is used for CNN layer and LSTM blocks.

$$f(x) = x^{+} = \max(0, x) \tag{16}$$

Scaled polynomial constant unit SPOCU is activation function define by (Kise) in 2020.The SPOCU activation function is given below.

$$s(x) = \alpha h\left(\frac{x}{r} + \beta\right) - \alpha h(\beta)$$

where $\beta \in (0,1)$, $\alpha, r > 0$ *and*

$$h(x) = \begin{Bmatrix} r(c), & x \geq c, \\ r(x), & x \in [0, c) \\ o & x < 0, \end{Bmatrix} \tag{17}$$

With $r(x) = r^3\left(x^5 - 2^{x^4} + 2\right)$ and $1 \leq c < \infty$. we admit c goes to infinity with $r(c) \to \infty$ cleary s is continuous $s(0) = 0$ *and* $\hat{s}(x) = \left(\frac{x}{r} + B\right)$. Notice that for c = 1 one has $h(1+) = h(1-) = 0$ and $h(0-) = h(0-) = 0$ which implies that s is continuous too. (This is not true for the second derivative) For c = 1 the range s is the

$$H_{s=}[s(-\beta r), s(1-\beta)r] = [-\alpha r(\beta), \alpha(1 - r(\beta))], r(B) \in [0, 1]$$

[47, 48]. Figure 6 presents activation function rule how it work.

Where x is the input to a neuron the problem of vanishing gradient can be greatly reduced using the ReLU family of activation functions. ADAM optimization algorithm is used for stochastic gradient descent for model training. The network is trained for 100 epochs with batch size of 1.

Algorithm 2: **CNN-LSTM model training algorithm**	
Input:	Residual: Residual values of the VAR model
	N_step:The lag step between each input and output
Output:	TrainPred, testPred: The predicted train and test data of the multivariate time series. {Phase1: Data preprocessing}
1	Normalize Residual data Convert into input/output with the percentage of 80%
2	Train_cl, test_cl = divide (Residual, 0.75)
3	X_train, y_train = split (train_cl, N_step)
4	X_test, y_test = split (test_cl, N_step)
5	Reshape X_train et X_test into (samples, subsequences, time steps, features)
6	**Define** model
7	**Add** TimeDistributed(Conv1D (filters = 64, kernel_size=1, activation = 'relu', input_shape = (None, N_steps, n_features)))
8	**add** TimeDistributed(MaxPooling1D (pool_size=2))
9	**add** TimeDistributed (flatten())
10	**add** LSTM (units = 50, activation = 'relu')
11	**add** Dense (n_features = 4) {Phase3: Model fitting & estimation}
12	**Repeat**
13	**Forward propagate** model **with** X_train
14	**Forward propagate** model **with** X_train
15	**Update** model parameters
16	MSE, MAE = evaluate_model (X_train, y_train)
17	**If** MSE converged**:**
18	**End Repeat**
19	MSEt, MAEt =evaluate_model (X_test, y_test){Phase4: Model prediction}
20	TrainPred = predict (X_train)
21	TestPred = predict (X_test)
22	**return** trainPred, testPred

Fig. 8 Pseudo-code of algorithm 2

5.1 Experimental evaluations

The current section presents the experimental evaluation of the proposed method, including, dataset collection,

experimental results and the evaluation of the efficiency of the proposed method.

5.2 Experiment dataset

The dataset contains the performance metrics of 1750 virtual machines in the Bit-brains distributed data center. (http://gwa.ewi.tudelft.nl/datasets/gwa-t-12-bitbrains). We choose the first trace; fast storage: The trace consists of 1250 virtual machines connected to storage area network (SAN) devices. Each file consists of a set of lines; each line represents an observation of the performance metrics of a virtual machine every 300 ms since 2018-01-01. We have selected 4000 observations for our workload prediction model.

5.3 Analysis of the results

The resulting time series Y1 and Y2 are stationary then we do not need to perform a difference operation. In the first step, 70% of the multivariate time series Y is used for training and 30% for testing as indicated in the algorithm 2. By choosing $P_{\max} = 4$ as the max lag order we have the following result table which are mention in Table 3.

Where BIC stands for bayesian information criterion, which also estimates the quality of a model. As the third lag order has the least value of AIC then it will be used for fitting the VAR model.In the second step, the residuals are calculated, 75% of data are used for training and 25% for testing propose. Figures 9, 10 and 11 show the result of CPU, RM and memory usage in the system are mention below.

For evaluation process of the proposed hybrid model these parameters are used which is Mean Squared Error (MSE), Root Mean Squared Error (RMSE) values and the Mean Absolute Error (MAE) values of the prediction were calculated.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2 \tag{18}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - y_i'\right|$$

Table 3 VAR model order selection

	AIC	BIC
0	38.60	36.61
1	36.31	36.32
2	36.05	36.08
3	36.00*	36.04*

where y_i presents the output value and y_i' presents the predicted value the MSE learning curve of both of the train and test data are close to each other, the same for the MAE learning curve (Figs. 12 and 13).The MSE values of train and test data have small variations, which means the model has been generalized adequately which are present Table 4.

Table 4 presents the result of test which are used for simulation process.

The result is present in Figs. 14, 15, and 16 performing simulation on the dataset show the comparison of actual load with prediction load present and Table 5 shows the important of feature show prediction. In this paper, bit-brains data set are used for evaluation of resource predication where different tests are performed. We calculate the consumption time, AIC and BIC method as highlighted in Table 4 we can see that the proposed method has less time for Computation time for implementation and forecasting. Runtime of proposed hybrid CNN- LSTM model compare with well knows techniques which are ARIMA-LSTM, VAR-GRU and VAR-MLP models. Based on Tables 5 and 6 the runtime of proposed model is batter then other models due to state of art time serious and latest activate function used for predication in this model so the proposed modes result are better.

Figures 9, 10, 11, 12, 13, 14, 15, and 16 present the run time usage of parameter (i-e) C.P.U, RAM, and Network for their result it can be observed that scale pattern is significant on three parameters and the proposed hybrid algorithm ratio is batter scale then actual parameter. Therefore, we can say that our proposed hybrid algorithm results are batter then the three other standard models. Based on Tables 4, 5 and 6 the proposed model indicates lower values of MSE and MAE compared to the ARIMA-LSTM, the VAR-GRU and the VAR-MLP models. The RMSE value is far lesser in ARIMA-LSTM model. However, the proposed model still additionally shows inferior

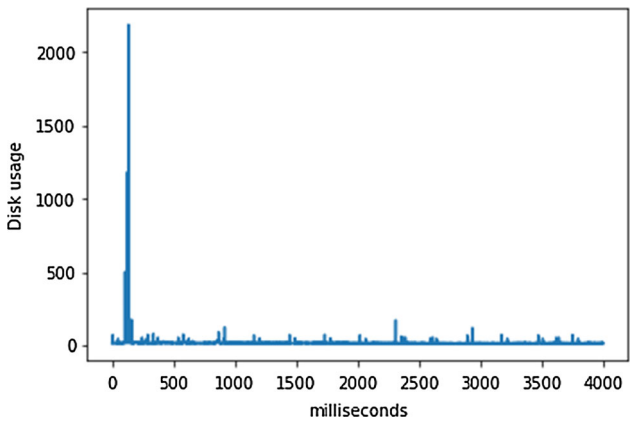


Fig. 9 RM/Disk usage by milliseconds

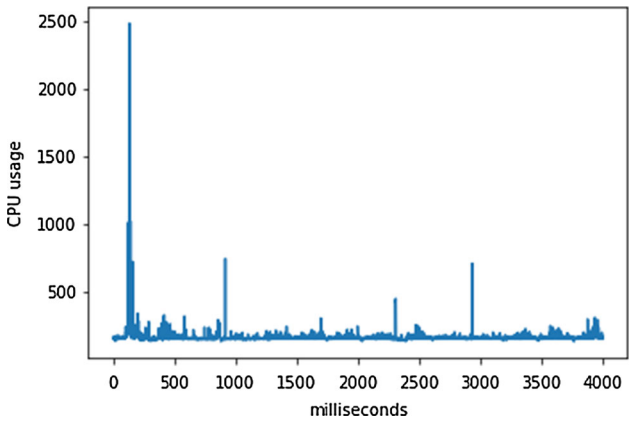


Fig. 10 CPU usage by milliseconds

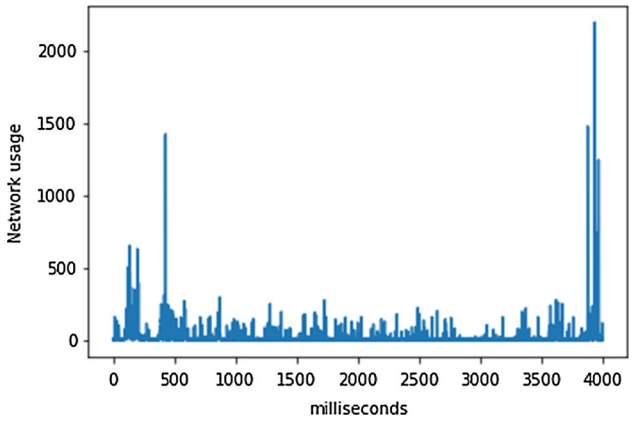


Fig. 11 Network usage by milliseconds

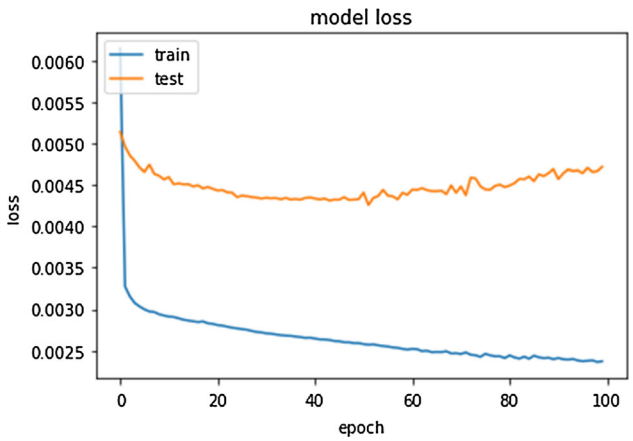


Fig. 12 MSE learning curve

values of RMSE compared to the remaining models. As mention in this study that all those VM that are considered as the MPE are below then 30% which means the result of predication performance as considering with dynamic

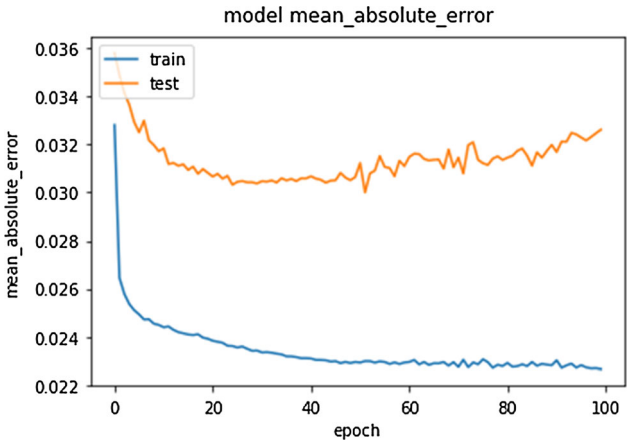


Fig. 13 MSE learning curve

Table 4 CNN-LSTM model performance results of test data

Method	Evaluation metrics		
	RMSE	MSE	MAE
VAR-MLP	0.3446	0.5870	0.03800
VAR-GRU	0.3295	0.5740	0.3575
ARIMA-LSTM	0.3111	0.0357	0.3665
Proposed model	0.3193	0.5650	0.3469

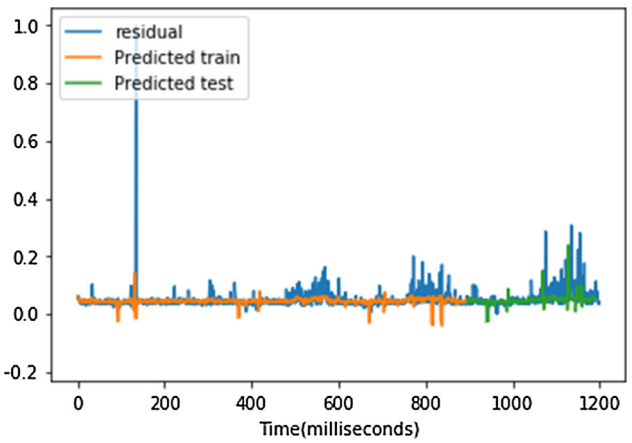


Fig.14 Predicted train and test data of CPU use

nature if the cloud. Where the MAPE for CPU on VM1 to VM12 are below then 15% which means the proposed hybrid algorithm result are good and impressive which are mention in Tables 4 and 5. While MAE for predication of memory shows that good result and value of memory and it measurement purpose millisecond and Kilobytes are used. From Figs. 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 and 16 and Tables 3, 4, 5, and 6 show the model performs from

Table 5 Run time of all models

Parameters	RMSE	MSE	MAE
CPU Utilization	0.03765	0.299769	0.137823
Memory Utilization	0.000004	0.008671	0.002587
VM Utilization	0.1838898	1.116114	0.733408
Network Utilization	0.1938893	1.216114	0.933412

Table 6 Run time of all models

Method	Evaluation metrics		
	RMSE	MSE	MAE
VAR-MLP	0.03765	0.299769	0.137823
VAR-GRU	0.000004	0.008671	0.002587
ARIMA-LSTM	0.03765	0.008023	0.002536
Proposed model	0.1838898	1.116114	0.733408

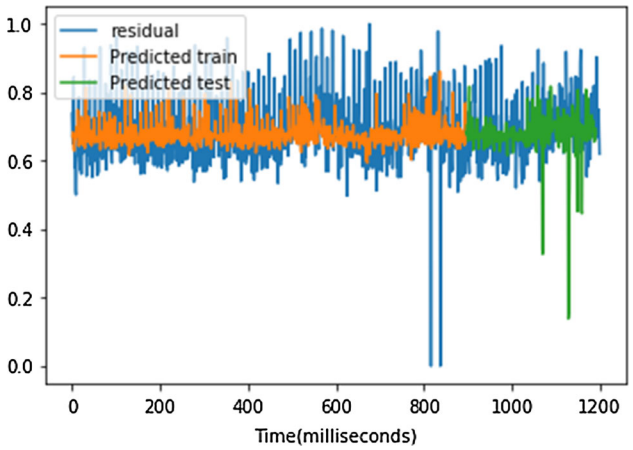


Fig.15 Predicted train and test data of RM

differently on the different VM, memory and CPU all the result show proper improvement in predication of resource utilization in cloud data center (Figs. 17, 18, 19).

6 Conclusions

An important feature of cloud computing is the ability to determine allocation of resource and application based on actual usage. However, for resource allocation operation required start-up time. For that reason, it needs plan in advance the amount of resource needed for future. For that reason in this paper we proposed hybrid CNN-LSTM

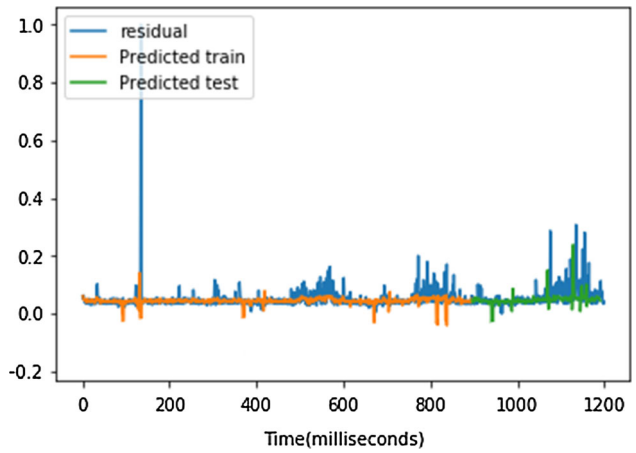


Fig. 16 Predicted train and test data network

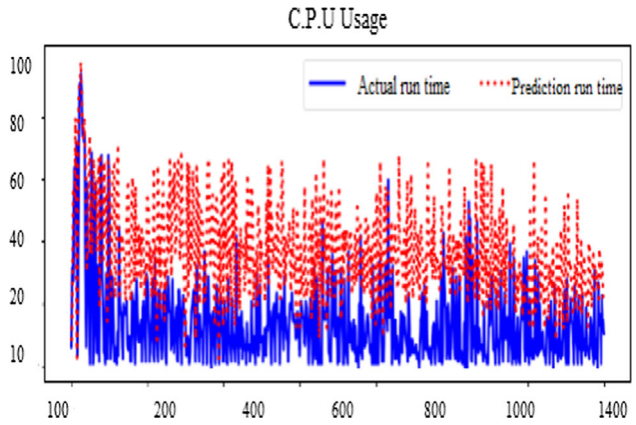


Fig. 17 C.P.U Run time

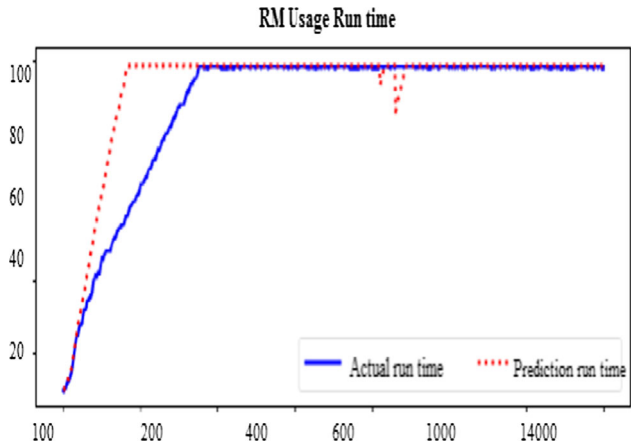


Fig. 18 Runtime of RM

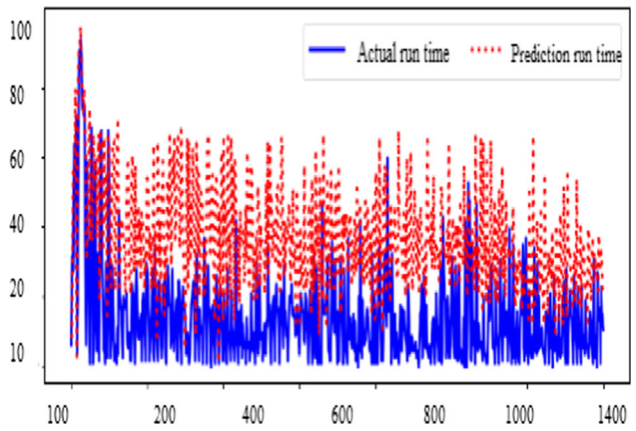


Fig. 19 Runtime of Network

model for multivariate workload prediction in an attempt to extract complex features of the VM usage components, then model temporal information of irregular trends in the time series components. The proposed approach was tested using actual data from Bit brains data. The results are positive and show that the proposed method is more effective and accuracy in CPU, RM and network utilization and prediction system than the other predictive models. The main aim of our future work is that we try to implement the proposed technique for VM replacement section and as well as load prediction in cloud datacenter taking different parameter in upcoming research.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest. We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

References

1. Zhu Y, Zhang W, Chen Y, Gao H (2019) A novel approach to workload prediction using attention-based LSTM encoder-decoder network in cloud environment. EURASIP J WirelCommunNetw 2019(1):274
2. Ullah A, Nawi NM (2020) Enhancing the dynamic load balancing technique for cloud computing using HBATAABC algorithm. Int J Model Simul Sci Comput 2050041
3. Ullah A (2019) Artificial bee colony algorithm used for load balancing in cloud computing. IAES Int J ArtifIntell 8(2):156
4. Tabrizchi, H., & Rafsanjani, M. K. (2020). A survey on security challenges in cloud computing: issues, threats, and solutions. J Supercomput, 1–40
5. Song G, Wang Z, Han F, Ding S, Gu X (2020) Music auto-tagging using scattering transform and convolutional neural network with self-attentionn. Appl Soft Comput 106702
6. Singh M, Kumar R, Chana I (2020) A forefront to machine translation technology: deployment on the cloud as a service to enhance QoS parameters. Soft Comput
7. Silvestrini A, Veredas D (2008) Temporal aggregation of univariate and multivariate time series models: a survey. J Econ Surveys 22(3):458–497
8. Shah, N. F., & Kumar, P. (2018). A comparative analysis of various spam classifications. In: Progress in intelligent computing techniques: theory, practice, and applications (pp 265–271). Springer, Singapore.
9. Saeedi A, Saeedi M, Maghsoudi A, Shalbaf A (2020) Major depressive disorder diagnosis based on effective connectivity in EEG signals: a convolutional neural network and long short-term memory approach. Cognit Neurodyn, 1
10. Reshmi R, Saravanan DS (2020) Load prediction using (DoG–ALMS) for resource allocation based on IFP soft computing approach in cloud computing. Soft Comput 1–9
11. Rana N, Latiff MSA, Abdulhamid SIM, Chiroma H (2020) Whale optimization algorithm: a systematic review of contemporary applications, modifications and developments. Neural Comput Appl, 1–33
12. Ralha CG, Mendes AH, Laranjeira LA, Araújo AP, Melo AC (2019) Multiagent system for dynamic resource provisioning in cloud computing platforms. Future GenerComputSyst 94:80–96
13. Phan TD, Zincir-Heywood N (2019) User identification via neural network based language models. Int J NetwManag 29(3):e2049
14. Nicanor LD, Aguirre HO, Moreno VL (2020) An assessment model to establish the use of services resources in a cloud computing scenario. In: High performance vision intelligence (pp 83–100). Springer, Singapore
15. Manvi SS, Shyam GK (2014) Resource management for Infrastructure as a Service (IaaS) in cloud computing: a survey. J NetwComputAppl 41:424–440
16. Malladi RK, Dheeriya PL (2020) Time series analysis of Cryptocurrency returns and volatilities. J Econ Finance, 1–20
17. Ma A, Gao Y, Huang L, Zhang B (2019) Improved differential search algorithm based dynamic resource allocation approach for cloud application. Neural ComputAppl 31(8):3431–3442
18. Liu R, Ye Y, Hu N, Chen H, Wang X (2019) Classified prediction model of rockburst using rough sets-normal cloud. Neural ComputAppl 31(12):8185–8193
19. Kisel'ák J, Lu Y, Švihra J, Szépe P, Stehlík M (2020) "SPOCU": scaled polynomial constant unit activation function. Neural Comput Appl, 1–17
20. Kholidy HA (2020) An intelligent swarm based prediction approach for predicting cloud computing user resource needs. Comput Commun, 151: 133–144
21. Jauro F, Chiroma H, Gital AY, Almutairi M, Shafi'i MA, Abawajy JH (2020) Deep learning architectures in emerging cloud computing architectures: recent development, challenges and next research trend. Appl Soft Comput 96: 106582
22. Iqbal W, Berral JL, Erradi A, Carrera D (2019) Adaptive prediction models for data center resources utilization estimation. IEEE Trans NetwServManag 16(4):1681–1693
23. Iqbal W, Berral JL, Carrera D (2020) Adaptive sliding windows for improved estimation of data center resource utilization. Future GenerComputSyst 104:212–224
24. Imdoukh M, Ahmad I, Alfailakawi MG (2019) Machine learning-based auto-scaling for containerized applications. Neural Comput Appl, 1–16
25. Hong CH, Varghese B (2019) Resource management in fog/edge computing: a survey on architectures, infrastructure, and algorithms. ACM ComputSurv (CSUR) 52(5):1–37
26. Hermida JRF, Villa RS, Seco GV, Pérez JME (2003) Evaluation of what parents know about their children's drug use and how

they perceive the most common family risk factors. J Drug Educ 33(3):337–353
27. Gupta S, Dileep AD, Gonsalves TA (2020) Online sparse BLSTM models for resource usage prediction in cloud datacentres. IEEE Trans Netw Serv Manag
28. Gopinath MP, Tamizharasi GS, Kavisankar L, Sathyaraj R, Karthi S, Aarthy SL, Balamurugan B (2019) A secure cloud-based solution for real-time monitoring and management of Internet of underwater things (IOUT). Neural ComputAppl 31(1):293–308
29. El Kafhali S, El Mir I, Salah K, Hanini M (2020) Dynamic scalability model for containerized cloud services. Arab J Sci Eng, 1–16
30. Damaševičius R, Sidekerskienė T (2020) Short time prediction of cloud server round-trip time using a hybrid neuro-fuzzy network. J Artif Intell Syst 2(1): 133–148
31. Canal R, Hernandez C, Tornero R, Cilardo A, Massari G, Reghenzani F, PiĄtek W (2020) Predictive reliability and fault management in exascale systems: State of the art and perspectives. ACM ComputSurv (CSUR) 53(5):1–32
32. Brave SA, Butters RA, Justiniano A (2019) Forecasting economic activity with mixed frequency BVARs. Int J Forecast 35(4):1692–1707
33. Baker LA (2020) Fidgetin-Like 2: a novel negative regulator of axonal growth and a promising therapeutic target for promoting nerve regeneration (Doctoral dissertation, Albert Einstein College of Medicine)
34. Ullah A, Nawi NM, Aamir M, Shazad A, Faisal SN (2019) An improved multi-layer cooperation routing in visual sensor network for energy minimization. Int J Adv Sci Eng Inf Technol, 9(2): 664–670, 2019. [Online]. Available: https://doi.org/10.18517/ijaseit.9.2.2957
35. Akça E, Yozgatlıgil C (2020) Mutual information model selection algorithm for time series. J Appl Stat, 1–16
36. Adi E, Anwar A, Baig Z, Zeadally S (2020) Machine learning and data analytics for the IoT. Neural Comput Appl, 1–29
37. Abdullah L, Li H, Al-Jamali S, Al-Badwi A, Ruan C (2020) Predicting multi-attribute host resource utilization using support vector regression technique. IEEE Access 8: 66048–66067
38. Sodhro AH, Malokani AS, Sodhro GH, Muzammal M, Zongwei L (2020) An adaptive QoS computation for medical data processing in intelligent healthcare applications. Neural ComputAppl 32(3):723–734
39. Sodhro AH, Pirbhulal S, de Albuquerque VHC (2019) Artificial intelligence-driven mechanism for edge computing-based industrial applications. IEEE Trans IndInf 15(7):4235–4243
40. Sodhro AH, Luo Z, Sodhro GH, Muzamal M, Rodrigues JJ, de Albuquerque VHC (2019) Artificial intelligence based QoS optimization for multimedia communication in IoV systems. Future GenerComputSyst 95:667–680
41. Khan SU, Baik R (2020) MPPIF-Net: identification of plasmodium falciparum parasite mitochondrial proteins using deep features with multilayer bi-directional LSTM. Processes 8(6):725
42. Ullah A, Nawi NM, Arifianto A, Ahmed I, Aamir M, Khan SN. Real-time wheat classification system for selective herbicides using broad wheat estimation in deep neural network
43. Gupta N, Jalal AS (2019) Integration of textual cues for fine-grained image captioning using deep CNN and LSTM. Neural Comput Appl 1–10
44. Jin Z, Yang Y, Liu Y (2019) Stock closing price prediction based on sentiment analysis and LSTM. Neural Comput Appl, 1–17
45. Ouhame S, Hadi Y, Arifullah A (2020) A hybrid grey wolf optimizer and artificial bee colony algorithm used for improvement in resource allocation system for cloud technology
46. Umar S, Baseer S (2016) Perception of cloud computing in universities of Peshawar, Pakistan. In: 2016 Sixth international conference on innovative computing technology (INTECH) (pp 87–91). IEEE
47. Ouhame S, Hadi Y (2019) Multivariate workload prediction using Vector Autoregressive and Stacked LSTM models. In Proceedings of the New Challenges in Data Sciences: Acts of the Second Conference of the Moroccan Classification Society (pp. 1–7)
48. Prakash RG, Shankar R, Duraisamy S (2020) FUPA: future utilization prediction algorithm based load balancing scheme for optimal VM migration in cloud computing. In: 2020 Fourth international conference on inventive systems and control (ICISC) (pp. 638–644). IEEE