National College of Ireland

A Proactive Mechanism To Improve Workload Prediction For Cloud Services Using Machine Learning

MSc Research Project
Cloud Computing

Sumedh Gursale
Student ID: x18208592

School of Computing
National College of Ireland

Supervisor: Manuel Tova-Izquierdo

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Sumedh Gursale
Student ID:	x18208592
Programme:	MSc in Cloud Comuting
Year:	2020
Module:	MSc Research Project
Supervisor:	Manuel Tova-Izquierdo
Submission Due Date:	17/08/2020
Project Title:	A Proactive Mechanism To Improve Workload Prediction For Cloud Services Using Machine Learning
Word Count:	6454
Page Count:	19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

I agree to an electronic copy of my thesis being made publicly available on TRAP the National College of Ireland's Institutional Repository for consultation.

Signature:	
Date:	17th August 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Proactive Mechanism To Improve Workload Prediction For Cloud Services Using Machine Learning

Sumedh Gursale
x18208592

Abstract

Service elasticity is an important enabler of cloud computing. It is nothing but the ability to adapt the system to inconsistent changes in workload by dynamic provisioning and de-provisioning services so that the unused resources suit the current demand at all times. This requirement of users is fulfilled by almost all cloud service providers. However, one of the major challenges cloud service providers face is the efficient resource allocation as per the demand changes and maintaining the quality of services (Qos) as per service level agreement (SLA's). The service providers don't fulfill the SLA and reason being is an unavailability of demand for workload that could lead to downtime because of heavy traffic across the network and to avoid this all cloud service providers offers standard solution of over-provisioning to support peak load and guarantee of maintaining QoS over lifetime of operation. This causes in resource consumption and is not cost-effective as the machine stays idle most of the time and contribute to greater power utilization. This paper focuses on implementing a hybrid prediction approach based on machine learning techniques. In first stage, the focus is to break down or split time series data input signal into two parts. Secondly to predict low frequency components, Support Vector Regression (SVR) is used on first part. Second part of time series is more likely noise and has high frequency, so for the prediction Artificial Neural Network (ANN) is used. Lastly, an inverse wavelet transformation is implemented to reconstruct these samples to original signal from two multi-scale predictions in order to achieve accurate workload prediction. Based on the overall results, the proposed approach has a relatively better predictability compared with competitive approach. The results are evaluated on two models and proposed model (A hybrid SVR + ANN) has outperformed the other model.

1 Introduction

Over the last couple of years cloud computing (CC) has developed tremendously. Cloud computing is an Internet-based measure that provides with information, software products and software service that is shared with an idea of pay as you use. While cloud computing has so many advantages, it also has certain issues which need to be addressed. Resource management is important and vast mechanism. The issues in resource management should be addressed on priority. Resource management is the process for assigning an on-demand services including storage services, servers where customers run the projects, computer networks, computer systems, virtual machines to wide variety of cloud applications (Madni et al.; 2017). By this, all services are shared among infrastructure service providers and cloud users. Cloud service providers effectively provides cloud customers all the solutions and services under the pact of Service Level Agreements (SLAs). To seamlessly perform resource mechanism without wasting extra computational cost, it is important to know precise demand to improve the accuracy of resource allocation process to allocate the resources as per the demand of the customer within the scope of the Service Level Agreements (Madni et al.; 2017).

The three main key providers associated with cloud computing are 1) Software as a Service (Saas) 2) Platform as a service (PaaS) 3) Infrastructure as a service (IaaS). Saas environment is conducive for many software applications which are offered to customers. Large number of customers can access simultaneously each applications without interfering each other. PaaS environment provide customers, a platform for building software and services. Physical resources or in some instances virtual computing resources are provided to customers in Infrastructure as a service. In IaaS environment, cloud computing efficiency is closely related to management of resources (Moreno-Vozmediano et al.; 2019). Cloud computing efficiency can be greatly enhanced by proactively predicting workloads and accordingly managing cloud resources. Computing resources are deployed as per the workload forecast and thus automatically scale up or down to match workload. Method of work load estimation is influential in defining efficiency of resource scaling. Statistical Method and Machine Learning Method are the two approaches for forecasting workloads. Comparison between present workload and related past workload is made and then workload is predicted in statistical method. In Machine Learning method, historical data related to prior workload is applied to design and predict the potential workload. Earlier, it was difficult to estimate workload for long term using statistical method. However researchers solved the issue with help of machine learning. Few approaches associated with machine learning taken from literature that have been applied to forecast workload are - support vector machine (SVM), regression tree (Messias et al.; 2015) and artificial neural network (ANN) (Kumar and Singh; 2018)

Experts are trying to concentrate on exploring new and novel solutions for tackling the problems efficiently to make cloud computing more secure, robust, safe and cost-effective. Cloud service providers often employ techniques of auto-scaling to implement elasticity. This empowers for automated scaling decision-making depending on the priorities on various performance measures such as hardware metrics (e.g. CPU usage, memory) or service metrics (e.g. response time, service efficiency, length of queue etc). Auto-scaling systems can be categorized as proactive and reactive. Reactive mechanisms continuously track the system and invoke a particular scaling action if a specific demand is met (e.g. delivering or removing a specified number of resources if a defined parameter is higher or

lower than a specified threshold) (Moreno-Vozmediano et al.; 2019). The main issue with reactive approach is that the reaction speed or re-flexing time (the time elapsed from the detection of the trigger event until the services are provided and available for use) may not be sufficient to prevent overloading the machine; however, due to frequent variation of the allocated resources these mechanisms may influence cognitive instability of the system (Varghese and Buyya; 2018). In contrast, predictive or adaptive methods try to estimate the number of resources needed over the next time window, based on statistical or computational methods of evaluated workloads and device parameters. Although most of the present cloud systems, services and networks using reactive approach. Number of research is being done on predictive models based on queuing theory and time series analysis (Moreno-Vozmediano et al.; 2019).

Quick run time and superior precision these are the two metrics on which existing prediction algorithms competing. The proposed research here aims to improve the current system by predicting workload with the help of Machine Learning techniques which can further be used in the process of auto scaling mechanism by improving prediction accuracy. Integrating Machine learning algorithm can help us to boost the overall cloud computing performance, resource allocation strategies to optimize the overall cost and response time. In this paper, workload prediction approach is proposed by dividing and merging generated time series and applying to the combination of SVR + ANN model. Please refer the figure 1, to see cloud computing architecture with proposed workload prediction system. (Sharifian and Barati; 2019).

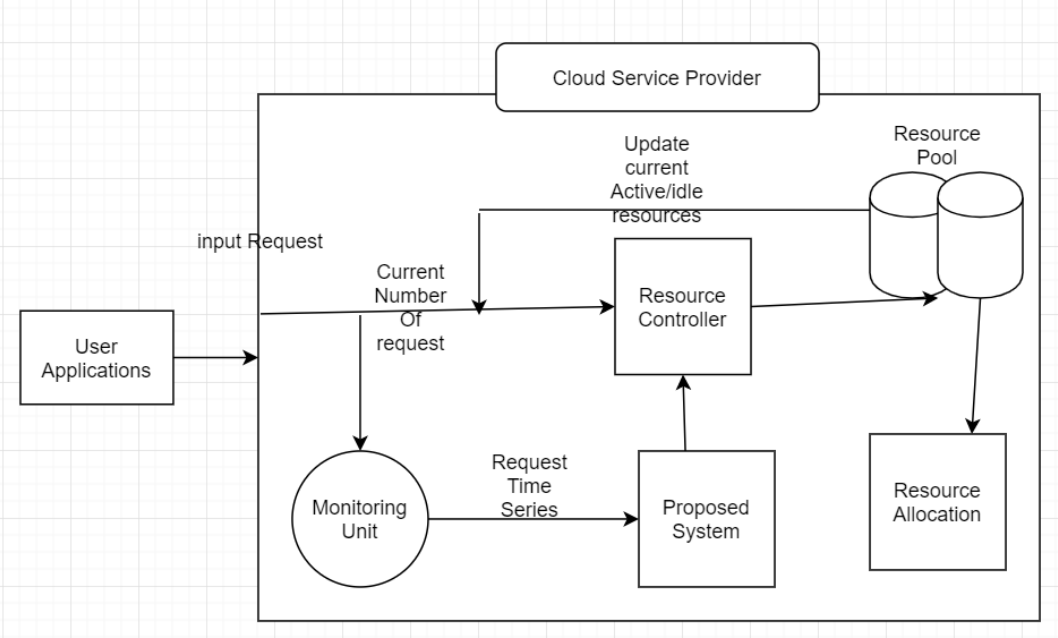

Figure 1: CC Architecture with proposed system

This paper proposes a novel method for predicting workload in cloud environment. Methodology is divided as follows, and used SVR + ANN hybrid algorithms to predict cloud workload accurately.

- Wavelet transformation is used to divide the input time series signals to two sub-

scales of different time frequency components as the nature of workload in the cloud setup is highly fluctuating over the time.

- In each sub-scale, one of the proposed a combination of SVR and ANN predictors is applied and trained the same to predict and again reconstructed workload using inverse wavelet transformation.
- An evaluation using generated synthetic data for this research.

This paper is structured further as follows: First section Related Works covers the work performed by researchers to workload estimation, resource allocation, auto-scaling using machine learning technologies on time series. The next section Methodology covers the proposed method with data generation, loading, pre-processing, transformation processes along with the detailed approach of proposed workload prediction method. The evaluation work evaluates how the proposed algorithm works on data used for training and testing to extract the expected results. Finally the conclusion sums up the paper as a whole.

Research Question - Can Machine Learning based predictive prowess improve workload prediction required for auto-scaling in cloud environment to attenuate the total cost of provided services and resource over-provisioning?

2 Related Work

2.1 Introduction

This section covers primarily the overview of similar research within the same technical area. There have been a range of different techniques, suggestions for predicting or forecasting the workload on Cloud. Nevertheless, this paper, addresses the subject of resource allocation and management in cloud setting, i.e. papers on workload prediction models using different methods and related fields together with dynamic resource allocation, resource auto scaling up/down etc.

2.2 Literature Review

The article (Li; 2015) proposes an auto-regression based system for predicting workload on the web server but the pattern of model is precisely linear in nature. A linear combination of previous values of the component under consideration is used to predict the value for future time instances, in auto regression. The authors used exponential smoothing for prediction of the seasonal time series data. To forecast the workload, two separate methods were used: additive seasonal model and multiplicative seasonal model. The model is ideal for time series representing seasonal behavior. In (Sun et al.; 2013), the authors of the paper implement a method which uses regression technique and service workloads of live sports events broadcast from commercial internet service provider is analyzed. The approach is simple and based on the statistical model perhaps that model will not capture patterns of the more rigorous data. This model is not efficient to predict

workload of more complex workload. The authors (Khan et al.; 2012) put forward a prediction on the workload, Model based approach with multiple time series. The model does a grouping of related programs to boost the predictions accuracy. The writers have also used hidden Markov Method (HMM) for the distinction of temporal correlations in obtained VM clusters. The information used by authors to define workload pattern variations over time. Some other methods have also been used for accurate workload prediction.

The author of (Tong et al.; 2014) suggested a exponentially segmented pattern and equivalent transformation to reconstruct prediction problem to conventional classification problem which means interval of prediction is subdivided into segments whose lengths extends exponentially. They (Tong et al.; 2014)achieve results using described method which comes under the category of long-term methods as it seems to be correct over longer period of time. But the problem with that it requires large computations and also requires samples in huge numbers in order to extract accurate workload prediction results. However, The average load estimation is a complicated task, since the average load can not accurately represent load fluctuations over some long time frame. Hence even though the workload is being predicted workload over long time frame, average methods in the proposed prediction algorithm is not used in this paper.

Rafael Moreno-Vozmediano1, Rubén S. Montero1,in (Moreno-Vozmediano et al.; 2019) evaluate and present a unique auto scaling mechanism based on the prediction. Machine learning techniques are used to predict the cloud workload and this workload prediction is then further being used to accurately predict the load of processing on a distributed servers and the appropriate number of resources are estimated that must be provisioned in order to avoid SLA violation and to optimise the response time while minimizing resource over provisioning that helps to mitigate energy consumption and equivalent infrastructure cost. They have used SVM regression model to accurately predict servers's processing load. Based on analytical method, they did optimal selection of SVM regression model. Further the prediction results are used to build queue based performance model to determine actual number of resources that must be provisioned. They carried out the results using real workload traces and produce the acceptable prediction result closer to the optimal case.

In paper (Yu et al.; 2018), authors suggest an idea based on a job-pool, where the awareness of the workloads of a huge pool of tasks is used to for the prediction of workload of new tasks. Based on their workloads the pool of jobs are clustered, and to help to learn the workload characteristics in each clusters, neuralnet is used. The authors of (Yu et al.; 2018) use preliminary workload pattern and submission parameters at the time when new job arrives which helps them discover the cluster it belongs to. finally, the respective neuralnet is used to predict workload of the new job in pipeline far to the future. The paper (Calheiros et al.; 2015) to predict cluster based long-term workload, basically they have used ARIMA machine learning algorithm. They test the task-pool systematic model experimentally and the findings lead to its efficacy and also managed to do experiments of clustering related learning with Non-clustering for predicting task-pool based workload. The test results suggest that the clustering based learning can mitigate the average prediction error significantly.

Researchers (Di et al.; 2014) use the Bayes model to create a new predictive approach to predict a long-term workload volatility pattern in Google data center. They attempted to predict: both of the mean workload over some future time period (up to 16 hours) and the mean load (which they refer to as a pattern) precisely over subsequent potential time periods. They build innovative features that are used to define the important and predictive statistical properties for the Bayesian analysis of the host work load. Those features include host load reliability, pattern and trends. Authors assessed whether any of these functions complement each other, and improve the predictive ability of the Bayesian model. A month load traces of a Google data center with over 10,000 machines is used to check them for assessment. Authors equate their Bayesian prediction methodology comprehensively with eight other models and stateoftheart methods using a variety of techniques such as self-regression, noise filters and moving averages. The MSE i.e mean-squared error of this approach is 0.0014 for only one interval, and is about 105 or less for a series. They confirm that pattern estimation accuracy could be increased by about 7 percent on average by detecting a collection of busiest or idlest hosts out of a total of ten k hosts across a load balancing instance. However, the approach proposed in this paper does not calculates average values over the time intervals because it can not accurately represents fluctuations in workload.

In paper (Calheiros et al.; 2015) time series datasets have been used to predict cloud workload by using very popular machine learning algorithm called ARIMA. The outcome of workload prediction is then used for resource allocation. Author (Calheiros et al.; 2015) apply ARIMA algorithm to time series historical data to pull out model parameters and use them to forecast future workload. We all know most of the cloud workloads are volatile and highly fluctuating with time and ARIMA is not good choice for predicting workloads with high volatility. On the contrary, to deal with the most fluctuating workload, in approach proposed in this paper the workload is sub-scaled by using wavelet transform and SVR + ANN one of the both is applied to best suited sub-scale. Decision is made after analysing characteristics of both the algorithms and the appropriate Algorithm is applied to that sub-scale in order improve Workload prediction accuracy.

In the paper (Jheng et al.; 2014), authors proposed a method named Gray Forecasting to allocate Virtual Machines. This Gray Forecasting model is basically workload prediction approach which is the 1st string of the given research field of area. At first they utilize the time dependent of workload at the same time period of every day and with help of that further forecasting VM workload inclination towards decreasing or increasing. In the next phase of research they compare earlier time period usage of workload with the value predicted in the previous stage. The downside of this given approach is that newly predicted data,they don't use to update a model. To put it another way, Rather than data dependency, time dependency has been used between predicted data and historical data. Contrary, it is proposed in this paper that in each sliding-window, it update model which helps to increase the accuracy of the model. In (Li et al.; 2016) an improved version of gray model is suggested, using particle swarm optimization algorithm, wherein the preliminary and background values, which significantly affect the performance or the output of the model are estimated using algorithm PSO. The grey model is based on distinction in numbers and one of the boons are that it needs little or very small amount of train data. Gray based models produce good results for high volatility loads, but due to the time dependence hypothesis in this model, it does not

produce good workload predictions on or for time series datasets with high fluctuations. The approach proposed in this paper, though, the method dose not make any time dependence hypothesis concerning to time series but predicts workload in each and every step is based on the corresponding previous sliding-window values.

In the paper (Kumar et al.; 2018) kumar proposed a cloud workload prediction approach using Long short Term Memory (LSTM). LSTM are basically special form of Recurrent Neural Network. In case of predicting time series which has long term dependencies LSTM is very useful because LSTM is well known for remembering long term information. They produce a model that tested on benchmark detasets of web server logs and the result indicate that the predicting workload using LSTM shows acceptable values for volatile workloads. But its prediction for workload with high fluctuations shows lower accuracy. Proposed method in this paper performs well on both with lower and higher fluctuations sub scales.

Authors of (Qazi and Aizenberg; 2018) proposed a workload prediction model build on the concept back propogation neural network. To improve the correctness of the prediction, it uses the stochastic model. The paper produces acceptable results. In (Caron et al.; 2011) caron at al. predicted future workload by using concept of pattern matching. The given model investigate the similar patterns in the past and to do that it uses KMP string matching algorithm in order to predict the upcoming workload based on matching patterns from the history. The disadvantage is that the model gets slower as the history of data workload increases.

The authors (Chen et al.; 2015) used fuzzy neural network and represented each resources as a period of fluctuations and flatness which means period of high fluctuation and low fluctuation workload. That makes system complex because authors use two layers two layers of predictions are involved that leads to delay in workload predictions. Furthermore in (Chang et al.; 2013), proposed a model using steepest descent learning algorithm and neural network to predict workload. Model produces results with improved prediction accuracy over time hindrance neural network and linear regression. But the given model also produces high prediction errors. There are certain problem with with gradient based approaches as they are sensitive towards learning rate choices and initial solution and to overcome this problem, (Lu et al.; 2016) implemented a back propagation learning algorithm based prediction model to predict workload in cloud environment. They used workload cluster traces produced by google and evaluated model over it by estimating task based on computational latency. However, the correctness of model goes down as the latency levels increases.

2.3 Gaps in Literature Review

To summarise, some works use existing approaches directly, and some work streamlines the models to improve accuracy of prediction. Many in the study work leverages resources utilization predictive methods for improving virtual machine consolidation, resource utilization and efficiency in resource management or usage that is changing over the time that leads to poor accuracy in prediction. This is the reason a hybrid method i.e combination of two approaches is used which improves the accuracy of workload prediction.

Considering high fluctuations in Cloud resource use over the short to medium term and short term prediction difficulty the multi-step predictive approach is used to solve this hurdle. Further elaborated on proposed method, evaluations and results in later chapters of the paper.

3 Methodology

There are many data mining methodologies available, out of which KDD is perfect match for the research, According to (Saltz et al.; 2017) KDD model works well with Machine learning technique and when the efforts are taken on large datasets and extract some information and pattern using ML. As KDD steps primarily concentrate on the execution phase Instead of an approach to project management, it is better suited for classification and Forecast.

3.1 Data Selection And Generation

Workload of cloud are usually with lot of fluctuations. Cloud workload is much more noisier than that of grid computing workload. In this project the synthetic data is created. The created data-set is resembling with the real time cloud workload traces. Behaviour and the complexity of generated synthetic data is similar as that of real workload traces of cloud environment. The data-set is generated programmatically and therefore no real-life survey or experiment gathers the given data used in this research. However, its primary objective is to be versatile and abundant enough to perform compelling experiments with proposed ML algorithm for classification, regression.

3.2 Data Preprocessing

The next step is to pre-process the data prior to the usage for the experiments to remove inconsistencies and major errors from the dataset. Data cleaning is the exercise of ensuring the correctness, consistency and accessibility of your data and in this project this process has been done using Python programming language. Data cleaning process involves steps like removing missing values, Null values, removing redundant values or repeated variables, etc.

3.3 Feature Extraction

In this process, found out some important characteristics from generated time series dataset to work on the proposed model for predicting cloud workload. This practice in this research is helping to enhance machine learning algorithm and ultimately leading to solve the given research problem.

3.4 Transformation

In this step, data has been transformed using Wavelet Transformation. To be precise discrete wavelet transformation is used and disintegrated the input signals into Low frequency and high frequency sub-scales. As per the requirement for the given research,

the transformed data then applied to the ML techniques to accurately predict the cloud workload.

3.5 Data Mining

In this step, different data mining techniques have been used to full fill the research requirements. In this research, patterns are found out from the data and applied to the machine learning to predict the workload and different models are created using SVR, ANN and from the combination of both.

3.6 Evaluation

In this stage of evaluation, correctness of prediction, precision and sensitivity as found are carried out from the results of different models created using data mining techniques. Further the results are compared and tested by understanding the curves and then results are plotted using python libraries.

4 Design Specification

The proposed prediction model is build using Python programming language and further divided into three stages.Input signal is divided time series through various sub-scales of time frequency using Wavelet transformation. Suggested hybrid algorithm SVR+ANN is applied on sub-scales to predict workload. In third stage the output is reconstructed and results are evaluated. Wavelet transformation and machine learning techniques used are discussed further in this paper.

4.1 Wavelet Transformation

Variety of factors are responsible for the noisy behaviour of time series cloud workload. Wavelet transform, in fact a multi scale time-frequency analysis model which decomposes input load time series to several sub-scales. In this way the input of pre-processing time series led to a reduction in the disturbed and irregular/unusual characteristics of time series analysis (Liu et al.; 2015). Similar with filter bank frequency in the signal processing, wavelet transformation also has capability to divide time series input data varying bandwidth sub-scales. Each of the sub-scale further applied to the one of the prediction ML algorithm.

In the propose approach in this paper, to be precise the discrete wavelet transformation is used. DWT is implementation of wavelet transformation which uses discrete set of the wavelet scales and translation following some fixed set of protocols. DWT named Haar is the simplest of wavelet transform Giorgi et al. (2014) and used in this paper for wavelet transformation using Python programming language. The selection of Haar DWT is done because of it is very good for compression and signal processing as it helps to reduce the redundancy after transformation. However DWT is used in this approach in order to accurately predict cloud workload.

4.2 Support Victor Regression - SVR

SVR is a binary classification extension of SVM i.e. Support Vector Machines with a difference that outputs take infinite values. SVR can also be used to estimate functions, to tilt curves and time series prediction. The regression problem is finding a function on the grounds of a training sample which approximates maps from an input domain to actual numbers. Using SVR researchers have proved that it could be used to increase the accuracy of prediction ref (Sharifian and Barati; 2019) as it can achieved by adjusting SVR parameters properly.

4.3 Artificial Neural Network - ANN

It is a nonlinear modeling method that is ideal for for modeling through a variety of applications. This is more architecturally flexible. It bears with high similarity to the neurons in the brain and hence it is called as Artificial Neural Network. ANN is a part of Deep Machine Learning Technique and in this paper it is implemented by using Keras Python library. Because of the regression problem, Adam optimizer is decided to use in the code where as other hyper- parameter used in the code are is loss = binary_crossentropy and metrics used is accuracy. In training scenario, the network is re-adjusted, and newly inserted neurons are discontinued but the amounts are increased with the check probability of decommissioning correlated with units (Kumar and Singh; 2018).

5 Implementation

The implementation workflow of entire research is illustrated in Figure 2. Entire process is divided into 4 sub-parts

- Data Generation
- Decomposition of Input signal using wavelet Transformation
- Proposed workload prediction approach (SVR+ANN)
- Reconstructing the signal to see the Predicted results.

5.1 Data Generation

As discussed above that the cloud workloads are volatile in nature. When in the day time usually the frequency of input requests are high and low during the night and also it depends upon the geographical location where that cloud services is being provided. To see the performance and to emphasise proposed hybrid approach, the high level data with local variations and randomness in the data is needed. So it is planned work on the input requests those are more noisier than usual. To fulfill our requirement synthetic testing data is created by combining the periodic wave function and pseudo randomness. The data generation script is created using python programming language in such a way that it data can further tweaked to model of data situations like high/low randomness, vary the period etc. Basically if you want to use more complex data you can generate data as per the requirement.

As discussed earlier, script is prepared using python programming language to create custom signal that can simulate diverse scenarios of workload variations and randomness. You can see in the figure 2, in the first phase of data generation, the time series data are generated with 40000 values in a series. Then further, values are divided into rows of 100 values by iterating through steps of 10 values. Then further in the process, split the 100 values into two parts 90 and 10 values respectively. The first part with 90 values that has been used for the input and second part with 10 values, which is used for the prediction. Later the data has been divided to training and testing dataset. In the next step, this generated input signal is used for the next step of wavelet transformation as shown in figure 2.

5.2 Decomposition of input signal using wavelet Transformation

This unique stage of research which includes data prepossessing. Wavelet transform has the ability to break down or disintegrate time series into sub-scales of different frequencies bandwidth more like a frequency filter bank in the signal processing. Input signals can also be subdivided by wavelet transformation to a prediction of the time series (overall signal shape) and the several time series with higher frequency. This is unique and important phase of research discussed in this paper. This process has direct impact on the proposed hybrid model's prediction accuracy. Discrete wavelet transformation is used in the process of wavelet transformation. DWT Haar is the simple way of implementing WT and as discuss earlier, it is very good for compression and signal processing as it helps to reduce the redundancy after transformation.

In our research using Discrete wavelet transformation the input signal is divided into 2 components one with low frequency and other with high frequency. Prior to that the input time series data is divided into small section of 100 values. Further data is disintegrate using Discrete wavelet transform into two sets named cA and cD having size of 50 values each. cA represents low frequencies component whereas cD represents high frequency component. As shown in the workflow, each component is then subdivided further into 2 parts 45 values for the input applied to the prediction technique and rest 5 for the prediction output.

5.3 Proposed workload prediction approach (SVR+ANN)

The given proposed approach in which a combination of Support Victor Regression(SVR) and Artificial Neural Network(ANN) is used to accurately predict cloud workload. Taking into account of advantages of both the machine learning techniques, this approach is proposed. SVR model works well and reliable to predict variables that changes slowly. As the name suggests this model is very accurate when estimating values using regression datasets. But SVR model faces difficulty while predicting highly fluctuating variables. When the high frequency variations combined with low frequency variations that occurs over a longer period of time SVR has limitations accurately predicting values. In this paper to accurately predict workload high frequency workload is applied to ANN then it is further combined with the SVR prediction model to get more accurate results. This composite model is applied on complex time series data where the data have variations with multiple frequencies and randomness.

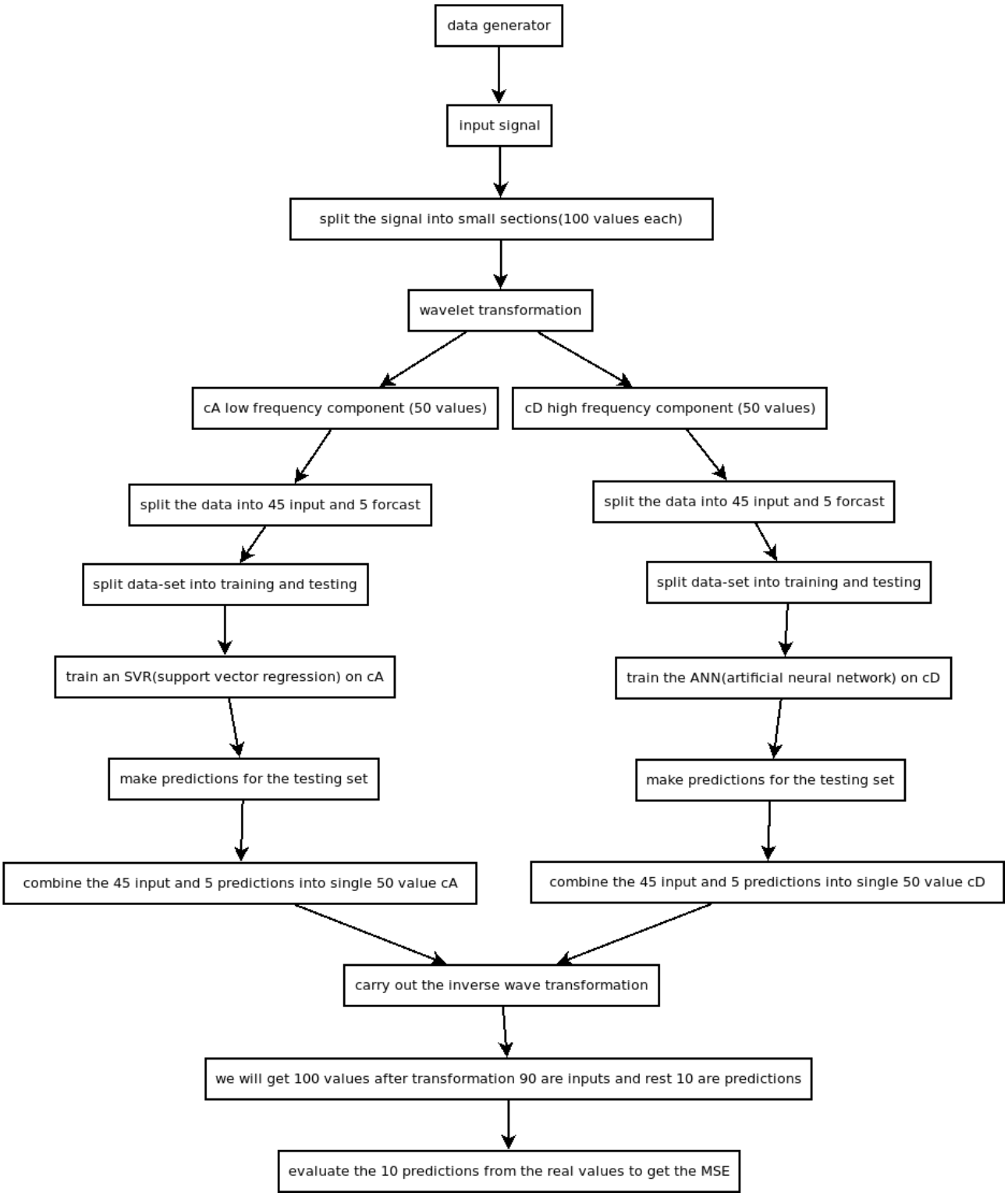

Figure 2: Proposed Workload Prediction Approach Workflow

After Wavelet Transformation phase, input signal is divided into two sets as discussed earlier. cA represents low frequency components whereas cD represent high frequency component. Then cA is again divided further into two parts with 45 value as input and 5 values for output. cD is split in similar way. As shown in figure 2 the cA training dataset is applied or trained on SVR algorithm. 45 values of input of cA are feed into the SVR algorithm. SVR forecast the next 5 values in cA. The input of cA component 45 values is then further combined with predicted output 5 values to get output component with 50 values. Similarly, cD training dataset is applied or trained on ANN algorithm. 45 values of input of cD are feed into the ANN algorithm. ANN forecast the next 5 values in cD. The input of cD component 45 values is then further combined with predicted output 5 values to get output component with 50 values. As shown in workflow cA and cD components combined to reconstruct the original signal on 100 values using inverse wavelet transformation. Inverse wavelet transformation works exactly opposite of Wavelet transformation. This gives the 100 values as the final predicted result in which 90 values are the original input and 10 values are predicted values. Further these predicted values are compared with the original signal to obtain the MSE and RMSE. The output showed using graphs with predictions and real values are discussed later in the evaluation section. The accuracy of SVR model is highly depends on the kernel used. For the best accuracy low frequency variations components 'rbf' kernel is used. For high frequency variation component, simple ANN model implemented in tensorflow 2x. This network is taking 45 values as input and has 2 hidden layers. ADAM optimiser has been used and MSE as the loss function.

6 Evaluation

The proposed hybrid model (Wavelet Transformation + combined SVR and ANN) implemented successfully along with SVR implemented individually. Thereafter results were evaluated by considering their Mean Square Error (MSE) and Root Mean Square Error (RMSE). The results produced by the proposed model and graphs are discussed below in detail.

6.1 Experiment 1: Simple SVR model

Separated time series data into small sections on 100 values each. In the data processing the origin is shifted each time by 10 values which means if the first row is 0-100 then the next row is 10-110 and so on. From each row of 100 values, first 90 values are used as input and remaining 10 values as output. Here in simple SVR model without using wavelet transformation to divide data into sub scales all 90 values are applied on SVR model to predict the next 10 values. Accuracy of prediction is evaluated by comparing the predicted 10 values with real values as shown in the output graph figure 3. Mean Square Error of SVR model is 0.02066 whereas Root mean square error of SVR is 0.1438.

There was a great deal of work on the dataset before it was used for predicting values. After carrying out various test on SVR only model the prediction results obtained are satisfactory. By observing the graph refer figure 3, SVR prediction vs real values, we can see the satisfactory curves of predicted values denoted by blue colour. Figure 4 represents results of calculated MSE and RMSE of the SVR only model on command line. Output

clearly shows that the predicted values are in line with the actual values for most of the plot story, but at various data points it goes out of control.

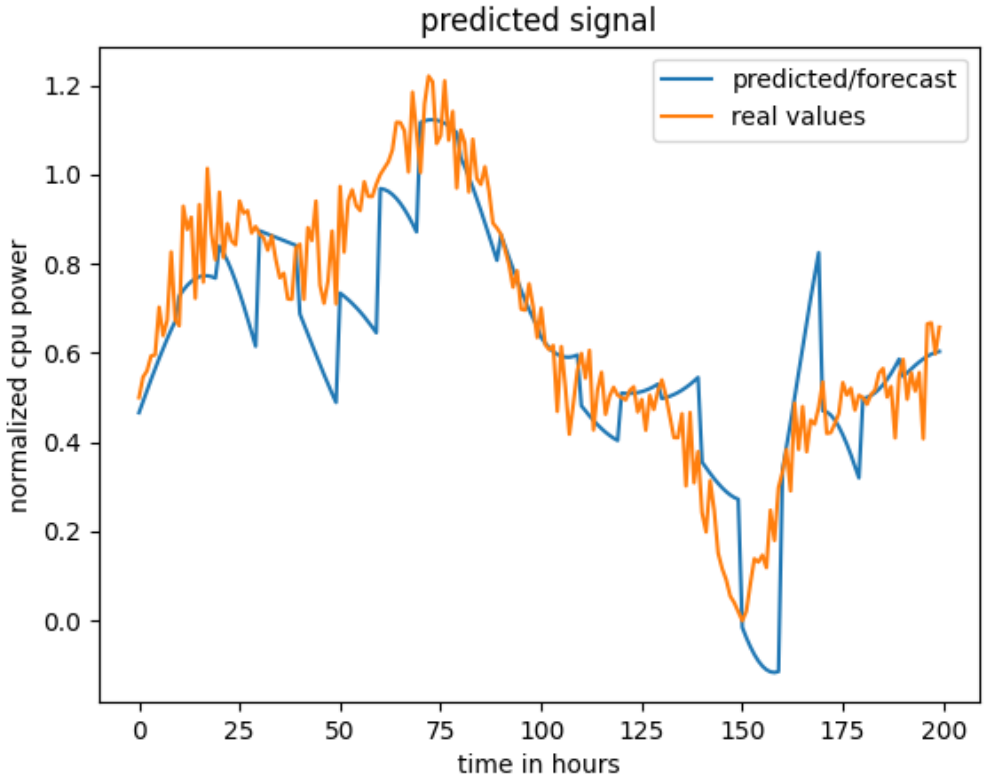

Figure 3: SVR Prediction Vs Real Values

```
C:\Users\Sumedh\Thesis\SVR+ANN>python benchmark_svr.py
(3990, 100)
(3990, 90)
mse of svr algorithm is: 0.02066195169772968
rmse of svr algorithm is: 0.14374265789155868

C:\Users\Sumedh\Thesis\SVR+ANN>
```

Figure 4: MSE and RMSE values for SVR

6.2 Experiment 2: Proposed Prediction Approach

This section gives insight of evaluation process of the proposed model. In the proposed model, data generation and wavelet transformation are the key phases. Wavelet transformation technique is used to spit the 90 values via wavelet into 2 sets of 45 each namely cA and cD. 45 values of low frequency components i.e cA are feed into SVR algorithm and SVR forecasts next 5 values. Inputs are combined with predicted output to get the cA output with 50 values. Similarly 45 values of high frequency components i.e cD are feed into ANN algorithm and then ANN forecasts next 5 values. Input combined with output to get cD with 50 values. Ultimately to get final output, cA and cD are combined to reconstruct original signal on 100 values using inverse wavelet transformation. Final output comprises of 90 values are of the original signal and rest 10 values are of the forecast.

The predicted results obtained from the proposed model are promising and more accurate predicted values as compared to the SVR only model. In the SVR+ANN plots, actual values shown by orange line and forecast values represented by blue line. In figure 5, output clearly shows the accuracy of proposed model is better than the SVR only model. MSE of proposed model is 0.01475 and RMSE of the model is o.1214 showed in figure 6 which proves the significant difference in accuracy as compared to simple SVR only model. It can be noticed that the predicted values are in sync with the actual values for most of the plot story.

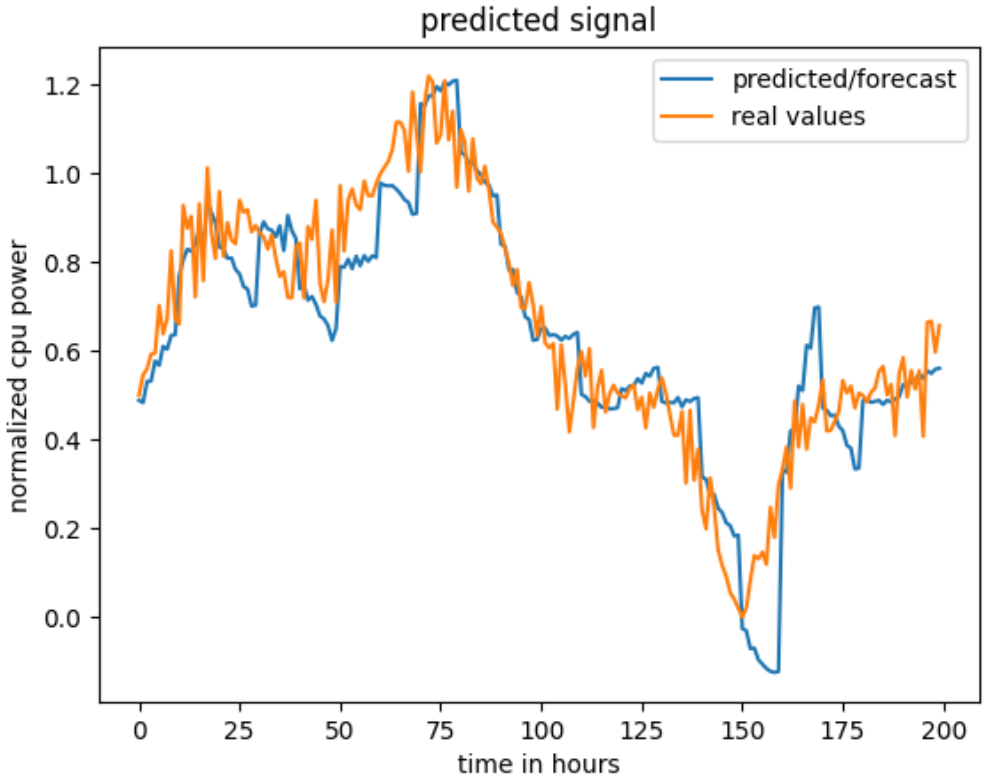

Figure 5: Proposed Model Prediction Vs Real Values

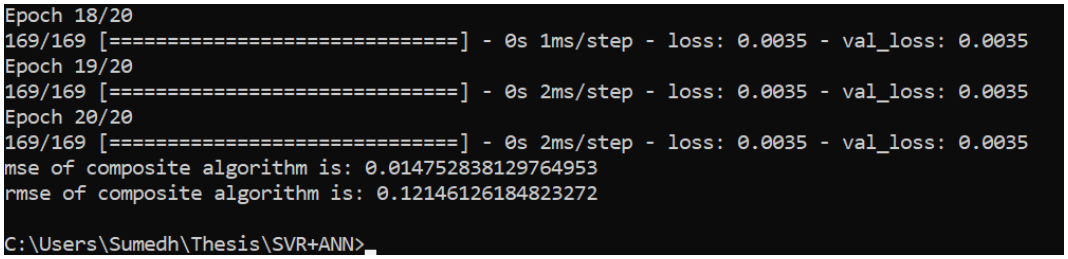

Figure 6: MSE and RMSE values for Proposed Model

6.3 Discussion

Above mentioned experiments addressed the cloud workload results predicted using two separate algorithms wherein the first algorithm, Simple SVR model, is a regression model, the other is SVR+ANN, a deep learning algorithm as well as a proposed time series forecast model. The nature of these models, advantages / disadvantages of each and respective predicted output of those experiments are varied. The proposed model was the best model in terms of performance and other metrics like MSE and RMSE. Proposed model produced satisfactory results. The results obtained from the experiments are shown in the table 1. It can be therefore concluded that developed SVR+ANN model predicts cloud workload with low values of MSE and RMSE.

Table 1: Comparison of prediction methods.

Models	**MSE**	**RMSE**
SVR Only	0.02066	0.1438
SVR+ANN	0.01475	0.1214

It is a novel approach towards cloud workload prediction and in terms of including Machine Learning into the cloud computing research. However, the experiments results obtained are satisfactory but the results are not extraordinary when comparing to the other researches in this field and state of arts of this paper. This could be because of tests are carried out on different kinds of training and testing datasets. Perhaps obtained results vary because of dataset is extremely volatile or very stable. It depends on the load on the cloud services and other factors like the geographical location of the cloud services being offered. It is always helpful to consider all the characteristics of algorithm that may be responsible for enhancing approach further to accurately predict the cloud workload. It's always a good approach for any research to consider all the factors of proposed technique that provides room for training models, and substantially reduces the error.

7 Conclusion and Future Work

This paper evaluates machine learning based predictive approach which is essential for dynamic auto-scaling in cloud infrastructure. This paper accurately predicted workload which helps to allocate resources as per the demand and that will eventually lead to attenuate the total cost of provided services and resource over-provisioning and to make sure that allocated resources utilized fully. Our proposed combined approach furthermore, to achieve the utmost accuracy, it works better in terms of computation cost. This unique approach consider synthetic data generation and decomposition of input signal i.e multi scale decomposed analysis and trained and tested for each sub-scale (cA and cD) purely different carefully configured / tuned prediction model (SVR / ANN) as per the requirements. Considering advantages and other factors and then sub divided input signals are applied on either SVR or ANN module. The final prediction is achieved by combining outputs of previous stages using inverse wavelet transformation which has higher accuracy of prediction. Accuracy has been improved by applying the complex time series of high variations with multiple frequencies and randomness input to the ANN model. However, this process is expensive in computational terms.

For the future works, one can try to decompose input signal in more that 2 sub-scales using wavelet transformation is suggested for future work, considering all the factors of chosen algorithm to help enhance the prediction accuracy. In order to improve accuracy of workload prediction, Using other standard rivalry algorithms and the family of Artificial neural network ANN is recommended.

References

Calheiros, R. N., Masoumi, E., Ranjan, R. and Buyya, R. (2015). Workload prediction using arima model and its impact on cloud applications' qos, *IEEE Transactions on Cloud Computing* **3**(4): 449–458.

Caron, E., Desprez, F. and Muresan, A. (2011). Pattern matching based forecast of non-periodic repetitive behavior for cloud clients, *Journal of Grid Computing* **9**(1): 49–64.

Chang, Y.-C., Chang, R.-S. and Chuang, F.-W. (2013). A predictive method for workload forecasting in the cloud environment, *Lecture Notes in Electrical Engineering Advanced Technologies, Embedded and Multimedia for Human-centric Computing* p. 577–585.

Chen, Z., Zhu, Y., Di, Y. and Feng, S. (2015). Self-adaptive prediction of cloud resource demands using ensemble model and subtractive-fuzzy clustering based fuzzy neural network, *Computational Intelligence and Neuroscience* **2015**: 1–14.

Di, S., Kondo, D. and Cirne, W. (2014). Google hostload prediction based on bayesian model with optimized feature combination, *Journal of Parallel and Distributed Computing* **74**(1): 1820–1832. JCR Impact Factor:1.815 (2018).

Giorgi, M. D., Campilongo, S., Ficarella, A. and Congedo, P. (2014). Comparison between wind power prediction models based on wavelet decomposition with least-squares support vector machine (ls-svm) and artificial neural network (ann), *Energies* **7**(8): 5251–5272.

Jheng, J.-J., Tseng, F.-H., Chao, H.-C. and Chou, L.-D. (2014). A novel vm workload prediction using grey forecasting model in cloud data center, *The International Conference on Information Networking 2014 (ICOIN2014)* .

Khan, A., Yan, X., Tao, S. and Anerousis, N. (2012). Workload characterization and prediction in the cloud: A multiple time series approach, *2012 IEEE Network Operations and Management Symposium* .

Kumar, J., Goomer, R. and Singh, A. K. (2018). Long short term memory recurrent neural network (lstm-rnn) based workload forecasting model for cloud datacenters, *Procedia Computer Science* **125**: 676–682.

Kumar, J. and Singh, A. K. (2018). Workload prediction in cloud using artificial neural network and adaptive differential evolution, *Future Generation Computer Systems* **81**: 41–52.

Li, K., Liu, L., Zhai, J., Khoshgoftaar, T. M. and Li, T. (2016). The improved grey model based on particle swarm optimization algorithm for time series prediction, *Engineering Applications of Artificial Intelligence* **55**: 285–291.

Li, T.-H. (2015). A hierarchical framework for modeling and forecasting web server workload, *Journal of the American Statistical Association* **100**(471): 748–763.

Liu, S., Hu, Y., Li, C., Lu, H. and Zhang, H. (2015). Machinery condition prediction based on wavelet and support vector machine, *Journal of Intelligent Manufacturing* **28**(4): 1045–1055.

Lu, Y., Panneerselvam, J., Liu, L. and Wu, Y. (2016). Rvlbpnn: A workload forecasting model for smart cloud computing, *Scientific Programming* **2016**: 1–9.

Madni, H., Shafie, A. L., Yahaya, C. and Abdulhamid, S. (2017). Recent advancements in resource allocation techniques for cloud computing environment: A systematic review, *Cluster Computing* **20**: 1–45.

Messias, V., Estrella, J., Ehlers, R., Santana, M., Santana, R. and Reiff-Marganiec, S. (2015). Combining time series prediction models using genetic algorithm to autoscaling web applications hosted in the cloud infrastructure, *Neural Computing and Applications* . JCR Impact Factor:4.664 (2018).

Moreno-Vozmediano, R., Montero, R., Huedo, E. and Llorente, I. (2019). Efficient resource provisioning for elastic cloud services based on machine learning techniques, *Journal of Cloud Computing* **8**.

Qazi, K. and Aizenberg, I. (2018). Cloud datacenter workload prediction using complex-valued neural networks, *2018 IEEE Second International Conference on Data Stream Mining Processing (DSMP)* .

Saltz, J., Shamshurin, I. and Connors, C. (2017). Predicting data science sociotechnical execution challenges by categorizing data science projects, *Journal of the Association for Information Science and Technology* **68**(12): 2720–2728.

Sharifian, S. and Barati, M. (2019). An ensemble multiscale wavelet-garch hybrid svr algorithm for mobile cloud computing workload prediction, *International Journal of Machine Learning and Cybernetics* **10**(11): 3285–3300. JCR Impact Factor:3.844 (2018).

Sun, Y. S., Chen, Y.-F. and Chen, M. C. (2013). A workload analysis of live event broadcast service in cloud, *Procedia Computer Science* **19**: 1028–1033.

Tong, J.-J., E, H.-H., Song, M.-N. and Song, J.-D. (2014). Host load prediction in cloud based on classification methods, *The Journal of China Universities of Posts and Telecommunications* **21**(4): 40–46.

Varghese, B. and Buyya, R. (2018). Next generation cloud computing: New trends and research directions, *Future Generation Computer Systems* **79**: 849–861. JCR Impact Factor:4.639 (2018).

Yu, Y., Jindal, V., Bastani, F., Li, F. and Yen, I.-L. (2018). Improving the smartness of cloud management via machine learning based workload prediction, *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)* .