An Approach for Optimizing Resource Allocation and Usage in Cloud Computing Systems by Predicting Traffic Flow

Sekwatlakwatla, Sello Prince; Malele, Vusumuzi

Sello Prince Sekwatlakwatla
sek.prince@gmail.com
North-West University, Sudáfrica

Vusumuzi Malele
Vusi.Malele@nwu.ac.za
North-West University, Sudáfrica

Abstract: The cloud provides computing resources as a service (scalable and cost-effective storage, management, and accessibility of data and applications) through the Internet. Even though cloud computing offers many opportunities for ICT (information and communication technology), many issues still remain, and the increasing demand for resource management and traffic flow is also becoming increasingly problematic. The amount of data in the cloud computing environment is increasing on a daily basis, which increases data traffic flow. Due to this problem, clients complained about the network speed. Autoregressive Integrated Moving Average (ARIMA), Monte Carlo, Extreme gradient boosting regression (XGBoost), is used in this paper for predicting traffic flow. A Monte Carlo prediction of 84% outperformed ARIMA's prediction of 79.8% and XGBoost's prediction of 71.5%, indicating that Monte Carlo is more accurate than other models when predicting traffic flow in organizational cloud computing systems. A machine learning model will be used for future studies, along with hourly monitoring and resource allocation.

Keywords: Monte Carlo technique, Autoregressive integrated moving average (ARIMA) and Extreme gradient boosting regression.

I. INTRODUCTION

As telecommunication networks play an increasingly important role in deploying cloud applications and services in data centers, converged network traffic is growing each year as the cloud computing concept becomes more widely used as a means of providing access to applications and services[1].

Recently, cloud computing has emerged as an innovative way to deliver and host services via the Internet. As a result, business owners are attracted to cloud computing because they do not need to plan ahead for provisioning, and they can start with the fewest resources and expand them only when demand increases[1-3].

Introducing advanced technologies like cloud computing has revolutionized businesses and industries worldwide. It has created a new way of storing and accessing information online[4].Despite offering tremendous opportunities for the IT industry, cloud computing technologies are still in their infancy, with many issues yet to be resolved. It is an increasingly Thoughtful problem for the ICT (Information and Communication Technology) infrastructure organization to ignore the growing traffic flow and resource management demands.

Managing and maintaining ICT resources is easier and more efficient thanks to cloud computing [6]. The use of cloud computing by many organizations is growing, so if a tool does not exist to forecast cloud

computing traffic, resource allocation to clients will be inefficient [8].Cloud computing is complicated by the inconsistent flow of network traffic, which makes it difficult to predict which network resources will meet the needs of all network clients at a particular moment. Clients complained about slow system times, application timeouts, and high bandwidth usage due to inconsistencies in traffic flow [9].

In recent literature, many cloud computing service providers have been finding it hard to allocate the resources they need to meet their clients' needs, causing system bottlenecks, especially during peak periods [5].

The main contributions of this paper are the simulation experiments conducted and the bibliometric analysis conducted. Monte Carlo, Extreme Gradient Boosting (XGBoost), and Autoregressive Integrated Moving Average (ARIMA) were used to make predictions.

Using a cloud network services traffic flow dataset, this paper compares three prediction techniques used to overcome this challenge. It is structured as follows: an introduction, related work, methods ,simulations , results, and conclusions

II. RELATED WORK

A. Bibliometric analysis of the proposed techniques

For exploring and analyzing large amounts of scientific data, bibliometric analysis has become a popular and rigorous method [1]. As a result, analysis is able to uncover the evolutionary nuances of specific fields. Hence, this paper examines the proposed prediction technique to examine the main key area of this approach.
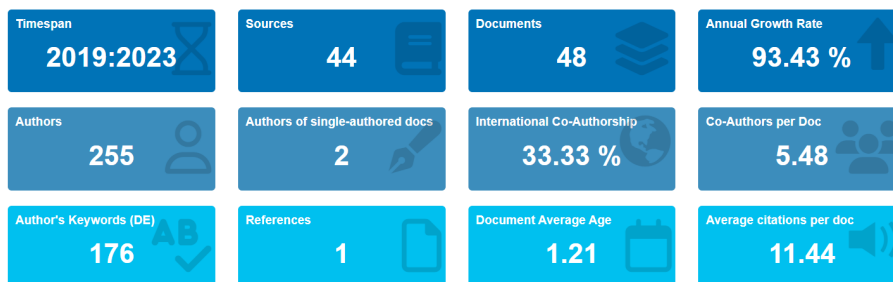

FIG. 1.
Summary of the scopus view

As a result of bibliometric analysis techniques, researchers can gain insights into the selected research area very quickly. Search queries were executed on the Scopus database, and 44 documents were extracted between 2019 and 2023. According to Fig 1, out of 255 authors and 176 author keywords (DE), there is a high prediction rate of 93.43%.

FIG. 2.
Most Frequent Words

With reference to the Monte Carlo technique, Autoregressive integrated moving average (ARIMA) and Extreme gradient boosting regression ((XGBoost), Fig 2 presents the word cloud generated using the authors' keywords. Forecasting words are shown in a larger size, such as "machine learning," "learning systems," and "time series" and "time series analysis". Prediction was the area of focus for three proposed models.

B. Gap analysis on the related techniques

A cloud computing service provider must provide adequate and accurate traffic prediction to meet customer needs and support organizations effectively [2-3]. Cloud computing service providers estimate requests and the amount of data they must move frequently to predict system traffic flow. Intelligent transportation cyber-physical cloud control system (ITCPCCS) improved accuracy in cloud control traffic; however, it was difficult to predict in real time since the technique is in the development stage [2].

In recent years, Monte Carlo simulation Vibrational mode decomposition (VMD) has been shown to be a more accurate tool for predicting short-term conditions than long-term models. Limit forecasts to short-term deterministic and probabilistic models [5]. The accuracy of forecasts is tested using publicly available Google trace data and Bit Brain data, but workloads in dynamic clouds may experience substantial forecast errors [13].

There is an overlap between the highest values of subseries, models orders, and weights that are time-dependent and auto-regressive. Auto-regressive integrated moving average models (ARIMAs) [14] produce more accurate forecasts and are more efficient to compute. As most applications and information will have to be moved repeatedly in the future, it is best to move them daily. The activity stream's predictions have been validated in numerous studies [13] and 85% of the predictions have been met. In spite of this, there has been no progress in implementing a framework that predicts traffic using an application framework. Since cloud computing is a continuous process, it is difficult to predict traffic 100% accurately [5-6].

The stationary time series are assumed to remain unchanged over time in this method, which is a Markov chain Monte Carlo approach and a Seasonal Autoregressive Integrated Moving Average (SARIMA) model [6]. In comparison to the existing method, the proposed method has a lower error rate. Artificial Neural Network (ANN) [7] has the ability to predict accurately with a small error or small neurons. It is concise and straightforward to perform Monte Carlo simulations of GP distributions, and errors increase and decrease with sample size [8].

Traffic flow predictions can be enhanced with Multiple Linear Regression Unit-M4ulti Region Correlation (MRC-MLRU) models [1], which are useful when traffic data is limited. eXtreme Gradient Boosting (XG Boost) [9] this method provides better prediction performance with an R2 of 0.992 and 0.949, respectively. The baseline model tends to be over fitted. This is an improved multivariate linear regression variable parameter spatiotemporal (MLR-VPST) zoning model [11]. Both quality and accuracy have improved, and errors are no longer beyond the process limits [15].

III. METHODS

A. Probabilistic Models

To quantify uncertainty, a probabilistic model integrates the first principle knowledge with data to capture a distribution of state transitions between samples in a batch run by making predictions based on the model predictions. A probability model calculates the probability of certain events occurring rather than monitoring actual data to look for events and data points that conform to a set of rules defined by historical analysis-Monte Carlo.

Monte Carlo technique is a computer-based mathematical technique that accounts for risk quantitatively in forecasting and decision-making [13].
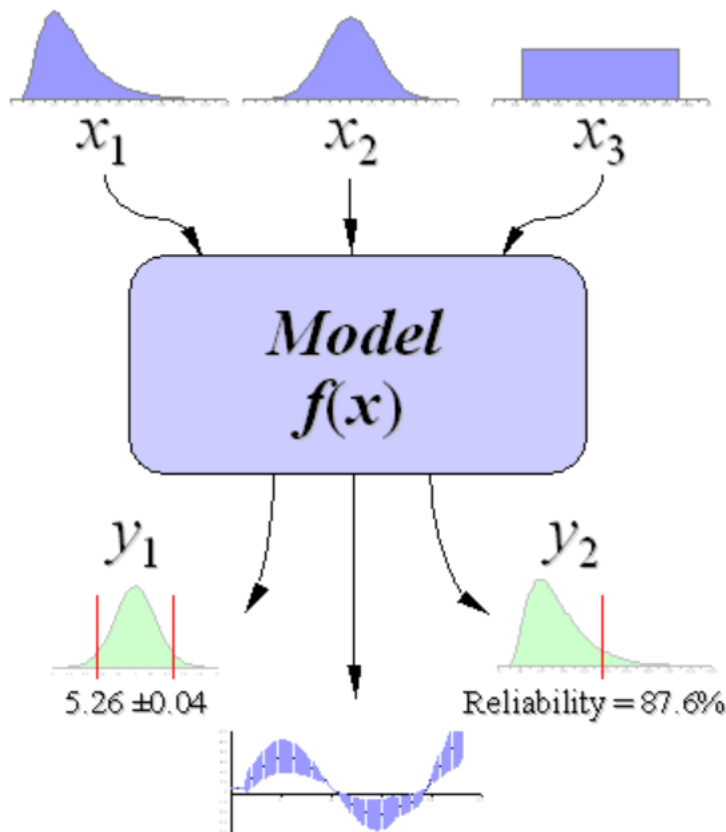
FIG. 3
The simple principle of the Monte Carlo simulation [9]

In order to predict the outcome of a Monte Carlo simulation, the following steps are proposed Fig 3:
Process 1: The first process is to generate a regression parametric:
Technique, $p=f(x\_1,x\_2.......x\_r)$.

Process 2: Set up input generator, x_(c1,) x_(c2,).......x_cq
Process 3: Model evaluation and data storage as y_c
Process 4: Process 2 and 3 must be repeated i=1 to n
Process 5: Calculate confidence intervals, complete statistics,and histograms based on the results.

B. Machine Learning/Data Mining

In machine learning, data are used to create mathematical models that predict or decide without explicit programming. The algorithms use "training data" to make predictions Extreme gradient boosting regression (XGBoost), in time series forecasting, models like Extreme Gradient Boosting (XGBoost) can provide reasonable forecasts without tweaking hyper parameters [9]. This is a machine learning algorithm that uses gradient boosting to solve a variety of problems in machine learning.
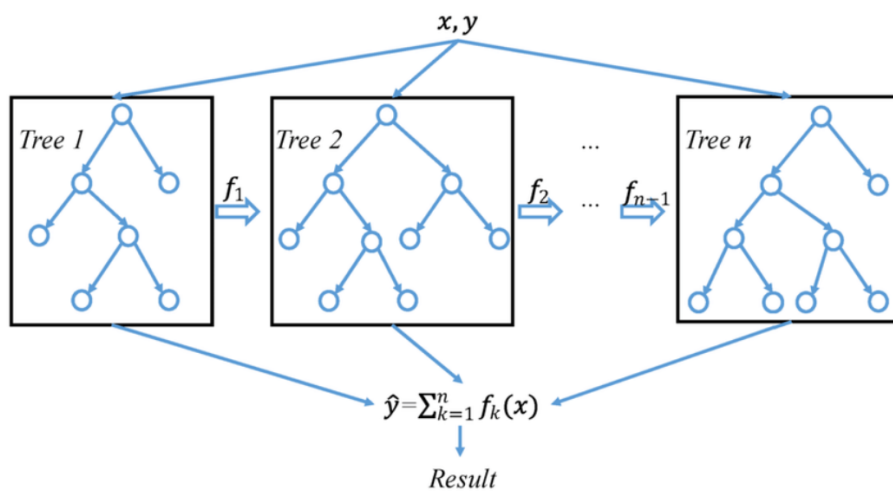
FIG. 4.
The simple principle of the XGBoost [11]

It is shown in Fig 4 that XGBoost is an the algorithm moves through iterations, it learns from the residuals of its neighbors. Rather than accepting the majority of the prediction results in Random Forest, it produces a more accurate prediction in this algorithm [1].

$$\hat{y_i} = \sum_{k=1}^{n} f_k(x_i), \qquad f_k \in F, \tag{1}$$

Where f_k In regression trees, the space of trees is defined as,
f_k Assumes the form of a tree, so $f_k(x_i)$ is the outcome of tree k, and $\hat{y_i}$ Evaluate predicted i request (x_i).

C. Statistical Analysis

The Autoregressive integrated moving average (ARIMA) is an advanced analytics technique that uses historical data and statistical models to predict future outcomes. Serial correlation is used to forecast or predict future outcomes with autoregressive integrated moving averages (ARIMA) [4].In equation 2, the process of calculating autoregressive integrated moving averages (ARIMA) is illustrated by explaining past observations and random errors.

$$K_t = \theta_0 + \varphi_1 k_{t-1} + \varphi_2 y_{t-2} \dots + \varphi_c y_{t-c} + \rho_t + \theta_1 \rho_{t-1} + \theta_2 \rho_{t-2} \dots + \theta_h \rho_{t-h} \tag{2}$$

k_t and ρ_t in Equation. (2) this is a representation of the real value and the error at a particular time interval t, respectively, and $\varphi_i(i = 1,2....p)$ and $\theta_i(i = 1,2.....h)$ During auto regression, time series is a linear regression of the p previous values accompanied by the error. Moving average MA(q) describes the current rate of a time series in terms of an error at time t, and q earlier errors.

IV. DATASET

As cloud computing grows, its traffic flow increases every day. As well as uncoordinated traffic signal control and a lack of real-time data, the constant availability of the system is a critical element that cannot be ignored. Nowadays, traffic congestion has a huge impact.

In this study, data were obtained from a South African, Sandton, and automotive industry company with an IT department. As the company outsources other IT infrastructure companies, we are dealing with Cloud computing systems within the organization that are experiencing resource allocation problems. This study collected comprehensive data from January 1, 2017, to December 31, 2022 (representing daily page loads, unique visitors, new visitors, and returning visitors).

V. SIMULATED SCENARIOS

As illustrated in Fig 5, a three-stage research design is proposed in this study as well as three predictive methods
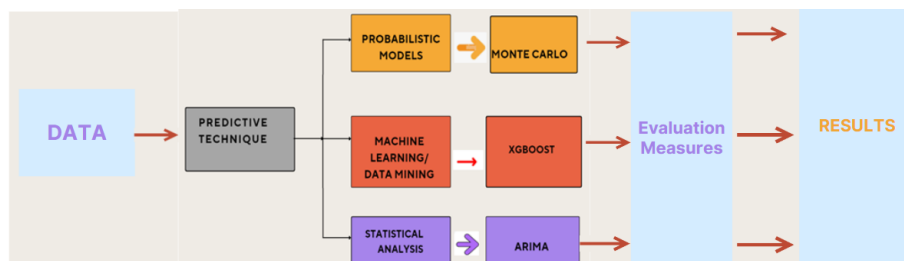
FIG. 5.
Proposed technique

The simulation was performed using the Time Series Lab and Crystal Ball.

A model receives data and predicts it, then the model is evaluated and the results are presented in Fig 5.

Method 1 Data > Prediction Techniques > Probabilistic Models > Monte Carlo Techniques > Evaluations > Results.

Method 2 Data > Prediction Techniques > Machine Learning > XGBoost Techniques > Evaluations > Results.

Method 3 Data > Prediction Techniques > Statistical Analysis Techniques> Arima> Evaluations > Results.

VI. RESULTS

A. Evaluation Measures

In evaluation measures, errors between expected and real qualities are defined numerically [15]. The difference between them indicates how well a model did. The data set presented to the network was used to evaluate the performance of the matched network. Mean absolute percentage error (MAPE) was applied to Monte Carlo, ARIMA, and XGBoost techniques in this regard. These measures are defined by equations.

$$\text{MAPE} = \frac{1}{N} \sum_{i=1}^{N} \frac{|S_i - V_I|}{V_I} \times 100 \tag{3}$$

Where

S_i Value predicted for observation i,
V_i The actual observation value i, and
N Number of observations.

These prediction analysis techniques have been used to improve cloud network traffic flow predictions. Time Series Lab, Crystal Ball, and real data were used to simulate traffic flow and determine the effectiveness of the proposed model.
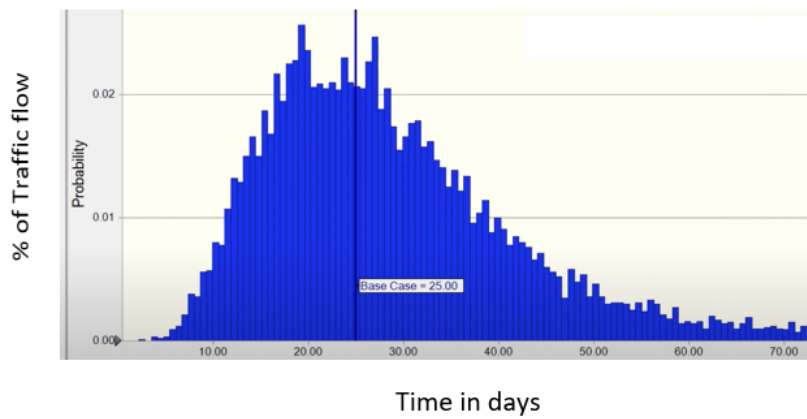
B. Monte Carlo technique results

FIG. 6.
The normal distribution of the observed data was calculated every 20 seconds according to a 2.2 minute interval.
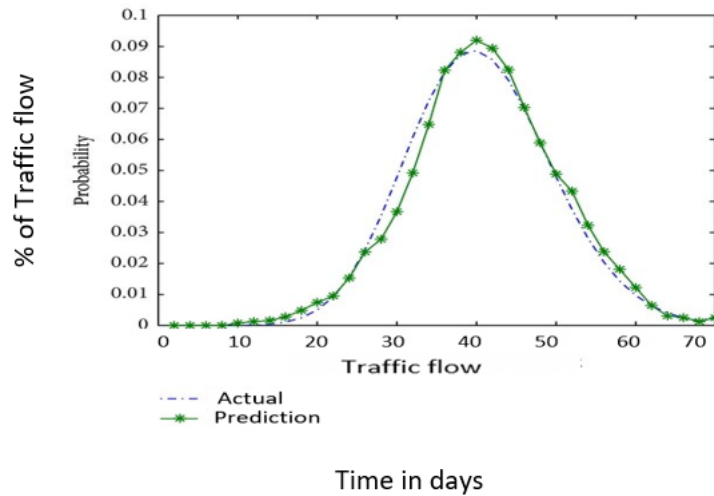
FIG. 7.
Observed value

In Fig. 7, when traffic volume was low, the model accuracy increased; however, when traffic volume increased, the model accuracy also decreased, and the prediction was 84%, Fig 7 illustrates Monte Carlo results for the observed data. The normal distribution for the observed data was calculated every 20 seconds according to a 2.2 minute interval as explained in Fig 8. The MAPE prediction errors is 4.49 %.

C. Autoregressive Integrated Moving Averages (ARIMA)

As shown in Fig 8, the results of the Autoregressive Integrated Moving Averages (ARIMA) model were produced within 0.40 seconds, which was 0.40 seconds longer than the previous model. The MAPE prediction error is 5.59 %.
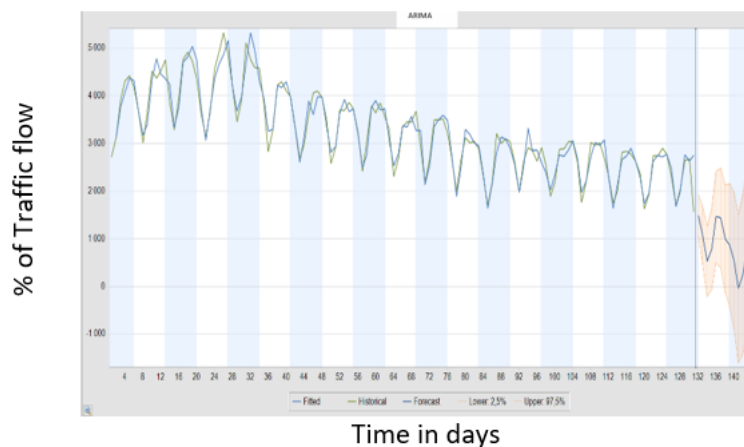
FIG. 8.
Autoregressive Integrated Moving Averages (ARIMA)

In Fig. 8, when traffic volume increased, the model accuracy increased; however, when traffic volume decreased, the model accuracy also decreased, and the prediction was 79,8%.

D. Extreme Gradient Boosting Regression (XGBoost)

As shown in Fig 9, extreme gradient boosting (XGBoost) reduced training time to 10 seconds, however, prediction accuracy decreased with a minimum in error evaluation of MAPE 5.88
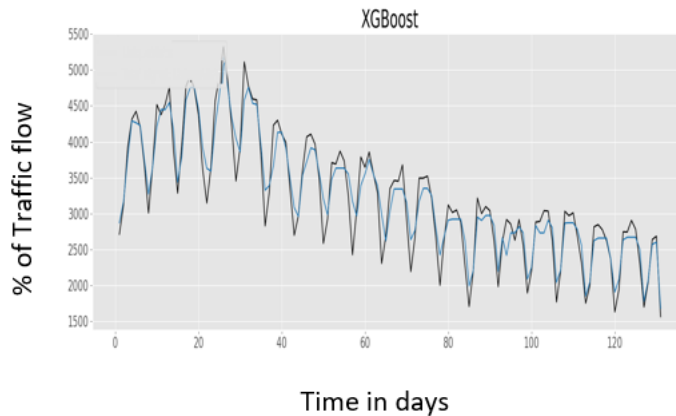

FIG. 9.
The XGBoost prediction.

In Fig. 9, when traffic volume increased, the model accuracy increased; however, when traffic volume decreased, the model accuracy also decreased, and the prediction was 71,5%.

VII. DISCUSSION

TABLE I.
Comparison of Prediction Technique Errors

Comparison of Techniques Errors			
Evaluation Items	Monte Carlo techniques	ARIMA	XGBoost
MAPE	4.49	4.59	5.88
Training time	~120 obs/sec	~0.40 obs/sec	10.43 obs/sec

In order to identify the best performing techniques, prediction analysis techniques, Monte Carlo techniques, ARIMA techniques, and XGBoost techniques were applied to a real dataset from an online cloud networking application system experiencing traffic congestion.

From 2017 to 2022, a comprehensive set of data was collected on , unique visitors, first-time visitors, and returning visitors on a daily basis.In Table 2, the Monte Carlo training method showed better performance than the ARIMA and XGBoost techniques. It improves prediction accuracy with the minimum amount of errors. Hence, Monte Carlo techniques should be incorporated into traffic prediction models or architectures.

VIII. CONCLUSION

The quality of service can only be improved if future workloads are accurately forecasted and resources allocated optimally. By using this predictive analysis, cloud providers can prevent a wide range of losses, such as service outages, excessive or inadequate provisioning of cloud resources, and customer losses.

This paper compares three predictive data analytics techniques to determine how well a cloud computing system predicts future traffic parameters based on their data analytics. A bibliometric analysis review is also conducted to analyses the focus area of the proposed technique, and the findings will contribute to the design of a predictive model for managing cloud computing resources. As a result of traffic control systems' ability to predict future values of traffic parameters, their performance can be improved.

Based on bibliometric analysis, it has been found that the proposed methods are relevant for predicting cloud computing traffic flow, and Monte Carlo techniques are more efficient, than Extreme Gradient Boosting (XGBoost) and Autoregressive integrated moving averages (ARIMA) for analyzing data at a yearly level. Using hourly, daily, weekly, and monthly traffic predictions, future work will determine the most efficient time to allocate cloud computing resources. In addition, other predictive techniques will be included in order to determine which ones will be the most effective in determining the best real-time resource management in cloud computing systems.

A Monte Carlo prediction of 84% outperformed ARIMA's prediction of 79.8% and XGBoost's prediction of 71.5%, indicating that Monte Carlo is more accurate than other models when predicting traffic flow in organizational cloud computing systems. A machine learning model will be used for future studies, along with hourly monitoring and resource allocation.

References

[1] Zheng, M., Huang, R., Wang, X., Li, X. Do firms adopting cloud computing technology exhibit higher future performance? A textual analysis approach. Journal of International Review of Financial Analysis, 90, (2023) . [Online]. Available:

[2] Apat, H. K., Nayak, R., Sahoo, B. A comprehensive review on Internet of Things application placement in Fog computing environment.Journal of Internet of Things, 23,(2023).[Online]. Available:https://doi.org/10.1016/j.iot.2023.100866.

[3] Cheng, M., Qu, Y., Jiang, C., Zhao, C. Is cloud computing the digital solution to the future of banking? Journal of Financial stability,63,(2022).[Online]. Available:https://doi.org/10.1016/j.jfs.2022.101073.

[4] Luo, J., Gong, Y. Air pollutant prediction based on ARIMA-WOA-LSTM model. journal of Atmospheric Pollution Research,14,(2023).

[5] Jing, J., Magnin, I.E., Frindel, C. Monte Carlo simulation of water diffusion through cardiac tissue models. journal of Medical Engineering and Physics.120,(2023).[Online]. Available:

[6] Alés, A., Lanzini, F. Modelling of chemical and magnetic order in Ni-Mn-Al shape memory alloys using Monte Carlo simulations.Journal of Journal of Magnetism and Magnetic Materials, 285,(2023). [Online]. Available:https://doi.org/10.1016/j.jmmm.2023.171110

[7] Meerasri , J., Sothornvit, R. Artificial neural networks (ANNs) and multiple linear regression (MLR) for prediction of moisture content for coated pineapple cubes. Journal of Case Studies in Thermal Engineering,33,(2022). [Online]. Available: https://doi.org/10.1016/j.csite.2022.101942

Deng, T., Wu, J. Efficient graph neural architecture search using Monte Carlo Tree search and prediction network.Journal of Expert Systems With Applications,213, (2023). [Online]. Available: https://doi.org/10.1016/j.eswa.2022.118916

[9] Lee, K., Im, S., Lee, B. Prediction of renewable energy hosting capacity using multiple linear regression in KEPCO system.Journal of Energy Reports,9 , pp.: 343-347 (2023). [Online]. Available:https://doi.org/10.1016/j.egyr.2023.09.121.

[10] Afrasiabian, B., Eftekhari, M. Prediction of mode I fracture toughness of rock using linear multiple regression and gene expression programming.Journal of Rock Mechanics and Geotechnical Engineering, 14 , pp.: 1421-1432 (2023). [Online]. Available: https://doi.org/10.1016/j.jrmge.2022.03.008.

[11] Silagyi, D.V, Liu, D. Prediction of severity of aviation landing accidents using support vector machine models. Journal of Accident Analysis and Prevention,187, (2023) [Online]. Available:https://doi.org/10.1016/j.aap.2023.107043.

[12] Sharma, R., Awasthi, A. An embedded element based 2D finite element model for the strength prediction of mineralized collagen fibril using Monte-Carlo type of simulations.Journal of Biomechanics,108,(2020)[Online]. Available:https://doi.org/10.1016/j.jbiomech.2020.109867.

[13] Nadjafi, M., Gholami,P. Probability fatigue life prediction of pin-loaded laminated composites by continuum damage mechanics-based Monte Carlo simulation. Journal of Composites Communications, 32,(2022). [Online]. Available: https://doi.org/10.1016/j.coco.2022.101161

[14] He, H., Fan, Y. A novel hybrid ensemble model based on tree-based method and deep learning method for default prediction. journal of Expert Systems With Applications, 176,(2021) .[Online]. Available :https://doi.org/10.1016/j.eswa.2021.114899

[15] Santhusitha, D., Karunasingha, K.: Root mean square error or mean absolute error? Use their ratio as well: Information Sciences, 585, pp.: 609-629 (2022). [Online]. Available: https://doi.org/10.1016/j.ins.2021.11.036.