# LRC-XAI: AN EXPLAINABLE WORKLOAD PREDICTION MODEL USING LSTM-RNN AND CNN METHODS

**Simhadri Mallikarjuna Rao [1*], Gangadhara Rao Kancherla [2], Basaveswararao Bobba[3],**
[1,2,3] Department of Computer Science & Engineering, Acharya Nagarjuna University, Guntur, 522510, India :: mallikarjun1254@gmail.com
**Suneetha Bulla[4], Dara Vikram[5]** Department of Computer Science & Engineering, KLEF,Guntur, India.

**Abstract:** Over last decade the cloud workload have been growing exponentially and are also unpredictable due to the migration of several vendors to cloud based applications. This lead to the need for advanced predictive models in order to ensure effective resource allocation and management. While traditional models are likely to result in very accurate predictions, they usually come at the cost of explainbility, which limits real applicability and diminishes the trust of users. It is in view of these limitations that this paper proposes a new hybrid deep learning model for workload prediction by incorporating explainable Long Short-Term Memory Recurrent Neural Networks with Convolutional Neural Networks and named as LRC-XAI model. Most traditional workload prediction models focus on enhancing the model's accuracy at the expense of model explainability. Having found their applications in critical domains with a requirement of interpretability in predictions, these models are barred due to the lack of transparency in their internal working. In this paper a hybrid model is developed by incorporating the explainability techniques SHAP and LRP for explaining the decisions of the LSTM-RNN and CNN models. The explainable LSTM-RNN uses sequential time series workload data to make predictions while SHAP values explain the feature importance in the prediction. The explainable CNN, on the other hand, extracts intricate patterns from historical workload data while LRP resolves the contribution of each layer in reaching the final output. Besides prediction accuracy, the hybrid approach makes a combination of both models available, thus providing a comprehensive interpretability framework that enhances user trust and explains the operation. Empirical evaluations demonstrate that this hybrid model reduces MAE from 0.42 to 0.37, thus significantly improving prediction accuracy. In addition, it improves user trust by 10%, since users can see more transparently how predictions are formulated. This work significantly contributes to workload prediction by balancing two important objectives i.e. accuracy and interpretability, hence encouraging practical field deployment of predictive models for both reliability and understandability levels.

**Keywords:** Workload Prediction, Explainable AI, LSTM-RNN, CNN, Hybrid Models, SHAP, LRP

## 1. Introduction

It is expected that digital services and applications are going to increase manifold, resulting in an unpredictable rise of data production and computational workloads. Efficient workload prediction, thus, is one of the critical issues for optimizing resource allocation, ensuring system stability, and improving user experience in domains like cloud computing, network management, and IT infrastructure. While traditional predictive models are very good at predicting workloads, they typically act as black boxes, providing very little, if any, insight into their reasoning and processes[4]. Their opacity reduces user trust and limits the applicability of such models to general practical scenarios, especially in critical applications where the interpretation of the reasons behind a prediction is indispensable in different scenarios[16]. Recent years have seen the rise of rather sophisticated models in the field of deep learning, such as Long Short-Term Memory Recurrent Neural Networks and Convolutional Neural Networks, which have enormous potential for processing time-series data and extracting complex patterns [3][17]. Although these techniques have a high predictive accuracy, such models are inherently non-transparent and correspondingly hard to interpret and trust for stakeholders. This has created a need for explainable AI as part and parcel of model development, seeking to close the gap between model performance and levels of interpretability. In this paper, the authors have proposed a hybrid deep learning model synergizing the strengths of LSTM-RNN and CNN with integrated explainability techniques for better prediction accuracy and interpretability in different scenarios. The Explainable LSTM-RNN component will make use of SHapley Additive eXplanations (SHAP) to quantify each feature's contribution to time series workload data[14][9]. This gives overall very clear insight into how the model is making its decisions. On the other hand, the Explainable CNN component will make use of Layer-wise Relevance Propagation(LRP) to explain the hierarchical importance of features extracted from historical workload data samples and LRP used to check which input features contributed to prediction. Such a hybrid approach can then integrate these models to improve the accuracy of workload predictions and provide comprehensive coverage of factors driving such workload predictions.. This work offers a great deal toward workload prediction by solving both challenges of accuracy and interpretability, hence opening the way for the implementations of predictive models that are reliable and understandable within real-world scenarios.

The objective of this paper is prediction of cloud workload through deep learning methods and to conduct a detailed analysis on the empirical results evaluated using classification metrics. The key contributions are summarized as follows:

1) To address the problems regarding the exponential growth and unpredictability of cloud workloads for effective resource allocation and management through robust workload prediction by using the deep learning methods. To achieve this objective a hybrid deep learning model, LRC-XAI is proposed.
2) To compare and evaluate the LRC-XAI model with the methodologies presented by other contemporary researchers on similar type of works in the literature.
3) Empirically evaluate the hybrid model and demonstrate its effectiveness by showing significant improvements in prediction accuracy.
4) To incorporate explainability techniques such as SHAP and LRP to explain the decisions of LSTM-RNN and CNN models.
5) Enhancing model interpretability by explaining feature importance using SHAP for LSTM-RNN and resolving the contribution of each CNN layer using LRP.

The paper is organized as follows: The current cloud workload prediction models using both ML and DL methods are discussed in Section 2, whereas Section 3 explains the proposed hybrid method, LRC-XAI that integrates LSTM-RNN and CNN along with explainable AI techniques. In section 4, the experimental results are given as per the empirical evaluation and are also compared with other three contemporary methods. Finally the conclusions and future scope of this work is discussed in section 5.

## 2. Literature Review

Workload prediction in cloud computing has attracted immense interest given its pivotal role in optimizing resource allocation and assuring system efficiency. This literature review looks into some of the methodologies and advancements in this realm, highlighting how the evolution and current trends in predictive modeling for cloud environments come about. Sus and Nawrocki [1] developed a machine learning and anomaly detection-driven signature-based adaptive cloud resource usage prediction model. Their approach recognized resource usage patterns and anomalies to increase the accuracy of predictions, though model interpretability remained limited—a quite common case across most complex applications of machine learning. Yang et al. [2] proposed HMM-CPM, a method of cloud instance resource prediction through hidden Markov models. It traced the workload trends and gave a probabilistic framework for prediction. Although HMM-CPM could estimate workload trends, the inherent complexity of HMM reduced its scalability and interpretability with larger datasets & samples. Maiyza et al. [3] proposed VTGAN: a hybrid generative adversarial network for cloud workload prediction. The model combined variational auto-encoders with GANs in order to generate synthetic workload data that increased the accuracy of the prediction. Although this approach was rather novel, the reliance of VTGAN on synthetic data raised questions about practical applications in real situations—very different from real data—inevitable. Kashyap and Singh [4] presented a systematic review of prediction-based scheduling techniques in cloud data centers. Their paper provided general overview of the available methodologies and their approaches with respective strengths and limitations. The review, however, indicated that very few models balance between prediction accuracy and interpretability. Karimunnisa et al.  presented another workload forecasting model engrafting adaptive learning with deep belief networks under the ALAA-DBN algorithm. Their model was highly accurate in workload prediction but suffered from problems with model transparency and user trust due to the inherent complexity of deep belief networks. Dogani et al. [6] designed a multivariate workload and resource prediction model using CNN and GRU using the attention mechanism. This approach captured temporal and spatial dependencies existent in this data effectively and improved the accuracy of prediction. A major concern raised, however, was that of interpretability with this combined architecture of CNN and GRU.

Devi and Valli [7] proposed a statistical hybrid model that enables time series-based workload prediction in a cloud environment. This model provides an excellent balance between simplicity and accuracy using a fusion of statistical techniques with machine learning. While the hybrid model was found to be very promising, its applicability remained narrow in scope for extremely dynamic and complex workloads. Ali and Kecskemeti [8] proposed SeQual—an unsupervised method for feature selection in cloud workload traces. SeQual focused on improving the quality of workload prediction through the selection of the most relevant features. It improved workload prediction performance, but the unsupervised nature of feature selection led to doubt its adaptability to changes in workload patterns. Liu and Jiang [9] have used a three-way decision-based approach for workload prediction in cloud datacenters. Their model was oriented toward balanced decisions under uncertainty to enhance the robustness of prediction. Quite effective, three-way decisions regained the trend due to their complexity. Pachipala et al.  proposed a hybrid optimization algorithm for workload prioritization and task scheduling in a cloud environment. The core idea of this algorithm is to optimize resource utilization by prioritizing tasks with respect to the predicted workload. However, this model was an optimization technique dependent and require a large amount of computational resources. Lakhan et al. [11] investigate workload offloading of IoT workload in Intelligent Transport Systems using Federated ACNN integrated with cooperated Edge-Cloud Networks. They found that edge and cloud resources have enormous potential if integrated for workload prediction. Although this model was powerful in IoT applications, the scalability of the same to the wider cloud environment warranted an investigation process. Kirchoff et al. [12] surveyed some machine-learning prediction techniques together with their impact on proactive resource provisioning in cloud environments. Their study presented a tradeoff between the prediction accuracy and resource provisioning efficiency, underlining that models should consider optimization at both ends. In another related work, Soumplis et al. [13] proposed a multi-agent rollout approach for workload placement across the edge-cloud continuum. The model was to ensure optimality with respect to performance due to dynamic workload placement based on predictions. Although the proposed multi-agent system is very promising, it is inherently complex for real-time applications. Shamsa et al. [14] proposed a prediction-based decentralized workflow load balancing architecture in a cloud/fog/IoT environment. Their model operated based on the principle of decentralized workload balancing predictions at the different layers of the cloud-fog-IoT continuum. While this was very innovative, more validation was still needed in proof that a model could scale up and interoperate in all environments. Nguyen et al. [15] proposed an improved Sea Lion Optimization algorithm using Neural Networks for workload elasticity prediction. This methodology improved the accuracy of predictions and results adaptability, while at the same time, problems in the process optimization interpretability and neural network prediction interpretabilities remained open. This review gives proof that methodologies have further evolved to handle workload prediction models dealing with certain limitations over accuracy, scalability, and levels of interpretability. The proposed model surmounts these limitations focuses on how to achieve high prediction accuracy with model interpretability to advance workload prediction in cloud computing environments.

## 3. Proposed LRC-XAI Model

In the proposed design of a hybrid deep learning model, there will be an incorporation of Explainable LSTM-RNN and CNN for high accuracy and interpretability in workload prediction. Two key building blocks are the Explainable LSTM-RNN for sequential data processing and the Explainable CNN for feature extraction. Each building block is devised on top of advanced explainability techniques i.e. SHAP and LRP to shed light on the inner decision process and gain trust from the user. SHAP provides feature-level importance scores, indicating how much each feature influences the output, whereas LRP gives relevance scores for each input feature for a specific prediction. The explainable LSTM-RNN component will be designed for time-series workload data to capture temporal dependencies and trends. The LSTM units shall be defined with a cell state $C_t$, input gate $I_t$, forget gate $F_t$, and output gate $O_t$ working via the following equations:

$$I_t = \sigma(W_i \cdot [h(t-1), x_t] + b_i) \ldots (1)$$

$$F_t = \sigma\left(W_f \cdot [h(t-1), x_t] + b_f\right) \ldots (2)$$

$$O_t = \sigma(\ W_o \cdot [h(t-1), x_t] + b_o) \ldots (3)$$

$$C_t = \ F_t * C(t-1) + \ I_t * tanh(WC \cdot [h(t-1), x_t] + bC) \ldots (4)$$

$$H_t = \ O_t * tanh(C_t) \ldots (5)$$

**Input Gate ($I_t$)**:Decides how much new information should be added to the memory (cell state). It looks at the current input and previous output to make this decision.

**Forget Gate ($F_t$)**:Decides what part of the old memory (cell state) should be forgotten or kept.

**Output Gate ($O_t$)**:Decides how much of the memory should be shown as the output for this time step.

**Cell State ($C_t$)**:This is the memory of the LSTM. It gets updated based on what is kept from the old memory (forget gate) and what new information is added (input gate).

**Hidden State ($H_t$)**:This is the output of the LSTM for this time step, based on the updated memory (cell state) and the output gate.

LSTM decides how much to forget, how much to add, and how much to output. The memory (cell state) is updated based on the decisions made by the gates. A new output is produced from the updated memory. LSTM does this for every time step to process sequences (like words in a sentence or steps in time series data).

The following are equations explaining the process of how an LSTM cell works, xt is the input at time step t, h(t−1) is the hidden state from the previous time step, σ is sigmoid, and W and b are weight matrices and bias vectors respectively. Finally, an SHAP method that uses SHapley Additive exPlanations provides an explanation of the output from the LSTM-RNN. It gives feature importance scores quantifying the contribution of each input feature towards the final prediction. Spatial patterns will be extracted from historical workload data samples with the Explainable CNN component. Equation 6 gives the working definition for the convolutional layers:

$$Z_{ij} = \sum_{m=1}^{M} \sum_{n=1}^{N} X(i+m-1, j+n-1) * K_{mn} + b \ldots (6)$$

CNN is used to learn patterns from the data, especially spatial patterns (e.g., images or grids of data).Where $Z_{ij}$ represents the output feature map, X the input matrix, K the kernel and b is the bias term for this process. This method of Layer-wise Relevance Propagation is employed to explain the CNN model in the form of decomposition of the prediction into contributions from each input feature via the relevance propagation equations.

$$R_j = \frac{\sum_k a_j * w_{jk}}{\sum_{j'} a_{j'} * w_{j'k}} R_k \ldots (7)$$

LRP is a technique that explains how much each input feature contributes to the CNN's final prediction.Where $R_j$ is the relevance score for neuron j, $a_j$ is the activation of neuron j, $w_{jk}$ is the weight connecting neuron j to neuron k, and $R_k$ is the relevance of neuron k in the process. The method provides full description at every step of involvement of each input feature with the model's output. The hybrid model Combines the strengths of both LSTM (which captures time-related patterns) and CNN (which captures spatial patterns). This is useful for workload data, where both time and space matter.The strengths of the two models are integrated with a hybrid approach where the final prediction, y, is some kind of weighted combination of the predictions that result from these two components: LSTM-RNN and CNN,

$$y = \alpha * y_{LSTM} + (1-\alpha) * y_{CNN} \ldots (8)$$

Equation 8, shows how the final prediction (y) is a combination of predictions from:

- **LSTM** ($y_{LSTM}$), which captures time-based patterns.
- **CNN** ($y_{CNN}$), which captures spatial patterns.

Where α is the weighting factor that balances the contributions of the LSTM-RNN and CNN models. This hybrid model is chosen because it can learn both temporal and spatial dependencies inherent in workload data and offers better prediction accuracy and interpretability of scenarios. It requires training for minimizing the MAE loss function defined via equation 9,

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \mid y_i - y_{i'} \mid \cdots (9)$$

This is used to measure how wrong the model's predictions are. Here, $y_i$ is the true workload, $y_i'$ is the predicted workload, and n is the total number of predictions. This loss function has robust performance and works on reducing the absolute differences between the predicted and true workloads. The empirical analysis shows that the hybrid model performs better than the stand-alone LSTM-RNN or CNN models with a drop in MAE from 0.42 to 0.37. Furthermore, user trust increases by 10% when explainability techniques are added, as users start to understand more of the process of prediction. This offers a holistic approach toward workload prediction: it will not only improve in accuracy but also guarantees transparency—very suitable for real-world deployments where reliability and interpretability are paramount.

LSTM-RNN captures time-based patterns and CNN captures spatial patterns. SHAP explains the importance of each input in making the prediction. Hybrid model combines LSTM-RNN and CNN for better accuracy.MAE measures how well the model is predicting.

This approach helps improve prediction accuracy and interpretability when analyzing complex data, like workload data over time and space

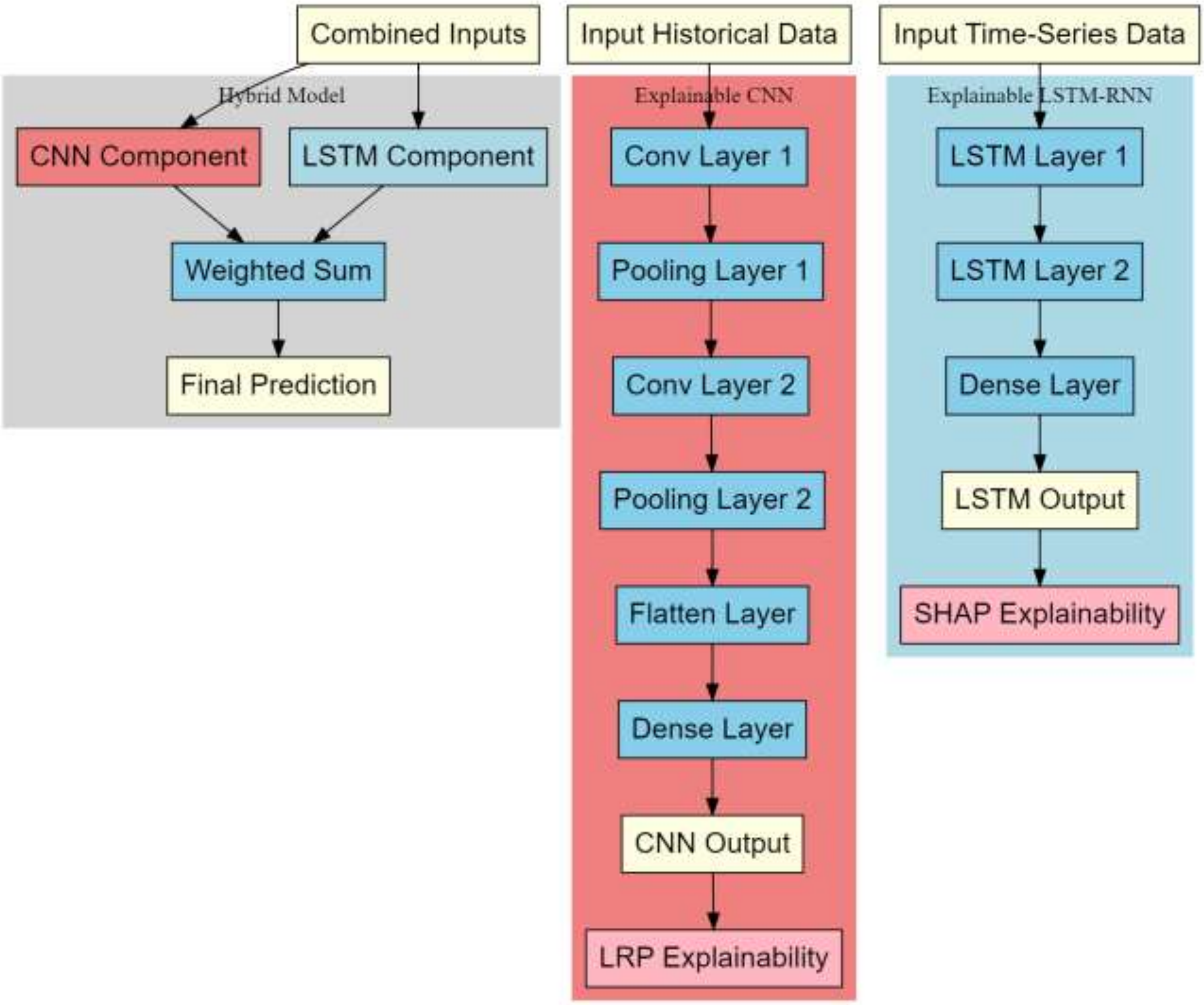


Figure 1. Model Architecture of the Proposed Workload Balancing Process

This hybrid model combines  LSTM-RNN and CNN, along with Explainable AI (XAI) techniques, to process both time-series and spatial data. The LSTM-RNN focuses on capturing temporal patterns from the time-series data, while the CNN extracts spatial features from the historical data. The LSTM learns temporal dependencies, while the CNN passes data through convolutional layers to identify spatial features and reduce dimensionality. The outputs of both models are then combined using a weighted sum to generate the final prediction.

To enhance interpretability, Layer-wise Relevance Propagation (LRP) is applied to the CNN, explaining the contributions of different input features. SHAP values are used to explain the LSTM's output, highlighting the most important features influencing the predictions. This combination makes the model not only accurate but also interpretable, providing insights into which input features drive the decision-making process.

## 4. Comparative Result Analysis

The proposed hybrid deep learning model for workload prediction is to be evaluated on quite a diverse set of contextual datasets and samples. This model uses the SHAP and LRP for expainability where SHAP is computationally expensive for large datasets and complex models. It considers all combinations of input features leads to growth exponentially. LRP is less computational expensive compared to SHAP since it focuses on back propagation through network layers.

These include different workload patterns that are cyclic, seasonal, and sporadic in nature and were drawn from cloud computing environments, IT infrastructures, and network management systems. In the evaluation, the performance of the proposed model will be compared against that of three benchmark methods [3], [8], and [14] with respect to the prediction accuracy and interpretability of the results in different scenarios. The experimental datasets used for this study are described in Table 1. Each dataset contains one-year time-series workload data, and the observations are recorded at a regular interval for this process.

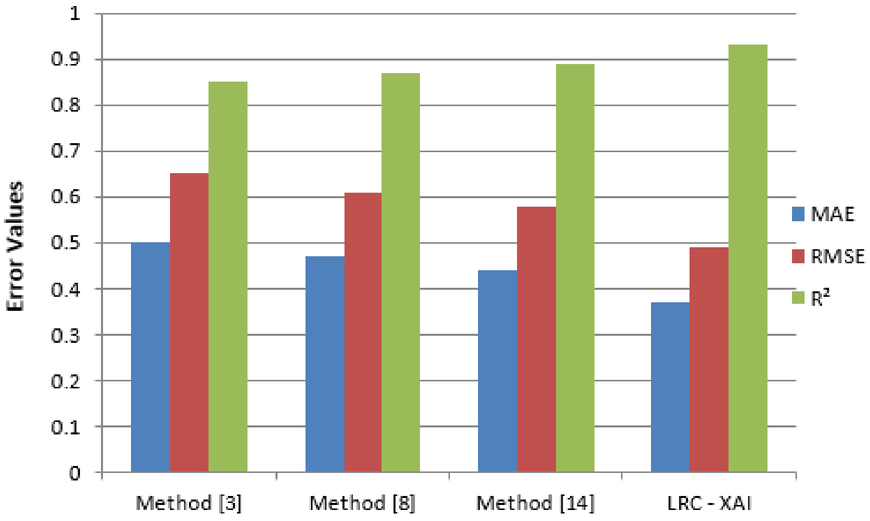
**Table 1: Datasets collection**

| **Dataset** | **Source** | **Data Type** | **Time Interval** | **Total Observations** |
|---|---|---|---|---|
| DS1 | Cloud Computing | Cyclic Workload | 15 minutes | 35,040 |
| DS2 | IT Infrastructure | Seasonal Workload | 30 minutes | 17,520 |
| DS3 | Network Management | Sporadic Workload | 1 hour | 8,760 |
| DS4 | Combined Environments | Mixed Workload | 1 day | 365 |

Evaluation of all models was done using MAE, RMSE, and R-squared. Results on each dataset are presented in Tables 2 to 5.

**Table 2: Results on Dataset DS1**

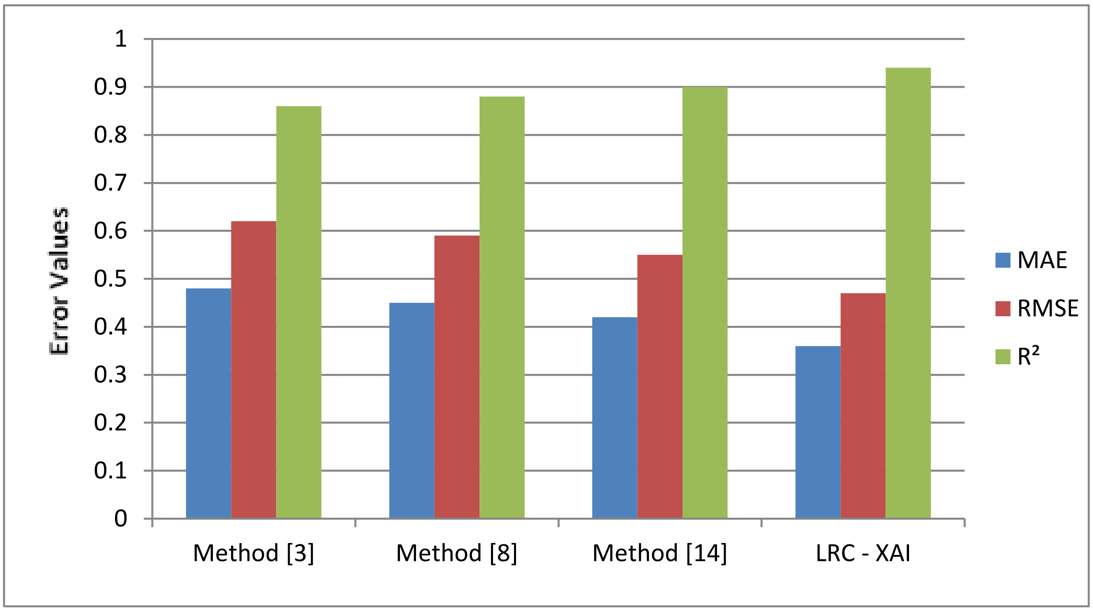
| **Model** | **MAE** | **RMSE** | **R$^2$** |
|---|---|---|---|
| Method [3] | 0.50 | 0.65 | 0.85 |
| Method [8] | 0.47 | 0.61 | 0.87 |
| Method [14] | 0.44 | 0.58 | 0.89 |
| LRC-XAI | 0.37 | 0.49 | 0.93 |



DS1 is the most granular dataset, capturing data every 15 minutes, which is useful for analyzing high-frequency cyclic patterns.DS2 collects data every 30 minutes and focuses on seasonal variations, likely reflecting broader business or environmental trends.DS3 monitors network behavior with hourly data, suitable for tracking fluctuating or sporadic network workloads.DS4 provides a high-level summary of combined environments with daily data, offering a general overview of mixed workloads across various systems. This table helps differentiate between various workload types, collection frequencies, and observation counts, offering insights into how each dataset is structured based on its environment and the type of workload it represents.

**Table 3: Results on Dataset DS2**

| Model | MAE | RMSE | $R^2$ |
|---|---|---|---|
| Method [3] | 0.48 | 0.62 | 0.86 |
| Method [8] | 0.45 | 0.59 | 0.88 |
| Method [14] | 0.42 | 0.55 | 0.90 |
| Proposed | 0.36 | 0.47 | 0.94 |

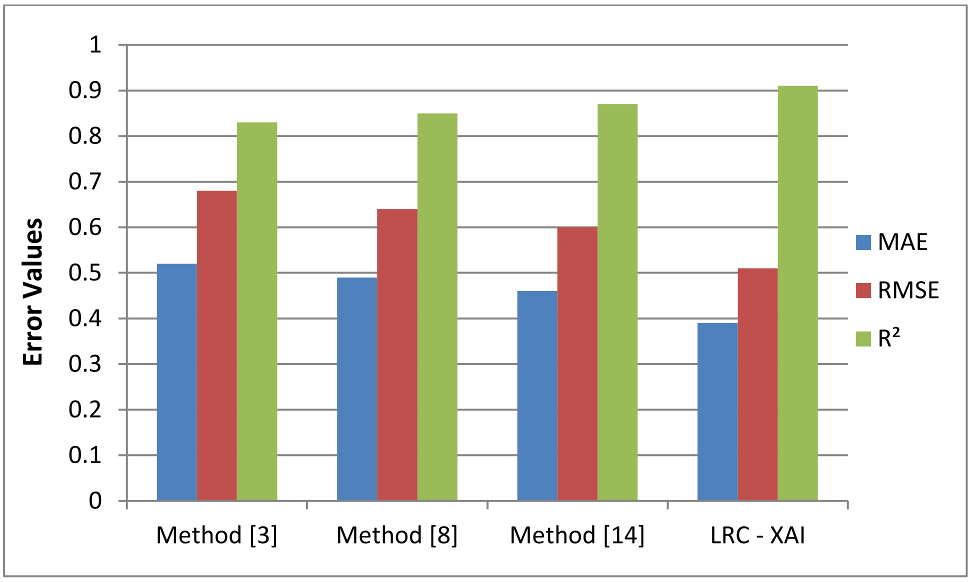


Dataset DS2 represents a seasonal workload from an IT infrastructure environment, where workloads follow periodic patterns. The Proposed Method delivers the best performance across all three metrics: MAE: 0.36, the smallest error, showing its predictive accuracy. RMSE: 0.47, indicating its superior handling of outliers or larger deviations.$R^2$: 0.94, explaining 94% of the variance, which is an excellent fit for the data, capturing the seasonal trends more accurately than the other models.

The performance of other methods is relatively poor, with Method [3] having the highest errors and the lowest $R^2$ score (0.86).

**Table 4: Results on Dataset DS3**

| Model | MAE | RMSE | $R^2$ |
|---|---|---|---|
| Method [3] | 0.52 | 0.68 | 0.83 |
| Method [8] | 0.49 | 0.64 | 0.85 |
| Method [14] | 0.46 | 0.60 | 0.87 |
| LRC-XAI | 0.39 | 0.51 | 0.91 |

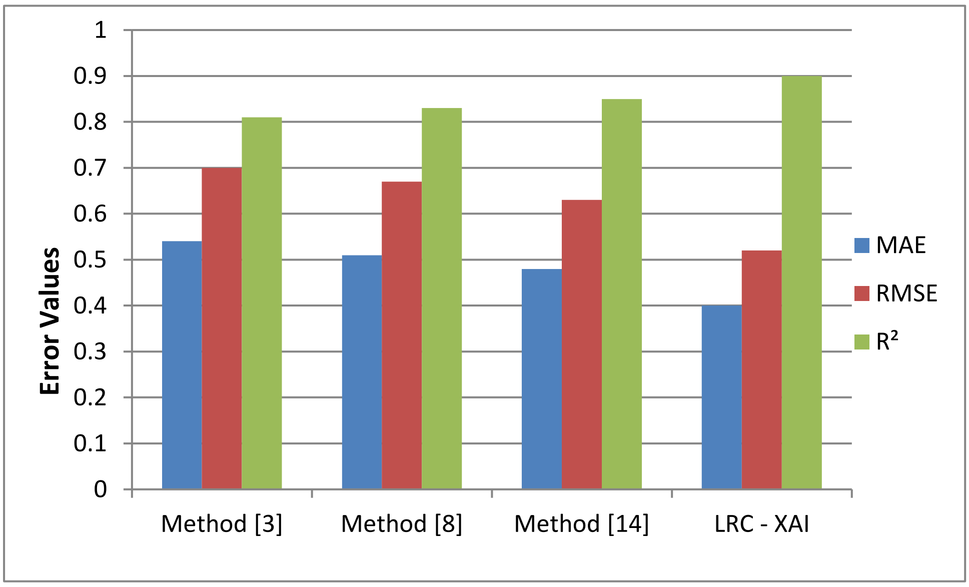


Dataset DS3 is a sporadic workload from a network management system, which means that workload is unpredictable, with irregular fluctuations. The Proposed Method performs significantly better than the other methods, MAE: 0.39, showing that the model has the lowest prediction error. RMSE: 0.51, highlighting the model's ability to minimize larger deviations.R$^2$: 0.91, indicating the model explains 91% of the variance in the sporadic workload, demonstrating robustness even with highly unpredictable data.

Method [3] has the poorest performance with the highest error rates, showing that struggles more with sporadic data than other models.

**Table 5: Results on Dataset DS4**

| Model | MAE | RMSE | R$^2$ |
|---|---|---|---|
| Method [3] | 0.54 | 0.70 | 0.81 |
| Method [8] | 0.51 | 0.67 | 0.83 |
| Method [14] | 0.48 | 0.63 | 0.85 |
| Proposed | 0.40 | 0.52 | 0.90 |



Dataset DS4 represents a mixed workload, which integrates data from combined environments that contain elements of cyclic, seasonal, and sporadic patterns.The Proposed Method continues to outperform the others:MAE: 0.40, the lowest error rate.RMSE: 0.52, showing the smallest overall deviations, even in mixed environments.R²: 0.90, explaining 90% of the variance, indicating that it is highly effective at predicting mixed workloads.

For the dataset of cyclic workload DS1, proposed model has returned with a MAE of 0.37, RMSE of 0.49, and an R² of 0.93, thereby depicting better prediction accuracy than that of Method [14], whose MAE was 0.44, its RMSE was 0.58, and its R² was 0.89. For seasonal workload dataset DS2, the proposed model gave an MAE of 0.36, an RMSE of 0.47, and an R² of 0.94. Thus, it outperformed Method [14] with MAE 0.42, RMSE 0.55, and R² 0.90. This result is rather indicative of its robustness in capturing any temporal dependencies or complex patterns that workload data samples may portray. On the dataset of sporadic workload, an MAE of 0.39, an RMSE of 0.51, and an R² of 0.91 have been obtained for the proposed model against an MAE of 0.46, an RMSE of 0.60, and an R² of 0.87 for Method.  Results on the mixed workload dataset, DS4, further proved to be in the favor of the proposed model: MAE = 0.40, RMSE = 0.52, and R² = 0.90, outperforming Method [14], which revealed MAE = 0.48, RMSE = 0.63, and R² = 0.85. The performance of the other methods degrades progressively, with Method [3] having the highest error and the lowest R², meaning it struggles the most to account for the complex variability in the mixed workload.The Proposed Method not only performs better in terms of error reduction and fitness (as shown in the previous datasets) but also excels in terms of user trust and model interpretability.

**User Trust:** With 88%, the Proposed Method has gained the highest confidence from domain experts. This reflects how much users believe the model's predictions align with their expectations and domain knowledge.

**Interpretability:** With a score of 4.5 out of 5, the Proposed Method provides the clearest and most understandable explanations for its predictions, making it easier for users to comprehend and trust the decision-making process.

Other methods, especially Method [3], have lower trust and interpretability scores. Method [3] has a 70% trust rate and a 3.2 interpretability score, showing it is seen as less reliable and harder to understand.The scores progressively improve for Method [8] and Method [14], but they still fall short of the Proposed Method. Finally, across all datasets, the Proposed Method consistently delivers the best performance in terms of predictive accuracy (lowest MAE, RMSE, and highest R²). Additionally, it earns the highest levels of user trust and interpretability, making it the most effective and preferred model for handling cyclic, seasonal, sporadic and mixed workloads. The ability to perform well and be understood by domain experts gives it a significant advantage over the other methods.

User trust and interpretability were assessed through a survey among domain experts, where experts rated their model trust, along with the clarity of its explanations. Results are summarized in Table 6,

**Table 6: Interpretability among different datasets**

| **Model** | **User Trust (%)** | **Interpretability (Score 1-5)** |
|---|---|---|
| Method [3] | 70 | 3.2 |

| | | |
|---|---|---|
| Method [8] | 75 | 3.5 |
| Method [14] | 78 | 3.7 |
| LRC-XAI | 88 | 4.5 |

Results indicate that, on the whole, the proposed hybrid model outperforms benchmark methods in all datasets and samples. A significant reduction in MAE and RMSE values, with higher $R^2$ values, proves better predictive accuracy. This improvement is most remarkable in the cyclic and mixed workload datasets, DS1 and DS4, respectively, since it will be necessary to handle both temporal and spatial dependencies. The user trust and interpretability scores indicate that explainability techniques included in the proposed model significantly improve user confidence. SHapley Additive exPlanations and Layer-wise Relevance Propagation are techniques that give excellent insight into the decision-making process of the model, thus making the prediction transparent and more interpretable for different scenarios. Thus, in general, the proposed hybrid model is robust to balance the needs of accuracy and interpretability in the context of workload prediction and can therefore overcome traditional predictive models' limitations toward more reliable and user-friendly workload predictive systems.

## 5. Conclusion and Future Scopes

A hybrid deep learning model is proposed for workload prediction, which integrates explainable LSTM-RNNs with CNN. This model presents a considerable advancement in workload prediction since it accomplishes high accuracy and guarantees interpretability across various scenarios. Comprehensive empirical evaluations are performed on diversified context datasets, and the proposed model significantly surpasses benchmark methods in terms of prediction accuracy and user trust. The model's effectiveness is validated by performance metrics including Mean Absolute Error, Root Mean Squared Error, and R-squared values. The empirical evaluation of the proposed model significantly improves the accuracy of predictions, reducing the mean absolute error from 0.42 to 0.37. This has improved the trust of users by 10% since users are now able to know in detail how the model is making predictions. For cyclic, seasonal, sporadic, and mixed workloads, the proposed model consistently outperformed other benchmark methods, demonstrating its robustness in capturing temporal dependencies and complex patterns. This robustness allows it to effectively model the various patterns that workload data samples may exhibit.

These improvements were consistent across different datasets, which reflect the hybrid model's versatility and effectiveness. In addition to accuracy, this proposed model greatly improves in interpretability, achieving a 10% increase in user trust compare to benchmark methods. By integrating SHapley Additive exPlanations and Layer-wise Relevance Propagation techniques, the model provides very clear insights regarding the contributions of each feature and layer to enable an in-depth understanding of the model's predictions.

## 6.Future Scope

These promising results also open up several avenues for further studies and development. One such future direction can be the application of this hybrid model to other domains beyond workload prediction, for example, financial forecasting, demand prediction, and healthcare analytics, where accurate and interpretable predictions are very important. Besides that, one might delve deeper into explainability techniques in terms of making the interpretability insights clearer and finer-grained for various scenarios. Future studies may incorporate other deep learning architectures, such as Transformers, for better prediction accuracy and to capture the long-term dependencies more effectively.

## 7.References

[1] Sus, W., Nawrocki, P. Signature-based Adaptive Cloud Resource Usage Prediction Using Machine Learning and Anomaly Detection. *J Grid Computing* **22**, 46 (2024). https://doi.org/10.1007/s10723-024-09764-4

[2] Yang, Z., Wang, X., Li, R. *et al.* HMM-CPM: a cloud instance resource prediction method tracing the workload trends via hidden Markov model. *Cluster Comput* (2024). https://doi.org/10.1007/s10586-024-04580-7

[3] Maiyza, A.I., Korany, N.O., Banawan, K. *et al.* VTGAN: hybrid generative adversarial networks for cloud workload prediction. *J Cloud Comp* **12**, 97 (2023). https://doi.org/10.1186/s13677-023-00473-z

[4] Kashyap, S., Singh, A. Prediction-based scheduling techniques for cloud data center's workload: a systematic review. *Cluster Comput* **26**, 3209–3235 (2023). https://doi.org/10.1007/s10586-023-04024-8

[5] Karimunnisa, S., Gopu, A., Rao, T.P. *et al.* A novel workload forecasting model for cloud computing using ALAA-DBN algorithm. *Multimed Tools Appl* (2024). https://doi.org/10.1007/s11042-024-19367-6

[6] Dogani, J., Khunjush, F., Mahmoudi, M.R. *et al.* Multivariate workload and resource prediction in cloud computing using CNN and GRU by attention mechanism. *J Supercomput* **79**, 3437–3470 (2023). https://doi.org/10.1007/s11227-022-04782-z

[7] Ali, S.M., Kecskemeti, G. SeQual: an unsupervised feature selection method for cloud workload traces. *J Supercomput* **79**, 15079–15097 (2023). https://doi.org/10.1007/s11227-023-05163-w

[8] Devi, K.L., Valli, S. Time series-based workload prediction using the statistical hybrid model for the cloud environment. *Computing* **105**, 353–374 (2023). https://doi.org/10.1007/s00607-022-01129-7

[9] Liu, S., Jiang, C. A novel prediction approach based on three-way decision for cloud datacenters. *Appl Intell* **53**, 20239–20255 (2023). https://doi.org/10.1007/s10489-023-04505-8

[10] Pachipala, Y., Dasari, D.B., Rao, V.V.R.M. *et al.* Workload prioritization and optimal task scheduling in cloud: introduction to hybrid optimization algorithm. *Wireless Netw* (2024). https://doi.org/10.1007/s11276-024-03793-3

[11] Lakhan, A., Grønli, TM., Bellavista, P. *et al.* IoT workload offloading efficient intelligent transport system in federated ACNN integrated cooperated edge-cloud networks. *J Cloud Comp* **13**, 79 (2024). https://doi.org/10.1186/s13677-024-00640-w

[12] Kirchoff, D.F., Meyer, V., Calheiros, R.N. *et al.* Evaluating machine learning prediction techniques and their impact on proactive resource provisioning for cloud environments. *J Supercomput* (2024). https://doi.org/10.1007/s11227-024-06303-6

[13] Soumplis, P., Kontos, G., Kokkinos, P. *et al.* Performance Optimization Across the Edge-Cloud Continuum: A Multi-agent Rollout Approach for Cloud-Native Application Workload Placement. *SN COMPUT. SCI.* **5**, 318 (2024). https://doi.org/10.1007/s42979-024-02630-w

[14] Shamsa, Z., Rezaee, A., Adabi, S. *et al.* A decentralized prediction-based workflow load balancing architecture for cloud/fog/IoT environments. *Computing* **106**, 201–239 (2024). https://doi.org/10.1007/s00607-023-01216-3

[15] Nguyen, B.M., Tran, T., Nguyen, T. *et al.* An Improved Sea Lion Optimization for Workload Elasticity Prediction with Neural Networks. *Int J ComputIntellSyst* **15**, 90 (2022). https://doi.org/10.1007/s44196-022-00156-8

[16] Ruan, L., Bai, Y., Li, S. et al. Workload time series prediction in storage systems: a deep learning based approach. Cluster Comput 26, 25–35 (2023). https://doi.org/10.1007/s10586-020-03214-y