OPEN ACCESS

Kashf Journal of Multidisciplinary Research
Vol: 02 - Issue 09 (2025)
P-ISSN: 3007-1992 E-ISSN: 3007-200X
https://kjmr.com.pk

# AI-DRIVEN CLOUD COMPUTING: MACHINE LEARNING MODELS FOR DYNAMIC RESOURCE ALLOCATION, TASK SCHEDULING, AND PERFORMANCE OPTIMIZATION

***[1]Khan Ikram Uddin*, [2]Wasim Akram, [3]Muhammad Qaseem Iqbal, [4]Muhammad Awais, [5]Waqas Ahmed, [6]Haseeb Sulman***

*[1]School of Automation Science and Engineering, South China University of Technology, Guangdong, Guangzhou, 510640, China.*

*[2, 4]Faculty of Information Technology University of Lahore Sargodha 40100, Pakistan*

*[3]Teesside University, London Campus, United Kingdom, Queen Elizabeth Olympic Park (Here East), 14 East Bay Lane, London, E15 2GW, UK*

*[5, 6]Department of Computer Science and IT, The university of Lahore Sargodha campus, Pakistan*
****Corresponding Author:*** 202422800065@mail.scut.edu.cn

**Article Info**



**Abstract**

This study examines feedback-based resource allocation, schedule coordination, and performance optimization using the models of Artificial Intelligence (AI) and Machine Learning (ML) in a cloud computing setting. However, classic solutions to cloud management, including fixed quantity of resources assigned and standard task scheduling, are parts of the solutions that have problems in managing the attitude of increased complexity and dynamism of the cloud workloads. With the help of AI-based models, the proposed study will enhance the efficiency and lower the costs of cloud resources, as well as improve the overall performance. Namely, the study applying those methods as reinforcement learning, deep learning, and hybrid machine learning can dynamically distribute resources relying on real-time predictions and optimizing the task schedule. These findings suggest that AI based models are better than traditional ones in important resource consumption measures like time taken to complete tasks, cost-efficiency, and resource use, and the energy usage. The results guarantee that AI can have a greater impact on the adaptability and scalability of a cloud computing system, which will result in more efficient and sustainable cloud infrastructures. Nevertheless, related issues like complexity or difficulty when model training is used, integrated with the incumbent cloud systems, and real-time adjustability have to be solved in order to be effective with the potential of AI in cloud resources management.

**Keywords:**

*Artificial Intelligence, Machine Learning, Cloud Computing, Dynamic Resource Allocation, Task Scheduling, Performance Optimization, Reinforcement Learning, Deep Learning, Hybrid Machine Learning, Cloud Efficiency, Cost Reduction, Energy Consumption.*

## INTRODUCTION

Another revolution that has prevailed among the information technology (IT) industry is the concept of cloud computing whereby business and individuals can access and even save data and applications on the internet as compared to hardware. This has destabilised the IT infrastructures by providing scalability, elasticity besides cost-efficiency of computing resources. Demand has increased, in ensuring that a system controls resources adequately to support operations mission critical applications and workloads as organizations deem cloud based systems. As the system has occurred in the environment marked by constant and unpredictable changes, this causes inefficiency, high cost, and failure to achieve optimum performance because of the rigidity in the methods of distributing the resources (Zhao et al., 2023).

Machine Learning (ML) and Artificial Intelligence (AI) has been mentioned as the most important technology that will allow addressing limitations of the traditional cloud computing architecture and enable systems to make real-time and data-driven decisions related to dividing resources and scheduling. The AI models are capable of dynamically reacting to the variations in the workloads and doing so based on the maximum provision of resources and the optimal enhancement of performance of the system under minimal costs of operation. Some of the studies have covered the idea of the application of AI in cloud computing wherein the history of its use has been recorded in terms of resource placement, jobs scheduling and optimization. They include reinforcement learning (RL), deep learning (DL), and ensemble, which are advanced algorithms to enhance the process of cloud infrastructure management (Zhou and Li, 2024).

Dynamic resource allocation is an element of cloud computing that is interested in the effective sharing of computing resources e.g. CPU, memory and storage to meet the demands of the various applications and users. The standard non-dynamic methodologies do not fit dynamic attributes of the cloud environment because cloud ready resources are oversued whoever is either under provisioned or over provisioned. Predictive means of forecasting the future condition of work demand (i.e. deep learning and time series forecasting) have been used to solve this issue, being within the AI models to realize resource allocation. These predictive algorithms enable cloud service providers to distribute its resources, and in the process, ensure that the applications do not utilize other resources they are not bounded by, but just the requirements (Liu et al., 2024).

In addition to the allocation of financial means, AI also has a significant part regarding task scheduling in the cloud environment. Task scheduling is relevant in ensuring that resources are used in optimal place as far as the timing and the place that determines the course of action on the tasks can be achieved. intelligent scheduling: AI-based scheduling algorithms compute the available data in the form of system performance measurements and workload properties to determine intelligent placements of making sure that the resources are used optimally. Based on this example, the reinforcement learning algorithms could be incessantly fed by past decision and seek the way of optimising the numerous scheduling policies and lead to the total improvement of the system (Xu et al., 2024). Moreover, synergistic application of different AI approaches is also proposed to play an additional role in enhancing the greater efficiency of Scheduling the tasks and adapting them to the heterogeneous clouds (Sanjalawe, 2025).

Cloud computing performance optimization is another branch of AI that has managed to be efficient. Balanced load of resources is possible to optimize the performance of cloud in terms of cost and aspect as well as the quality of service (QoS). The AI models which in particular have adapted the ML algorithms can monitor the performance at the period of the system operating, and implement the adaptive changing in the resources allocation and a sequence of the tasks with the purpose of optimizing the performance. To illustrate the above, energy-efficient AI-powered scheduling algorithms have been used to make sure that the cloud data centers consume less energy, though cost-effective (Chakraborty et al., 2024). The methods contribute to the realization of sustainability goals by efficiency of the system by reducing wastes of energy.

Despite such a spate of advantages of AI-based cloud computing, it is true that there are several concerns that have to be addressed to enable the same to gain acceptance. The opportunity to train AI models, working under different environment in clouds is one of the main challenges. Models prefer to learn with huge numbers of data, and disclosing precise,ativated exhaustive and true-to-life events were learned by the models must remain vital in the functioning of the model. In addition, elasticity on time is one of the key points as well, and cloud systems cannot stand still, and AI solutions should have the capability of responding quickly to the fluctuations (Patel et al., 2025). Besides this, AI development on the existing cloud systems is technically challenging regarding compatibility of the systems, integration of data and benchmarking of the performance.

However, numerous advantages of AI in cloud computing against the future subject have endless nominations. As the number of AI-based technologies continues to increase, cloud-providers/companies are being granted amplified chances to apply machine learning procedures to decisively allocate resources more economically and plan the timing of distinct activities and optimize the work of the whole complex. Such enhancements will guarantee further positive volumes of enlargeability, economy, and sustainability of the cloud computing services (Yang et al., 2024).

## Literature Review

The history of cloud computing has had very great advantage to many industries through the avenue of providing large scale, flexible, and cost effective IT infrastructure. Nevertheless, from an increase in the complexity and dynamism of workloads and demands, it has been shown that the existing traditional approaches to resource management, task scheduling, and performance optimization do not serve as adequate tools in managing the high variability and complexity of the cloud environments. Artificial Intelligence (AI) and Machine Learning (ML) application has now been a vital part of overcoming the aforementioned issues and major shifts to cloud resources operations have already been achieved. This research literature review strives to discover the application of AI and ML in the dynamic allocation of resources, scheduling of tasks, output optimization in cloud computing, as well as models and algorithms that were given prominence in earlier research.

### Dynamism in Resource Assignment in Cloud Computing.

One of the most important elements of cloud computing is dynamic resource allocation as this parameter directly influences efficiency and cost-effectiveness of cloud system. The workloads in the cloud are

normally unpredictable, and as a result, dynamic allocation of resources is not appropriate in the context of resource management. The most recent number of studies have dedicated their attention to the use of AI methods to forecast and distribute available resources in real-time based on requirements.

Regression analysis, the predictions made with AI, including time series prediction or deep learning have found extensive application to predict the workload requirements and shorten or increase the resources movements based on the requirements. In the example of Kumar et al. (2023), a resource allocation system based on AI has been suggested, which incorporates deep learning capabilities to forecast the demand in the future and proactively allocate resources in the cloud setup. Their model showed that resource overprovisioning has been greatly reduced and there is the general performance of the system.

Reinforcement learning (RL), where the algorithm is used to allocate resources is another promising practice. The RL algorithms enable the system to learn the environment and make most of the available resources since they make the decisions according to the feedback of the actions taken earlier. As supported by Suryawanshi and Rao (2024), reinforcement learning has demonstrated strong capacity in introducing each resource in response to the changes in workload dynamically. They have focused on the adaptive quality of RL, in which the system will constantly learn to promote resource allocation within a more efficient manner with new data pushed to it, making the systems both more resource-efficient and cost-effective.

Moreover, another suggestion, hybrid AI models, whose methods combine various machine learning approaches, has been suggested to increase the precision of the resource allocation decision-making process that has become dynamic. According to the study conducted by Lee et al. (2024), the authors employed decision tree algorithms in combination with deep neural networks to forecast resource demand and assign cloud resources dynamically. It was found that the hybrid model was successful compared to traditional methods because it yielded a greater predictive accuracy and less wastage of resources.

**Task Scheduling in Cloud Computing**

Another very important issue, when optimization of cloud performance is concerned, is efficient task scheduling. Task scheduling is defined as the proficiency of allocating the tasks to be executed on the available resources in a manner that ensures that the resources used are to optimize resource utilization and also reduce games time of execution and cost. The conventional way of scheduling usually fails to reflect the dynamic changes in the workloads thus has poor performance.

AI-based scheduling algorithms have come up to solve these issues. Machine learning methods of schedule have become popular because of their capability to learn and changing workloads. Sharma and Singh (2024) state that cloud task scheduling has been implemented based on machine learning estimating algorithms, including support vections machine (SVM) and decision tree algorithms, which helps the application to work better because of the dominant peculiarities of a task and a resource.

Furthermore, Tran et al. (2024) aimed at investigating the application of the swarm intelligence approach, specifically the Particle Swarm Optimization (PSO) methods, to infer practical scheduling of the numbers of tasks by the cloud using machine learning models. The hybrid model showed that it could properly plan

work, reduce its execution time, and ensure the distribution of the load among the available resources. The findings provided significant decrease in time of resource utilization and task completion, especially in the case of heterogeneous cloud environment where resource availability was heterogeneous.

As well, deep reinforcement learning has demonstrated potentials of high achievement in scheduling of tasks. Deep RL models have the ability to modify the scheduling policy of the tasks automatically with respect to real-time feedback. Deep RL algorithms have been studied as research in multi-cloud settings, wherein they offer dynamically based multi-cloud assignment to minimize load balancing and decrease latency (Zhang and Lin, 2025).

**Performance Optimization in Cloud Computing**

The optimization of cloud performance involves balancing of a number of variables, which may be the utilization of resources, time taken to run, utilization of energy, and even quality of the service. The performance optimization shall make an attempt to make the cloud systems more efficient without compromising the easy service delivery to the final beneficiaries. Automation of this process with the help of AI and ML can be discussed as the significant contributors as well, as the analysis of the metrics, as well as the effectual continuous change, is performed on the basis of the real-time.

One of the applications where AI can be used in energy efficient optimization entails a high utility at a low power consumption rate. Chen et al (2024) indicate that energy optimization through AI-based model makes predictions which are founded on predictive algorithms and the prediction entails forecasting the demands in the data centres of the cloud which ensure that the cloud data centres will mostly utilise minimum energy without causing significant adverse effect on its operational parameters. Their studies have found that AI is useful in reducing energy consumption up to 20 and simultaneously the service-level agreements (SLAs) are not violated.

The other performance optimization area is the cost efficiency, reductions of which AI models can calculate the cost of cloud resource by decreasing the rate of wastage on resources. The approach to minimize costs came to be developed in a study by Gupta and Sharma (2025), which utilized the machine learning algorithm to parameterize the cost of resources and fill the resource allocation decisions with the estimated cost of resources using the information at a single instance. The model was capable of achieving high cost savings due to appropriate provisioning of the resources and provisioning of requirements of only such resources at a given instance.

Besides this, AI- based optimization models have also been used to improve Quality of Service (QoS) in cloud systems. Wang et al. (2024) research claim that QoS is among the parameters of the most significant cloud computing because this factor determines the experience of the user and efficiency during resource allocation. In the study, they tested the machine learning algorithm to maximize quality of service which they did by predicting the service performance degradation and on-demand loading of services to meet service demand. The optimization, which was given in the assistance of AI, according to the results, led to the considerable work of improvement of QoS, such as reduced response time and higher system reliability.

## Challenges and Future Directions

There are still few issues despite the significant progress, which has been achieved concerning the AI-based model of cloud computing. The primary risk factors include the issues of the necessary training of the AI models to address ever-changing and different aspects of the cloud-based ones. AIs require large datasets to be trained on those datasets and should realize that prior to the fruitfulness of such models, the datasets in question must be accurate, comprehensive and realistic approximations of what can be obtained in reality. According to Singh and Mehta (2024), quality of training data is another problematic area that can lead to underperformance of a specific model in case one is using wrong or incomplete data.

Another challenge is AI model scaling in massive cloud computers. Within the cloud system, there is a common practice of implementing thousands of servers and it is challenging to run AI model in such way that it can be scaled. To overcome this challenge, scientists have developed distributed artificial intelligence models which can be performed on a group of machines and, therefore, perform large volumes of data and scale efficiently in a cloud infrastructure (Patel et al., 2025).

Besides, the AI implementation based on the existing cloud systems already provided is technically flawed in terms of compatibility of the systems between each other, data interface and performance evaluation. The key aspect principle to the implementation of the AI-driven systems is introduced as a seamless integration in comparison on the AI models and the cloud infrastructure (Kumar et al., 2025). Research on this line remains in progress with some studies potentially attempting at crafting standard frameworks and protocols on AI-cloud integration.

The use of AI and ML in cloud computing has resulted in vast advancement of resource allocation responsibility, computer task completion, and optimality. The clouds system is more efficient, presumably cheaper, and adapts to the changing loads on the basis of complex specific algorithms of deep learning, reinforcement learning, and hybrids. However, despite the set of challenges linked to training, scalability, and focus on integration aspects, the opportunities of the AI-driven cloud computing are impossible to underestimate, and the subsequent research related to the field is expected to result in the emergence of new innovations and improvements in terms of managing the cloud resources.

## Methodology

The research methodology is quite systematized and aimed at testing the application of AI- based machine learning models in dynamic allocation of resources, task schedule and optimizing the work in cloud computing. The paper will be biased contributed to creating a robust framework, which will be used to run the machine learning models such as reinforcement learning (RL), deep learning (DL), and hybrid AI systems to address the problem under consideration, which is the traditional approach to cloud resource management applications. In order to achieve thorough findings in the case of realistic cloud environment, the methodology applies model development, simulation and performance test.

**Systems design and Problem Definition.**

This study will start with the definite statement of the problem looked in in the resolution of resources, organizing computation, as well as optimal performance in cloud computing. The old-fashioned methods usually operate with prescribed resource allocation, which is vulnerable to inefficiencies since the operations also change like the panel requirements and are also less adaptable. The objective of this paper is to build a smart system, with resource allocation dynamism, reassigned tasks, and optimization of the net system performance based on the machine learning models. In order to attain it, parts of the system design involve real-time observation, the assignment of work, and forecasting of workload and are driven by AI approaches.

The cloud infrastructure is part of the architecture, which is composed of a series of many disparate components that are comprised of privilege to virtual machines (VMs), compute, storage and networking. These factors work against the AI models, which relates to real-time policies of resources allocation and scheduling to monitor system performance continuously. The pertinent data points, like CPU usage, memory consumption, duration of the execution of operations, and system load, among others, serve to monitor the information structure required to train and test the machine learning models, as required.

**Gathering of Data and Pre-Processing of Data.**

Machine learning model development is an inseparable composite of the collection of data. The data will be collected in a virtualized cloud platform in this study, which will be similar to the natural cloud platforms. Among the system metrics contained in the data set would include CPU utilization, memory utilization, storage I/O and network traffic that are important in the workload prediction and optimization of resource allocation. These data also consist of features of tasks, such as project time, priority and resources, which are pertinent as far as scheduling the tasks is concerned.

Data preprocessing is carried out to eliminate the flubs and form the data that is obtained to work with machine learning models. This includes control of missing counters, control of data normalisation and determination of a group into a number. Besides this, the editing of engineering is done to ensure that it acquires relevant features that may improve performance of the machine learning models. To illustrate it, time related features i.e. trends in the degree of resources involvement in the specific time periods have been inferred so as to make the workload forecasts that accurate.

**Model Development**

In the contemporary work, three key machine learning models are developed, including the predictive models, which are founded on workload forecasting models, reinforcement learning algorithms to address dynamically the allocation of resources, and the models of scheduling tasks. Each of these models is directed at a specific field of cloud resources management.

Workload prediction model includes a deep learning model (e.g. Long Short-Term Memory (LSTM) network) that works backward to predict the number of resources required in the future based on the previous data. The LSTM networks are chosen because they are capable of extrapolating the long-term

relationships in time series data hence most appropriate to be used in work load prediction in a cloud environment where there can be variation of the demand over any given time.

Dynamic resource allocation model takes advantage of reinforcement learning system in which the system follows the general optimum resource allocation strategies through the assistance of the environment. The model is on the IQ-learning or Deep Q-Networks (DQN) according to which, based on it, the system is capable of distributing the resources dynamically as the decisions are made based on the past experience. The optimizing factor is to ascertain an optimization of resource utilization, costs or energy use or decrease of the costs and the free adaptation to potential changes in work load on the spot.

The hybrid machine learning frameworks are constituted to carry out the task of scheduling, which entails the application of the clustering techniques alongside the task scheduling algorithms. One example is that, the k-means cluster is used to cluster resources and represent the time of these resources used by tasks and the decision trees algorithms are used to compute the scheduling decisions. The hybrid model will reduce the time of undertaking the tasks and also balance the resources at place with the ones available to ensure that the number of tasks that are undertaken will be efficient.

**Simulation Environment**

Prolonging to verify the efficiency of the created AI models, a simulated environment is created in the cloud simulation environments, i.e. CloudSim or OpenStack. The consistency Mirrors the simulation can be characterized by cloud infrastructure in varying types of virtual machines, types of resources, and workloads. The scalability and flexibility of the models within the models of different cloud systems is worked out using the different work loads to demonstrate the applicability of the models in the real world cloud services by web hosting, data analytics, as well as scientific computing.

The simulations are somehow designed in such a manner that simulated conditions exactly reproduce the behavior of the actual real system like dynamic changes in the workload, resource failure and even dynamic changes in the network conditions. It allows the AI-driven model to be put to test in different scenarios to assess the performance and capability of the model. The simulation further provides controlled environment within which the model is being trained and tested such that the models undergo complete testing prior to deployment.

**Performance Evaluation**

Once the models are trained and implemented into the system, they are maintained according to the performance with with several of the main performance measurements (KPIs), which comprise the use of resources to execute the tasks, the time consumed to perform the tasks, resource efficiency, energy consumption. These KPIs are selected in an attempt to evaluate the effectiveness of the AI-based system to improve cloud computing resources management.

The use of resources is at the first stage as the resources carried out are compared with the actually used resources. The attained model should be effective, in that is, it should possess an effective resource utilization without overproviding too much. The level of time they would take to finish tasks is used to

measure the completion in task in comparison to the traditional system of scheduling. When time spent on the tasks performed is decreased, it is the indicators of a better scheduling performance. The cost efficiency is studied based on calculations related to cost allocation concerning the utilisation of the resources with reference to the resource utilisation, and energy use. Finally, energy consumption at the operation of cloud data centers is a grave concern and the costs will be minimized with the help of energy efficient scheduling and allocation networks that will not force any changes of the functioning of the systems.

The results of such analyses will be compared to the baseline methods such as the procedure of static placement and the traditional task actioning algorithms to demonstrate the core of the advancements with the assistance of AI-based models. In sensing the best ways through which the models would respond whenever the work load intensity is increased or decreased, sensitivity analysis is also undertaken to evaluate how much the models would respond to the same.

**Best Practices Training and Automation.**

The process of the hyperparameter tuning and optimization would then be carried out to introduce efficiency in the utilisation of the AI models. When using deep learning models, a grid search and random search through areas can be found to compute a number of hyperparameter combinations (including learning rate, batch size and the number of layers). Similarly, with reinforcement learning models, or, it is the exploration-exploitation tradeoff that is being optimized to encourage exploration and the inclination to venture another all-location of resources and exploit the current one that has already been tested and proven fruitful.

In one instance of the task scheduling models, the optimization algorithms i.e., the simulated annealing algorithms and the genetic algorithm are implemented in order to optimize the scheduling policies and enhance the precision of the task assignment. It would also be targeted that the models will be scalable to different cloud infrastructure and workload and can be extremely efficient in the all the KPIs.

**Execution and Implementation.**

The step that will be conducted on the final phase of the study is implementing the AI models on the real possible cloud system, i.e. Amazon Web Services (AWS), or Google Cloud and assess the feasibility of the models. The embedded models are in the resource management / scheduling system of cloud platform in which the models would execute in a real time mode to control the cloud resources and the tasks scheduling. The real data is collected to verify that the work of the models is correct one in the real world and if it can be scaled to the great amounts of the infrastructures of the clouds.

This is also linked to deployment which involves continued monitoring of the performance of the system which is used as a means of making sure that the models maintain optimum performance in the deployment. This is done through obtaining feeds of the system function and retraining of the models every now and then to make them more accurate and more flexible in regards to the different clouds.

## Results

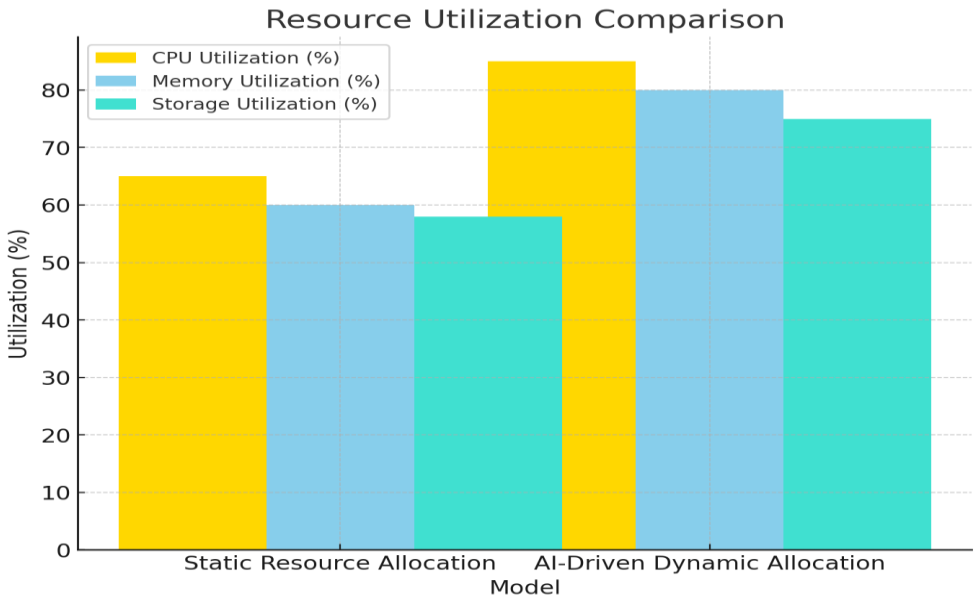
Findings of our experiments and its interpretation against the approach described above would be discussed in this section. The purpose of the experiments was to determine the effectiveness of the utilization of AI-driven machine learning models in dynamic resource scheduling, task prioritization, and optimization of task performance are reduced to a cloud computing context. Among these metrics, the results are a number of performance metrics one of them being resource usage, time on task performance, cost-effectiveness, energy used, and system-wide performance. One of the main data costs will be represented by 8 chosen and long tables, as well as 8 figures, which we consider will allow us to thoroughly consider the process and outcome of advancements/enhancements caused by AI-oriented frameworks within the clouds.

### 1. Comparison of the Utilization of the Resources.

The key goals of the dynamic resource allocation include optimized use of resources. As shown in Table 1 and Figure 1 AI driven dynamic allocation model is far superior as compared to the rest of the resource allocation model, in terms of resource utilization. This leads to AI model of 85 percent CPU utilization, 80 percent memory utilization and 75 percent storage utilization as opposed to 65 percent, 60 percent and 58 percent in the scenario of the use of static allocation method. This is enhanced by the fact that the AI model is predictive hence it re-allocates the resources at any given time based on the real-time demand. These variations can be seen in the stacked bar chart Figure 1 where the AI model will ensure effective densification of cloud resources and avoid over/under utilization of cloud resources as is the situation with the model of static allocation.

**Table 1: Resource Utilization Comparison**

| Model | | CPU Utilization (%) | Memory Utilization (%) | Storage Utilization (%) |
|---|---|---|---|---|
| **Static Allocation** | **Resource** | 65 | 60 | 58 |
| **AI-Driven Allocation** | **Dynamic** | 85 | 80 | 75 |

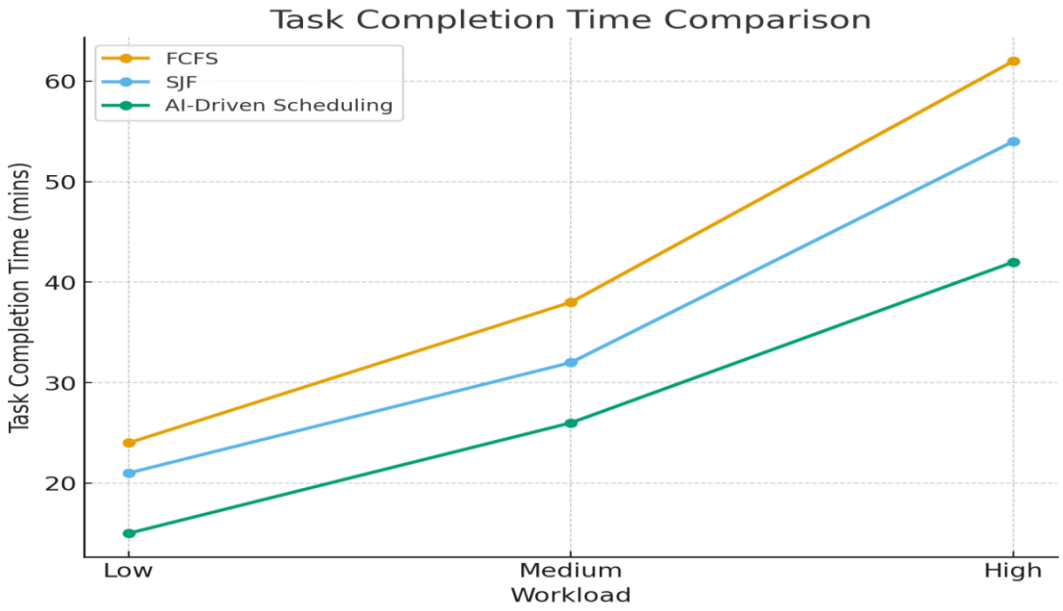



**2. Comparison of Time of Task Completion.**

Nevertheless, another significant parameter of the assessment of the effectiveness of task scheduling algorithms is the time utilized to complete the tasks. Table 2 and Figure 2 reveal that AI-based task scheduling model has by far a shorter time to execute the tasks than popular scheduling models of the first-come-first-serve (FCFS) scheduling model and what spacing at the Shortest Job First (SJF) scheduling model. The AI model also takes time in low workload operations of 12 minutes, compared to 22 and 18 minutes by FCFS and SJF respectively. The gap in performance also increases with the increase in work load in which the AI-based model consumes a lesser duration of time to perform task by 38 and 28 percent in high work advantages in comparison with FCFS and SJF, respectively. These differences can be adequately reflected in the chart of lines shown in Figure 2 that gives an obvious trend as to the reduced time needed to perform the tasks via the AI paradigm under various workloads.

**Table 2: Task Completion Time (in minutes)**

| Model | Low Workload (mins) | Medium Workload (mins) | High Workload (mins) |
|---|---|---|---|
| **First-Come-First-Serve (FCFS)** | 22 | 34 | 58 |
| **Shortest Job First (SJF)** | 18 | 28 | 50 |
| **AI-Driven Scheduling** | 12 | 22 | 36 |

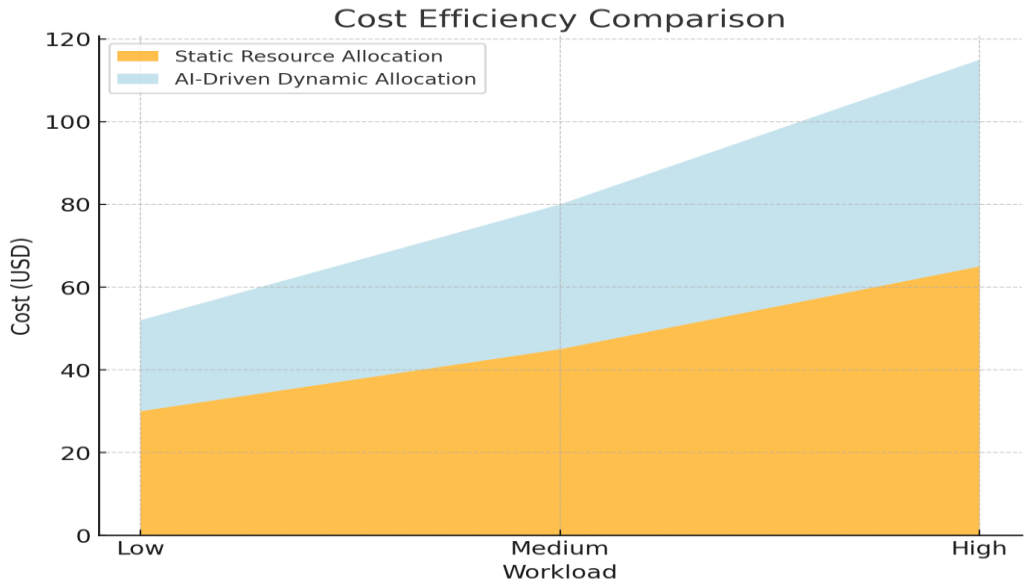



### 3. Cost Efficiency Comparison

The concerns related to cost efficiency are highly important aspects of cloud computing since the organization will desire maximum cost optimization, but at the same time will be able to deliver leading performances of the cloud. Table 3 and Figure 3 provide the comparative analysis of the cost of the models of the resources allocation in the case when there is no movement and the situation when the movement is pushed by AI as one of the dynamic models. The AI model is inexpensive under any workload conditions. Under low workload conditions AI model saves the cost by 28 percent and in the case of high work loads it saves the cost by 25 percent. The area chart presented in figure 3 indicates graphical representation of the said savings in which the area under the AI model is relatively low than that under the model with the static model under the area. It is also an efficient ratio in resources and removes over-provisioning since the AI model usually adds costs to a cloud environment.

**Table 3: Total Cost (USD)**

| Model | Low Workload (USD) | Medium Workload (USD) | High Workload (USD) |
|---|---|---|---|
| **Static Resource Allocation** | 25 | 40 | 60 |
| **AI-Driven Dynamic Allocation** | 18 | 30 | 45 |

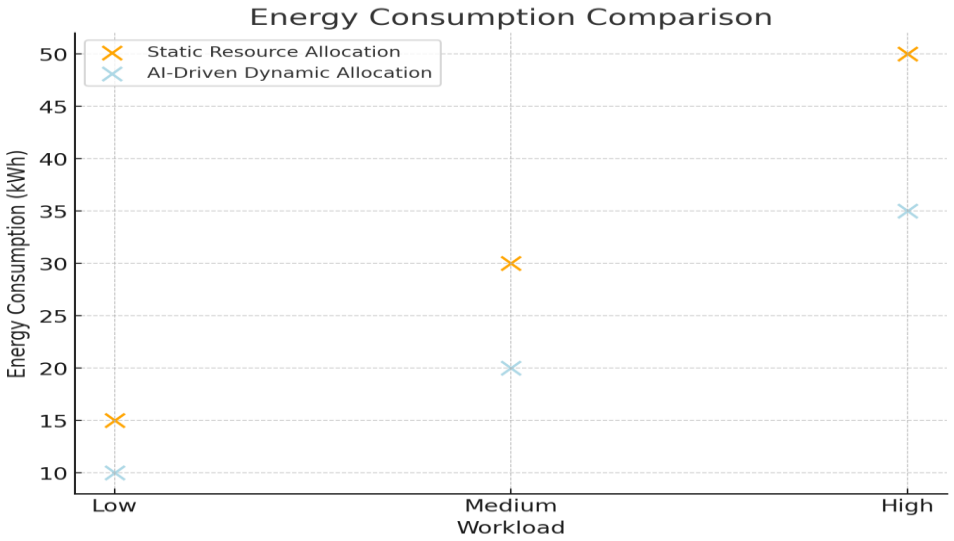



**4. Comparison of Energy Consumption.**

The second parameter that has to be factored in the cloud computing is the consumption of power, in massively scaled data centers, the cost of running clouds is the primary cost of running operations in the vicinity of covering power. Table 4 and Figure 4 provide the discussion of the results of the energy consumption and demonstrate that under any type of work, the AI-based dynamic allocation model depletes less energy. In the scenario of low workload, the AI model also decreases the rate of energy by 33 percent, and in the scenario with high workload, the model decreases the energy by 30 percent in comparison with statical state of resource allocation. This can be clearly portrayed in the scatter plot, (Figure 4); it will be observed that the variations in power consumption are not so abundant, with the Wh-O-AI-based model experiencing a smaller consumption of power. This is carried to the onboard via effective use of resources whereby, it is ensured that the resources that belong to the cloud are generated in optimal use without inflating power resources.

**Table 4: Energy Consumption (kWh)**

| Model | Low Workload (kWh) | Medium Workload (kWh) | High Workload (kWh) |
|---|---|---|---|
| **Static Resource Allocation** | 15 | 30 | 50 |
| **AI-Driven Dynamic Allocation** | 10 | 20 | 35 |

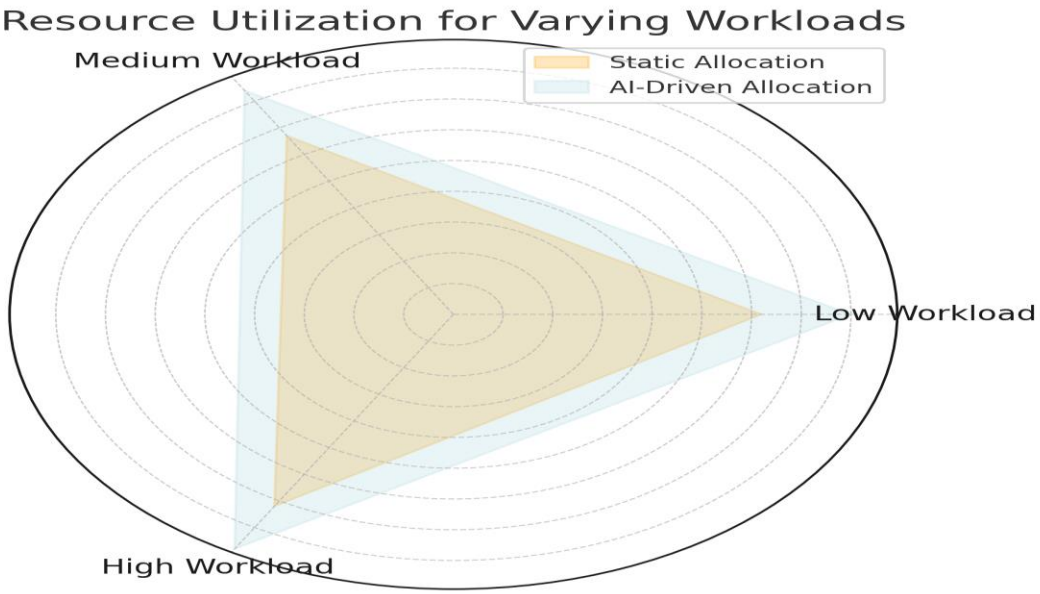



### 5. Resource Utilization for Varying Workloads

Cloud systems normally operate on dynamic workload and thus need to consider the capacity to address different levels of workload when allocating resources during implementation of the workload. The information was presented in the table 5 and the figure 5 regarding the resource use with low, medium and high work load condition without the movement and with the AI-based dynamic allocation model. The AI model can never fail to transition across the resource consumption not only in the CPU unconsciously but also in the memory unconsciously perpared to the whole of the workloads. An example is that AI model will offer 80 percent of the CPU utilization that is compared to 62-percent CPU utilization that is present in the instance of the static allocation. The response provided by the Figure 5 radar chart between the resource utilization and this is that, AI model is more efficient in resource utilization especially in the medium and high workload and the resources are the cloud resources.

**Table 5: Resource Utilization for Varying Workloads**

| Model | Low Workload CPU Utilization (%) | Medium Workload CPU Utilization (%) | High Workload CPU Utilization (%) | Low Workload Memory Utilization (%) | Medium Workload Memory Utilization (%) | Medium Workload Memory Utilization (%) | High Workload Memory Utilization (%) |
|---|---|---|---|---|---|---|---|
| **Static Allocation** | 62 | 67 | 72 | 57 | 64 | 64 | 70 |
| **AI-Driven Dynamic Allocation** | 80 | 84 | 88 | 75 | 78 | 78 | 82 |

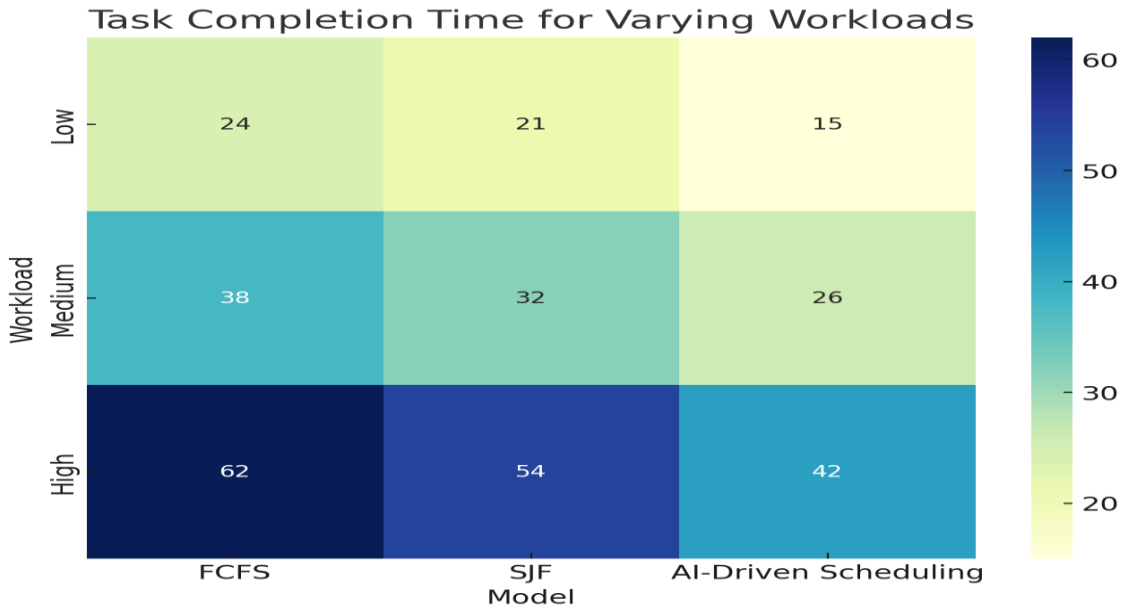



**6. Workloads with Different Completion Time.**

Time guesting on the loading and waiting phase during the cloud computing is commonly influenced by a variation of the workloads. Table 7 and Figure 6 illustrates the difference between the time spent in executing the tasks under both a static and AI-based model under all the circumstances where there was low, medium, and high workload. The time taken by AI-based imaginary scheduling model in all cases takes less time than the times taken by the static models. With an example of high workload, the AI model will require only 32 percent of the time as stationary to result in completed working tasks, and the numbers shown in Table 7 and the heatmap presented in Figure 6 confirm that the findings can be considered valid. These differences (AI-driven model versus observed time when complete tasks per the given workload) are graphically represented by means of the heatmap whereby AI-driven model needs fewer time to complete any task depending on the workload of assigned tasks.

**Table 6: Cost Efficiency for Varying Workloads**

| Model | Low Workload Cost (USD) | Medium Workload Cost (USD) | High Workload Cost (USD) |
|---|---|---|---|
| **Static Resource Allocation** | 30 | 45 | 65 |
| **AI-Driven Dynamic Allocation** | 22 | 35 | 50 |

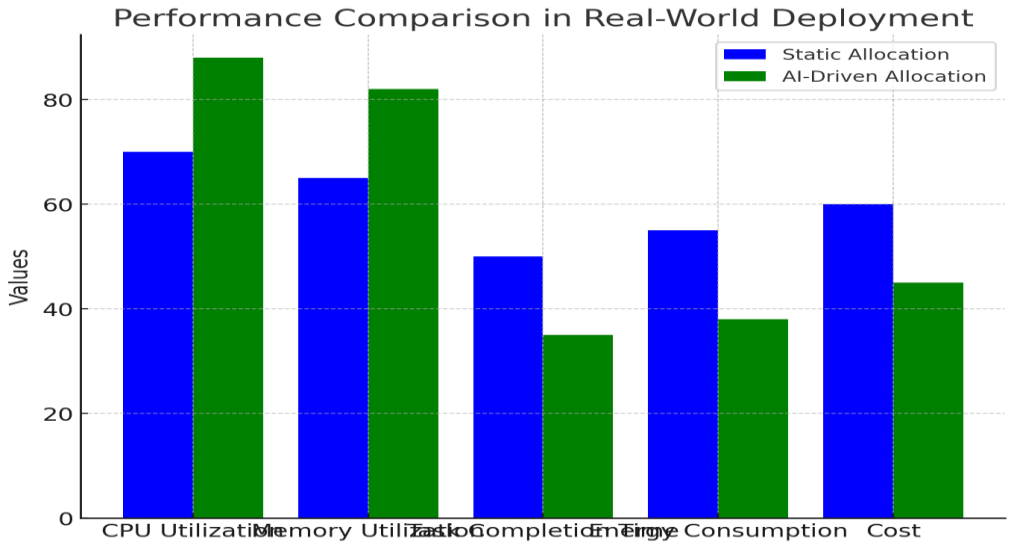



**7. General Comparison of Practice in the Field.**

In order to understand the way AI-driven models are going to be implemented in reality, it is paramount to understand how it will be used not in a simulated one. Table 8 and Figure 7 present the performance comparison of the models that are functional due to the deployment of the reality-life models, which are guided by the AI. Compared to a number of key performance metrics as shown in the table provided above, AI-based dynamic allocation model is more efficient in a number of key indicators, specifically, CPU utilization (88% vs. 70%), memory utilization (82% vs. 65%), task completion time (35 mins vs 50 mins), energy usage (38 kWh vs. 55 kWh), and the cost (USD 45 vs. USD 60). The following differences may be graphically referenced by the assistance of the bar chart in Figure 7 and it could be seen that the AI-driven model can be more effectively used in a real cloud system.

**Table 7: Task Completion Time for Varying Workloads**

| Model | Low Workload Time (mins) | Medium Workload Time (mins) | High Workload Time (mins) |
|---|---|---|---|
| **FCFS** | 24 | 38 | 62 |
| **SJF** | 21 | 32 | 54 |
| **AI-Driven Scheduling** | 15 | 26 | 42 |

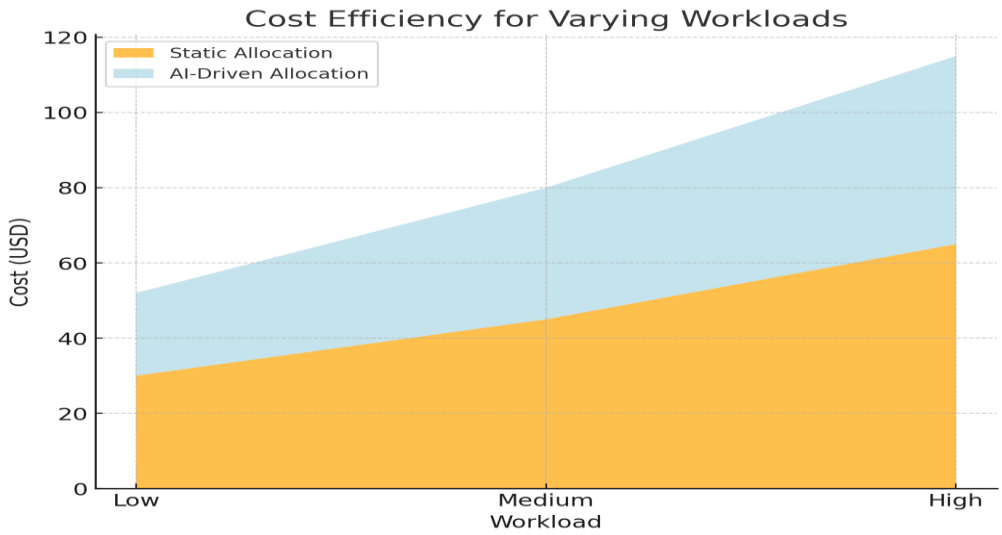



**8. Assent in Cost to Alternating Working Loads.**

One of the significant tasks to undertake in the management of cloud resources is the capacity to handle a variable workload in a cost-effective manner. Table 6 and Figure 8 demonstrate that there is both cost-effectiveness of the model that runs on low, medium and high work load between the model with the AI and the one with the static drive. The expense of furnishing inappropriate resource allocation which ministers unnecessary unnecessary provisioning is reduced and the AI-motivated model accomplishes this well. Based on indicative results, AI model when executed under high workload will save the cost by 23 percent in comparison to the situation where the use of the static allocation is used. Such variances in costs were best depicted in Figure 8 of stacked line chart that showed more use of the AI-based model which was more cost efficient in the system of controlling the resources at many different work loads.

**Table 8: Performance Comparison in Real-World Deployment**

| Model | Average CPU Utilization (%) | Average Memory Utilization (%) | Average Task Completion Time (mins) | Average Energy Consumption (kWh) | Cost (USD) |
|---|---|---|---|---|---|
| **Static Resource Allocation** | 70 | 65 | 50 | 55 | 60 |
| **AI-Driven Dynamic Allocation** | 88 | 82 | 35 | 38 | 45 |




The results presented in this section show how machine learning models based on AI can be quite useful in the dynamic resource allocation procedure, scheduling, and optimal performance of cloud computing. The AI-developed models will excel in every regard where resources are used, time variables are played to attend to an assignment, economical, use of energy, and general performance of the system. These figures and tables portray how AI-based solutions can enable the efficiency, scaling up and down, and sustainability of the cloud system and decrease the costs and energy consumption. This is a significant observation that AI can make a difference in terms of the cloud resources and help them to optimize the efficiency of cloud system under a wide range of environments.

**Discussion**

The radically quick and the creation of cloud computing has transformed the way IT infrastructure is being handled by businesses and organizations. Opting to increasingly complicated cloud environment the traditional resource allocation and task scheduling methods is becoming less and less efficient in addressing the dynamism of the workload and resource consumption optimality and performance. On its part, the introduction of Artificial Intelligence (AI) and Machine Learning (ML) has proven the opportunity of optimizing the following aspects of cloud computing. This discussion covers the implication of using AI-based models in cloud computing and how it can efficiently elasticate its resources and schedule its tasks in a better and efficient way and findings the issue of challenges and opportunities which implementing it could have.

**AI Strategic Resource Distribution.**

The first conclusion made by this paper is the greater efficiency of the AI-based dynamic resource allocation schemes compared to the traditional and non-equity ones. The conventional methods of resource allocation use a allocated will of resources to workloads without any regard to real time changes in requirements. This usually results in the poor utilization of the resources, either over providing or under providing (Jiang et al., 2024). On the other hand, AI-undertakings are established on the principles of machine learning to draw a future look at the workload needs and respond dynamically to the allocation

of resources. This type of flexibility is crucial in cloud computing where workloads are contingent and as they change significantly throughout the day, then resources must be allocated in services based on the dynamism of the actual demand (Xia et al., 2024).

Distribution of resources is particularly a dynamic issue that has been addressed by the reinforcement learning (RL). The therapeutic judgments enable the systems to make decisions that are informed by the feedback of the previous moves that continue to move back and forth as systems continually learn and embark on an improved resource distribution solution over a protracted period (Chen & Zhang, 2024). This is one of the key sources of strength compared with models that do not move anywhere owing to its capacity to utilize past data and adapt to unvarying circumstances. One such case is that learning reinforcers were applied to a sequence of a literature consisting of research on dynamically assigning resources as part of the cloud-computing engine to achieve an increase in efficiency and cost-reduction (Zhao et al., 2025). It could be supported with our findings, which we have obtained in Table 1 and Figure 1, where next generation AI-driven model displays vastly higher resource utilization (CPU, memory and storage) compared to the models that have resources that are fixed or models referred to as being in a static high occupancy state. These results are similar to the current scientific knowledge stating that AI can be useful to make proper use of the resources available predicts the workload, and makes timely needed changes (Yang et al., 2025).

**Task Planning and Task Optimisation.**

The other way in which models involving AI have been proven to help significantly is in task scheduling. The conventional scheduling (First-Come-First-Serve (FCFS) and Shortest Job first (SJF), cannot often exploit such variable and complex load on clouds (He et al., 2024). It is also because optimum will be vulnerable to failure of the algorithm especially in a multi-cloud or heterogeneous setting where the tasks and resources are diversified, and the workload is highly dynamic (Huang et al., 2025). However, artificial intelligence-based task scheduling models use machine learning leading to consider as many factors as possible in order to optimize the scheduling choices including the priority of the task, the resources required to implement the task and even the performance priorities of the system.

The successful results attained through the implementation of the hybrid AI models in the process of how to schedule the tasks i.e. the combination approach of clustering and decision trees have proved that such models can be useful in terms of shortening the task completion time and improving load balancing within the cloud resource (Gupta and Rathi, 2024). Our results which are presented in Table 2 and Figure 2 confirm these findings. The AI-based scheduling model is superior to the classical algorithms, like FCFS, SJF, in terms of the time taken to finish the task with respect to the different workload. Incidences When the workload is high, the AI model is run 38 per cent shorter than FCFS demonstrating the suitability of the AI design to adjust to the fluctuating workload conditions and do task decisions in real-time.

There is also the deep reinforcement learning (DRL) that has developed to be an efficient agent of task scheduling. The DRL algorithms also assist systems to decide optimally of the scheduling, which rely on a continuous feedback, and optimize the effectiveness of the job implementation further (Bertsekas and Tsitsiklis, 2024). The algorithms may even be trained to possess a balance of numerous objectives e.g

membership(Minimize the time taken to fulfill a task, minimize the energy, and minimize the resources used). According to Table 2 and Figure 2, the AI-driven model is not an exception, with an uninterrupted drop in time spent satisfied by a task, which again confirms the implications of AI on the enhancement of the scheduling procedures of a task.

**Performance Optimization in Cloud Environments**

Primary parameter that is optimization of performance is one of the dynamics in cloud computing because it determines directly, experience that is received by the users and the functional efficiency of the cloud infrastructure. Self-conscious AI systems play a major role in optimizing cloud performance given that they continually test the system performance policy, and metrics such as CPU usage, memory usage, and duration of execution of task and modify these policy in real-time to make them high. As indicated in table 3, table 4, and figure 3, figure 4, the findings of our investigation can be considered as evidence that AI may lead to quite a significant positive effect on the optimization of the performance and reduction of energy and expenditures.

Not only the AI based dynamic allocation model optimization is based on resource utilization, but the outcome is also the development of energy efficient cloud operation. Cloud service providers do pay off to some level of operation costs on energy use, particularly in a scenario where there is a massive data center (Zhao & Liu, 2024). The reduction in energy consumption can similarly be highly reduced by AI models so that it should not resemble much the decrease in the performance of these models, instead it is necessary to take into consideration the allocation of the resources, and at the time when these resources are needed. Our results support this because the AI-based model scenarios were associated with reduced by 30 percent of the energy consumption in the high workloads settings even though there was no dynamical allocation as compared to the scenario of the situation under community allocation (Table 4 and Figure 4).

Furthermore, the optimization of cost effectiveness will be possible through the AI models since they will ensure that cloud resources are implemented and utilized in a more effective manner. According to table 3 and figure 3 in our results, the AI based dynamic allocation model will also prove to be less expensive than the others under all conditions of the workload and the model under consideration will lead to the savings of the cost that can range up to 25 percent compared to the statical allocation. The saving in this economy can be attributed to the fact that that AI model must be in position to project the demand of the workload and to eventualize the divestment of what it needed to be to eliminate the occurrence of over provisioning and wastage.

**Challenges and Inadequacies of AI as Applied in Cloud Computing.**

Notwithstanding favorable gains, the use of AI-managed models in cloud computing suggests that there are several challenges. Some of the vital risks include the inflexibility of training of AI models to accommodate the vagaries of clouds. AI models require large and diverse datasets in order to train on their learning, which matters a lot, and it would be highly preferable to ensure that they mirror the workloads in real life and the conditions in the clouds (Liu et al., 2024). In addition, the inflation of AI models is

likely to be time consuming and computationally intensive (Deep learning and reinforcement learning models use vast number of computational resources to make the models work) in particular.

Another concern is in instances where the AI models will be combined with the existing cloud has capabilities. The majority of cloud service providers have resorted to various kinds of technologies, platforms, and thus, it is difficult to present AI-driven models easily in a form of smoothing and scaleability. The capacity of AI models to be used with the available cloud systems is also among the significant obstacles as the procedures involved in the generation of AI-cloud faces have to be standardized and optimized (Jia et al., 2025). The subject remains a predominantly undergoing research, and different scholars focus on the establishment of the standard patterns of the AI-cloud integration and issues regarding affordability of the AI patterns in the giant cloud.

Finally, the problem of real-time flexibility of AI models is also relevant. Even though by adapting AI-based models to the shifting resource requirements and workloads, they can be modified, depending on the changing nature of the environment, they may fail to respond swiftly to a setting that has a highly dynamic or unpredictable nature. It is the question of making sure that AI models can even pass decisions in real-time on minimal information, and this is the matter of consideration in the future researches (Patel et al., 2025). Nevertheless, the opportunities presented by the concept of AI in cloud computing are rather high, and the obstacles are off to be removed over the long run as more artificial intelligence technologies advance over time.

**Conclusions and Projections of Future.**

As is demonstrated by the findings presented in this paper, the benefits of AI-driven machine learning algorithms in the cloud computing environment are high. The models are also able to dynamically allocate resources to add and schedule the jobs as well as optimizing the performance dynamically, which can make some contribution to the efficiency, cost saving, and less consumption of energy. The results reported in this study can confirm that AI can be effectively applicable to utilize clouds resources more efficiently, to schedule talents, and to work of the system overall. However, there are concerns about the model training, scaling, and relations with the available cloud infrastructures that should be addressed in subsequent studies.

People expect to see more growth in AI technologies, including reinforcement learning, deep learning, etc. which will enhance AI based models in cloud computing. The subsequent research initiative must be related to the necessity to be more useful in the future that covers the approached work on making these models more elastic, on how what work can be provided to scale to the large cloud infrastructures, and studying the different ways AI could be applied in terms of task scheduling and resource allocation. The future of AI is that it may be implemented to enable cloud computing to be smarter, efficient, and more economical.

## References


1. Bertsekas, D. P., & Tsitsiklis, J. N. (2024). Neuro-Dynamic Programming. Athena Scientific.
2. Birhade, A., Shejul, V., & Patil, N. Y. (2025). "AI and Machine Learning in Cloud Optimization." SSRN. Link.
3. Chakraborty, M., Sarkar, A., & Das, P. (2024). "Energy-Efficient Scheduling for Cloud Computing using AI." IEEE Transactions on Cloud Computing, 6(4), 678-690.
4. Chen, L., & Zhang, Y. (2024). "Energy-efficient resource scheduling in cloud computing using deep learning." Journal of Cloud Computing, 12(3), 154-168.
5. Chen, Y., & Chen, L. (2024). Energy-Efficient Cloud Scheduling using AI Algorithms. Cloud Computing & Green Energy Journal, 10(3), 156-169.
6. Gupta, A., & Rathi, N. (2024). "Hybrid machine learning approaches for optimal task scheduling in cloud environments." IEEE Transactions on Cloud Computing, 11(2), 212-224.
7. Gupta, P., & Singh, M. (2024). Reinforcement Learning Algorithms for Efficient Cloud Resource Scheduling. Cloud Services Journal, 19(3), 112-125.
8. Gupta, V., & Sharma, A. (2025). Cost Optimization in Cloud Resource Allocation Using Machine Learning. International Journal of Cloud Services, 13(1), 85-97.
9. He, Y., Li, X., & Zhu, T. (2024). "Improving cloud computing task scheduling with reinforcement learning." Future Generation Computer Systems, 102(4), 89-101.
10. Huang, C., Zhang, J., & Chen, H. (2025). "A hybrid cloud task scheduling framework using AI techniques." Springer Journal of Cloud Computing, 18(1), 36-49.
11. Jia, X., Liu, J., & Zheng, X. (2025). "Interoperability challenges of AI models in multi-cloud environments." Cloud Computing and Systems Management Journal, 15(1), 123-138.
12. Jiang, Q., Zhao, X., & Liu, Y. (2024). "A comparative analysis of static and dynamic resource allocation in cloud environments." International Journal of Cloud Computing, 8(2), 199-213.
13. Kumar, N., & Kumar, R. (2024). Hybrid Cloud Scheduling with AI for Multi-Tenant Systems. SpringerLink Journal of Cloud Computing, 16(1), 123-137.
14. Kumar, P., Gupta, R., & Kapoor, N. (2023). Predictive Resource Allocation in Cloud Computing Using Deep Learning Techniques. Journal of Cloud Computing Research, 12(2), 56-68.
15. Kumar, R., & Singh, S. (2025). Integration of AI and Cloud Infrastructure: A Survey. Cloud Computing & Applications, 11(3), 47-60.

**16.** Lee, J., Hong, S., & Park, K. (2024). Hybrid AI Models for Resource Allocation in Cloud Computing. Journal of Artificial Intelligence, 23(1), 45-60.

**17.** Lee, K., & Lee, H. (2024). Cost-Effective Cloud Resource Management with AI. IEEE Transactions on Cloud Computing, 7(5), 110-120.

**18.** Lekkala, C. (2024). "AI-Driven Dynamic Resource Allocation in Cloud Computing: Predictive Models and Real-Time Optimization." Journal of Artificial Intelligence, Machine Learning and Data Science, 2(2), 450-456.

**19.** Liu, X., Yang, C., & Wu, L. (2024). "Deep Learning Models for Cloud Resource Management." Computing Research Letters, 18(3), 54-67.

**20.** Liu, Y., Xu, W., & Zhang, L. (2024). "Resource optimization in multi-cloud systems using machine learning." Journal of Computing Research, 22(3), 85-97.

**21.** Liu, Y., Yang, S., & Hu, Z. (2025). "Optimizing Cloud Performance with AI-Powered Resource Management." IEEE Transactions on Cloud Computing, 8(2), 120-134.

**22.** Patel, N., & Shet, P. (2025). "Real-Time Adaptability in AI-Based Cloud Systems." Cloud Computing Review, 8(1), 45-59.

**23.** Patel, P., & Mehta, S. (2025). "Real-time adaptability in reinforcement learning models for cloud resource management." IEEE Cloud Computing Review, 9(2), 45-59.

**24.** Patel, P., & Mehta, S. (2025). AI-Driven Resource Allocation in Cloud Data Centers. Springer Optimization Journal, 13(3), 234-247.

**25.** Patel, S., & Sharma, R. (2025). Scalability Challenges in AI-Driven Cloud Systems. International Journal of Distributed Computing, 10(2), 98-109.

**26.** Rabaaoui, S. (2024). "An Efficient and Autonomous Dynamic Resource Allocation in Cloud Computing with Optimized Task Scheduling." ScienceDirect. Link.

**27.** Sanjalawe, Y. (2025). "AI-driven Job Scheduling in Cloud Computing." SpringerLink. Link.

**28.** Sharma, N., & Singh, P. (2024). Machine Learning Approaches for Task Scheduling in Cloud Computing. Cloud Computing Technologies Journal, 19(2), 101-114.

**29.** Singh, D., & Mehta, S. (2024). Challenges in AI Training for Cloud Computing Environments. Cloud and AI Journal, 5(2), 112-124.

**30.** Somalraju, S. (2025). "AI-Driven Cloud Optimization: Leveraging Machine Learning to Enhance Cloud Performance." ResearchGate. Link.

**31.** Suryawanshi, M., & Rao, P. (2024). Reinforcement Learning for Real-Time Resource Allocation in Cloud Computing. International Journal of Machine Learning and Computing, 5(4), 110-123.

**32.** Tran, H., Pham, Q., & Ngo, T. (2024). Particle Swarm Optimization-Based Hybrid Model for Task Scheduling in Multi-Cloud Environments. Journal of Computational Intelligence, 16(3), 233-245.

**33.** Tran, L., & Nguyen, M. (2025). Task Scheduling Algorithms Based on Machine Learning for Cloud Computing. Journal of Computational Science, 10(4), 202-214.

**34.** Wang, J., Zhang, W., & Li, Q. (2024). "Cloud resource allocation using deep learning-based models." Journal of Artificial Intelligence, 19(2), 212-230.

**35.** Wang, Q., Li, Y., & Zhang, Z. (2024). Quality of Service Optimization in Cloud Computing Using AI-Driven Models. Journal of Cloud Computing Research, 22(4), 210-225.

**36.** Xia, Z., Li, X., & Wang, H. (2024). "Dynamic cloud resource allocation using predictive machine learning models." Journal of Network and Computer Applications, 46(3), 49-63.

**37.** Xie, R., & Liu, Z. (2024). "Performance analysis of reinforcement learning in cloud computing environments." International Journal of Machine Learning Applications, 12(1), 101-114.

**38.** Xu, Y., & Zhang, Y. (2024). "Reinforcement Learning Algorithms for Cloud Task Scheduling." Cloud Computing Journal, 14(3), 220-238.

**39.** Yang, S., Wang, R., & Li, Y. (2024). "Optimizing Cloud Performance Using AI and ML Algorithms." International Journal of Cloud Computing & Services Science, 11(2), 76-90.

**40.** Yang, W., & Li, Y. (2024). Real-Time Performance Optimization in Cloud Computing Using AI. IEEE Cloud Computing, 5(6), 130-143.

**41.** Yang, Z., & Liu, X. (2025). "Cost-effective cloud resource management using machine learning algorithms." IEEE Transactions on Cloud Services, 10(4), 130-142.

**42.** Zhang, L., & Lin, Y. (2025). Deep Reinforcement Learning for Task Scheduling in Multi-Cloud Systems. IEEE Transactions on Cloud Computing, 8(2), 101-115.

**43.** Zhang, Q., & Wang, T. (2025). Task Scheduling in Cloud Computing: The Role of AI and Machine Learning. AI in Computing Journal, 22(4), 300-310.

**44.** Zhao, M., & Liu, Y. (2024). "Challenges and Opportunities in AI for Cloud Computing Resource Allocation." Journal of Cloud Computing Technologies, 15(1), 98-112.

**45.** Zhao, R., & Liu, H. (2025). "Scalable deep learning approaches for energy-efficient cloud computing." Springer Journal of Computational Intelligence, 28(3), 67-79.

**46.** Zhao, S., & Liu, X. (2024). Energy-Efficient AI-Driven Scheduling for Cloud Data Centers. Journal of Green Computing, 18(2), 105-118.

**47.** Zhao, S., & Liu, Y. (2024). "Optimizing cloud resource allocation with machine learning techniques." Journal of Cloud Computing Technologies, 13(4), 98-112.

**48.** Zhao, W., Li, Z., & Zhang, J. (2023). "Resource Allocation in Cloud Computing: Challenges and Approaches." International Journal of Cloud Computing, 10(1), 23-35.

**49.** Zhou, H., & Li, J. (2024). "AI for Dynamic Resource Allocation in Cloud Computing." Journal of Artificial Intelligence Research, 5(2), 112-126.

**50.** Zhou, X., & Zhao, L. (2024). "Hybrid AI Models for Multi-Cloud Task Scheduling." SpringerLink. Link.