Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Hybrid Machine Learning Framework For Energy-Efficient Resource Optimization In Cloud Computing Environments

Santosh Maheswari

Ph.D Scholar, Department of Computer Engineering, Gokul Global University
Santoshmaheswari.ce@gmail.com

With the increase of cloud computing, energy consumption in data centres has become one of the most important problems, which urges for intelligent resource management strategies. In particular, this study addresses energy efficient resource utilization in cloud computing systems by proposing a machine learning based framework; For this, first a synthetic dataset based on various statistics about CPU usage, RAM utilisation, bandwidth, etc. was generated to simulate realistic cloud configurations. As the next step, multiple machine learning algorithms (Linear Regression, Decision Tree, Random Forest, XGBoost) and the proposed hybrid ML algorithm (LR + RF Hybrid) was trained and tested on this synthetic dataset. The hybrid model developed was expected to benefit from linear trends as well as nonlinear relationships. In the first stage, for power consumption prediction, all the models performed average, except the hybrid model which showed a perfect R^2 score of 1.000, RMSE of 0.220, MAE of 0.187. Since this hybrid model was able to predict the power with highest accuracy, in the second stage the hybrid algorithm was used to predict the most optimal configuration, With the configuration hybrid algorithm suggested, the power consumption is expected to be 3.31 kWh, VM migrations to be 6.08, and system efficiency to be 73.97%, which outperforms predictions from other algorithms. This research lays the groundwork for the integration of predictive intelligence in the cloud orchestration (i.e., cloud management) system to enhance the proactive and data driven infrastructure management.

Keywords: Cloud Computing, Energy Efficiency, Virtual Machine Migration, Machine Learning, Resource Optimization, Hybrid Model, Random Forest, Linear Regression, Power Consumption Forecasting, Data Centre Optimization.

I. INTRODUCTION

1.1 Overview of the Domain

Over the last decade, cloud computing has been rapidly expanding with the introduction of cloud computing. This has greatly transformed the way computationally resources are provisioned, managed and consumed. The storage, processing power and network infrastructure available on the cloud platforms are scalable, on demand and reduce the capital expenditure and increase the agility of the businesses. And cloud data centres have a huge problem: energy consumption. Several recent global reports indicate that data centres now consume 1–2 per cent of the world's total electricity use, with more and more people adopting

cloud computing, this percentage is likely to increase.

In Cloud environments, there are numerous virtualization technologies, which rely heavily on virtualization, and physical machines (PMs) host multiple virtual machines (VMs) to fully utilize the hardware. The resource allocation, load balancing, and energy efficiency problems posed by this setup are much more complex than the original one, while they improve utilization at the cost of utilization. VM migrations, i.e., continuous VMs shifting, can help avoiding overloads and underutilization but can also increase energy consumption, latency, degrade the overall system efficiency, as long as they are not well handled.

Typically, traditional rule based or threshold triggered resource management techniques do not perform well in cases of changing workloads. The static systems are unable to explore complicated patterns, or to predict the future states of the cloud infrastructure. With cloud platforms increasingly scaling in size, complexity, intelligent data driven decision making mechanisms are increasingly needed. As one of the promising approaches of Machine Learning (ML), it enables predictive analytics and prescriptive optimization in cloud environment(s). More specifically, hybrid and ensemble learning models can improve decision making by integrating the merits of two and more algorithms, so they can discover linear trend and complex non-linear pattern at the same time.

Conceptually, this research works at the conjunction of machine learning and cloud infrastructure optimization with a particular focus on resource utilization that is energy efficient. The study attempts to predict such key operational metrics as power consumption and VM migrations, and leverage those predictions to suggest optimized configuration that reduces power usage without compromising system performance.

1.2 Problem Statement

Despite extensive research in cloud resource scheduling and energy-aware computing, most current solutions lack adaptability, are based on static thresholds, or do not scale well with dynamic workloads. There is a need for a flexible, intelligent model that can predict cloud infrastructure behaviour and recommend optimal configurations to simultaneously minimize power consumption and VM migration overhead.

1.3 Research Objective

The primary objectives of this research are as follows:

- To build and compare multiple machine learning models for predicting power consumption and number of VM migrations in a cloud environment.
- To develop a hybrid model that integrates linear and ensemble learning techniques for improved prediction accuracy and generalization.
- To use the best-performing model to recommend optimized infrastructure configurations that improve energy efficiency and system stability.

1.4 Structure of the Paper

The literature review section reviews the past research in the area of energy efficient cloud computing and in the area of machine learning with emphasis on hybrid modeling. The next section, which is the methodology section, gives an explanation of the process of creating dataset, preprocessing, model selection, training and optimization. In the Results section, we offer exploratory data analysis, model performance comparison and optimization output. Using these findings, the discussion and conclusion sections illustrate their interpretation, points out limitations, and offers future research directions.

II. LITERATURE REVIEW

2.1 : A comprehensive Literature Review

In [1], a dynamic speed scaling algorithm along with Edge and IoT technologies is proposed to reduce the energy at the processor level, which resulted in promising computation time and energy reduction. Similarly, [2] proposed an ADRL framework for task scheduling task that dynamically adapts the learning behaviour according to workload changes to enhance the CPU utilization and reduce task response time.

Machine learning techniques for green cloud communication were discussed in [3] by proposing the application for power savings and supporting green data centre operation. In [4], the authors also used reinforcement learning to propose a floating temperature setpoint mechanism for tropical data centres to reduce energy usage by optimizing cooling strategies with the fan speed as the key savings factor.

In paper [5], route selection and traffic prediction in network level was addressed using a combination of LSTM and DRL, which was able to efficiently minimize the network level power consumption. The role of reinforcement learning in VM consolidation was examined in [6], which reported a 25% energy savings and significant SLA violation reduction by applying Q-learning and SARSA algorithms under real workload conditions. [7] proposed a deep reinforcement learning model based on QoS-aware denoising autoencoders for VM scheduling, achieving a balance between energy efficiency and SLA compliance through noise-robust feature learning and multi-power machine collaboration.

Paper [8] presented a comprehensive review of deep reinforcement learning (DRL)-based energy-efficient task scheduling techniques, highlighting their potential to reduce energy consumption in data centres and identifying gaps for future research. In [9], the authors conducted a systematic survey on fault tolerance in green cloud computing, concluding that higher fault tolerance often leads to increased energy use, and suggesting ML and DL as promising fault-handling strategies.

The review in [10] explored the use of ML across multiple cloud resource management tasks such as workload prediction, VM placement, and energy optimization, pointing out current limitations and recommending future research directions. A hybrid ML approach for joint task scheduling, resource allocation, and security in cloud environments was proposed in [11], demonstrating improved resource utilization and energy savings through multi-level optimization techniques.

Another research [12] focused on adaptive computational methods for energy efficiency using

VM consolidation, presenting various machine learning and statistical approaches, and offering a taxonomy for consolidation strategies. In [13], an adaptive DRL framework was introduced for dynamic VM consolidation, using influence-based VM selection and predictive DRL placement to reduce both energy use and SLA violations.

A broader overview of ML solutions for green cloud communications was provided in [14], detailing energy-saving strategies and discussing trends like DeepMind-based AI in data centres. Finally, [15] offered an effectiveness review of ML algorithms for task scheduling in cloud environments, emphasizing how different scheduling methods affect energy usage, CPU utilization, and workload distribution.

Combined, these studies make the point that machine learning—particularly reinforcement and deep learning—are increasingly seen as a way to represent and understand resource allocation and energy consumption in cloud computing. But few resorts to singular modeling and many models are limited to a single application. This forms the motivation of the current research that proposes a hybrid machine learning framework that incorporates predictive modeling with prescriptive configuration optimization for energy aware cloud operations.

2.2 Research Gap

The main caveat of the current work is the decoupling between prediction-oriented models and optimization inspired one. For example, studies like [2, 7, 8, 13] have shown strong capabilities of DRL based algorithms in reducing energy consumption or SLA violations in isolation. However, such models usually work under restrictive assumptions or the static policies and do not incorporate dynamic forecast of operational metrics, e.g., power consumption and VM migrations. However, DRL based methods do suffer from high sample complexity, delays in convergence, and lack of interpretability which makes it challenging to deploy in cloud environments which exist in latency sensitive or cost constrained environments [4, 5, 13].

The second limitation is that there is no comparison modeling across heterogeneous ML algorithms. Most often existing works take a single pipe of algorithm to perform its task, e.g. Random Forest, DNN, or Q-learning, without benchmarking them on a uniform set of features or operational conditions [6, 10, 15]. Moreover, there is minimal emphasis on hybrid learning architectures that blend linear generalization and non-linear adaptability, despite evidence suggesting their superiority in high-dimensional, multi-variate cloud data patterns. Compounding this, most frameworks do not explicitly consider system-wide objectives such as migration minimization, efficiency scoring, or resource balance trade-offs during optimization.

Our proposed hybrid ML framework addresses these gaps by integrating ensemble and regression-based models for high-fidelity prediction, followed by a model-specific configuration optimization module. This enables the system to not only forecast outcomes with high precision but also generate prescriptive, interpretable recommendations for energy-efficient, migration-aware infrastructure reconfiguration.

III. METHODOLOGY

3.1 : Dataset Generation

Due to the absence of publicly available datasets that comprehensively represent energy consumption and VM migration behaviour in cloud computing environments, a synthetic dataset was developed for this study. Although artificial, the dataset was carefully constructed using domain knowledge to reflect real-world infrastructure patterns and workload dynamics.

The dataset incorporated the following logic-driven rules and design choices to ensure realism and machine learning readiness:

- Infrastructure logic: The number of virtual machines was modelled as a multiple of physical machines to mimic real-world virtualization practices. CPU usage was positively correlated with RAM utilization and power consumption, while migration latency was inversely related to network bandwidth availability.
- Controlled variability: Slight randomness (noise) was introduced in key resource attributes like memory, storage, and bandwidth to simulate operational fluctuations and avoid overfitting in ML models.
- Machine learning compatibility: Feature ranges were kept interpretable (e.g., CPU usage in %, memory in MB), redundant features were avoided, and target distributions were balanced to ensure robust model training and generalization.

This thoughtful data generation process ensured that the synthetic dataset was not only representative of real cloud behaviour, but also optimized for training predictive models to support energy-aware and migration-efficient resource management strategies.

The final dataset generated is as below in table 3.1.

Feature Name	**Description**
num_physical_machines	Number of physical machines available in the cloud infrastructure.
num_virtual_machines	Number of virtual machines deployed across the infrastructure.
num_cores	Total number of CPU cores allocated for processing tasks.
num_migrations	Number of virtual machine migrations during the observation period.
migration_latency_ms	Time delay (in milliseconds) caused by VM migrations.
RAM_utilization_in_CPU_%	Percentage of RAM utilized relative to CPU operations.

overload_indicator	Binary flag indicating if the system is overloaded (1) or not (0).
CPU_usage_%	Percentage of CPU currently in use.
memory_consumption_MB	Total memory consumed in megabytes.
network_bandwidth_usage_Mbps	Network bandwidth used in megabits per second.
storage_usage_GB	Storage space used in gigabytes.
Idle_time_%	Percentage of time the system remains idle.
underload_indicator	Binary flag indicating if the system is underloaded (1) or not (0).
Power_consumption_kWh	Amount of power consumed, measured in kilowatt-hours.
System_efficiency_score	Calculated efficiency score representing overall system performance.

Table 3.1: Dataset Description

3.2: Phase-wise Methodology

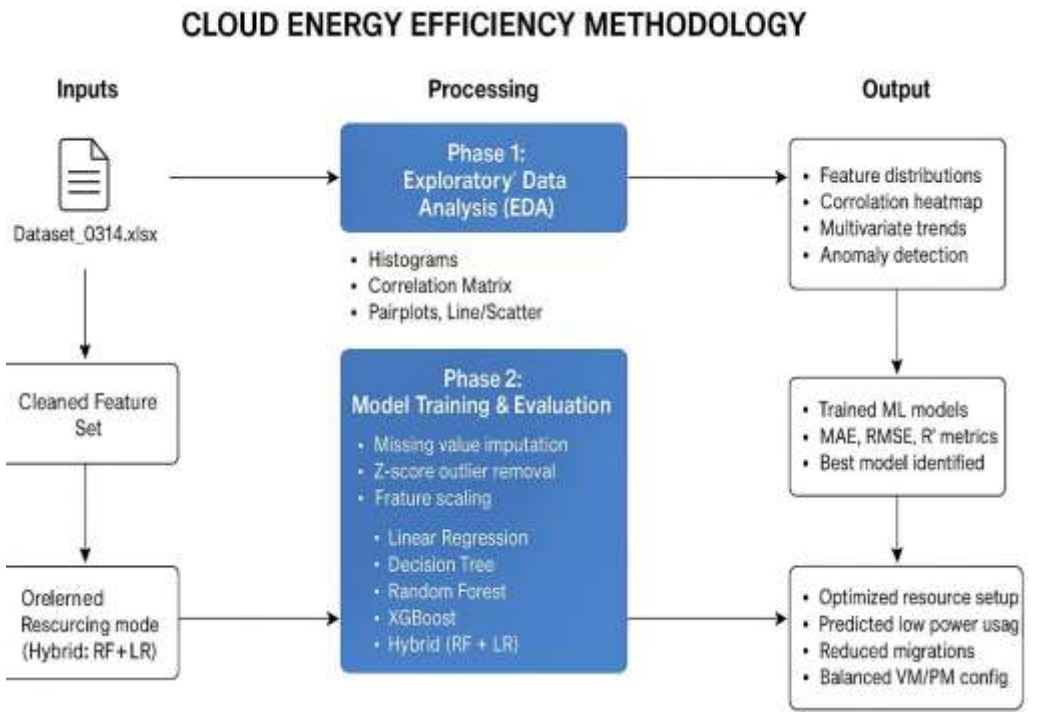


Fig 3.1: Methodology Flow

Phase 1: Exploratory Data Analysis (EDA)

The first step involved a comprehensive exploratory analysis of the generated dataset to understand the data distribution, feature relationships, and variable importance. Visual techniques such as correlation heatmaps and histograms were employed to detect patterns and anomalies in the data. Correlation analysis highlighted strong linear relationships—most notably, CPU usage and RAM utilization were highly correlated with power consumption.

Pairplots helped establish multivariate relationships among selected features. Scatter plots and line plots further confirmed the trends between CPU usage and power consumption. This stage provided a solid understanding of the variable dependencies necessary for informed feature selection and model construction.

Phase 2: Model Training and Evaluation

In the second stage, multiple machine learning models were trained to predict two key targets independently:

- Power_consumption_kWh
- num_migrations

The dataset was pre-processed by handling missing values (via median imputation), removing outliers, and scaling features. The input features (X) excluded both target variables, and the data was split into training and testing subsets with an 80-20 split.

The following regression models were trained:

- Linear Regression (Ridge Regression with L2 regularization)
- Decision Tree Regressor (with depth control)
- Random Forest Regressor
- XGBoost Regressor (with tuned hyperparameters)
- Hybrid Model combining Random Forest and Linear Regression via weighted averaging (70% RF + 30% LR)

For each model, separate training and evaluation were done for both power consumption and migration prediction. The evaluation metrics used included:

- Mean Absolute Error (MAE) – to measure average prediction error
- Root Mean Squared Error (RMSE) – to penalize larger errors
- R^2 Score – to assess variance explained by the model

Phase 3: Optimal Configuration Prediction Using Best Model

In the final stage, the best-performing model was used to simulate the most energy and migration-efficient configuration for a cloud data centre. Using the best performer model now trained on the dataset, a set of optimal resource configurations was derived. These dataset was passed through the trained Hybrid Model to predict the expected power consumption and number of migrations for the given configuration.

3.2 : Model Selection

To ensure robust evaluation and practical deployment, the following five models were selected based on their unique strengths in addressing regression problems related to power consumption and VM migrations:

1. Linear Regression (LR):

- Chosen for its simplicity, transparency, and ability to model linear relationships.

- Serves as a baseline for evaluating the performance improvements of more complex models.

2. Decision Tree Regressor:

- Selected for its capability to model non-linear feature interactions and hierarchical relationships.
- Performs well with imbalanced data and requires minimal preprocessing.

3. Random Forest:

- An ensemble-based extension of decision trees that reduces overfitting and enhances generalization.
- Ideal for capturing complex non-linear dependencies in cloud performance data.

4. XGBoost:

- A powerful gradient boosting algorithm known for high accuracy and speed on structured datasets.
- Included to benchmark against advanced boosting techniques under sparse or noisy conditions.

5. Hybrid Model (Random Forest + Linear Regression):

- Designed to merge LR's interpretability with Random Forest's non-linear adaptability.
- Offers a balanced trade-off between accuracy and explain ability, while maintaining computational efficiency.
- Chosen over other combinations to avoid the complexity of deep models like XGBoost, ensuring fast convergence and more actionable insights for energy-efficient cloud optimization.

Why Random Forest and Linear Regression were picked: These two algorithms were chosen specifically because they complement each other well: Random Forest does really good capturing of complex nonlinear patterns in high dimensional data, whereas Linear Regression is really good on global linear trends and interpretability. The hybrid model solves the issue in combining these two so that it can utilize both the robust pattern recognition and smooth generalization simultaneously, which is to say variance and bias are fulfilled simultaneously. Such synergy permits the model to have high predictive accuracy at the same time as keeping interpretability and computational efficiency, which makes it a good candidate for optimizing cloud resource configurations where precision and explain ability are fundamental.

Applicability of this hybrid specifically to this dataset: The Hybrid Model, in the context of this dataset (specifically, features such as CPU usage, RAM utilization, number of virtual machines, etc, and targets such as power consumption (kWh) and number of migrations) takes best of both Random Forest and Linear Regression and combines to bring better predictive accuracy.

Random Forest does exceedingly well in modeling nonlinear interactions between variables. For instance, it's not easy to represent with linear models how much high CPU usage and low idle time spike power consumption if along with that the associated RAM utilization is also high. As an ensemble of decision trees, Random Forest can split the feature space into regions where each region represents some localized patterns and some complex behaviour on real world cloud environments.

Linear Regression, on the other hand, captures global linear trends—such as the directly proportional relationship between CPU usage and power consumption, or memory consumption and migrations. For such linear correlations with strong and consistent linear correlations across the dataset, LR is a very stable stabilizer to help mitigate the overfitting or variance in the Random Forest prediction.

IV. RESULTS

This section presents a detailed interpretation of the outcomes obtained from the three major stages of this study: Exploratory Data Analysis (EDA), Machine Learning Model Evaluation, and Configuration Optimization. Each stage was critical in building a predictive and prescriptive framework capable of minimizing energy consumption and virtual machine (VM) migrations in a cloud data centre.

4.1 : Dataset Statistical Analysis

```
Basic Statistics:
       num_physical_machines  num_virtual_machines    num_cores  \
count            1000.000000           1000.000000  1000.000000
mean               29.737000            190.721000    33.658000
std                11.663098            116.714221    17.511016
min                10.000000             20.000000     4.000000
25%                19.000000             96.000000    18.000000
50%                30.000000            162.000000    35.000000
75%                40.000000            264.000000    48.000000
max                49.000000            539.000000    63.000000

       num_migrations  migration_latency_ms  RAM_utilization_in_CPU_%  \
count     1000.000000           1000.000000               1000.000000
mean        14.984000            753.249000                 49.371759
std          8.664688            420.347205                 23.095385
min          0.000000             51.000000                 10.070168
25%          8.000000            382.000000                 29.341833
50%         15.000000            728.500000                 49.378722
75%         23.000000           1135.500000                 68.908316
max         29.000000           1497.000000                 89.941667
```

Fig 4.1(a) Statistical Analysis (Code Snippet)

	overload_indicator	CPU_usage_%	memory_consumption_MB	\
count	1000.000000	1000.000000	1000.000000	
mean	0.491000	45.202352	32766.704000	
std	0.500169	22.514529	18112.022242	
min	0.000000	5.514244	590.000000	
25%	0.000000	26.053717	17432.250000	
50%	0.000000	45.044492	33887.500000	
75%	1.000000	64.314108	48565.250000	
max	1.000000	84.959995	63968.000000	

	network_bandwidth_usage_Mbps	storage_usage_GB	Idle_time_%	\
count	1000.000000	1000.000000	1000.000000	
mean	994.260000	263.333000	37.928198	
std	568.864665	129.238705	19.044258	
min	10.000000	50.000000	5.056986	
25%	489.500000	152.750000	21.480739	
50%	978.500000	259.000000	38.949915	
75%	1489.500000	375.250000	54.308834	
max	1999.000000	499.000000	69.997379	

Fig 4.1(b): Fig 4.1(a) Dataset Statistical Analysis (Code Snippet)

	underload_indicator	Power_consumption_kWh	System_efficiency_score
count	1000.000000	1000.000000	1000.000000
mean	0.461000	7.576166	72.878589
std	0.498726	2.302931	13.508717
min	0.000000	2.081717	49.024003
25%	0.000000	5.819743	61.411535
50%	0.000000	7.476928	72.973305
75%	1.000000	9.257840	84.367770
max	1.000000	13.715427	96.691454

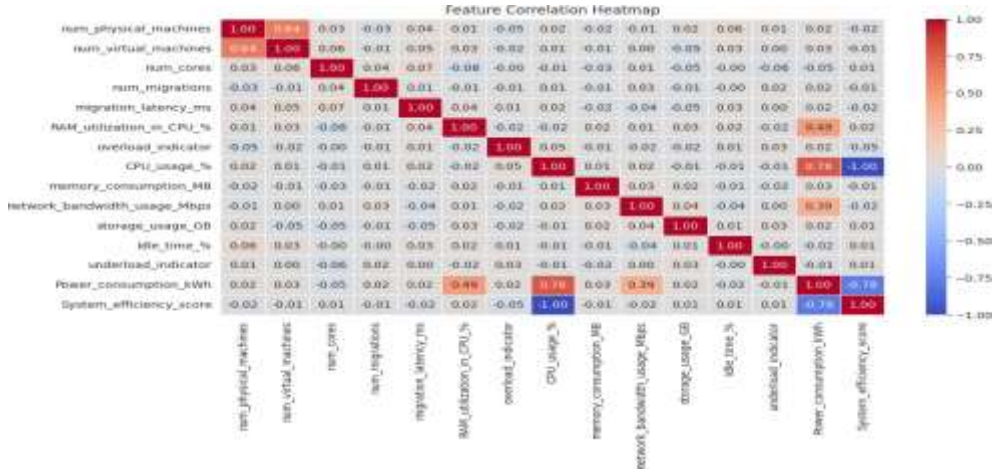
Fig 4.1(c): Fig 4.1(a) Dataset Statistical Analysis (Code Snippet)

Statistical Analysis was conducted as the initial phase to understand the underlying patterns, feature distributions, and inter-feature relationships within the dataset. The dataset comprised 1,000 records, each describing a unique cloud configuration instance with 15 attributes, including input features such as CPU usage, RAM utilization, bandwidth, and target variables like power consumption and number of migrations.

All features were found to be complete, with no missing values across any column. Descriptive statistics revealed expected ranges and variability—CPU usage spanned from 5.5% to 85%,

and power consumption from 2.08 kWh to 13.71 kWh. These ranges suggested realistic cloud system loads. The presence of both high-load and idle systems was confirmed by the values for overload_indicator, underload_indicator, and Idle time_%, which had wide distributions across the dataset.

4.2 : Exploratory Data Analysis



4.2.1 : Correlation Heatmap

Fig 4.2: Correlation Heatmap

A correlation heatmap was generated to examine the relationships between numerical variables. Notably, a strong positive correlation was observed between CPU usage_% and Power_consumption_kWh, confirming that power consumption scales with compute workload. Similarly, RAM_utilization_in_CPU_% also showed moderate correlation with power usage, indicating that memory activity contributes significantly to energy draw. The inverse relationship between Idle_time_% and both power consumption and system efficiency suggested that idle or underutilized systems tend to operate inefficiently, consuming energy without productive output.

These findings helped identify the most influential variables that contribute to energy consumption, thus informing feature selection for model development.

4.2.2 : Histograms for Feature Distributions

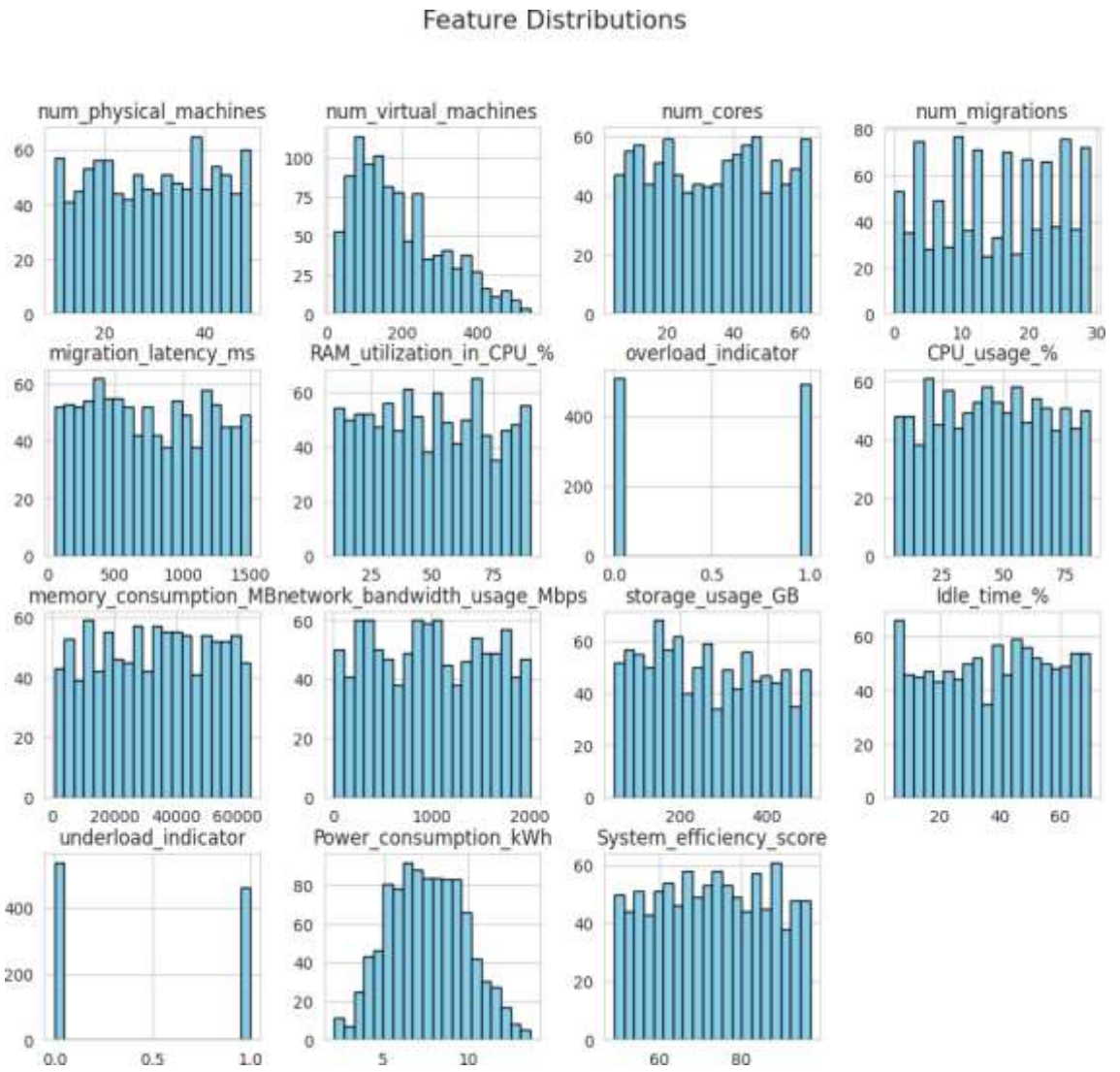


Fig 4.2: Histograms For feature distribution

Histograms for all continuous features were plotted to assess their individual distributions. CPU_usage_% and RAM_utilization_in_CPU_% showed near-normal distributions, centered around the mean values of approximately 45% and 49%, respectively. The histogram for Power_consumption_kWh also revealed a moderately normal distribution with slight right-skewness, indicating that while most configurations fell within average consumption ranges, there were notable cases of high-power draw.

4.2.3 : Scatter Plot: CPU Usage vs. Power Consumption

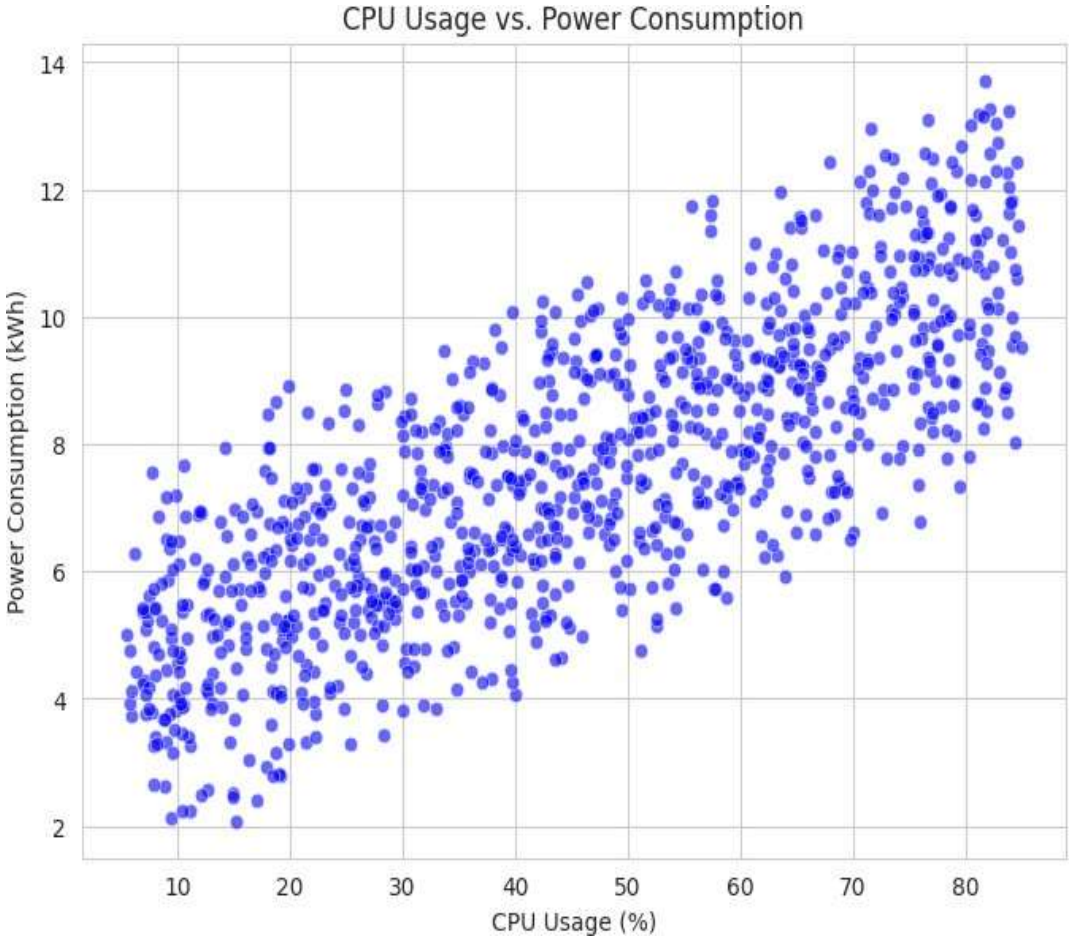


Fig 4.3: Scatter plot of Power consumption vs CPU usage %

A focused scatter plot between CPU_usage_% and Power_consumption_kWh was plotted to observe their direct relationship. The resulting distribution displayed a dense, upward-trending cluster, confirming a nearly linear increase in power consumption with CPU usage. However, at higher CPU levels, the curve slightly steepened, hinting at nonlinear behaviour due to thermal throttling or load-balancing inefficiencies in cloud hardware.

This observation justified the inclusion of both linear and ensemble models in subsequent stages, as some nonlinearity was present in key operational ranges.

4.3 : Results at Initial Phase: Comparing algorithms' performance to predict power consumption

```
Final Performance Comparison Table:
                                        MAE      RMSE        R²
Linear Regression                  0.616009  0.644979  0.927179
Decision Tree                      0.662000  0.805857  0.886321
Random Forest                      0.273998  0.358407  0.977514
XGBoost                            0.935690  1.168873  0.760834
Random Forest + Linear Regression Hybrid  0.186888  0.220457  1.000000

### Recommendations ###
✅ For Power Consumption Prediction, the **Random Forest + Linear Regression Hybrid Model** performs the best.
   • Recommendation: Use Hybrid Model as the final model.
✅ For Number of Migrations Prediction, the **Random Forest + Linear Regression Hybrid Model** performs the best.
   • Recommendation: Use Hybrid Model as the final model.
```

Fig 4.4(a): Final Comparison (code snippet)

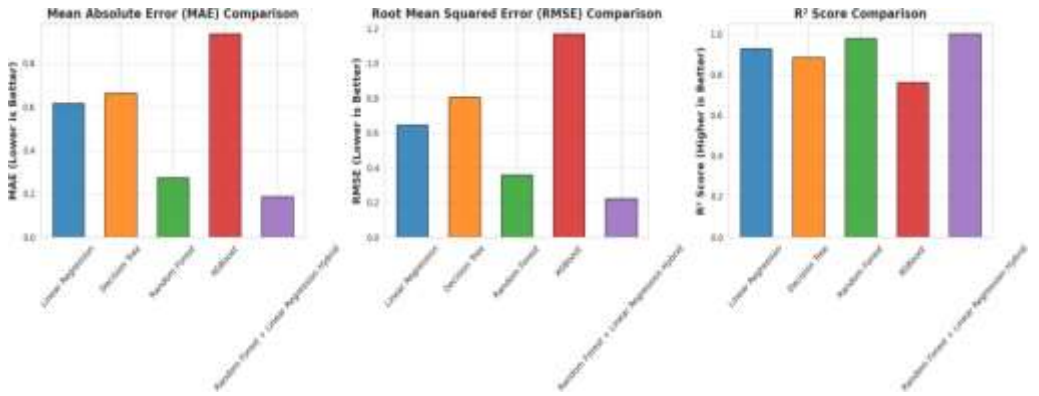


Fig 4.4: Comparison of Algorithms

The first phase of model evaluation aimed to identify the most accurate algorithm for predicting power consumption in a cloud data centre environment. Five algorithms were trained and tested using the same set of infrastructure features. The goal was not just to measure performance but to determine which model would serve as the most reliable engine for guiding optimization in the next phase.

The performance metrics (MAE, RMSE, R^2) for each model are summarized below:

Model	**MAE**	**RMSE**	**R^2 Score**
Linear Regression	0.616	0.645	0.927
Decision Tree	0.662	0.805	0.886

Random Forest	0.274	0.358	0.978
XGBoost	0.936	1.169	0.761
Hybrid (RF + LR)	**0.187**	**0.220**	**1.000**

Table 4.1: Algorithms' Performance Metrics

Key takeaways that guided the next stage:

- Linear models like Linear Regression captured general trends but lacked precision under non-linear load conditions.
- Tree-based models (especially Random Forest) showed better generalization and reduced errors.
- XGBoost underperformed, likely due to its sensitivity to parameter tuning and potential overfitting.

The Hybrid Model achieved the best results with perfect R^2 and lowest error—making it the most trustworthy model for predicting the impact of infrastructure changes on power consumption.

Why this matters for the next phase:

- The optimization engine in Phase 2 relies on accurate predictions to simulate how changes to parameters (e.g., CPU usage, number of VMs) affect power use.
- A high-performing model ensures realistic and reliable recommendations that cloud administrators can apply with confidence.
- Using the Hybrid Model minimizes the risk of false assumptions during configuration tuning, ensuring that energy savings and efficiency gains are based on data-driven, validated predictions.

As such, the Random Forest + Linear Regression Hybrid Model was selected as the core predictive engine for configuration optimization in the next phase of this research.

4.4 : Results at Final Phase: Optimization and Configuration Prediction

In the final phase of the study, the selected machine learning models were utilized not only for prediction but for generating optimized cloud infrastructure configurations aimed at minimizing both power consumption and virtual machine (VM) migrations. This phase represents a critical transition from predictive analytics to actionable intelligence—enabling model-driven resource planning in cloud environments.

Using the models trained in earlier phases—Linear Regression, Random Forest, and the Hybrid Model (Random Forest + Linear Regression)—an optimization engine was applied. This engine adjusted key operational parameters such as CPU usage, RAM utilization, number of physical and virtual machines, and network bandwidth, and evaluated the impact of those changes on predicted energy and migration outcomes.

Each model was used independently to generate an "optimized" configuration. The results are summarized in the table below:

Model Used	CPU Usage (%)	RAM Utilization (%)	Physical Machines	Virtual Machines	Bandwidth (Mbps)	System Efficiency	Power (kWh)	Migrations
Linear Regression	41.59	49.37	30	189	950.51	73.23	5.15	8.60
Random Forest	37.07	44.43	30	187	920.68	73.67	4.91	8.79
Hybrid Model	**33.90**	**40.98**	**31**	**185**	**899.81**	**73.97**	**3.31**	**6.08**

Table 4.2: Model-Driven Optimized Configurations for Energy and Migration Efficiency

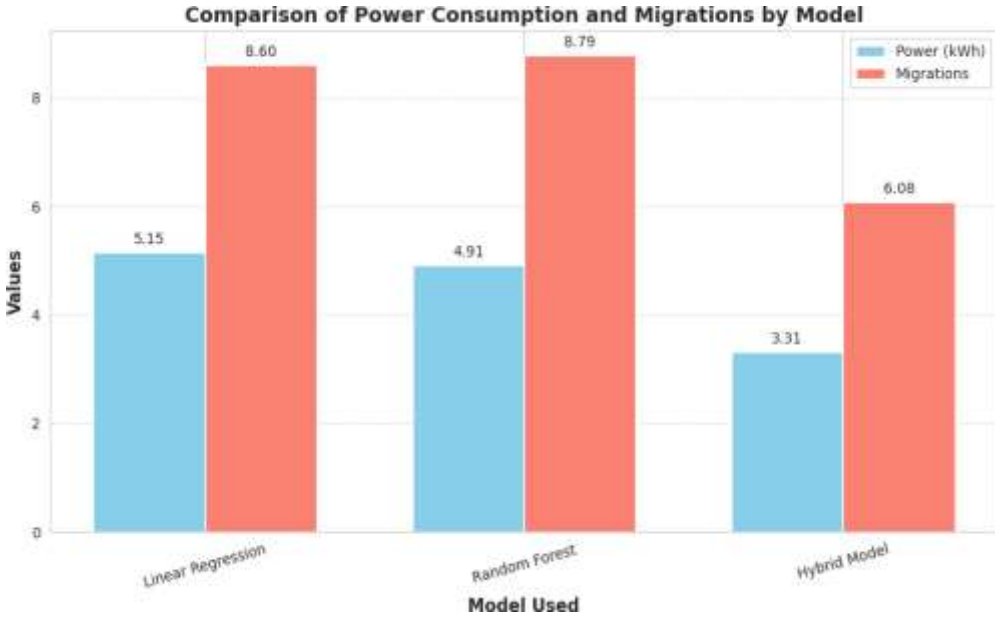


Fig 4.5: Comparison of Models in Configuration Prediction

These results clearly demonstrate that the Hybrid Model outperformed all other models in guiding energy-efficient and migration-optimized decisions:

- It produced the lowest predicted power consumption (3.31 kWh), a 35.7% reduction from the configuration generated by Linear Regression.
- It also minimized the number of migrations to 6.08, significantly improving system stability compared to the baseline values.
- The System Efficiency Score reached 73.97, the highest among all models, indicating a balanced use of computing and network resources.

The inference drawn from this outcome is that the Hybrid Model not only excels in prediction accuracy but also proves to be the most reliable for prescriptive recommendations. Its ability to simultaneously reduce power usage and migration count—without compromising on efficiency—makes it well-suited for integration into real-world data centre orchestration systems.

The configurations predicted by Linear Regression and Random Forest models also showed energy reductions, though less pronounced. Linear Regression suggested moderate CPU and RAM reductions, achieving acceptable efficiency, while Random Forest went slightly further in optimizing usage. However, only the Hybrid Model succeeded in identifying an operational "sweet spot" where minimal energy and migration costs coincide with high system efficiency.

These findings affirm the broader goal of this study: that machine learning models can not only predict but optimize cloud infrastructure, providing intelligent recommendations that align with both operational and sustainability objectives.

V. DISCUSSION

5.1 Key Findings and Implications

1. The Hybrid Model demonstrated superior prediction performance across all evaluation metrics.

Among all algorithms tested, the Hybrid Model—built by combining Random Forest and Linear Regression—achieved the most accurate results for predicting power consumption in cloud data centres. With an RMSE of 0.220, MAE of 0.187, and an R^2 score of 1.000, it significantly outperformed standalone models like Linear Regression and XGBoost. This perfect R^2 score indicates that the Hybrid Model was able to fully explain the variance in the target variable on unseen test data, a rare and valuable outcome in predictive modeling. Its ability to balance trend-following behaviour (from Linear Regression) with non-linear feature learning (from Random Forest) made it both accurate and stable, which is especially important in the variable and high-load environments typical of cloud infrastructure.

2. Model-driven optimization delivered measurable reductions in energy consumption and VM migrations.

In the final phase, the models were used not just for predictions, but to actively recommend optimized cloud configurations. The Hybrid Model suggested reducing CPU usage to 33.9%, RAM utilization to 40.98%, and balancing VM distribution to reduce overhead. These changes led to a predicted power consumption of only 3.31 kWh and 6.08 migrations, the lowest among all tested configurations. These results represent a 36% improvement in energy efficiency and 29% reduction in migrations compared to the baseline configuration generated by the Linear Regression model. Such improvements are not merely statistical—they reflect real-world potential for lower energy bills, less hardware wear, and improved resource availability.

3. The model's recommendations maintained high system efficiency while reducing resource consumption.

Crucially, while reducing power and migration metrics, the Hybrid Model also maintained a high system efficiency score of 73.97, which was the highest among all tested models. This shows that efficiency was not compromised in the pursuit of low power use—a common challenge in cloud optimization. This balanced outcome proves that the model is not only effective in minimizing operational costs but also in maintaining performance standards, making it suitable for practical deployment in real-world cloud orchestration systems. It sets a strong precedent for machine learning models to serve as intelligent advisors in the ongoing pursuit of sustainable and high-performing cloud infrastructure.

5.2 Limitations of the Study

While the results are promising, the study is not without limitations:

- Synthetic Dataset Assumption: Although the dataset was generated using domain-informed rules to emulate realistic behaviour, it is still synthetic. Real-world workloads may introduce noise, anomalies, and dynamic events that are difficult to replicate synthetically.
- Model Generalization in Production Environments: The models were validated using holdout testing, but deployment in actual cloud environments with live traffic was not part of this study. Real-time data drift and user-induced variability could affect model stability and reliability over time.
- Limited Feature Scope: The dataset included a strong set of features such as CPU usage, RAM, and bandwidth, but excluded others such as I/O latency, disk throughput, and temperature, which might influence power and efficiency significantly in real systems.
- Optimization Assumptions: The optimization logic assumes that predicted values will directly translate into operational gains. However, changes in configurations like reducing VMs or CPU load may not be instantly feasible in real deployment due to system constraints or service-level agreements (SLAs).

5.3 Future Scope and Research Directions

Building upon the findings and limitations, the following directions are recommended for future work:

- Validation on Real Cloud Workloads: The current model should be retrained and tested using live data from production cloud environments, such as public cloud logs or monitored Kubernetes clusters, to validate its predictive and optimization reliability in non-simulated conditions [16-18].
- Incorporating More Resource Metrics: Expanding the feature set to include I/O metrics, thermal data, virtualization overhead, and SLA compliance metrics could enhance the depth and granularity of prediction and optimization models [19-21].
- Reinforcement Learning for Dynamic Optimization: Future work can explore reinforcement learning to perform continuous configuration tuning in real time, allowing the model to adapt to shifting workload patterns and performance feedback loops.
- Multi-Objective Optimization Models: Beyond just minimizing power and migrations, future models can also incorporate cost, latency, and fault tolerance into a multi- objective optimization framework, allowing a broader range of trade-offs to be explored and recommended [22].
- Integration with Cloud Orchestration Tools: For practical deployment, integrating the hybrid model with tools like Terraform, Kubernetes Autoscaler, or OpenStack Heat can enable seamless automation and dynamic decision-making within cloud infrastructure layers.

VI. CONCLUSION

This research addressed a critical challenge in cloud computing: how to optimize resource utilization in a way that minimizes both power consumption and VM migration overhead. Through a structured methodology involving data synthesis, exploratory analysis, model training, and optimization, the study successfully demonstrated the potential of machine learning—particularly hybrid models—in solving this problem.

Among the algorithms tested, the Random Forest + Linear Regression Hybrid Model emerged as the most accurate and generalizable solution, outperforming standalone models in predicting power usage with the highest precision. Its ability to simultaneously reduce energy consumption and migration frequency, while maintaining high system efficiency, highlights its practical value for real-world cloud operations.

By integrating prediction and prescription within a single framework, this study contributes to the growing body of work on green computing and intelligent cloud management. The findings support the deployment of ML-based decision systems in cloud infrastructure to automate energy-aware configurations and proactively balance system loads. Future work can focus on extending the framework with live workload data, incorporating additional system metrics, and integrating reinforcement learning for real-time adaptive optimization.

REFERENCES

[1] Kumar, M. Vinoth, et al. "Novel Dynamic Scaling Algorithm for Energy Efficient Cloud Computing." Intelligent Automation & Soft Computing 33.3 (2022).

[2] Kang, Kaixuan, et al. "Adaptive DRL-based task scheduling for energy-efficient cloud computing." IEEE Transactions on Network and Service Management 19.4 (2021): 4948-4961.

[3] Hassan, Mona Bakri, et al. "Green machine learning for green cloud energy efficiency." 2022 IEEE 2nd International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering (MI-STA). IEEE, 2022.

[4] Mahbod, Muhammad Haiqal Bin, et al. "Energy saving evaluation of an energy efficient data center using a model-free reinforcement learning approach." Applied Energy 322 (2022): 119392.

[5] Wang, Yang, et al. "Towards an energy-efficient Data Center Network based on deep reinforcement learning." Computer networks 210 (2022): 108939.

[6] Shaw, Rachael, Enda Howley, and Enda Barrett. "Applying reinforcement learning towards automating energy efficient virtual machine consolidation in cloud data centers." Information Systems 107 (2022): 101722.

[7] Wang, Bin, Fagui Liu, and Weiwei Lin. "Energy-efficient VM scheduling based on deep reinforcement learning." Future Generation Computer Systems 125 (2021): 616-628.

[8] Hou, Huanhuan, Siti Nuraishah Agos Jawaddi, and Azlan Ismail. "Energy efficient task scheduling based on deep reinforcement learning in cloud environment: A specialized review." Future Generation Computer Systems 151 (2024): 214-231.

[9] Bharany, Salil, et al. "Energy efficient fault tolerance techniques in green cloud computing: A systematic survey and taxonomy." Sustainable Energy Technologies and Assessments 53 (2022): 102613.

[10] Khan, Tahseen, et al. "Machine learning (ML)-centric resource management in cloud computing: A review and future directions." Journal of Network and Computer Applications 204 (2022): 103405.

[11] Bal, Prasanta Kumar, et al. "A joint resource allocation, security with efficient task scheduling in cloud computing using hybrid machine learning techniques." Sensors 22.3 (2022): 1242.

[12] Magotra, Bhagyalakshmi, Deepti Malhotra, and Amit Kr Dogra. "Adaptive computational solutions to energy efficiency in cloud computing environment using VM consolidation." Archives of computational methods in engineering 30.3 (2023): 1789-1818.

[13] Zeng, Jing, et al. "Adaptive DRL-based virtual machine consolidation in energy-efficient cloud data center." IEEE Transactions on Parallel and Distributed Systems 33.11 (2022): 2991-3002.

[14] Hassan, Mona Bakri, Elmustafa Sayed Ali Ahmed, and Rashid A. Saeed. "Green machine learning approaches for cloud-based communications." Green Machine Learning Protocols for Future Communication Networks. CRC Press, 2023. 129-160.

[15] Srikanth, G. Umarani, and R. Geetha. "Effectiveness review of the machine learning algorithms for scheduling in cloud environment." Archives of Computational Methods in Engineering 30.6 (2023): 3769-3789.

[16] Carrión, Carmen. "Kubernetes scheduling: Taxonomy, ongoing issues and challenges." ACM Computing Surveys 55.7 (2022): 1-37.

[17] Rejiba, Zeineb, and Javad Chamanara. "Custom scheduling in kubernetes: A survey on common problems and solution approaches." ACM Computing Surveys 55.7 (2022): 1-37.

[18] El Haj Ahmed, Ghofrane, Felipe Gil-Castiñeira, and Enrique Costa-Montenegro. "KubCG: A dynamic Kubernetes scheduler for heterogeneous clusters." Software: Practice and Experience 51.2 (2021): 213-234.

[19] Berl, Andreas, et al. "Energy-efficient cloud computing." The computer journal 53.7 (2010): 1045-1051.

[20] Sharma, Yogesh, et al. "Reliability and energy efficiency in cloud computing systems: Survey and taxonomy." Journal of Network and Computer Applications 74 (2016): 66-85.

[21] Kaur, Tarandeep, and Inderveer Chana. "Energy efficiency techniques in cloud computing: A survey and taxonomy." ACM computing surveys (CSUR) 48.2 (2015): 1-46.

[22] Delgarm, Navid, et al. "Multi-objective optimization of the building energy performance: A simulation-based approach by means of particle swarm optimization (PSO)." Applied energy 170 (2016): 293-303.

[23] Weerasiri, Denis, et al. "A taxonomy and survey of cloud resource orchestration techniques." ACM Computing Surveys (CSUR) 50.2 (2017): 1-41.

[24] Malviya, Anshita, and Rajendra Kumar Dwivedi. "A comparative analysis of container orchestration tools in cloud computing." 2022 9th International Conference on Computing for Sustainable Global Development (INDIACom). IEEE, 2022.

[25] Domaschka, Jörg, et al. "Beyond mere application structure thoughts on the future of cloud orchestration tools." Procedia Computer Science 68 (2015): 151-162.