Indonesian Journal of Electrical Engineering and Informatics (IJEEI)
Vol. 11, No. 3, September 2023, pp. 834~853
ISSN: 2089-3272, DOI: 10.52549/ijeei.v11i3.5037

Machine Learning Centered Energy Optimization In Cloud Computing: A Review

Nomsa Puso[1], Tshiamo Sigwele[2], Oba Zubair Mustapha[3]
[1,2]Department of Computer Science and Information Systems, BIUST University, Botswana
[3]Institute of Technology, Kwara State Polytechnic, Nigeria

Article Info

Article history:

Received Aug 19, 2023
Revised Sep 3, 2023
Accepted Sep 22, 2023

Keywords:

Energy Efficiency
Cloud Computing
Machine Learning
Reinforcement Learning
Virtual Machines

ABSTRACT

The rapid growth of cloud computing has led to a significant increase in energy consumption, which is a major concern for the environment and economy. To address this issue, researchers have proposed various techniques to improve the energy efficiency of cloud computing, including the use of machine learning (ML) algorithms. This research provides a comprehensive review of energy efficiency in cloud computing using ML techniques and extensively compares different ML approaches in terms of the learning model adopted, ML tools used, model strengths and limitations, datasets used, evaluation metrics and performance. The review categorizes existing approaches into Virtual Machine (VM) selection, VM placement, VM migration, and consolidation methods. This review highlights that among the array of ML models, Deep Reinforcement Learning, TensorFlow as a platform, and CloudSim for dataset generation are the most widely adopted in the literature and emerge as the best choices for constructing ML-driven models that optimize energy consumption in cloud computing.

Corresponding Author:

Tshiamo Sigwele,
Department of Computer Science and Information Systems,
Botswana International University of Science and Technology (BIUST),
Plot 10071, Boseja, Palapye, Botswana.
Email: sigwelet@biust.ac.bw

1. INTRODUCTION

Cloud computing is a fast-growing technology combining two significant trends: Information Technology (IT) efficiency and business agility [1], [2] . The National Institute of Standards and Technology (NIST) defines Cloud computing as a unique model for enabling convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with the least service provider interactions and management efforts. These resources are networks, servers, storage, applications, and services. The cloud offers the ability to store data without restrictions and to hide a vast amount of data from other users. The users can access the required files, documents, and applications on demand. Users only pay for the services provided by the cloud vendors instead of buying the expensive Infrastructure.

A data centre's power consumption is divided into three categories: cooling systems, data centre networks and servers [1]. Cooling systems takes 15% to 30% of the power, and servers consume 40% to 55% [1]. The network consumes 10% to 25% of the power and it was also reported that cloud data centres alone consume approximately 7% of global electricity and are expected to rise to 13% by 2030 [3]. This sector is also responsible for an estimated 2% of global emissions, comparable to the aviation industry. The energy consumption of data centres is expected to exceed 140 billion kilowatt-hours per year [4]. The energy-saving scheduling of data centres is critical for cloud service providers and will also contribute to environmental sustainability. The major problem in the currently in cloud computing energy optimization research is that there are limited critical reviews articles for machine learning based approaches in cloud computing energy

optimization to direct and help researchers. There are to the best of our knowledge no reviews that can suggest the best ML model, tools, datasets, and evaluation metrics for current and future research. This study provides a comprehensive review of energy efficiency in cloud computing using ML techniques and extensively compares and contrasts different ML approaches in terms of the learning model adopted, ML tools used, model strengths and limitations, datasets used, and evaluation metrics. Below are the contributions of this review article:

1. To the best of our knowledge, this study has made its ~~first~~ attempt to extensively compare and construct Machine Learning (ML)-based energy efficiency optimization approaches in cloud computing in terms of the approach's learning models adopted, ML tools used, datasets used, and evaluation metrics.
2. The study investigated various ML-driven models that optimize energy consumption in cloud computing and then recommended to future researchers that, Deep Reinforcement Learning, TensorFlow as a platform, and CloudSim for dataset generation are the best options and most suitable.
3. The study discussed the challenges and limitations of the current approaches and highlighted future research directions.

This article is organized as follows. Section 2 presents related works on non-machine learning based energy efficiency techniques in cloud computing, categorized into Dynamic Voltage Frequency Scaling (DVFS), Aware consolidation, VM placement, VM migration, and VM selection. Section 3 compares and contrast the latest ML based energy efficiency techniques in cloud computing. Section 4 presents the results and discussions of the study by performing a detailed analysis of related works to deduct the most adopted models, objectives, tools, datasets, etc. Section 5 presents the future directions and Section 6 concludes the research study.

2. RESEARCH METHOD

The objective of the review paper is to provide a comprehensive overview of the state-of-the-art research and advancements in machine learning techniques for energy optimization in cloud computing environments. The review paper will focus on recent research published in peer-reviewed journals, conference proceedings, and other reputable sources. The time frame for selecting literature will be from 2017 to 2023. The review will follow a systematic approach to literature review, involving the following key steps: (1) Literature Searching using online academic databases including ScienceDirect, SpringerLink, IEEE Xplore, ACM Digital Library, Google Schoolar and Elsevier's Cloud Computing Journal to identify relevant articles using keywords such as "machine learning," "energy optimization," "cloud computing,", "power-efficient," "resource allocation". (2) Inclusion Criteria of articles that specifically address machine learning techniques applied to energy optimization in cloud computing between 2017 and 2023. (3) Exclusion Criteria of articles that are not focused on machine learning or energy optimization, or those that are not related to cloud computing. (4) Data Extraction of key information from selected articles, including title, authors and years, machine learning models, research objectives, evaluation metrics, model limitations, datasets and tools used. (5) Synthesis: Organizing the extracted information into a summarized table then extracting trends and patterns of the most adopted models, objectives, metrics, tools, and datasets. For a detailed overview of cloud computing, the reader is directed to [5][6][7] [8].

2.1. Non-Machine Learning Based Cloud Computing Energy Optimization Approaches

Figure 1 shows several non-machine learning based energy efficiency techniques in cloud computing, categorized into Dynamic Voltage Frequency Scaling (DVFS), Aware consolidation, VM placement, VM migration, and VM selection. These techniques are described in the subsequent sections.

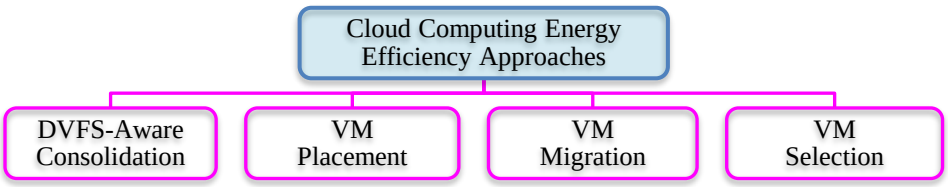

Figure 1. Energy-Efficient Approaches for Cloud Computing.

(1) DVFS-Aware Consolidation Energy Efficiency Methods: Dynamic Voltage Frequency Scaling (DVFS) is a state-of-the-art energy-saving technique for reducing the power consumption in current computer systems. This technique enables CPUs to run at various combinations of clock frequencies and voltages based on system performance requirements at a given time [9]. Consolidation refers to the live migration of virtual machines from one host to another with minimum execution-related delays [10]. DVFS-Aware Consolidation changes

the voltage and frequency automatically to cut down on CPU heat generation and power consumption. Furthermore, less heat generated enables cooling systems to be switched off while saving more energy [9]. The author in [9] proposed an energy efficiency heuristic using VM consolidation (EEHVMC) to minimize energy or power consumption, VM migrations and reduces service level agreement violations. The fundamental concept categorizes host computers according to their CPU and memory consumption. The host machines are divided into three primary groups by defining two criteria related to CPU and memory utilization: host overloaded (HOL) and host medium loaded (HML). Lastly, the host underloaded (HUL) machines. VMs were moved to the HML as of HOL to reduce the power consumption in cloud data centers. The suggested method reassigned the VMs to the HML from the HUL hosts and switched the inactive hosts into power-saving mode [9]. The author [11] created an adaptive VM consolidation mechanism based on the DRL method called ADVMC for energy-efficient cloud data centers. It implements both the VM placement and selection for reducing energy consumption and eliminating the SLA violations of users as compared to many other VM consolidation strategies. Based on workload detection, hosts in a cloud data center are divided into three types: overloaded hosts, underloaded hosts, and regular hosts whose workload falls between overloaded hosts and underloaded hosts [11].

The author in [12] proposed an energy-efficient and quality of service aware VM consolidation strategy for improving resource utilization and saving energy on the cloud data centers. Dynamic VM consolidation is a highly effective strategy for increasing resource utilization and maintaining a regular operational state while upholding SLAs for all hosts. To efficiently reallocate VMs to hosts, this dynamic VM consolidation strategy is divided into several steps which are: identifying overloaded host, choosing the migrated VM from the overloaded host, identifying the underloaded host then placing the VM. Figure 2 shows a summary of the evaluation metrics, strength's and limitations of DVFS Aware consolidation method for cloud computing energy optimization.

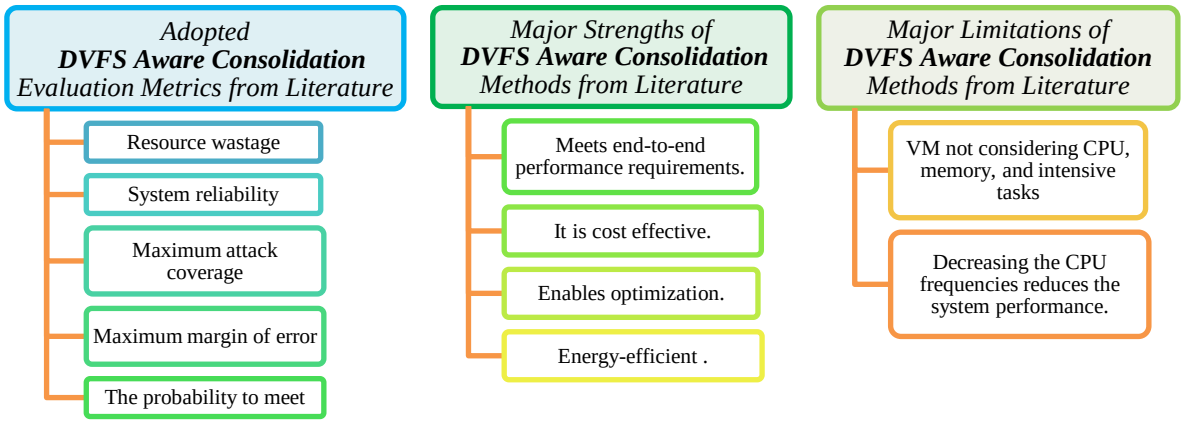

Figure 2. Literature Summary of evaluation metrics, strength's and limitations of DVFS Aware consolidation method for cloud computing energy optimization.

(2) VM Placement Energy Efficiency Methods: VM placement methods are sometimes called VM Allocation or VM Scheduling. In large cloud data centers, VM placement (VMP) is the process of choosing the best suitable Physical Machine (PM) to host the user's requested virtual machines [13]. It is a technique that executes VMs to enhance power efficiency and maximize resource utilization. The author in [13] proposed a holistic resource management strategy named bio-inspired virtual machine (Bio-VMP) for cloud environment that is both sustainable and energy efficient. This technique used a novel flower pollination-based non-dominated sorting optimization (FP-NSO) method which maximizes resource utilization while minimizing energy consumption and carbon emissions (CE) of the data center. It is based on the concepts of non-dominated sorting technique-based genetic method (NSGA-II) and flower pollination optimization (FPO). It also explores the most possible and optimal virtual machine placement allocations subjected to energy consumption, resource utilization, and CE [13]. A prediction-aware deep reinforcement learning (DRL) based VM placement technique (PADRL) was created by the author in [11], in order to find acceptable hosts for VMs to be migrated in an efficient way. The PADRL scheme is comprised of two primary components, which are long short-term memory (LSTM) based state prediction and deep Q-learning (DQN) based VMP. The VMs that needed to be migrated were set up on suitable hosts, including all the VMs from underloaded hosts and a few from overloaded hosts. This method is achieved in terms of energy saving and reduction of service level agreement violations in cloud data centers [11].

Using the dominance-based multi-objective artificial bee colony (MOABC) technique, the author in [14] presented a multi-objective VMP for achieving the best VMs to PM mapping. This technique balanced overall energy consumption, resource waste, and system reliability to meet QoS and SLA requirements. The list of migrating virtual machines, destination physical machines, number of solutions as well as the maximum number of repetitions is the first input parameters used by the suggested algorithm [14]. The author in [12] proposed a virtual machine placement heuristics strategy called CUECC for choosing a targeted host with the greatest reward in combination with both real-time CPU utilization and energy consumption changes when the virtual machine is deployed. The global manager (GM) gathered the running status of hosts in the data center while local manager (LM) monitored the status of hosts to check if they are underloaded or overloaded. Therefore, the most underloaded host migrated all the VMs to some hosts then switched the host to inactive state for saving energy. Figure 3 shows a summary of the evaluation metrics, strength's and limitations of VM Placement method for cloud computing energy optimization.

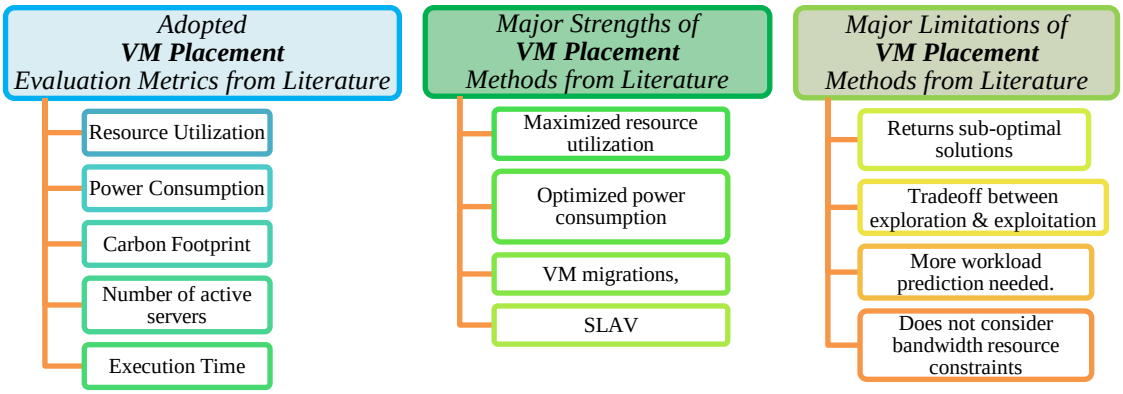

Figure 3. Literature Summary of evaluation metrics, strenghths and limitations of VM Placement method for cloud computing energy optimization.

(3) VM Migration Energy Efficiency Methods: Virtual Machine (VM) migration is an enchanting feature of virtualization technology that plays an essential role in the administration of cloud data centers, it shifts the running virtual machines from a physical machine to another [15]. Author in [15] proposed a plethora of VM migration strategies, aiming at serving the QoS-driven user requirements and also reviewing the most recent and cutting-edge load balancing, energy-aware, SLA-aware, and network-aware live VM methods based on artificial intelligence (AI). Several methods have been proposed to achieve effective VM migration and these techniques are divided into non-live and live VM migration. The live VM migration as the name implies keeps providing services to users while the virtual machine is being moved in order to ensure continuous connectivity and efficient resource utilization. Whereas the non-live VM migration stops all the VM services while it is in the process and resumes them when it is done. The VM migration controller manages all aspects of the VM migration process, whether it is non-live or live [15]. To address the issues faced due to the increment of cloud data centers, the author in [16] offered an energy aware VM allocation and migration approach. ML based artificial bee colony (ABC) is used for ranking the VMs with respect to the load while considering the energy efficiency as a vital metric. The most efficient VMs were further selected based on the dynamics of the load and energy, applications were migrated from one VM to another. Also, active VM servers prevented the resource under-utilization by reducing the resource idling time [16].

The author [17] in proposed a VM migration method called V2PQL which is based on Markov decision process and Q-learning technique. One of the main benefits of this virtual technology is that it freed VMs from the underlying hardware, allowing flexibly management of resources and maximizing the use of shared resources. By shifting VMs from overloaded PMs to light-loaded ones, a cloud system achieved load balancing. VMs that were operating on a light load PM were merged into another physical machine for saving energy by lowering the number of running PMs. A cross-data center virtual machine migration approach called EVMA was suggested by author in [18] considering the issue of energy consumption in multi-data center environment. The destination data center for the VM migration was selected first based on the bandwidth between data centers and this approach of choosing an overload host as well as VM was chosen based on the historical CPU load. Studies revealed that the algorithm performed well in lowering the data center's energy consumption while maintaining QoS [18]. Figure 4 shows a summary of the evaluation metrics, strength's and limitations of VM Migration method for cloud computing energy optimization.

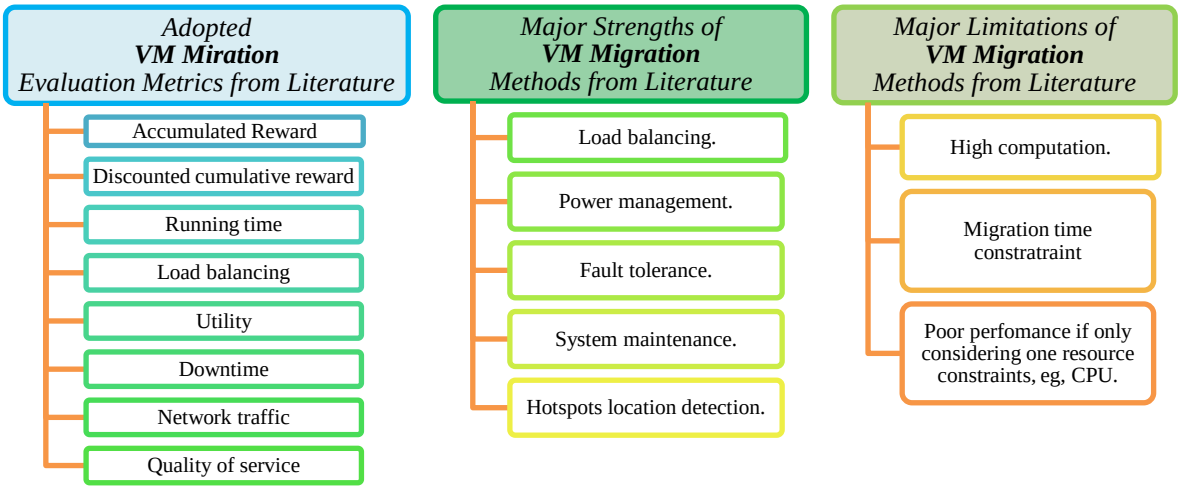

Figure 4. Literature Summary of evaluation metrics, strength's and limitations of VM Migration method for cloud computing energy optimization.

(4) VM Selection Energy Efficiency Methods: Virtual Machine Selection is a crucial strategy for enhancing power proficiency and asset usage in cloud infrastructures [19]. It is concerned with assigning computational tasks to VMs. Varying central processing unit (CPU) utilization of virtual machines is one of the leading causes of fluctuating CPU utilization of hosts. The virtual machine monitor (VMM) takes over when the host is identified as overloaded. The author in [12] used VM selection strategy called AUMT to reduce the amount of energy consumption from loud data centers, the number of migrations and to also improve the quality of service (QoS). It is based on average CPU utilization of the VMs as well as migration time. The author in [14] introduced a novel heuristic VM selection policy that considers migration time and total number of migrations to reduce the harmful impacts of migrations on QoS. Investigating task completion times while selecting VMs for migration is one of this policy's most vital points. In this regard, a constraint was defined for selecting only those VMs that could not complete their tasks before the mean first passage time in the current physical machine (PM) to avoid further VM migration that leads to the improvement of QoS [14].

The author in [11] explored a dynamic influence coefficient based VM selection algorithm (ICVMS) to preferentially select those VMs with the largest influence coefficient (IC) value, that is the VMs with the most impact for migration, which helps to remove excessive workloads rapidly and accurately for overloaded hosts. A list of VMs were selected for migration to restore the overloaded host to a normal state. A new VMs selection method named MRCU (maximum ratio of CPU utilization to memory utilization) was proposed by author in [9], to choose VMs for migration when CPU intensive tasks overload a host. The MRCU approach considers both the central processing unit and memory components. Since a higher CPU workload result in a larger power consumption, this algorithm selects a virtual machine with the highest CPU value for migration and its objective is to save energy [9]. Figure 5 shows a summary of the evaluation metrics, strengths, and limitations of VM Selection method for cloud computing energy optimization.

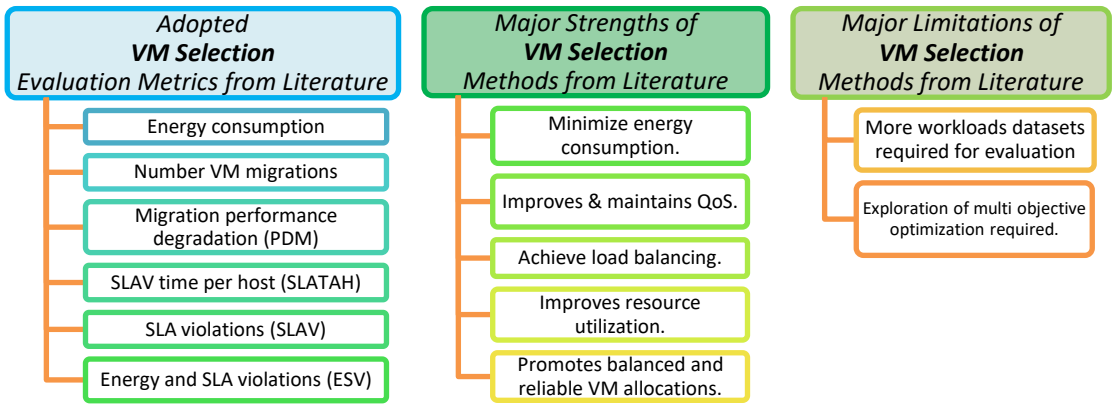

Figure 5. Literature Summary of evaluation metrics, strength's and limitations of VM Selection method for cloud computing energy optimization.

2.2. Machine Learning Based Cloud Computing Energy Optimization Approaches

Machine Learning (ML) is a kind of artificial intelligence that simulates human learning [20]. It enables computers to develop their predictive abilities up to the point where they can carry out tasks on their own without having to be programmed. Based on historical training data, ML-driven software applications may forecast future results [21]. It takes a lot of data, computing power, and Infrastructure to train an accurate ML model. ML models can be trained using a cloud ML platform, which offers the computation, storage, and services needed. ML is made more affordable, flexible, and accessible due to cloud computing, which also enables faster ML algorithm development. The benefits of Machine Learning in the Cloud are as follows; According to authors in [22], many businesses have the resources to develop ML models internally using open-source frameworks like Scikit Learn, TensorFlow, or PyTorch. Even if internal teams are competent at creating algorithms, they frequently struggle to scale models for usage with real-world workloads and deploy them in production, which frequently calls for sizable computing clusters. ML workloads that come in bursts function well on the pay-per-use cloud [22]. Enterprises can quickly scale up as initiatives go into production and demand rises due to the cloud's ease of ML experimentation. The Limitations of Machine Learning in the Cloud are as follows; Although ML has transformed several industries, it frequently falls short of expectations. This can be attributed to several factors, including a lack of appropriate data, a lack of data availability, data bias, challenges with confidentiality, poorly planned goals and algorithms, improper tools and personnel, a shortage of resources, and evaluation concerns [23]. It can be challenging to switch systems across clouds or services when running ML models on the cloud. To accomplish this, the data must be moved in a way that preserves model performance. Little changes in the input data can often have a big impact on ML models [24]. The same security issues apply to cloud-based ML as they do to any cloud computing platform. Attackers may hack cloud-based ML systems, which are frequently exposed to public networks and susceptible to manipulation of ML results or infrastructure cost increases.

Nevertheless, there are several applications for ML. ML algorithms are essential in situations when deployment is required for development. One of the main factors influencing ML solutions' widespread acceptance is their dynamic nature. These algorithms can be used to replace some human tasks because of how flexible they are. The finest illustration of this is the fact that chatbots that use natural language processing are replacing customer service representatives. These chatbots operate by evaluating consumer inquiries and responding to them automatically [23]. ML focuses on the process through which computers learn how to carry out activities without having received specialized training [23]. Computers are used to analyze and identify trends from data. Figure 6 shows various ML types and techniques or algorithms. ML algorithms can be trained using various techniques, each of which has advantages and disadvantages of its own. The ML algorithms are generally divided into three groups: supervised, unsupervised, and reinforcement learning.

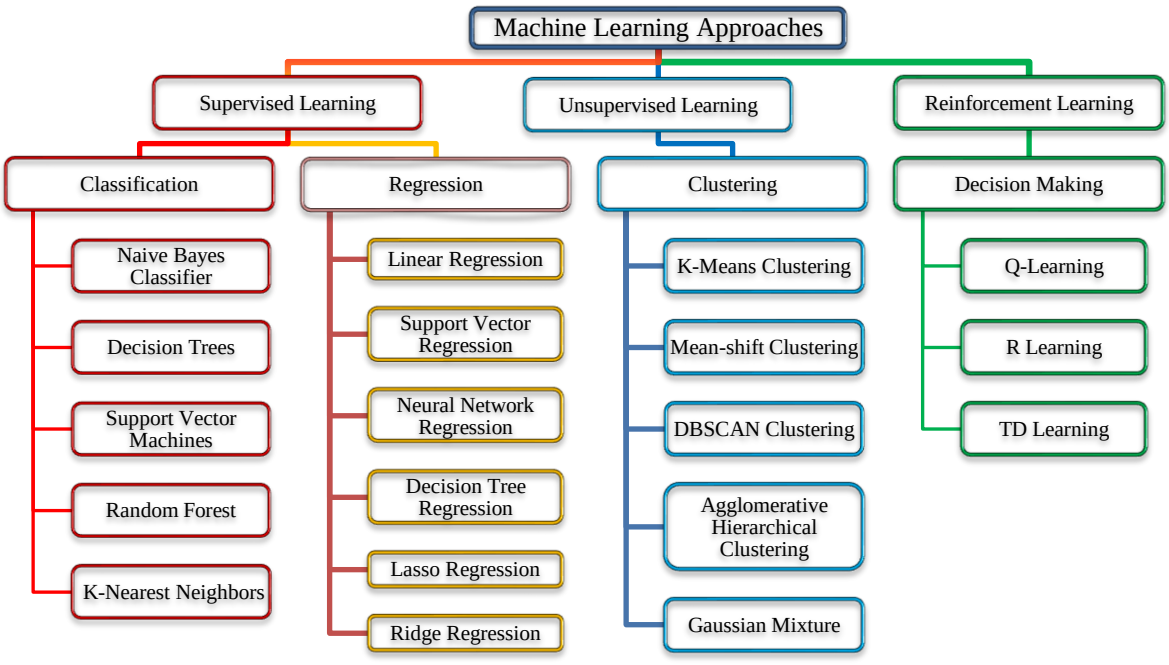

Figure 6. ML types and techniques or algorithms.

(1) Supervised Learning: The first category in the ML hierarchy is supervised learning. When given a dataset, which includes both the inputs and the matching labelled outputs, then it is modeled by supervised algorithms [23]. The training samples in the dataset have inputs that have been mapped to the matching outputs. Each training example in the mathematical model is described by an array called a feature vector, and the training data is represented by a matrix through an iterative process. A function that is created by supervised learning algorithms can be used to forecast the results of brand-new inputs. That task has been learned by an algorithm that gradually improves the precision of its predictions over time. Classification and regression are other divisions of supervised learning [23]. Supervised ML allows the collection or production of data output from previous experience and helps optimize performance criteria [25]. Also, it helps solve various types of real-world computation issues. A lot of computation time is required for supervised learning training. Unlike unsupervised learning, it cannot cluster or classify data by independently determining its features [25].

(2) Supervised Learning: The second category in the ML hierarchy is unsupervised learning. Unsupervised learning techniques start with a data set that only contains the inputs and looks for patterns, such as grouping or clustering of data points [23]. In this type of learning, algorithms are trained on test data that has not been labelled. Unsupervised learning algorithms operate by finding commonalities rather than responding to feedback. It is used in the identification of anomalies, clustering, and determination of the probability density function. It deals with clustering. It is utilized to uncover hidden patterns that are of utmost value to the industry and have several real-time applications since it can perceive things that human minds cannot [20]. Compared to the supervised learning task, it is less complex. As there is no label or output measure to prove its usefulness, it is not always certain that the results will be valuable. An unsupervised task's output and sorting cannot be precisely defined. It strongly depends on the model and on the machine. Results are often less accurate [20].

(3) Reinforcement Learning: The third category in the ML hierarchy is Reinforcement Learning (RL). According to the author in [26], an agent is trained in RL (a decision-making technique) to learn a desired behaviour in an interactive environment based on the events it encounters. In essence, the agent gets rewarded for every action it does in a specific state, and this reward represents the success of the selected action in that state. With enough practice, the agent learns which activities produce the highest cumulative reward over time. The environment in which the RL agent functions is frequently modeled using the Markov Decision Process (MDP). As a result, it is believed that state transitions and rewards are governed by the Markov property, which holds that the future state and reward only depend on the present state and the action the agent takes in the current. Reinforcement learning can be used to address immensely challenging issues that are unsolvable using conventional methods [26]. Long-term outcomes, which are very difficult to accomplish, are best achieved with this technique. The model can correct errors made during the training process. For real physical systems, the curse of dimensionality severely restricts reinforcement learning, and it requires a lot of data and computation [26]. A state overload brought on by excessive reinforcement learning may have a negative effect on the results.

This section will present a review of related works which adopted or developed ML models to solve the energy efficiency optimization problem in cloud computing. This is a review of articles over the last five years. The extracted information will be presented in the form of a table, as shown in Table 1. Each column in the table will represent the features of the related work in question. The table rows represent several related works (the authors) and the years they were published. The table will present the reviewed article with its author and the years, the ML model that was proposed to solve the energy efficiency problem in cloud computing, the proposed objective, the evaluation metrics, e.g., accuracy, the limitation of the proposed model, and the datasets and tools adopted. Table 1 shows the reviewed related works on ML models for energy efficiency in cloud computing. A detailed analysis of related works in Table 1 will be presented at the end of the table, showing graphically the most adopted ML models, the most adopted objectives, the most adopted ML evaluation metrics with ranges, e.g., ML model accuracy ranges, prevalent ML model limitations, and most adopted ML datasets and tools. From Table 1, trends and patterns on the most adopted features will be drawn. This review will then specify and suggest the discovered trends and patterns of the reviewed features as a recommendation to current and future researchers.

Table 1. Related Works on Machine Learning based Energy Efficiency in Cloud Computing

	Authors & Years	Machine Learning Models	ML Research Objectives	ML Evaluation Metrics & Performance	ML Model Limitations	ML Datasets and Tools
1.	Liu et al., (2017) [27]	Deep Reinforcement Learning (DRL)	To adaptively allocate resources in cloud computing system.	- Power Consumption (16% reduction) - Latency (10% reduction)	Partially solves the resource allocation problem	Dataset: Actual Google cluster-usage traces ML Tool: Keras
2.	Cai et al., (2017) [20]	K-Means and Page Rank	To minimize the energy consumption of data centre.	- Energy consumption (25% reduction)	High Computation	Dataset: Used the customized dataset ML Tool: Hadoop
3.	Zhong et al., (2017) [25]	Wavelet Support Vector Machine (WSVM)	To analyze the cycle and frequency of the input signal by replacing the kernel function of SVM by a wavelet function.	- CPU Usage (3% increase) - Mean absolute error (6.5% minimized) - Root mean square error (30% minimized)	Does not address workload prediction.	Dataset: Google cloud computing center dataset. ML Tool: TensorFlow
4.	McGough et al., (2018) [28]	Random Forest (RF) And Multilayer Perceptron (MLP)	To minimize the energy consumption of data center.	- Accuracy (70% achieved) - Energy consumption (51.4% reduction) - Overhead (4.9% increase)	Uses real trace-logs for complex. situations to occur in the presented platform.	Dataset: 2010 exemplar dataset. ML Tool: Scikit-Learn
5.	Li et al., (2018) [29]	Deep Reinforcement Learning	To reduce power/energy consumption by enhancing computation speed and hardware footprint reduction	- Power usage (54.1% reduced) - Average job latency (18.7% increase)	The action space in the DRL framework needs to be reduced.	Dataset: Google cluster workload traces ML Tool: TensorFlow
6.	Zhang et al., (2018) [30]	Linear and Logistic Regression	To model and analyze the multi-dimensional cloud resource allocation problem.	- Accuracy (98% achieved) - CPU utilization (100% achieved) - Storage utilization (100% achieved) - Memory utilization (100% achieved)	The resource allocation algorithm does not satisfy the strategy proof of the auction mechanism.	Dataset: Grid Workloads Archive. ML Tool: GNU Octave 4.2.1
7.	Jararweh et al., (2018) [31]	Logistic Regression Model	To predict physical machine overloading	- Energy consumption (Reduced by 67%) - No. of VM migrations (Reduced by 91%) - No. of host shutdowns (reduced by 86%) - SLA violation (Reduced by 45%)	The proposed algorithm was not applied with varying workloads that represent different cloud customers.	Dataset: CloudSim 4.0 dataset. ML Tool: TensorFlow
8.	Moreno-Vozmediano et al., (2019) [32]	Support Vector Machine (SVM) Regression	To optimize the latency of the service and reduce VM over-provisioning.	- Mean Absolute Error (94% MAE reduction) - Root Mean Squared Error (11% RMSE reduction) - SLA Violation (8% reduction)	Only based on normalized polynomial kernels.	Dataset: Real web service logs from Complutense University of Madrid. ML Tool: Weka
9.	Rajalakshmi et al., (2019) [33]	Reinforcement Learning	To improve the quality of the VM consolidation algorithm for energy consumption.	- SLA violation (8.5% reduction) - Energy consumption (32% reduction)	The number of hosts can be increased to simulate the behaviour of the proposed work.	Dataset: CloudSim PLANET LAB workload. ML Tool: TensorFlow
10.	Sui et al., (2019) [34]	K-Means Clustering Algorithm	To reduce cost and response times	- Average utilization, (100% Improvement) - VM migration number (Reduced by 94.5%) - Energy consumption (Reduced by 49.13%)	Only addressed the prediction of the distribution of resource demand.	Dataset: CloudSim 4.0 dataset. ML Tool: Eclipse 4.5.1
11.	Thein et al., (2020) [35]	Reinforcement Learning And Fuzzy Logic	To provide the effective management of physical resources.	- Power Usage Effectiveness (between 1.78 and 1.96) - Resource utilization (Above 50%) - SLA Violation (24% reduction)	For a very large number of infrastructure resources, the scheduling process may become slow.	Dataset: PlanetLab Virtualized Research datasets. ML Tool: MLBox

Table 1 (Continued) Related Works on Machine Learning based Energy Efficiency in Cloud Computing

	Authors & Years	Machine Learning Models	ML Research Objectives	ML Evaluation Metrics & Performance	ML Model Limitations	ML Datasets and Tools
12.	Tong et al., (2020) [36]	Deep Q Learning	To solve the problem of scheduling tasks in cloud computing environment.	- Makespan (5% Improvement) -Load Standard deviation (32% Improvement)	Does not cut costs when after reaching task deadlines	Dataset: CloudSim 4.0 dataset ML Tools: TensorFlow
13.	Ding et al., (2020) [24]	Q Learning	Proposed a Q-learning based task scheduling framework for energy-efficient cloud computing.	- Energy consumption (25% reduction) - Average response time (68.7% reduction) - CPU utilization rate (20% increase)	The model evaluation did not consider scalability with the increase in the number of VMs.	Dataset: CloudSim 4.0 dataset. ML Tool: TensorFlow
14.	Asghari et al., (2020) [37]	Cooperative Reinforcement Learning (RMFW Model)	To reduce user costs, energy consumption, and perform load balancing of resources.	- Scheduling time (61% reduction) - Makespan (20% reduction) - Resource Utilization (29% improvement) - Cost (32% reduction) - Energy (50% reduction)	Loss of accuracy due to discretization of state space.	Dataset: CloudSim 4.0 dataset. Tool: TensorFlow
15.	Madhusudha et al., (2021) [38]	Random Forest (GA-RF Model)	To minimize power consumption while maintaining load balance yet maximizing resource utilization.	- Energy Consumption (39% reduction) - Execution Time (37% reduction) - Resource Utilization (11% improvement) - Average Start Time (46% reduction) - Average Finish Time (43% reduction)	The model was not tested with various ML and deep learning approaches for the better solutions	Dataset: Real workload traces from PlanetLab. ML Tool: TensorFlow
16	Yan et al., (2021) [39]	Deep Q-Learning (Dueling-DDQN)	To reducing power consumption, ensuring resource load balance, and improving user service quality.	- Average Reward (20% Improvement) - Power Consumption (30% reduction)	Do not have specific optimization goals.	Dataset: CloudSim 4.0 dataset. ML Tool: TensorFlow (TF) A Jararweh gents
17.	Wang et al., (2021) [4]	Deep Reinforcement Learning	Proposed a DRL model based on QoS feature learning to optimize data centre resource scheduling.	- Energy consumption (22% improvement) - SLA rate (4.5%–26% improvement)	Affected by response time, reliability, and other parameters.	Dataset: 1998 World Cup website workload ML Tool: TensorFlow
18.	Caviglione et al., (2021) [40]	Deep Reinforcement Learning (DRL VMP)	To selects the most suitable PMs to deploy VMs requested by users	- QoE (50% improvement) - Consumed power (7% reduction) - Total reward (75% improvement)	Optimization problem (indirectly acted by means of heuristics).	Dataset: Traces taken from Netalia (www. Netalia.it). ML Tool: TensorFlow
19.	Shaw et al., (2022) [3]	Reinforcement Learning	To solve the VM consolidation problem for improved cloud resource management.	- Energy Consumption (47% reduction) - Energy efficiency (Improved by 25%) - Service violations (Reduced by 63%)	A limitation of the tradeoff between exploration and exploitation.	Dataset: Real workload data from PlanetLab. ML Tool: TensorFlow
20.	Jayanetti et al., (2022) [26]	Deep Reinforcement Learning	Energy-efficient resource scheduling in edge-cloud environment.	- Energy consumption (56% improvement) - Execution time (46% improvement)	Only operates in a centralized manner.	Dataset: CloudSim 4.0 dataset. ML Tool: Keras
21.	Chen et al., (2023) [41]	Deep Reinforcement Learning (DQTS)	To dynamically and collaboratively schedule high-dimensional cloud objectives.	- Makespan 2484 (16.6% improvement) - Fairness (Improve by 5.3%.)	Performance degradation in larger-scale workflow scheduling	Dataset: CyberShake, Epigenomics, ML Tools: PyTorch
22.	Cui et al., (2023) [42]	Deep Reinforcement Learning (DQN)	Optimizing the computing, caching, and communication resources	- Network Delay (44% reduction) - Reward per episode (39% improvement)	A tradeoff between energy consumption and network not considered	Dataset: Not Specified ML Tools: Not Specified

Table 1 (Continued) Related Works on Machine Learning based Energy Efficiency in Cloud Computing

	Authors & Years	Machine Learning Models	ML Research Objectives	ML Evaluation Metrics & Performance	ML Model Limitations	ML Datasets and Tools
23.	Sen et al, (2019) [43]	Reinforcement Learning based Task Assignment approach (RILTA)	Ensures the timeliness guaranteed execution of intelligent cognitive assistants' tasks with high energy efficiency.	- Guarantee ratio (13%-22% improvement) - Average energy consumption (25% reduction of 1MJ) -Task allocation time (25% improvement) - Task running time (34%-51% faster) - Percentage deadline (63%-68% decrease)	State space only consists of processor capacity and available bandwidth. Cannot always find the optimal solution due to limitation in finding the best policy using Q-learning, with high variance error bars.	Dataset: Generated via uniform distribution from iFogSim Tool: iFogSim
24.	Kumar et al., (2022) [44]	ML frameworks among them: Supervised learning, neural networks, support vector regression, etc	A Review: Reviews the ML based cloud improvement through dynamic load allocation, task scheduling, energy optimization, live migration, etc	The review table presented compares reviewed literature based on the year, authors, learning model, benefits, limitations, platforms adopted and the performance.	No dataset comparison. No analysis of most adopted datasets, tools, metrics, performance, etc. Outdated articles up to 2021.	Dataset: The review does not compare any datasets. ML Tool: several tools compared.
25.	Demirci, (2016) [45]	Presenting several ML schemes: Supervised, Unsupervised, Reinforcement and Hybrid	A review: Provides a comparative classification of ML models in the cloud for energy reduction.	The review table presented compares reviewed literature based on the year, authors, learning model and objectives.	Does not compare based on metrics, performance, tools limitation, datasets. Outdated articles up to 2016.	Dataset: No comparison. ML Tools: No comparison.
26.	Khan et al., (2022) [46]	Presents several ML centric resource management in cloud computing	A review: The review compares several ML approaches based on experiments configuration, datasets, performance, limitations.	Energy Consumption (Best reviewed article has 38% energy reduction) Several other performance metrics of articles were presented; accuracy, latency, etc.	No analysis of most adopted datasets, tools, metrics, performance, etc. Outdated with articles up to 2021. Missing values on performance for some article	Dataset: Several datasets were compared with. ML tools: Several tools were compared with.
27.	Soni et al., (2022) [47]	Presents several ML schemes generally in cloud computing	A review: Compares machine learning techniques in emerging cloud computing integrated paradigms	Several performance metrics; power consumption, latency, delay, accuracy, cost, utilization, etc.	No comparison on; performance but only accuracy performance presented. No comparison based on ML tools.	Dataset: MNIST, LSTM, Cifar10, etc. ML tools: No comparison.

3. RESULTS AND DISCUSSIONS

This section presents a detailed analysis of the literature review presented previously in Table 1. This section will provide a detailed analysis of the most adopted ML models, the prevalent objectives of current approaches, the most adopted evaluation metrics used in current approaches with ranges, e.g., the minimum and maximum accuracy that has ever been achieved in current works, the most prevalent limitations of current related works, the most adopted ML datasets and tools.

4.1. The Most Adopted ML Models from Literature

Figure 7 shows an analysis of the most adopted ML models from the literature for cloud computing energy optimization. The analysis shows that Deep Reinforcement Learning (DRL), Support Vector Machine (SVM) Regression, K-means, Random Forest (RF), and Linear and Logistic Regression were among the ML models adopted. Figure 6 shows that 60% of related works adopted the DRL model to solve energy efficiency problems on the cloud. The reason why DRL is the most adopted model is because, compared to other models, the DRL algorithm [43]–[45]:

- Can handle complex and high-dimensional state and action spaces, which are common in many real-world applications.
- Learn complex policies that are difficult to specify manually by learning to maximize a reward signal over a long-time horizon.
- Can learn online and adapt to changing environments, making them suitable for applications where the environment is dynamic or uncertain.
- Can leverage large amounts of data to improve performance by training on massive datasets or using techniques such as data augmentation or transfer learning.

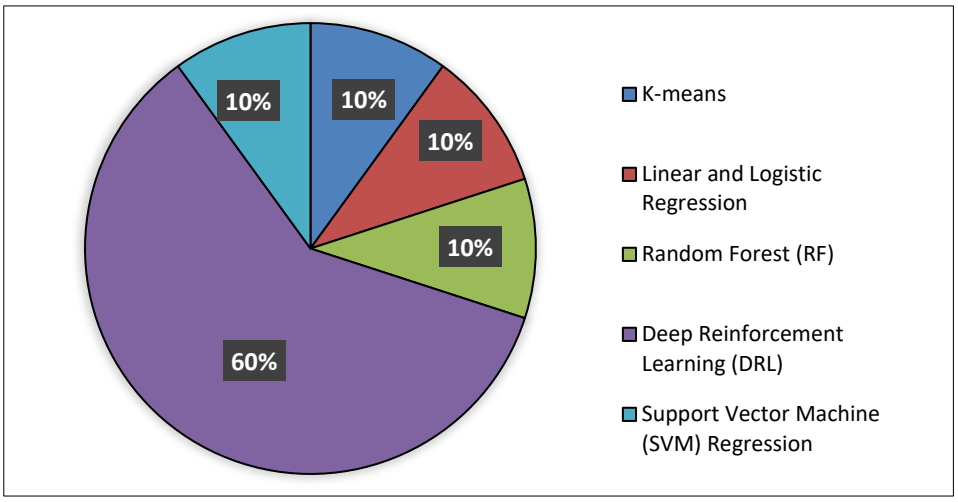

Figure 7. Most adopted ML model from literature for cloud computing energy optimization.

4.2. The Most PREVALENT Objectives from Current Approaches

Figure 8 shows an analysis of research articles from the literature with the most prevalent objectives of using ML for cloud energy minimization. It can be observed from Figure 8 that the most dominant research objective is power or energy consumption reduction at 28%. Other objectives in addition to energy reduction are load balancing at 14%, task/resource scheduling at 10%, VM consolidation at 10%, response time at 7%, resource allocation at 7%, reduce costs at 7%, SLA and resource utilization at 4%, low overhead at 4% and many others. The main reasons for energy consumption reduction as the dominant research objective in literature are because:

- By minimizing energy usage, data centers can reduce their operating expenses and pass on some of the savings to their customers.
- By reducing energy consumption, data centers can operate with fewer resources and plan for future expansion more efficiently.
- By minimizing energy consumption, data centers can operate more sustainably and reduce their overall carbon footprint.
- Minimizing energy consumption can also improve the reliability of cloud computing services. By reducing the load on servers and other hardware components, data centers can extend their lifespan and reduce the risk of hardware failures.

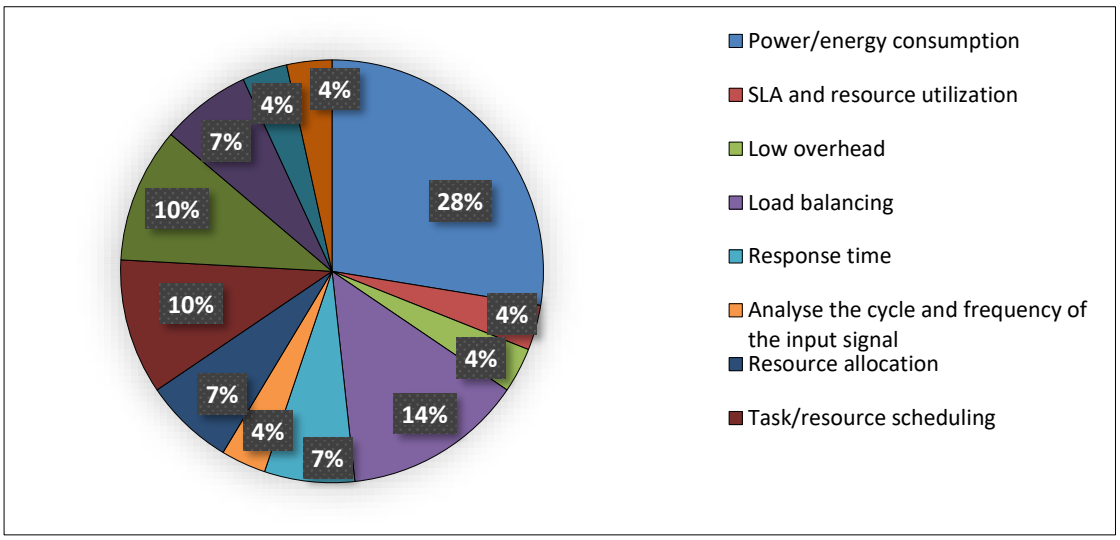

Figure 8. Most adopted research objectives from literature for cloud computing energy optimization

4.3. The Most Adopted ML Datasets from Literature

Figure 9 shows a summary analysis of the most adopted ML datasets from the literature, with a larger percentage of 40% being dominated by the CloudSim datasets simulation toolkit, whilst other datasets were taken from different sources and databases. The main reasons for the CloudSim dataset being dominant in literature is because:

- CloudSim is a widely used simulation tool for modeling and simulating cloud computing environments and is used to simulate a variety of cloud computing scenarios. The dataset includes parameters such as the number of hosts, the number of virtual machines, the amount of memory and storage, and the network bandwidth.
- CloudSim dataset is based on real-world cloud computing environments, which makes the simulations more realistic and accurate.
- CloudSim dataset is flexible and can be customized to suit specific research needs. Researchers can modify the input parameters to simulate different cloud computing scenarios and evaluate the performance of different cloud computing architectures and algorithms.

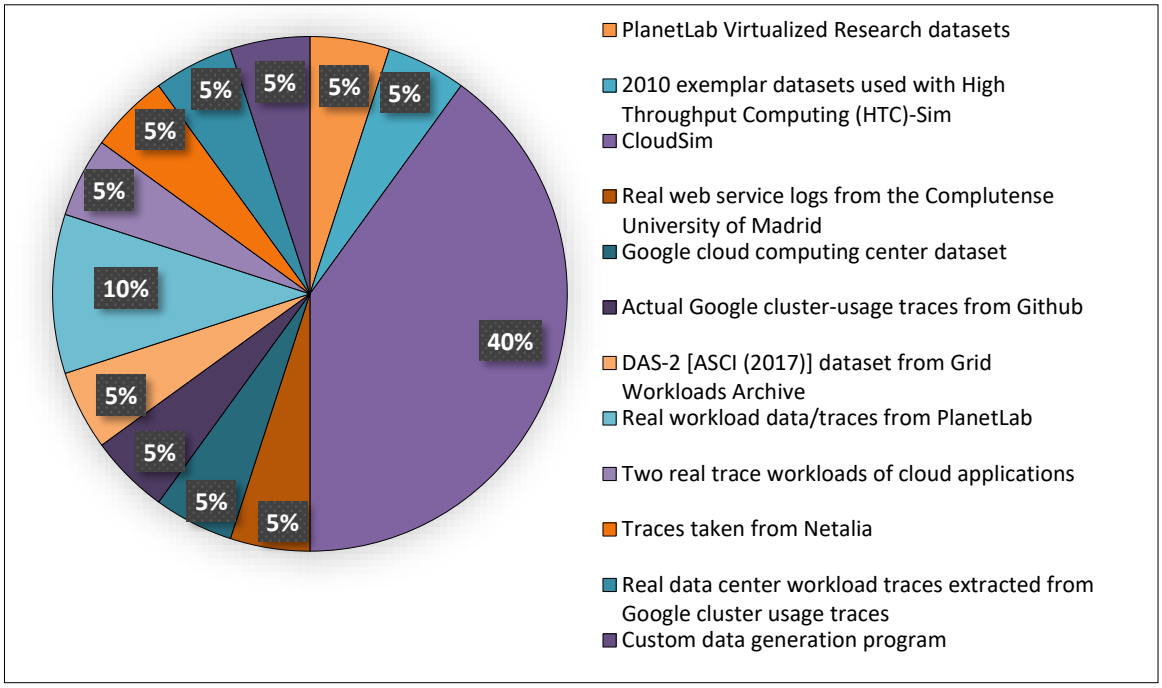

Figure 9. Most adopted ML datasets from literature for cloud computing energy optimization

4.4. The Most Adopted ML Model Evaluation Metrics from Literature.

Figure 10 shows a summary of the most adopted ML evaluation metrics for cloud energy minimization from literature, which is comprised of energy consumption at 13%, accuracy at 9%, power consumption at 4%, execution time at 4%, resource utilization at 4%, average reward at 4%, Makespan at 4%, power usage at 4%, average job latency at 4% and many others. The most adopted metric is energy consumption (28%) followed by accuracy (9%), and the explanation for this is that; for energy consumption, the reasons are similar to why energy consumption is the dominant objective, and this was presented in the previous sections. The reason for accuracy being the next dominant metric is that:

- The accuracy of a model is a primary measure of its performance. High accuracy means that the model is making accurate predictions, while low accuracy indicates that the model is making errors.
- Accuracy is a convenient metric for comparing the performance of different models. It allows us to determine which model is better suited for a particular task based on their accuracy scores.
- Accuracy is also used in the training process to optimize the model's parameters. By monitoring the accuracy of the model during training, we can adjust the parameters to improve the model's performance.

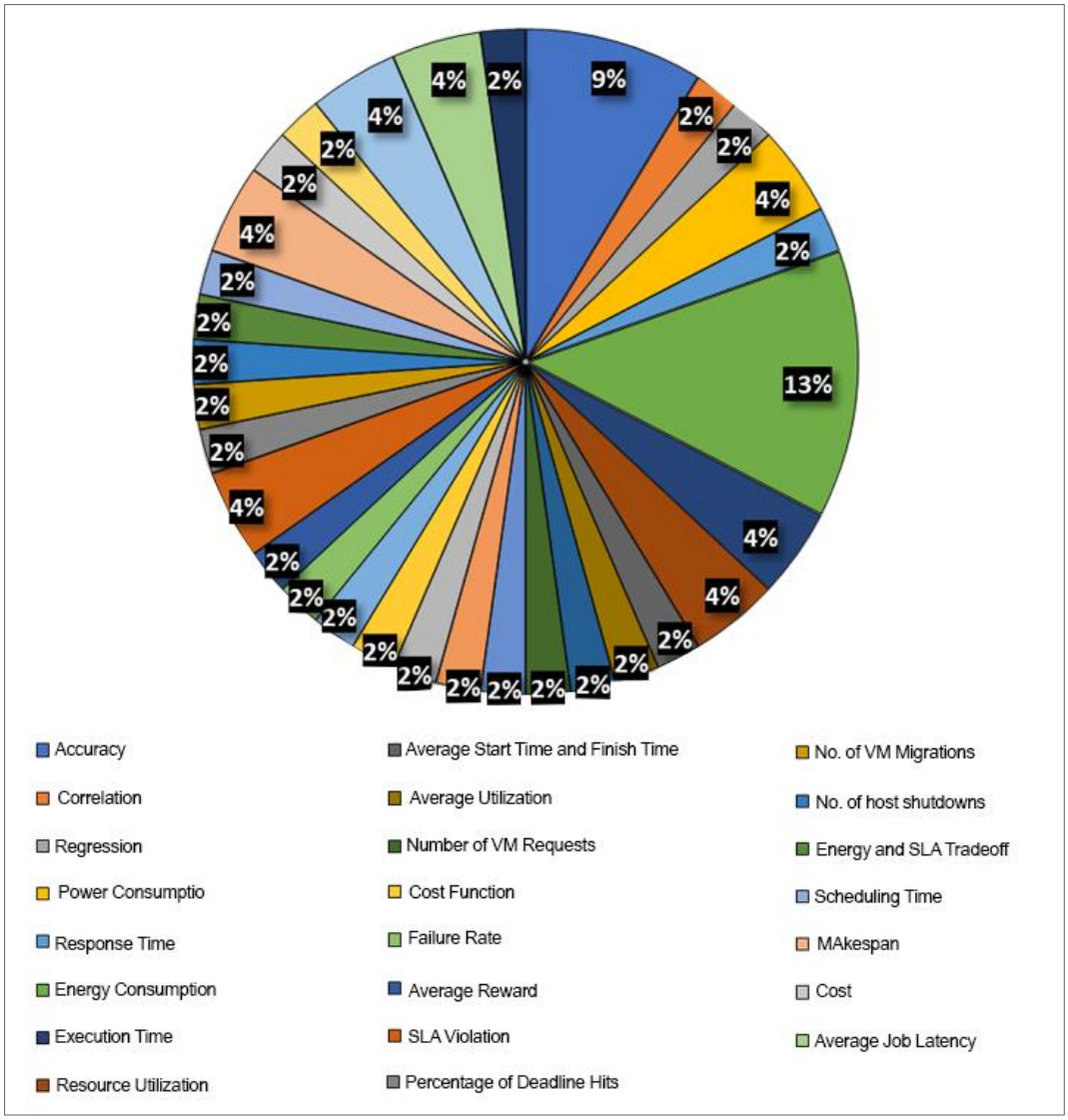

Figure 10. Most adopted ML metrics from literature for cloud computing energy optimization

4.5. The Most Prevalent Limitations with the Current Approaches

Figure 11 shows a percentage summary of the most prevalent limitations of current ML-based approaches for cloud computing energy reduction. A larger portion of current literature approaches is being dominated by poor performance or poor model accuracy or loss of accuracy (15%), model optimization problems (15%) followed by resource allocation problems at 10% and other limitations like high computation (5%), centralized approaches (5%) which are a single point of failure, etc. The poor performance/ loss of accuracy is a major limitation because.

- When a model loses accuracy, it is less effective at making predictions. This can result in reduced performance and lower quality results.
- A model with reduced accuracy is more likely to misclassify data. In some applications, misclassification can have serious consequences.
- Loss of accuracy can also be a symptom of overfitting or underfitting. Overfitting occurs when the model is too complex and fits the training data too closely, leading to poor performance on new data. Underfitting occurs when the model is too simple and does not capture the underlying patterns in the data resulting in poor accuracy. When a model loses accuracy, its predictions become less reliable. This can make it difficult to trust the model's outputs and can lead to poor decision-making.

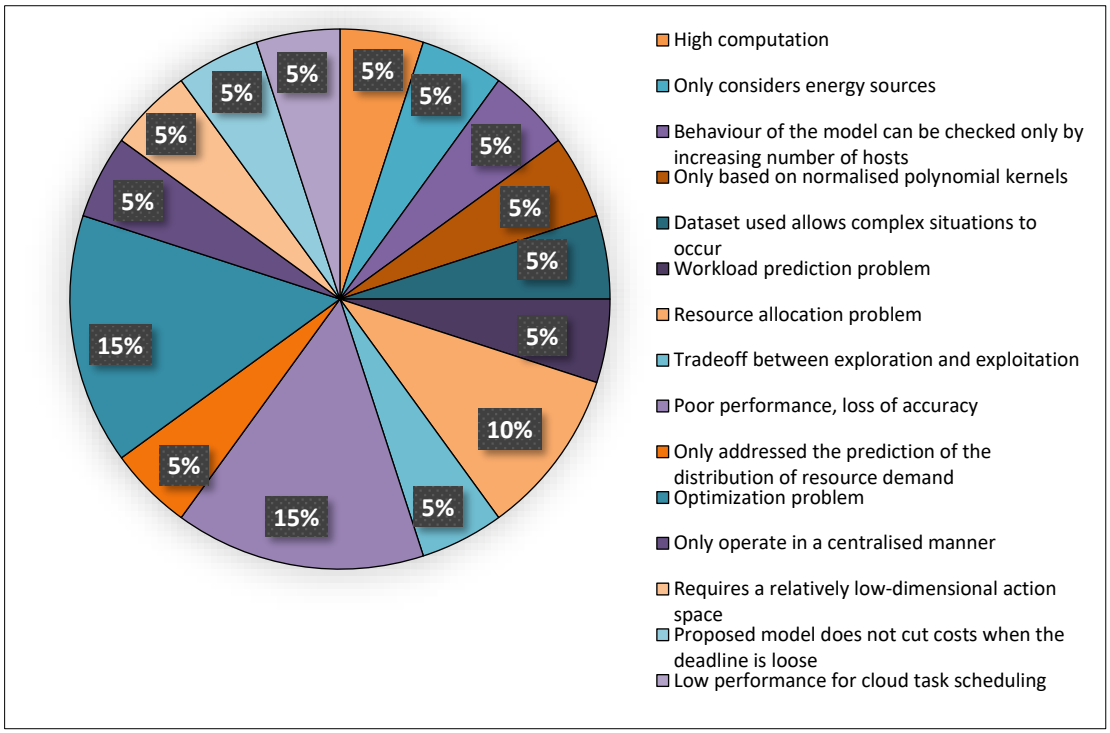

Figure 11. Most prevalent limitations with current approaches for cloud computing energy reduction

The prevalence of optimization problems at 15% of the literature as shown in Figure 11 are also major limitations in ML models for cloud computing energy minimization due to the following.

- Time Complexity: Optimization problems can be very time-consuming to solve, especially for large datasets and complex models. This can make it difficult to train models in a reasonable amount of time.
- Resource Requirements: Optimization problems require significant computational resources, such as memory and processing power. This can limit the scalability of machine-learning algorithms and make them difficult to implement on low-power devices.
- Local Minima: Optimization problems can have multiple local minima, which can make it difficult to find the global minimum. This can result in suboptimal solutions that do not fully capture the underlying patterns in the data.
- Sensitivity to Initialization: Optimization problems can be sensitive to the initialization of the model's parameters. This can lead to different solutions for different initializations, which can make it difficult to compare the performance of different models.

In addition, the resource allocation problem, which is prevalent in 10% of current literature, is also a major limitation in ML models for cloud computing energy minimization because it can significantly impact the performance and cost of cloud services. Resource allocation refers to the process of distributing computing resources, such as processing power, memory, and storage, to different applications and users in a cloud environment. If resources are not allocated efficiently, it can result in wasted resources and increased costs for cloud providers and their customers. If resources are not allocated optimally, it can result in degraded performance, which can lead to user dissatisfaction and loss of business. Resource allocation is particularly challenging in cloud environments with dynamic workloads. In such environments, the demand for resources can vary significantly over time, making it difficult to allocate resources effectively.

4.6. Most Adopted ML Tools from Literature

Figure 12 shows a summary of the most adopted ML tools for cloud energy minimization from the literature. As shown in Figure 12, 52% of the current literature has adopted the TensorFlow ML tool. TensorFlow is one of the most popular open-source machine learning frameworks available today, and it has several advantages over other reviewed ML tools hence its prevalent adoption in literature because it offers high flexibility, scalability, and seamless integration with other software libraries. It supports various tasks, is

scalable, and integrates with popular programming languages. Its active developer community provides resources, and it supports production environments for easy deployment and scaling.

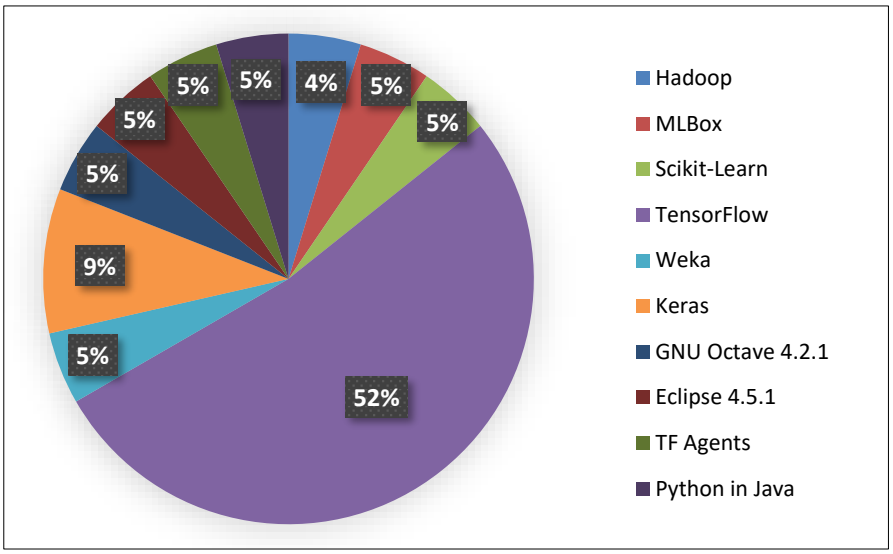

Figure 12. Most adopted ML tools from literature for cloud computing energy optimization

Even though DRL has been the most adopted in literature, from Table 1, Jararweh et al [31] achieved an energy reduction performance of 67% using Logistic Regression Model which is the best and maximum achieved from the reviewed literature showing that adoption and performance are different. The DRL algorithms balance exploration (trying out new tactics) and exploitation (leveraging current techniques), which is why many authors have adopted them. DRL can capture highly complex and dynamic cloud settings with a huge number of variables and interactions that logistic regression and other models may struggle with. DRL can adapt to the challenges of cloud environments, which include resource allocation, workload scheduling, and server provisioning. This is critical in the cloud setting, where the ideal trade-off between energy efficiency and performance must be found. DRL might assist with near-real-time decision-making. DRL can learn from historical data and previous experiences, which is useful for recognizing workload patterns and trends and making more informed decisions.

4.7. Related Works Performance

The performance metrics of related works and their values are clearly presented in Table 1 under the "ML Evaluation Metrics and Performance" column. This section will present the literature performance results and give a critical performance analysis to deduce any trends or patterns that emerge. The performance of related works was presented in the form of a percentage improvement for every metric in Table 1, which can either be a percentage increase (if the metric is maximized) or a percentage decrease (if a metric is minimized). The values of the performance percentages were calculated by critically examining the findings found in the article's results section on tables, graphs, figures, or clearly stated in the abstract, discussion, and conclusion sections. In any article reviewed, the proposed framework was compared with the baseline framework within that article for any metric, and the percentage improvement was then calculated using equation (1).

$$Perfomance = \frac{ProposedFramework_{Metric\_M\_Value} - BaselineFramework_{(Metric\_M\_Value)}}{Baseline\ Framework_{Metric\_M\_Value}} \times 100 \quad (1)$$

Performance refers to measurable outcomes in the form of a percentage to show either improvement or degradation in performance. The *Metric_M_Value* refers to the exact value of metric *M* from the results. refers to the exact value of metric *M* for the proposed framework of that article. The *refers* to the exact value of metric *M for* the baseline framework of that article. If the *performance* value is a positive number and the objective is to maximize *M*, then there is good performance. If the *performance* value is a negative number and the objective is to maximize *M*, then there is poor performance. If the *performance* value is a negative number and the objective is to minimize *M*, then there is good performance. If the *performance* value is a positive number and the objective is to minimize *M*, then there is poor performance. If *performance* in *(1)* gives a value of zero, then both the baseline and the proposed frameworks have the same performance.

Figure 13 shows the performance analysis of reviewed frameworks that considered energy consumption as a metric, and the results show that 67% has been the maximum energy reduction performance in literature from [31], and the lowest performance was from [35] at 2%. The average energy consumption performance based on Fig. 13 is 38%. Figure 14 shows the performance analysis of reviewed frameworks that considered accuracy as a metric, and the results show that 98% has been the maximum achieved accuracy performance in literature from [30], and the lowest performance was from [39] at 20%. The average accuracy achieved based on Fig. 14 is 55%. Figure 15 shows the performance analysis of reviewed frameworks that considered latency, delay, and response time as metrics, and the results show that 68.7% has been the maximum achieved value performance in literature from [24], and the lowest performance was from [27] at 10%. The average accuracy achieved based on Fig. 15 is 41%. Figure 16 shows the performance analysis of reviewed frameworks that considered SLA as a metric, and the results show that 63% has been the maximum achieved performance in literature from [3], and the lowest performance was from [32] at 8%. The average SLA achieved based on Fig. 16 is 45%. In Figure 17, the maximum resource utilization was achieved by Sui [34] and Zhang [30] at 100% performance. Nevertheless, the lease performance from reviewed articles was attained by Zhong in [25] at 2%.

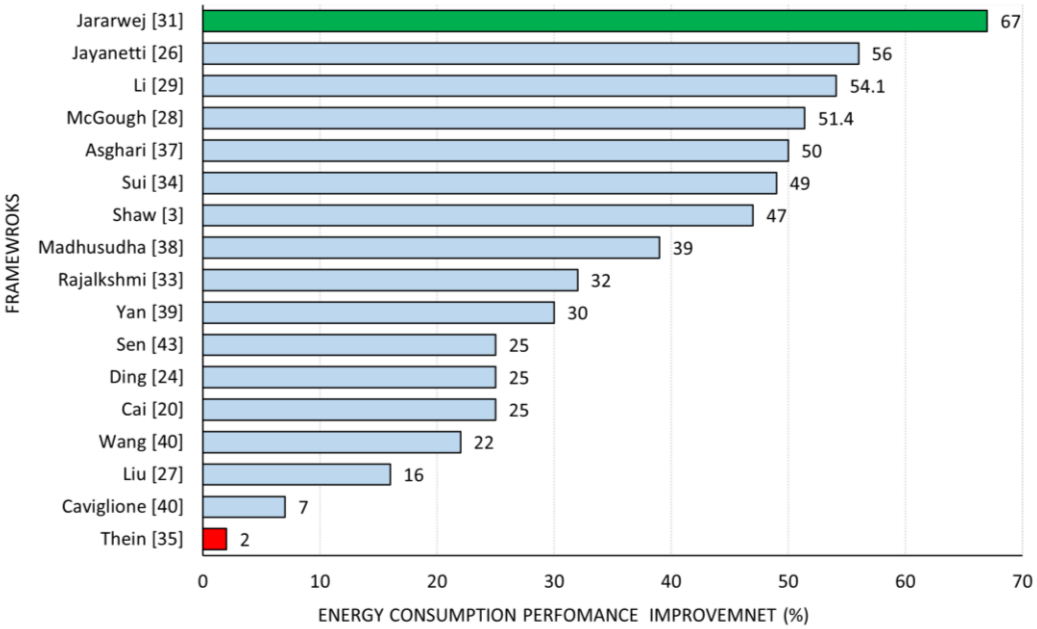

Figure 13. Performance analysis of related works based on energy consumption metric.

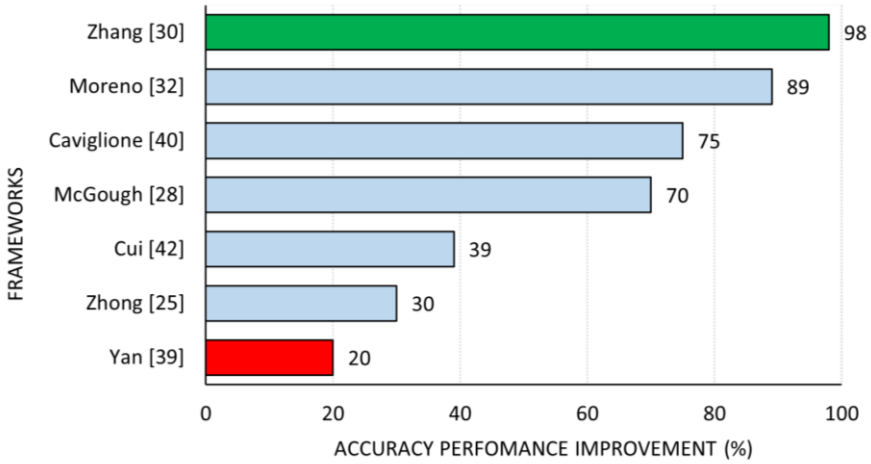

Figure 13. Performance analysis of related works based on accuracy metric

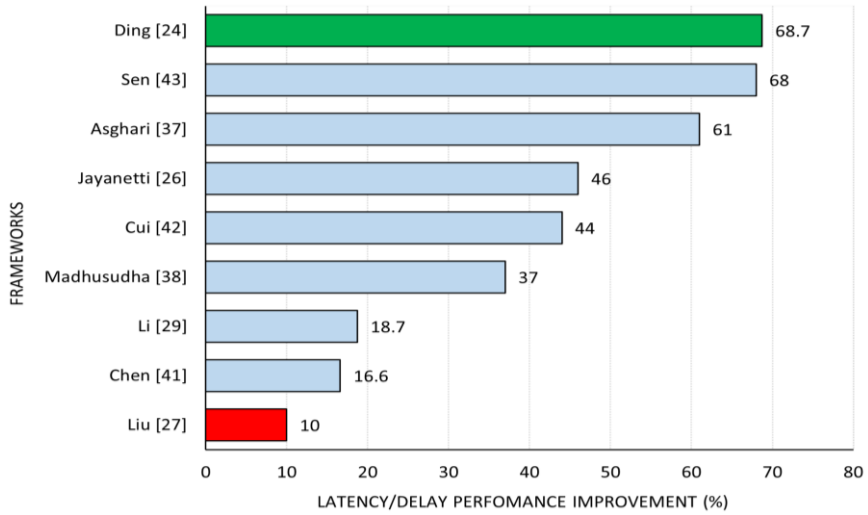

Figure 15. Performance analysis of related works based on delay/latency/response time metric.

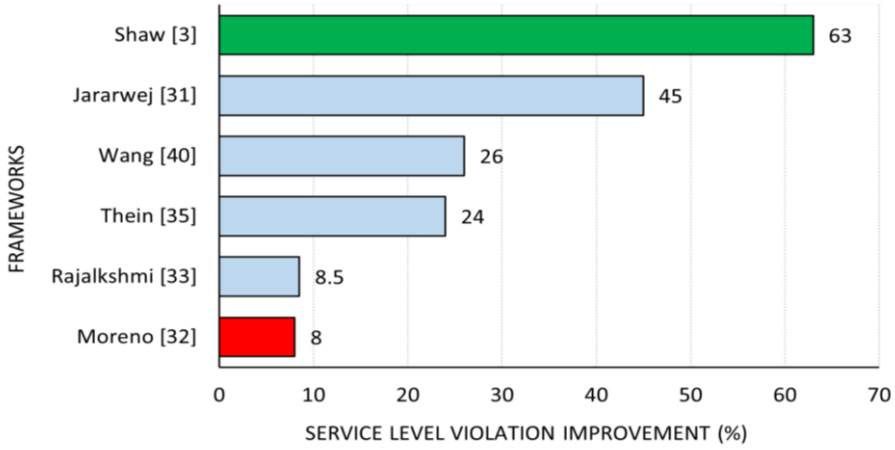

Figure 16. Performance analysis of related works based on Service Level Violation.

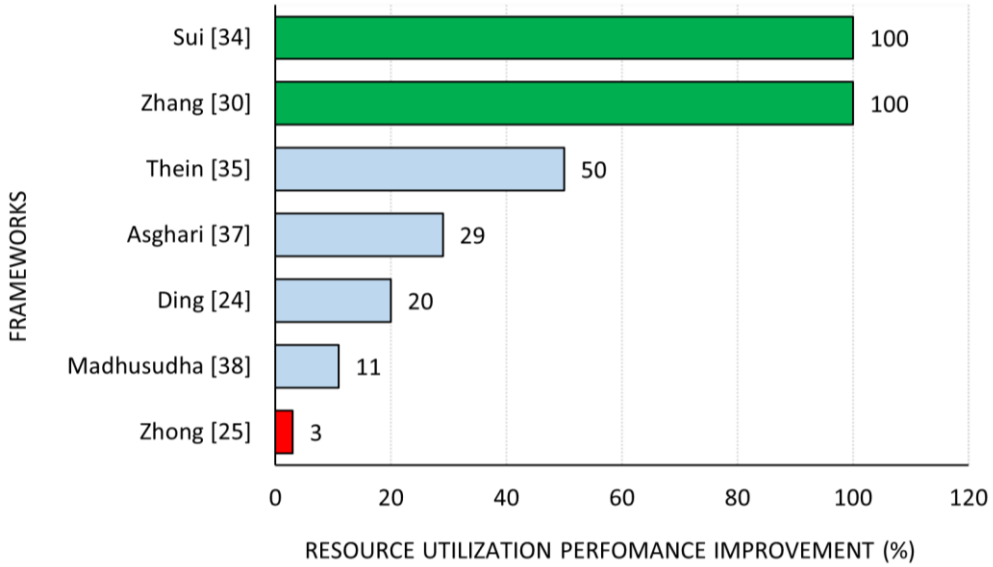

Figure 17. Performance analysis of related works based on Resource Utilization.

5. CONCLUSION

In conclusion, machine learning-based energy minimization techniques in cloud computing have shown significant promise in reducing energy consumption while maintaining performance. The results show that 67% has been the maximum energy reduction performance from literature with an average performance at 38%. These techniques leverage various algorithms, such as Deep Reinforcement Learning (DRL), Reinforcement Learning (RL), Random Forest (RF), and Support Vector Machines (SVM), to optimize resource utilization in cloud environments. The review analysis revealed that 60% of related works adopted the DRL model to solve energy efficiency problems on the cloud because DRL can adapt to changing environments and leverage large amounts of data to improve performance. The review analysis has also shown that the most dominant research objective is power or energy consumption optimization in cloud computing because high energy consumption leads to high costs and degraded performance. In addition, 40% of the literature has used the CloudSim datasets for their ML models because they are flexible, open source, and can be customized to suit specific research needs. The review also concluded that the most adopted or dominant evaluation metrics are energy consumption (28%), followed by accuracy (9%). A larger portion of current literature approaches are dominated by the limitations of poor performance, poor model accuracy, or loss of accuracy (15%), model optimization problems (15%), resource allocation problems (10%), and other limitations like high computation (5%), centralized approaches (5%) that are a single point of failure. Furthermore, 52% of the current literature has adopted the TensorFlow ML tool for cloud computing ML-based optimization models due to its flexibility, scalability, large and active community support, etc. This review therefore concludes and recommends DRL as the best ML model to optimize energy in cloud computing, CloudSim as the best tool to generate the dataset, and TensorFlow as the best ML platform for building the DRL model. The future direction of cloud computing energy optimization using machine learning algorithms is likely to focus on several key areas like exploring the combination of multiple optimization techniques to overcome the limitations of individual techniques, dynamically assigning resources in real time, and incorporating additional data sources such as weather data, occupancy patterns, and workload characteristics. Future research may focus on developing algorithms that simultaneously optimize multiple objectives, such as energy consumption, cost, and performance. Machine learning-based energy optimization may be integrated with other technologies, such as blockchain, edge computing, and IoT, to enhance the efficiency and scalability of cloud systems. As renewable energy sources become more prevalent, machine learning-based energy optimization algorithms may be developed to take advantage of these energy sources, further reducing the carbon footprint of cloud computing.

REFERENCES

[1] Nagaraju K, "Cloud Computing-An Overview & Evolution," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 3, no. 1, pp. 2456–3307, 2018.

[2] J. Surbiryala and C. Rong, "Cloud computing: History and overview," *in Proceedings - 2019 3rd IEEE International Conference on Cloud and Fog Computing Technologies and Applications*, Cloud Summit 2019, 2019. doi: 10.1109/CloudSummit47114.2019.00007.

[3] R. Shaw, E. Howley, and E. Barrett, "Applying Reinforcement Learning towards automating energy efficient virtual machine consolidation in cloud data centers," *Inf Syst*, vol. 107, p. 101722, Jul. 2022. doi: 10.1016/j.is.2021.101722.

[4] B. Wang, F. Liu, and W. Lin, "Energy-efficient VM scheduling based on deep reinforcement learning," *Future Generation Computer Systems*, vol. 125, pp. 616–628, Dec. 2021. doi: 10.1016/j.future.2021.07.023.

[5] N. Al Mudawi, N. Beloff, and M. White, "Issues and challenges: Cloud computing e-government in developing countries," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 4, 2020. doi: 10.14569/IJACSA.2020.0110402.

[6] I. Odun-Ayo, O. Ajayi, and C. Okereke, "Virtualization in cloud computing: Developments and trends," *in Proceedings - 2017 International Conference on Next Generation Computing and Information Systems*, ICNGCIS 2017, doi: 10.1109/ICNGCIS.2017.10.

[7] N. Jain and S. Choudhary, "Overview of virtualization in cloud computing," *in 2016 Symposium on Colossal Data Analysis and Networking*, CDAN 2016, 2016. doi: 10.1109/CDAN.2016.7570950.

[8] M. Abu-Alhaija, N. M. Turab, and A. R. Hamza, "Extensive study of cloud computing technologies, threats and solutions prospective," *Computer Systems Science and Engineering*, vol. 41, no. 1, pp. 225–240, 2022. doi: 10.32604/csse.2022.019547.

[9] U. Arshad, M. Aleem, G. Srivastava, and J. C.-W. 'Lin, "Utilizing power consumption and SLA violations using dynamic VM consolidation in cloud data centers," *Elsevier*, pp. 1–14, Jul. 2022.

[10] S. Supreeth and K. Patil, "VM Scheduling for Efficient Dynamically Migrated Virtual Machines (VMS-EDMVM) in Cloud Computing Environment," *KSII Transactions On Internet And Information Systems*, vol. 16, no. 6, pp. 1892–1912, Jun. 2022.

[11] J. Zeng, D. Ding, K. Kang, H. M. Xie, and Q. Yin, "Adaptive DRL-Based Virtual Machine Consolidation in Energy-Efficient Cloud Data Center," *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, no. 11, 2022. doi: 10.1109/TPDS.2022.3147851.
[12] J. Wang, H. Gu, J. Yu, Y. Song, X. He, and Y. Song, "Research on virtual machine consolidation strategy based on combined prediction and energy-aware in cloud computing platform," *Journal of Cloud Computing: Advances, Systems and Applications*, pp. 1–18, 2022.
[13] A. K. Singh, S. R. Swain, D. Saxena, and C.-N. 'Lee, "A Bio-Inspired Virtual Machine Placement Toward Sustainable Cloud Resource Management," *IEEE Systems Journal (Early Access)*, pp. 1–12, Mar. 2023.
[14] M. H. Sayadnavard, A. Toroghi Haghighat, and A. M. Rahmani, "A multi-objective approach for energy-efficient and reliable dynamic VM consolidation in cloud data centers," *Engineering Science and Technology, an International Journal*, vol. 26, 2022. doi: 10.1016/j.jestch.2021.04.014.
[15] M. 'Imran, M. 'Ibrahim, M. S. U. 'Din, M. A. U. 'Rehman, and B. S. 'Kim, "Live virtual machine migration: A survey, research challenges, and future directions," *Elsevier*, pp. 1–18, Aug. 2022.
[16] S. Talwani, K. Alhazmi, J. Singla, H. J. Alyamani, and A. K. Bashir, "Allocation and migration of virtual machines using machine learning," *Computers, Materials and Continua*, vol. 70, no. 2, 2022. doi: 10.32604/cmc.2022.020473.
[17] C. H. Tran, T. K. Bui, and T. V. Pham, "Virtual machine migration policy for multi-tier application in cloud computing based on Q-learning algorithm," *Computing*, vol. 104, no. 6, 2022. doi: 10.1007/s00607-021-01047-0.
[18] H. Li, J. Liu, and Q. Zhou, "Research on energy-saving virtual machine migration algorithm for green data center," *IET Control Theory and Applications*, 2022. doi: 10.1049/cth2.12401.
[19] B. B. Naik, D. Singh, and A. B. Samaddar, "Multi-objective Virtual Machine Selection in Cloud Data Centers Using Optimized Scheduling," *Wirel Pers Commun*, vol. 116, no. 3, 2021. doi: 10.1007/s11277-020-07807-z.
[20] X.-B. Cai, Y.-X. Ji, and K. Han, "Energy Efficiency Optimizing Based on Characteristics of Machine Learning in Cloud Computing," *ITM Web of Conferences*, vol. 12, p. 03047, Sep. 2017. doi: 10.1051/itmconf/20171203047.
[21] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets*, vol. 31, no. 3, 2021. doi: 10.1007/s12525-021-00475-2.
[22] J. Schmitt, J. Bönig, T. Borggräfe, G. Beitinger, and J. Deuse, "Predictive model-based quality inspection using Machine Learning and Edge Cloud Computing," *Advanced Engineering Informatics*, vol. 45, 2020. doi: 10.1016/j.aei.2020.101101.
[23] A. Trisal and D. Mandloi, "Machine Learning: An Overview," *International Journal of Research - GRANTHAALAYAH*, vol. 9, no. 7, 2021. doi: 10.29121/granthaalayah.v9.i7.2021.4120.
[24] D. Ding, X. Fan, Y. Zhao, K. Kang, Q. Yin, and J. Zeng, "Q-learning based dynamic task scheduling for energy-efficient cloud computing," *Future Generation Computer Systems*, vol. 108, pp. 361–371, Jul. 2020. doi: 10.1016/j.future.2020.02.018.
[25] W. Zhong, Y. Zhuang, J. Sun, and J. Gu, "The cloud computing load forecasting algorithm based on wavelet support vector machine," *in Proceedings of the Australasian Computer Science Week Multiconference*, New York, NY, USA: ACM, Jan. 2017, pp. 1–5. doi: 10.1145/3014812.3014852.
[26] A. Jayanetti, S. Halgamuge, and R. Buyya, "Deep reinforcement learning for energy and time optimized scheduling of precedence-constrained tasks in edge–cloud computing environments," *Future Generation Computer Systems*, vol. 137, pp. 14–30, Dec. 2022. doi: 10.1016/j.future.2022.06.012.
[27] N. Liu et al., "A Hierarchical Framework of Cloud Resource Allocation and Power Management Using Deep Reinforcement Learning," *in 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, IEEE, 2Jun. 2017, pp. 372–382. doi: 10.1109/ICDCS.2017.123.
[28] A. S. McGough, M. Forshaw, J. Brennan, N. Al Moubayed, and S. Bonner, "Using Machine Learning to reduce the energy wasted in Volunteer Computing Environments," Oct. 2018, [Online]. Available: http://arxiv.org/abs/1810.08675
[29] H. Li, R. Cai, N. Liu, X. Lin, and Y. Wang, "Deep reinforcement learning: Algorithm, applications, and ultra-low-power implementation," *Nano Commun Netw*, vol. 16, pp. 81–90, Jun. 2018. doi: 10.1016/j.nancom.2018.02.003.
[30] J. Zhang, N. Xie, X. Zhang, K. Yue, W. Li, and D. Kumar, "Machine learning based resource allocation of cloud computing in auction*," Computers, Materials and Continua*, vol. 56, no. 1, 2018. doi: 10.3970/cmc.2018.03728.
[31] Y. Jararweh, M. B. Issa, M. Daraghmeh, M. Al-Ayyoub, and M. A. Alsmirat, "Energy efficient dynamic resource management in cloud computing based on logistic regression model and median absolute deviation," *Sustainable Computing: Informatics and Systems*, vol. 19, pp. 262–274, Sep. 2018. doi: 10.1016/j.suscom.2018.07.005.
[32] R. Moreno-Vozmediano, R. S. Montero, E. Huedo, and I. M. Llorente, "Efficient resource provisioning for elastic Cloud services based on machine learning techniques," *Journal of Cloud Computing*, vol. 8, no. 1, 2019. doi: 10.1186/s13677-019-0128-9.
[33] N. R. Rajalakshmi, G. Arulkumaran, and J. Santhosh, "Virtual machine consolidation for performance and energy efficient cloud data center using reinforcement learning," *Int J Eng Adv Technol*, vol. 8, no. 3 Special Issue, 2019.
[34] X. Sui, D. Liu, L. Li, H. Wang, and H. Yang, "Virtual machine scheduling strategy based on machine learning algorithms for load balancing," *EURASIP J Wirel Commun Netw*, vol. 2019, no. 1, p. 160, Dec. 2019. doi: 10.1186/s13638-019-1454-9.
[35] T. Thein, M. M. Myo, S. Parvin, and A. Gawanmeh, "Reinforcement learning based methodology for energy-efficient resource allocation in cloud data centers," *Journal of King Saud University - Computer and Information Sciences*, vol. 32, no. 10, 2020. doi: 10.1016/j.jksuci.2018.11.005.
[36] Z. Tong, H. Chen, X. Deng, K. Li, and K. Li, "A scheduling scheme in the cloud computing environment using deep Q-learning," *Inf Sci (N Y)*, vol. 512, pp. 1170–1191, Feb. 2020. doi: 10.1016/j.ins.2019.10.035.

[37] A. Asghari, M. K. Sohrabi, and F. Yaghmaee, "A cloud resource management framework for multiple online scientific workflows using cooperative reinforcement learning agents," *Computer Networks*, vol. 179, p. 107340, Oct. 2020. doi: 10.1016/j.comnet.2020.107340.
[38] M. H. Madhududhan, S. Kumar T, S. M. F. D. S. Mustapha, P. Gupta, and R. P. Tripathi, "Hybrid Approach for Resource Allocation in Cloud Infrastructure Using Random Forest and Genetic Algorithm," *Sci Program*, vol. 2021, pp. 1–10, Oct. 2021. doi: 10.1155/2021/4924708.
[39] J. Yan, J. Xiao, and X. Hong, "Dueling-DDQN Based Virtual Machine Placement Algorithm for Cloud Computing Systems," *in 2021 IEEE/CIC International Conference on Communications in China (ICCC)*, IEEE, Jul. 2021, pp. 294–299. doi: 10.1109/ICCC52777.2021.9580393.
[40] L. Caviglione, M. Gaggero, M. Paolucci, and R. Ronco, "Deep reinforcement learning for multi-objective placement of virtual machines in cloud datacenters," *Soft comput*, vol. 25, no. 19, pp. 12569–12588, Oct. 2021. doi: 10.1007/s00500-020-05462-x.
[41] G. Chen, J. Qi, Y. Sun, X. Hu, Z. Dong, and Y. Sun, "A collaborative scheduling method for cloud computing heterogeneous workflows based on deep reinforcement learning," *Future Generation Computer Systems*, vol. 141, pp. 284–297, Apr. 2023. doi: 10.1016/J.FUTURE.2022.11.032.
[42] T. Cui, R. Yang, C. Fang, and S. Yu, "Deep Reinforcement Learning-Based Resource Allocation for Content Distribution in IoT-Edge-Cloud Computing Environments," *Symmetry (Basel)*, vol. 15, no. 1, 2023. doi: 10.3390/sym15010217.
[43] T. Sen and H. Shen, "Machine learning based timeliness-guaranteed and energy-efficient task assignment in Edge Computing Systems," *in Proceedings - IEEE 3rd International Conference on Fog and Edge Computing (ICFEC)*, 2019, doi:10.1109/cfec.2019.8733153.
[44] Y. Kumar, S. Kaul, and Y. Hu, "Machine Learning for Energy-resource allocation, workflow scheduling and live migration in cloud computing: State-of-the-art survey," *Sustainable Computing: Informatics and Systems*, vol. 36, pp. 100780, 2022. doi:10.1016/j.suscom.2022.100780
[45] M. Demirci, "A survey of machine learning applications for energy-efficient resource management in Cloud Computing Environments," *in Proceedings 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA*), 2015. doi:10.1109/icmla.2015.205
[46] T. Khan, W. Tian, G. Zhou, S. Ilager, M. Gong, and R. Buyya, "Machine learning (ML)-centric resource management in cloud computing: A review and future directions," *Journal of Network and Computer Applications*, vol. 204, pp. 103405, Aug. 2022. doi: 10.1016/J.JNCA.2022.103405.
[47] D. Soni and N. Kumar, "Machine learning techniques in emerging cloud computing integrated paradigms: A survey and taxonomy," *Journal of Network and Computer Applications*, vol. 205, pp. 103419, Sep. 2022. doi: 10.1016/J.JNCA.2022.103419.

BIOGRAPHY OF AUTHORS

Nomsa Puso is currently an MSc student in the department of computer science and information systems at Botswana International University of Science And Technology (BIUST), Botswana. Her MSc thesis is in the area of energy efficiency in cloud computing, adopting machine learning technology, specifically Deep Reinforcement Learning. She is currently involved extensive research and writing several publications at BIUST.

Dr. Tshiamo Sigwele is currently a lecturer in the Department of Computer Science and Information Systems at Botswana International University of Science and Technology (BIUST) with research interests in cloud computing, machine learning, and wireless communication. Dr. Sigwele graduated in 2017 with a Ph.D. in cloud computing and telecommunications from the University of Bradford, UK. He has over 20 internationally recognized publications. He worked as a researcher from 2017 to 2018 in a British Council-funded project, BLESS U: Bandar Lampung Enhanced Smart Health Services with Smart Ubiquity, with a grant total of €89,937 and published several high-quality publications. He is currently supervising PhD and MSc students in the areas of cloud computing and machine learning. He is involved in several research projects at BIUST.

Dr. Oba Zubair Mustapha has received his PhD from the University of Bradford UK. In 1997 and 2011, respectively, he earned his electrical engineering bachelor's degree and master's degree from the University of Ilorin in Nigeria. From 2000 to 2003, he was a System Engineer at Cyberaccess & Communication Ltd. in Nigeria. He subsequently began working as a lecturer in the department of electrical and electronic engineering at Kwara State Polytechnic, Institute of Technology, Nigeria, in 2007. His areas of interest include resource management, artificial intelligence, wireless mobile networks, control systems, and telecommunications. He is a member of IEEE, IAENG, and COREN.