2021 National Computing Colleges Conference (NCCC) | 978-1-7281-6719-0/20/$31.00 ©2021 IEEE | DOI: 10.1109/NCCC49330.2021.9428825

Machine Learning Approaches for Load Balancing in Cloud Computing Services

Dalia Abdulkareem Shafiq
School of Computer Science & Engineering (SCE)
Taylor's University
Subang Jaya, Malaysia
daliakareem7@gmail.com

NZ Jhanjhi
School of Computer Science & Engineering (SCE)
Taylor's University
Subang Jaya, Malaysia
noorzaman.jhanjhi@taylors.edu.my

Azween Abdullah
School of Computer Science & Engineering (SCE)
Taylor's University
Subang Jaya, Malaysia
Azween.Abdullah@taylors.edu.my

***Abstract*—As the demand for cloud services increases, optimization of resources becomes essential. Static algorithms are no longer sufficient to solve cloud-related challenges such as imbalanced workload distribution in Virtual Machines or improper resource allocation to cloud users. Thus, the need to explore other rich approaches can greatly improve cloud applications' performance and tackle the above challenges. This research investigates the latest Machine Learning approaches that can tackle the above challenges in cloud environment. A comparison of these approaches included highlighting their strengths and weaknesses to induce a research gap useful for upcoming researchers in the field.**

Keywords—Cloud Computing; Machine Learning; Regression; Classification; Virtualization; Optimization.

I. Introduction

Cloud Computing has been around since 1997 and the demand for cloud services is in the rise. The technology offers wide range of online services through multiple delivery models namely: Platform as Service (PaaS), Software as Service (SaaS) and Infrastructure as service (IaaS). Cloud users and organizations [1] mostly benefit from SaaS model whereby it can eliminate hardware cost by providing accessible software online such as Google Drive, Gmail, YouTube etc. This is one of the simplest delivery models in the cloud as it provides flexible services with less cost and physical storage. However, the increase in the services' everyday usage and handling the large data sets behind it can cause some challenges to the cloud services' backend.

IaaS model handles the backend (server-end) of the cloud services. Cloud computing relies heavily on virtualization to create an abstract layer between software and hardware [2][3]. In a typical cloud environment, users submit their requests, and these are converted into Virtual Machines. Cloud service providers in this model are responsible for ensuring efficient allocation of resources [4] to clients with optimal cloud services performance. Performance has been stated among the top three challenge in cloud computing [5], hence it becomes important to research efficient methods to improve the utilization of cloud resources and increase the satisfaction of users.

There has been extensive research on many objective-oriented algorithms to address optimization and resource allocation challenges in the cloud environment. However, the existing Load Balancing algorithms are either static where fewer parameters are considered or dynamic, where performance can easily degrade due to sudden failures in the nodes [6]. Thus, such approaches may not be suitable for use in in such environments where the load is constantly changing [7]. Therefore, it becomes important to discover other efficient and intelligent approaches to tackle cloud-related challenges.

Machine learning is a subcategory of artificial intelligence that has been an active topic in IT and intelligent systems. Data is stored in large quantities in the cloud machines, and it can be trained to make precise predictions and evaluations based on analysis to perform tasks more efficiently. It is the fastest-growing field nowadays. According to a survey done by RightScale in 2019, recent research shows that Machine Learning plays a vital role in Cloud computing. It represents a figure of 786 professionals (48% of respondents) among different expertise are considering using Machine learning services in the future [8]. Machine learning is being offered as a cloud service. Thus, the combination of such technologies can be established as architectures in the future to cover different layers from business workflow to software. Machine Learning approaches can be divided into two main groups [9]:

- **Supervised Learning**: this type of learning provides a precise classification of a labelled data sample with a defined output. This means the algorithm has a specific outcome that can be predicted from a set of independent variables. Examples of algorithms include Linear Regression, Support Vector Machine, Neural Networks and Naïve Bayes classifiers.
- **Unsupervised Learning**: unlike supervised learning, this type of learning does not provide a clear pattern in the dataset; the data sample are unlabelled. Hence a model is trained to have minimum errors when learning the categorization of such information. Such training can be used mainly to solve clustering scenarios such as classification—for example, K-Means clustering and Fuzzy Clustering algorithms.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

The rich techniques that Machine learning offers can highly enhance the capabilities of cloud services. As the cloud handles and stores numerous amounts of data in the cloud, it can provide intelligent predictions and decisions-based solutions [10] to utilize cloud resources and perform huge tasks efficiently. This survey aims to introduce recent Machine Learning approaches leveraged by researchers in the cloud computing field.

A. *Challenges in Cloud Computing*

Despite the many benefits cloud services offers to businesses and end users, cloud providers still few challenges face. Before we review the current Machine Learning approaches, we should identify some challenges in the Cloud Computing domain. These challenges are summarized as below:

1) Task Scheduling: this process is responsible for assigning tasks with certain constraints. When users submit many requests from separate locations, the assignment of tasks may be inefficient. This can result in low user satisfaction as user requests may not be prioritized correctly [11]. For a better response time, high priority jobs should be given a chance to run first.

2) Load Balancing: The cloud deals with many storage and retrieval operations at faster rate. Therefore, it becomes essential to provide a proper workload distribution method. There are two main goals in load balancing as listed and explained below:

a) *Resource allocation:* The demand for cloud services is increasing; if there is no fairness in the distribution of workload among clients (end users) and cloud servers, it may cause inefficiency and unavailability of resources.

b) *Server resource optimization:* if resources are not appropriately utilized, this could lead to low Data Centres' optimization. Dealing with load balancing should avoid overloading situations, and in turn, energy consumption should be less. Additionally, energy consumption has been a challenge in Vehicular Cloud Computing as well as it becomes harder to reduce for VMs on overloaded or underloaded hosts [12].

B. *Survey Plan & Organization*

This survey has been done to include several up-to-date articles in the field of Load Balancing in Cloud Computing. The plan of this survey is summarized in the following points:

- **Selection of the reviewed research articles:** articles included in the survey are selected based on the use of Machine Learning techniques in Cloud Computing environment to address the challenges listed in the previous section. The articles are acquired from several reputed sources such as IEEEXplore, ScienceDirect, ResearchGate and so on. Filtration is carried out first based on the appropriateness of article title to article, followed by the abstract. Lastly, the article is selected after analysis of the quality of the content presented, the objectives and challenges addressed using Machine Learning is carefully done.
- **Presentation of the survey paper:** the paper is organized and presented to make it easier for readers to understand how to leverage Machine Learning techniques to solve cloud-related challenges. The articles are surveyed as the following: Initially, the approach's objective is identified, for example, the proposed method is to address Load Balancing issues in cloud environments. Then, we identify the proposed Machine Learning technique used to tackle the challenge along with the parameters. Finally, the outcome of the approach is explained along with graphical representations where applicable.
- **Analysis of reviewed approaches:** then the reviewed approaches are summarized to include the strength and weakness. The approaches are also examined based on the strategy used in Machine learning and their types (supervised or unsupervised), and finally the problem area that is targeted by such approaches in cloud environments. This will benefit the readers to choose an appropriate approach to address cloud-related challenges.
- **Organization of the paper:** The rest of the paper is organized as follows. Section II presents the literature review where multiple Machine Learning approaches are introduced to solve cloud-related challenges such as Load Balancing. Section III provides a discussion of this review's findings, the comparison of the reviewed approaches and the research gap. Finally, Section IV concludes the review and suggests points for future research.

II. Literature Review

This section introduces a review of recent literature associated with machine learning and cloud computing.

Current research focuses on two main aspects of Cloud Computing: Load Balancing and Task Scheduling. Researchers have proposed numerous recent approaches as a hybrid approach to enhance cloud-based applications' performance considering these two aspects. In [13], researchers have proposed a hybrid approach of Task Scheduling and Load Balancing based on an improved weighted RR algorithm, as shown in figure 1 below. The algorithm considers both static and dynamic scheduling in order to complete longer tasks with higher priority. The scheduler logically finds the most fitting VM to allocate the tasks to it using a proposed algorithm, as seen in the figure below. Whereas the load balancer makes migration decisions from heavy VM to light VM. The cloud resource manager communicates with all the VMs resources to collect information such as VM capabilities, current load and the number of tasks in the execution/waiting queue.

In [14], authors have developed a scheduling algorithm to achieve better load balancing and high QoS by decreasing the Response Time and Makespan in comparison to MaxMin, Shortest Job First and Round Robin. The algorithm is able to decide on the distribution of tasks to VMs based on the state of the VM and task timing. While both above approaches consider each VM's capabilities and task length and provide efficient utilization of resources through task migration process, however, they still lack in an intelligent approach to prioritize user tasks. As the cloud environment serves multiple users simultaneously, prioritization of important requests should be considered to reduce the waiting time to efficiently serve the cloud users.

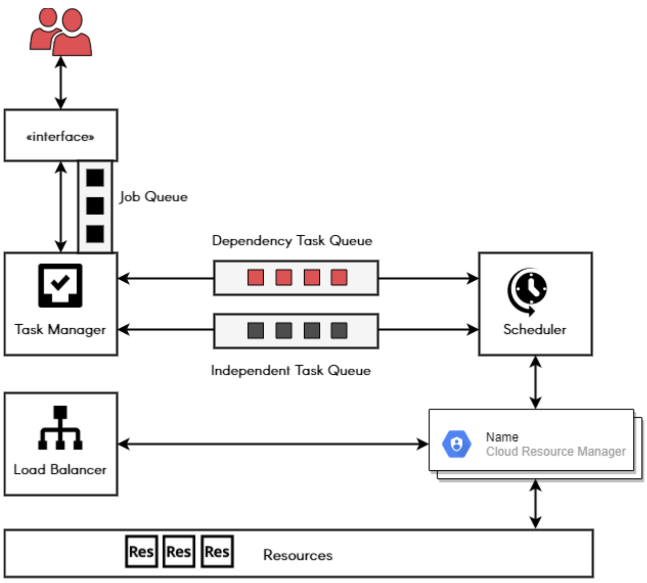

Fig. 1. Revised Load Balancing & Task Scheduling Model Adopted From [13].

Multiple approaches utilized machine learning techniques in to further optimize the cloud resources for optimal performance as introduced in this literature. Authors in [15] applies classification for Virtual Machines and users' tasks. Dataset is generated from various log files as it provides a clear understanding of online users' behaviour. Such logs offer accurate information for task parameters such as size. The Machine Learning approach classifies tasks in three groups based on their sizes. The resource requirement is then calculated. Given that if task is light it will require less resources this will results in efficient usage of resources. Virtual Machines are also classified into five groups based on their calculated CPU and RAM utilization. The tasks are mapped to VMs in the pattern that light tasks run on heavy VM and so on. The approach still lacks in addressing load balancing challenges. Tasks might still be allocated to VM when it is in overloaded state and hence might increase latency since task age, arrival time are not considered.

In classification Machine Learning approach, authors in [16] proposed a new C-Rule Algorithm to make predictions of the maximum workload that can be handled in the cloud. This approach is useful to Cloud Service Providers to dynamically assign workload and prevent overloading situations in the network. Based on the predictions obtained from Cicada tool, the proposed algorithm is applied for better resource allocation. The traffic data is gathered from SFlow-enabled devices and exported to a simulation framework (CloudSim). The algorithm is used to decide whether there is an overloaded workload scenario by comparing the predictions gathered against historical data. The algorithm classifies the prediction in two categories: reliable and unreliable to look for least and fast computational of workload balancing. The approach manages to reduce waiting time significantly as the hosts increases.

In [17] authors proposed a hybrid approach in classification using swarm intelligence to enhance the scheduling process in the cloud environment. The proposed MLCCSI) model utilizes classifier chains method to select an order on the labelled data set. The training for each label is done. The approach selects the best scheduling technique using multiple rules. Results shows that the approach takes less execution time and Makespan of 7% and 75%.

Another classification approach is presented in [18] to sort the user tasks by considering the task parameter: deadline and utilizing the priority queue to determine the learning algorithm. Sometimes, in Machine Learning, finding a labelled dataset becomes a challenge. Several online sources provide public data for web logs such as "Workers Browser Activity in CrowdFlower Tasks" provided by Kaggle or "Logs of Real Parallel Workloads from Production Systems". However, sometimes this information might not be sufficient to identify user tasks such as deadline, arrival time etc. or VMs information such as utilization %. In this approach, the authors assign values randomly to tasks for the implementation of the data set. The approach applied multiple classifier models (such as linear SVM). Experiment results reveal that Boosted Tree is among the best and accurate classifiers of 94.3% rate for 400 tasks. However, the approach mainly address task scheduling and hence balancing challenge is still not solved.

Another Machine Learning classification approach to select the best scheduling algorithm is proposed in [19]. Among five scheduling algorithms, the most suitable one will be selected using the trained classifier. After allocation process, VM and Cloudlet (task) attributes are extracted and stored in a text file. This is taken as a dataset which is used in the training process. The training model explains the selection of the algorithm in figure 2 below. The input to Machine Learning algorithm is attributes of tasks and VMs. The approach uses types of classification algorithm such as: tree based, rule based, Bayes and lazy classification to produce classification rules.

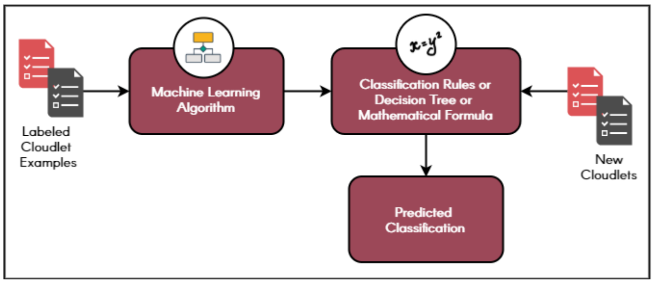
Labeled Cloudlet Examples
Machine Learning Algorithm
$x=y^2$
Classification Rules or Decision Tree or Mathematical Formula
New Cloudlets
Predicted Classification

Fig. 2. Revised Machine Learning Training Model Adopted From [19].

To solve task scheduling challenge in cloud computing, users' requests should be prioritized considering various task parameters. In this research [20] authors have tackled the problem of priority by using classification approach. It considers three main task parameters: deadline (specified time to execute tasks); task length (workload) and finally task age. It utilizes queues concept to allow the task with the highest priority to first execute. Doing so can minimize the waiting time however, it does not address overloading situations of Virtual Machines and hence it does not address load balancing and migration challenges. Additionally, the approach doesn't specify which type of classifier is used for priority.

Researchers in [21] presents a prediction approach to determine the availability of a Virtual Machine. It assigns static weight to each parameter of the tasks such as completion time, type of task (requires high bandwidth or not), latency and the urgency of tasks (high, medium, or low priority). These values are computed to assign priority to tasks. Besides that, it classifies VMs based on bandwidth and CPU (Million Instruction Per Second). VM that result in low execution time. Based on the minimum execution time, the task is assigned to the suitable VM; if the VM is not found, then the task completion time is predicted to estimate the availability of the VM. The approach in [21] has been further extended to include Global and Local Queues by researchers in [22]. A Global queue is used to receive the new requests and categorize the jobs based on the computational (such as processing speed) and communication (such as MIPS). After that, jobs are moved to the local queue, and the static task scheduling algorithm, Min-Min, is applied to select the top priority task. Results show that the method achieves a resource utilization of 7.73% in contrast to the traditional min-min algorithm. Another supervised Machine Learning technique is proposed in [23], whereby regression is used. Using regression will give better analytic data to make an efficient load balancing mechanism. It sets an upper and lower threshold to put the VMs in the queues based on CPU utilization. This can work in the cloud environment where dynamic load occurs.

Authors apply Linear Regression in [24] to minimize the amount of workload migration. The approach calculates threshold value to make sure each Virtual Machine has a minimum amount of workload. First, the data is collected and pre-processed to obtain a log file containing number of servers, virtual machines, and number of jobs. Using regression, with minimum error the accuracy of the proposed algorithm is obtained. The threshold is calculated using this algorithm, then it is compared to the server load, if it's greater then, it'll not update the load else, it goes to the coordinator. This is responsible to adjust the workload.

Clustering is another Machine Learning approach that falls in the unsupervised category. Researchers in [25] presented a comprehensive approach addressing both concept of Task Scheduling and Load Balancing. It makes use of clustering algorithms such as Mean Shift (MSC) and Dominant Sequence (DSC). In scheduling, it address the task priority challenges by considering two main parameters: Makespan and Deadline which is an important parameter in Service Level Agreement document [26]. Based on priority, the DSC algorithm clusters tasks. Another algorithm is used to rank the tasks after the clustering process called as MHEFT, selecting the high-ranking priority task. Besides that, Virtual machines are also put in clusters based on MSC algorithm's outcome utilizing the kernel functions. To address Load Balancing, a load is distributed by considering the weight of the server, the connections and finally the capacity. For this purpose, WLC algorithm is applied, however, in this algorithm the number of jobs is used to predict the server load. This may cause inaccurate results due to high variations of load in task allocation process. Additionally, once the server weight of server is determined it may be difficult to modify it instantaneously [27].

Researchers in [28] provide another Clustering approach. The proposed MLBL method uses both Support Vector Machine (SVM) and K-method for clustering of Virtual Machines for dynamic resource mapping. SVM is used put the jobs in three clusters: minimum, average, and maximum time (resources used). Then K-Mean is used to classify VMs based on CPU and RAM. The K value is calculated using the total number of data centers or available hosts. After this, mapping of jobs depends on the availability of the VM. Experiment results shows the approach improves on QoS and reduce task rejections.

III. Discussion

In this section, a thorough investigation of the algorithms is discussed whereby it includes tables and figures to summarize the reviewed algorithms, the available implementation tools based on review, and finally, suggestions for future research.

A. *Summary of the reviewed algorithms*

After we discussed the approaches in the literature review section, now the analysis of these Machine learning approaches in presented in this section. The comparative analysis is compiled in table I as seen below. The table represents the references, the authors' names and article publication year, advantages, the disadvantages of the proposed approach, and finally the evaluation tools used by researchers for ML and simulation.

TABLE I. Comparative Summary of Literature review

S/N	Ref.	Authors & Year	Approach	Strengths	Weaknesses	Evaluation Tool
1.	[15]	(Elrotub & Gherbib, 2018)	Classification ML technique of tasks based on information from log files and VM based on CPU & RAM calculated value.	Utilizes the log files from users as dataset; obtains optimal resource utilization.	Task scheduling issues such as task priority is not addressed; May cause latency as jobs can still be assigned to VM with heavy workload; the approach's performance is not evaluated.	Weka for training; Discussion Scenario for evaluation
2.	[16]	(Jodayreea et al, 2019)	A Predictive approach to dynamic resource distribution of cloud services	Better allocation of resources; use SFlow-enabled devices for the prediction of incoming load; reduce waiting time; require less computational power.	Not Available.	Cicada for load prediction; CloudSim for simulation.
3.	[17]	(Rjoub & Bentahar, 2017)	Multi Label Classifier Chains Swarm Intelligence (MLCCSI)	Minimize the execution time & Makespan; better task scheduling.	Does not consider other VM metrics such as CPU capacity, Bandwidth and RAM.	CloudSim
4.	[18]	(Er-raji & Benabbou et al, 2018)	Supervised classification ML to assign priority to tasks using different classifiers models.	Better scheduling and less Response Time; Finds the best classifier model with high accuracy rate.	Does not address challenges related to Load Balancing and task migration; the approach's performance is not evaluated.	Classification Learner APP for training; No evaluation

S/N	Ref.	Authors & Year	Approach	Strengths	Weaknesses	Evaluation Tool
5.	[19]	(Tikar & Jaybhaye, 2015)	ML trained classifier to select the best scheduling algorithm to execute tasks	Better utilization of resources and low Response Time.	Does not address load balancing challenges.	Weka for classification; CloudSim for evaluation
6.	[20]	(Er-raji & Benabbou, 2017)	Priority task scheduling using queues for task classification based on parameters e.g., deadline, age, and length.	Less waiting time required for tasks with low priority.	Does not address load balancing challenges	No Evaluation
7.	[21]	(Khurana & Singh, 2018)	Prediction and categorization of VM's availability to enhance scheduling method	High performance as less Makespan & better resource utilization is achieved.	May cause high no. of task rejections; Non-preemptive scheduling; high failure rate.	Cybershake & Montage Workflow
8.	[22]	(Miglani & Sharma, 2018)	Categorization of tasks in VMs using Global and Local Queues for improved load balancing	Avoids overloading of resources (balanced utilization)	Uses non-preemptive scheduling method.	CloudSim
9.	[23]	(Panchal & Parida, 2018)	Supervised Regression ML for task transfer from overloaded VM to less loaded VM	Efficient method for load balancing.	Not Available.	AWS EC2
10.	[24]	(Padmavathi, 2020)	MLB algorithm using linear regression approach to avoid VM migration	Minimize the no. of workload migration and job rejections.	Aggressive consolidation may cause the performance to degrade.	CloudSim
11.	[25]	(Al-Rahayfeh & Atiewi, 2019)	Clustering algorithms to cluster tasks based on priority (deadline & makespan) and VM into 3 clusters (overloaded, underloaded & idle).	Addresses task priority; lower Makespan & high throughput.	Shortcomings in load balancing method since of WLC is applied.	CloudSim
12.	[28]	(Lilhore et al, 2020)	MLBL approach using SVM classification & K-method clustering techniques to enhance scheduling of tasks.	Enhanced the quality of services; better resource utilization & less task rejection.	Other VM parameter such as bandwidth is not considered for clustering.	CloudSim

Table II below provides a review of the presented literature that addressed the cloud computing challenges such as load balancing and task scheduling or hybrid using Machine Learning techniques. The authors have classified them based on the underlying ML technique/algorithm used. This classification is carried out based on the information from tables I and it is useful to identify different approaches to tackle such challenges in this domain. As we can see from the table, most of the approaches utilizes classification Machine Learning techniques, which has gained high attention in the recent research field. However, few approaches consider hybrid approaches which can be addressed in the future research to improve the overall performance of cloud-based applications.

TABLE II. RECENT APPROACHES BASED ON THEIR STRATEGY AND THEIR CORRESPONDING PROBLEM AREA

S/N	Approach	Strategy Applied				Problems solved		
		Classification	*Support Vector Machine*	*Regression*	*Clustering*	*Task Scheduling*	*Load Balancing*	*Hybrid (TS & LB)*
1.	(Elrotub & Gherbib, 2018)	✓	×	×	×	×	×	✓
2.	(Jodayreea et al, 2019)	✓	×	×	×	×	✓	×
3.	(Rjoub & Bentahar, 2017)	✓	×	×	×	✓	×	×
4.	(Er-raji & Benabbou et al, 2018)	✓	✓	×	×	✓	×	×
5.	(Tikar & Jaybhaye, 2015)	✓	×	×	×	✓	×	×
6.	(Er-raji & Benabbou, 2017)	✓	×	×	×	✓	×	×
7.	(Khurana & Singh, 2018)	×	✓	×	×	✓	×	×
8.	(Miglani & Sharma, 2018)	×	✓	×	×	×	×	✓
9.	(Panchal & Parida, 2018)	×	×	✓	×	×	✓	×
10.	Padmavathi, 2020	×	×	✓	×	×	×	✓
11.	(Al-Rahayfeh & Atiewi, 2019)	×	×	×	✓	×	×	✓
12.	(Lilhore et al, 2020)	×	✓	×	✓	×	×	✓

[a] TS denotes Task Scheduling; LB denotes Load Balancing.

The authors also present some statistics based on the reviewed approaches. Figure 3 (a) below depicts the distribution of the reviewed articles by the year of publication from 2015 until 2020. The authors focused on reviewing recent approaches in the year of 2018 being the highest portion as seen in the figure below. The second highest is in the years of 2019 and 2020 making this review recent and useful for richer research gap.

Figure 3 (b) below categories the reviewed approaches based on two main training method in Machine Learning: supervised and unsupervised. As seen from the figure below, most authors utilize supervised training as it provides a clearer outcome of the variables trained from previous scenarios.

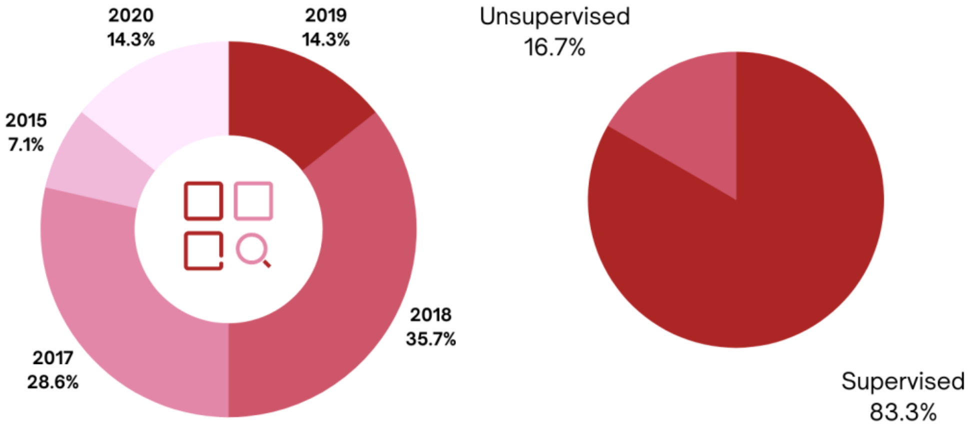

Fig. 3. (a) Percentage of Year of Publications of Surveyed Articles; (b) Supervised or Unsupervised Approach.

B. *Findings*

As can be seen from the above literature, recent approaches are mostly using classification supervised ML which indicates that the dataset used to train such models should be clearly labelled as found from literature [15]. There are four main algorithms [18] used in supervised classification technique listed below:

1) Decision Tree [29]*:* This is considered as one of the most useful algorithm for supervised learning. It is in a flowchart structure and a decision is taking after navigating through the whole tree to execute its features. These types of algorithms can solve quantitative problems and determines a predictive value after answering a set of questions and conditions.

2) Support Vector Machines [30]*:* This is known to be a linear model. In classification, it creates a line called a hyperplane, to split the data into multiple classes. There could exists multiple lines that separate the data, however SVM finds the best line by looking at the closest points to the line.

3) K Nearest Neighbours [31]*:* This algorithm finds similar data points nearest to each other, assumes it is the right value, K for the learning. However, it finds the best results by inputting different values for K.

4) Ensemble Classifier [32]*:* The combination of multiple base classifiers intends to generate accurate results than the single classifier. For example, when using a decision tree, more and more questions are added for classification, this is where these algorithms will become handy to utilize.

Table III below depicts a comparison of the above Machine Learning classifiers. Research claims that the most accurate and appropriate classifier can be identified by estimating the algorithm's accuracy in contrast to the problem [33] the model is trying to solve. In Cloud Computing, recent literature focused on J48 Tree or boosted Tree [15], [18] to solve task classification problems and it has been proven that it provides the best performance compared to KNN and SVM.

TABLE III. MACHINE LEARNING ALGORITHMS COMPARISON [33]

	Decision Trees	**Neural Networks**	**Naïve Bayes**	**kNN**	**SVM**	**Rule-learners**
Accuracy in general	**	***	*	**	****	**
Speed of learning with respect to number of attributes and number of instances	***	*	****	****	*	**
Speed of classification	****	****	****	*	****	****
Tolerance to missing values	***	*	****	*	**	**
Tolerance to irrelevant attributes	***	*	**	**	****	**
Tolerance to redundant attributes	**	**	**	**	***	**
Tolerance to highly interdependent attributes (e.g., parity problems)	**	***	*	*	***	**

	Decision Trees	Neural Networks	Naïve Bayes	kNN	SVM	Rule-learners
Dealing with discrete/binary/continuous attributes	****	***(not discrete)	***(not continuous)	***(not directly discrete)	**(not discrete)	***(not directly continuous)
Tolerance to noise	**	**	***	*	**	*
Dealing with danger of overfitting	**	*	***	***	**	**
Attempts for incremental learning	**	***	****	****	**	*
Explanation ability/transparency of knowledge/classifications	****	*	****	**	*	****
Model parameter handling	***	*	****	***	*	***

[b.] (**** stars represent the best and * star represents the wort performance).

Many researchers have taken new approaches to enhance Cloud Computing's performance, such as utilizing ML techniques or using clustering algorithms. However, these approaches often tend to focus on one aspect either task scheduling or load balancing. Thus, to produce an effective algorithm for better performance, hybrid approach where both task scheduling and load balancing should be considered. It is concluded from this review that machine learning approaches can be used to solve the following challenges in the cloud domain:

- Decision on migration of VM
- Classification of user tasks
- Classification of Virtual Machines

As the literature concluded, current approaches may result in limitations to migration of VM. This can be a challenge as it can result in uneven workload distribution and inefficient resource utilization. An efficient algorithm should migrate VM without exceeding the threshold given [12][34], Machine Learning prediction techniques can help to determine the threshold and makes migration decisions. Although there are many Machine Learning approaches such as regression, clustering, deep learning etc., the recent existing literature focused more on classification supervised Machine Learning technique to solve Balancing and Task Scheduling issues. And it has been shown that boosted tree [18] is the best among 22 classifiers for solving the problem of priority in task scheduling. Since it can schedule a large number of tasks with a high accuracy rate, less training time and memory is suitable for the dynamic nature of Cloud Computing technology.

IV. Conclusion & Future Work

The research aimed to introduce various Machine Learning approaches to address challenges such as Load Balancing and Task scheduling in the cloud computing domain. Using an appropriate classifiers and algorithms, a high-end model can be proposed addressing such challenges. It has been concluded that classification techniques are on the rise in cloud computing domain as it gives Cloud service provider the opportunity to prioritize user requests. It is always a challenge to migrate workload from one VM to another, with Machine Learning techniques an informed and intelligent decision can be made to find the suitable VM and lessen the number of migrations in the future.

In future work, more approaches will be reviewed considering different parameters for each challenge in cloud environment using novel machine learning techniques.

Acknowledgement

We want to express our most generous gratitude and appreciation to Taylor's University, School of Computer Science & Engineering (SCE) for giving us this fantastic opportunity to conduct our research.

References

[1] IDG, "Cloud Computing Survey: 2018," 2018.

[2] M. Agarwal and D. G. M. S. Srivastava, "Cloud Computing: A Paradigm Shift in the Way of Computing," *Int. J. Mod. Educ. Comput. Sci.*, vol. 9, no. 12, pp. 38–48, 2017, doi: 10.5815/ijmecs.2017.12.05.

[3] B. Singh and G. Singh, "A Study on virtualization and hypervisor in cloud computing," *Int. J. Comput. Sci. Mob. Appl.*, vol. 6, no. 1, pp. 17–22, 2018.

[4] M. Adhikari and T. Amgoth, "Heuristic-based load-balancing algorithm for IaaS cloud," *Futur. Gener. Comput. Syst.*, vol. 81, pp. 156–165, 2018, doi: 10.1016/j.future.2017.10.035.

[5] N. Zanoon, "Toward Cloud Computing: Security and Performance," *Int. J. Cloud Comput. Serv. Archit.*, vol. 5, no. December, pp. 1–9, 2015, doi: 10.5121/ijccsa.2015.5602.

[6] D. A. Shafiq, N. Jhanjhi, and A. Abdullah, "vLoad balancing techniques in cloud computing environment: A review," *J. King Saud Univ. - Comput. Inf. Sci.*, Mar. 2021, doi: 10.1016/j.jksuci.2021.02.007.

[7] W. Hashem, H. Nashaat, and R. Rizk, "Honey Bee Based Load Balancing in Cloud Computing," *KSII Trans. Internet Inf. Syst.*, vol. 11, no. 12, pp. 5694–5711, 2017, doi: 10.3837/tiis.2017.12.001.

[8] RightScale, "RightScale 2019 State of the Cloud Report from Flexera," pp. 1–50, 2019.

[9] "Machine Learning - GeeksforGeeks," Jan-2020. .

[10] "How machine learning and cloud computing is transforming the world," Jun-2018. .

[11] D. A. Shafiq, N. Jhanjhi, A. Abdullah, and M. A. AlZain, "A Load Balancing Algorithm for the Data Centres to Optimize Cloud Computing Applications," *IEEE Access*, pp. 1–1, 2021, doi: 10.1109/ACCESS.2021.3065308.

[12] S. Kumar Pande *et al.*, "A Resource Management Algorithm for Virtual Machine Migration in Vehicular Cloud Computing," *Comput. Mater. Contin.*, vol. 67, no. 2, pp. 2647–2663, 2021, doi: 10.32604/cmc.2021.015026.

[13] C. D. Devi and R. V Uthariaraj, "Load Balancing in Cloud Computing Environment Using Improved Weighted Round Robin

Algorithm for Nonpreemptive Dependent Tasks," *Sci. World J.*, 2016, doi: 10.1155/2016/3896065.

[14] A. Dhari and K. I. Arif, "An Efficient Load Balancing Scheme for Cloud Computing," *Indian J. Sci. Technol.*, vol. 10, no. 11, pp. 1–8, 2017, doi: 10.17485/ijst/2017/v10i11/110107.

[15] M. Elrotub and A. Gherbi, "Virtual Machine Classification-based Approach to Enhanced Workload Balancing for Cloud Computing Applications," *Procedia Comput. Sci.*, vol. 130, pp. 683–688, 2018, doi: 10.1016/j.procs.2018.04.120.

[16] M. Jodayree, M. Abaza, and Q. Tan, "A predictive workload balancing algorithm in cloud services," *Procedia Comput. Sci.*, vol. 159, pp. 902–912, 2019, doi: 10.1016/j.procs.2019.09.250.

[17] G. Rjoub and J. Bentahar, "Cloud task scheduling based on swarm intelligence and machine learning," in *2017 IEEE 5th International Conference on Future Internet of Things and Cloud, FiCloud 2017*, 2017, vol. 2017-Janua, pp. 272–279, doi: 10.1109/FiCloud.2017.52.

[18] N. Er-raji, F. Benabbou, M. Danubianu, and A. Zaouch, "Supervised Machine Learning Algorithms for Priority Task Classification in the Cloud Computing Environment," *IJCSNS Int. J. Comput. Sci. Netw. Secur.*, vol. 18, no. 11, pp. 176–181, 2018.

[19] A. P. Tikar, S. . Jaybhaye, and G. . Pathak, "Task Scheduling in the Cloud Using Machine Learning Classification," no. March, pp. 1–6, 2015.

[20] N. Er-raji and F. Benabbou, "Priority Task Scheduling Strategy for Heterogeneous Multi-Datacenters in Cloud Computing," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 2, pp. 272–277, 2017, doi: 10.14569/ijacsa.2017.080236.

[21] S. Khurana and R. Kumar Singh, "Virtual machine categorization and enhance task scheduling framework in cloud environment," *2018 Int. Conf. Comput. Power Commun. Technol.*, pp. 391–394, 2018, doi: 10.1109/GUCON.2018.8675020.

[22] N. Miglani and G. Sharma, "An adaptive load balancing algorithm using categorization of tasks on virtual machine based upon queuing policy in cloud environment," *Int. J. Grid Distrib. Comput.*, vol. 11, no. 11, pp. 1–12, 2018, doi: 10.14257/ijgdc.2018.11.11.01.

[23] B. Panchal and S. Parida, "An Efficient Dynamic Load Balancing Algorithm Using Machine Learning Technique in Cloud Environment," *Int. J. Sci. Res. Sci. Eng. Technol.*, vol. 4, no. 4, pp. 1184–1186, 2018.

[24] M. Padmavathi, S. Mahaboobasha, and V. V. J. Ramakrishnaiah, "A Novel Approach for Cloud Computing Load Balancing with Machine Learning," *Int. J. Adv. Sci. Technol.*, vol. 29, no. 11, pp. 1209–1214, 2020.

[25] A. Al-Rahayfeh, S. Atiewi, A. Abuhussein, and M. Almiani, "Novel Approach to Task Scheduling and Load Balancing Using the Dominant Sequence Clustering and Mean Shift Clustering Algorithms," *Futur. Internet*, vol. 11, no. 5, p. 109, 2019, doi: 10.3390/fi11050109.

[26] D. A. Shafiq, N. Jhanjhi, and A. Abdullah, "Proposing A Load Balancing Algorithm for the Optimization of Cloud Computing Applications," *MACS 2019 - 13th Int. Conf. Math. Actuar. Sci. Comput. Sci. Stat. Proc.*, 2019, doi: 10.1109/MACS48846.2019.9024785.

[27] L. Zhou, X. Cui, and S. Wu, "An optimized load-balancing scheduling method based on the WLC algorithm for cloud data centers," *J. Comput. Inf. Syst.*, vol. 9, no. 17, pp. 6819–6829, 2013, doi: 10.12733/jcis6513.

[28] U. K. Lilhore, S. Simaiya, K. Guleria, and D. Prasad, "An Efficient Load Balancing Method by Using Machine Learning-Based VM Distribution and Dynamic Resource Mapping," *J. Comput. Theor. Nanosci.*, vol. 17, no. 6, pp. 2545–2551, 2020, doi: 10.1166/jctn.2020.8928.

[29] P. Yadav, "Decision Tree in Machine Learning - Towards Data Science," 2018. .

[30] R. Pupale, "Support Vector Machines(SVM) — An Overview - Towards Data Science," 2018. .

[31] O. Harrison, "Machine Learning Basics with the K-Nearest Neighbors Algorithm," 2018. .

[32] E. Lutins, "Ensemble Methods in Machine Learning: What are They and Why Use Them?," 2017. .

[33] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," *Inform.*, vol. 31, pp. 249–268, 2007, doi: 10.1007/s10751-016-1232-6.

[34] H. Babbar, S. Rani, M. Masud, S. Verma, D. Anand, and N. Jhanjhi, "Load Balancing Algorithm for Migrating Switches in Software-Defined Vehicular Networks," *Comput. Mater. Contin.*, vol. 67, no. 1, pp. 1301–1316, 2021, doi: 10.32604/cmc.2021.014627.