International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249 – 8958 (Online), Volume-9 Issue-5, June 2020

# An optimized Load Balancing Algorithm in Cloud Computing

**Pooja Arora, Anurag Dixit**

***Abstract:*** ***The Cloud Computing is the new advanced field with the architecture of service-oriented. As there is increase in the application of cloud computing which lead to rise in the number of task and workload. This workload makes unbalanced load due to uneven distribution of task on different nodes with different capabilities of each node. This makes some nodes overloaded and some underloaded in cloud and lead to unbalanced load. To make efficient utilization of resources of cloud it becomes necessary to balance this load to satisfy the user. The load balancing algorithm is used for the redistribution of task from overloaded node to underloaded node. This paper discusses about the matrices that affect the load balancing algorithm in cloud. It also present the analysis of different load balancing algorithm and related work of different authors. In this paper an algorithm is proposed on the basis of study for optimized load balancing algorithm in cloud computing.***

***Keywords : Cloud Computing, Load Balancing, Resource utilization, Energy Efficiency***

## I. INTRODUCTION

### This Cloud Computing

Cloud computing is the modern internet based service provider which provide on-demand services to clients to access the shared group of resources. The resources can be hardware or software. The cloud computing gets attention from both academic and industrial communities due to its paradigm that "Everything as a Service".Due to the advancement of cloud computing, many enterprises and individuals are using itfor the storage of large data instead of building and maintaining their own local data centres. In today's world, the services on cloud are provided same as utility services like water and electricity. You need to pay as much as you used these services. The cloud users may also enjoy various types of computing services offered by public cloud [12]. The users now focus on their main objective without worrying about computing resources requirement. With time there is continues increase for the cloud computing resources and it lead to more efficient utilization of computing resources. To provide the hassele free services to its client it is necessary to improve the efficiency which further lead to increase the throughput.Because of the implications for greater flexibility and availability at lower cost, cloud computing is a subject that has been receiving a good deal of attention [11].

In Figure 1 there are basically three layers in the cloud computing architecture. Every layer has its own importance.The basic services that the user get from cloud computing architecture are:

1. Software as a Service(SaaS)–In this service end user don't need to install and run different types of applications on their local machine. They can directly use these applications over a cloud network as a service.example is SalesForce.com.
2. Platform as a Service(PaaS)- In this service user can get platform for development and management of Software. In this the software developer can develop and deploy different types application without worrying about tools, languages and API. All these services are provided by cloud service providers. Example is Google App Engine.
3. Infrastructure as a Service(IaaS)-In this service the users are provided storage space in cloud and resources for computation as per their demand and pay as per their usage. Amazon EC2 is an example for the same.

There are four types of cloud infrastructure i.e Public Cloud, Private Cloud, Hybrid Cloud, Community Cloud. The Public cloud is open for all users i.e general pubic and is available all time and user can pay as per usage eg Amazon and Google Cloud. The private cloud is for the particular organization for their use only. This cloud is not available for general public. The Hybrid Cloud is the combination of public and private cloud. This is used for commercial purpose. The Community cloud is the hybrid form of private cloud i.e. it is used for a particular group of users. A particular community form this type of cloud.

### 1.2 Load Balancing

The main purpose of load balancing is to provide optimize usage of computing resources to reduce the deployment and operational cost for users. The importance of load balancing in cloud computing is to provide efficient solution to handle multiple request of multiple client by effective use of computing resources available. The Load Balancing is one of the major problems in cloud computing. Now a days load balancing becomes a major challenge and concerns because of the following reasons:

Revised Manuscript Received on June 22, 2020.
* Correspondence Author
**Pooja Arora*,** IT Dept., BCIIT, Delhi, India.
**Anurag Dixit,** SCSE,Galgotias University, Delhi, India.

1. Improvement of performance in cloud environment
2. Maximize the throughput
3. Minimize Response Time
4. Minimize the Cost incurred to user
5. Optimize the resource utilization
6. Save Energy

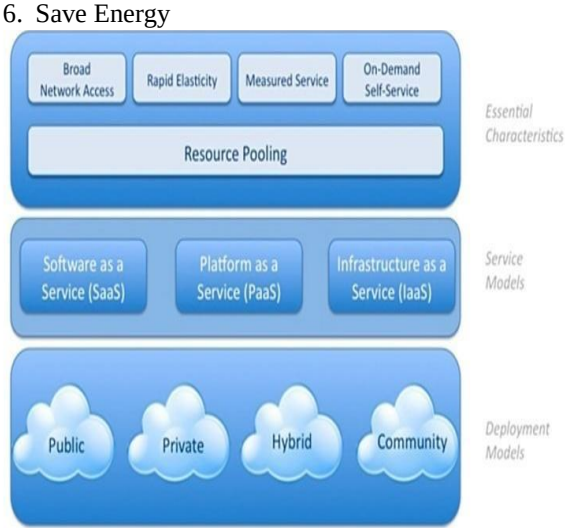


**Figure 1 : Cloud computing architecture[25]**

Without load balancing there can be delays in user response, resources cannot be optimally utilized and resultant is fall in the performance of system. The load in the cloud computing can be load on network, memory and CPU etc. The Load Balancing is the method of allocating and reallocating the load among the available nodes in such a manner that no node will remain overloaded in order to increase the performance in cloud computing environment. Load Balancing calculates various terms like minimizing communication delays, minimizing execution time, maximizing throughput and maximizing resource utilization [10].

The purpose of this paper is to give a literature on the load balancing algorithms and present the categorization of these algorithms. The rest of this paper is organized as follows. Section 2 discusses classification of load balancing algorithm in Cloud Computing. The load balancing metrics and policies are provided in Section 3. Finally the Section 4 concludes this paper.

## II. LOAD BALANCING ALGORITHM IN CLOUD COMPUTING

The algorithm balances the load in the cloud by transferring the load of overloaded machine to under-loaded machine in an efficient manner. The Load Balancing algorithms were categorized on the basis of two factors:

1. Static algorithm
2. Dynamic algorithm

The Static algorithm first needs to get the information of all resources of system. In this algorithm no run-time status of resources are required. It works at the start before execution of task. No changes can be done at the execution time for balancing the load. This type of algorithm is suitable in homogenous process. In this the decision for distributing the load is depend on the prior information of system resources and task. It is used with predicting processing loads in advance. The Dynamic algorithm works at the run-time. In this the decision for distributing the load is depend on the current state of system. If there is no overloading of machine at run-time then no redistribution of task will take place. The redistribution decision will take place at run-time only on the basis of current status of system resources. This algorithm gives better performance than static algorithm. It works well for heterogeneous system. It is used with unpredictable processing loads.

### 2.1 Matrices that effects load balancing

Throughput (TP):
Throughput is the total tasks completed per unit time by a machine. The system performance value is evaluated by throughput. The System performance is high if throughput value is high. The System performance is low if throughput value is low.
Thrashing (TH):
Due to limited memory or other resources in cloud setup there may cause thrashing. It occurs in the cloud setting when VM is busy in migration instead of processing the task. This is due to incorrect scheduling algorithm.
Therefore, the correct algorithm for load balancing used to manage this aspect.
Reliability (R):
Reliability will perform as per specification of system. If any failure occurs during the runtime of task then that task will be shifted to any other Virtual Machine and thus create a failure free System. The stability of the system depends on reliability of the system.
Accuracy (A):
It evaluates the correctness of result of executed task. It measures the output of executed task and compare it with actual values.
Predictability (PR):
It is the rate for predicting of number of tasks allocated , executed and completed with the available resources in cloud. The value is predicted on the basis of previous pattern of the arrived task, the assigned task and executed task in the cloud system. The balancing of the load improves with improved predicted values of task which further improves makespan.
Makespan (MS):
It means the time needed for completion of all submitted tasks to the cloud system. The time of resources allocation to the task is makespan.
Scalability(S):
It means the capability of system using load balancing method to perform in the situation of overloaded task. It means how the system works under unexpected situation. The resources which are available will be rescale from time to time.
Fault Tolerance (FT):
It means ability of the system to work in case of failure. To provide uninterrupted service in case of one or more failure of system elements is the ability of fault-tolerant method. The extra resources and Virtual Machines are needed to handle some types of failure. For this we need to incur extra cost.
Associated Overhead (AO):

The overhead is the additional cost which is incurred for executing the load balancing algorithm. The associated overhead is overhead incurred associated with implementing the algorithm.

Migration time (MT):

It is the time needed to migrate a task or Virtual Machine from one resource to another. The migration of task may take place from one VM to another VM under single or multiple host. In the same way, the migration of VM will take place from one host to another host within same or different data centers. The more migration of VMs results in more migration time which affect the makespan and load balancing of the system.

Response time (RT):

It is the time required by the system to respond a task. It is the total time consists of transmission time, waiting time, and service time. Thus, the system performance is inversely related to the response time. The minimum response time makes better makespan value.

**2.2 Analysis of different load balancing algorithm**

| S.No. | Authors | Methods | Advantages | A. Disadvantages | B. Parameters | C. Experimental /Simulaion Tool used |
|---|---|---|---|---|---|---|
| 1 | Shang-Liang Chen et al. [1] | Cloud Load Balance (CLB) algorithm | balance the loading performance when users logged in at the same time | **D.** The method failed to consider different load approach | **E.** Online users, Time-cost, Priority Service(PS) value | **F.** testing server uses OS Windows7; the Programming Language is MS Visual Studio 2010 C# with the MS SQL Server 2005 database. The Server is Internet Information Services 2.0 with a RAM of 4,096 MB and an Intel T2390/1.86 GHz CPU |
| 2 | Qi Liu et al. [2] | Hadoop-LB with prediction model based on K-ELM (PMK-ELM) | reduce the running<br>time, high accuracy | **G.** waste more storage space | **H.** execution time | **I.** The server is equipped with 288 GB of<br>**J.** memory and a 10 TB hard driver on<br>**K.** Hadoop 2.6.0 |
| 3 | Jia Zhao et al. [3] | Load Balancing based on Bayes and Clustering (LB-BC) | improved throughput | The method failed to apply LAN | **L.** MakeSpan, Throughput | **M.** CloudSim platform |
| 4 | Shiva Razzaghzadeh et al. [4] | load balancing strategy | improves the makespan and cost | The method failed to extend load balancing for dependent tasks and failed to consider<br>more factors such as fault tolerance for HR label | **N.** Makespan, execution time | **O.** Cloudsim and validate it in Amazon EC2 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 5 | Ranesh Kumar Naha and Mohamed Othman [5] | Cost aware brokering and Load aware brokering algorithm | minimized the cost | The method failed to consider efficient load balancing algorithm for minimizing the execution time | **P.** virtual machine cost, Response time and processing time | **Q.** CloudAnalyst |
| 6 | Weihua Huang et al. [6] | Fuzzy clustering method with Feature Weight Preferences for Load Balancing (FWPFC-LB) | achieve better load balancing performance | The method failed to consider adaptive parameter optimization for data fusion and convergence rate acceleration in clustering | **R.** Throughputs, Makespans, Migration times | **S.** |
| 7 | Narander Kumar and Diksha Shukla [7] | fuzzy row penalty method | increase performance and scalability, minimize associated overheads, takes less execution time | workload is not distributed properly | **T.** Execution Time, response time and space complexity | **U.** CloudSim |
| 8 | Santanu Dam et al. [8] | Ant-Colony-Based Meta- Heuristic Approach | minimize the make span as well as the number of Virtual Machine (VM) is also reduced | The method failed to include fault tolerance and priority of the job | **V.** response time | **W.** CLOUD ANALYST |

**2.3 Related Work**

In this section, the survey of eight existing techniques based on LB is elaborated along with their drawbacks.

Shang-Liang Chen et al. [1] developed a dynamic annexed balance method for solving the problem caused by uneven distribution of loads. Here, the LB in cloud is taken into consideration for processing server power and for computer loading. Thus, the server can manage extreme computational requirements. At last, two algorithms in load balancing are tested with experiments for proving the approach to be innovative. The method can balance the load performance when the users are logged at same time, but the method is not applicable for different load methods.

Qi Liu et al. [2] devised an adaptive scheme, named Hadoop-LB using Prediction Model based on Kernel function-Extreme Learning Machine (PMK-ELM) algorithm to attain time efficiency by offering heterogeneous cloud infrastructure. A dynamic speculative execution strategy on real-time management for cluster resources is derived for optimizing the execution time of map phase and a prediction model is utilized for fast prediction with minimum task execution time. An adaptive solution is devised for optimizing the performance of space-time by combining the prediction model with a multi-objective optimization algorithm, but the method requires more storage space.

Jia Zhao et al. [3] devised an advanced approach, named Load Balancing based on Bayes and Clustering (LB-BC), for deploying requested tasks to the hosts. LB-BC uses limited constraint for physical hosts to attain a task deployment approach with global search capability using performance function for computing the resources. Here, the Bayes theorem is integrated with the clustering process for obtaining optimal clustering set with a physical host. However, the method is unable to work on real computing environment.

Shiva Razzaghzadeh et al. [4] developed a method for distributing the dynamic load using distributed queues in cloud infrastructure. This method maps the tasks and HRs by allocating label to each HR. The load balancing and mapping process are devised on the basis of Poisson and exponential distribution. This method permits the task allocation process to execute with high power by adapting distributed queues aware of the service qualities. The method did not use genetic algorithm for determining the optimal HR by resolving the faults. Ranesh Kumar Naha and Mohamed Othman [5] developed an algorithm, named Cost aware brokering and Load aware brokering algorithm, for balancing the loads in data centers and the VM.

This algorithm minimizes the overall processing and response times as the tasks are assigned to the available physical resources in an effective manner. The algorithm did not consider real-world cloud brokering for the evaluation. Weihua Huang et al. [6] developed a fuzzy clustering method using feature weight preferences for overcoming load balancing issues in multiclass system resources and can attain optimal solution for load balancing by load data fusion. Here, feature weight preferences are utilized for establishing the relationship between specific cloud scenarios and LB procedure. The method was not applicable for optimizing the adaptive parameter in load data fusion and to improve the convergence rate acceleration in clustering.

Narander Kumar and Diksha Shukla [7] devised a method, named fuzzy row penalty method, to solve the issues of LB in a cloud computing environment based on fuzzy. Here, the fuzzy technique is utilized for addressing uncertain response time in fuzzy cloud environment. Here, the fuzzy row penalty method was utilized to address the balanced fuzzy LB problem and unbalanced fuzzy LB problem in a cloud infrastructure. The generated result is utilized to solve the LB issues based on response time and space complexities for maximizing the performance, minimizing the scalability, overheads, but workload distribution is improper. Santanu Dam et al. [8] utilized an advanced computational intelligence technique, named Ant Colony Optimization (ACO), to balance the loads among VM in cloud computing. Here, the ACO is utilized for designing an intelligent multi-agent system by collective behavior of ants. Here, the ACO is utilized for addressing the LB problem in cloud computing and to minimize the makespan and to minimize the VM usage, but fault tolerance and handling high priority job are still considered as major issues. Mahfooz Alam and Mohammad Shahid [26] proposed a Load Balancing Strategy with Migration cost LBSM to execute an independent batch of tasks on various heterogeneous MINs viz. Here, MetaCube, X-Torus and Folded Crossed Cube having the objective of minimizing the load imbalance on processors. Raza Abbas Haidri et al. [27] developed a Capacity based Deadline Aware Dynamic Load Balancing (CPDALB) algorithm to address the load balancing problem. CPDALB focuses on the selection of VM for allocation of tasks to ensure customer satisfaction in terms of meeting deadline constraints and cost of running applications. It uses the concept of deploying requests to VMs based on their processing capacities to minimize load imbalance with satisfying the deadline.

## III. ALGORITHM FOR OPTIMIZED LOAD BALANCING IN CLOUD COMPUTING

The pseudocode of the load balancing algorithm using the proposed EHGWO algorithm is presented in Algorithm 1.

| Algorithm 1. Proposed Load balancing Algorithm | |
|---|---|
| 1 | Initialize the tasks, $Y_s = 100$ |
| 2 | For each task |
| 3 | Assign the task to VMs based on round robin scheduling |
| 4 | Compute $\ell\left(M_g^P\right)$ using equation (8). |
| 5 | If $\ell\left(M_g^P\right) > 0.8$ |
| 6 | Call load balancing algorithm using the proposed EHGWO algorithm |
| 7 | Else |
| 8 | No load balancing |
| 9 | End if |
| 10 | End for |
| 11 | Repeat |

In this section, the proposed Elephant Herding-based Grey Wolf Optimizer (EHGWO) proposed for balancing the loads to enhance the efficiency of the cloud computing system is explained. The proposed EHGWO is designed by integrating EHO [21] and GWO [20] for selecting the optimal VM. The goal is to balance the loads present in the VM in such a way that no system is overloaded by removing the tasks from overloaded VMs and assign them to underloaded VMs. Thus, it is essential to choose the tasks, which needs to be reallocated in an effective manner. Initially, the capacity of loads occupied by the VM is determined using the executed tasks and then, the balance is evaluated. Whenever the load of VM is unbalanced, the capacity with load is evaluated to take the decision for the load balancing. The proposed EHGWO is adapted for reallocating the tasks by balancing the loads, whenever the VM is overloaded considering certain factors, which involve bandwidth, execution time, priority and communication cost. Once the tasks are removed, it is added to other VMs for task execution.

The proposed model performs the reallocation of tasks for balancing the load by allocating the tasks from underloaded VM to overloaded VM by enhancing the cloud system efficiency.

Here, the load status is checked once the tasks are assigned to the VMs and then, the reallocation is done. Thus, the load of PM and VM, and capacity of VM are evaluated for allocating the tasks. These evaluations are utilized for migrating the tasks from the overloaded VM and the VM that can manage the loads are determined using the load balancing algorithm.

Here, two pick factors, namely Task pick Factor (TPF) and VM pick Factor (VPF) are designed, which decide the selection of tasks. These factors determine the suitable under loaded VMs for reallocating the tasks that are taken from the overloaded VM. Meanwhile, the VPF is computed using the capacity and the load of VMs, for finding the under loaded VMs and can be utilized for reallocating the tasks.

## IV. CONCLUSION

This paper proposes the load balancing algorithm using EHGWO in a cloud computing model using two pitch factors, named Task Pick Factor (TPF) and Virtual Machine Pick Factor(VPF). For initiating load balancing, the tasks assigned to the overloaded VM are reallocated to under loaded VMs. Here, the proposed load balancing algorithm adapts capacity and loads for the reallocation. Based on TPF and VPF, the tasks are reallocated from VMs using the proposed EHGWO. The proposed EHGWO is developed by integrating EHO and GWO algorithm using a new fitness function formulated by load of VM, migration cost, load of VM, capacity of VM, and makespan.

## REFERENCES


1. Shang-Liang Chen, Yun-Yao Chen , Suang-Hong Kuo," CLB: A novel load balancing architecture and algorithm for cloud services", Computers & Electrical Engineering,Vol:58, pp:154-160, February 2017.
2. Qiiu, Weidong Cai, Jian Shen, Xiaodong Liu, Nigel Linge, "An Adaptive Approach to Better Load Balancing in a Consumer-centric Cloud Environment," IEEE Transactions on Consumer Electronics, Vol: 62, no: 3, pp: 243 - 250, August 2016.
3. Jia Zhao, Kun Yang, Xiaohui Wei, Yan Ding, Liang Hu, and Gaochao Xu, "A Heuristic Clustering-based Task Deployment Approach for Load Balancing Using Bayes Theorem in Cloud Environment", IEEE Transactions on Parallel and Distributed Systems, Vol: 27, no: 2, pp: 305 -316, February 2016.
4. Shiva Razzaghzadeh, Ahmad Habibizad Navin, Amir Masoud Rahmani, and Mehdi Hosseinzadeh, "Probabilistic Modeling to Achieve Load balancing in Expert Clouds", Ad Hoc Networks, January 2017.
5. Ranesh Kumar Naha and Mohamed Othman, "Cost aware service brokering and performance sentient load balancing algorithms in the cloud", Journal of Network and Computer Applications, Vol: 75, pp: 47–57, November 2016.
6. Weihua Huang, Zhong Ma, Xinfa Dai, Mingdi Xu and Yi Gao, "Fuzzy Clustering with Feature Weight Preferences for Load Balancing in Cloud", International Journal of Software Engineering and Knowledge Engineering,Vol: 28, no:5,pp:593–617,2018.
7. Narander Kumar and Diksha Shukla," Load Balancing Mechanism Using Fuzzy Row Penalty Method in Cloud Computing Environment", Information and Communication Technology for Sustainable Development, pp: 365-373,2017.
8. Santanu Dam, Gopa Mandal, Kousik Dasgupta and Parmartha Dutta," An Ant-Colony-Based Meta-Heuristic Approach for Load Balancing in Cloud Computing", Applied Computational Intelligence and Soft Computing in Engineering, pp: 29,2018.
9. Shalini Joshi and Uma Kumari, "Load Balancing in Cloud Computing:Challenges & Issues," In proceedings of 2nd International Conference on Contemporary Computing and Informatics (IC3I), December 2016.
10. Mahfooz Alam and Zaki Ahmad Khan," Issues and Challenges of Load Balancing Algorithm in Cloud Computing Environment", Indian Journal of Science and Technology, Vol: 10, no:25, July 2017.
11. Wayne Jansen and Timothy Grance, "Guidelines on security and privacy in public cloud computing",pp: 800-144, 2011.
12. Jianting Ning, Zhenfu Cao, Xiaolei Dong, Kaitai Liang, Hui Ma and Lifei Wei," Auditable -Time Outsourced Attribute-Based Encryption for Access Control in Cloud Computing",IEEE Transactions on Information Forensics And Security, 2017.
13. Zenon Chaczko, Venkatesh Mahadevan, Shahrzad Aslanzadeh and Christopher Mcdermid," Availability and Load Balancing in Cloud Computing", In proceedings of International Conference in Computer and Software Modeling, IPCSIT, Vol:14 , 2011.
14. Daraghmi, E.Y. and Yuan, S.M.,"A small world based overlay network for improving dynamic load-balancing", Journal of Systems and Software, Vol: 107, pp:187-203,2015.
15. A. S. Milani and N. J. Navimipour," Load balancing mechanisms and techniques in the cloud environments: Systematic literature review and future trends", J. Netw. Comput. Appl. Vol:71, no:1, pp: 86–98,2016.
16. Y. C. Jiang, "A survey of task allocation and load balancing in distributed systems", IEEE Trans. Parallel Distrib. Syst.,Vol: 27no:2,pp: 585–599,2016.
17. G. Mateusz, G. Alicja and B. Pascal," Cloud brokering: Current practices and upcoming challenges", IEEE Cloud Comput.,Vol: 2,no:2,pp: 40–47,2015.
18. D. Falco, E. Laskowski, R. Olejnik, U. Scafuri, E. Tarantino and M. Tudruj, "Extremal optimization applied to load balancing in execution of distributed programs", Appl. Soft Comput.,Vol: 30,no:3,pp: 501–513,2015.
19. Y. Jiang, "Concurrent collective strategy diffusion of multi agents: The spatial model and case study", J. Netw. Comput. Appl.,Vol: 39,no:4,pp: 448–458,2009.
20. Mirjalili, S., Mirjalili, S.M. and Lewis, A.,"Grey wolf optimizer" Advances in engineering software,Vol: 69, pp.46-61,2014.
21. G. Wang, S. Deb and L. d. S. Coelho, "Elephant Herding Optimization, "In proceedings of 3rd International Symposium on Computational and Business Intelligence (ISCBI),pp:1-5,2015.
22. Singh, A., Juneja, D., Malhotra, M., 2015. Autonomous agent based load balancing algorithm in cloud computing. Procedia Comput. Sci. 45, 832–841.
23. Singh, A., Juneja, D., Malhotra, M., 2015. A novel agent based autonomous and service composition framework for cost optimization of resource provisioning in cloud computing. J. King Saud Univ. Comput. Inf. Sci.
24. Moganarangan, N., Babukarthik, R.G., Bhuvaneswari, S., Basha, M.S., Dhavachelvan, P., 2016. A novel algorithm for reducing energy-consumption in cloud computing environment: Web service computing approach. J. King Saud Univ. Comput. Inf. Sci. 28 (1), 55–67.


## AUTHORS PROFILE

**Ms. Pooja Arora ,** Assitant Professor in Banarsidas Chandiwala Institute of Information Technology(affiliated to GGSIPU, Delhi). She is pursuing Ph.D in Computer Science from Galgotias University.She has done M.Tech(CSE) from School of Computer Science & Engineering from G.G.S.I.P.U Delhi. She has total experience of 11 years.

**Dr. Anurag Dixit**, Professor in School of *Computer Science and Engineering Galgotias University.He has done Ph.D in Computer Science from Jawaharlal Nehru University, New Delhi and M.Tech in Communication Engineering.His Research Area is Soft Computing E-learning and Software Engineering. He has total experience of 17+ yearS.