

IJEAST

INTERNATIONAL JOURNAL
OF ENGINEERING APPLIED SCIENCE
AND TECHNOLOGY

VOLUME : 1 ISSUE : 4 Print / Issue Publication Date: 08-Oct-2016

ISSN : 2455-2143

Indexed In

Google SCHOLAR OPEN ACCESS doi Crossref

WWW.IJEAST.COM

editor@ijeast.com

International Journal of Engineering Applied Sciences and Technology, **2016**
Vol. 1, Issue 4, ISSN No. 2455-2143, Pages 1-7
Published Online January-February 2016 in IJEAST (http://www.ijeast.com)



AN EFFICIENT TREE-BASED CLASSIFIERS ENSEMBLE WITH ANT COLONY OPTIMIZATION USING ROUGH SET THEORY

Surekha Samsani
Department of CSE
JNTUK-UCEV, Vizianagaram, AP, India

Dr. G. Jaya Suma
Department of IT
JNTUK-UCEV, Vizianagaram, AP, India

***Abstract*— Feature Subset Selection is an essential preprocessing step used for classification applications to improve accuracy and speed. In this paper, the popular Ant Colony algorithm, which is swarm intelligence technique, is integrated with Rough Set Theory for feature subset selection and the obtained feature subset is submitted to the Decision Tree Classification technique. The performance of the RST based ACO feature selection technique is evaluated by submitting the reduced dataset to two different decision tree classification techniques one is the C4.5 and the other is the Variable Precision Rough Set based Decision Tree classification. Experiments have been conducted on the medical datasets taken from the UCI machine learning repository and the results revealed that the RST based ACO generated feature subsets with an improvement in accuracy of the classification techniques.**

Keywords— **Feature Subset Selection, Ant Colony Algorithm, Rough Set Theory, Variable Precision Rough Set Model based Decision Tree.**

I. INTRODUCTION

The dimensionality of the real world datasets is increasing day by day and this huge dimensionality of the datasets is increasing the search space which further affects the speed and computational complexity of the classification task and this curse of dimensionality is becoming a major problem in machine learning and data mining. Huge dimensional datasets are consisting of redundant and inappropriate features; these irrelevant and redundant features are increasing the risk of the model over fitting in the decision tree classification task, which also affects the accuracy of the decision tree classification. Hence, from these huge set of features there is a need to identify the most promising ones that contribute maximum for the classification.

Feature Subset Selection [1] is the process of identifying the most relevant set of features from the datasets. During past two decades, swarm-based optimization techniques are being applied successfully in many applications for solving complex optimization problems. In this paper, the popular Ant Colony Algorithm, which is based on the pheromone concepts of the ant colonies, is integrated with the rough set theory for selecting the most relevant set of features. Since, Rough Set Theory [2] is a mathematical tool for dealing with uncertainty and vagueness and incorporating the concepts of RST in feature selection process increases the efficiency of the underlying feature selection technique. After identifying the most relevant set of features, the reduced dataset is submitted to the two different decision tree classification techniques. One is the traditional entropy based C4.5 decision tree classification technique and the other is the RST based Variable Precision Rough Set Model based Decision tree classification technique. The performance of the RST based ACO feature subset selection technique is evaluated by submitting the reduced dataset to the C4.5 and VPRSM based DT classification techniques and the results are tested using 10-fold cross validation test.

The rest of the paper is structured as follows: section II discusses the related work, section III gives the methodology of the work, and in section IV experimental results are given and finally section V concludes the paper.

II. RELATED WORK

Chen *et al*[3] proposed a new feature selection technique which integrates rough set theory with Ant Colony Optimization for selecting the most relevant set of features from several medical data sets and compared the obtained results with the traditional rough set based Quick reduct algorithm for feature subset selection and the results proved that the new hybrid approach i.e., RSTACO generated most promising features. Maryam *et al*[4] proposed a new feature selection method which uses the output of Genetic Algorithm as the initial step for the Ant Colony Algorithm for Persian font recognition. And the results proved that the hybridization of GA with ACO achieved better results than GA and ACO when applied alone. Chen *et al*[5] proposed a modified ACO algorithm for selecting the important features from several images and compared the results with the traditional GA and ACO algorithms. Jensen *et al*[6] presented a Rough Set Approach for Attribute reduction and tested with approach on

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 4, ISSN No. 2455-2143, Pages 1-7
Published Online January-February 2016 in IJEAST (http://www.ijeast.com)

13 datasets and compared the results with the conventional hill climbing approaches for feature selection and the results proved that the new RSAR approach for feature selection resulted in finding minimal reducts. Mao *et al*[7] proposed a new approach for inducing decision trees based on the concepts of Variable Precision Rough Set Theory and they applied the Variable Precision Explicit region concept for the for node selection in the process of decision tree induction and the results showed that the new approach for decision tree induction results in minimum tree size while maintaining the accuracy rate.

III. METHODOLOGY

The methodology of the proposed work is given in figure 1.

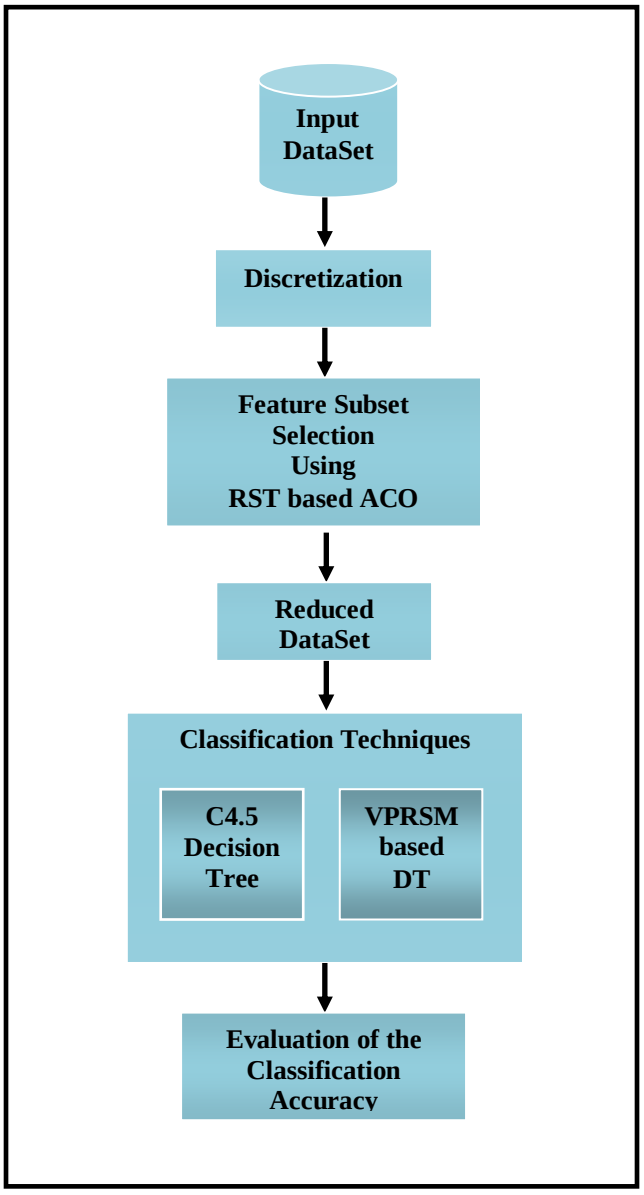


Figure 1. Methodology of the proposed work

The methodology of the proposed work can be divided into two major phases namely Feature Subset Selection using RST based ACO and Tree-based Classification. At first, the input dataset should be discretized and then identify the most relevant set of features by applying the RST based ACO technique and then reduce the size of the dataset by keeping only those features generated by the feature selection technique and then submit the reduced dataset to the two different decision tree based classifiers namely C4.5 and VPRSM based Decision Tree and finally evaluate the performance of the FSS using RST based ACO by 10-fold cross validation examination.

Phase 1: The RST based ACO algorithm for Feature Subset Selection

The Ant Colony Optimization algorithm is the popular swarm based optimization algorithm, which is based on the natural behavior of the real ant colony societies. The ACO algorithm mimics the pheromone depositing principle followed by the real ant colony societies and has been applied successfully for selecting the most relevant set of features. In this work, the concepts of Rough Set Theory have been incorporated with the swarm based optimization algorithm ACO for generating the most promising features.

The basic ACO algorithm[3] starts with a set of artificial ants, where each ant corresponds to a possible solution to the problem and these set of ants try to perform better solution in some predetermined number of iterations. Each ant starts from a random feature, and then selects the best route and the pheromone is updated for that route, this process repeats for a predetermined number of times and finally the route with highest pheromone will be selected as the best one. But, the problem with the basic ACO algorithm for optimization is, it searches for all possible paths and experiences more time complexity because of the random selection of starting point. So, to improve the performance by reducing the number of starting points, the concepts of Rough Set Theory have been integrated with the basic ACO algorithm for selecting the most promising starting points and then searches for the best solution by starting only from these points and trying out of selected paths only. Hence, the Rough Set based approach to feature selection based on ACO starts from the core features and thus reduces the effort required to try out all possible paths and these core features can be generated by the feature core algorithm [3]. The basic concepts of the RST can be found in the literatures [8, 9, and 10].

Probability of selecting a feature:

$$p_{ij}^{n}(k)=\frac{\tau_{ij}^{\alpha}\eta_{ij}^{\beta}(k)}{\Sigma_{l\varepsilon allowed_n}\tau_{il}^{\alpha}\eta_{il}^{\beta}(k)} \quad , \quad j \varepsilon allowed_n \quad \ldots\ldots \quad (1)$$

Where

n represents the number of ants, k is the number of iterations, $allowed_n$ is the set of unselected features, τ (i, j) is the

International Journal of Engineering Applied Sciences and Technology, **2016**
Vol. 1, Issue 4, ISSN No. 2455-2143, Pages 1-7
Published Online January-February 2016 in IJEAST (http://www.ijeast.com)



pheromone value and η (i, j) is the heuristic information for choosing feature j when ant is at i^{th} feature. α and β are the parameters used to determine the relative importance of the pheromone and heuristic information and their value is in between 0 and 1.

Pheromone Updating Rule:

$$\tau_{ij}(k+1) = \rho\tau_{ij}(k+1) + \Delta\tau_{ij}(k) \ldots (2)$$

Where

$\tau_{ij}(k)$- is the amount of pheromone on a given path from i to j at k^{th} iteration

$\tau_{ij}(k+1)$ - is the amount of pheromone at next iteration k+1

ρ - is a decay constant used to simulate the evaporation of pheromone,

$\Delta\tau_{ij}(k)$- is the amount of pheromone deposited, given by

$$\Delta\tau_{ij}(k) = \begin{cases} \Sigma \frac{q}{|R(k)|} & \text{if the path from } i \text{ to } j \text{ has been traversed} \\ 0 & \text{Otherwise} \end{cases}$$

Where

q is constant and

R (k) is the minimal feature reduct at k^{th} iteration.

Heuristic Information:

$$\eta(r,p) = sgn(r,\{Core \cup p\},D) \ldots . (3)$$

Where Core(C) denotes the set of core C with respect to D in DT, and the given ant is currently at node p and r ∈ C – {Core ∪ p}. ***sgn***[3] denotes the function of significance.

Algorithm of Ant Colony Optimization for Feature Subset Selection

1. Compute the core features by feature core algorithm[3] and let it be Core
2. Repeat
3. For each ant
4. Select a feature randomly from the unselected feature set {C-Core} by equation(1)
5. Find the solution by computing heuristic information by equation (3)
6. Update pheromone by equation (2)
7. Until (Max number of cycles)

Phase 2: Tree-based Classification Techniques

In this paper, the reduced features obtained from the RST based ACO are submitted to the two different types of decision tree classification techniques namely C4.5 and VPRSM based Decision tree classification.

In C4.5 [11] Decision tree classification, the selection of the splitting attribute is based on the fundamental entropy based measures and in the VPRSM based decision tree classification technique [12,13,14] the selection of splitting attribute is based on the variable precision explicit region. Variable Precision Rough Set model is an extension to the RST with the addition of a new parameter β called the misclassification degree [15], and with the introduction of this misclassification degree VPRSM based DT classifies the data more accurately and also addresses the problem of model over fitting caused the traditional decision tree construction methods.

The performance of the RST based ACO feature selection algorithm is evaluated by 10-fold cross validation test [16].

IV. EXPERIMENTAL ANALYSIS

For experimental analysis, three bench mark standard medical data sets Chronic Kidney Disease, Diabetes and Vertebral Column are taken from UCI repository [17] and are listed below in the Table 1.

Table 1 Description of the Data sets

Data Set	Classes	Instances	Total features	Continuous Features
Chronic Kidney Disease	Chronic disease, Not Chronic Disease	769	24	11
Diabetes	Tested Positive, Tested Negative	400	8	6
Vertebral column	Hernia, Spondylolithesis, Normal	301	6	6

The observed accuracies of the two classification techniques, C4.5 and VPRSM based DT are given in Table 2.

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 4, ISSN No. 2455-2143, Pages 1-7
Published Online January-February 2016 in IJEAST (http://www.ijeast.com)

Table 2 Performance of the Classification techniques on Undiscretized Data

	C4.5			VPRSM based DT		
	Chronic Kidney Disease	**Diabetes**	**Vertebral Column**	**Chronic Kidney Disease**	**Diabetes**	**Vertebral Column**
Size	11	39	23	10	24	***5***
Leaves	6	20	12	5	13	***3***
Accuracy	95.25%	73.82%	81.61%	95.45%	78.23%	***83.21%***

After discretization of the continuous valued attributes, the accuracy of the classifications techniques has been increased. The results of VPRSM based DT for different β-values is given in tables 3,4 and 5.

Table 3 Results of VPRSM based DT on Chronic Kidney disease with different β values

β value	**0.1**	**0.15**	**0.2**	**0.25**	**0.3**	**0.35**	**0.4**	**0.45**
Size	12	11	***9***	3	3	4	4	4
Leaves	8	8	***7***	2	2	3	3	3
Accuracy	93.75	77	***97.25***	85.75	85.75	69.75	69.75	69.75

Table 4 Results of VPRSM based DT on Diabetes dataset for different values of β

β value	**0.1**	**0.15**	***0.2***	**0.25**	**0.3**	**0.35**	**0.4**	**0.45**
Size	18	18	***16***	16	9	2	2	3
Leaves	5	2	***6***	6	4	1	1	2
Accuracy	71.95	38.76	***81.26***	81.26	69.37	99.08	99.08	77.70

Table 5 Results of VPRSM based DT on Vertebral Column for different values of β

β value	**0.1**	**0.15**	***0.2***	**0.25**	**0.3**	**0.35**	**0.4**	**0.45**
Size	3	3	***3***	3	3	3	4	4
Leaves	2	2	***2***	2	2	2	3	3
Accuracy	86.77	86.77	***86.77***	86.77	86.77	86.77	66.77	86.77

From the results of tables 3,4 and 5, the VPRSM based DT model yields best results for **β=0.2.** The observed accuracies of the C4.5 and VPRSM based DT on discretized datasets is given in table 6.

Table 6 Classification accuracies on Discretized Data

	C4.5			VPRSM		
	Chronic Kidney Disease	**Diabetes**	**Vertebral Column**	**Chronic Kidney Disease**	**Diabetes**	**Vertebral Column**
Size	9	22	5	9	16	3
Leaves	5	13	3	7	6	2
Accuracy	95.25	78.25	81.61	97.25	81.23	86.77

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 4, ISSN No. 2455-2143, Pages 1-7
Published Online January-February 2016 in IJEAST (http://www.ijeast.com)

The graphical representation of the comparison of experiment results of accuracy of the two classification techniques on discretized data of various data sets are shown in figure 2.

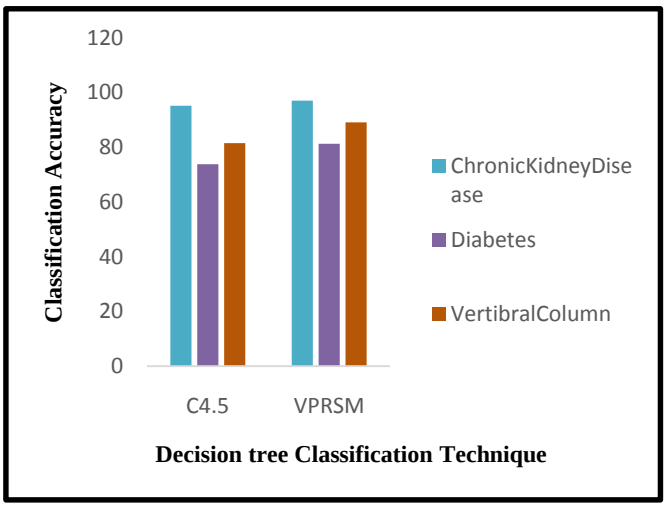


Figure 3 Comparison of accuracies of C4.5 and VPRSM based DT on Discretized data

The feature selection technique is employed on the data set that results in feature subset and the experimental results are shown in table 7

Table 7 Feature Subset generated for the RST based ACO

Data Set	Total Number of features	Reduced Number of features	Feature Sub Set
Chronic Kidney Disease	24	10	Age, Specific gravity, Albumin, Serum creatinine, Sodium, Potassium, Hemoglobin, Red blood cell count, Diabetes mellitus, Pedal edema, Anemia
Diabetes	8	6	Pregnant, Plasma, Pressure, Insulin, Mass, Pedi
Vertebral Column	6	6	Pelvic incidence, Pelvic tilt, Lumbar lordosis angle, Sacral lope, Pelvic radius, Grade of spondylolithesis

The experimental results obtained from various data sets of reduce data for C4.5 and VPRSDT are listed in Table 8 and Table 9 respectively.

Table 8 Classification accuracies for C4.5 on Reduced Data

	C4.5 Decision Tree		
	Chronic Kidney Disease	**Diabetes**	**Vertebral Column**
Size	11	22	23
Leaves	9	13	3
Accuracy	95.25	78.25	81.61

Table 9 Classification accuracies for VPRSDT on Reduced Data

	VPRSM based Decision Tree		
	Chronic Kidney Disease	**Diabetes**	**Vertebral Column**
Size	9	16	3
Leaves	7	6	2
Accuracy	97.75	81.23	86.77

The comparison of experiment results of accuracy and size of the trees obtained for reduced features of various data sets is shown in figure 3.

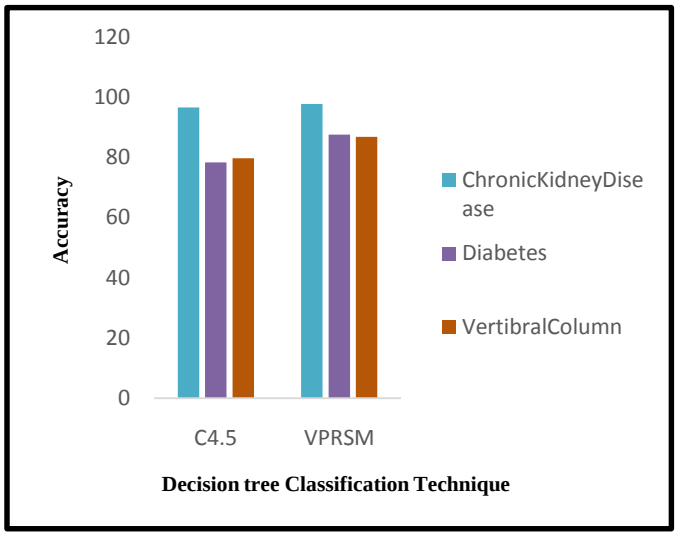


Figure 4 Comparison of accuracies of C4.5 and VPRSM based DT on reduced data

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 4, ISSN No. 2455-2143, Pages 1-7
Published Online January-February 2016 in IJEAST (http://www.ijeast.com)



The resulted accuracies obtained by applying different classification techniques on various benchmark standard data sets collected from UCI repository are shown in in table 10.

Table 10. Performance of the RST based ACO feature selection technique

Data Set	C4.5 DT Classifier			VPRSM based DT Classifier		
	Un Discretized	Discretized	Reduced	Un Discretized	Discretized	Reduced
Chronic Kidney	95.25	95.25	95.25	95.45	97.25	***97.75***
Diabetes	73.82	78.25	78.25	78.23	81.23	81.23
Vertebral Column	81.61	81.61	81.61	83.21	86.77	86.77

V. CONCLUSION

In this paper, the Rough Set based ACO approach with heuristic information is used for selecting a subset of features from the datasets taken from UCI repository. On applying the feature selection technique ACO, the Chronic kidney disease data set with ***24*** attributes are reduced to ***10*** attributes, and for Diabetes dataset number of features reduced from ***8*** to ***6*** and there is no reduction of features for the Vertebral column disease dataset. When submitting these reduced features to C4.5 and VPRSM based Decision tree classification models the observed accuracies have been increased. The experimental results proved that the performance of the RST based ACO for feature selection and VPRSM based decision tree classification technique outperformed the traditional methods for classification.

VI. REFERENCE

[1] L.Ladha and T.Deepa, Feature Selection Methods and Algorithms, International Journal on Computer Science and Engineering (IJCSE), vol.3, no.5, pp. 1787-1797,2011.

[2] Z.Pawlak, Rough sets, International Journal of Computer and Information Science11, 341-356, 1982.

[3] Yumin Chen, Duoqian Miao, Ruizhi Wang, A rough set approach to feature selection based on ant colony optimization : Department of Computer Science and Technology, Tongji University, Shanghai 201804, PR China.

[4] Maryam Bahojb Imani, Tahereh Pourhabibi, Mohammad Reza Keyvanpour, and Reza Azmi A New Feature Selection Method Based on Ant Colony and Genetic Algorithm on Persian Font Recognition, International Journal of Machine Learning and Computing, Vol. 2, No. 3, June 2012

[5] Chen, Ling, Bolun Chen, and Yixin Chen. "Image feature selection based on ant colony optimization." AI 2011: Advances in Artificial Intelligence. Springer Berlin Heidelberg, 2011. 580-589.

[6] Jensen, Richard, and Qiang Shen. "Finding rough set reducts with ant colony optimization." Proceedings of the 2003 UK workshop on computational intelligence. Vol. 1. No. 2. 2003.

[7] Jin-Mao Wei, Ming-Yang Wang, Jun-Ping You, VPRSM Based Decision Tree Classifier, Institute of Computational Intelligence, Key Laboratory for Applied Statistics of MOE, Northeast Normal University Changchun, Jilin 130024, P.R. China.

[8] Z. Pawlak, "On rough dependency of attributes in information systems", Bulletin of the Polish Academy of Sciences, vol.33, pp. 551–599,1985.

[9] Z.Pawlak, "Rough Sets. Theoretical Aspects of Reasoning about Data," Kluwer Academic Publications, Netherlands,1995.

[10] Z.Pawlak, "Rough Set approach to knowledge-based decision support," European Journal of Operational Research, 99, pp. 48-57,1997.

[11] J.R.Quinlan , C4.5: Programs for Machine Learning, Morgan Kaufmann, 1993.

[12] J Wei , Rough Set based approach to selection of node, International Journal of Computational Cognition 1(2), 25-40, 2003.

[13] W.Ziarko, Variable Precision Rough set model, Journal of Computer and System Sciences46 (1), 39-59, 1993.

[14] Zengtai Gong , Variable Precision Rough Set Model For Incomplete Information Systems And Its β-Reducts , Computing and Informatics, Vol. 31, 1385–1399, 2012

[15] Jin-Mao Wei, Shu-Qin Wang, Ming-Yang Wang, Jun-Ping You, Da-You Liu, Rough set based approach for inducing decision trees –Institute of Computational Intelligence, Northeast Normal University, Changchun,



Jilin130024,PRChina,
[16] Son, C.S., "Decision-making model for early diagnosis of congestive heart failure using rough set and decision tree approaches", Journal of Biomedical Informatics, vol.45, pp.999-1008,2012.
[17] http://archive.ics.uci.edu/ml/datasets

IJEAST

INTERNATIONAL JOURNAL
OF ENGINEERING APPLIED SCIENCE AND TECHNOLOGY

AUTHORS

Surekha Samsani

Dr. G. Jaya Suma

ABOUT IJEAST

International journal of Engineering Applied Science and Technology (IJEAST) is a peer-reviewed, open access journal that publishes high - quality research paper in the field of Engineering, Applied Science and Technology.

IJEAST aims to provide a platform for researchers, academicians, and professionals to share their innovative ideas, research findings, and practical experiences with the global scientific community.

JOURNAL METRICS

H-INDEX	i10-INDEX	IJEAST CITATIONS	PEER-REVIEWED & OPEN ACCESS
33	154	9,647 (IJEAST CITATIONS)	PUBLISHED MONTHLY

PEER REVIEWED

All submission are rigorously peer reviewed to ensure quality.

AUTHORS ARE OUR PRIORITY

We value our authors and recognize their contribution to the advancement of research and knowledge.

OPEN ACCESS

Free and unrestricted access to research for all.

GLOBAL REACH

Connecting researchers and professionals worldwide.

TIMELY PUBLICATION

We ensure a swift and efficient publication process.

For more information, visit our website
www.ijeast.com

editor@ijeast.com

www.ijeast.com

India

2455-2143