International Journal of Research Studies in Computer Science and Engineering (IJRSCSE)
Volume 11, Issue 1, 2025, PP 29-36
ISSN 2349-4840 (Print) & ISSN 2349-4859 (Online)
DOI: https://doi.org/10.20431/2349-4859.1101004
www.arcjournals.org

Multi-Objective Feature Selection for Distributed Systems in Three Stages

Malepati Venkata Sai Chaitrika Chowdary*, N Naresh, Vunnava Dinesh Babu

Department of CSE, R V Institute of Technology, Guntur

***Corresponding Author:** ***Malepati Venkata Sai Chaitrika Chowdary**, Department of CSE, R V Institute of Technology, Guntur*

Abstract: *Data mining researchers must deal with the challenges of large-scale data analysis in the modern environment. Due to the strategic and operational limitations of learning methodologies, big data management is extremely challenging. Distributed systems are a state-of-the-art method for analyzing large volumes of data. Large data sets could be handled by distributed learning, which could result in reliable and efficient solutions. Distributed learning refers to multi-node learning techniques that are highly efficient, more precise, and scalable to larger input data quantities. By simultaneously distributing tasks over multiple devices, it enables users to greatly increase their productivity.In this paper uses a three-stage multi-objective feature selection (TMFS) method to choose the best features and boost the effectiveness of distributed systems. The TMFS technique employs the correlation coefficient (CC), Fisher score (FS), information gain (IG), mean absolute deviation (MAD), and min-max normalisation (MMN) among five feature selection strategies in three steps. The TMFS method reduces the dataset size and the classification error rate efficiently.*

Keywords: *Clustering, Orchestra Clustering, Distributed systems, multi-objective feature selection*

1. INTRODUCTION

This paper focuses on tackling the difficulties of classifying large amounts of data and investigating practical methods to complete the task. The two phases of classification are training and testing (Yousafzai, Hayat, and Afzal, 2020). During the training phase, a classification algorithm uses class labels to identify particular patterns in training data. During the testing phase, a trained classification algorithm predicts a class label for a test dataset. Additionally, big data is a vast collection that continues to grow quickly over time. The magnitude and complexity of this data made it impossible for any of the traditional data processing tools to store or handle.The magnitude of Big Data is a factor in how long it takes the classifier to categorise it and how much memory it uses. Distributed classifier training can potentially process massive data rapidly and effectively.

The scalability and efficiency limitations of learning approaches make handling large volumes of data extremely difficult (Thudumu, Branch, Jin, and Singh, 2020). For example, when the computational difficulty exceeds the primary memory, the method cannot scale efficiently because of memory constraints. Distributed learning techniques could be used to scale and handle large data sets. Because distributed learning techniques can divide learning operations across multiple processors, they could make processing large data sets easier. Distributed learning makes use of conventional machine learning techniques. A unique ensemble machine and deep learning approaches are required to improve the precision of distributed learning.

Ensemble strategies employ many classification techniques to improve predictive accuracy over a single classification strategy. Therefore, the distributed ensemble machine and deep learning technique were suggested in this paper.

According to Kumar, Bashir, Rashid, and Kharel (2021), machine-generated datasets have become larger in recent years. Massive amounts of information that exceed the capabilities of traditional database management systems are comprised of vast quantities of data. Big data is pervasive because modern data-intensive systems can handle a variety of data sources and formats (Erraissi and Belangour, 2018).

The necessity for effective partitioning strategies that meet demands for memory use, processing speed, and implementation time has increased as datasets continue to grow. The problem with large data is that

object grouping increases the comparability of data from different groups. Big data is used in practically every business (Desai, 2018). Big data can be divided into three categories: somewhat organized, disorganized, and organized. Any information that is saved, accessed, and handled according to a predetermined format is considered organized data.

Computer science expertise has improved over time in its ability to put these ideas into practise.

Any data with an unclear structure or form is considered unorganized data. Unorganized data must overcome a number of obstacles in order to be analyzed and its value retrieved because of its enormous bulk. An appropriate example would be a heterogeneous data set consisting of a combination of text documents, photos, and video files. Even though organisations today have a massive data, they are unsure how to utilize it since it is unorganized. 'Semi-organized' data contains both kinds of data.

Big data mining is a technique used to gather and analyze vast amounts of information to extract essential insights. Different data mining procedures are frequently used to process big data. But, the dimensionality could not be reduced using the procedures. Hence, clustering is a vital data mining technique for performing big data analytics. Toalleviate the effects of dimensionality, associated data points are grouped through clustering (Amutha and Sharma, 2021).

The technique of grouping related objects together is called clustering. With tiny datasets, conventional clustering approaches are efficient. Every object inside a cluster ought to be identical to every other object within the cluster when clustering huge datasets (Sreedhar, Kasiviswanath and Chenna Reddy, 2017). The capacity to autonomously group identical objects enables the detection of unseen patterns and critical aspects when a big volume of data is sorted into categories. Users are capable of comprehending a lot of data due to it.

Five categories of clustering methods could be employed to group data into clusters. These include algorithms based on partitions, hierarchies, densities, grids, and models (Oyelade, Isewon, Oladipupo, Emebo, Omogbadegun, Aromolaran and Olawole, 2019). Numerous methods of each category have been effectively used in actual data mining scenarios (Ahmad and Khan, 2019).

Due to its simplicity in learning and execution and low temporal complexity compared to other approaches, partition-based approaches are unquestionably the most well-known and often employed approaches (Ansari, Afzal and Sardar, 2019). A dataset is split into k parts representing a cluster in partition-based clustering approaches. Starting with a collection of information items, a typical data clustering technique splits them into k groups using similarity measures like Euclidean distance. The succeeding few conditions could be satisfied by partition-based clusters: At least one data item should be present in each cluster, and each item can only belong to one cluster (Niroomand, Bach and Elser, 2021).

2. Related Work

Five phases of the comprehensive system for managing extensive datasets are portrayed in Figure 1.1.

The first stage is gathering the Higgs boson dataset from the Kaggle data repository (Kaggle.com, 2022). This dataset includes 33 features, including a target feature and 2 50 000 instances. This paper, for the sake of simplicity, uses 10% of the Higgs boson dataset to assess a huge dataset processing system. It has 25,000 instances and 33 features, one of which is a target feature.

The second stage involves clustering, a popular method for partitioning any dataset. Employing the combination of K-Means, K-Medoids, Fuzzy C-means, Expectation-Maximization (EM), and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) approaches, this phase presents an ensemble partition-based clustering with a majority voting technique for partitioning huge datasets. Initially, the substantial dataset is segmented separately using these five clustering techniques. Subsequently, a majority voting technique is employed to select the ultimate clusters among the five clustering methods. The data instances are allocated to the cluster that garnered the highest number of votes among these five clustering algorithms.

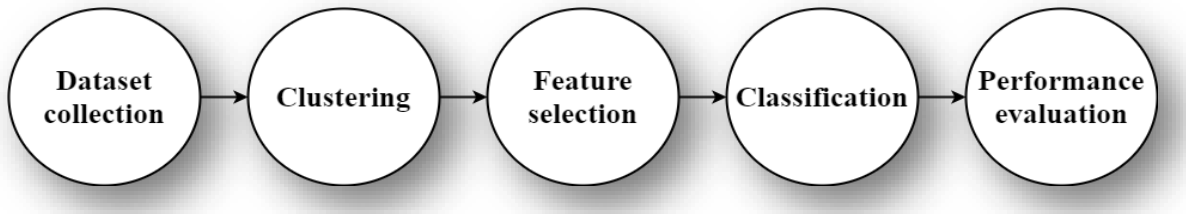


Figure 1.1. *Overview of the processing of huge datasets*

In this paper, feature selection, is employed to identify the optimal features for creating usable models. The performance of distributed systems is improved in this stage by introducing a novel three-stage multi-objective feature selection (TMFS). The TMFS algorithm utilizes the correlation coefficient, Fisher score, information gain, absolute deviation, and min-max normalisation as five feature selection strategies at three levels. They help in decreasing the error rate in classification, minimizing the number of features, and condense the dataset.

3. PROPOSED METHOD

A huge dataset has a high degree of dimension. A high-dimensional dataset must be classified in a very difficult and time-consuming manner. Additionally, it offers less accuracy. a feature selection strategy is required to boost the effectiveness of the classification algorithm via eliminating redundant features and decreasing the dataset dimensionality. However, the accuracy of currently used feature selection techniques is lower. This chapter suggests a new three-stage multi-objective feature selection (TMFS) method to provide greater accuracy than current feature selection strategies.

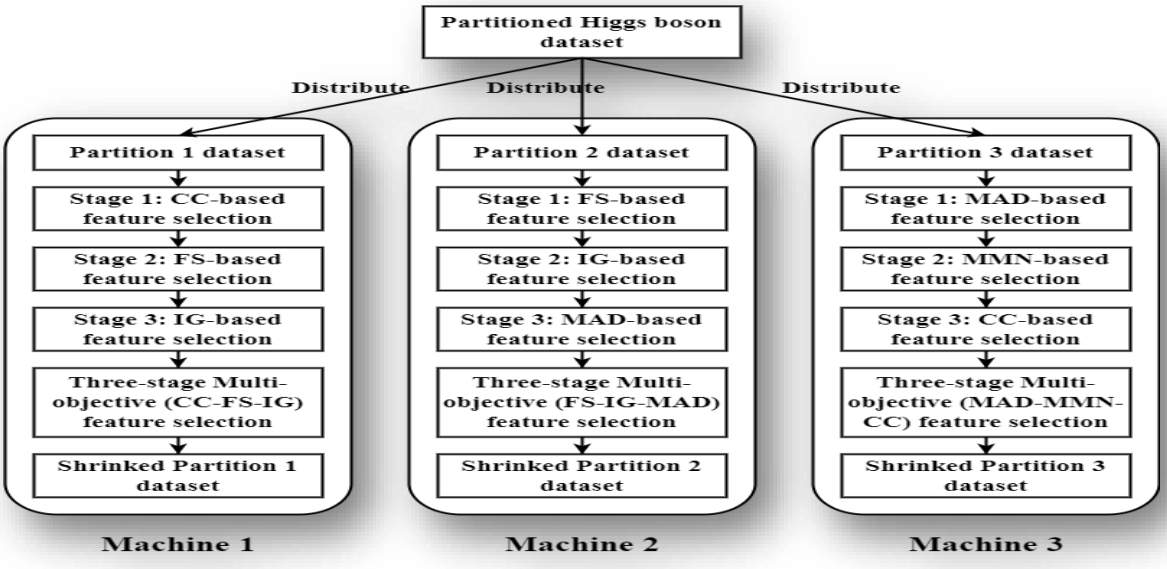
3.1 Feature Selection in Distributed Systems

Researchers working in data mining should overcome the challenges of big data analysis. Therefore, they rely on distributed systems, which split enormous datasets into smaller ones and distribute them among many processing devices. Distributing one device's large data processing load across multiple devices reduces processing time. But, on the other hand, the learning approach requires more time and space because the dataset has many redundant and dimensional features. Therefore, it will make the learning strategy less effective. Therefore, a feature selection strategy is crucial for choosing the required features while removing unwanted ones (El-Hasnony, Barakat, Elhoseny, and Mostafa, 2020). But, the majority of feature selection methods choose various feature subsets for a similar data set. Therefore, the accuracy of the classification may change as a result. To deal with this problem, this chapter proposed a new three-stage multi-objective feature selection (TMFS) method to boost the effectiveness of distributed systems.

3.2 Three-Stage Multi-Objective Feature Selection for Distributed Ensemble Machine and Deep Learning

Large datasets from the real world have a variety of features in their information representation. However, only a few of these might be pertinent to the target feature. Many features may be unrelated to the target feature; this section suggests an ensemble-based feature selection technique to choose a fixed feature set relating to the target feature that increases classification accuracy.

This section created a novel three-stage multi-objective feature selection (TMFS) technique to boost distributed learning efficiency. The suggested TMFS technique combines many subsets of features to choose the optimal subset. The TMFS methodutilizes five feature selection methods at three levels: Min-Max Normalization (MMN), Information Gain, Fisher Score, Mean Absolute Deviation, and Correlation Coefficient.



To enhance efficiency, the dataset needs to be divided into smaller pieces and distributed across numerous machines. Initially, the dataset is clustered using ensemble partition-based clustering with a majority vote. Then, the dataset is partitioned into K parts utilizing this approach (with K set to 3 in this instance). Each partition is subsequently distributed to individual machines. Finally, the TMFS algorithm is used on each machine to select the most appropriate features. The flowchart of the proposed TMFS method is illustrated in Figure 3.1.

4. Experimental Results

4.1 Dataset Description

The proposed TMFS technique was evaluated on the Higgs boson dataset, which is accessible through the Kaggle data repository (Kaggle.com, 2022). The dataset contains 33 features and 250,000 events. To simplify the evaluation process, we utilized only 10% of the Higgs boson dataset. However, there are 25,000 events in it. Two labels in the target class served as a unique identifier for each event: signal = 1; background = 0.

4.2 Statistical Metrics

Accuracy:

The proportion of accurate forecasts to all instances is known as accuracy (AC), which is shown in Eq. (4.1):

$$\mathbf{AC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{4.1}$$

In this context, TP denotes the count of instances accurately classified as 0, FP denotes the count of instances incorrectly classified as 0, TN denotes the count of instances accurately classified as 1, and FN denotes the count of instances incorrectly classified as 1.

Sensitivity:

Sensitivity (SE) is determined by its capacity to accurately predict 0 events (TP cases) which are showed Eq. (4.2):

$$\boldsymbol{SE} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{4.2}$$

Specificity:

Specificity (SP) is determined by its capacity to accurately predict 1 event (TN cases), which is shown in Eq. (4.3):

$$\boldsymbol{SP} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \tag{4.3}$$

Precision:

Positive predictive value is called the precision (PR), which is depicted in Eq. (4.4):

$$\boldsymbol{PR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{4.4}$$

4.1 Accuracy

Table4.1. Presents a comparison of the accuracy of various feature selection techniques for partition 1 of the Higgs boson dataset on machine 1.

Table4.1. *Comparing the accuracy of several feature selection techniques for Higgs boson Dataset Partition 1 on Machine 1*

Techniques	Accuracy
CC	0.837
FS	0.878
IG	0.756
TMFS (CC-FS-IG)	0.911

Additionally, Figure 4.1 displays the accuracy comparison. This comparison shows that the best feature selection technique is TMFS (CC-FS-IG).

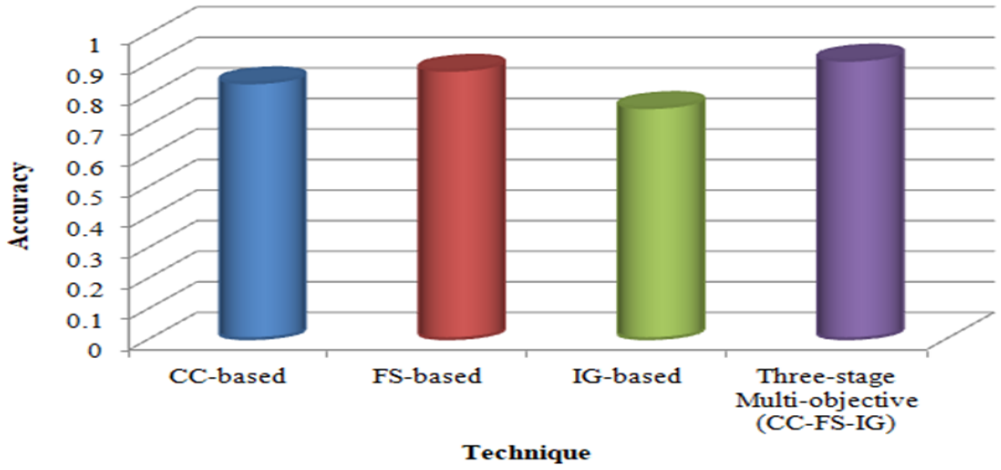


Figure4.1. *The comparison of accuracy among different feature selection methods for Partition 1 of the Higgs boson dataset on Machine 1*

The accuracy of the IG technique is one of the major drawbacks. The CC technique, however, offers the maximum accuracy compared to the IG technique. On the other hand, the FS technique offers the most accuracy compared to the CC technique. However, the TMFS (CC-FS-IG) technique provides a very high accuracy compared to the FS technique.

Table 4.2 presents a comparison of the accuracy of different feature selection techniques for Partition 2 of the Higgs boson dataset on Machine 2.

Table4.2. *Comparing the accuracy of several feature selection techniques for Higgs boson Dataset Partition 2 on Machine 2*

Techniques	Accuracy
FS	0.899
IG	0.844
MAD	0.769
TMFS (FS-IG-MAD)	0.916

Additionally, Figure 4.2 displays the accuracy comparison. This comparison shows that the best feature selection technique is TMFS (FS-IG-MAD).

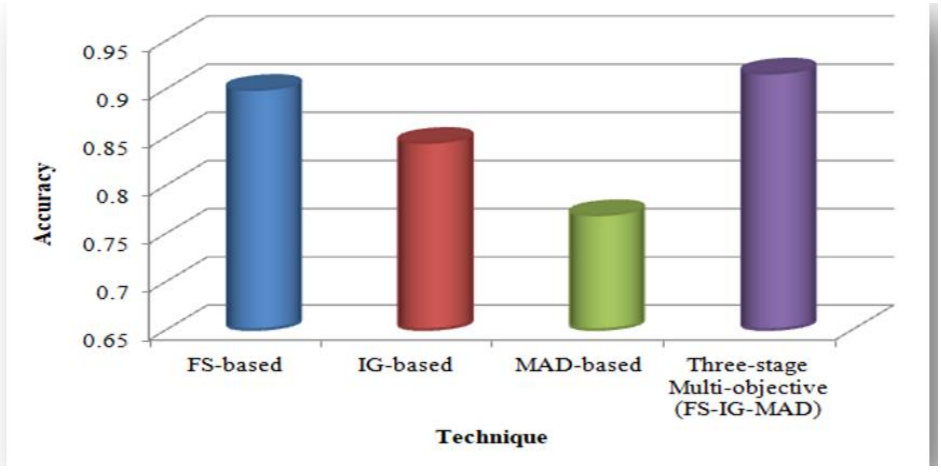


Figure4.2. *The comparison of accuracy among different feature selection methods for Partition 2 of the Higgs boson dataset on Machine 2*

Among other techniques, the MAD technique's accuracy is noticeably worse. However, the IG technique offers the maximum accuracy compared to the MAD technique. However, the FS technique is more accurate than the IG technique. But when compared to the FS technique, the TMFS (FS-IG-MAD) technique provides a high level of accuracy.

Table 4.3 compares the accuracy of different feature selection techniques for partition 3 of the Higgs boson dataset on machine 3.

Table4.3. *Comparing the accuracy of several feature selection techniques for Higgs boson Dataset Partition 3 on Machine 3*

Technique	Accuracy
MAD	0.702
MMN	0.74
CC	0.773
TMFS (MAD-MMN-CC)	0.842

Additionally, Figure 4.3 displays the accuracy comparison. This comparison shows that the best feature selection technique is TMFS (MAD-MMN-CC).

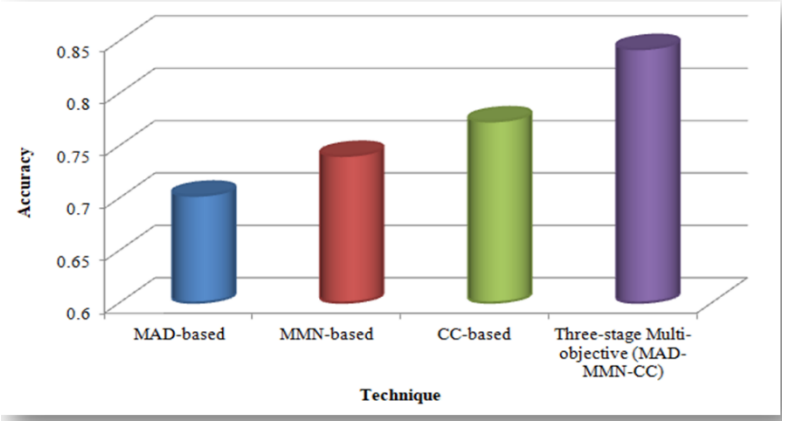


Figure4.3. *Comparing the accuracy of different feature selection techniques for partition 3 of the Higgs boson dataset on machine 3*

Among other techniques, the MAD technique's accuracy is noticeably worse. However, the MMN technique offers the maximum accuracy compared to the MAD technique. However, the CC technique offers the maximum accuracy compared to the MMN technique. However, the TMFS (MAD-MMN-CC) technique provides significantly higher accuracy when compared to the CC technique.

5. CONCLUSION

Partition-based clustering is the partitioning method that is most frequently and extensively employed. However, due to big datasets' explosive growth, existing clustering algorithms are inadequate to extract knowledge from huge datasets. Conventional proposed ensemble partition-based clustering using the majority voting technique for huge dataset partitioning using K-Means, K-Medoids, Fuzzy C-means, EM, and DBSCAN algorithms to cluster such large datasets. Accuracy and execution time were the primary measures to measure the efficiency of the ensemble clustering method. The experimental findings demonstrated that, compared to the other five clustering strategies, the suggested ensemble clustering technique delivered the greatest accuracy and clustered data more quickly.

Each partition in the distributed system is distributed to each machine by the DL-BPF framework. Each machine chooses the best features once it has received the partition. Techniques for feature selection are typically used to improve classifier performance. The accuracy of the results varied at different levels because different feature selection methods selected different feature subsets. To address this issue and improve the efficiency of distributed systems, a new three-stage multi-objective feature selection (TMFS) method was introduced in this paper. The TMFS technique chooses the best feature subset by integrating various feature subsets. The experimental results showed that the proposed TMFS method reduces the size of the dataset and the number of features while achieving the highest levels of accuracy, sensitivity, specificity, and precision.

REFERENCES

[1] Madhuri, C. R., Jandhyala, S. S., Ravuri, D. M., & Babu, V. D. (2024). Accurate classification of forest fires in aerial images using ensemble model. Bulletin of Electrical Engineering and Informatics, 13(4), 2650–2658. https://doi.org/10.11591/eei.v13i4.6527

[2] Venugopal, N. L. V., Sneha, A., Babu, V. D., Swetha, G., Banerjee, S. K., & Lakshmanarao, A. (2024). A Hybrid Model for Heart Disease Prediction using K-Means Clustering and Semi supervised Label Propagation. 2024 3rd International Conference for Advancement in Technology, ICONAT 2024. https://doi.org/10.1109/ICONAT61936.2024.10774787

[3] Venugopal, N. L. V., Sneha, A., Babu, V. D., Swetha, G., Banerjee, S. K., & Lakshmanarao, A. (2024). A Hybrid Model for Heart Disease Prediction using K-Means Clustering and Semi supervised Label Propagation. 2024 3rd International Conference for Advancement in Technology, ICONAT 2024. https://doi.org/10.1109/ICONAT61936.2024.10774787

[4] Babu, V. D., & Malathi, K. (2023). Large dataset partitioning using ensemble partition-based clustering with majority voting technique. Indonesian Journal of Electrical Engineering and Computer Science, 29(2), 838–844. https://doi.org/10.11591/ijeecs.v29.i2.pp838-844

[5] Kavya, K., Sree, R., Dinesh Babu, V., Vullam, N., Lagadapati, Y., & Lakshmanarao, A. (n.d.). Integrated CNN and Recurrent Neural Network Model for Phishing Website Detection.

[6] Babu, V. D., & Malathi, K. (2023). Three-stage multi-objective feature selection for distributed systems. Soft Computing. https://doi.org/10.1007/s00500-023-07865-y

[7] Babu, V. D., & Malathi, K. (2023). Three-stage multi-objective feature selection for distributed systems. Soft Computing. https://doi.org/10.1007/s00500-023-07865-y

[8] Vunnava, D. B., Popuri, R. B., Daruvuri, R. K., & Anusha, B. (2023). An Automated Epilepsy Seizure Detection System (AESD) Using Deep Learning Models. International Conference on Self Sustainable Artificial Intelligence Systems, ICSSAS 2023 - Proceedings, 454–461. https://doi.org/10.1109/ICSSAS57918.2023.10331731

[9] Ashok, D., Nirmala, N. M. V., Srilatha, D., Rao, K. V., Babu, V. D., & Basha, S. J. (2023). Leveraging CNN and LSTM for Identifying Citrus Leaf Disorders. 2nd International Conference on Automation, Computing and Renewable Systems, ICACRS 2023 - Proceedings, 730–735. https://doi.org/10.1109/ICACRS58579.2023.10404123

[10] Babu, V. D., & Malathi, K. (2023). Three-stage multi-objective feature selection with distributed ensemble machine and deep learning for processing of complex and large datasets. Measurement: Sensors, 28. https://doi.org/10.1016/j.measen.2023.100820

[11] C. N. Phaneendra, P. Rajesh, C. M. Kumar, V. A. Koushik and K. K. Naik, "Design of Single Band Concentric Square Ring Patch Antenna for MIMO Applications," 2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), Mandya, India, 2022, pp. 1-5, doi: 10.1109/ICERECT56837.2022.10060285.

[12] C. N. Phaneendra, K. V. V. Ram, D. Naveen, L. Sreekar and K. K. Naik, "Design a Multi-Band MIMO Patch Antenna at X, K, and Ku Band for Wireless Applications," 2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), Mandya, India, 2022, pp. 1-6, doi: 10.1109/ICERECT56837.2022.10060667.

[13] K. K. Naik, V. Lavanya, B. J. Reddy, M. Madhuri and C. N. Phaneendra, "Design of Sloted T-Shape MIMO Antenna at X-Band for 5G and IoT Applications," 2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), Mandya, India, 2022, pp. 1-6, doi: 10.1109/ICERECT56837.2022.10060565.

[14] Jagadeeswari, C., Naga, S. G., & Dinesh Babu, V. (2020). Statistical Analysis Proving COVID-19's Lethalty Rate for the Elderly People-Using R. International Journal of Advanced Science and Technology, 29(11s), 1366–1370.

[15] Roja, D., & Dinesh Babu, V. (2018). A Survey on Distributed Denial-of-Service Flooding Attacks with Path Identifiers (Vol. 3, Issue 11). www.ijrecs.com

[16] Shini, S., Gudise, D., Dinesh, V., & Bu, B. A. (n.d.). International Journal of Research Availa bl e Detect Malevolent Account In Interpersonal Union. https://edupediapublications.org/journals/index.php/IJR/

[17] V. Jyothsna, B. N. Madhuri, K. S. Lakshmi, K. Himaja, B. Naveen and K. D. Royal, "Facemask detection using Deep Learning," 2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS), Coimbatore, India, 2023, pp. 533-537, doi: 10.1109/ICISCoIS56541.2023.10100472.

[18] J. S. Shankar and M. M. Latha, "Troubleshooting SIP Environments," 2007 10th IFIP/IEEE International Symposium on Integrated Network Management, Munich, Germany, 2007, pp. 601-611, doi: 10.1109/INM.2007.374823.

[19] S. Velan et al., "Dual-Band EBG Integrated Monopole Antenna Deploying Fractal Geometry for Wearable Applications," in IEEE Antennas and Wireless Propagation Letters, vol. 14, pp. 249-252, 2015, doi: 10.1109/LAWP.2014.2360710.

[20] S. Holm, T. M. Pukkila and P. R. Krishnaiah, "Comments on "On the use of autoregressive order determination criteria in univariate white noise tests" (reply and further comments)," in IEEE Transactions

on Acoustics, Speech, and Signal Processing, vol. 38, no. 10, pp. 1805-1806, Oct. 1990, doi: 10.1109/29.60113.

[21] Z. . -D. Bai, P. R. Krishnaiah and L. . -C. Zhao, "On rates of convergence of efficient detection criteria in signal processing with white noise," in IEEE Transactions on Information Theory, vol. 35, no. 2, pp. 380-388, March 1989, doi: 10.1109/18.32132.

Citation: *Malepati Venkata Sai Chaitrika Chowdary (2025).“ Multi-Objective Feature Selection for Distributed Systems in Three Stages”. International Journal of Research Studies in Computer Science and Engineering (IJRSCSE), vol 11, no. 1, 2025, pp. 29-36. DOI: https://doi.org/10.20431/2349-4859.1101004.*