International Journal of Data Informatics and Intelligent computing (IJDIIC)
Vol. 1, No. 2, December 2022, pp. 16~22
DOI: 10.59461/ijdiic.v1i2.17

Clustering Techniques for Recommendation of Movies

Mahesh T R[1], V Vinoth Kumar[1]
[1]Department of Computer Science and Engineering, Jain (Deemed-to-be University), Bangalore, India

Article Info

Article history:

Received September 14, 2022
Revised October 04, 2022
Accepted November 06, 2022

Keywords:

Collaborative filtering
Recommendation Algorithm
Recommender system
Principle Component Analysis
Hierarchical Clustering
Big Data

ABSTRACT

A recommendation system employs a variety of algorithms to provide users with recommendations of any kind. The most well-known technique, collaborative filtering, involves users with similar preferences although it is not always as effective when dealing with large amounts of data. Improvements to this approach are required as the dataset size increases. Here, in our suggested method, we combine a hierarchical clustering methodology with a collaborative filtering algorithm for making recommendations. Additionally, the Principle Component Analysis (PCA) method is used to condense the dimensions of the data to improve the accuracy of the outcomes. The dataset will receive additional benefits from the clustering technique when using hierarchical clustering, and the PCA will help redefine the dataset by reducing its dimensionality as needed. The primary elements utilized for recommendations can be enhanced by applying the key elements of these two strategies to the conventional collaborative filtering recommendation algorithm. The suggested method will unquestionably improve the precision of the findings received from the conventional CFRA and significantly increase the effectiveness of the recommendation system. The total findings will be applied to the combined dataset of TMDB and Movie Lens, which is utilized to suggest movies to the user in accordance with the rating patterns that each individual user has generated.

Corresponding Author:

Mahesh T R
Department of Computer Science and Engineering
Jain (Deemed-to-be University)
Bangalore
India
Email: trmahesh.1978@gmail.com

1. INTRODUCTION

Big data, which is generated by internet users and is referred to as such, is a crucial area for research. It is crucial to conduct more study and offer solutions for the issues in the current systems because technology and the services offered to users are always evolving [1-3]. It is a challenging effort to produce appropriate recommendations; hence almost every organization belonging to any industry needs an effective method for its users to submit recommendations [4-6].

At a very complex information platform, a recommendation system is a means to offer consumers suggestions based on the anticipated user preferences. The three filtering algorithms that are utilized to create any recommendation system are content-based filtering, collaborative filtering, and hybrid filtering [7-9]. The collaborative filtering recommendation algorithm (CFRA), which is more well-known and has been utilized by large big data producers and consumers like eBay, Amazon, and Facebook, is the algorithm that is most frequently used among these three methods. Collaborative filtering bases its recommendations on user

ratings and preferences that are similar to those of the user for whom they are being created. Numerous studies have demonstrated that adding K means clustering features to the conventional CFRA will increase its effectiveness. To further increase this effectiveness, we are now using hierarchical clustering in place of the conventional K means clustering [10-12]. Additionally, it is demonstrated that hierarchical clustering is more effective than K means because it does not require an initial cluster count.

Additionally, it has been demonstrated that applying more PCA on this will increase the recommendation's accuracy. It should be emphasized that the PCA should be used before applying the clustering technique because the reduced number of dimensions will make clustering simpler [13-16].

2. RELATED WORK

Authors have suggested a suggestion algorithm that is more effective than the conventional collaborative filtering technique. To enhance the conventional CFRA, they applied two extra methods, K means clustering and PCA [17-19]. They put forth two techniques, one of which combines K means clustering with PCA and the other with just K means clustering. Compared to the conventional CFRA, both algorithms performed admirably. The experiment made use of the Netflix dataset. The Frechet distance is the first measure of similarity of AIS trajectories developed by the authors. Second, to determine the number of clusters at the final level of clustering, the distance matrix acquired in the first stage is decomposed using principal component analysis (PCA). Finally, the conventional hierarchical clustering process is combined with the aforementioned distance matrix and clustering number.

The author of this study put up novel techniques for agglomerative hierarchical clustering that automatically mention the number of clusters [20-22]. The author of this study put up novel techniques for agglomerative hierarchical clustering that automatically mention the number of clusters. Here, they presented two methods for achieving this goal: using a variant of the cluster validity measure and using a statistical model selection method like BIC. The authors attempted to address the issue of choosing the right number of clusters through this. A formal approach for counting clusters has been described by the authors and is based on the agglomerative hierarchical clustering (AHC) algorithm. The novel index and method can calculate the AHC-generated clustering results and establish the ideal number of clusters that can be used to various dataset types, including linear, manifold, annular, and convex structures [23-25].

The K-means strategy to clustering and how the choice of main seeding impacts the outcome are both topics covered by the authors in this study. In order to compare with a data collection, hierarchical methods are utilized as a baseline. For a pure numerical synthetic data set, the authors have assessed the k-means clustering and hierarchical clustering algorithms. The several methods of clustering that are regularly utilized on the internet have been described by authors. Here, the numerous types of hierarchical clustering are also presented, along with a detailed explanation of hierarchical clustering itself. Additionally, a comparison of the clustering methods is completed. Three factors were used by the authors to gauge the quality of clusters: cohesion measurement, silhouette index, and elapsed time [6]. Six techniques were listed by all the authors in this research as potential ways that collaborative filtering recommender systems could learn more about some new users. In these methods, a set of items is chosen so that the collaborative filtering system can present it to each subsequent new user for rating purposes.

A strategy put out by the authors to fix the collaborative filtering method's scalability and sparsity issues. They utilized a novel CF model that is based on the Artificial Immune Network Algorithm in an effort to tackle both issues at once (aiNet). The rationale for choosing this is that by specifying the data structure, including their spatial distribution and cluster interrelations, aiNet is possible to reduce sparsity and offer the feature of scalability of the dataset. Additionally, they have decreased the sparsity rate and used the k-means clustering technique using aiNET.

Researchers on the profitability of online sales shared their findings. Typically, suggestions are produced based on the preferences and interests of the user, without taking the seller's profit into account. They suggested the CPPRS (Convenience plus Profitability Perspective Recommender System) and HPRS as two profitability-based recommender systems to address this (Hybrid Perspective Recommender System). The suggested technique is successfully put into practice. The authors built a recommendation system called WebPUM, an online prediction that makes use of a Web usage mining system, and they all suggested a method for categorizing a user's navigational patterns in order to forecast that user's future goals. Authors have provided a straightforward analysis of the Amazon website. The expansion of Amazon's recommendation algorithm has been discussed. It has also been demonstrated that Amazon employs collaborative filtering to provide accurate recommendations. They researched the Amazon for two decades before offering any recommendations.

The sparsity and scalability issues, which are the two main drawbacks of collaborative filtering, were addressed by the authors. They have suggested a brand-new technique for solving these issues called the neighbor user method, which builds on the subspace clustering strategy. This approach is effective because the authors identify various subspaces of the objects classified as Interested, Neither Interested, Nor

Uninterested, nor Uninterested, and Uninterested. The suggested method was evaluated using the MovieLens 100K, MovieLens 1M, and Jester datasets in order to compare it to more established methods. The findings show that the suggested strategy can improve the Recommender Systems' operational capacity. In unsupervised learning tasks like PCA and K-means clustering, they have proposed a compression technique for large-scale data sets that results in both computational speed and memory gains. They employed randomized preconditioning transformation to do this, which is the key component of their approach and makes it possible to use it in streaming and distributed data environments.

An overview of numerous distinct machine-learning algorithms employed in the context of big data analytics is provided in a paper by the authors. They made an effort to close the knowledge gap amongst researchers by providing an overview of a few distinct machine learning algorithms in order to make it evident to researchers. The paper gives a straightforward analysis of big data analytics, with a special emphasis on distributed machine learning techniques for large data that are data-intensive. This review is purely theoretical.

This work, which deals with the similarity measure technique, was presented by the author. In this case, he used various similarity measurements to verify their effectiveness. The performance of the various similarity metrics used in collaborative filtering is examined in this research in order to compare them. They employed Apache Mahout for this, which made accurate analysis possible in a time-effective manner. Authors used the Hadoop framework to analyze data that was available on the internet. They used the Hadoop framework to create suggestions for customers based on user ratings, likes, and reviews. They first applied the data using the Mahout interface after filtering it. This entire project was carried out for a movielens dataset-based movie recommendation.

The MyMediaLite library/API, which offers support for the algorithms, has been examined by the authors in this. This library deals with collaborative filtering for two types of data: prediction of ratings (for example, on a scale of 1 to 5 stars) and prediction of items from implicitly positive feedback alone. They also discussed some potential future development on this library. The writers provided a recommendation based on keyword knowledge that utilizes the keywords. The investigation made advantage of collaborative filtering, which is accomplished in Hadoop. In this experiment, real-world data is used, and by KSAR, the recommendation accuracy is much enhanced.

3. PROPOSED METHODOLOGY

Since the traditional collaborative filtering recommendation algorithm employs a formal way of filtering, which renders it ineffective when used alone, it lacks accuracy and efficiency. It is possible to draw the conclusion that the algorithm still requires significant improvement based on the recommendations offered by the collaborative filtering algorithm. Traditional collaborative filtering recommendation algorithms are less accurate when used, which renders them ineffective when applied to large datasets, or Big Data. Dealing with enormous data results in reduced accuracy, which is unsuitable. The algorithm will not be effective for providing recommendations to consumers if it is applied to a large amount of data in real-world applications. The amount of attributes that are available is entirely taken into account when information is extracted to recommend products to users, rendering the collaborative filtering recommendation algorithm ineffective. Additionally, more comparisons must be conducted and take more time to compute the more attributes that are used to produce suggestions. The overall dimensions that were considered when generating recommendations should be removed if necessary .

In addition, a different clustering technique can be used to replace the k-means clustering that was previously used with the collaborative filtering algorithm. The k-means clustering method has various downsides, but these flaws could be overcome by switching to a more recent clustering method. K-means clustering is inefficient to utilize if the numbers of the clusters are not properly defined since the numbers of the clusters that should be made need to be defined at the beginning of the process. The accuracy decreases and the time needed to provide recommendations to the user increases as the number of dimensions used to evaluate the results increases. Therefore, one of the main tasks is to lower the dataset's dimensionality or amount of attributes.

By substituting more modern techniques for the older ones, the issue with the prior algorithm can be fixed. Similar to before, the approach combines the PCA dimensionality reduction technique with the K-means clustering technique. The authors have previously suggested combining two of these strategies in the collaborative filtering algorithm. While preserving the PCA as it was previously employed, we have now provided a better clustering method in comparison to the k-means clustering. Since hierarchical clustering is a superior clustering approach to work with, k-means clustering can be replaced. The data's dimensionality will be reduced using the PCA as a dimensionality reduction technique.

Given that there is no requirement to specify the number of clusters at the outset of the clustering in hierarchical clustering, the results will be superior to those of k-means clustering. It will be possible to divide the clusters according to the dataset after using hierarchical clustering and defining the necessary number of

clusters. But the dataset needs to be enhanced before the clustering algorithm is applied. The output of the algorithm will be more effectively produced if the input is accurate. Therefore, dimensionality reduction should be performed on the input dataset to improve it, and to do this, PCA must be applied to the dataset. In conclusion, we will do PCA on the dataset before providing it as input, and once the principle components have been obtained, this data is provided as input to hierarchical clustering. Prior to applying hierarchical clustering and making final recommendations, the collaborative filtering method will first perform the PCA. Therefore, the recommendations can be made more accurate and the collaborative filtering algorithm can be enhanced in this way. The proposed method flowchart is shown in figure 1.

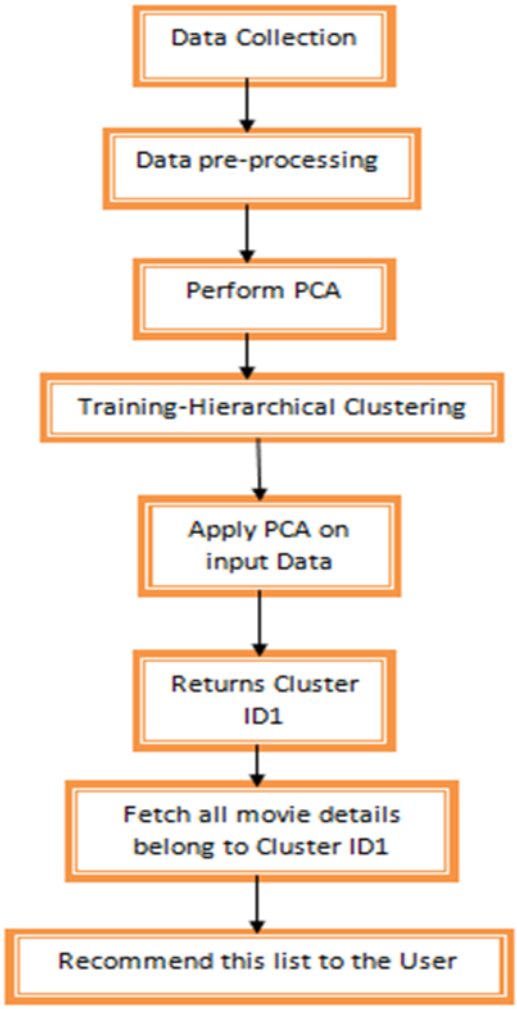

Figure 1. Flowchart of the proposed method

The proposed algorithm works as shown below.

Algorithm 1. Proposed Algorithm

Step 1: Data collection - collect the movie related data like name, rating etc. in the form of csv file.

Step 2: Data pre-processing - perform manual data analysis and eliminate the feature which is less correlate to other feature.

Step 3: Perform PCA (principal component analysis) on the data and save the data into csv file.

Step 4: Define hierarchical clustering (agglomerative) model.

Step 5: Train the hierarchical clustering (agglomerative) model on the data.

Step 6: Take the one user input and apply PCA on that.

Step 7: Perform the prediction in the input it give the cluster id.

Step 8: Fetch all the movie detail which belong to this cluster id and make the list of it.

4. EXPERIMENTAL RESULTS

We used the TMDB dataset and the MovieLens Dataset to conduct the analysis of the suggested approach. Movie information is derived from the TMDB dataset, and user information and rating trends are combined from the MovieLens dataset. For the sake of analysis, both datasets are combined. There are scores available from 1 to 5. The experiment was run to assess the reliability of the recommendations generated by the algorithm we suggested in our study. In this experiment, the accuracy term is determined in order to compare the suggested and the existing algorithms.

As a result of applying this data to the prior pca and k-means collaborative method, we are able to calculate the accuracy of the earlier approach. The accuracy is computed as shown in equation (1).

$$\text{Accuracy} = \frac{(\{\text{Relevent Documents}\}\text{Intersection }\{\text{Retrieved Documents}\})}{\{\text{Relevent Documents}} *100 \quad (1)$$

The suggested approach that uses hierarchical clustering is now being examined. Over the same data, the collaborative filtering technique, pca, and hierarchical clustering are examined. The accuracy of this algorithm is compared to the one that already exists.

The experiment demonstrates a clear improvement in the precision of the recommendations generated by our suggested method. The following graph compares the accuracy of the results for both approaches utilizing k-means clustering with pca and hierarchical clustering with pca (2).

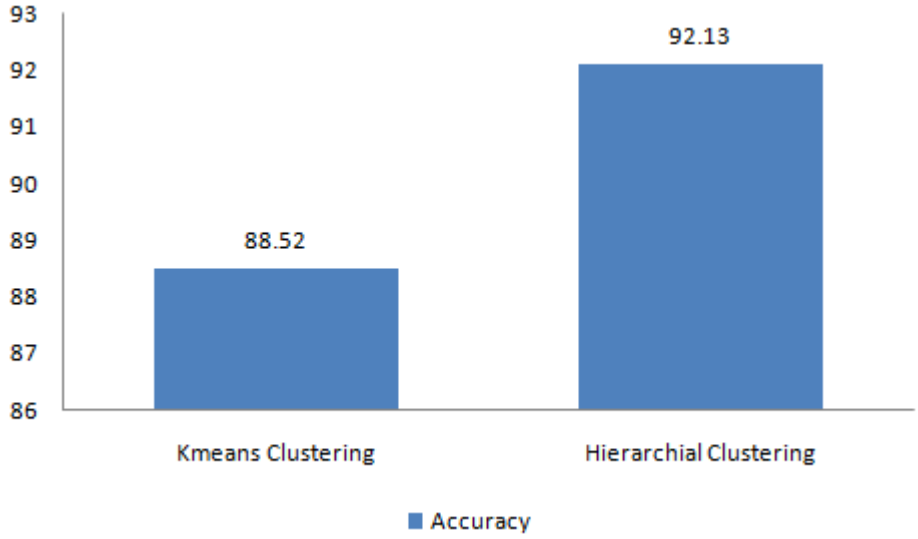

Figure 2. Accuracy of different algorithms

The suggested hierarchical clustering clearly outperforms the k-means clustering that was previously in use, according to figure 2. The accuracy of the suggested algorithm's output is better than that of the older clustering method. Therefore, compared to the preceding technique, it is preferable to utilize Hierarchical clustering with PCA.

5. CONCLUSION

The proposed research study looks at the user recommendations that the system makes. The hierarchical clustering method and PCA are used throughout the entire project to assess the system's accuracy. The intersection of the suggested movies and the user-rated movies from before is used to assess the system's accuracy. The experiment demonstrates improved outcomes over the preceding algorithms.

REFERENCES

[1] Mahesh, T.R., Ram, M.S., Ram, N.S.S., Gowtham, A., Swamy, T.V.N. (2022). Real-Time Eye Blinking for Password Authentication. In: García Márquez, F.P. (eds) International Conference on Intelligent Emerging Methods of Artificial Intelligence & Cloud Computing. IEMAICLOUD 2021. Smart Innovation, Systems and Technologies, vol 273. Springer, Cham. https://doi.org/10.1007/978-3-030-92905-3_52

[2] Mahesh, T.R., Krishna, G.V., Sathwik, P., Chowdary, V.A., Hemchand, G. (2022). Providing Voice to Susceptible Children: Depression and Anxiety Detected with the Help of Machine Learning. In: García Márquez, F.P. (eds) International Conference on Intelligent Emerging Methods of Artificial Intelligence & Cloud Computing. IEMAICLOUD 2021. Smart Innovation, Systems and Technologies, vol 273. Springer, Cham. https://doi.org/10.1007/978-3-030-92905-3_5

[3] Mounica, Yasaswi, Vandana, & Sai Joshna. (2022). Melanoma classification using deep transfer learning. International Journal of Data Informatics and Intelligent Computing, 1(1), 11–20. https://doi.org/10.5281/zenodo.7101199

[4] N. Thangarasu, R. Rajalakshmi, G. Manivasagam, & V. Vijayalakshmi. (2022). Performance of re-ranking techniques used for recommendation method to the user CF- Model. International Journal of Data Informatics and Intelligent Computing, 1(1), 30–38. https://doi.org/10.5281/zenodo.7108931.

[5] V. Vivek; T. R. Mahesh; C. Saravanan; K. Vinay Kumar, "A Novel Technique for User Decision Prediction and Assistance Using Machine Learning and NLP: A Model to Transform the E-commerce System," in Big Data Management in Sensing: Applications in AI and IoT, River Publishers, 2021, pp.61-76.

[6] S. Roopashree, J. Anitha, T.R. Mahesh, V. Vinoth Kumar, Wattana Viriyasitavat, Amandeep Kaur, An IoT based Authentication System for Therapeutic Herbs measured by Local Descriptors using Machine Learning Approach, Measurement, 2022,111484, ISSN 0263-2241, https://doi.org/10.1016/j.measurement.2022.111484

[7] T. R. Mahesh, V. Dhilip Kumar, V. Vinoth Kumar, Junaid Asghar, Banchigize Mekcha Bazezew, Rajesh Natarajan V Vivek, " Blended Ensemble Learning Prediction Model for Strengthening Diagnosis and Treatment of Chronic Diabetes Disease", Computational Intelligence and Neuroscience, vol. 2022, Article ID 4451792, 9 pages, 2022. https://doi.org/10.1155/2022/4451792

[8] T. R. Mahesh, V. Dhilip Kumar, V. Vinoth Kumar, Junaid Asghar, Oana Geman, G. Arulkumaran, N. Arun, "AdaBoost Ensemble Methods Using K-Fold Cross Validation for Survivability with the Early Detection of Heart Disease", Computational Intelligence and Neuroscience, vol. 2022, Article ID 9005278, 11 pages, 2022. https://doi.org/10.1155/2022/9005278

[9] Karthick Raghunath K. M, V. Vinoth Kumar, V.Muthukumaran, krishna kant singh, Mahesh T R, Akansha Singh, "DETECTION AND CLASSIFICATION OF CYBER ATTACKS USING XGBOOST REGRESSION AND INCEPTION V4", Journal of Web Engineering, RIVER PUBLISHERS, Vol 21, Issue 4, 2022 https://doi.org/10.13052/jwe1540-9589.21413

[10] Mahesh, T.R., Vinoth Kumar, V., Vivek, V. et al. Early predictive model for breast cancer classification using blended ensemble learning. Int J Syst Assur Eng Manag (2022). https://doi.org/10.1007/s13198-022-01696-0

[11] A. Srivastava, V. V. Kumar, M. T. R and V. Vivek, "Automated Prediction of Liver Disease using Machine Learning (ML) Algorithms," 2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), 2022, pp. 1-4, doi: 10.1109/ICAECT54875.2022.9808059

[12] T. R. Mahesh, V. Vivek, V. V. Kumar, R. Natarajan, S. Sathya and S. Kanimozhi, "A Comparative Performance Analysis of Machine Learning Approaches for the Early Prediction of Diabetes Disease," 2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), 2022, pp. 1-6, doi: 10.1109/ACCAI53970.2022.9752543

[13] S. Surana, K. Pathak, M. Gagnani, V. Shrivastava, M. T. R and S. Madhuri G, "Text Extraction and Detection from Images using Machine Learning Techniques: A Research Review," 2022 International Conference on Electronics and Renewable Systems (ICEARS), 2022, pp. 1201-1207, doi: 10.1109/ICEARS53579.2022.9752274.

[14] D. S. Chaithanya, K. L. Narayana and M. T R, "A Comprehensive Analysis: Classification Techniques for Educational Data mining," 2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON), 2021, pp. 173-176, doi: 10.1109/CENTCON52345.2021.9688070

[15] M. R. Sarveshvar, A. Gogoi, A. K. Chaubey, S. Rohit and T. R. Mahesh, "Performance of different Machine Learning Techniques for the Prediction of Heart Diseases," 2021 International Conference on Forensics, Analytics, Big Data, Security (FABS), 2021, pp. 1-4, doi: 10.1109/FABS52071.2021.9702566

[16] P. Shrestha, A. Singh, R. Garg, I. Sarraf, T. R. Mahesh and G. Sindhu Madhuri, "Early Stage Detection of Scoliosis Using Machine Learning Algorithms," 2021 International Conference on Forensics, Analytics, Big Data, Security (FABS), 2021, pp. 1-4, doi: 10.1109/FABS52071.2021.9702699

[17] A. Srivastava, V. V. Kumar, M. T. R and V. Vivek, "Automated Prediction of Liver Disease using Machine Learning (ML) Algorithms," 2022 Second International Conference on Advances in Electrical, Computing,

Communication and Sustainable Technologies (ICAECT), 2022, pp. 1-4, doi: 10.1109/ICAECT54875.2022.9808059

[18] V, K., Dr Savita Chaudhary, & Radhika A D. (2022). Feature Extraction in Music information retrival using Machine Learning Algorithms. International Journal of Data Informatics and Intelligent Computing, 1(1), 1–10. https://doi.org/10.5281/zenodo.7093881

[19] S. Surana, K. Pathak, M. Gagnani, V. Shrivastava, M. T. R and S. Madhuri G, "Text Extraction and Detection from Images using Machine Learning Techniques: A Research Review," 2022 International Conference on Electronics and Renewable Systems (ICEARS), 2022, pp. 1201-1207, doi: 10.1109/ICEARS53579.2022.9752274

[20] D. S. Chaithanya, K. L. Narayana and M. T R, "A Comprehensive Analysis: Classification Techniques for Educational Data mining," 2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON), 2021, pp. 173-176, doi: 10.1109/CENTCON52345.2021.9688070

[21] S. R. A, M. R, R. N, S. L and A. N, "Survey on Malicious URL Detection Techniques," 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), 2022, pp. 778-781, doi: 10.1109/ICOEI53556.2022.9777221.

[22] P. Shrestha, A. Singh, R. Garg, I. Sarraf, T. R. Mahesh and G. Sindhu Madhuri, "Early Stage Detection of Scoliosis Using Machine Learning Algorithms," 2021 International Conference on Forensics, Analytics, Big Data, Security (FABS), 2021, pp. 1-4, doi: 10.1109/FABS52071.2021.9702699

[23] Rajesh, N., Selvakumar, A.A.L. Association rules and deep learning for cryptographic algorithm in privacy preserving data mining. Cluster Comput 22 (Suppl 1), 119–131 (2019). https://doi.org/10.1007/s10586-018-1827-6

[24] P. Chaitanya Reddy, R. M. S. Chandra, P. Vadiraj, M. Ayyappa Reddy, T. R. Mahesh and G. Sindhu Madhuri, "Detection of Plant Leaf-based Diseases Using Machine Learning Approach," 2021 IEEE International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), 2021, pp. 1-4, doi: 10.1109/CSITSS54238.2021.9683020

[25] K. K. Jha, R. Jha, A. K. Jha, M. A. M. Hassan, S. K. Yadav and T. Mahesh, "A Brief Comparison On Machine Learning Algorithms Based On Various Applications: A Comprehensive Survey," 2021 IEEE International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), 2021, pp. 1-5, doi: 10.1109/CSITSS54238.2021.9683524

BIOGRAPHIES OF AUTHORS

T. R. Mahesh is serving as Associate Professor and Program Head in the Department of Computer Science and Engineering at Faculty of Engineering and Technology, JAIN (Deemed-to-be University), Bengaluru, India. Dr. Mahesh has to his credit more than 40 research papers in Scopus and SCIE indexed journals of high repute. He has been the editor for books on emerging and new age technologies with publishers like Springer, IGI Global, Wiley etc. Dr. Mahesh has served as reviewer and technical committee member for multiple conferences and journals of high reputation. His research areas include image processing, machine learning, Deep Learning, Artificial Intelligence, IoT and Data Science. He can be contacted at email: trmahesh.1978@gmail.com

V. Vinoth Kumar is an Associate Professor at Department of Computer Science, JAIN (Deemed-to-be University), Bangalore, India. His current research interests include Wireless Networks, Internet of Things, machine learning and Big Data Applications. He is the author/co-author of papers in international journals and conferences including SCI indexed papers. He has published as over than 35 papers in IEEE Access, Springer, Elsevier, IGI Global, Emerald etc. He is the Associate Editor of International Journal of e-Collaboration (IJeC), International Journal of Pervasive Computing and Communications (IJPCC) and Editorial member of various journals. He can be contacted at email: pvkumar243@gmail.com