JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Segmenting Customers Using K-Means Clustering

M. Anusha Rani[1]*, Danda Bindu Sree[2] , Pakalapati Himasri[3], Shaik Najiya[4] , Nimma Gayathri[5]
[1*]Assistant Professor: Vignan's Nirula Institute of Technology and Science for Women.
[2,3,4,5] B.Tech Scholar: Vignan's Nirula Institute of Technology and Science for Women.

Abstract

To perform well in future company needs good revenue and that can be achieved by effective decisions. There is much competition going on nowadays and every company have their unique ways of near payoff. Decision using data is the correct way. Everyone is diverse and we cannot predict what their purchasing or taste preferences are. Effective decisions are mandatory for any company to generate good revenue. However, we can use machine learning to organize and process the dataset then apply different algorithms to it in order to identify its target behavior. Without this, it is a lot difficult and there are no best methods available to locate the alike natured and group of people in a huge dataset. So, it will help to buyer assignment using classification of K-Means clustering to group the same data together. Data analysis technique can help is by revealing these customer segments. Clustering algorithms applied to customer level data provide businesses with a way of organizing their strategy around these rich customer insights. This work embodies the real life value of clustering in getting to know how the customer segments which enables companies to channel their resources appropriately, build loyal customers and grow their revenue.

Key Words: K-Means, un-supervised learning, segmentation, cluster

1. Introduction

In the recent years, [1-5] as business grows competition also grew and increased data that supported this sophisticated mechanisms of Mining where useful information can be extracted from the available database means collections of big data were manipulated using Data mining.

Data mining is the process where methods are applied to extract data patterns in order to present it in the human readable format which can be used for the purpose of decision support. Customer segmentation is described as a crucial strategy for [5-8] businesses to target specific customer groups, improve product offerings, and enhance customer relationships.

Customer Segmentation is the process of division of customer base into several groups called as customer segments such that each customer segment consists of customers who have similar characteristics. The segmentation is based on the similarity [8-10] in different ways that are relevant to marketing such as gender, age, interests, and miscellaneous spending habits.

The customer segmentation has the importance as it includes, the ability to modify the programs of market so that it is suitable to each of the customer segment, support in business decision; identification of products associated with each customer segment and to manage the demand and supply of that product; identifying and targeting the potential customer base, and predicting customer defection, providing directions in finding the solutions.

First [10-17] words has been used for data mining applied methods to extract the patterns from data (information) and convert in human readable format so it will work as decision support. Customer segmentation is mentioned multiple times as a critical approach to help businesses better pivot their products, services and develop stronger customer relationships.

The introduction also touches upon the rapid growth of the market, the shift to online platforms, and the generation of large amounts of customer data, emphasizing the benefits of utilizing customer segmentation to achieve company goals and increase profits. The algorithm is used when you have unlabeled data (i.e. data without defined categories or groups).

2. Literature Survey

Aman Banduni, Prof Ilavedhan A, [1] in studies [18-25] customer segmentation using machine learning. In this paper, they explained the concept of customer segmentation.

Customer Segmentation using K-Means Algorithm is based on the Analysing such data is an important need. In the modern era of innovation, where there is a large competition to be better then everyone, the business strategy needs to be according [26-30] to the modern conditions (**YashKushwaha, Deepak kumarSrinivastva ,2020**) [2].

"Customer Segmentation Based on Purchasing [32-35] Behavior Using K-Means Clustering" by **M.A.Hossain and Mohammad Shorif Uddin (2016)** [3]. This research paper focuses on a practical application of clustering, specifically K-means clustering, for customer segmentation based on purchasing behavior. The study demonstrates the effectiveness of clustering [36-42] in identifying distinct customer groups and proposes strategies to target these segments. The findings suggest that personalized marketing approaches derived from clustering analysis can significantly improve sales.

D. P. Yash Kushwaha, Deepak Prajapati [4] in studied [43-47] customer segmentation in detail and also studied in detail about k-means clustering algorithm and performed customer segmentation using K-means clustering algorithm and observed the clusters formed and compared the results with the other clustering algorithms.

Concept decompositions [48-51] for large sparse text data using clustering, is to analyse high dimensionality of text can be a deterrent in applying complex learners such as Support Vector Machines to the task of text classification(**S. Dhillon and D. M. Modha ,2001**) [5].

Juni Norma Sari, Ride Dedriana, Lukito Nugroho, Paulus Insap Santosa [6] in reviewed all customer segmentation techniques.

Ina Maryani; Dwiza Riana; [52] Rachmawati Darma Astuti; Ahmad Ishaq; Sutrisno; Eva Argarini Pratama [7] in studied different clustering techniques.

"Enhancing Customer Segmentation with Machine Learning" by **Abhijit J. Patil and Prashant R. Nair(2021)** [8]. This paper discusses [53-56] the integration of machine learning techniques, including clustering algorithms, for customer segmentation. It emphasizes the benefits of utilizing advanced algorithms to uncover complex patterns in customer data. The study showcases real-world examples of businesses that have successfully improved sales by adopting machine learning-driven segmentation strategies.

In these days competition is huge [57] and all companies are moving forward with their own different strategies. We should use data and take a proper decision. Every person is different from one another and we don't know what he/she buys or what their likes are (**D. Aloise, A. Deshpande, P. Hansen, and P. Popat,2009**) [9].

Shi Na; Liu Xumin; Guan Yong [10] in studied [58] in detail about k means clustering algorithm and observed its pros and cons.

3. Proposed Methodology

In the topic of Segmenting Customers using K-Mean Clustering, we have choosen the model of K-Mean Algorithm and the model is explained as follows:

The process for customer segmentation analysis with clustering is a number of steps that transform raw customer data into an output designed to provide you sales-enhancing insights. The major reason of the technique is Clustering, that helps in forming groups or segments of customers who share something in common as they move towards customers with individualised marketing communications. Let me review the approach for using clustering to drive your sales with customer segmentation.

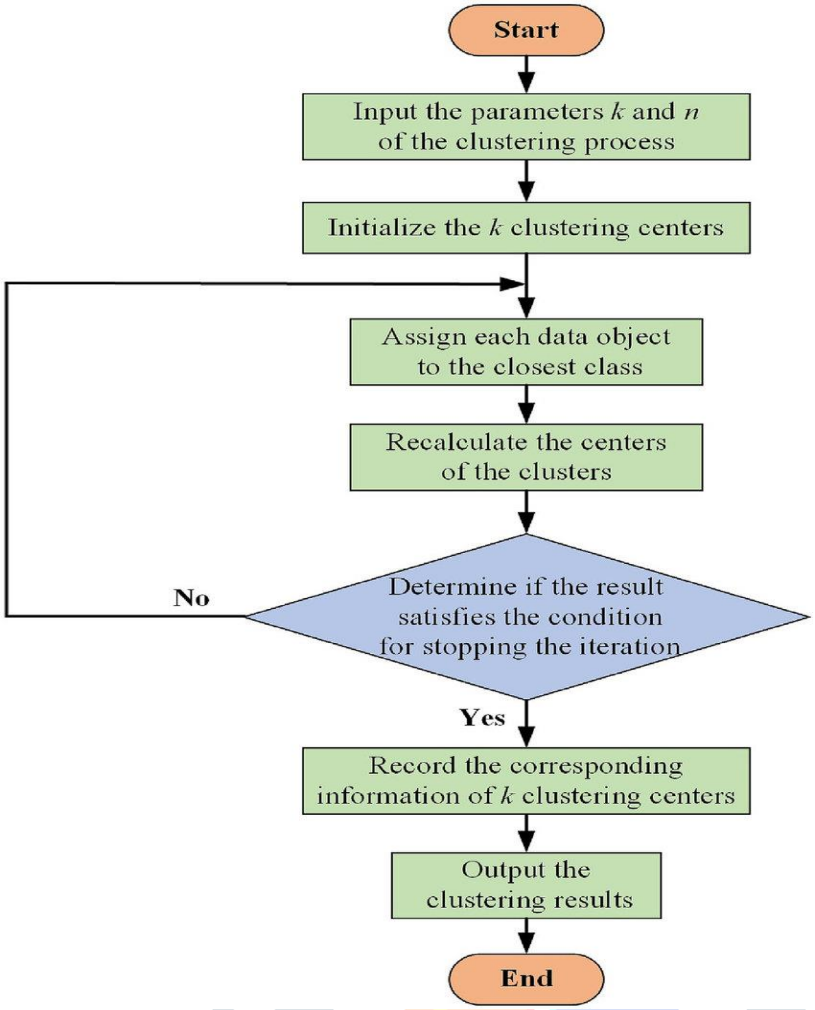

Fig. 1.1. Architecture of K-Mean Algorithm

Step-1: Select the number K to decide the number of clusters.
Step-2: Select random K points or centroids. (It can be other from the input dataset).
Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.
Step-4: Calculate the variance and place a new centroid of each cluster.
Step-5: Repeat the third steps, which means reassigning each datapoint to the new closest centroid of each cluster.
Step-6: If any reassignment occurs, then go to step-4 else go to FINISH.
Step-7: The model is ready.

Formula:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\left\| x_i - v_j \right\| \right)^2$$

where,

'||x_i - v_j||' is the Euclidean distance between x_i and v_j.
'c_i' is the number of data points in i^{th} cluster.
'c' is the number of cluster centers.

4. Results

4.1. Visualize age of customers

The image shows a histogram plot generated using R, which provides a graphical representation of the distribution of the Age variable from the customer_data dataset. The histogram reveals how customer ages are distributed across different age classes. The highest frequency (30) is observed in the age class between 30 and 35. The lowest frequency is in the 50-55 age class. Most age classes have around 10–25 individuals. This kind of histogram helps in understanding the distribution of age within the dataset, identifying the most common age groups, and seeing if the data is skewed toward younger or older individuals.

```
hist(customer_data$Age,
     col="blue",
     main="Histogram to Show Count of Age Class",
     xlab="Age Class",
     ylab="Frequency",
     labels=TRUE)
```

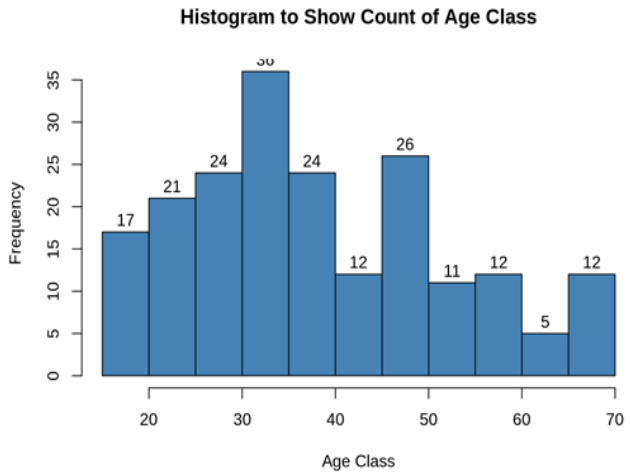

Fig. 1.2. Histogram to show count of age class

4.2. Visualize the gender of customers

A bar plot is being used to compare the count of males and females in a dataset. This creates a frequency table of the Gender variable from the customer_data dataset. It counts how many males and females are present in the dataset and stores these values. The red bar represents the count of females, and the cyan bar represents the count of males.

```
a=table(customer_data$Gender)
barplot(a,main="Using BarPlot to display Gender Comparision",
        ylab="Count",
        xlab="Gender",
        col=rainbow(2),
        legend=rownames(a))
```

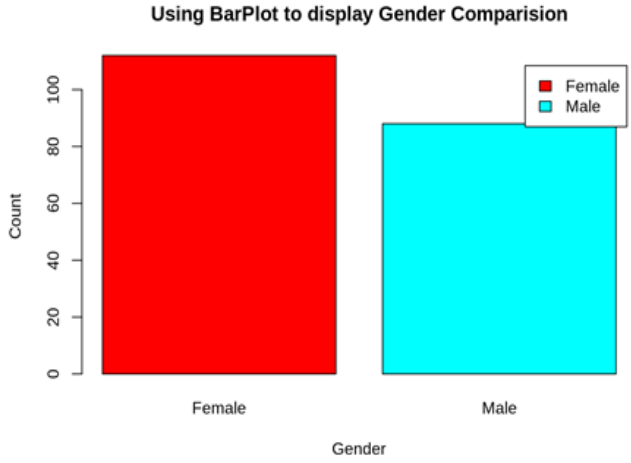

Fig. 1.3. Using Bar Plot to display Gender comparison

5. Conclusion

This customer segmentation analysis using clustering can be an important sales improvement that can lead to a beneficial increase in business growth. This is been to spread the information that how important and needed to understand wide ranging customer requirements and behaviours through the use of cluster techniques in order to execute effective marketing strategies. The results show how tailored strategies and resource allocation can be optimized by focusing on customer segments. Although this machine learning application is very useful in the market, a manager can pay full attention to all clusters that he have found and adjust these requirements as it was shown in this study with client segmentation in shopping malls. To meet the customers, the mall managers have to know what its customers must be needing. study their consumption patterns, and create seamless customer experiences that feel familiar to meet their needs.

6. References

[1] Aman Banduni, Prof Ilavedhan A, "Customer Segmentation using machine learning," School of Computing Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India.

[2] MuhammadSufyan, Deepak KumarSrivastava, "Understanding the customers' shopping experience on online buying","Journel of retailing and consumer service",vol.68,pp.111-123,2022.

[3] Hossain, M. A., & Uddin, M. S. (2016). Customer Segmentation Based on Purchasing Behavior Using K-Means Clustering. *International Journal of Computer Applications*, 149(1), 14-18.

[4] D. P. Yash Kushwaha, Deepak Prajapati, "Customer Segmentation using K- Means Algorithm," 8th Semester Student of Beach in Computer Science and Engineering, Galgotias University, India.

[5] I. S. Dhillon and D. M. Modha, "Concept decompositions for large sparse text data using clustering," Machine Learning, vol. 42, issue 1, pp. 143-175, 2001.

[6] Juni Nurma Sari,Ridi Ferdiana,Lukito Nugroho,Paulus Insap Santosa,"Review on Customer Segmentation Technique",Department of Electrical Engineering and Information Technology, University of Gadjah Mada, Jogjakarta, Indonesia, Department of Informatics Technology, Polytechnic Caltex Riau, Pekanbaru, Indonesia.

[7] Maryani, Ina et al. "Customer Segmentation based on RFM model and Clustering Techniques with K-Means Algorithm." 2018 Third International Conference on Informatics and Computing (ICIC) (2018): 1-6.

[8] Patil, A. J., & Nair, P. R.(2021). Enhancing Customer Segmentation with Machine Learning. *Journal of Business Research*, 134, 139-149.

[9] D. Aloise, A. Deshpande, P. Hansen, and P. Popat, "The Basis Of Market Segmentation" Euclidean sum-of-squares clustering," Machine Learning, vol. 75, pp. 245-249, 2009.

[10] S. Na, L. Xumin and G. Yong, "Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm," 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, 2010, pp. 63-67, DOI: 10.1109/IITSI.2010.74.

[11] Patibandla, R. S. M. L., Rao, B. T., Narayana, V. L., & Srinivas, V. S. (2021). An overview of ontology-based artificial intelligence services in health care systems. In *Proceedings of the International Conference on Health Care Systems* (pp. 47–63).

[12] Maddumala, V. R., Maha Lakshmi, K., Anusha, P., & Lakshman Narayana, V. (2020). Enhanced morphological operations for improving the pixel intensity level. *Journal of Computational and Theoretical Nanoscience*, 29(3), 9191–9201.

[13] Narayana, V. L., Malleswari, K. S. N., Divyanjali, M., Nandini, S., & Purnima, G. (2023). Video frame based prompt compression model with steganography for secure data transmission. In *Proceedings of the International Conference on Advanced Intelligent Systems* (pp. 373–377). https://doi.org/10.1109/ICAIS56108.2023.10073883

[14] Naresh, A., Pavani, V., Meghana Chowdary, M., & Lakshman Narayana, V. (2020). Energy consumption reduction in cloud environment by balancing cloud user load. *Journal of Critical Reviews*, 7(7), 1003–1010. https://doi.org/10.31838/jcr.07.07.184

[15] Patibandla, R. S. M. L., & Vejendla, L. N. (2022). Significance of blockchain technologies in industry. In *Advances in Blockchain Technologies* (pp. 19–31). https://doi.org/10.1007/978-3-030-70501-5_2

[16] Pasala, S., Pavani, V., Lakshmi, G. V., & Narayana, V. L. (2020). Identification of attackers using blockchain transactions using cryptography methods. *Journal of Critical Reviews*, 7(6), 368–375. https://doi.org/10.31838/jcr.07.06.65

[17] Mounika, B., Anusha, P., Narayana, V. L., & Lakshmi, G. V. (2020). Use of blockchain technology in providing security during data sharing. *Journal of Critical Reviews*, 7(6), 338–343. https://doi.org/10.31838/jcr.07.06.59

[18] Narayana, V. L., Vinayaki, K. V., Swetha, P. A., Sri, K. D., & Chaithanya, G. (2024). Superior attribute weighted set for object skeleton detection using ResNet50 with edge-based segmentation model. In *Proceedings of the International Conference on Smart Computing and Systems* (pp. 1132–1139). IEEE. https://doi.org/10.1109/ICSCSS60660.2024.10624879

[19] Gopi, A. P., & Naik, K. J. (2022). An IoT model for fish breeding analysis with water quality data of pond using modified multilayer perceptron model. *2022 International Conference on Data Analytics for Business and Industry (ICDABI)*, 448-453. https://doi.org/10.1109/ICDABI56818.2022.10041617

[20] Arepalli, P. G., & Naik, K. J. (2024). A deep learning-enabled IoT framework for early hypoxia detection in aqua water using lightweight spatially shared attention-LSTM network. *Journal of Supercomputing*, *80*(2), 2718-2747. https://doi.org/10.1007/s11227-023-05580-x

[21] Arepalli, P. G., & Naik, K. J. (2023). An IoT-based water contamination analysis for aquaculture using lightweight multi-headed GRU model. *Environmental Monitoring and Assessment*, *195*(12), Article 1516. https://doi.org/10.1007/s10661-023-12126-4

[22] Gopi, A. P., Gowthami, M., Srujana, T., Gnana Padmini, S., & Durga Malleswari, M. (2023). Classification of denial-of-service attacks in IoT networks using AlexNet. In *Smart Innovation, Systems and Technologies* (Vol. 316, pp. 349-357). https://doi.org/10.1007/978-981-19-5403-0_30

[23] Bikku, T., Gopi, A. P., & Prasanna, R. L. (2019). Swarming the high-dimensional datasets using ensemble classification algorithm. In *Advances in Intelligent Systems and Computing* (Vol. 815, pp. 583-591). https://doi.org/10.1007/978-981-13-1580-0_56

[24] Arepalli, P. G., & Khetavath, J. N. (2024). Water quality classification using multi-cell RNN in aquaculture ponds for Catla fish. In *Lecture Notes in Networks and Systems* (Vol. 897, pp. 363-370). https://doi.org/10.1007/978-981-99-9704-6_34

[25] Arepalli, P. G., & Naik, K. J. (2024). Water contamination analysis in IoT-enabled aquaculture using deep learning-based AODEGRU. *Ecological Informatics, 79*, Article 102405. https://doi.org/10.1016/j.ecoinf.2023.102405

[26] Arepalli, P. G., & Naik, K. J. (2024). An IoT-based smart water quality assessment framework for aqua-ponds management using Dilated Spatial-temporal Convolution Neural Network (DSTCNN). Aquacultural Engineering, 104, Article 102373. https://doi.org/10.1016/j.aquaeng.2023.102373

[27] Gopi, A. P., Narayana, V. L., & Kumar, N. A. (2018). Dynamic load balancing for client-server assignment in distributed systems using genetic algorithm. Ingenierie des Systemes d'Information, 23(6), 87-98. https://doi.org/10.3166/ISI.23.6.87-98

[28] Sarada, K., Narayana, V. L., Gopi, A. P., & Pavani, V. (2020). An iterative group based anomaly detection method for secure data communication in networks. Journal of Critical Reviews, 7(6), 208-212. https://doi.org/10.31838/jcr.07.06.39

[29] Narayana, V. L., Gopi, A. P., & Chaitanya, K. (2019). Avoiding interoperability and delay in healthcare monitoring system using blockchain technology. Revue d'Intelligence Artificielle, 33(1), 45-48. https://doi.org/10.18280/ria.330108

[30] Gopi, A. P., Jyothi, R. N. S., Narayana, V. L., & Sandeep, K. S. (2023). Classification of tweets data based on polarity using improved RBF kernel of SVM. International Journal of Information Technology, 15(2), 965-980. https://doi.org/10.1007/s41870-019-00409-4

[31] Narayana, V. L., Gopi, A. P., Khadherbhi, S. R., & Pavani, V. (2020). Accurate identification and detection of outliers in networks using group random forest methodology. Journal of Critical Reviews, 7(6), 381-384. https://doi.org/10.31838/jcr.07.06.67

[32] Rao, B. T., Patibandla, R. S. M. L., Narayana, V. L., & Gopi, A. P. (2021). Medical data supervised learning ontologies for accurate data analysis. In Semantic Web for Effective Healthcare Systems (pp. 249-267). https://doi.org/10.1002/9781119764175.ch11

[33] Patibandla, R. S. M. L., Gopi, A. P., Narayana, V. L., & Rao, B. T. (2023). Decentralized smart healthcare systems using blockchain and AI. In Blockchain applications in healthcare: Innovations and practices (Vol. 1, pp. 139-154). DOI: 10.1002/9781394229512.ch8

[34] Lakshman Narayana, V., & Gopi, A. P. (2020). Enterotoxigenic Escherichia coli detection using the design of a biosensor. Journal of New Materials for Electrochemical Systems, 23(3), 164-166. DOI: 10.14447/jnmes.v23i3.a02

[35] Rani, B. M. S., Majety, V. D., Pittala, C. S., Vijay, V., Sandeep, K. S., & Kiran, S. (2021). Road Identification Through Efficient Edge Segmentation Based on Morphological Operations. *Traitement du Signal*, *38*(5).
[36] Kanumalli, S. S., Ch, A., & Murty, P. S. R. C. (2020). Secure V2V Communication in IOV using IBE and PKI based Hybrid Approach. *International Journal of Advanced Computer Science and Applications*, *11*(1).
[37] Kiran, S., Kanumalli, S. S., Krishna, K. V. S. S. R., & Chandra, N. (2021). WITHDRAWN: internet of things integrated smart agriculture for weather predictions and preventive mechanism.
[38] Prathipati, Silpa Chaitanya, and Susanta Kumar Satpathy. "A Multilevel De-Noising Approach for Precision Edge-Based Fragmentation in MRI Brain Tumor Segmentation." *Traitement du Signal* 40.4 (2023): 1715.

[39] Krishna, Komanduri Venkata Sesha Sai Rama, et al. "Classification of Glaucoma Optical Coherence Tomography (OCT) Images Based on Blood Vessel Identification Using .CNN and Firefly Optimization." Traitement du Signal 38.1 (2021).
[40] Prathipati, Silpa Chaitanya, and Susanta Kumar Satpathy. "Transforming 3D Brain Tumour Image Segmentation: An Enhanced V-Net Approach for Precise Diagnosis and Treatment Planning." *2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*. IEEE, 2024.

[41] Ekkurthi, A., Sujatha, V., Kumar, K.V.,"**Effective Moving Object Tracking Using Adaptive Background Subtraction with Advanced Probability Evolutionary Algorithm",*International Journal on Recent and Innovation Trends in Computing and Communication***, 2023, 11, pp. 1–3
[42] Sujatha, V., Yaddala, M., Kollipara, V., Shaik, K., Burri, R.K(23),"Movie reviews data classification using convolution neural networks",***AIP Conference Proceedings***. 2023, 2724, 030009
[43] Godavarthi, B., Majety, V. D., Mrudula, Y., & Nalajala, P. (2019). Fault identification in power lines using GSM and IoT technology. *Advances in Intelligent Systems and Computing, 815*, 647–655. https://doi.org/10.1007/978-3-319-91117-2_70
[44] Majety, V. D., & Murali, G. (2018). A remote epileptic patient supervising system. *Advances in Modelling and Analysis B, 61*(4), 207–210. https://doi.org/10.18280/ama_b.610402
[45] Naresh, A., TSLP, H., Ch, G., & Kumari, G. R. P. (2023, July). Early Prophecy of Low-Birth-Weight Babies Using BM Error Rate Classifier. In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1-6). IEEE.

[46] Kumari, G. R. P., Reddy, A. H., Lakshmi, K., Abhinaya, B., Sanjana, S., & Naresh, A. (2024, March). Time-Frame-Based Drowsiness Detection System Using CNN. In *2024 2nd International Conference on Disruptive Technologies (ICDT)* (pp. 711-716). IEEE.
[47] V. Pavani, K. Divya, V. V. Likhitha, G. S. Mounika and K. S. Harshitha, "Image Segmentation based Imperative Feature Subset Model for Detection of Vehicle Number Plate using K Nearest Neighbor Model," *2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, Coimbatore, India, 2023, pp. 704-709, doi: 10.1109/ICAIS56108.2023.10073848.

[48] V. Pavani, M. N. Swetha, Y. Prasanthi, K. Kavya and M. Pavithra, "Drowsy Driver Monitoring Using Machine Learning and Visible Actions," *2022 International Conference on Electronics and Renewable Systems (ICEARS)*, Tuticorin, India, 2022, pp. 1269-1279, doi: 10.1109/ICEARS53579.2022.9751890.
[49] Sri, Kurra Santhi, et al. "Advanced system control with traffic handling for secure communication in IoT routing protocol." *Journal Européen des Systèmes Automatisés* 54.2 (2021): 229-233.
[50] Arumugham, Vinothini, et al. "An explainable deep learning model for prediction of early-stage chronic kidney disease." *Computational Intelligence* 39.6 (2023): 1022-1038.
[51] Majety, Vasumathi Devi, et al. "Enhanced secure communication AODV routing protocol using SVM in MANETS." *AIP Conference Proceedings*. Vol. 2724. No. 1. AIP Publishing, 2023.
[52] Krisha, P.S., Peram, S.R. (2023). CT image precise denoising model with edge based segmentation with labeled pixel extraction using CNN based feature extraction for oral cancer detection. Traitement du Signal, Vol. 40, No. 3, pp. 1297-1304. https://doi.org/10.18280/ts.400349
[53] P. S. Krishna, V. R. Aparna, V. Priyanka, P. T. Niharika and T. Shivangi, "Convolution Neural Network Model with Feature Linked Vector for Oral Cancer Detection," *2023 IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT)*, Bhopal, India, 2023, pp. 304-308, doi: 10.1109/CSNT57126.2023.10134660.
[54] Rayachoti, Eswaraiah, Sudhir Tirumalasetty, and Silpa Chaitanya Prathipati. "SLT based watermarking system for secure telemedicine." *Cluster Computing* 23.4 (2020): 3175-3184.
[55] Eswaraiah, Rayachoti, Tirumalasetty Sudhir, and Prathipati Silpa Chaitanya. "Curvelet transform based watermarking for telemedicine." *Wireless Personal Communications* 122.1 (2022): 309-329.
[56] Varshini, Y., Mounika, T., Kumari, G. R. P., Sirisha, G., & Deepthi, Y. (2023, March). Crop Yield Forecast Using Machine Learning. In *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)* (Vol. 1, pp. 2310-2315). IEEE.
[57] B. Aruna Kumari "**Time Series Data Classification for Precise Stock Market Price Prediction using ML" ICICACS International Conference on Integrated Circuits and Communication Systems, Scopus indexed, ISBN:**979-8-3503-1755-8/ **https://ieeexplore.ieee.org/document/10498248,18 April 2024**
[58] B.Aruna Kumari "HumanvAction Recognition From Video Frames Using Recurrent Neural Networks" ICDT 2nd International Conference on Disruptive Technologies (ICDT),Scopis indexed , ISBN:979-8-3503-7105-5/ https://ieeexplore.ieee.org/document/10489658, 11 April 2024.