Journal of Business and Management Studies
ISSN: 2709-0876
DOI: 10.32996/jbms
Journal Homepage: www.al-kindipublisher.com/index.php/jbms

JBMS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Optimizing Marketing Strategies with RFM Method and K-Means Clustering-Based AI Customer Segmentation Analysis

Malay Sarkar[1] ✉ Aisharyja Roy Puja[2] and Faiaz Rahat Chowdhury[3]
[12]Management Science and Quantitative Methods, Gannon University, Erie, PA, USA
[3]MBA Business Analytics, Gannon University, Erie, PA, USA
Corresponding Author: Malay Sarkar, **E-mail**: sarkar002@gannon.edu

| ABSTRACT
Retrospectively, an organization's capacity to comprehend the distinct needs of its clients will undoubtedly provide it with a competitive advantage in terms of delivering targeted client services and tailoring personalized marketing initiatives. This research investigated the efficiency of the k-means clustering algorithm as a technique for efficient consumer segmentation. The k-Means algorithm consolidated with RFM analysis is globally accredited as a profound partitioning clustering technique that has proven to be highly efficient in various business settings. The experimental outcomes provided persuasive evidence of the algorithm's performance in terms of consumer segmentation. The overall cluster purity evaluation was computed to be 0.95. This value demonstrated that the k-Means clustering algorithm incorporated with the RFM analysis attained a relatively high accuracy rate of 95% in terms of precisely and accurately segmenting the consumers based on their shared behaviors and characteristics. The high purity value of 0.95 illustrated the efficiency of the k-Means clustering algorithm in terms of accurately segmenting and categorizing the clients. This showcased that the algorithm efficiently organized and pinpointed consumers into distinct clusters based on their similarities, facilitating targeted marketing strategies and personalized approaches.

| KEYWORDS
K-mean algorithm; RFM analysis; customer segmentation; marketing strategies; clustering

| ARTICLE INFORMATION
ACCEPTED: 01 March 2024 **PUBLISHED:** 07 March 2024 **DOI:** 10.32996/jbms.2024.6.2.5

1. Introduction

The presence of tremendous competitors in the business domain has resulted in exponential rivalries between organizations in acquiring new clientele and retaining existing ones. As a consequence, the significance of excellent customer service has become pivotal, regardless of the company's size. Besides, a company's capability to understand the unique needs of its clients will provide it with a competitive advantage in terms of delivering targeted client services and tailoring personalized marketing initiatives (Kansal & Choudhury, 2018). This comprehension can be attained through systematic customer segmentation, where clients with similar market attributes are categorized into segments. The concepts of machine learning Big Data have immensely led to the widespread employment of automated approaches to client segmentation, surpassing the mainstream and frequently ineffective tactics of market analysis, particularly when dealing with a large client base (Ishaq, 2018). This study will explore the k-means clustering algorithm as the proposed methodology to accomplish efficient customer segmentation.

According to Singh (2023), *Customer Relationship Management* (CRM) technology acts as a mediator between client management activities across all phases of a relationship - initiation, maintenance, and termination - and company performance. Client Segmentation provides a measurable tactic for assessing client data and distinguishing clients based on their purchasing behavior. Consequently, this allows clients to be categorized into various groups, enabling marketing professionals to execute targeted marketing tactics and efficiently retain clients. Once clients are segmented, rules can be developed to portray the characteristics

of clients within each category based on their purchasing behavior. These rules can then be employed to categorize new clients into a suitable group that shares similar purchasing characteristics.

Jadhav (2021) indicated that customer segmentation can be efficiently performed when k-means are consolidated with the RFM technique. In this technique, R denotes recency, which computes the time elapsed between a client's most current transaction and their previous transaction. F represents frequency, showcasing the number of transactions a client has made within a specific timeline. M stands for monetary, reflecting the overall value of a client's transaction amount. Massive evidence has proven that the values of R, F, and M play a pivotal role in terms of ascertaining the characteristics of consumer behavior (Singh, 2023). In this setting, data mining methods, specifically clustering, are employed to establish clusters according to the input data. Within every cluster, the data points illustrate greater similarity when contrasted to data points in other clusters. The measure of similarity is ascertained by computing the distance between the data points, which can be conducted using the Manhattan distance formula.

$$d(x,y)\hat{} \ = \sum\_i\hat{}n▒(\mathrm{xi}-\mathrm{yi})$$

In the above equation, the variable "n" denotes the number of dimensions in the provided dataset, while "x" and "y" denote the data points within the dataset. The distance across data points "x" and "y" is portrayed by "d (x, y)". Jadhav (2021) indicates that for consumer segmentation purposes, the attributes R, F, and M are employed as the three dimensions in the clustering method. Nevertheless, there are substitute Artificial Intelligence (AI) techniques that can be adopted for consumer segmentation, such as the Genetic Algorithm (GA), Self-Organizing Map (SOM), Artificial Bee Colony (ABC), and Particle Swarm Optimization (PSO). It is worth mentioning that Genetic Algorithm is a specific kind of evolutionary computational technique. This study presents an elevated clustering algorithm crafted to efficiently segment consumers into respective groups. The algorithm's performance is assessed by contrasting it to other clustering algorithms, such as single link, K-means, and complete link methods.

2. Literature Review

2.1 Customer Segmentation

Jadhav (2021) indicated that customer segmentation revolves around the classification of consumers into distinct categories according to their shared characteristics. This process frequently comprises employing cluster analysis to obtain a set of segments. Once consumer identification is made, the subsequent phase entails consumer attraction, which targets to motivate and engage each segment of consumers in unique ways. Consumer retention concentrates on maintaining the loyalty of existing clients, while consumer development aims to maximize the value of their purchases. Customer segmentation is a prominent technique utilized during the consumer attraction stage to pinpoint and choose clients for every segment (Kamthania, 2021). The RFM evaluation is employed to pinpoint and exhibit characteristics employing three characteristics, most notably, Recency (R), Monetary (M), and Frequency (F).

2.2 Clustering Algorithm

Clustering refers to an unsupervised classification technique employed when there are no preset classes. In this method, data points in a dataset are allocated to output classes according to their respective proximity to other data points. Every class forms a cluster, with the qua of clusters being equivalent to the quantity of output classes (Kamthania, 2021). The principal goal of clustering is to establish clusters where the data within every cluster portrays high similarity (intra-class similarity) and low similarity with data in other clusters (inter-class similarity).

As per Kansal and Choudhury (2018), there are two major groups of clustering methods: partitioning and hierarchical algorithms. Hierarchical algorithms can be further categorized into agglomerative and divisive methods. Agglomerative clustering begins by handling every data point as a distinct cluster and then continuously merges clusters until all points are consolidated into a single cluster. By contrast, divisive clustering first considers all data points as a portion of a single cluster and then subsequently categorizes clusters until every cluster contains only one data point. On the other hand, partitioning algorithms categorize the dataset into a predefined quantity of clusters, entitled "k." These algorithms target to effectively partition the data set into k clusters according to the specific procedure or optimization objectives.

2.3 K-Means Algorithm

The k-Means algorithm is widely renowned as a prevailing partitioning clustering method. It functions on the premise of centroids, where every data point is allocated to one of the K non-overlapping clusters predefined before running the algorithm (Muhidin, 2018). The k-Means algorithm functions as follows: when presented with a set of d-dimensional training input vectors {x1, x2, ..., xn}, it categorizes the n training illustrations into k clusters or sets of data points portrayed as S = {S1, S2, ..., Sk}, with k being equal to or less than n. The goal is to reduce the within-cluster sum of squares, making sure that data points are categorized together in clusters that reduce the variance within each cluster. The K-means algorithm intends to categorize a database D comprising n objects into k clusters, enhancing the selected partition criterion. Each item is designated to one of the k non-overlapping clusters

(Kholief, 2021). This algorithm functions according to the partitioning approach to clustering. The phases involved in the K-means algorithm are as follows:

$$\text{Argmin S } \sum\_(i = 1)\text{^}k \sum\_xgsi\ X^{n}\text{-U}_{i}$$

Essentially, k-means clustering Algorithm steps can be articulated as follows:

1) Determine the number of clusters, k.
2) Launch the k cluster centroids.
3) Compute the distance between every item to k-cluster centers utilizing the Manhattan distance formula provided by Eq. 1
4) Allocate the n data points to the nearest clusters.
5) Updating the centroid of every cluster utilizing the data points therein.

2.4 Proposed Model

This study proposes the K-means clustering model, which is one of the widely adopted unsupervised learning models utilized by data scientists to automatically divide datasets into groups or clusters based on the similarity between data points. This clustering method, notably K-means, resolves the challenge of clustering unlabeled datasets by pinpointing intrinsic categories within the data. The process of classifying n observations into K clusters is represented as K-means clustering. It is a centroid-oriented clustering technique where data points are assigned to clusters based on the computation of the distance between each point and a centroid. The similarity between data points is normally measured using the Euclidean distance metric.

3. Methodology

3.1 Data Description

The dataset encompassed 40,244 consumer transaction records extending a one-month duration between Amazon, an online retail organization, and its distributors in the B2B domain. The data was collected from a diverse range of avenues, such as social media platforms, consumer relationship management (CRM) systems, Google Analytics, surveys, and other relevant resources. Three values, most notably R, F, and M, are described based on the party ID, purchase amount, and purchase date for each transaction. R denotes the time frame in days between a consumer's current transaction and their previous transaction. F represents the number of transaction records related to the client, while M corresponds to the overall purchase amount for that client within the dataset. Subsequently, the dataset consisted of four attributes: party ID, R, F, and M.

3.2 Feature Normalization

This phase entails data preparation, where feature normalization is conducted to bring data components to a constant scale and reinforce the clustering algorithm's performance. Every data point is modified to fit within the range of -2 to +2. Prevalent normalization methods include decimal scaling, Min-max scaling, and z-score normalization. In this scenario, the z-score normalization method was adopted to normalize the features before implementing the k-means algorithm on the dataset. The formulae for normalization utilizing the z-score method are provided in Equation.

$$Xnorm = (X - uf)/af$$

In Equation (2), $xnorm$ portrays the normalized value of the feature vector component, "x," within the feature vector "f." This value is computed by employing a normalization method to rescale the original value of the feature element. μf represents the mean of the feature vector "f." It is computed by adding up all the values in the feature vector and grouping the sum by the overall number of elements in the vector. in contrast, σf corresponds to the standard deviation of the feature vector "f." It is a statistical measure that measures the variability or dispersion of the values in the feature vector around the mean.

3.3 Centroids Initialization

In the clustering procedure, the preliminary centroids or means play a pivotal role. Figure 1 showcases the initialization and selection of cluster centers. The Forgy technique was adopted to choose the initial centroids, resulting in the selection of four separate cluster centers portrayed by different shapes. The Forgy technique is a renowned method for starting cluster centroids. In this technique, k (in this case, k=4) data points are randomly selected from the dataset as the initial centroids. These randomly chosen data points serve as the starting points for the clustering algorithm and act as the initial representatives for every cluster.

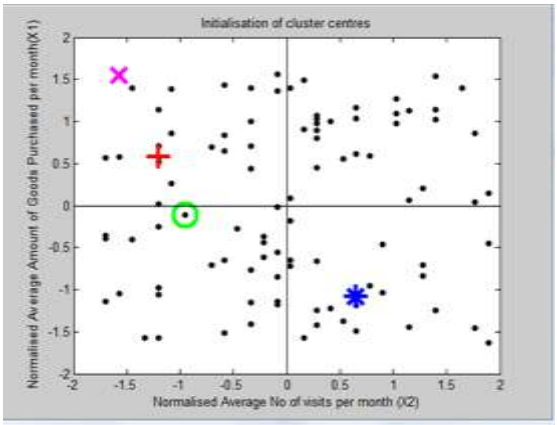


Fig 1: Displays the Initialization phase of the K-mean algorithm

Assignment Phase

In the assignment phase, every data point was assigned to the cluster in which the centroid generates the least within the cluster sum of squares contrasted with other clusters. Particularly, the square Euclidean norms of every data point from the present centroids are computed. Subsequently, the data points are assigned affiliations to the cluster that provides the minimum square Euclidean norm. This can be mathematically articulated in the equation below:

$$S_j^{(t)}=\{X_p:||X_p-\mu_i^{(t)}||^2\leq||X_p-\mu_j^{(t)}||^2\ j,\ 1\leq j\leq k\}(3)$$

During the assignment phase of the clustering protocol, every single data point, represented as *xp*, is allocated to a single cluster or set $s(t)$ at a specified iteration, represented as t. This assignment makes sure that every data point is specifically correlated with one cluster. In the setting of clustering algorithms like k-means, the assignment phase is conducted iteratively until convergence is attained. At every iteration, the data points are reallocated to the clusters premised on specific criteria, such as reducing the within-cluster sum of squares or amplifying the similarity between data points within similar clusters.

After every iteration of the clustering procedure, new centroids are computed for every single cluster. The centroid denotes the central point or center of a cluster and acts as a representative for that cluster. The calculation of the new centroids is conducted by computing the mean of all the data points that belong to the particular cluster. Mathematically, the formula for measuring the new centroid for a cluster can be showcased as:

$$Cj^{(t+1)} = 1/nj \sum_{i=1}^{nj} \mathrm{Xi}$$

Where:

- $Cj^{(t+1)}$ stands for the novel centroid for the j-th cluster at iteration t+1.
- **x***i* represents a data point that is appropriate to the j-th cluster.
- *nj* denotes the overall number of data points in the j-th cluster.

The calculation comprises taking the summation of all the data points within the cluster and allotting them by the overall number of data points in that cluster (denoted by *nj*). This computation generates the mean of the data points, which subsequently becomes the upgraded centroid for the cluster. By recomputing the centroids at every iteration according to the mean of the data points in the cluster, the centroids progressively shift towards the center of the clusters. This iterative protocol proceeds until convergence is attained, where the centroids become steady, and no further substantial changes happen. The upgraded centroids reflect the advancing position of the clusters premised on the distribution of the data points, facilitating a more accurate representation and interpretation of the clusters.

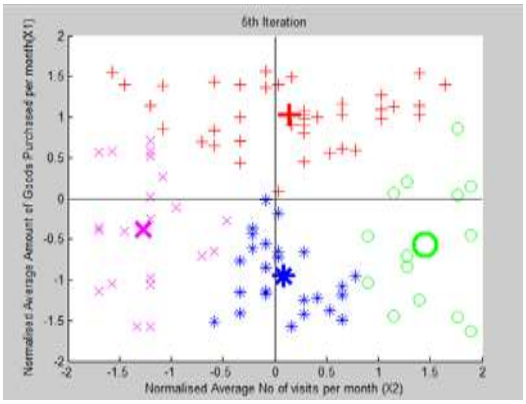
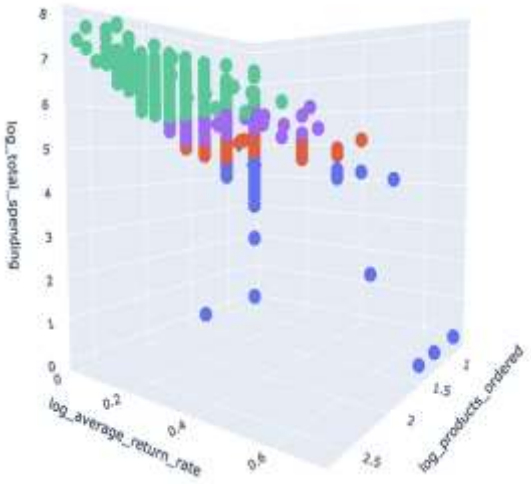


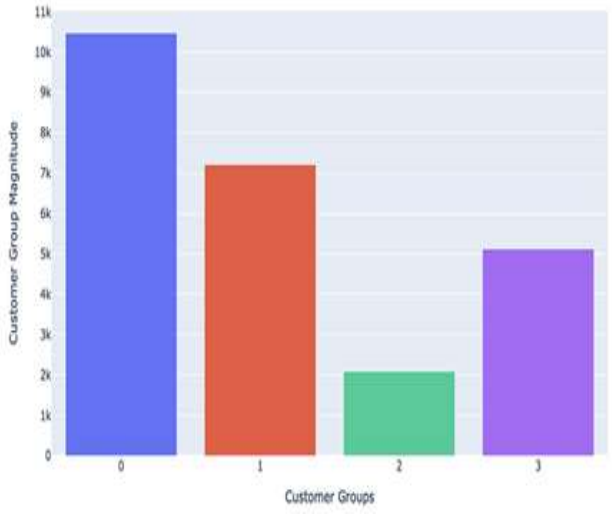
Fig 2: Showcases the location of the centroids and cluster categories after the 30th iteration

Iteration	*Cluster Centre +*		*Cluster Centre **		*Cluster Centre* **O**		*Cluster Centre* **X**	
	INTIALISED CLUSTER CENTROIDS:							
0	-0.0892	1.3654	0.6541	-1.0856	-0.2131	-0.3669	-0.2131	-0.3669
	UPDATED CLUSTER CENTROIDS:							
1	0.5656	1.0971	0.8733	-0.9508	-0.6306	-0.6728	-0.6306	-0.6728
2	0.5798	1.0456	0.9976	-0.9639	-0.5466	-0.8295	-0.5466	-0.8295
3	0.5502	1.0346	1.0376	-0.9348	-0.5600	-0.9284	-0.5600	-0.9284
4	0.5502	1.0346	1.0376	-0.9348	-0.5641	-0.9557	-0.5641	-0.9557
5	0.5502	1.0346	1.0376	-0.9348	-0.5901	-0.9894	-0.5901	-0.9894

Figure 2 portrays the locations of the centroids and the modified assignments of their respective cluster categories after the fulfillment of the 30th iteration. Every cluster member presumes a similar shape as their equivalent cluster centroid, visually displaying their association. Furthermore, Table I presents an extensive overview of the adjustments in the cluster centroids from the first phase (0th iteration) up to the 5th iteration. The table pinpoints the advancement of the centroids as the clustering algorithm proceeds, exhibiting the alterations made to their locations over multiple iterations. This info presents insight into how the centroids progressively converge toward steady positions, illustrating the refinement and convergence of the clustering algorithm.

4. Experimental Results





As showcased in Figure [3], the centroids of every consumer group are visualized using cubes, while the independent data points are displayed as spheres. The above four consumer groups can be pinpointed according to their respective characteristics:

- ✓ ***Blue Group***: This group comprises clients who have made at least one purchase, with an approximate expenditure of up to 100 dollars, and had the greatest average return rate. These consumers were inferred to be potentially brand-new users of the online store.
- ✓ ***Purple Group***: Consumers belonging to this category had placed orders for 1 to 4 items, with an approximated overall expenditure of 300 dollars and an average return rate of 0.5. They exhibited moderate purchasing activities and had a moderate level of returns.
- ✓ ***Red Group***: The red category represented clients who had ordered 1 to 4 products, with an estimated overall expenditure of 150 dollars and a return rate of 0.5. These clients exhibited similar purchasing patterns and return behavior as the purple group.
- ✓ ***Green Group:*** Consumers in the green category had purchased 1 to 13 products, spending an approximation of 600 dollars and returning 0 items on average. This category represented the most beneficial customer base for the online retail company since they exhibited a high level of purchasing activity without any significant returns.

By classifying consumers into these respective categories based on their expenditure, purchasing behavior, and return trends, retail online companies such as Amazon can gain valuable insights and tailor their marketing strategies to efficiently target and engage each customer segment.

4.1 Performance Evaluation

To determine the level to which a cluster comprises a specific class of data points, a purity measure was adopted. This measure, represented as $purity(D\text{i})$, evaluates the purity of every cluster. The computation of cluster purity is expressed by Equation (5):

$$purity(D\text{i}) = \max j(P\text{i}(C\text{j})) \quad (5)$$

In this equation, $P\text{i}(C\text{j})$ denotes the quantity of data points ascribed to class $C\text{j}$ within cluster I, also referred to as $D\text{i}$. The expression $P\text{i}(C\text{j})$ corresponds to the ratio of data points in cluster i that belong to class $C\text{j}$. To ascertain the purity of a cluster, the maximum portion of any class $C\text{j}$ within that cluster is considered. Particularly, the cluster purity is ascertained by the greatest proportion of any specific class present in that cluster. The confusion matrix is displayed in the table below:

Cluster	HBRV	HBIV	LBRV	LBIV	Purity
Cluster X	1	0	0	0	1.000
Cluster+	28	21	0	0	0.954
Cluster*	0	0	22	1	0.957
Cluster O	0	2	22	1	0.889
Total	29	23	23	25	**0.950**

Table 2: Displays the Confusion Matrix

According to Table 2 above, it was evident that the overall cluster purity expressed as $Puritytotal(D)$ was computed to be 0.95. This value illustrates that the K-mean clustering algorithm attained a relatively high accuracy rate of 95% in terms of segmenting the customers. The relatively high purity value of 0.95 indicated that the K-mean clustering algorithm was effective in terms of accurately segmenting and grouping the clients based on their respective shared behaviors and characteristics. This signified that there was a high level of homogeneity within every cluster, with the majority of consumers in each cluster exhibiting similar attributes or patterns.

By attaining a 95% accuracy rate in consumer segmentation, the K-mean clustering algorithm provided valuable info and a reliable tenet for online retail companies such as Amazon to comprehend and target specific client segments effectively. This accurate segmentation can assist in terms of crafting tailored marketing strategies, individualized recommendations, and consumer-centric tactics, ultimately leading to enhanced customer satisfaction and business performance.

5. Conclusion

This study explored the k-means clustering algorithm as the proposed methodology to accomplish efficient customer segmentation. The k-Means algorithm is widely renowned as a prevailing partitioning clustering method. It functions on the premise of centroids, where every data point is allocated to one of the K non-overlapping clusters predefined before running the algorithm. From the experimental result, it was evident that the overall cluster purity expressed as $Puritytotal(D)$ was computed to be 0.95. This value illustrates that the K-mean clustering algorithm attained a relatively high accuracy rate of 95% in terms of segmenting the customers. The relatively high purity value of 0.95 indicated that the K-mean clustering algorithm was effective in terms of accurately segmenting and grouping the clients based on their respective shared behaviors and characteristics.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Charles, P. (2023). An improved clustering algorithm for customer segmentation. *www.academia.edu*. https://www.academia.edu/107949086/An_Improved_Clustering_Algorithm_for_Customer_Segmentation

[2] Hasan, R. (2024) Revitalizing the Electric Grid: A Machine Learning Paradigm for Ensuring Stability in the USA." *Journal of Computer Science and Technology Studies* 6.1 (2024): 141-154.

[3] Ishaq, A. I. (2018). Customer Segmentation based on RFM model and Clustering Techniques With K-Means Algorithm. *www.academia.edu*. https://www.academia.edu/111027525/Customer_Segmentation_based_on_RFM_model_and_Clustering_Techniques_With_K_Means_Algorithm

[4] Jadhav, S. (2021). Customer Segmentation using the RFM Model and K-Means Clustering. *www.academia.edu*. https://www.academia.edu/103419318/Customer_Segmentation_using_RFM_Model_and_K_Means_Clustering

[5] Kamthania, D. (2021). Market Segmentation Analysis and Visualization using K-Mode Clustering Algorithm for E-Commerce Business. *www.academia.edu*. https://www.academia.edu/53604628/Market_Segmentation_Analysis_and_Visualization_using_K_Mode_Clustering_Algorithm_for_E_Commerce_Business

[6] Kansal, T., & Choudhury, T. (2018, December 1). *Customer Segmentation using K-means Clustering*. IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/8769171

[7] Kholief, M. (2021). Automated Market Analysis by RFMx Encoding Based Customer Segmentation using Initial Centroid Selection Optimized K-means Clustering Algorithm. *Aast*. https://www.academia.edu/84861176/Automated_Market_Analysis_by_RFMx_Encoding_Based_Customer_Segmentation_using_Initial_Centroid_Selection_Optimized_K_means_Clustering_Algorithm

[8] Muhidin, A. (2018). ANALYSIS OF HIERARCHICAL CLUSTERING AND K-MEAN METHODS WITH LRFMP MODEL ON CUSTOMER SEGMENTATION. *www.academia.edu*. https://www.academia.edu/35955183/ANALYSIS_OF_HIERARCHICAL_CLUSTERING_AND_K_MEAN_METHODS_WITH_LRFMP_MODEL_ON_CUSTOMER_SEGMENTATION

[9] Singh, S. (2023, January 19). *Customer Segmentation of E-commerce data using K-means Clustering Algorithm*. IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/10048834