Institute of Research and Publication Indonesia (IRPI)
Public Research Journal of Engineering, Data Technology and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/predatecs
Vol. 1 Iss. 2 January 2024, pp: 102-108
ISSN(P): 3024-921X | ISSN(E): 3024-8043

Comparison of K-Means, BIRCH and Hierarchical Clustering Algorithms in Clustering OCD Symptom Data

Alika Rahmarsyarah Rizalde[1*], Haykal Alya Mubarak[2], Gilang Ramadhan[3], Mohd. Adzka Fatan[4]

[12]Department of Information System, Faculty of Science and Technology,
Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia
[3]Department of Automotive Engineering, Faculty of Engineering,
Kasetsart University, Thailand
[4]Department of Dirasat Islamiyah, Faculty of Al Qur'an Al Karim,
University of Qur'an Al-Karim and Islamic Sciences, Sudan

E-Mail: [1]rahmarsyaraha@gmail.com, [2]haykalalya82@gmail.com,
[3]gilang.r@ku.th, [4]akafatan05@gmail.com

Received Dec 27th 2023; Revised Jan 20th 2024; Accepted Jan 30th 2024
Corresponding Author: Alika Rahmarsyarah Rizalde

Abstract

The hallmarks of Obsessive-Compulsive Disorder (OCD) are intrusive, anxiety-inducing thoughts (called obsessions) and associated repeated activities (called compulsions). To understand the patterns and relationships between OCD data that have been obtained, data will be grouped (clustering). In clustering using several clustering algorithms, namely K-Means, BIRCH, In this work, hierarchical clustering was used to identify the optimal cluster value comparison, and the Davies Bouldin Index (DBI) was used to confirm the results. Then the results of the best cluster value in processing OCD data are using the BIRCH algorithm in the K10 experiment which gets a value of 1.3. While the K-Means algorithm obtained the best cluster at K10 with a value obtained of 1.36 and the Hierarchical clustering algorithm also at the K10 value of 2.03. Thus in this study, the comparison results of the application of 3 clustering algorithms obtained results, namely the BIRCH algorithm shows the value of the resulting cluster is the best in clustering OCD data. This means that the BIRCH algorithm can be used to cluster OCD data more accurately and efficiently.

Keyword: BIRCH, Clustering, Hierarchical Algorithm, K-Means, OCD

1. INTRODUCTION

In the new era of digitalization, the collection of health data is growing rapidly, opening up opportunities to expand insights into various medical conditions. Related to mental conditions today can affect the mental health of individuals, people affected by mental health disorders today are also easily obtained and accessed on the available pages by displaying increasingly large data and stored in databases to be able to increase the need for efficient and effective analysis [1]. A certain level of anxiety can make a person more productive, more so in psychiatric disorders or mental conditions characterized by excessive anxiety or a persistent thought (obsession) [2]. Obsessive-Compulsive Disorder (OCD) is the name given to this disorder, which is frequently characterized by intrusive, anxiety-inducing thoughts (called obsessions) and associated repetitive activities (called compulsions) [3].

One such aspect that has been the focus of research is OCD. With the growth in the amount of OCD data available, there is a need for effective data analysis methods to identify patterns and relationships that may not be directly visible. OCD is characterized by recurrent intrusive thoughts. OCD is currently diagnosed based on subjective clinical interviews and scale evaluations, which often result in inconsistent diagnostic outcomes among psychiatric, religious and community professionals [4]. OCD is the name given to this disorder, which is frequently characterized by intrusive, anxiety-inducing thoughts (called obsessions) and associated repetitive activities (called compulsions) [5]. The categorization of OCD symptoms into categories helps in deep understanding of symptom variability, response to treatment, and may provide new insights into OCD subtypes. To understand the patterns and relationships between the OCD data that have been obtained, clustering will be performed.

The term "clustering" is used interchangeably to describe how search groups can collect disaggregated data. Depending on the community setting, presumptions made about the assembly process's component parts,

and the context in which the assembly is employed during the data gathering process, several labels emerge [6]. In this study, clustering analysis is one approach that can be used to group OCD data into categories that have similar characteristics. To build clusters, a variety of clustering algorithms and methods are employed. This data clustering focuses on three main cluster algorithms that belong to unsupervised learning algorithms, namely K-means, Hierarchical Clustering, and Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH).

K-Means algorithm is one of the simplest unsupervised learning algorithms and plays an important role in the field of data mining. With the simplicity of the application of the K-Means algorithm, it is usually used to handle very large data sets [7]. Known as the most prominent clustering method, K-Means has the advantage of being an algorithm that is straightforward to implement, relatively fast in terms of computing time, and has been widely utilized to tackle many computational issues [8][9].While cluster analysis evolved into a subset of statistical multivariate analysis, hierarchical clustering the first clustering technique to emerge—was employed by social scientists and biologists [8]. An integrated hierarchical clustering algorithm is what is known as the BIRCH algorithm. It is highly adapted to address clustering problems of discrete and continuous attribute data because it makes use of feature clustering (CF) and cluster feature tree (CF Tree), two concepts for broad cluster description [10]. The other part of the BIRCH algorithm is a clustering algorithm referred to as tree-BIRCH [11].

The data that has been clustered using these three clustering algorithm methods will be identified based on their respective groups. The K-means, BIRCH, and Hierarchical clustering formulas will then be used to calculate the accuracy of each algorithm, and the results can be compared with existing groups that can be examined based on their supporting characteristics. Based on research by U. R. Gurning, Mustakim, I. Permana, and I. Maita (2023), this study used clusters made up of two to ten trials, while earlier studies have applied and contrasted a number of clustering methods to see which ones work best for grouping COVID-19 data. With a value of k = 4, the outcomes show the best cluster produced by the k-means algorithm. However, for k = 9, the K-Medoid method has the best cluster. The Davies Bouldin Index is used to validate the clustering results (DBI). This study presents a more up-to-date understanding of the dynamics of the condition by using a more recent and pertinent dataset of OCD symptoms than earlier research. Furthermore, using a wide range of techniques and more recent data, this study applies several clustering algorithms, including K-Means, BIRCH, and Hierarchical Clustering. The reason this study was conducted is that, by using cluster analysis of symptom data, it offers a comprehensive understanding of the features of OCD. It also directs the choice of the best algorithm for processing data related to OCD symptoms. The findings have the potential to greatly advance the field of mental health diagnosis and treatment approaches. Furthermore, this research offers helpful guidelines for data analysis in a variety of scenarios and is not limited to OCD alone. It may also be applied to mental health contexts in general. Thus, the purpose of this study was to evaluate each algorithm's accuracy. Then, it used the Davies Bouldin Index (DBI) to confirm the clustering findings and compare the clusters that could be evaluated based on the supported attributes.

2. MATERIAL AND METHOD

2.1. Stage of Research

In this study, several stages were carried out, namely collecting data, pre-processing data, clustering process using three algorithms and testing cluster quality using DBI and ending with results and analysis. as seen in Figure 1.

2.2. Obsessive-Compulsive Disorder (OCD)

People with OCD are mentally ill and must engage in activities to calm their intrusive thoughts [12]. Higher OCD levels may be associated with more psychological discomfort in a person [13]. The hallmark of OCD is recurrent intrusive thoughts that drive an individual to repeatedly carry out the same action. Genetic studies have shown that abnormalities in immunologic mechanisms can cause OCD [5]. OCD symptoms are diseases that fall under a number of diagnostic categories but share characteristics common to OCD, including anxiety, compulsive actions, and obsessive thoughts [14]. OCD is also an under-treated disease due to the great contrast in OCD, etiology, symptoms, subtypes, and treatment response [15].

2.3. K-Means

K-Means is a partitioning method that is able to apply data analysis as objects based on the location and distance between various data points [16]. One of the most used methods for cluster analysis is the k means clustering algorithm since it produces reliable and quick result [17]. The purpose of this algorithm is to be able to break an object into k clusters, then an analysis is carried out where each cluster object is obtained through the closest average [18]. For that, it can be applied with the formula 1.

$$m\sum_{k}^{k} \quad dik = \sqrt{\sum_{j=1}^{m} \quad (C_{i}j - X_{i}j)^{2}} \tag{1}$$

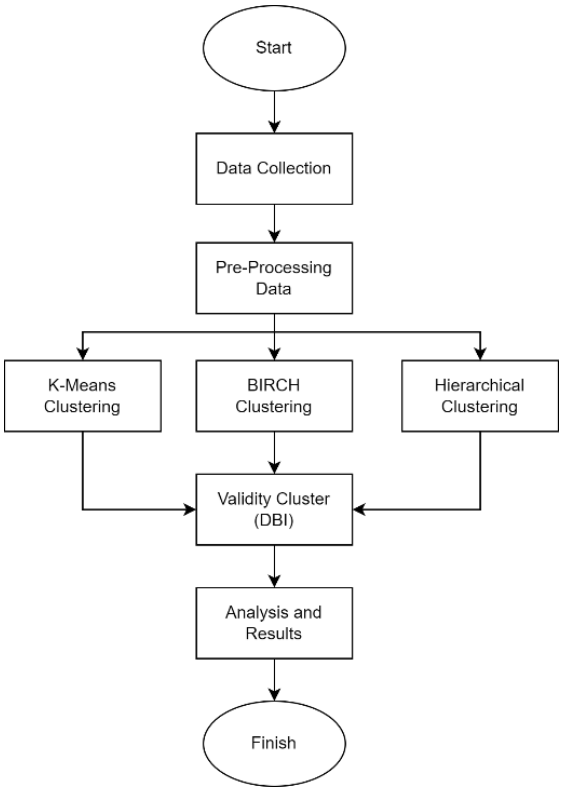


Figure 1. Research Methodology

2.4. BIRCH

Birch's cluster feature tree (CF Tree) and feature clustering (CF) are used in this approach. While BIRCH can solve big data sets with faster execution times, the quality of the generated clusters remains subpar [10]. The benefit of BIRCH is that it can infer the best subcluster that can be obtained while limiting input/output while gradually producing the highest quality clusters, which is useful when other methods struggle to handle outliers and huge datasets [19]. The steps of the BIRCH algorithm are as follows [20].

Algorithm 1 : BIRCH Algorithm

```
Input: D= {t1, t2, t3 … … tn } // Set of elements
T // Threshold for CF tree construction
Output:
K // Set of clusters.
BIRCH clustering algorithm:
    For  each, tii ϵϵ D   do
    Determine correct leaf node for ti insertion;
    If
    threshold condition is not violated then Add ti to cluster and update CH triples;
    else
    if room to insert ti then
    insert ti as singe cluster and update CF triples;
    else
    split leaf node and redistribute CF features;
end
```

2.5. Hierarchical Clustering

This type of algorithm divides the data into different levels, eventually forming a tree-like cluster structure [21]. With minimal assumptions about the overall distribution of data points, hierarchical clustering can create a hierarchical structure among data points by spontaneously defining clusters based on the branches in the tree hierarchy. It is hence appropriate for a variety of data types. The depiction of all data points with hierarchical relationships for result interpretation is another significant benefit [22]. Hierarchical clustering algorithms use centroids as cluster centers to minimize the Euclidean distance, and choose medoids only as cluster representations. This algorithm is deterministic, meaning that it is reproducible. However, it is also greedy, meaning it generates local solutions with a collection of cluster assignments [23].

2.6. Davies-Bouldin Index (DBI)

When using clustering techniques, the validity of a cluster is assessed using the Davies-Bouldin Index. In measuring, When DBI reaches the highest point in a cluster, it indicates that the cluster type or the differences between clusters are more apparent. DBI maximizes the distance between points in a cluster [24]. Since external validity and internal validity are the two primary categories used in validation to evaluate the effectiveness of clustering results, DBI is utilized as a clustering matrix [25]. The number of clusters and the intra- and inter-cluster distances are used to generate the DBI, which will show the degree of similarity across clusters regardless of the number of clusters and partitioning method utilized [26]. It is possible to obtain more patterned and comprehensive information in addition to determining the ideal number of clusters by using DBI to approximate the intra-cluster distance [27]. To calculate the DBI value can be applied with the formula 2.

$$DB = \sum_{i=1}^{p} \quad \left(\frac{\sigma i + \sigma i}{p}\right) \tag{2}$$

3. RESULTS AND ANALYSIS

In the analysis and results obtained using the application of 3 clustering algorithms with the Python language executed with Google Colab to be able to implement the model that will be presented in this section. The following process will be carried out at the analysis and results stage.

3.1. Collecting and Preprocessing Data

The data used in this study are patient data detected as having OCD symptoms obtained on the kaggle website with a time span of November 2013 - November 2022. The data obtained from this data collection is 1,500 data. The categories used are Age, Gender, Education Level, Duration of Symptoms (months), Previous Diagnoses, Family History of OCD, and other categories. Next, preprocessing the data using excel to determine whether the attributes to be used still have a lot of missing values, noise and are still in the form of qualitative data so that they must first be converted into numerical data. After making sure there is no missing value or cleaning data in the data, it can be continued with data transformation and data normalization. And determine the clustering results obtained by utilizing DBI.

3.2. K-Means Clustering

The implementation for clustering in this study was carried out using three models, namely K-Means, BIRCH, and Hierarchical Clustering. This model processes 1500 OCD patient data that has been processed previously.

In clustering with the K-Means algorithm, 9 trials were conducted, after which each trial was tested for cluster validity using the Davies-Bouldin Index technique. The test results as seen in Figure 2.

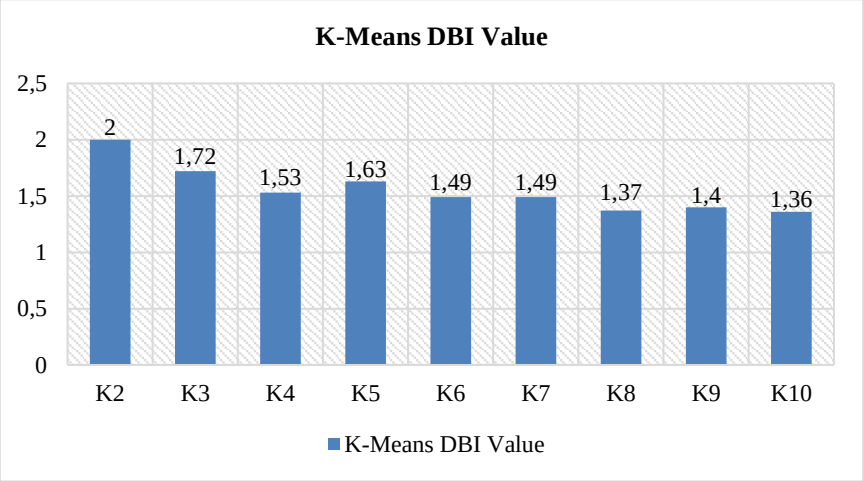


Figure 2. K-Means Validity Results

In Figure 2 on this K-Means algorithm, it can be seen that the best cluster from the validity test lies in the K10 experiment with a value of 1.36.

3.3. BIRCH

The clustering results with the BIRCH algorithm were also conducted with 9 trials, and then each trial was tested for cluster validity by applying the Davies-Bouldin Index technique. The test results as seen in Figure 3.

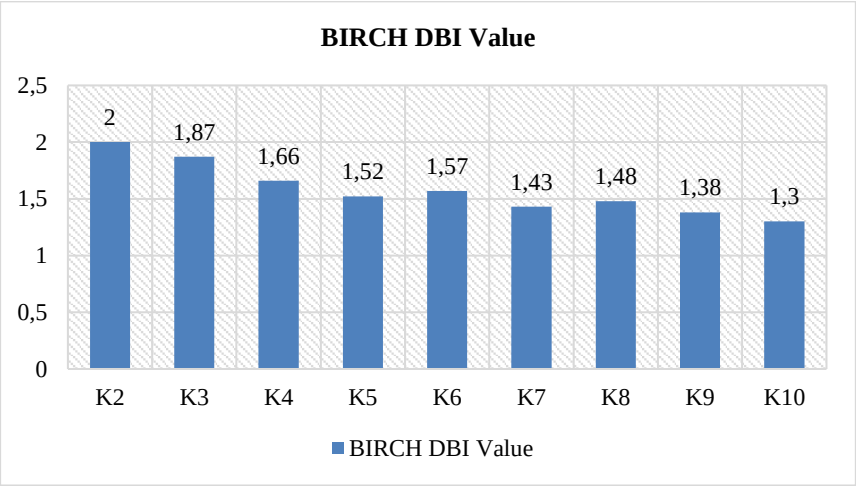


Figure 3. BIRCH Validity Results

In Figure 3 on this BIRCH algorithm, it can be seen that the best cluster of the validity test lies in the K10 experiment with a value of 1.3.

3.4. Hierarchical Clustering

In the clustering process with the Hierarchical Clustering algorithm, the same is done with 9 trials, after which each trial is tested for cluster validity making use of the Davies-Bouldin Index method. Figure 4 displays the test findings.

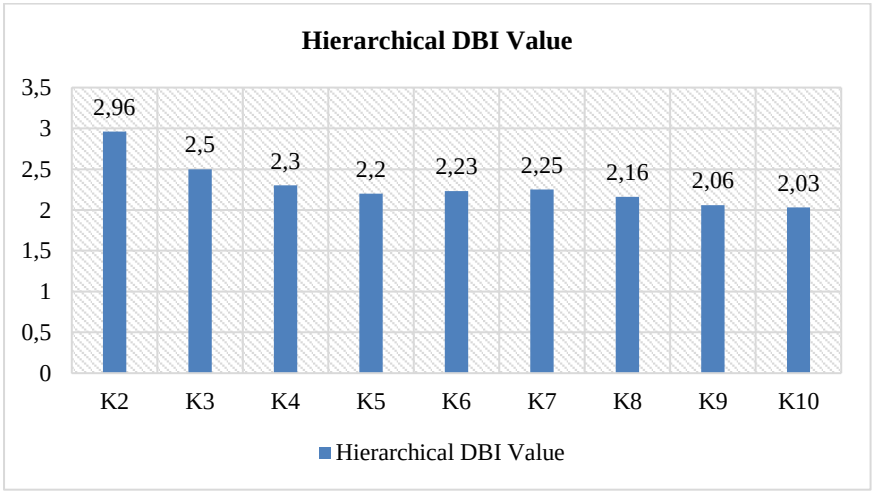


Figure 4. Hierarchical Clustering Validity Results

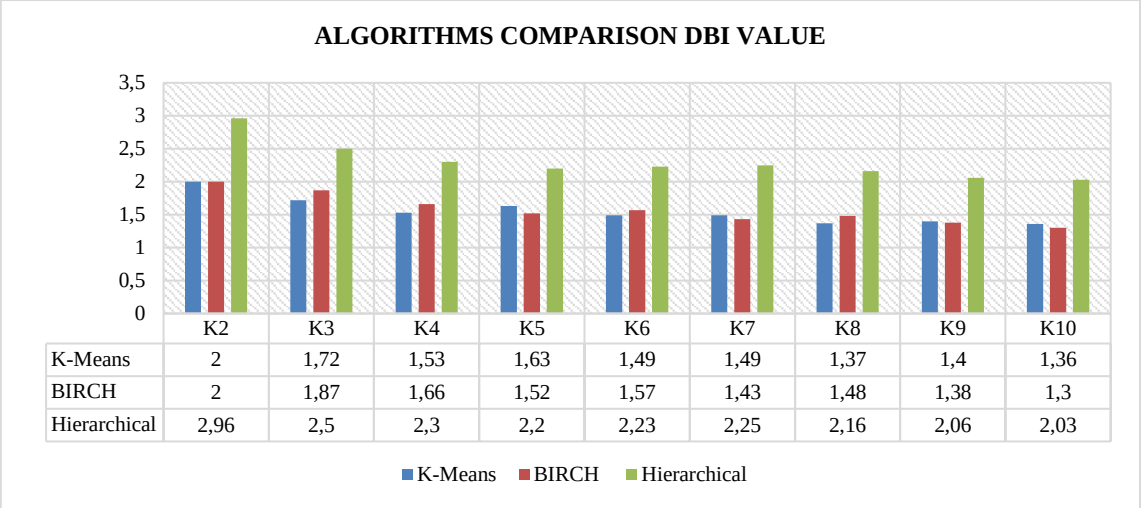
With a value of 2.03, the K10 experiment is the best cluster for the validity test, according to Figure 4 of this Hierarchical Clustering technique.

3.5. Comparison DBI Results of Algorithms

After clustering and testing the validity of clusters in each algorithm, the next step is to compare which algorithm has the best validity test value, which can be seen in Figure 5.

The best cluster of the K-Means algorithm's validity test on K10 may be observed based on Figure 5 for these three algorithms with a value of 1.36, the BIRCH algorithm on K10 with a value of 1.3, and the Hierarchical algorithm on K10 with a value of 2.03. Furthermore, the three best are compared, which validity test value is the best among the best, can be seen in Figure 6.

Based on Figure 6 on these three algorithms,As can be observed, the K10 experiment of the BIRCH algorithm, which has a value of 1.3, is the best cluster for the validity test. Thus, in this study, the most optimal cluster was obtained among the three algorithms, namely the BIRCH algorithm with trial K10 with a validity test value of 1,3.



	K2	K3	K4	K5	K6	K7	K8	K9	K10
K-Means	2	1,72	1,53	1,63	1,49	1,49	1,37	1,4	1,36
BIRCH	2	1,87	1,66	1,52	1,57	1,43	1,48	1,38	1,3
Hierarchical	2,96	2,5	2,3	2,2	2,23	2,25	2,16	2,06	2,03

Figure 5. DBI Value Comparison

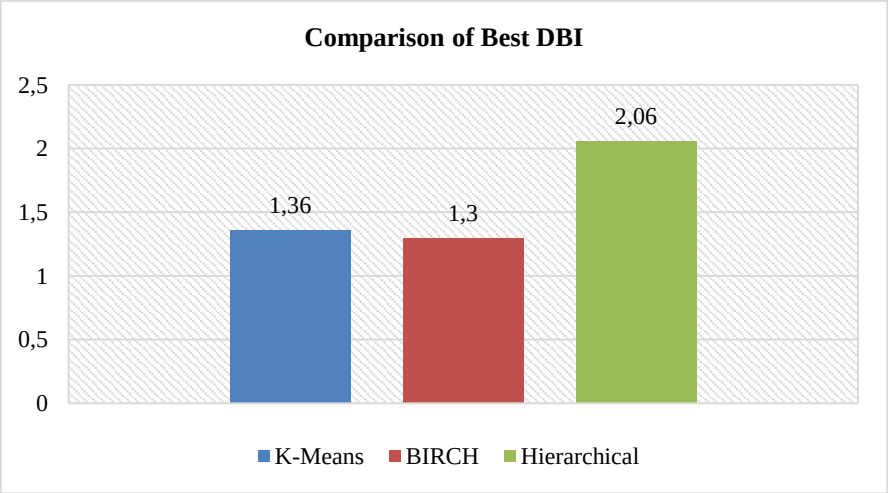


Figure 6. Comparison of Best DBI

4. CONCLUSION

According to the results of data analysis and processing by applying clustering algorithms, namely K-Means, BIRCH, and Hierarchical Clustering, it can be concluded that the best cluster value in OCD data processing using the BIRCH algorithm in the K10 experiment obtained a value of 1.3. While the K-Means algorithm obtained the best cluster with K10 with a value obtained of 1, 36 and the Hierarchical clustering algorithm also at the K10 value of 2.03. The clustering results from the application of several algorithms are validated using the Davies Bouldin Index (DBI). Thus in this study the BIRCH algorithm outperforms the K-Means and Hierarchical algorithms in clustering OCD data in accordance with the purpose of this study which is to be able to determine the accuracy of each algorithm.To be able to know the validity test by getting the best clustering results, further research can be done by applying several other clustering algorithms.

REFERENCES

[1] J. Bruder *et al.*, "MusicCohort: Pilot feasibility of a protocol to assess students' physical and mental health in a Canadian post-secondary school of music," *BMC Res. Notes*, vol. 14, no. 1, pp. 1–7, 2021, doi: 10.1186/s13104-021-05829-9.

[2] W. Novia Annisa *et al.*, "Suicidal Risk in People with Obsessive Compulsive Disorder," 2023, [Online]. Available: http://dx.doi.org/10.29303/jbt.v23i4.5602.

[3] P. Morgado, "What Is Obsessive Compulsive Disorder?," *Front. Young Minds*, vol. 7, 2019, doi: 10.3389/frym.2019.00138.

[4] X. Yang *et al.*, "Multivariate classification of drug-naive obsessive-compulsive disorder patients and healthy controls by applying an SVM to resting-state functional MRI data," *BMC Psychiatry*, vol. 19, no. 1, pp. 1–8, 2019, doi: 10.1186/s12888-019-2184-6.

[5] H. Lamothe, J.-M. Baleyte, P. Smith, A. Pelissolo, and L. Mallet, "Individualized Immunological Data

for Precise Classification of OCD Patients," *Brain Sci.*, vol. 8, no. 8, p. 149, 2018, doi: 10.3390/brainsci8080149.

[6] A. Ali *et al.*, "Systematic Review: A State of Art ML Based Clustering Algorithms for Data Mining," *Proc. - 2020 23rd IEEE Int. Multi-Topic Conf. INMIC 2020*, 2020, doi: 10.1109/INMIC50486.2020.9318060.

[7] M. A. Ahmed, H. Baharin, and P. N. E. Nohuddin, "Analysis of K-means, DBSCAN and OPTICS Cluster algorithms on Al-Quran verses," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 8, pp. 248–254, 2020, doi: 10.14569/IJACSA.2020.0110832.

[8] K. P. Sinaga and M. S. Yang, "Unsupervised K-means clustering algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

[9] Z. Lv, T. Liu, C. Shi, J. A. Benediktsson, and H. Du, "Novel Land Cover Change Detection Method Based on k-Means Clustering and Adaptive Majority Voting Using Bitemporal Remote Sensing Images," *IEEE Access*, vol. 7, pp. 34425–34437, 2019, doi: 10.1109/ACCESS.2019.2892648.

[10] F. Ramadhani, M. Zarlis, and S. Suwilo, "Improve BIRCH algorithm for big data clustering," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 725, no. 1, 2020, doi: 10.1088/1757-899X/725/1/012090.

[11] "Evaluation of BIRCH Clustering Algorithm for Big Data," vol. 4, no. 4, 2019.

[12] A. Guazzini, M. C. Gursesli, E. Serritella, M. Tani, and M. Duradoni, "Obsessive-Compulsive Disorder (OCD) Types and Social Media: Are Social Media Important and Impactful for OCD People?," *Eur. J. Investig. Heal. Psychol. Educ.*, vol. 12, no. 8, pp. 1108–1120, 2022, doi: 10.3390/ejihpe12080078.

[13] E. Fontes-Perryman and R. Spina, "Fear of missing out and compulsive social media use as mediators between OCD symptoms and social media fatigue.," *Psychol. Pop. Media*, vol. 11, no. 2, pp. 173–182, 2022, doi: 10.1037/ppm0000356.

[14] G. Martinotti *et al.*, "Therapeutic potentials of ketamine and esketamine in obsessive–compulsive disorder (Ocd), substance use disorders (sud) and eating disorders (ed): A review of the current literature," *Brain Sci.*, vol. 11, no. 7, 2021, doi: 10.3390/brainsci11070856.

[15] R. Rostami *et al.*, "Efficacy and clinical predictors of response to rTMS treatment in pharmacoresistant obsessive-compulsive disorder (OCD): A retrospective study," *BMC Psychiatry*, vol. 20, no. 1, pp. 1–13, 2020, doi: 10.1186/s12888-020-02769-9.

[16] M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means algorithm: A comprehensive survey and performance evaluation," *Electron.*, vol. 9, no. 8, pp. 1–12, 2020, doi: 10.3390/electronics9081295.

[17] H. S. Kim, S. K. Kim, and L. S. Kang, "BIM performance assessment system using a K-means clustering algorithm," *J. Asian Archit. Build. Eng.*, vol. 20, no. 1, pp. 78–87, 2021, doi: 10.1080/13467581.2020.1800471.

[18] I. Kamila, U. Khairunnisa, and M. Mustakim, "Perbandingan Algoritma K-Means dan K-Medoids untuk Pengelompokan Data Transaksi Bongkar Muat di Provinsi Riau," *J. Ilm. Rekayasa dan Manaj. Sist. Inf.*, vol. 5, p. 119, Feb. 2019, doi: 10.24014/rmsi.v5i1.7381.

[19] M. C. Nwadiugwu, "Gene-Based Clustering Algorithms: Comparison Between Denclue, Fuzzy-C, and BIRCH," *Bioinform. Biol. Insights*, vol. 14, pp. 1–6, 2020, doi: 10.1177/1177932220909851.

[20] M. Arora, S. Agrawal, and R. Patel, "User Location prediction using Hybrid BIRCH clustering and Machine Learning approach," *J. Integr. Sci. Technol.*, vol. 12, no. 1, pp. 1–7, 2023.

[21] Q. Qi and Y. Wang, "Motor Group Aggregation of Refinery and Chemical Enterprises Based on Hierarchical Clustering Algorithm," vol. 4, pp. 13–22, 2021, doi: 10.23977/jaip.2020.040102.

[22] R. Petegrosso, Z. Li, and R. Kuang, "Machine learning and statistical methods for clustering single-cell RNA-sequencing data," *Brief. Bioinform.*, vol. 21, no. 4, pp. 1209–1223, 2019, doi: 10.1093/bib/bbz063.

[23] H. Teichgraeber and A. Brandt, "Clustering methods to find representative periods for the optimization of energy systems: An initial framework and comparison," *Appl. Energy*, vol. 239, pp. 1283–1293, Apr. 2019, doi: 10.1016/j.apenergy.2019.02.012.

[24] M. Mughnyanti, S. Efendi, and M. Zarlis, "Analysis of determining centroid clustering x-means algorithm with davies-bouldin index evaluation," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 725, no. 1, 2020, doi: 10.1088/1757-899X/725/1/012128.

[25] Y. A. Wijaya, D. A. Kurniady, E. Setyanto, W. S. Tarihoran, D. Rusmana, and R. Rahim, "Davies Bouldin Index Algorithm for Optimizing Clustering Case Studies Mapping School Facilities," *TEM J.*, vol. 10, no. 3, pp. 1099–1103, 2021, doi: 10.18421/TEM103-13.

[26] H. Barros *et al.*, "Nutritional Clustering of Cookies Developed with Cocoa Shell, Soy, and Green Banana Flours Using Exploratory Methods," *Food Bioprocess Technol.*, vol. 13, Sep. 2020, doi: 10.1007/s11947-020-02495-w.

[27] I. F. Ashari, R. Banjarnahor, D. R. Farida, S. P. Aisyah, A. P. Dewi, and N. Humaya, "Application of Data Mining with the K-Means Clustering Method and Davies Bouldin Index for Grouping IMDB Movies," *J. Appl. Informatics Comput.*, vol. 6, no. 1, pp. 07–15, 2022, doi: 10.30871/jaic.v6i1.3485.