Received: June 7, 2023. Revised: September 8, 2023.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

# ImpClust: An Algorithm to Cluster Chemical Datasets for Drug Discovery

**Hutashan V. Bhagat**[1]* **G. Uday Kiran**[1] **Manminder Singh**[2]

*[1]B V Raju Institute of Technology, Narsapur, Hyderabad, India*
*[2]Sant Longowal Institute of Engineering and Technology, Longowal, Punjab, India*
* Corresponding author's Email: hutashan20@gmail.com

**Abstract:** Data clustering, an unsupervised machine learning technique, plays a critical part in the process of drug discovery in chemoinformatics. Researchers have come up with numerous clustering algorithms over the past decades that are well suited to analyze large chemical datasets of high dimensionality. The applications of clustering algorithms can be seen in lead compound selection which is the process of identifying the chemical compound that helps in the treatment of disease and results in the development of a new drug in the drug discovery process. The quantitative structure-property relationship (QSPR) in the drug discovery process identifies the compounds having similar properties using clustering algorithms over the structural descriptors of the chemical compounds. The quantitative structure-activity relationship (QSAR) process uses cluster analysis to identify the empirical relationships between the chemical structure and biological activities among similar compounds. The acute toxicity of the chemical compound is controlled by the chemists in the drug discovery process using cluster analysis. Considering the numerous applications of data clustering in the drug discovery process, in this paper, an improved clustering algorithm ImpClust is proposed to cluster similar compounds based on chemical composition. Five benchmark datasets are considered to evaluate the performance of the proposed ImpClust algorithm. The experimental results obtained are compared with the five commonly used clustering algorithms. A total of five cluster validation indexes (DI-Index, COP-Index, DB-Index, CH-Index and Silhouette Index) are used to evaluate the clusters formed utilizing the different clustering algorithms. The experimental findings show that the proposed ImpClust algorithm achieves a significantly high score for Silhouette Index, DI-Index, and CH-Index whereas for COP-Index and DB-Index the proposed ImpClust algorithm achieves a significantly low score in comparison to the five existing clustering techniques.

**Keywords:** Drug discovery, Clustering, Lead compound, Screening, Chemical structures and chemoinformatics.

## 1. Introduction

The drug discovery process (DDP) is the process by which new medications are discovered and developed. It is a complex and lengthy process that can take many years to complete [1]. The DDP includes the following stages:

- *Target identification:* The identification of a particular molecule or biological process that plays a role in the disease is involved in this phase. It also includes the identification of a particular protein or enzyme having a key role in the disease process.
- *Lead compound identification:* After identifying a target, the next phase is to find a lead compound that can interact with the target and potentially provide a therapeutic benefit. This may involve screening large libraries of compounds to identify those that have the desired properties.
- *Lead compound optimization:* After identifying a lead compound, it is optimized to improve features like as potency, selectivity, and pharmacokinetics. Modifying the chemical structure of the substance or testing it in animal models to assess its safety and efficacy may be involved.
- *Preclinical testing:* To ensure the safety and efficacy of the drug, several preclinical tests are performed before it is tested on humans. This often entails evaluating the drug's pharmacology, toxicity, and potential adverse effects in animal models.
- *Clinical trials:* Once a drug candidate passes preclinical testing, it may proceed to clinical trials,

Received: June 7, 2023. Revised: September 8, 2023.

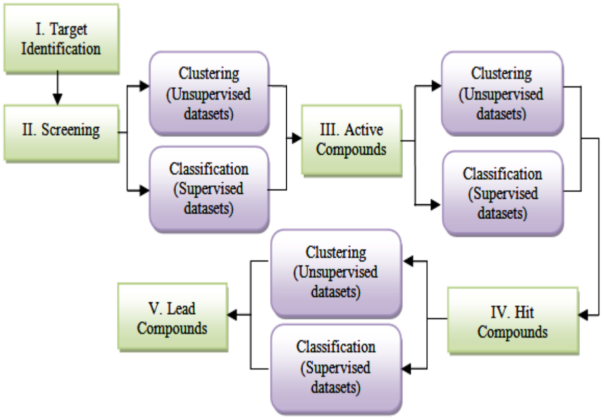
I. Target Identification
II. Screening
Clustering (Unsupervised datasets)
Classification (Supervised datasets)
III. Active Compounds
Clustering (Unsupervised datasets)
Classification (Supervised datasets)
IV. Hit Compounds
Clustering (Unsupervised datasets)
Classification (Supervised datasets)
V. Lead Compounds

Figure. 1 Determination of lead compound in the DDP

which entail testing the drug in humans to evaluate its safety and efficacy. Clinical trials are normally undertaken in three stages, with each comprising a greater number of participants.

- *Regulatory approval:* If a drug successfully completes all of these stages, it may be submitted to regulatory agencies for approval. Regulatory agencies will evaluate the safety and efficacy of the drug and decide whether to approve it for use in humans.
- *Post-marketing surveillance:* Even after a drug is approved for use, it continues to be monitored for safety and efficacy through post-marketing surveillance.

Clustering is a technique commonly used in DDP to identify and group similar molecules or compounds. This is a useful approach for analyzing large datasets of molecules and can be applied at different phases of the DDP.

In the lead identification stage, clustering can be used to group similar molecules based on their chemical structures, physicochemical properties, and biological activities. This helps to identify promising lead compounds that can be further optimized for potency and selectivity.

Clustering can also be used in virtual screening, where large databases of compounds are screened to identify molecules with potential activity against a target. Clustering can be used to group similar molecules together, which helps to identify common structural features and inform the design of new compounds [2].

Overall, clustering is a valuable tool in DDP that helps to identify and group similar molecules, leading to the discovery of new leads and the optimization of existing compounds.

The discovery of lead compounds is a vital stage in the DDP. It involves identifying a molecule or compound that has the potential to be developed into a drug that can effectively treat a specific disease or condition. Fig. 1 shows the various steps involved in the process of lead compound identification and these steps can be explained as follows [3].

1. *Target identification:* Determining a particular biological target implicated in a disease process, such as a protein or enzyme.
2. *Screening:* Conducting a high-throughput screening of vast libraries of compounds to discover those that interact with the target and have potential therapeutic activity.
3. *Hit-to-lead optimization:* Refining the initial hits to improve their potency, selectivity, pharmacokinetic properties, and safety profiles.
4. *Lead selection:* Selecting a few lead compounds with the best combination of potency, selectivity, pharmacokinetics, and safety for further development.
5. *Preclinical testing:* Preclinical investigations are being carried out in animal models to assess the safety and effectiveness of the lead compounds.

Once a lead compound has been identified and optimized, it can be developed into a drug candidate, which undergoes further testing and clinical trials before being approved for use in humans.

In the context of lead compound identification, clustering is typically performed on large libraries of compounds using various algorithms such as K-means or hierarchical clustering. The compounds are represented as vectors in a high-dimensional chemical space, where each dimension represents a structural feature or property of the compound. Once the compounds have been clustered, the next step is to select representative compounds from each cluster for further testing. The goal is to select a small number of compounds that are representative of the entire cluster and have the highest potential for further development [4].

This work's key contribution can be given as:

- A data splitting-based clustering algorithm, ImpClust, is proposed to cluster similar data items. The proposed ImpClust algorithm first determines the candidate medoid subset using Z_score. The data items within the candidate medoid subset are free from noises and outliers and hence, used as medoids for the formation of clusters.
- Comprehensive comparison tests on five benchmark datasets are carried out to validate the performance of the proposed ImpClust method.
- Standard cluster validation Indexes are used evaluate the new ImpClust algorithm's performance to the five existing clustering approaches.
- The proposed ImpClust algorithm achieves a significantly high score for Silhouette Index, DI-

Index, and CH-Index and a significantly low score for COP-Index and DB-Index.

- The proposed ImpClust can have applications for solving real-world quandaries such as drug identification, pattern recognition, social network analysis, market research and customer analysis, recommendation systems and so on.

The following sections of this manuscript are organized as follows: Section 2 discusses the existing clustering algorithms. Section 3 discusses the step-by-step explanation of the proposed algorithm. Section 4 discusses the methodology used to carry out the experimental work. Section 5 gives the analysis of the results and the last section 6 concludes the manuscript.

## 2. Related literature

Drug discovery is an important process for any pharmaceutical company for the discovery and development of new medication. Data clustering plays a significant part in the DDP for the identification of the lead compound. Although clustering can be a powerful tool for data analysis, there are several challenges that can arise during the clustering process. Some of these challenges are:

- *Determining the optimal number of clusters*
- *Handling high-dimensional data*
- *Dealing with noisy data*
- *Selection of initial cluster centers*

The researchers have come up with numerous data clustering techniques to efficiently cluster the unsupervised data thereby considering all the above challenges.

The Elbow method [5], the gap statistic method [6] and the average Silhouette score method [7] are the commonly used techniques that can efficiently determine the optimal number of clusters.

Any clustering algorithm's effectiveness is determined by how effectively the initial cluster centres are chosen. Selection of good initial cluster centers results in better performance of the clustering algorithm and poor selection of the initial cluster centers results in the worst performance of the clustering algorithm. The authors in [10] proposed the INCK algorithm, an improved version of the K-medoids clustering algorithm. Instead of a random selection of the initial medoids, the INCK algorithm selects the initial medoids from a set of chosen data objects that is free from noise and outliers. A step-increasing function that increases the number of clusters from 2 to the *k*-optimal value is used. The major limitation of the INCK algorithm is for every dataset requires manually determining the threshold value and the selection of the wrong threshold value results in poor cluster formation.

Similarly, the authors in [11] introduced two algorithms, ICCS_K-means and MNN (M nearest neighbours), both of which improved on the K-means method. The ICCS_K-means method is used to determine the *k*-optimal value for each dataset. The distance and density functions are used by the MNN algorithm to find out the initial cluster centers. The results show that both the ICCS_K-means and the MNN algorithms are able to form clusters of better quality as compared to the traditional K-means algorithm. To determine the *k*-optimal value for each dataset the ICCS_K-means algorithm needs to go through every single data point in the dataset which results in high time complexity of the algorithm.

The authors in [11A] proposed the similarity-based K-means clustering (SKC) approach, which depends on divergence distance for clustering attributes. The authors implemented clustering techniques based on the similarity of categorical data. It identified similarities between attributes of inter and intra-clusters to improve the performance of the proposed method. Pre-processing techniques were used to remove noise from the data and estimate similarities between noise-free elements. Insignificant attributes were removed, and relevant attributes were chosen from the pre-processed elements. The major limitation of the SKC algorithm is it performs well for datasets in which the data items are highly correlated and the performance degrades for datasets having low connectivity and cohesion.

Even if the clustering techniques are adequate, the presence of noise or outliers in the dataset always leads to insignificant cluster formation. The KNN algorithm is the best algorithm to group together similar data by removing noise and outliers but, for high dimensional datasets the algorithm cannot perform well. The authors of [13] introduced the POD (Parallel outlier detection) algorithm, which employs a weighted KNN approach to locate the *k*-nearest neighbours based on the Z-order curve. Regardless of the size of the datasets, the POD method performs better. Generally, in categorical or mixed type datasets the categorical features are converted into numerical values. This conversion will result in huge information loss within the dataset.

The problem of determining the clusters of arbitrary shapes and boundaries is solved by the algorithms based on the minimum spanning tree (MST). The authors in [15] proposed CTCEHC (MST-based hierarchical clustering algorithm) algorithm to handle the dataset having clusters of arbitrary shapes and boundaries. The CTCEHC is a three-stage algorithm that makes use of centroid to calculate the inter-cluster similarity of the minimum spanning tree, the geodesic distance between the centroids of the minimum spanning tree and to merge all the small sub-clusters, cut-edge constraint is used. The CTCEHC algorithm has good performance but it has a high time complexity of $O(n^2)$.

Connectivity and cohesion within the dataset are the two important factors to find out the similarity values. For the datasets having convex-shaped clusters, cohesion is an important factor to be taken into consideration. The K-means and K-medoids algorithms have a complete dependency on the connectivity within the dataset and partial dependency on the cohesion within the dataset hence, not suitable for datasets having convex structures. The authors in [16] proposed a gravity center clustering (GCC) that takes into consideration a center point within a cluster to find out the similarity for each data item within the dataset that holds both connectivity and cohesion factor. As a result, the GCC algorithm shows better performance than the traditional K-means, K-medoids and K-medians algorithms. The major limitation of the GCC algorithm is for the high dimensional datasets or the datasets having low cohesion and connectivity within the data items, it results in poor cluster formation. Similarly, the authors in [17] proposed a metaheuristic-based K-medoids Crow Search Algorithm (KMCSA) hybrid algorithm. The KMCSA algorithm is based on the crow search optimization technique to balance the exploration and exploitation phase of the K-medoids clustering technique hereby improving the clustering performance of the K-medoids algorithm. The key limitation of the KMCSA is its high time complexity compared to the K-medoids algorithm.

Taking the aforementioned issues into account, a partitioning-based statistical method ImpClust is proposed to efficiently cluster the data without being affected by the noise or outliers. Moreover, an in-depth description of the proposed ImpClust algorithm is provided in section 3.

## 3. Proposed ImpClust algorithm

The ImpClust method introduced here is a statistical partitioning-based clustering approach. The ImpClust technique is based on the concept of performing efficient data clustering by picking initial cluster centres that are unaffected by noise or outliers. The working of the proposed ImpClust algorithm can be divided in four main parts which are discussed below.

- *Candidate medoid selection:* Consider a dataset $X_p$ containing $n$ data items such that $X_p = \{y_1, y_2, y_3 \ldots\ldots y_n\}$ where each data item $y_i$ is having $m$ dimensions. The pair-wise distance for the dataset $X_p$ can be calculated using the Eq. (1).

$$dist(y_i, y_j) = \sqrt{\sum_{t=1}^{m}(y_i^t - y_j^t)} \qquad (1)$$

Eq. (2) calculates the divergence between any data item $y_i$ and the other data items (object variance ($\sigma_i$)) in the dataset $X_p$ is given by:

$$\sigma_i = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n} dist(y_i, y_j)^2} \qquad (2)$$

The overall object variance ($\sigma_{x_p}$) for all the data items within the dataset $X_p$ can be given by the Eq. (3).

$$\sigma_{x_p} = \{\sigma_1, \sigma_2, \sigma_3, \sigma_4, \ldots\ldots\ldots..\sigma_n\} \qquad (3)$$

The data items that are having high variances are generally considered as noises or outliers. The selection of a data item to be a medoid decreases as the value of variance for that particular data item increases.

Z-score, commonly referred to as standard score, is a statistical metric that quantifies the number of standard deviations a data point is from the mean of a data collection and may be calculated using Eq. (4).

$$Z\_score_i = \frac{\sigma_i - \bar{\sigma}}{\sigma_{stdr}} \qquad (4)$$

where, $\sigma_i$ represents the object variance of data item $y_i$, $\bar{\sigma}$ represents the average value of $\sigma_{x_p}$ and $\sigma_{stdr}$ is the standard deviation of the $\sigma_{x_p}$.

The candidate medoid ($F_m$) subset can be formed using the Eq. (5).

$$F_{\boldsymbol{m}} = \{y_i | Z\_score_i < \delta, i = 1, 2 \ldots\ldots n\} \qquad (5)$$

where, $\delta$ is the cut-off value. The possible values for $\delta$ is 2.5 or -2.5. If the $Z\_score$ value is less than -2.5 or greater than 2.5, the data item is classified as noise or an outlier.

The candidate medoids subset $F_m$ now contains data items from the dataset $X_p$ that are free from noise or outlier and may be used to choose starting medoids, enhancing clustering performance.

- *Initial medoid selection:* The selection of first medoid ($M_1$) from the candidate medoid subset $F_m$ is done using the Eq. (6).

$$M_1 = \underset{y_i \in F_m}{\text{argmin}}\{T_i | i = 1, 2, 3 \dots . n\} \quad (6)$$

where, the distance $T_i$ of the data item $y_i$ is given by Eq. (7).

$$T_i = \sum_{j=1}^{n} dist(y_i, y_j) \quad (7)$$

The selection of second medoid ($M_2$) from the candidate medoid subset $F_m$ can be done using the Eq. (8).

$$M_2 = \underset{y_i \in F_m}{\text{argmax}} \{dist(y_{i,} M_1) | i = 1, 2, 3 \dots . n\} \quad (8)$$

When two medoids, $M_1$ and $M_{2,}$ are chosen, two clusters for the dataset $X_p$ are formed. The next stage is to increase the number of medoids until the $k$ optimal value is reached.

- *K-optimal medoid selection*: As the next possible medoid, a data item from the current two clusters with the greatest distance value among all computed distances is chosen. If $C_i$ is the $i^{th}$ cluster having $M_i'$ medoid the next feasible medoid can be selected using the Eq. (9).

$$M_i' = \underset{y_t \in C_i \cap F_m}{\text{argmax}} \{dist(y_{t,} M_1) | t = 1, 2, 3 \dots . n\} \quad (9)$$

Using the above equation (9), we have a feasible medoid subset $M' = \{M_1', M_2', \dots \dots M_{k-1}'\}$ containing medoids selected from each cluster. The selection of the next feasible medoid from this subset can be done by using the Eq. (10).

$$M_{i+1} = \underset{M_j' \in M'}{\text{argmax}} \{dist(M_{j,} M_j') | j = 1, 2, 3 \dots . k\} \quad (10)$$

The medoid obtained from the above equation (10) is added to the medoid set $M_i$ so that there are $i+1$ number of medoids. Eqs. (9) and (10) are both repeated until $k$ optimal medoids are produced.

- *Cost function:* If $\mathrm{M} = \{\mathrm{M}_1, \mathrm{M}_2, \dots \dots \mathrm{M}_\mathrm{k}\}$ represents the feasible medoids subset used to form k optimal clusters $\mathrm{C}_1, \mathrm{C}_2, \mathrm{C}_3 \dots \dots \mathrm{C}_\mathrm{k}$, then to minimize the total cluster cost $\mathrm{E}_{\mathrm{Total}}$ Eq. (11) can be used.

$$E_{Total} = \sum_{i=1}^{k} \sum_{y \in C_i} dist(M_k, y)^2 \quad (11)$$

The step-by-step explanation of the proposed ImpClust algorithm is given in the Algorithm 1.

Algorithm 1: ImpClust algorithm

```
Input: A dataset $X_p$, optimal value of k and threshold value $\delta$.
Output: Clusters $C_1, C_2, \dots \dots \dots C_k$
Begin:
1.  Calculate the pair-wise distance between each pair of data items using Eq. (1).
2.  Calculate the object variance $\sigma_i$ using the Eq. (2) and $Z\_score_i$ using the Eq. (4), then determine the feasible medoids subset $F_m$ using the Eq. (5).
3.  Make selection of the first two medoids M = {$M_1$, $M_2$} using Eq. (6) and Eq. (8).
4.  Assign each data item to the nearest medoids and using Eq. (11) calculates the total cost $E_{Total}$.
5.  for $i \rightarrow 2$ to $k - 1$
6.  Calculate the next feasible medoid $M_{i+1}$ using the Eq. (10) and append the medoid to the original medoids set M.
7.  Repeat.
8.  Assign each data item to the nearest medoid utilizing the nearest distance principle.
9.  To update the medoid set M, calculate the sum of distances from all data items to their medoids and find a new medoid in each cluster which is a data item having minimum total distance to all other items in its cluster.
10. Update the current medoid in M by replacing it with the new medoid obtained in the above step 9.
11. Calculate the total cost of the cluster $E_{Total}$ using Eq. (11).
12. Continue until the total cost of $E_{Total}$ cluster no longer changes.
13. end for
End
```

Table 1. Total number of datasets used

| Datasets | Features | Instances | Classes |
|---|---|---|---|
| Ceramics | 19 | 88 | 2 |
| Biodegradation | 41 | 1055 | 2 |
| Wine | 13 | 178 | 3 |
| Glass | 10 | 214 | 6 |
| Milk_Quality | 8 | 1059 | 3 |

## 4. Proposed methodology

The methodology used in this study can be explained using three subsections. The first subsection gives the explanation of the benchmark datasets used in the experimentation. The second subsection gives a brief of the hardware and the software used and the last subsection represents the evaluation metrics to compare the performance of the proposed ImpClust algorithm.

### 4.1 Datasets

Table 1 represents the five publicly available benchmark datasets used for experimental purposes. The dimensions of these datasets vary from 8 to 41 and the number of data items varies from 88 to 1055. These datasets are freely accessible through the "UCI machine learning repository" and the "Kaggle repository."

### 4.2 Experimental setup

The targets (labels) of the datasets mentioned in Table 1 are omitted and all the data items are shuffled to get a pure unsupervised dataset for the purpose of experimentation. Five state-of-the-art algorithms (INCK, *K*-medoids, partitioning around medoids (PAM), *K*-means, and mini batch *K*-means [18, 19]) are utilized to compare the performance of the proposed ImpClust algorithm. The experiments are carried out on a laptop PC DELL XPS 13 Plus with an Intel Core i9 processor, 16GB RAM, 1TB SSD, Windows 11 Home, and 8GB graphics NVIDIA GeForce RTX 2070/300 Hz running Python IDE spyder version 5.2.2 in anaconda navigator.

### 4.3 Evaluation metrics

There are distinct approaches to find the optimal number of clusters (*k-optimal*) for the unsupervised datasets. Some of the commonly used such techniques are elbow method [20], Silhouette method [21], gap static method [22], sum of squares method [23], Clustree [24] method and so on. Because of its simplicity and accuracy, the Silhouette approach is employed in this study to discover the *k-optimal* values for each dataset. The silhouette score is calculated using *K*-means clustering for each dataset to obtain the optimal number of clusters (*k*). The *k-optimal* values for ceramics, biodegradation, wine, glass and Milk_quality datasets are 2, 2, 3, 6 and 3 respectively. To compare the performance of the proposed ImpClust clustering algorithm with the five existing techniques (INCK, *K*-medoids, partitioning around medoids (PAM), *K*-means and mini batch *K*-Means), five standard CVIs named Dunn's index (DI-Index), Ibai Gurrutxaga index (COP-Index), Davies Bouldin Index (DB-Index), Calinski Harabasz index (CH-Index) and silhouette index ) are used.

## 5. Results and analysis

The Silhouette index score obtained for different clustering techniques is shown in Table 2.

The Silhouette index score ranges from -1 to +1. The -1 value indicates that the data item is incorrectly assigned to the cluster, whereas the +1 value shows that the data item is optimally assigned to a cluster. A data item on the decision line of the two clusters that are contiguous has a Silhouette index score value of zero. From Table 2, it is observed that the proposed ImpClust approach achieved the maximum value of 0.68 while the PAM technique achieved the minimum value of 0.484 for the Ceramics dataset.

The proposed ImpClust approach achieved the maximum Silhouette index value of 0.51 for the Biodegradation dataset, whereas the INCK technique achieved the minimum value of 0.381. The proposed ImpClust approach achieved the maximum Silhouette index value of 0.588 for the Wine dataset, whereas the PAM approach achieved the minimum value of 0.488. The proposed ImpClust approach achieved the maximum Silhouette index value of 0.446 for the Glass dataset, whereas the PAM approach achieved the minimum value of 0.22. The proposed ImpClust approach achieved the maximum Silhouette index value of 0.351 for the Milk_Quality dataset, whereas the K-medoids technique achieved the minimum value of 0.254.

Table 2 concludes that the proposed ImpClust algorithm outperforms all the five existing clustering techniques after evaluating using Silhouette index.

Similarly, Fig. 2 visually depicts the comparison of the proposed ImpClust approach with existing clustering approaches for five benchmark datasets using Silhouette index as a CVI.

The DB index score obtained for the different clustering techniques is shown in Table 3. The DB index score has the lowest possible value of zero. A value close to zero shows better cluster formation whereas a high value indicates worst cluster formation.

According to Table 3, for the Ceramics dataset, the proposed ImpClust approach achieved the minimum DB index value of 0.413, while the Mini Batch K-Means approach achieved the maximum value of 0.689.

The proposed ImpClust approach produced the minimum DB index value of 0.479 for the Wine dataset, whereas the INCK technique obtained the maximum value of 0.576. The proposed ImpClust approach produced the minimum DB index value of 0.809 for the Glass dataset, whereas the PAM technique obtained the maximum value of 1.126. The proposed ImpClust approach obtained the minimum

Received: June 7, 2023. Revised: September 8, 2023.

Table 2. Silhouette index scores achieved using various clustering approaches

| **Datasets** | **ImpClust** | **K-means** | **K-medoids** | **PAM** | **INCK** | **Mini Batch K-means** |
|---|---|---|---|---|---|---|
| Ceramics | 0.680 | 0.553 | 0.520 | 0.484 | 0.545 | 0.522 |
| Biodegradation | 0.510 | 0.417 | 0.385 | 0.502 | 0.381 | 0.389 |
| Wine | 0.588 | 0.571 | 0.567 | 0.488 | 0.522 | 0.569 |
| Glass | 0.446 | 0.421 | 0.310 | 0.220 | 0.273 | 0.411 |
| Milk_Quality | 0.351 | 0.346 | 0.254 | 0.255 | 0.279 | 0.311 |

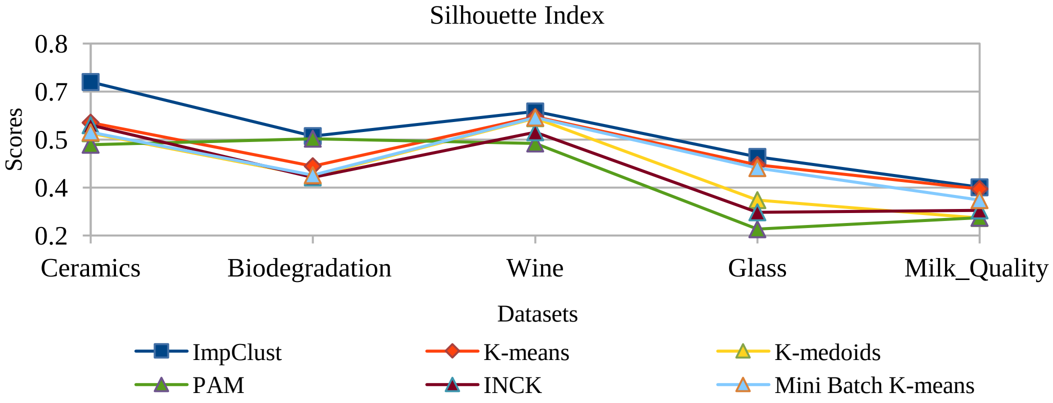

Figure. 2 Comparison of Silhouette index score obtained by the different clustering algorithms

Table 3. DB index scores obtained using different clustering techniques

| **Datasets** | **ImpClust** | **K-means** | **K-medoids** | **PAM** | **INCK** | **Mini Batch K-means** |
|---|---|---|---|---|---|---|
| Ceramics | 0.413 | 0.662 | 0.669 | 0.670 | 0.655 | 0.689 |
| Biodegradation | 0.653 | 0.957 | 0.964 | 0.612 | 1.019 | 0.973 |
| Wine | 0.479 | 0.534 | 0.529 | 0.546 | 0.576 | 0.537 |
| Glass | 0.809 | 0.891 | 1.011 | 1.126 | 0.827 | 0.904 |
| Milk_Quality | 0.920 | 1.258 | 1.314 | 1.395 | 1.349 | 1.243 |

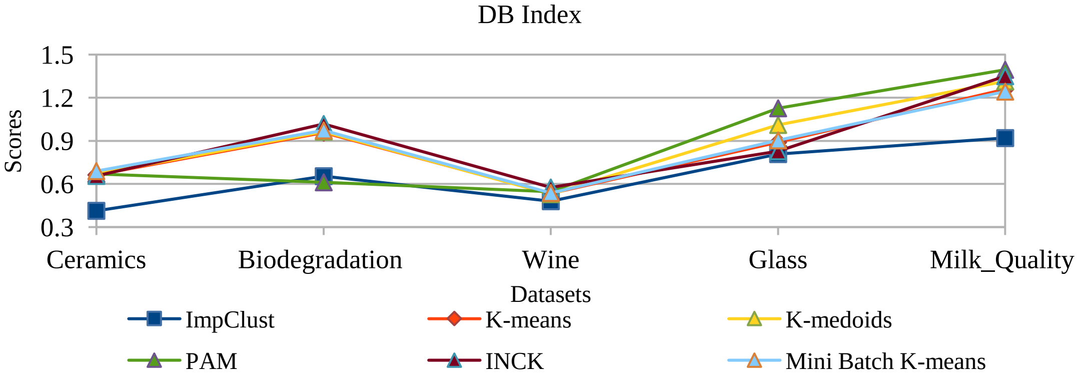

Figure. 3 Comparison of DB index score obtained by the different clustering algorithms

DB index value of 0.92 for the Milk_Quality dataset, whereas the PAM technique obtained the maximum value of 1.395.

Table 3 concludes that the proposed ImpClust algorithm outperforms four out of five datasets except for the Biodegradation dataset in which the PAM technique shows better cluster formation.

Similarly, Fig. 3 depicts a graphical comparison of the proposed ImpClust approach with known clustering algorithms for five benchmark datasets utilizing the DB index as a CVI.

The COP index score obtained for distinct clustering techniques is shown in Table 4. The COP index considers a cluster's cohesiveness, which is

Table 4. COP index scores obtained using different clustering techniques

| Datasets | ImpClust | K-means | K-medoids | PAM | INCK | Mini Batch K-means |
|---|---|---|---|---|---|---|
| Ceramics | 0.119 | 0.295 | 0.295 | 0.282 | 0.124 | 0.295 |
| Biodegradation | 0.202 | 0.209 | 0.206 | 0.203 | 0.210 | 0.206 |
| Wine | 0.129 | 0.118 | 0.123 | 0.134 | 0.122 | 0.116 |
| Glass | 0.180 | 0.214 | 0.257 | 0.261 | 0.307 | 0.218 |
| Milk_Quality | 0.206 | 0.338 | 0.263 | 0.361 | 0.396 | 0.359 |

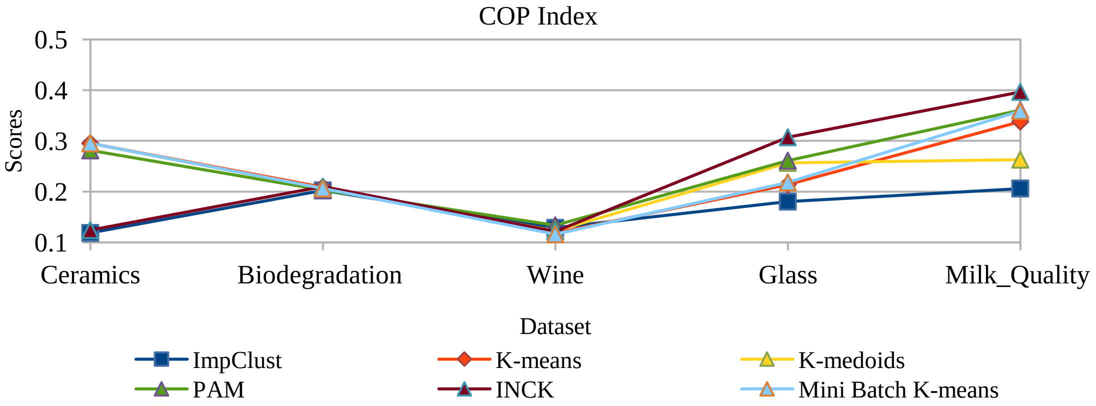

Figure. 4 Comparison of COP index score obtained by the different clustering algorithms

Table 5. Dunn index scores obtained using different clustering techniques

| Datasets | ImpClust | K-means | K-medoids | PAM | INCK | Mini Batch K-means |
|---|---|---|---|---|---|---|
| Ceramics | 0.500 | 0.365 | 0.365 | 0.010 | 0.054 | 0.365 |
| Biodegradation | 0.161 | 0.010 | 0.003 | 0.005 | 0.009 | 0.005 |
| Wine | 0.022 | 0.016 | 0.022 | 0.009 | 0.011 | 0.017 |
| Glass | 0.022 | 0.107 | 0.020 | 0.040 | 0.019 | 0.016 |
| Milk_Quality | 0.012 | 0.500 | 0.046 | 0.479 | 0.175 | 0.493 |

defined by the distance between the cluster centroid and all other data items, as well as its separation, which is calculated by the distance between the cluster's farthest neighbours. A clustering approach with the lowest COP index value results in superior cluster formation.

Table 4 demonstrates that the proposed ImpClust approach achieved the minimum value of 0.119 for the Ceramics dataset, while the K-means, K-medoids, and Mini Batch K-means algorithms obtained the maximum value of 0.295. The proposed ImpClust approach achieved the minimum COP index value of 0.202 for the Biodegradation dataset, while the INCK technique achieved the maximum value of 0.210. The Mini Batch K-means approach achieved the minimum COP index value of 0.116 for the Wine dataset, while the proposed ImpClust approach achieved 0.129 and the PAM approach achieved the maximum value of 0.134. For the Glass dataset, the proposed ImpClust approach achieved the minimum COP index value of 0.18, whereas the INCK technique achieved the maximum value of 0.307. For the Milk_Quality dataset, the proposed ImpClust approach achieved the minimum COP index value of 0.206, while the INCK approach achieved the maximum value of 0.396.

Table 4 concludes that the proposed ImpClust approach outperforms four out of five datasets except for the Wine dataset in which the Mini Batch K-means technique shows better cluster formation.

Similarly, Fig. 4 visually compares the proposed ImpClust approach to known clustering algorithms for five benchmark datasets using the COP index as a CVI.

The Dunn index score obtained for distinct clustering techniques is shown in Table 5. The Dunn index uses the ratio of inter-cluster distance to intra-cluster distance to get the value for a specific cluster. The Dunn index ranges from zero to infinity. A high Dunn index value implies that clusters develop more effectively.

According to Table 5, the proposed ImpClust approach achieved the greatest value of 0.5, while the PAM approach achieved the lowest value of 0.295 for

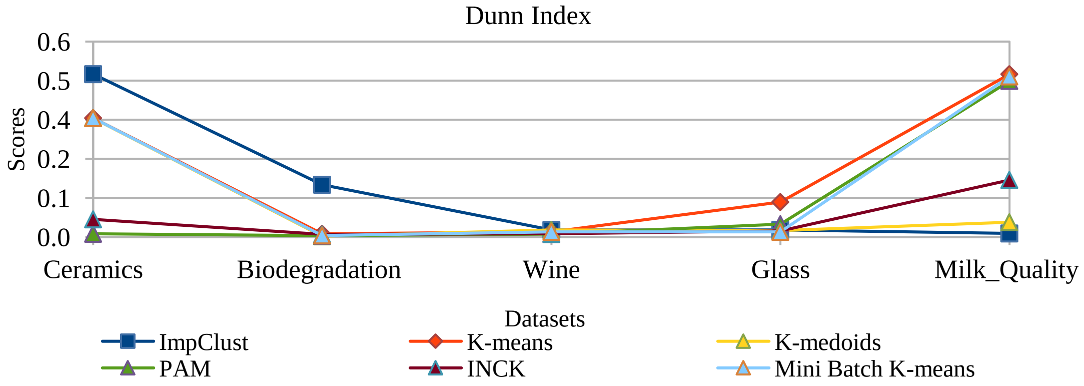

Figure. 5 Comparison of Dunn index score obtained by the different clustering algorithms

Table 6. CH index scores obtained using different clustering techniques

| Datasets | ImpClust | K-means | K-medoids | PAM | INCK | Mini Batch K-means |
|---|---|---|---|---|---|---|
| Ceramics | 500.679 | 59.022 | 59.022 | 12.757 | 420.595 | 59.022 |
| Biodegradation | 968.57 | 967.023 | 898.721 | 897.164 | 899.824 | 925.575 |
| Wine | 675.60 | 561.816 | 539.379 | 346.534 | 495.276 | 560.247 |
| Glass | 280.027 | 345.042 | 206.043 | 92.018 | 182.004 | 338.246 |
| Milk_Quality | 685.80 | 480.902 | 384.128 | 291.172 | 293.541 | 392.963 |

the Ceramics dataset. The proposed ImpClust approach achieved a maximum Dunn index value of 0.161 for the Biodegradation dataset, whereas the K-medoids technique obtained a minimum value of 0.003. The proposed ImpClust approach with the K-medoids technique produced a maximum Dunn index value of 0.022 for the Wine dataset, while the PAM technique obtained a minimum value of 0.009. For the glass dataset, the K-means approach achieved a maximum Dunn index value of 0.107, the proposed ImpClust approach achieved a minimum value of 0.016, and the mini batch K-means achieved a maximum value of 0.022. The K-means approach achieved a maximum Dunn index value of 0.5 for the Milk_Quality dataset, while the proposed ImpClust approach achieved a minimum value of 0.012.

Table 5 concludes that the proposed ImpClust algorithm outperforms three out of five datasets except for the Glass and Milk_quality dataset in which the K-means technique shows better cluster formation.

Similarly, Fig. 5 visually compares the proposed ImpClust approach to known clustering algorithms for five benchmark datasets using the Dunn index as a CVI.

The CH index score obtained for distinct clustering techniques is shown in Table 6. The Variance Ratio Criterion commonly referred to as the CH index, can be defined as the ratio of between-cluster to within-cluster dispersion. The greater the value of the CH index, the better the cluster formation.

According to Table 6, the proposed ImpClust approach achieved a maximum value of 500.7, while the PAM approach achieved a minimum value of 12.76 for the Ceramics dataset. The proposed ImpClust approach came up with the maximum CH index value of 968.57 for the Biodegradation dataset, while the PAM approach achieved the minimum value of 897.164. The proposed ImpClust approach achieved the maximum CH index value of 675.6 for the Wine dataset, while the PAM approach achieved the minimum value of 346.5. For the Glass dataset, the K-means approach achieved the maximum CH index value of 345.04, while the proposed ImpClust approach achieved the CH index value of 280.03, and the PAM technique achieved the minimum value of 92.02.

For the Milk_Quality dataset, the proposed ImpClust algorithm obtained the maximum CH index value of 685.8 while the PAM technique obtained the minimum value of 291.2. Table 6 demonstrates that the proposed ImpClust approach surpasses four of the five datasets, with the exception of the Glass dataset, where the K-means technique achieves better cluster formation.

Fig. 6 compares the proposed ImpClust approach to known clustering algorithms for five benchmark datasets using the CH index as a CVI.

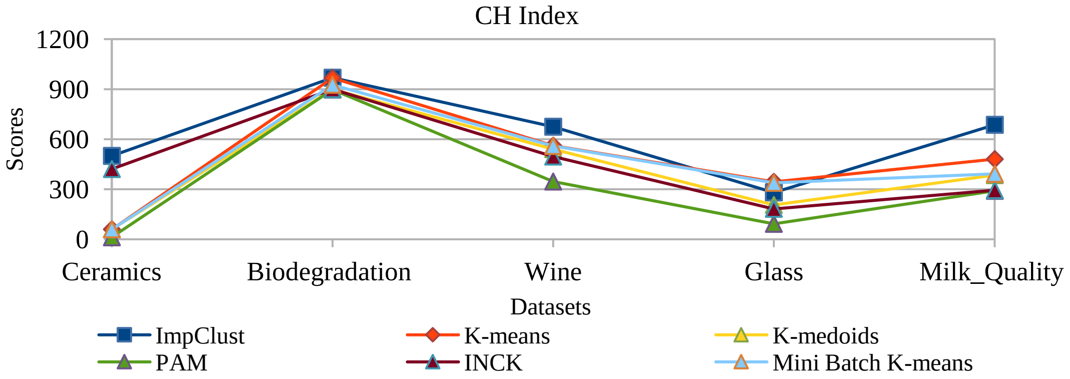

Figure 6. Comparison of CH index score obtained by the different clustering algorithms

Table 7. Number of datasets where various clustering techniques surpass one another

| Datasets | ImpClust | K-means | K-medoids | PAM | INCK | Mini Batch K-means |
|---|---|---|---|---|---|---|
| Silhouette Index | 5 | 0 | 0 | 0 | 0 | 0 |
| DB Index | 4 | 0 | 0 | 1 | 0 | 0 |
| COP Index | 4 | 0 | 0 | 0 | 0 | 1 |
| Dunn Index | 3 | 2 | 0 | 0 | 0 | 0 |
| CH Index | 4 | 1 | 0 | 0 | 0 | 0 |

Table 7 summarizes the above results achieved with different CVIs. It clearly shows how many datasets various clustering techniques outperform. According to Table 7, the proposed ImpClust approach achieves the best cluster formation for the maximum number of datasets.

## 6. Conclusion

Data clustering plays an important role in the DDP to identify and group similar molecules or compounds. In this study, an improved clustering algorithm ImpClust is proposed to cluster.

The proposed ImpClust algorithm takes into consideration the candidate medoids subset for the selection of initial medoids to form clusters of better quality. The performance of the proposed ImpClust algorithm is compared with the five existing clustering techniques (K-means, K-medoids, PAM, INCK and mini batch K-means) by using five benchmark chemical datasets (ceramics, biodegradation, wine, glass and Milk_Quality).

To validate the clusters formed using different clustering techniques, five CVIs (silhouette index, DB index, COP index, Dunn index, CH index) are used. The experimental findings show that the proposed ImpClust algorithm achieves a significantly high score for silhouette index, DI-index, and CH-index whereas for COP-index and DB-index the proposed ImpClust algorithm achieves a significantly low score in comparison to the five existing clustering techniques. The results obtained from the experimentation prove that the proposed ImpClust algorithm perform much well as compared to five existing clustering techniques.

## Conflicts of interest

The authors declare no conflict of interest.

## Author contributions

The contributions of authors are as follows: Conceptualization, Hutashan V. Bhagat; Methodology, Hutashan V. Bhagat; Software, Hutashan V. Bhagat; Validation, Hutashan V. Bhagat; Formal analysis, Hutashan V. Bhagat; investigation, Hutashan V. Bhagat; Data Curation, Dr. G. Uday Kiran; Writing-original draft preparation, Hutashan V. Bhagat; validation, Dr. G. Uday Kiran; supervision, Dr. Manminder Singh; project administration**,** Dr. Manminder Singh.

## References

[1] W. Chen, X. Liu, S. Zhang, and S. Chen, “Artificial intelligence for drug discovery: Resources, methods, and applications”, *Mol. Ther. Nucleic Acids*, Vol. 31, pp. 691–702, 2023.

[2] Z. Liu, R. A. Roberts, M. L. Nag, X. Chen, R. Huang, and W. Tong, “AI-based language

models powering drug discovery and development", *Drug Discov. Today*, Vol. 26, No. 11, pp. 2593–2607, 2021.

[3] J. Yu, X. Li, and M. Zheng, "Current status of active learning for drug discovery", *Artif. Intell. Life Sci.*, Vol. 1, No. 100023, p. 100023, 2021.

[4] P. Carracedo-Reboredo, J. Liñares-Blanco, N. Rodríguez-Fernández, F. Cedrón, F.J. Novoa, A. Carballal, V. Maojo, A. Pazos and C. Fernandez-Lozano, "A review on machine learning approaches and trends in drug discovery", *Comput. Struct. Biotechnol. J.*, Vol. 19, pp. 4538–4558, 2021.

[5] C. Yuan and H. Yang, "Research on K-value selection method of K-means clustering algorithm", *J*, Vol. 2, No. 2, pp. 226–235, 2019.

[6] T. V. S. Krishna, A. Y. Babu, and R. K. Kumar, "Determination of optimal clusters for a non-hierarchical clustering paradigm K-means algorithm", In: *Proc. of International Conference on Computational Intelligence and Data Engineering*, Singapore: Springer Singapore, pp. 301–316, 2018.

[7] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic", *J. R. Stat. Soc. Series B Stat. Methodol.*, Vol. 63, No. 2, pp. 411–423, 2001.

[8] D. Yu, G. Liu, M. Guo, and X. Liu, "An improved K-medoids algorithm based on step increasing and optimizing medoids", *Expert Syst. Appl.*, Vol. 92, pp. 464–473, 2018.

[9] Y. Li, J. Cai, H. Yang, J. Zhang, and X. Zhao, "A novel algorithm for initial cluster center selection", *IEEE Access*, Vol. 7, pp. 74683–74693, 2019.

[10] P. Kumar and A. Kanavalli, "A similarity based K-means clustering technique for categorical data in data mining application", *Int. J. Intell. Eng. Syst.*, Vol. 14, No. 2, pp. 43–51, 2021, doi: 10.22266/ijies2021.0430.05.

[11] Y. Ma and X. Zhao, "POD: A parallel outlier detection algorithm using weighted *k*NN", *IEEE Access*, Vol. 9, pp. 81765–81777, 2021.

[12] Y. Ma, H. Lin, Y. Wang, H. Huang, and X. He, "A multi-stage hierarchical clustering algorithm based on centroid of tree and cut edge constraint", *Inf. Sci. (Ny)*, Vol. 557, pp. 194–219, 2021.

[13] F. H. Kuwil, Ü. Atila, R. A. Issa, and F. Murtagh, "A novel data clustering algorithm based on gravity center methodology", *Expert Syst. Appl.*, Vol. 156, No. 113435, p. 113435, 2020.

[14] N. Sureja, B. Chawda, and A. Vasant, "An improved K-medoids clustering approach based on the crow search algorithm", *J. Comput. Math. Data Sci.*, Vol. 3, No. 100034, p. 100034, 2022.

[15] D. Sculley, "Web-scale k-means clustering", In: *Proc. of the 19th International Conference on World Wide Web*, 2010.

[16] E. Schubert and P. J. Rousseeuw, "Faster k-medoids clustering: Improving the PAM, CLARA, and CLARANS algorithms", *Similarity Search and Applications*, pp. 171–187, 2019.

[17] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, "Integration K-means clustering method and elbow method for identification of the best customer profile cluster", *IOP Conf. Ser. Mater. Sci. Eng.*, Vol. 336, p. 012017, 2018.

[18] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis", *J. Comput. Appl. Math.*, Vol. 20, pp. 53–65, 1987.

[19] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic", *J. R. Stat. Soc. Series B Stat. Methodol.*, Vol. 63, No. 2, pp. 411–423, 2001.

[20] R. Nainggolan, R. P. Angin, E. Simarmata, and A. F. Tarigan, "Improved the performance of the K-means cluster using the Sum of Squared Error (SSE) optimized by using the elbow method", *J. Phys. Conf. Ser.*, Vol. 1361, No. 1, p. 012015, 2019.

[21] P. Kranen, I. Assent, C. Baldauf, and T. Seidl, "The ClusTree: indexing micro-clusters for anytime stream mining", *Knowl. Inf. Syst.*, Vol. 29, No. 2, pp. 249–272, 2011.