EPiC Series in Computing

Volume 101, 2024, Pages 39–48

Proceedings of the 16th International Conference on Bioinformatics and Computational Biology (BICOB-2024)

Disease Clustering with Process Annotations from Gene Ontology

Drew Brady and Hisham Al-Mubaid
University of Houston – Clear Lake, Houston, TX, USA.
Bradyd6193@uhcl.edu, Hisham@uhcl.edu

Abstract

This paper presents a disease clustering approach by utilizing the biological process annotations from the Gene Ontology as the only data source for clustering diseases. As a result, a disease within a cluster should be more similar to all other diseases in the same cluster than to any disease in other clusters. Essentially, the clustering task is an unsupervised machine learning technique that attempts to discover and learn some hidden patterns from the disease information to place similar diseases together in the same cluster. We used two independent validations to examine our results. We examined the path length between disease pairs in the same cluster versus pairs in two separate clusters by utilizing semantic relationships from the Disease Ontology. We also utilized recently published results on disease similarity from a comprehensive study. Our experimental results are highly encouraging and highly agree with both validation methods. Specifically, most diseases placed in one cluster by our method are more similar to one another than to any disease in the other cluster, according to the validation results.

1 Introduction

The clustering task is one of the most important data analysis and machine learning processes with many applications in the bioinformatics field [1-4]. Clustering is an unsupervised machine learning task that attempts to place data items that are most similar to one another in the same group or cluster based on shared attributes or characteristics. This paper focuses on disease clustering, which involves placing diseases with similar features (i.e., similar diseases) in the same cluster with a total of two or more clusters. Every disease in a cluster should be more similar to all other diseases within the same cluster than to any disease in other clusters. The similarity between diseases is assessed based on the data source and disease information utilized in the disease clustering task. One of the most commonly used sources of information for disease similarity and clustering tasks is the set of disease genes, which consists of all genes known to be associated with the given diseases [1, 2, 4, 5]. Disease clustering is a crucial task in bioinformatics for exploring and understanding disease mechanisms at the molecular and functional levels [3, 4, 6-9]. Additionally, disease clustering outcomes are commonly used to analyze

H. Al-Mubaid, T. Aldwairi and O. Eulenstein (eds.), BICOB-2024 (EPiC Series in Computing, vol. 101), pp. 39–48

the relationship between diseases or groups of diseases [2, 5, 6, 8, 10]. This analysis is particularly relevant for human diseases for applications like drug repurposing and precision medicine.

As an unsupervised learning task, the clustering process works with unlabeled data to identify similarities between data items (i.e., diseases) not directly indicated by the attributes [9, 11, 12]. This paper presents a disease clustering process using the *biological process* (bp) annotation data for diseases. In the proposed method, we employ the bp annotations assigned to the disease via their genes as a functional profile of diseases for the clustering task.

The previous literature on disease similarity and disease-disease association relies on various kinds of disease information and attributes like disease-genes, disease-symptoms, disease-chemicals, gene expression profiles, protein-protein interactions, gene pathways, and more [4, 5, 9, 12, 13]. As a data type for determining disease associations, the bp annotation data has not been investigated extensively in disease similarity and clustering research. In this paper, we used the *Disease Ontology* (DO) to analyze the semantic relationships among the diseases [14]. To validate the results of our proposed method, we obtained the semantic relationships between diseases from DO and compared them with our results. Additionally, we used the disease similarity results from the fusion multi-view human disease network (MV-HDN) by Yang et al. (2023) to compare and validate our results [4]. The clustering results of our method are highly encouraging. The experimental results show that the disease process annotation profiles, which utilize the *Gene Ontology* (GO) bp annotations, are reliable data sources for disease clustering [15]. The reported results and experiments were conducted with a good number of diseases and various evaluation settings for disease clustering, utilizing only the disease process annotations from GO [15].

2 Background and Related Work

The most widely used clustering algorithms for data clustering in similar applications include K-means clustering, Agglomerative Hierarchical clustering, and DBSCAN [10, 11, 16]. Agglomerative hierarchical clustering belongs to the hierarchical clustering family, whereas the K-means clustering method is centroid-based, and DBSCAN is a density-based clustering method [1, 17].

In [3], Mathur et al. (2012) propose three methods for evaluating disease similarity by comparing the use of disease-gene and disease-bp associations (GO biological processes) for estimating the similarity between disease pairs [3]. For gene-based similarity, they proposed a method called gene-identity based (GIB), which utilizes gene sets associated with both diseases [3]. For process-based similarity, they proposed two methods: process-identity based (PIB) and process-similarity based (PSB). Both approaches utilize the GO processes associated with both diseases, but PSB also incorporates semantic information.

A comprehensive study and survey on the various clustering algorithms and their applications in several fields is presented in [18] by Anand and Kumar (2022). In [1], Karim et al. (2021) also present a thorough review of clustering algorithms. Their work further analyzed certain clustering algorithms based on deep learning techniques, known as deep learning-based clustering, by applying them to data from three bioinformatics tasks: bioimaging, cancer genomics, and biomedical text mining [1]. In another project, Santamaria et al. (2021) present a disease similarity approach that analyzes various disease sets associated with different biological features like genes, proteins, metabolic pathways, and genetic variants to build comprehensive disease models [8]. Their study utilizes and compares various distance metrics and clustering algorithms to the given disease sets.

Another disease relationship and similarity method using the interactome network is presented in [6] by Menche et al. (2015). They found that the network-based location of a disease module determines its pathobiological relationship to other diseases [6]. Moreover, they found that diseases exhibit higher similarity of their associated genes if they are closer in the interactome using GO annotation-based

similarity [6]. In a very recent project, with in-depth disease similarity analysis, Yang et al. (2023) present a disease-disease association network based on multi-view fusion (MV-HDN: Multiview Human Disease network) [4]. They produced the fusion MV-HDN by applying a similarity network fusion (SNF) model to three single-view networks that represent biological processes (B-HDN), phenotypic characteristics (S-HDN), and gene expressions (M-HDN).

3 Methodology

In the unsupervised machine learning task of clustering, we are interested in gathering similar diseases together into two or more bins, where each bin is called a *cluster*. Clustering algorithms such as K-means or Hierarchical clustering will attempt to identify and learn underlying patterns to group similar items together in meaningful categories.

A. Notations and Problem Specification:

Given a set D of n disease vectors: $D = \{d_1, d_2, \ldots, d_n\}$, each d_i represents an m-dimensional feature vector for each disease I, and the total number of diseases is n. A process annotation term assigned to disease i is encoded as a component d_{ij} in the vector d_i. Therefore, the functional profile of disease i is the feature vector d_i, which is made from the process annotations of disease i. In this work, we cluster the diseases in set D into two or more clusters based on their encoded process annotations.

B. The Clustering Step:

In this work, disease clustering involves representing each disease as a feature vector of disease process annotations. The clustering method then assigns diseases to clusters based on these representations. This process is an unsupervised machine learning task that involves grouping data points into two or more clusters, where the data points in each cluster are similar to one another. The goal of disease clustering is to group n data samples into k clusters, where each data sample is a disease represented as a feature vector [1, 7, 18].

In the K-means clustering method, each cluster is represented by a centroid μ_j, where $j = 1, \ldots, k$. This clustering process places the data points d_i into k clusters. For our analysis, we limited the number of clusters to $k = 2$ and $k = 3$. Our goal is to focus on which diseases will be placed together by the clustering algorithm. We then assess whether the diseases assigned to the same cluster are semantically more similar according to an independent source (e.g., DO). In general, most research and studies use data integration approaches with multiple data sources (i.e., several data types) to evaluate disease similarity and clustering. However, we only utilize one data type in this study, the disease process annotations. Therefore, we examine the effectiveness of this single information source in assessing disease similarity for categorizing similar diseases in the disease clustering task.

4 Evaluation and Results

In our evaluation of the proposed method, we used the disease process annotation data with the K-means clustering algorithm [11, 17]. We conducted two sets of evaluations, each utilizing a distinct verification method. The first evaluation used the hierarchical classifications from DO, while the second used the disease similarity results from the fusion MV-HDN study [4, 14].

For the first evaluation, we used various diseases with corresponding associations in the OMIM and DO ontologies [14, 19, 20]. We conducted ten experiments, each with five sets of randomly selected diseases. We adjusted the K-means clustering configuration for each disease set, using $k = 2$ and $k = 3$ clusters. Then, we randomly selected diseases to form verification, or validation, pairs within a cluster (intra-cluster) and between clusters (inter-cluster) to evaluate the cluster quality. Each experiment included ten verification pairs, consisting of five intra-cluster and five inter-cluster comparisons. Similarly, 20 verification pairs refer to ten intra-cluster and inter-cluster comparisons each. In Table 1, the experimental configurations for these ten experiments are detailed. Further, each set of experiments utilizes the same diseases. For example, the first two experiments were conducted on the same set of 20 diseases (refer to Table 1).

The *Disease Ontology* (DO) uses the *is_a* hierarchy to establish semantic relationships among diseases [14]. This hierarchy is represented with edges in a hierarchical structure, as shown in Figure 1. We utilized the ontology hierarchical structure to obtain the path length between two disease nodes with edge counting. For example, in Figure 1, the hierarchical relationship is illustrated for an intra-cluster disease pair in the first experiment, Exp1-A: *autosomal dominant intellectual developmental disorder 40* (DOID:0070070) and *posterior polymorphous corneal dystrophy 2* (DOID:0110856). With edge counting, the shortest path length between these two diseases is *three*. After measuring the path length for each verification pair, we calculated the average path length for both intra-cluster and inter-cluster pairs for comparison. In Table 2, a summary of the mean path lengths for the first four experiments is provided.

Shorter path lengths within a cluster indicate a higher degree of similarity among diseases, according to the DO [14]. This is further demonstrated in Figures 2 and 3, which show the average path lengths for disease verification pairs within and between clusters in the first evaluation. As illustrated in Table 2 and Figure 2, in all five clustering tests with $k = 2$ clusters, all intra-cluster disease pairs have a shorter path length in DO, indicating that they are more similar compared to inter-cluster pairs.

Set Number	No. of Diseases	Experiment Number	No. of Clusters	No. of Verif. Pairs
Set 1	20	1	2	10
		2	3	10
Set 2	40	3	2	20
		4	3	20
Set 3	100	5	2	20
		6	3	20
Set 4	20	7	2	20
		8	3	20
Set 5	40	9	2	20
		10	3	20

Table 1: Overview of the experimental design for each experiment (1-10).

Experiment Number	Intra-Cluster Distance	Inter-Cluster Distance
1	3.4	5.8
2	3.2	4.6
3	4.0	4.6
4	3.4	3.8

Table 2: Mean intra-cluster and inter-cluster path lengths for the first four experiments.

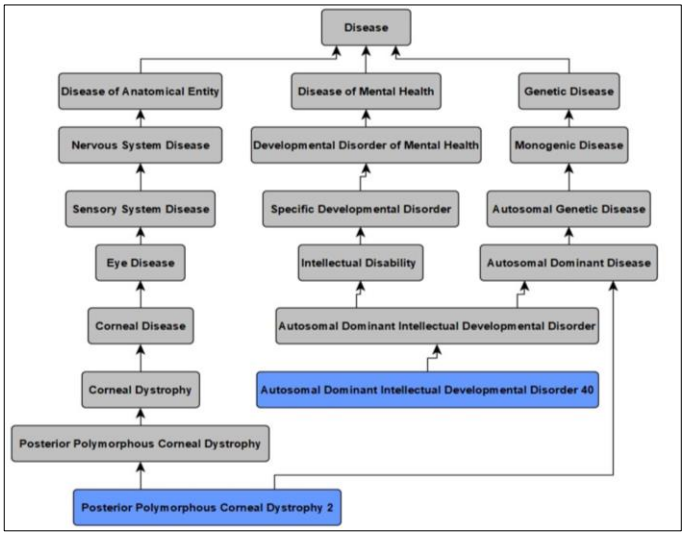

Figure 1: Illustration of the DO hierarchical relationship between *Autosomal Dominant Intellectual Developmental Disorder 40* (DOID:0070070) and *Posterior Polymorphous Corneal Dystrophy 2* (DOID:0110856).

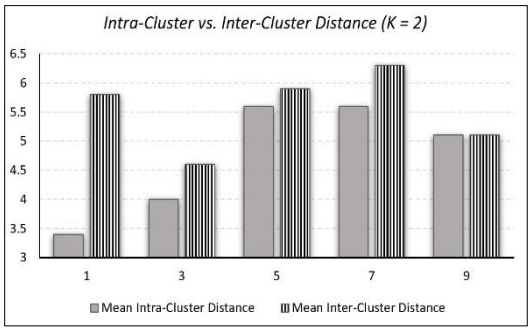

Figure 2: This graph displays a comparison of the average path lengths for intra-cluster and inter-cluster disease pairs for experiments using *k = 2* clusters in the first evaluation.

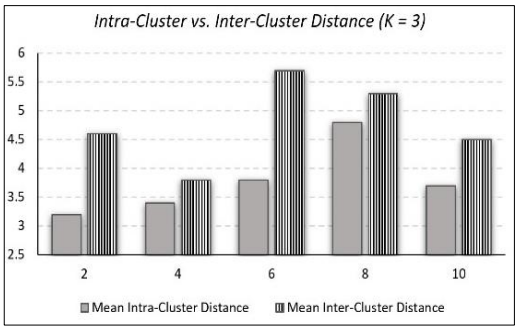

Figure 3: This graph displays a comparison of the average path lengths for intra-cluster and inter-cluster disease pairs for the experiments using *k = 3* clusters in the first evaluation.

In another evaluation setting to examine the effectiveness of the clustering based on disease process annotations, we conducted the second evaluation using a different form of validation for a total of 15 experiments. These experiments involved diseases that are either an exact match or closely related descendants of the diseases used to test the fusion MV-HDN, which have associations in both MeSH and DO ontologies [4, 14, 19]. The descendent diseases we used in these experiments share the same MeSH descriptors as their corresponding diseases in fusion MV-HDN to ensure a consistent basis for comparison. For this second evaluation, we followed a similar validation approach to the first evaluation. We conducted ten experiments using *k = 2* and *k = 3* clusters for each. Additionally, we included five more experiments with only *k = 2* clusters. Similar to the first evaluation, we randomly selected verification pairs for cluster quality assessments. The first set of ten diseases used in the first two clustering experiments is shown in Table 3, with the third column providing the number of process annotations for each disease. Furthermore, Table 4 contains the first six disease process annotations for the ten diseases used in experiments 1 and 2.

	Disease ID	No. of Process Annotations	Total Number of Components
1	DOID:11476	239	1241
2	DOID:2841	812	1241
3	DOID:9119	771	1241
4	DOID:0060060	277	1241
5	DOID:9970	545	1241
6	DOID:7148	463	1241
7	DOID:12449	467	1241
8	DOID:9744	540	1241
9	DOID:4450	57	1241
10	DOID:9074	455	1241

Table 3: The ten diseases used in the first and second (1-A and 1-B) experiments of the second evaluation.

Disease ID	Process Annotation Terms					
	GO:0000075	GO:0000077	GO:0000122	GO:0000165	GO:0000302	GO:0000723
DOID:11476	0	0	0	0	1	0
DOID:2841	0	0	1	1	0	0
DOID:9119	0	0	1	1	0	1
DOID:0060060	0	0	0	0	0	0
DOID:9970	0	0	1	0	1	0
DOID:7148	0	0	1	0	0	0
DOID:12449	1	1	1	0	0	1
DOID:9744	0	0	0	0	1	0
DOID:4450	0	0	0	0	0	0
DOID:9074	1	1	0	0	1	0

Table 4: The disease process annotations for the first set of 10 diseases that were used in the clustering experiments. Only the first six process annotations terms are shown here.

After clustering using the proposed method, we utilized the disease similarity results from a recent study by Yang et al. (2023) to obtain similarity values between disease verification pairs [4]. We used the similarity values computed between diseases in the MV-HDN similarity matrix to validate our cluster quality. In Tables 5 and 6, the validation results with the individual similarity scores of the verification pairs in experiment 1 are detailed.

With this verification method, a higher mean similarity value indicates a stronger correlation between diseases within a cluster. The following figures, Figures 4 through 7, further demonstrate these findings. Figure 4 presents the clustering results for the first ten experiments of the second evaluation using $k = 2$ clusters. Similarly, Figure 5 displays the clustering results for the first ten experiments of the second evaluation using $k = 3$ clusters. Figure 6 illustrates the results for the last five experiments of the second evaluation, which only employ $k = 2$ clusters. Finally, Figure 7 illustrates a comparison between the mean intra-cluster and inter-cluster similarity scores for all 15 experiments using $k = 2$

clusters in the second evaluation. These results demonstrate that diseases in the same cluster have higher similarity values than those in different clusters (i.e., inter-clusters).

5 Conclusion

This paper presents a simple method for disease clustering by utilizing a single data source, which is the disease process annotations from the Gene Ontology. Disease clustering outcomes are important for understanding the disease mechanisms at various molecular levels and for disease relationship analysis studies. One of the goals of this paper is to examine the effectiveness of a single source of data, which is the biological process taxonomy from the Gene Ontology, and the results are encouraging. The experimental results demonstrate that similar diseases are effectively grouped together. Our proposed method placed diseases that were closer and more similar to each another in the same cluster, as demonstrated by disease similarity in the fusion MV-HDN study and based on closeness (i.e., proximity) in the Disease Ontology.

Acknowledgements

This paper is based upon work supported by the National Science Foundation under grant no. 1928622.

Pair No.	Disease Pair	Cluster Type	Similarity Score
1	Osteoporosis, Aplastic Anemia	Intra-Cluster	0.000271948
2	Acute Myeloid Leukemia, Asthma	Intra-Cluster	0.000268082
3	Type 1 Diabetes Mellitus, Systemic Lupus Erythematosus	Intra-Cluster	0.016875417
4	Non-Hodgkin Lymphoma, Renal Cell Carcinoma	Intra-Cluster	0.009335265
5	Rheumatoid Arthritis, Type1 Diabetes Mellitus	Intra-Cluster	0.001084925
Mean Intra-Cluster Similarity Score			0.0055671

Table 5: The similarity values for the intra-cluster groupings of experiment 1 of the second evaluation. The last row details the average similarity score for intra-cluster pairs.

Pair No.	Disease Pair	Cluster Type	Similarity Score
6	Aplastic Anemia, Asthma	Inter-Cluster	0.000233477
7	Acute Myeloid Leukemia, Systemic Lupus Erythematosus	Inter-Cluster	0.000273618
8	Obesity, Type 1 Diabetes Mellitus	Inter-Cluster	0.007875158
9	Rheumatoid Arthritis, Osteoporosis	Inter-Cluster	0.003203446
10	Renal Cell Carcinoma, Systemic Lupus Erythematosus	Inter-Cluster	0.000265503
Mean Inter-Cluster Similarity Score			0.0023702

Table 6: The similarity values for the inter-cluster groupings of experiment 1 of the second evaluation. The last row details the average similarity score for inter-cluster pairs.

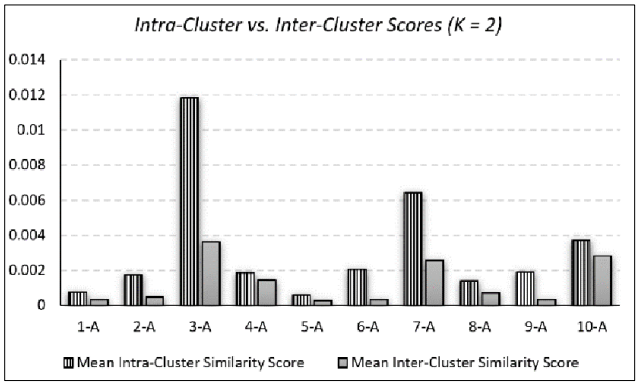

Figure 4: Illustration of the average intra-cluster and inter-cluster similarity scores of disease pairs for the second evaluation experiments using $k = 2$ clusters.

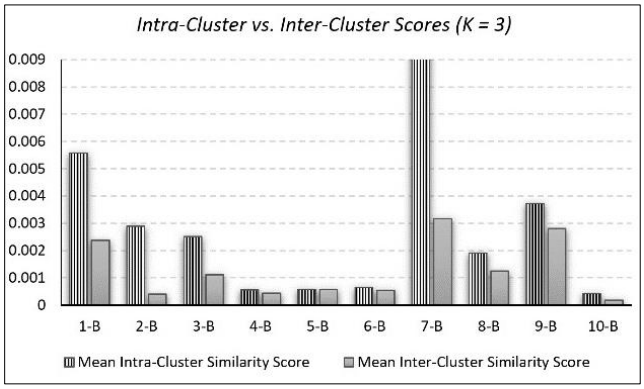

Figure 5: Illustration of the average intra-cluster and inter-cluster similarity scores of disease pairs for the second evaluation experiments using $k = 3$ clusters.

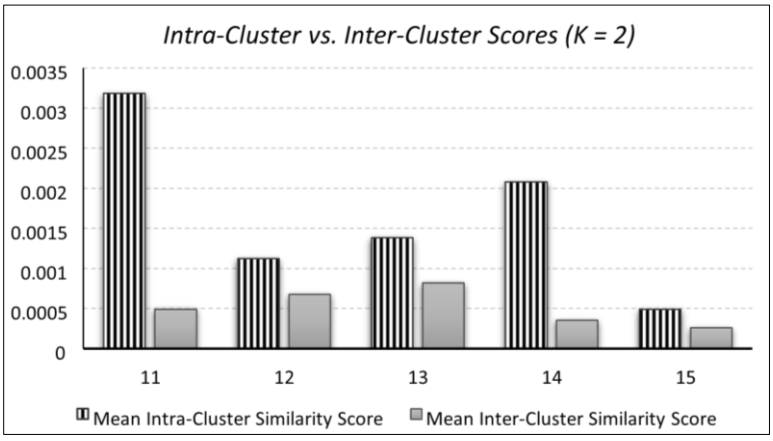

Figure 6: This figure compares the average intra-cluster and inter-cluster similarity scores for the final five experiments (21-25) that only use $k = 2$ clusters.

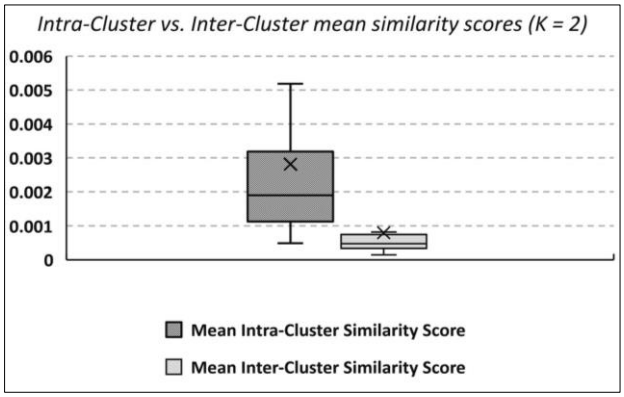

Figure 7: Illustration of the mean intra-cluster and inter-cluster similarity scores for the second evaluation experiments using $k = 2$ clusters.

References

[1] M. R. Karim *et al.*, "Deep learning-based clustering approaches for bioinformatics," *Briefings in Bioinformatics*, vol. 22, no. 1, pp. 393–415, Feb. 2020, doi: 10.1093/bib/bbz170.

[2] L. P. Santamaría *et al.*, "Analysis of new nosological models from disease similarities using clustering." *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 183-188, Apr. 2020. doi: 10.1101/2020.04.10.035394.

[3] S. Mathur and D. Dinakarpandian, "Finding disease similarity based on implicit semantic similarity," *Journal of Biomedical Informatics*, vol. 45, no. 2, pp. 363–371, Apr. 2012, doi: 10.1016/j.jbi.2011.11.017.

[4] X. Yang *et al.*, "Exploring novel disease-disease associations based on multi-view fusion network," *Comput Struct Biotechnol J*, vol. 21, pp. 1807–1819, Feb. 2023, doi: 10.1016/j.csbj.2023.02.038.

[5] E. Rojano *et al.*, "Evaluating, Filtering and Clustering Genetic Disease Cohorts Based on Human Phenotype Ontology Data with Cohort Analyzer," *J Pers Med*, vol. 11, no. 8, p. 730, Jul. 2021, doi: 10.3390/jpm11080730.

[6] J. Menche *et al.*, "Uncovering disease-disease relationships through the incomplete human interactome," *Science*, vol. 347, no. 6224, p. 1257601, Feb. 2015, doi: 10.1126/science.1257601.

[7] E. Min, X. Guo, Q. Liu, G. Zhang, J. Cui, and J. Long, "A Survey of Clustering With Deep Learning: From the Perspective of Network Architecture," *IEEE Access*, vol. 6, pp. 39501–39514, 2018, doi: 10.1109/ACCESS.2018.2855437.

[8] L. Prieto Santamaría, E. P. García Del Valle, M. Zanin, G. S. Hernández Chan, Y. Pérez Gallardo, and A. Rodríguez-González, "Classifying diseases by using biological features to identify potential nosological models," *Sci Rep*, vol. 11, no. 1, p. 21096, Oct. 2021, doi: 10.1038/s41598-021-00554-6.

[9] J. Sánchez-Valle *et al.*, "Interpreting molecular similarity between patients as a determinant of disease comorbidity relationships," *Nat Commun*, vol. 11, no. 1, p. 2854, Jun. 2020, doi: 10.1038/s41467-020-16540-x.

[10] J. Xie *et al.*, "Prediction of cardiovascular diseases using weight learning based on density information," *Neurocomputing*, vol. 452, pp. 566–575, Sep. 2021, doi: 10.1016/j.neucom.2020.10.114.

[11] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-Means Clustering Algorithm," *Applied Statistics*, vol. 28, no. 1, p. 100, 1979, doi: 10.2307/2346830.
[12] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised Deep Embedding for Clustering Analysis," 2015, doi: 10.48550/ARXIV.1511.06335.
[13] P. Dahal, "Learning Embedding Space for Clustering From Deep Representations," in *2018 IEEE International Conference on Big Data (Big Data)*, Seattle, WA, USA: IEEE, Dec. 2018, pp. 3747–3755. doi: 10.1109/BigData.2018.8622629.
[14] L. M. Schriml *et al.*, "The Human Disease Ontology 2022 update," *Nucleic Acids Research*, vol. 50, no. D1, pp. D1255–D1261, Jan. 2022, doi: 10.1093/nar/gkab1063.
[15] S. A. Aleksander *et al.*, "The Gene Ontology knowledgebase in 2023," *Genetics*, vol. 224, no. 1, p. iyad031, Mar. 2023, doi: 10.1093/genetics/iyad031.
[16] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, in KDD'96. Portland, Oregon: AAAI Press, Aug. 1996, pp. 226–231.
[17] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011, doi: 10.48550/ARXIV.1201.0490.
[18] S. K. Anand and S. Kumar, "Experimental Comparisons of Clustering Approaches for Data Representation," *ACM Comput. Surv.*, vol. 55, no. 3, pp. 1–33, Mar. 2023, doi: 10.1145/3490384.
[19] L. Leydesdorff, J. A. Comins, A. A. Sorensen, L. Bornmann, and I. Hellsten, "Cited references and Medical Subject Headings (MeSH) as two different knowledge representations: clustering and mappings at the paper level," *Scientometrics*, vol. 109, no. 3, pp. 2077–2091, Dec. 2016, doi: 10.1007/s11192-016-2119-7.
[20] J. S. Amberger, C. A. Bocchini, A. F. Scott, and A. Hamosh, "OMIM.org: leveraging knowledge across phenotype–gene relationships," *Nucleic Acids Research*, vol. 47, no. D1, pp. D1038–D1043, Jan. 2019, doi: 10.1093/nar/gky1151.
[21] M. Liu and P. D. Thomas, "GO functional similarity clustering depends on similarity measure, clustering method, and annotation completeness," *BMC Bioinformatics*, vol. 20, no. 1, p. 155, Dec. 2019, doi: 10.1186/s12859-019-2752-2
[22] H. Al-Mubaid and T. Aldwairi, "Utilizing Functional Annotation of Disease Genes for Disease Clustering," in EPiC Series in Computing, vol. 92. pp. 58–71. doi: 10.29007/zxxg.