Iraqi Journal for Computer Science and Mathematics

Journal Homepage: http://journal.esj.edu.iq/index.php/IJCM

e-ISSN: 2788-7421 p-ISSN: 2958-0544

Gene Expression Analysis via Spatial Clustering and Evaluation Indexing

Awrad D. Salman[1,*], Basaad Ali Hussain[1]

[1]University of Baghdad, College of Science, Computer Science Department, Iraq

*Corresponding Author: Awrad D. Salman

DOI: https://doi.org/10.52866/ijcsm.2023.01.01.004

Received June 2022; Accepted December 2022; Available online October 2022

ABSTRACT: The density-based spatial clustering for applications with noise (DBSCAN) is one of the most popular applications of clustering in data mining, and it is used to identify useful patterns and interesting distributions in the underlying data. Aggregation methods for classifying nonlinear aggregated data. In particular, DNA methylations, gene expression. That show the differentially skewed by distance sites and grouped nonlinearly by cancer daisies and the change Situations for gene excretion on it. Under these conditions, DBSCAN is expected to have a desirable clustering feature i that can be used to show the results of the changes. This research reviews the DBSCAN and compares its performance with other algorithms, such as the traditional number of clustering, K-mean particle swarm optimization (PSO), and Grey–Wolf optimization (GWO). This method offers high performance for improvement. The DBSCAN algorithm also offers better results of clusters and gives better performance assessment according to the results shown in this study.

Keywords: Clustering; Particle swarm optimization; Gene expression; Kmean Algorithm; DBSCAN algorithm; Grey-Wolf Optimization Algorithm

1. INTRODUCTION

Gene expression data are typically presented as a matrix, with genes arranged in rows and conditions, experiments, or time points arranged in columns. The content of the matrix is represented by the levels of gene expression in each scenario. These figures may be absolute, relative, or normalized in certain ways. The results of a specific array in a specific condition are contained in each column, referred to as the condition's profile [1]. Each row vector depicts a gene's expression pattern under all conditions. In the next installment, more formal definitions will be provided. Microarray technology has been utilized to explore the gene expression data on a number of occasions [2].

The goal is to evaluate the data and find patterns among the genes, which will lead to a better knowledge of cell processes and, as a result, a big step towards cell behavior modeling.

Gene expression data are commonly displayed using an expression matrix, which can also be annotated. The columns are used to determine the dataset's quality to represent the experimental conditions or time points. The rows of the expression matrix indicate the genes under investigation across all experimental conditions or time points. The expression levels of a gene across multiple experimental settings are referred to as a gene expression profile, whereas the expression levels of all genes in a sample are referred to as a sample expression profile. Researchers can evaluate the expression levels of a large number of genes in a variety of samples and settings by using microarrays [3]. The information gathered from them is referred to as gene expression data. When assessing gene expression data, a variety of objectives, such as combining subsets of genes expressed in subsets of scenarios or categorizing new genes based on the expression of previously classified genes, are considered. As a result, finding clusters of genes that are co-expressed under clusters of

situations simultaneously requires a clustering of the data matrix's rows (genes) and columns (conditions). "Biclustering" is the name for this sort of clustering [4, 5].

• The resultant clusters are known as a "bicluster," which is a group of genes that behave similarly under a subset of the expression data matrix's conditions. Notably, a gene/condition is not the same as a disease. Three basic procedures are utilized to acquire gene expression data from microarray experiments: image processing, transformation, and normalization. The techniques utilized in microarray research generate three types of data, which are curated in various internet databases, including the Gene Expression Omnibus. The scanned image of the Microarray chip.

• Expression data refer to the scanned image's normalized version, and they represent a matrix of integers to indicate the gene expression for a group of samples.

• Annotation data are those applied to the expression data (metadata). They consist of textual descriptors, such as gene functions and sample data (i.e., illness state or normal state) that aid in the interpretation of discovered gene expression levels [6]. In the study of gene expression data, identifying groups of genes with comparable expression patterns is a crucial first step, but it entails clustering algorithmic difficulty. The objective of clustering analysis (or clustering) is to arrange a set of items so that objects in the same group (called a cluster) are more comparable (in some sense) to each other compared with objects in other groupings (clusters). Machine learning, pattern recognition, image analysis, information retrieval, and bioinformatic all use exploratory data mining as a statistical data processing technique. The technique of grouping a set of data objects is known as the clustering of a gene in a certain experimental setting. With the development of microarray technology, there has been a boom in interest in extracting useful information from gene expression data that may be used to discover or validate biological processes [7]. In this case, a number of machine learning algorithms have proven to be beneficial [8].

2. LITERATURE REVIEW

Many studies have been conducted on density-based spatial clustering for noisy gene expression data.

Vijayalakshmi et al. (2020): Their method employed particle swarm optimization, non-dominated sorting, and multiclassifier approaches, such the k-nearest neighbor method, rapid decision tree, and kernel density estimation. To derive the best level of precision in breast cancer prediction, Bayes theorem was utilized to update the results. The proposed particle swarm optimization and non-domination sorting with the classifier approach model assisted in identifying the most essential breast cancer prediction variables. The purpose of the problem model is determined by the selected features. The proposed model was validated using data from the UCI machine learning data repository, including the WBCD and WDBC breast cancer datasets. The following criteria were taken into account: sensitivity, precision, accuracy, and time complexity [29].

Nawel Zemmal et al. (2020): A hybrid framework for integrating the active learning (AL) and particle swarm optimization (PSO) algorithms was suggested to reduce the cost of labeling while producing a more effective classifier. Eighteen benchmark datasets were used to compare the proposed solution to three well-known classifiers with different learning paradigms: support vector machine (SVM), extreme learning machine (ELM) with supervised learning, and transductive support vector machine (TSVM) with semi-supervised learning. All of them are active learning algorithms that use the Nave Base and Margi classifiers. Experiments showed that their proposed strategy could reduce the time and effort required by professionals to annotate medical data for producing an appropriate classifier. Furthermore, in view of speeding up the time-consuming marking process, the active learning method was adopted. PSO, a nature-inspired algorithm, was used to select the most informative medical examples from a huge number of unlabeled cases while enhancing the classifier's accuracy. A novel uncertainty measure was used in their method. [27].

Al-Obeidat et al. (2018): Principal component analysis, correlation, and spectral based feature selection were used in the first step of cancer sample classification. Then, the chromosome was tested using a genetic approach that employed autoencoder-based clustering. The produced function subset was applied to the classification problem. They used three classifiers (assist vector machine, k-nearest neighbors, and random forest) to avoid relying on a single one. Six benchmark gene expression datasets were employed in the performance evaluation, and four cutting-edge related techniques were compared with three sets of experiments to test the hypotheses related to the evaluation of selected features that would use sample-based clustering, modification of ideal parameters, and selection of a higher-performing classifier. The relationship was built using accuracy, recall, false positive rate, precision, F-measure, and entropy [30].

Chun Guan et al. (2019): Density-based spatial clustering of applications with noise (DBSCAN) is a dataset clustering algorithm for locating clusters of any shape. DBSCAN has three drawbacks: (1) the parameters are difficult to establish; (2) users cannot regulate the number of clusters; and (3) it cannot be used directly as a classifier. To address the drawbacks of DBSCAN, the authors presented a novel particle swarm optimized density-based clustering and classification (PODCC) approach. Particle swarm optimization (PSO), a well-known evolutionary and swarm algorithm (ESA), is used to solve a wide range of optimization issues, including data analytics. The proposed fitness function can be used to set the number of

clusters as an input for PODCC. To assess the suggested approach, ten synthetic datasets and ten benchmarking datasets from various available sources were employed **[26]**.

Asgarali Bouyer and Abdolreza Hatamlou (2018): In partition data clustering, the K-means technique separates objects into smaller, disjoint groups with the highest similarity to objects in the same category and the most dissimilar items in other groupings. K-harmonic means (KHM) emerged with the improved cuckoo search (ICS) and particle swarm optimization (PSO) to create a new clustering methodology. By dynamically and shrewdly modifying the radius, ICS could identify the global optimum solution by applying the Levy fly method. The proposed approach was less time-consuming than the traditional cuckoo hunt. With PSO, the ICS was prevented from sliding into the local optima. The suggested algorithm called ICMPKHM could more effectively and consistently solve the KHM local optima problem [23].

Santosh Kumar Majhi and Shubhra Biswal (2018): A hybrid clustering strategy based on K-means and ant–lion optimization was suggested for optimal cluster analysis. Ant–lion optimization (ALO) is a global optimization model that includes a stochastic component. The performance of the suggested approach was compared with those of the K-means, K-means–PSO, K-means–FA, DBSCAN, and revised DBSCAN clustering algorithms by using various performance measures. Eight datasets were used in the experiment, and the results were statistically examined. The combination of K-means and ant–lion optimization techniques outperformed the other three methods in terms of the number of intracluster distances and F-measure [24].

Chun Guan (2018): A type of clustering algorithm known as density-based clustering can locate clusters of any shape. A well-known density-based clustering technique is the density-based spatial clustering of applications with noise (DBSCAN). Genetic algorithms (GAs), particle swarm optimization (PSO), differential evaluation (DE), and artificial bee colony are a few of the different types of ESAs (ABC). To solve the shortcoming of DBSCAN, the hybrid ESA–DCC framework was used to discover the optimal parameters for density-based clustering and classification. The performance of the ESA–DCC method was compared with the usability of the K-means and DBSCAN algorithm **[31]**.

Limin Wang et al. (2018): A parameter called adaptive density development spatial clustering of application with noise was proposed to quickly solve the global optimization problem by using the cuckoo search method. The cuckoo search method can determine the best global parameter (Eps). The enhanced algorithm can remove human involvement in the clustering process and achieve clustering process automation. Their simulation results suggest that the algorithm can select an appropriate Eps parameter value and produce highly accurate clustering results [32].

3. METHODOLOGY

In this section, we will discuss all the aforementioned clustering and clustering algorithms and the research technique adopted in this study.

Clustering is an unsupervised learning task that uses similarity measures to group data items or patterns. In Rd space, objects exist as data points. Entities belonging to the same cluster have a higher degree of similarity than those belonging to different clusters. Cluster analysis is used to summarize or gain a better understanding of data in a certain context, such as the grouping of related documents for browsing, identifying protein structures and genes with similar roles, or compressing data. Several clustering techniques have been developed for pattern analysis, grouping, decision making, document retrieval, image segmentation, and data mining, but several significant challenges for correctly identifying the clusters still exist [7], as shown in Figure 1.

Clustering is a type of machine learning in which groups or classes are formed based on similar and shared characteristics among the objects being grouped. Clustering is the process of determining if two or more groups or clusters exist within a group of potential class members. These clusters are often discovered by trial and error, with items being grouped together based on similar characteristics. The derived results are analyzed to determine how successful they are. Then, the process is repeated until the groups are deemed suitable. As classes are undefined at the outset, clustering is an example of unsupervised learning. Clustering is the process of grouping a set of objects so that two objects in the same category are identical. Despite the well-defined principle of clustering, the properties of a real cluster are difficult to determine. The similarity measure used to group objects, which may differ between applications, plays a large role in determining the cluster. New algorithms, which are often extensions of existing algorithms, are created when new types of clusters are required. However, the selection of cluster features is not driven by any universal method or successful criteria. To ease the task of finding a successful algorithm, the algorithmic characteristics and properties of desirability must be sought. Clustering is the method that differentiates the different algorithms. Algorithms that use identical strategies are clustered together in the same model. The partitioning- and density-based models, in addition to the hierarchical model, are the most popular types. The prerequisites for the algorithms, apart from the artifacts to be clustered, distinguish these models from hierarchical models. One example is the mean and partitioning algorithm that requires the value of k for the number of clusters to be identified as an input parameter. The minimum number of elements per cluster and the density threshold are taken as the inputs to the density-based algorithm (DBSCAN) [8], as shown in Figures 2 and 3.

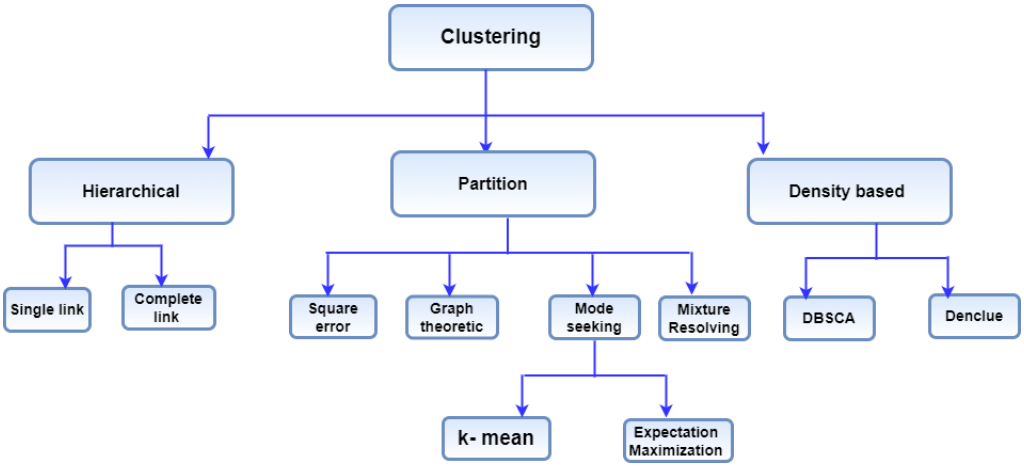

FIGURE 1. Clustering types

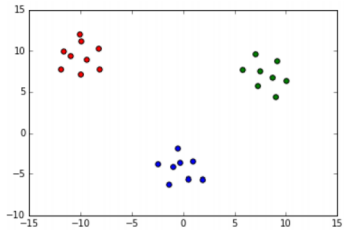

FIGURE 2. Clustering withthree clusters [9]

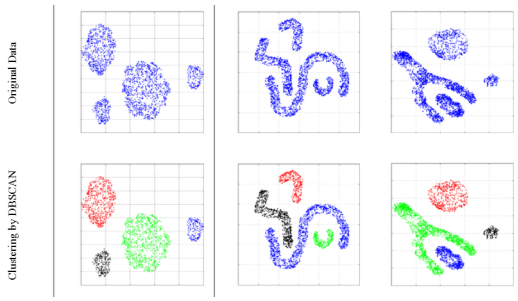

FIGURE 3. 3: DBSCAN's density-based clustering characteristics [10]

The clustering algorithm is a common data-analytic technique. Clustering methods can be applied to a variety of domains, including pattern recognition, machine learning, image processing, and information retrieval. It can also be used for data analysis. Each clustering algorithm has its own set of advantages and disadvantages. The density-based approach is simple and efficient, similar to other clustering techniques. By using the density-based clustering methods, which are meant to locate clusters of arbitrary form in databases containing noise, a cluster can be defined as a high-density zone partitioned by low-density regions in data space.

Density-based spatial clustering of applications with noise (DBSCAN) is an example of a density-based clustering algorithm. DBSCAN can detect clusters of any shape. Nevertheless, it is susceptible to input factors, particularly when data density is non-uniform. DBSCAN clustering algorithms are typically classified as follows: (a) DBSCAN clustering based on partitioning; (b) DBSCAN clustering based on grids; (c) hierarchical DBSCAN clustering; (d) DBSCAN clustering based on detection; (e) incremental DBSCAN clustering; and (f) spatial–temporal DBSCAN clustering [11].

4. ARCHITECTURE OF THE PROPOSED SYSTEM

4.1 DBSCAN CLUSTER

Clustering is an unsupervised learning technique that splits data points into several distinct bunches or groups, with data points in the same group sharing similar characteristics and data points in different groups differing in a certain way.

In general, all clustering methods begin by calculating data point similarities and then using that knowledge to generate clusters or batches of data. The focus of our analysis is the "density-based spatial clustering of applications with noise (DBSCAN)." DBSCAN entails a function for calculating the distance between two values, and it suggests how near they should be.

4.1.1. Density-Based Clustering Algorithm

Density-based clustering is an unsupervised learning strategy that uses the concept of a cluster in data space as a contiguous region of high-point density separated from other comparable clusters by contiguous regions of low-point density to identify various groups/clusters in the data.

DBSCAN is a density-based clustering method based on

The two parameters in the DBSCAN algorithm are as follows:

- "MinPts": The minimum number of points that must be grouped together for a region, which is also considered dense.
- "eps (ε)" distance: The algorithm utilizes this measure by looking for a nearby point.

4.1.2. Parameter Estimation

Each parameter has a different effect on the algorithm. As a rule of thumb, DBSCAN requires the aforementioned parameters, especially "MinPts", by using the number of dimensions D in the dataset. "MinPts" can be calculated as "MinPts" D + 1. If "MinPts" = 1, then each single point will form a cluster, which is an inefficient approach. If "MinPts" is set to 2, then the results will be identical to those obtained by using the single-link metric and cutting "a dendrogram" at height 1. Thus, "MinPts" must be set to no less than 3. Nevertheless, greater values are usually better for noisy datasets, and they can produce more significant clusters. As a general rule, "MinPts" = 2dim can be utilized, although greater values may be required for extensive data, noisy data, or data containing numerous duplicates.

The value of is selected by graphing the distance to the k = minPts-1 nearest neighbor and sorting them by size, starting with the greatest value and working your way down. If ε is too small, then most of the data will not be clustered. However, if ε is too high, then the clusters will merge, and most items will be in the same cluster. A modest value is therefore ideal. As a general rule, only a small number of points should be within distance of one another.

Distance function: The selection of the distance function is closely related to the selection of and has a significant impact on the results. Prior to parameter selection, an appropriate measure of similarity for the data collection is usually initially required. This parameter cannot be estimated; however, the distance functions should be chosen carefully based on the given data.

Our solution is composed of the following four major approaches:

1. **Preprocessing phase:** This phase is regarded as an integral step, as the data quality and the directly derived useful information would affect our model's capacity for learning. Therefore, our data must be preprocessed before incorporating them into our model.

2. **Feature selection phase:** Machine learning methodology is developed to find the clustering of the process for grouping similar objects into different groups, or more precisely, the partitioning of a dataset into subsets, so that the data in each subset would accord with some defined distance measure.

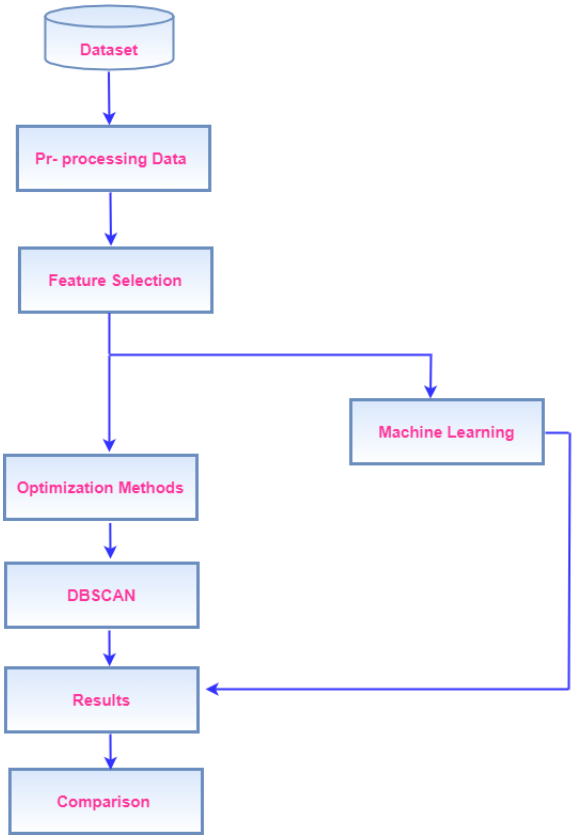

FIGURE 4. **Main block diagram of**

3. **Machine learning phase:** Machine learning is an artificial intelligence (AI) platform that gives systems the ability to learn and spontaneously generate a benign experience that had been explicitly programmed. However, machine learning is not a simple process. As the algorithm ingests training, more precise models can be produced based on the acquired data. A machine learning model is a production derived from using data to train the machine learning algorithm. After the training, users can obtain an output when provided a model with an input. Our work uses the k-means algorithm to cluster our dataset before its application to DBSCAN.

4**. Optimization phase:** Optimization algorithms are responsible for optimizing the cluster. Our work uses the Grey–Wolf optimization method and particle swarm optimization.

5. **DBSCAN clustering**: A method is developed to separate high-density groups from low-density groups. It can find an approach for clustering based on similarity density.

6. **Comprising phase:** A statistical measure is proposed to calculate the efficiency of the model.

5. DATASETS

At this stage, the input handled by the user is obtained, and the data in this step are standardized. First, collect from the public website the mammography dataset and yeast dataset, which data can be stored as CSV files and parsed by the method named parser data.

a- **Data cleaning:** A null cell is examined, and an intermediate value of the column is accounted for. Then, the missing values are imputed once the datasets are cleaned.

b- **Removal of N/ A:** The step handles the missing elements in the dataset. A null cell is checked, and the intermediate value of the column is compensated or filled in, or the missing values are imputed. The remainder of the data is used to estimate the missing values. Replacing the average by considering the missing predictor is a straightforward procedure, as shown in Algorithm 1.

c- **Scaling the data**: In this step, the mean from each column is subtracted to find the value of the standard deviation of the column. In this manner, the mathematical operations of clustering can be controlled. Assuming that the data are

Algorithm (1): Cleaning and treatment miss values
Input : Load Dataset
Output: Complete dataset
Step 1 : Read dataset's
Step 2: For I= 2 to N do
Step 3: check each column for features and compute the missing data Step 4: If missing data <25%
Step 5: Delete Column
Step 6: Compute the mean of features'
• Step 7: substitute feature with mean
• Step 8: End

distributed normally, then mean = 0 and variance = 1. This operation is used in deep clustering because it depends on probability. Furthermore, the scale process is performed on a standard normal distribution scale or a Gaussian scale.

5.1 PREPROCESSING THE DATASET

In this phase, the dataset is considered to be complete after the missing value is removed, as presented in Figure 5.

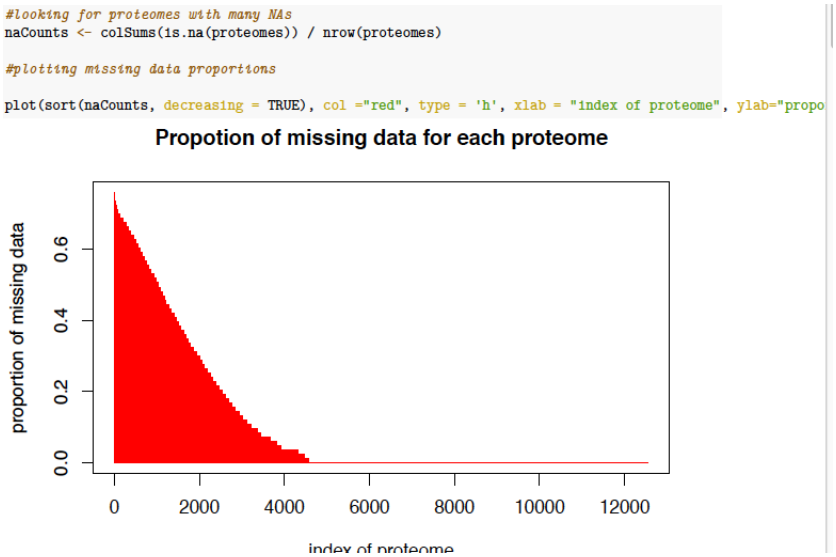

FIGURE 5. **missing data**

6. MAMMOGRAPHY DATASETS

The second dataset is the mammography dataset, which can be obtained from "Mammography - Breast Cancer | Kaggle) (46)."

The dataset included 24 mammograms with a known scanned cancer diagnosis. Then, the images were pre-processed by image segmentation computer vision algorithms to extract candidate objects from the mammogram images. Once segmented, the objects could be manually labeled by an experienced radiologist.

A total of 29 features were extracted from the segmented objects and considered to be most the relevant features for pattern recognition. The number was reduced to 18 and finally to 6 as follows (quoted directly from the paper):

- Area of object (in pixels).
- Average gray level of the object.
- Gradient strength of the object's perimeter pixels.
- Root mean square noise fluctuation in the object.
- Contrast, average gray level of the object minus the average of a two-pixel wide border surrounding the object.
- A low order moment based on shape descriptor. There are two classes, and the goal is to distinguish between microcalcifications and non-microcalcifications using the features for a given segmented object.
- Non-microcalcifications: negative case, or majority class.

RefSeq_ac...	gene_sym...	gene_name	AO-A12D.0...	C8-A131.0...	AO-A12B.0...	BH-A18Q.0...	C8-A130.0...	C8-A138.0...	E2-A154.0...
NP_958782	PLEC	plectin isoform 1	1.096131182	2.609942977	-0.659828047	0.195340653	-0.494059581	2.765080659	0.862659263
NP_958785	NA	plectin isoform 1g	1.111370375	2.650421787	-0.648742151	0.215412944	-0.503899187	2.779709215	0.870186035
NP_958786	PLEC	plectin isoform 1a	1.111370375	2.650421787	-0.654285099	0.215412944	-0.500619319	2.779709215	0.870186035
NP_000436	NA	plectin isoform 1c	1.107560577	2.646373906	-0.632113308	0.205376799	-0.510458925	2.797994911	0.866422649
NP_958781	NA	plectin isoform 1e	1.115180173	2.646373906	-0.64042773	0.215412944	-0.503899187	2.787023494	0.870186035
NP_958780	PLEC	plectin isoform 1f	1.107560577	2.646373906	-0.654285099	0.215412944	-0.503899187	2.779709215	0.870186035
NP_958783	PLEC	plectin isoform 1d	1.111370375	2.650421787	-0.648742151	0.215412944	-0.500619319	2.783366355	0.870186035
NP_958784	NA	plectin isoform 1b	1.111370375	2.650421787	-0.648742151	0.215412944	-0.500619319	2.783366355	0.870186035
NP_112598	NA	epiplakin	-1.517390359	3.909312755	-0.618255939	-1.035759872	-1.845365554	2.205538366	1.920170648
NP_001611	AHNAK	neuroblast differentiation -associated protein AHNAK isoform 1	0.482753678	-1.045293501	1.222002715	-0.517225683	-0.405503121	0.749996978	2.34919662
NP_076965	AHNAK	neuroblast differentiation -associated	0.261785384	-0.03737115	1.019685122	-0.724639359	-0.703971188	-0.156973535	1.581465934

FIGURE 6. Sample dataset (mammography)

- Microcalcifications: positive case, or minority class. Figure 6 shows the samples of mammography dataset.

7. IMPLEMENTATION AND RESULTS

The results of the implementation can be viewed as different plots representing the outputs.

7.1 FEATURE SELECTION

This dataset includes a list of proteins associated with PAM50 genes. Therefore, the dataset can represent the best variables for use as clustered breast cancer subtypes. However, we want to check whether DBSCAN can identify a set of proteins that has good or even better predictive power for cancer subtype classification.

The least absolute shrinkage and selection operator (LASSO) was used for feature selection. Lasso regression shrinks the dataset and generates a sparse set of predictor variables, but a stochastic element needs to be reduced, and the results are not always stable. This issue was addressed by repeating the lasso regression 10 times and ranking the variables in the order of number of times they had been selected in the final model. Figure 7 illustrates the obtained LASO results.

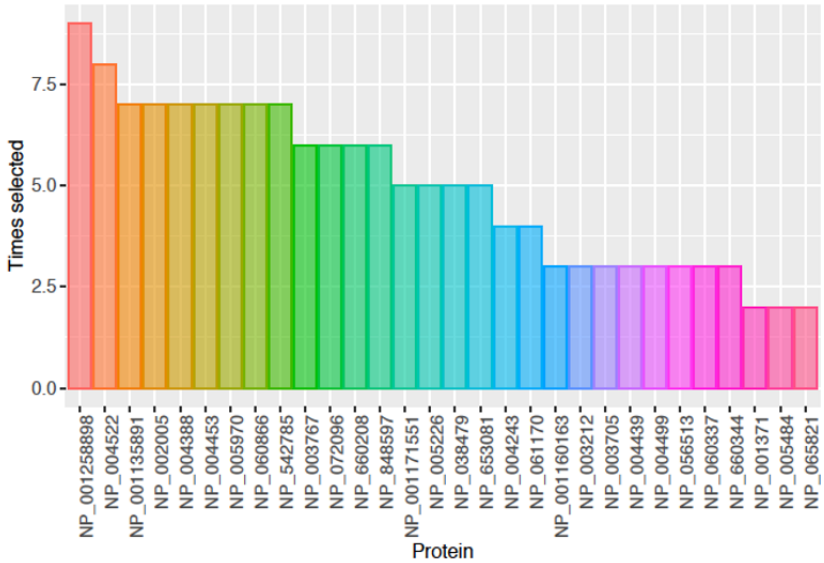

FIGURE 7. LASO results

The set of variables were repeatedly selected via lasso regression. The findings indicate a very wide dataset. Thus, even with 10 iterations, instability arises, and only approximately 30 variables could be selected more than 20% of the time.

7.2 OPTIMIZATION METHODS AND DBSCAN

The mammography dataset was used to fit many optimization methods. Implementing this phase was challenging. Thus, we first operated the GWO and PSO algorithms in the same manner as what we did in Experiment 1.

Figures 8 and 9 show the outcome of the approaches. The Grey–Wolf optimizer obtained one cluster, whereas the PSO had no clusters.

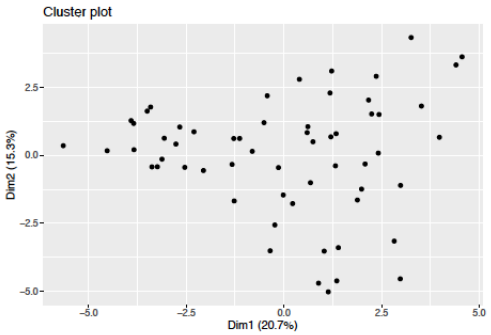

FIGURE 8. Grey–Wolf with DBSCAN

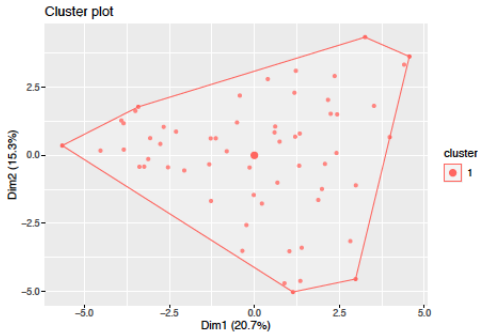

FIGURE 9. Grey–Wolf with DBSCAN

Our results indicate a low-level number of clusters. Consequently, we applied gradient analysis based on the optimum epsilon method.

The optimum epsilon parameter, or the best distance that can fit the DBSCAN, on the mammography dataset was measured via iterative clustering. For epsilon from 1 to 10, the value that gives the maximum clusters is 5, as shown in Figure 10.

Once DBSCAN was applied, the numbers of clusters reached 5 (Figure 11).

8. RESULT VALUE OF THE PROPOSED SYSTEM FOR DATASETS 1 AND 2

The comparative results of the two proposed systems are shown in Table 1.

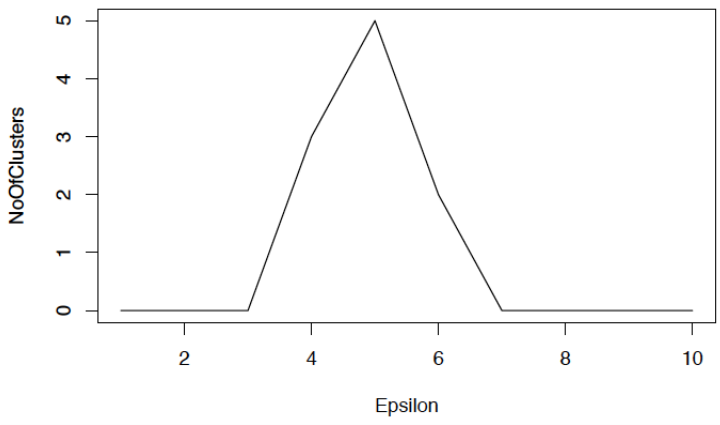

FIGURE 10. Epsilon from 1 to 10(Optimal Value = 5)

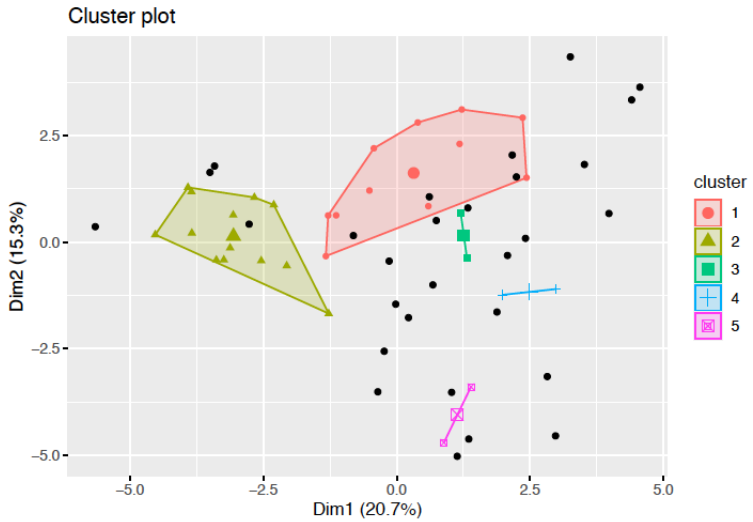

FIGURE 11. Number of Cluster with DBSCAN

Table 1. ClusteringResults of the Two Datasets with Different Proposed Systems

ID	Dataset	Approaches	Structures of Components	Number of Clusters
1	Yeast	Proposed system 1	K- means DBSCAN PSO+ DBSCAN GWO+ DBSCAN	K-means = 3 DBSCAN =3 PSO = 14 GWO = 8
2	mammog-raphy	Proposed system 2	gradient analysis based on the optimum epsilon method + DBSCAN	gradient analysis = 5 kmean=10
3	mammog-raphy	Challenges in dataset2	PSO + DBSCAN GWO+ DBSCAN	PSO = 0 GWO = 1

PSO optimization with DBSCAN can find better clustering results than K-means and GWO.

• PSO is recommended for balancing the clustering and classification problems. For the simulation of certain balanced datasets, gradient analysis offers better performance in clustering, as shown in Experiment 2.

9. CONCLUSIONS

The actual outcomes acquired during system implementation and analysis indicate that numerous conclusions may be drawn from the study. The results can be summarized as follows:

1. This study developed a model to integrate the optimization methods and DBSCAN. Two approaches with different datasets were proposed. In case we have found the best impact level in PSO Yeast dataset and optimum epsilon in Mammography dataset.

2. DBSCAN is not applicable in the unsupervised learning scenario, as the training data could not be used to drive clustering.

3. In proposed system 1, particle swarm optimized density-based clustering was able to address the drawbacks of the classical density-based spatial clustering of applications with noise (DBSCAN).

4. Proposed system 1 was evaluated on the Yeast dataset. Comparisons were undertaken to analyze the operation of the proposed optimization method and analyze the operation number of the clusters in contrast to those of DBSCAN+K-means, PSO, and GWO. The experimental results indicate that PSO performs better than K-means and GWO.

5. In proposed system 2, gradient analysis performs better than PSO and GWO when the number of clusters is high.

ACKNOWLEDGEMENT

The first author would like to thank the reviewers for providing useful suggestions, allowing for the improved presentation of this paper.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] A. Bouyer and A. Hatamlou, "An efficient hybrid clustering method based on improved cuckoo optimization and modified particle swarm optimization algorithms," *Applied Soft Computing*, vol. 67, pp. 172–182, 2018.

[2] C. Guan, K. K. F. Yuen, and F. Coenen, "Particle swarm Optimized Density-based Clustering and Classification: Supervised and unsupervised learning approaches," *Swarm and evolutionary computation*, vol. 44, pp. 876–896, 2019.

[3] X. Hu, L. Liu, N. Qiu, D. Yang, and M. Li, "A MapReduce-based improvement algorithm for DBSCAN," *Journal of Algorithms & Computational Technology*, vol. 12, pp. 53–61, 2018.

[4] S. K. Majhi and S. Biswal, "Optimal cluster analysis using hybrid K-Means and Ant Lion Optimizer," *Karbala International Journal of Modern Science*, vol. 4, pp. 347–360, 2018.

[5] S. Mohan, S. Bhattacharya, R. Kaluri, G. Feng, and U. Tariq, "Multi-modal prediction of breast cancer using particle swarm optimization with non-dominating sorting," *International Journal of Distributed Sensor Networks*, vol. 16, 2020.

[6] N. Zemmal, N. Azizi, M. Sellami, S. Cheriguene, A. Ziani, and M. Aldwairi, "Particle swarm optimization based swarm intelligence for active learning improvement: Application on medical data classification," *Cognitive Computation*, vol. 12, pp. 991–1010, 2020.

[7] P. Deepthi and S. M. Thampi, "PSO based feature selection for clustering gene expression data," *2015 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*, pp. 1–5, 2015.

[8] Y. Kravitz, 2017.

[9] O. Carlsson, 2014.

[10] K. Mumtaz and K. Duraiswamy, "An analysis on density based clustering of multi dimensional spatial data," *Indian Journal of Computer Science and Engineering*, vol. 1, pp. 8–12, 2010.

[11] X. Li, P. Zhang, and G. Zhu, "DBSCAN Clustering Algorithms for Non-Uniform Density Data and Its Application in Urban Rail Passenger Aggregation Distribution," *Energies*, vol. 12, pp. 3722–3722, 2019.

[12] D. Wang, D. Tan, and L. Liu, "Particle swarm optimization algorithm: an overview," *Soft Computing*, vol. 22, pp. 387–408, 2018.

[13] S. Sharma, A. K. Sharma, and D. Soni, "Enhancing DBSCAN algorithm for data mining," *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing*, pp. 1634–1638, 2017.

[14] A. Karami and R. Johansson, "Choosing DBSCAN parameters automatically using differential evolution," *International Journal of Computer Applications*, vol. 91, no. 7, pp. 1–11, 2014.