IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 August 2022, accepted 18 August 2022, date of publication 26 August 2022, date of current version 22 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3202187

RESEARCH ARTICLE

Consensus Nature Inspired Clustering of Single-Cell RNA-Sequencing Data

AMANY H. ABOU EL-NAGA[1], SABAH SAYED[2], AKRAM SALAH[2], AND HEBA MOHSEN[1]
[1]Department of Computer Science, Faculty of Computers and Information Technology, Future University in Egypt, Cairo 11835, Egypt
[2]Department of Computer Science, Faculty of Computers and Artificial Intelligence, Cairo University, Giza 12613, Egypt

Corresponding author: Amany H. Abou El-Naga (amany.abouelnaga@fue.edu.eg)

ABSTRACT Single-cell RNA sequencing (scRNA-seq) enables quantification of mRNA expression at the level of individual cells. scRNA-seq uncovers the disparity of cellular heterogeneity giving insights about the expression profiles of distinct cells revealing cellular differentiation. The rapid advancements in scRNA-seq technologies enable researchers to exploit questions regarding cancer heterogeneity and tumor microenvironment. The process of analyzing mainly clustering scRNA-seq data is computationally challenging due to its noisy high dimensionality nature. In this paper, a computational clustering approach is proposed to cluster scRNA-seq data based on consensus clustering using swarm intelligent optimization algorithms to accurately recognize cell subtypes. The proposed approach uses variational auto-encoders to handle the curse of dimensionality, as it operates to create a latent biologically relevant feature space representing the original data. The new latent space is then clustered using Particle Swarm Optimization Algorithm, Multi-Verse Optimization Algorithm and Grey Wolf Optimization Algorithm. A consensus solution is found using solutions returned by the swarm intelligent algorithms. The proposed approach automatically derives the number of clusters without any prior knowledge. To evaluate the performance of the proposed approach a total of four datasets have been used then a comparison against the existing methods in literature has been performed. Experimental results show that the proposed approach performs better than widely most used tools, achieving an adjusted rand index of .95, .75, .88,.9 for Biase, Goolam, Melanoma cancer and Lung cancer datasets respectively.

INDEX TERMS Single-cell RNA-seq, automatic clustering, unsupervised learning, swarm intelligence, metaheuristic algorithms, consensus clustering.

I. INTRODUCTION

Single cell RNA Sequencing enables quantification of gene expression at the cell level unlike Bulk RNA sequencing that averages the gene expression across a population of cells [1]. In other words, single cell RNA sequencing attempts to represent the distribution of expression level within each subpopulation per transcript for each individual cell in the sample, while bulk RNA sequencing only measures the average expression level per gene for a large population of cells [2].

Accordingly, bulk RNA sequencing utilizes studying comparative transcriptomics, disease studies where it could quantify expression signatures. Nevertheless, it is inadequate for studying complex heterogeneous systems [3]. scRNA-seq analysis gives insight about cellular heterogeneity by identifying cell types and detecting of rare cell types by studying cells behavior in its microenvironment which is applied in fields like cancer treatment, tumors composition, embryonic development etc. [4].

Single cell RNA sequencing was first introduced by Tang *et al.* [2], though it gained its wide popularity around 2014 as the sequencing cost due to new protocols made the sequencing process easier. Until now there is an urge to do modifications on existing computational tools and implementing new ones to perform single cell RNA sequencing analysis in order to study the biological inquiries about a particular cell type behavior etc. [5].

The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Tang.

Advances in sequencing technologies allows extracting genetic information from hundreds to thousands of cells, resulting in what is known by Curse of Dimensionality [6]. Since each cell (Sample) is expressed across hundred thousand of genes (Features), datasets resulting from scRNA-seq has a large number of genes (features) and each gene has an expression value in each cell sampled in the scRNA-seq experiment done. Single cell datasets size ranges from 10^2 to 10^6 cells with increase in size every year. Most single cell data consists of zeros since RNA molecules amount in sequenced cells is very low which causes dropout events [7]. However, not all these zeros are insignificant to be neglected in downstream analysis hence the higher the dropout level the more difficult it is to process the data. Consequently, more competent computational tools are required to efficiently perform a sequence of tasks such as processing the data, feature selection or dimensionality reduction and most importantly clustering the data [8].

In scRNA-seq analysis, clustering is considered the main step of the downstream analysis; it uses the transcriptomic relations between cells to group the cells into clusters where each cluster represents a cell type, or a cell lineage based on the type of analysis performed [9], [10]. Because of problems mentioned before such as dropout events and curse of dimensionality, the high dimensional space makes it difficult to use distance measures like Euclidean distance, Manhattan distance and such methods without proper data preprocessing techniques [11]. Not tackling such problems in scRNA-seq analysis may cause the problem of overfitting; as several features may not be important in the analysis. However, the model is trained in its learning process taking into consideration that these features matter, hence a major step before performing clustering is to get rid of un-informative genes (features) to get more robust well-defined clusters.

In this paper a new approach to find better clustering performance for scRNA-seq data based on consensus clustering using Swarm Intelligence techniques (Particle Swarm Optimization, Grey Wolf Optimization and Multi-Verse Optimization) is proposed. Each technique iterates to find an optimal solution in its search space then the solutions introduced by these techniques are fused into a consensus solution resulting in better clustering performance.

The major contributions of this research are:

- Proposing a new unsupervised clustering approach that automatically detects the number of clusters k, overcoming shortcomings of other methods that needs k to be predefined.
- Utilizing variational auto-encoders (VAE) as part of the proposed model, to create a biologically relevant latent feature space representing the original feature space. Yet, the latent features are less in number making it easier to deal with. Hence, overcoming the curse of dimensionality of the scRNA seq data.
- Accurately clustering the scRNA seq data in consensus clustering approach using metaheuristic techniques Particle Swarm optimization algorithm, Grey Wolf optimization algorithm and Multi-Verse optimization algorithm.
- Achieving an adjusted rand index (ARI) of .95 for Biase dataset, an ARI of .75 for Goolam dataset, an ARI of .88 for Melanoma cancer dataset and an ARI of .9 for Lung cancer dataset, with stable clusters regardless the number of samples in datasets.

II. RELATED WORK

Considering the many challenges that faces clustering single cell data in terms of execution time or clustering accuracy and stability, many methods have been developed to overcome these limitations compared to traditional methods. In this section, a brief background about clustering, its categories and the mechanism of the used metaheuristic techniques is presented, then a review of the recently developed tools based on the challenges is listed.

A. CLUSTERING

Clustering is a process to find natural grouping of data objects such that each cluster has similar data objects but simultaneously differs from data objects in other clusters depending on the similarity measure used [12]. That is why clustering is considered an unsupervised learning problem since prior information about the data groups is unavailable unlike classification problems where labels/groups of data are already known [13]. Clustering algorithms tries to minimize the intra-cluster distance between data objects of the same cluster and maximize inter-cluster distance between data objects of different clusters to find accurate grouping of these objects.

There are numerous algorithms and techniques to solve the clustering problem, and all can be categorized as follows:

- Hierarchical clustering [14]: it attempts to build a hierarchical tree by finding structure among data points using either agglomerative or divisive approach.
- Partition based clustering [15]: it attempts to partition the data into k-clusters by identifying the best k-centers; those centers are either centroids like in K-means algorithm or medoids.
- Graph based clustering [16]: it represents the data points as nodes in a graph then uses pairwise similarities to compose the edges between the nodes based on the similarity measured between those nodes.
- Ensemble clustering [17]: it utilizes finding clusters using different clustering algorithms then fuses all suggested solutions by these algorithms using consensus function to find better clusters representation for the data. Ensemble clustering shows better clustering performance than traditional clustering techniques.
- Search based clustering [18]: it tries to solve the clustering problem automatically by inspiring a natural or physical phenomena. The advantage of these methods that it solves the problem of the traditional clustering

methods that usually converges towards the nearest local optima. Search based algorithms include evolutionary algorithms and swarm intelligence algorithms.

The clustering problem also can be introduced as an optimization problem, where the data required to be clustered form the search space while exploring the optimal grouping of data points existing in the search space is the solution to be optimized [19]. Nature inspired Metaheuristic techniques are part of computational intelligence used in solving optimization problems. Their main concept is exploring a search space attempting to find the optimal solution for the problem in the space. These techniques are inspired from natural theories or natural behavior of intelligent organisms in nature [20]. Metaheuristics techniques can be categorized as either single solution-based search (e.g., Simulated annealing, Tabu search) or population-based search and the later consists of two subcategories Evolutionary algorithms (e.g., Genetic Algorithms, Differential Evolution) and Swarm Intelligence algorithms (e.g., Particle Swarm Optimization, Ant colony Optimization, Firefly Optimization etc.).

Evolutionary algorithms mimic biological evolution in nature, they start by initializing a population of random solutions then these solutions are evaluated using a fitness function and the best solution is chosen through several iterations of searching the space of solutions and evaluating them till stability is reached.

On the other hand, Swarm Intelligence algorithms are inspired by the social and collective intelligent behavior of organisms in nature for example Cuckoo search algorithm take after some cuckoo species and their strategy in egg laying, Ant colony Optimization take after ant and their strategy in finding the nearest path from a nest to a food source, etc. [21].

B. METAHEURISTICS TECHNIQUES

1) PARTICLE SWARM OPTIMIZATION ALGORITHM (PSO) [22]

a meta-heuristic, population-based algorithm that is a part of the swarm intelligence family of algorithms. PSO mimics a social behavior of organisms' swarm, such as a flock of birds or a species of fish. PSO randomly initializes a population of particles, where each particle represents a candidate solution to the optimization problem. Each particle has a velocity that helps in the movement of the particle that differs from other particles velocities, and a position in the swarm. The search space consists of the swarm of particles, and particles move randomly in that search space according to an update in its velocities.

2) GREY WOLF OPTIMIZATION ALGORITHM (GWO)

Another meta-heuristic algorithm that inspires its behavior from grey wolves is Grey Wolf Optimization algorithm, it is first introduced by Mirjalili *et al.* [23]. GWO follows the leadership hierarchy of the grey wolves mimicking their hunting mechanism in nature. The classification of the grey wolves are as follows α, β, δ, ω each representing a position within the hierarchy of the grey wolves. The α wolves take the responsibility of deciding during the hunting process, leading and tracking other wolves to keep up social equality within their groups. The β wolves comes after α wolves in hierarchy and are considered the advisors of the α wolves. Once α wolves die or become too old; β wolves ascend and become α wolves. The ω wolves may be the children of the group and are controlled by δ wolves. The δ wolves are also responsible for providing information to the α and β wolves. The hunting process of grey wolves starts with searching for and tracking the prey. After that, grey wolves start to surround their prey until it can no longer move. Lastly, grey wolves start attacking the pray.

3) MULTI-VERSE OPTIMIZATION ALGORITHM (MVO)

MVO is first introduced by Mirjalili *et al.* [24], it is a swarm intelligent optimization algorithm that is inspired from the theory of multi-verse in astrophysics. The multi-verse adopts the concept of the multiple universes created by the big bang. It also adapts the concept of interaction between these universes which takes place through divergent variety of holes. These holes are either black, white or worm holes. A transfer between any two pair of universes happens when a black and white hole interact through what is called a tunnel. Worm holes create those tunnels whereas black holes absorb everything, and white holes emit everything. Each universe represents a solution to the optimization problem and each feature represents an object within a universe. The fitness value per universe is calculated by some objective function and is known by the inflation rate, which controls the expansion through space. The fitness of the solution depends on the existence of white holes which leads to better fit solution. However, the existence of black holes leads to poor solutions.

Many tackled the problem of clustering scRNA-seq data using different clustering approaches; however, some proposed methods faced limitations.

Xu and Su [25], introduced SNN-Cliq in this approach the SNN graph is constructed by computing a similarity matrix that depends on Euclidean distance as a similarity measure between data points. For each data point, a list of the k-nearest neighbors is made based on the similarity matrix. The graph is then constructed by considering each cell to be a data point and a weighted edge between any two points is created if the two points have at least one common nearest neighbor. The maximal quasi-clique per node is found using a greedy algorithm fed by SNN graph as an input. All possible quasi-cliques are found then an elimination process is performed to remove any sub partial existence. Clusters are then identified by merging quasi-cliques based on overlapping rate. Iterative merging is later performed until there is no pair of clusters that have a greater overlapping rate than a certain threshold. Although this approach succeeds in having a clear definition of clusters, it all depends on how the single cell data is represented as a graph when a graph is too sparse it fails to detect the clusters.

Guo *et al.* [26], introduced SINCERA for analyzing single cell RNA-Sequencing data. In this pipeline, a gene filtering step is utilized first, as the genes expressed in less than a certain number of cells are filtered out. Normalization methods on both gene and cell level are later performed. For clustering, two-dimensional unsupervised hierarchical clustering is used, since it requires no prior information about the number of clusters however it faces a high time complexity limitation. It is worth mentioning that centered Pearson's correlation is used as the default similarity measure throughout the pipeline.

Satija *et al.* [27], introduced Seurat a tool that incorporates unsupervised clustering algorithm. Seurat combines dimensionality reduction methods with graph based partitioning techniques. Seurat identifies the set of most variable genes across the dataset to enhance the dimensionality reduction methods performance. The gene set is then used as input to principal component analysis (PCA) followed by graph-based clustering. Seurat has a low time complexity however the iterative process may hide small communities.

Yau *et al.* [28], introduced Pca-Reduce an approach that projects a gene expression matrix with dimensions ($n \times d$) representing number of cells and number of expressed genes across the cell respectively, into a score matrix. At this stage, K-means clustering is performed as an initial step on the score projected matrix with k as a large value ensuring all cell types are seized. Subsets of resulting initial clusters are then taken, and a probability of merging for all possible pairs is calculated using multivariate Gaussian with mean and covariance matrix. Two clusters are merged if the pair with the highest probability belongs to them, or by sampling a pair of clusters based on their merged probabilities. The process of projecting and merging clusters are repeated until one single cluster remains. Though PcaReduce has a low time complexity, it is sensitive to outliers.

Kiselev *et al.* [29], proposed SC3 a single cell data analyzing tool. SC3 takes gene expression matrix with rows representing the genes and columns representing the cells as input. Firstly, gene filtering process takes place in which genes expressed in less than $X\%$ of cells are filtered out as well as genes expressed in (100-X) % of cells. The filtering is done to remove non informative genes reducing the dimensionality of the data. A distance matrix is constructed for each similarity measure using Euclidean distance, Pearson and spearman. To transform all distance matrices, principal component analysis or calculating eigenvectors of associated Laplacian graph is used. K-means clustering algorithm is later performed on transformed distance matrices according to Hartigan and Wong algorithm [30]. Then clustering is performed to find a consensus matrix according to cluster-based similarity partitioning algorithm (CSPA) [31]. A binary similarity matrix is later generated per clustering result with cells being both dimensions of the new matrix. Two cells have a similarity of 1 if they belong to the same cluster and 0 otherwise. All similarity matrices generated based on the CSPA clustering results are averaged forming a new consensus matrix. Hierarchal clustering is then performed on the consensus matrix using an agglomerative approach. Though SC3 is scalable to large datasets, it is also sensitive to outliers. SC3 requires the user to determine k.

Lin *et al.* [32] introduced CIDR, an approach that performs dimensionality reduction using principal co-ordinate analysis (PCoA) on a dissimilarity matrix. The number of co-ordinates is determined based on variation of the scree algorithm. Hierarchical clustering is applied after determining the number of clusters according to the Calinski–Harabasz index [33]. CIDR provides hierarchical relationship among datapoints, but it faces a high time complexity limitation and the clusters given are not explicit.

Wang *et al.* [34], introduced a framework named SIMLR. Given an expression matrix, SIMLR calculates a symmetric matrix that seizes pairwise similarities between cells. Gaussian kernels with multiple different hyper-parameters are used by utilizing Euclidean distance between pairs of cells and K-nearest neighbors. An optimization algorithm is used to improve some of the hyper parameters. SIMLR then uses stochastic neighbor embedding (t-sne) method for dimensionality reduction. K-means is then performed on the resulting latent data. SIMLR makes no assumption about data distribution, but it faces computational challenges when it comes to large datasets.

Gan *et al.* [35] introduced conCluster, a model that performs consensus clustering. The model performs filtering out on the gene level to eliminate genes that are either expressed in $r\%$ of cells or (100-r) % of cells. Such genes hold no valuable genetic information defined as rare and ubiquitous genes respectively. A gene set of the most variable genes across the cells is identified. For dimensionality reduction, stochastic neighbor embedding (t-sne) is used with a perplexity set to 30 to reduce the number of dimensions to two. K-means is later performed T times utilizing multiple initial parameters for basic clustering. A binary matrix is generated per each clustering output using the resulting cluster labels. The generated binary matrices from the multiple clustering trials are concatenated into one large binary matrix. K-means is performed once again on the concatenated binary matrix using Calinski-Harabaz Index [33] to determine the number of clusters. The results of clustering trials are fused into a consensus one. Concluster provides robust clustering however it relies on combining other algorithms for ensemble.

Yang *et al.* [36] embeds four clustering methods SC3, Seurat, CIDR and t-sne+K-means. Gene expression matrix is taken as input after adjusting it to be suitable for all methods. Clustering using the four methods is performed individually. The results obtained are later used to construct an overall hyper graph which combines all hypergraphs resulting from individual results. For ensemble clustering one of partitioning algorithms, HGPA, MCLA and CSPA is used. Performance is evaluated using average normalized mutual information (ANMI), for ensemble solution and individual ones and the clustering result with highest ANMI is selected as the final result.

Nguyen *et al.* introduced MKGA [37], a clustering technique based on Genetic Algorithm. The chromosome was separated into two segments, the first portion is made up of a series of binary numbers ranging from 0 to 1, indicating whether the cluster is active or not. The clusters' cores are represented in the second part. Three separate objective functions are used to assess the fitness of a chromosome: Sum of squares with cluster, Davies-Bouldin index, and Silhouette index. For crossover, randomly selected parents with the same number of clusters are chosen. For mutation, Gaussian noise is introduced to the clusters' centers. The number of clusters can also be altered by enabling or removing a center. Each generation is subjected to the K-means operator, which is applied based on a user-defined probability. 16 illness data sets and 5 single-cell data sets were used to test the suggested technique. However, the technique's accuracy could be inefficient in case of large datasets.

Geddes *et al.* [38], propose an auto-encoder ensemble clustering framework. The framework starts with randomly projecting of single cell data into sub-spaces. Auto-encoders trains on the sub-spaces to compress the data into a lower dimensional space. Several experiments were made on different datasets to test the effects and to optimize the hyper parameters: the projection size, auto-encoder learning rate and feature space size. To prove the enhancement in clustering performance using the ensemble approach, standard k-means algorithm and kernel-based clustering algorithm SIMLR are used on random encoded data and raw data, then the performance is evaluated showing that ensemble clustering of encoded data is highly effective for single cell data. This framework incorporates relation among clusters however, it is sensitive to the parameters that should be determined.

Hua *et al.* [39], introduced LAK a computational pipeline for single cell data analysis. LAK uses Linnorm [40] to normalize the gene expression data by default. The number of clusters k is determined using the Gap statistics [41] but it also allows users to define k or choose other normalization methods. LAK does not have a gene filtering step instead; it uses other quality control measures as mentioned in [42]. LAK utilizes K-means with Euclidean distance as a similarity measure maximizing the between-cluster sum of squares (BCSS). Lasso and L2 penalty are introduced to the clustering process according to Sparse K-means. LAK faces a limitation of slow calculation of k when the number of cells exceeds 10000.

Vans *et al.* [43], proposed FEATS a pipeline for clustering single cell data. To preprocess the original expression data, lowly and highly expressed genes are filtered then the data is normalized. Agglomerative hierarchal clustering is later performed on normalized data utilizing Ward linkage criterion. The step of hierarchical clustering is needed so that ANOVA test can be applied. ANOVA test is used as a feature selection step based on the F-value. Genes with higher F-values are selected, PCA is performed to reduce the dimensionality, then Gap statics [41] is used to determine the number of clusters. A silhouette coefficient is measured to make sure samples are correctly grouped and only groupings that maximize the clustering score are kept. FEATS allows fitting to flexible cluster shapes; however, it suffers from high time complexity.

Cui *et al.* [44], proposed SCENA an unsupervised method for clustering single cell data. SCENA takes a gene expression matrix as input; where rows represent the genes and columns represent the cells. SCENA performs preprocessing in three stages: gene filtering, log transformation and normalization. Similarity matrices between cells based on Euclidean distance and K-nearest neighbors are generated per each highly variable feature set. For each final similarity matrix, spectral clustering is performed resulting in a binary matrix that indicates whether two cells belong to the same cluster or not. A consensus matrix is generated from the binary matrices and spectral clustering is used again to perform consensus clustering. A strength point of SCENA is it makes no prior assumption about data distribution; however, spectral clustering is computationally intensive for large datasets.

It can be concluded that clustering single cell data is a challenging computational problem that needs further exploration. Methods that follow partitioning clustering and methods using neural networks or affinity propagation are sensitive to outliers. Other methods like hierarchical clustering suffer from high time complexity. While density-based clustering suffers from both back draws, ensemble-based clustering shows robust clustering performance as it integrates multiple methods.

III. PROPOSED METHODOLOGY

The proposed system preserves the same workflow of scRNA-seq analysis with the following steps: First, datasets are preprocessed in two steps; gene filtering and matrix normalization. Secondly, a variational auto-encoder (VAE) is used for dimensionality reduction followed by a clustering step. In the clustering step, cells are clustered into groups where each cluster represents a unique type of cells. In the clustering step, the latest feature space resulting from dimensionality reduction step is used as input to three different algorithms (PSO, GWO, MVO). The best solutions resulting from these algorithms are used to generate a binary matrix that is clustered again by GWO, and the consensus solution is returned. Fig.1 illustrates the workflow of the proposed CNIC approach.

A. DATA PREPROCESSING

The proposed approach takes an expression matrix of dimensions ($n \times m$); where n represents the number of cells (samples), and m represents the number of genes/transcripts (features). Firstly, the genes (features) are filtered out to eliminate non-informative genes. The elimination process is required to handle the dropout problem. Dropout is a term used to indicate an expression of zero of a certain gene in a cell. This happens since RNA amounts extracted during the sequencing from the cells is almost insignificant. The elimination is decided upon the zero ratio; a gene is eliminated if

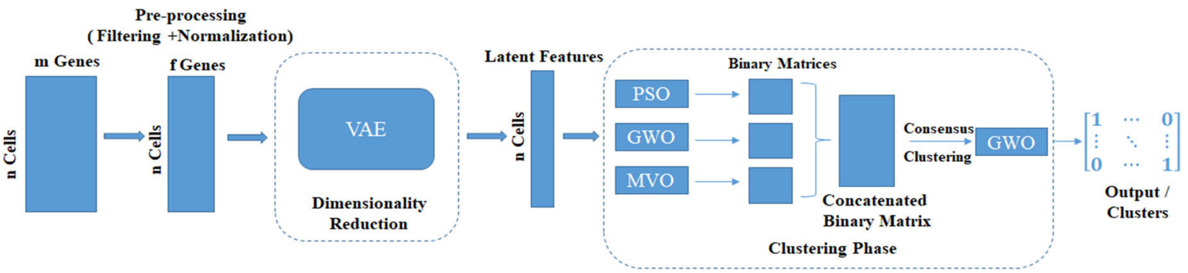

FIGURE 1. Proposed CNIC clustering workflow.

it has zero expression in all cells. Genes expressed in number of cells less than a certain threshold are also eliminated. The threshold is set that a gene must be expressed in more than $q\%$ of cells and no more than (100-q) % of cells. The threshold q is set here to 6 based on experiment of different values for q, best results are acquired when q is equal to 6.

After the elimination process, the new expression matrix is normalized according to equation (1).

$$\mathrm{dprocessed}_{\mathrm{nm}} = \frac{\mathrm{d}_{\mathrm{nm}} - \mathrm{d}_{\mathrm{m}}^{\mathrm{min}}}{\left(\mathrm{d}_{\mathrm{m}}^{\mathrm{max}} - \mathrm{d}_{\mathrm{n}}^{\mathrm{min}}\right) + \delta} \tag{1}$$

where d_{nm} is the original gene m expression at cell n, d_m^{min}, d_m^{max} are the minimum and maximum gene m expression amidst all cells respectively. δ is a relatively insignificant number as 10^{-10} to avoid dividing null values.

Single cell RNA sequencing data suffer from the curse of dimensionality since the number of features (genes/transcripts) exceeds the number of samples (cells). The process of clustering single cell sequencing data is computationally expensive. Hence, dimensionality reduction is a must to save both time and memory. Many dimensionality reduction methods were introduced to the single cell sequencing data analysis like principal component analysis (PCA), principal co-ordinate analysis (PcoA), stochastic embedding neighbor (t-SNE), uniform manifold approximation and projection (UMAP), etc [45]. However, such methods, for example PCA, assumes the linearity of the data which might not be the case. Instead, in the proposed approach a variational auto-encoder (VAE) structure is presented to encode the data into a latent feature space. VAEs compress the high dimensional space into a latent space of fewer features however, the latent feature space preserves the biological information of the original space. VAEs differ from normal reduction methods that it can uncover nonlinear features. VAEs also differ from regular auto-encoders in its stochastic nature. Instead of just deterministically encoding and decoding the data based on the construction error like the case of auto-encoders, VAEs assimilate the distribution of features over samples through learning mean and variance of data. Kullback-libler divergence is added to the reconstruction loss so that the latent features match a Gaussian distribution. VAEs use re- parameterization trick to grant back propagated gradient so that representation is learnt simultaneously. VAE consists of an encoding phase and a decoding phase; the encoding phase is responsible for compression of the data. During decoding phase, the data is reconstructed into its original shape and a loss value is calculated. The aim of decoding in this approach is to ensure the compressed latent space correctly represents the data. Accordingly, VAE in this context could be summarized to an input layer, an encoding phase and a decoding phase as follows:

1. Input layer holds the input shape of the single cell sequencing data after preprocessing; a normalized expression matrix of dimensions (m) representing the samples (cells) number and (p) representing the features (genes/transcripts) number after the filtering process.
2. Encoding phase consists of a dense layer, a batch normalization layer, an activation layer, another dense layer followed by another batch normalization layer and finally an activation layer. The result from the encoding phase is an expression matrix of dimensions ($m \times l$) where m represents the samples (cells) number and (l) represents the latent features after compression.
3. Decoding phase is a single layer of sigmoid activation.

Fig. 2 shows the architecture of the VAE used.

B. CLUSTERING

Given a dataset

$$\boldsymbol{S} = \{\boldsymbol{S_1}, \boldsymbol{S_2}, \ldots\ldots.., \boldsymbol{S_n}\} \tag{2}$$

where each instance needs to be assigned to only one of non-overlapping clusters. Suppose a set of subsets representing the clusters

$$\boldsymbol{C} = \{\boldsymbol{C_1}, \boldsymbol{C_2}, \ldots\ldots\ldots\ldots\ldots, \boldsymbol{C_k}\} \tag{3}$$

such that

$$\boldsymbol{S} = \coprod\nolimits_{i=1}^{k} \boldsymbol{C_i} \tag{4}$$

and

$$\boldsymbol{C_i} \cap \boldsymbol{C_j} = \boldsymbol{\varphi} \tag{5}$$

while $i \neq j$.

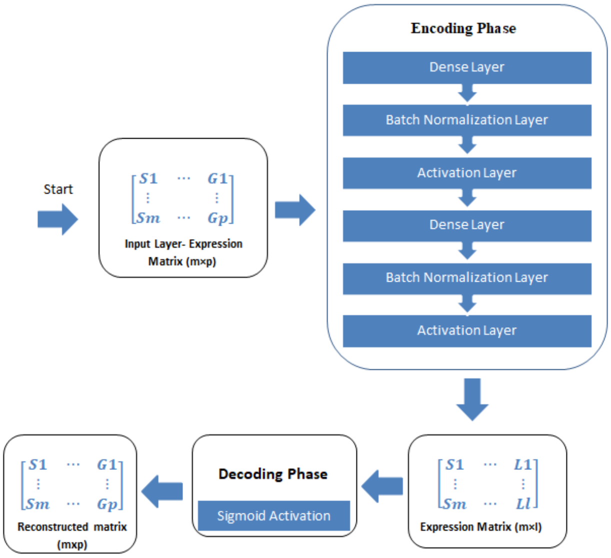

FIGURE 2. VAE architecture.

Three different algorithms are used in this approach to solve the clustering problem Particle Swarm Optimization algorithm, Multi-Verse Optimization algorithm and Grey Wolf Optimization algorithm. Each algorithm starts with a random initialization of a population of size X where each individual represents a solution to the clustering problem. The individuals' structure is shown in fig.3 where each individual consists of k centroids with f features.

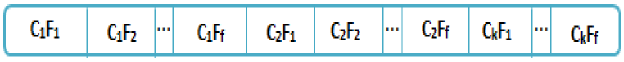

FIGURE 3. Individuals structure.

A binary matrix of the best solutions per algorithm according to the fitness function is then constructed. Then, the binary matrix is fed to the Grey Wolf algorithm once again to be clustered to find a better solution. The fitness function used by all algorithms is Silhouette Coefficient (SC) [46] calculated as mentioned in equations (6,7).

$$\mathrm{SCK} = \frac{1}{\mathrm{P}} \sum\nolimits_{\mathrm{i}=1}^{\mathrm{P}} \frac{\mathrm{bs(i)} - \mathrm{ds(i)}}{\max(\mathrm{bs(i)}, \mathrm{ds(i)})} \tag{6}$$

$$\mathrm{SC} = \frac{1}{\mathrm{k}} \sum\nolimits_{\mathrm{i}=1}^{|K|} \mathrm{SCK} \tag{7}$$

Such that: $s(i)$ represents an instance in cluster i where $i = 1, 2, 3, \ldots., k$. $bs(i)$ represents the average distance between instance $s(i)$ and all remaining instances in the exact cluster. $ds(i)$ represents the minimum distance between instance $s(i)$ and all other instances in all clusters. K, P represent the number of clusters and the number of all instances in a given cluster respectively.

The SC value is maximized to find better solution however, in the proposed approach the SC value is normalized, and a reversed value of SC is used as mentioned in equation (8).

$$\mathrm{SCrev} = 1 - \mathrm{Norm}\,(SC) \tag{8}$$

Phase 1 (Finding Optimal Number of Clusters k Automatically): To determine the optimal number of clusters without any prior knowledge, three methods are used Bayesian Information Criterion score (BIC) [47], Calinski–Harabasz(CH) [48], and Gap Statistic [41]. A median value of the three methods' results is calculated and used as the number of clusters.

Phase 2 (Clustering With Individual Algorithms):

1) CLUSTERING WITH PARTICLE SWARM OPTIMIZATION (PSO) ALGORITHM

PSO solves the clustering problem through the movement of the particles in the search space.

The goal of each particle's movement is to gain optimum velocity according to its local best (*Plocal_best*) value, and its neighbor's global best (*Pglobal_best*). A particle's position changes according to its current position, its current velocity, its distance from the (*Plocal_est*), and its distance from (*Pglobal_best*). All particles update their positions and velocities based on equations (9) and (10) respectively.

$$\mathrm{P_i}\,(\mathrm{t}+1) = \mathrm{P_i}\,(\mathrm{t}) + \mathrm{v_i}(\mathrm{t}+1) \tag{9}$$

$$\begin{aligned}\mathrm{V_i}\,(\mathrm{t}+1) = {} & \mathrm{Wv_i}\,(\mathrm{t}) + \mathrm{a_1r_1}\left(\mathrm{P_{Local\_best}}\,(\mathrm{t}) - \mathrm{P_i}\,(\mathrm{t})\right) \\ & + \mathrm{a_2r_2}(\mathrm{P_{global\_best}}\,(\mathrm{t}) - \mathrm{P_i}(\mathrm{t}))\end{aligned} \tag{10}$$

where Pi (t) , V_i (t) indicates the particle's position and velocity at iteration t respectively. The terms a_1, a_2 are acceleration coefficients while w is the inertia weight and r_1, r_2 are random numbers.

2) CLUSTERING WITH GREY WOLF OPTIMIZATION ALGORITHM (GWO)

The mathematical model of the GWO in this approach is as follows: The α wolf represents the fittest solution while β and δ wolves represent the second and third best solutions, respectively. All other solutions represent the ω wolves who follow the α, β, δ wolves leading the hunting process. Search agents known by the grey wolves, not including the fittest ones, encircle the pray according to equations (11), (12).

$$\mathrm{P}\,(\mathrm{t}+1) = \mathrm{Xp}\,(\mathrm{t}) - \mathrm{A} \times \mathrm{D} \tag{11}$$

$$\mathrm{D} = |\mathrm{C} \times \mathrm{Xp}\,(\mathrm{t}) - \mathrm{P}\,(\mathrm{t})| \tag{12}$$

where D represents the distance between position of the prey Xp and position of the search agent P at iteration t. Fittest grey wolves represented by α, β, δ wolves adjust their positions according to the prey's position according to the search agents' positions to start the hunting process modeled by equations (13-21).

$$\mathrm{A} = 2 \times \mathrm{a} \times \mathrm{r1} - \mathrm{a} \tag{13}$$

$$\mathrm{C} = 2 \times \mathrm{r2} \tag{14}$$

$$\mathrm{D} = |\mathrm{C1} \times \mathrm{P}\alpha - \mathrm{P}| \tag{15}$$

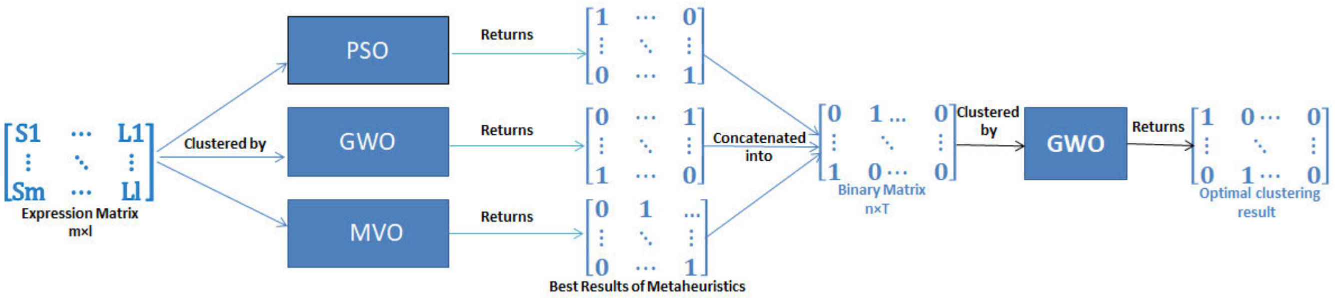

FIGURE 4. Consensus clustering process.

$$D = |C2 \times P\beta - P| \tag{16}$$

$$D = |C3 \times P\delta - P| \tag{17}$$

$$X1 = P\alpha - A1 \times D\alpha \tag{18}$$

$$X2 = P\beta - A2 \times D\beta \tag{19}$$

$$X3 = P\delta - A3 \times D\delta \tag{20}$$

$$P\,(t+1) = \frac{X1 + X2 + X3}{3} \tag{21}$$

where *D* represents the distance between the fittest grey wolves and best search agents, *A* and *C* are control coefficients to maintain exploration, r_1 and r_2 are random numbers in range [0, 1].

3) CLUSTERING USING MULTI-VERSE OPTIMIZATION (MVO) ALGORITHM

The MVO solves the clustering problem by forming universes such that each universe represents a solution to the clustering problem. Each solution consists of clusters centroids.

$\mathrm{Ui} = \{c_{i1}, c_{i2}, \ldots\ldots, c_{ik}\}$, and each centroid's dimensions are the features of the dataset initialized randomly. Then, for each universe the fitness of the universe known by inflation rate is calculated using an objective function. The best solution according to the objective function is obtained, and all universes are updated to move towards the best solution. The inflation rate is recalculated per universe, and the parameters maintaining the exploitation around the best solution are updated. The process is iteratively repeated until it reaches max number of iterations. Finally, the best universe and the cluster labels according to that universe formation is returned.

Phase 3 (Consensus Clustering): Last phase in the clustering process is finding a consensus solution that outperforms other solutions found by the clustering algorithms.

In this phase, the best solutions returned by the three clustering algorithms are used to generate a binary matrix. The binary matrix is of dimensions $N \times T$ such that N represents the number of cells to be clustered and T represents the cluster labels that resulted from the clustering algorithm per run. Since the optimization algorithms are run multiple times to ensure its performance; best solutions according to the fitness function are returned and the cluster labels results are used to generate a binary matrix per best solution. The matrix consists of all cells (samples) in the dataset as rows and the number of columns in the matrix is equal to the number of clusters k. All entries per row are zeros except for the cluster number that the cell (sample) belongs to is indicated by 1. All binary matrices are concatenated into one matrix. The new binary matrix is used as input to the GWO algorithm once again to be clustered and the final solution is found. Fig.4 illustrates the consensus clustering process.

C. EVALUATION MEASURES

To assess the performance of the clustering process, different evaluation measures are used to further prove the superiority of consensus nature inspired approach to other existing approaches in literature. Since the true labels and number of clusters are publicly available by the original authors of the datasets; evaluation measures such as Adjusted Rand Index, Completeness score, Homogeneity score and V-measure score are used.

1) ADJUSTED RAND INDEX (ARI) [49]

ARI is a measure that evaluate the similarity of two clustering results the ground truth and the predicted labels. ARI values range from −1 to 1 such that lower values specify poor clustering results while higher values closer to 1 specifies similar clustering result to the ground truth. 1.0 indicates perfect matching score between the predicted results and the ground truth. ARI makes no prior assumptions on the cluster structure hence it could be used to compare different algorithms. ARI is calculated according to equation (22).

$$\mathrm{ARI} = \frac{\mathrm{RI} - \mathrm{E[RI]}}{\max\,(\mathrm{RI}) - \mathrm{E[RI]}} \tag{22}$$

Such that *RI* is the Rand Index, *E[RI]* is the Expected Rand Index and *max (RI)* is the maximum Rand Index. Rand Index is calculated according to equation (23) as follows:

$$\mathrm{RI} = \frac{\sum_{p,t}\binom{n_{pt}}{2}}{\binom{N}{2}} \tag{23}$$

TABLE 1. Datasets description.

Dataset	Source	Accession number	#Cells	#Features	REF
Biase	NCBI	GSE7249	90	25737	3
Goolam	ArrayExpress	E-MTAB-3321	124	41480	5
Melanoma cancer	NCBI	GSE72056	4645	23686	2/7*
T cells in NSCLC	NCBI	GSE99254	12346	23458	16

TABLE 2. Estimation of k by CNIC versus other methods.

Dataset	Ref	CNIC	SC3	SINCERA	SNN-Cliq
Biase	3	**3**	3	5	6
Goolam	5	**5**	6	4	21

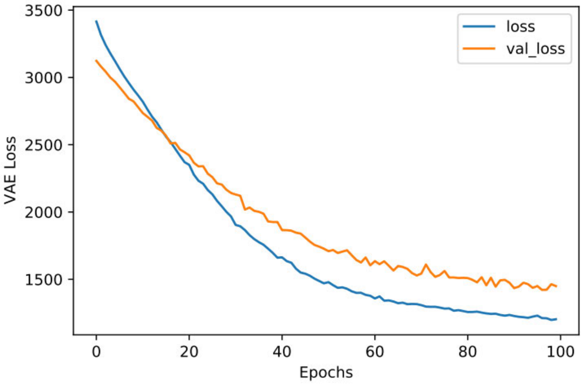

FIGURE 5. Loss Vs. Val_loss for Biase dataset.

Expected Rand Index and maximum Rand Index are calculated using equations (24), (25) respectively,

$$E(RI) = E\left(\sum\nolimits_{p,t}\binom{n_{pt}}{2}\right) \tag{24}$$

$$\max[RI] = \frac{1}{2}\left[\sum\nolimits_{p=1}^{|p|}\binom{n_{pt}}{2} + \sum\nolimits_{t=1}^{|T|}\binom{n_t}{2}\right] \tag{25}$$

where *p* represents the predicted clusters, *t* represents the true clusters and *n* is the number of data points.

2) COMPLETENESS, HOMOGENEITY, AND V-MEASURE

All three measures are used as intuitive metrics that uses conditional entropy analysis, on condition of having prior knowledge of the ground truth assignments.

a: HOMOGENEITY SCORE (HS)

Homogeneity indicates that each cluster includes only points (samples) of a single class. It is bounded by 0.0 and 1.0, where 0 specifies random clustering and 1 specifies perfect score.

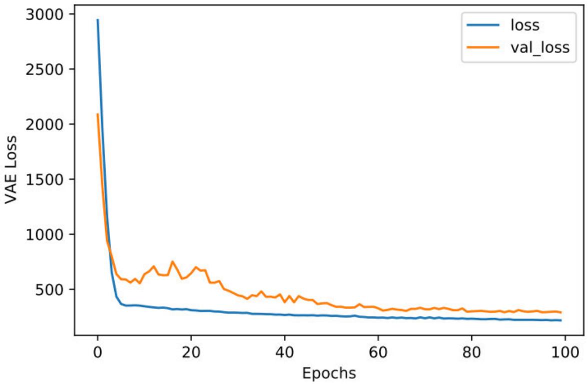

FIGURE 6. Loss Vs. Val_loss for Goolam dataset.

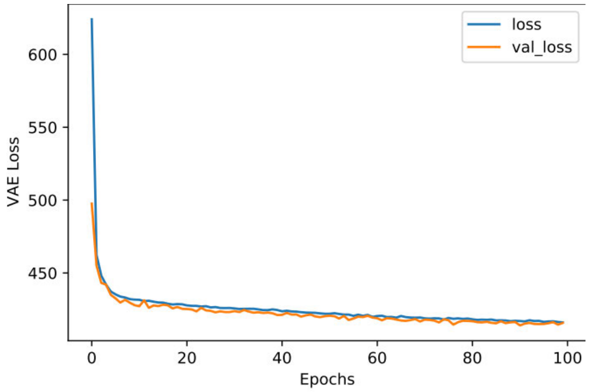

FIGURE 7. Loss Vs. Val_loss for Melanoma cancer dataset.

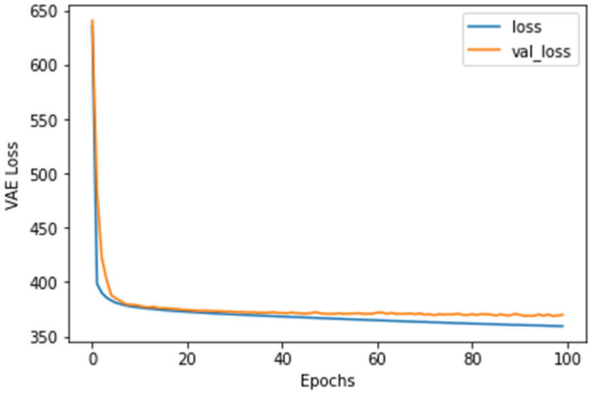

FIGURE 8. Loss Vs. Val_loss for Lung cancer dataset.

HS is calculated using equation (26):

$$HS = 1 - \frac{H(T\,|\,P)}{H(T)} \tag{26}$$

b: COMPLETENESS SCORE (CS)

Completeness means that all members of a certain class are allocated to the same cluster. This score is also bounded by 0.0 and 1.0 such that 0 indicates random clustering and

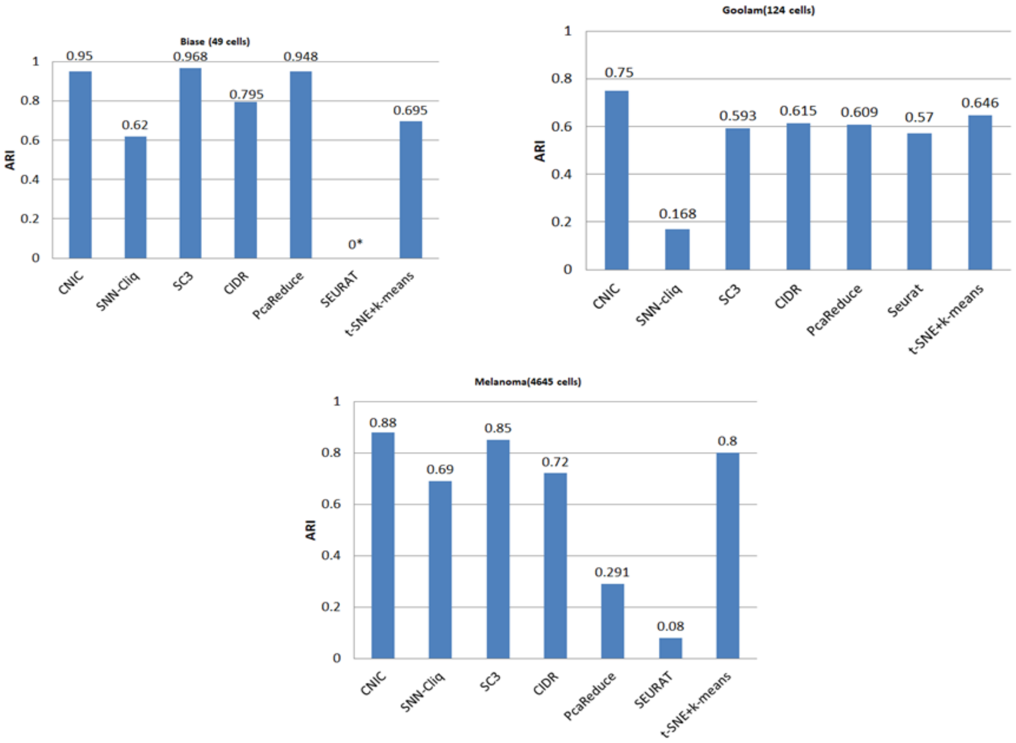

FIGURE 9. Performance comparison of CNIC and other six tools. Comparisons are based on ARI values. ∗ SEURAT operates on datasets with more than 100 cells.

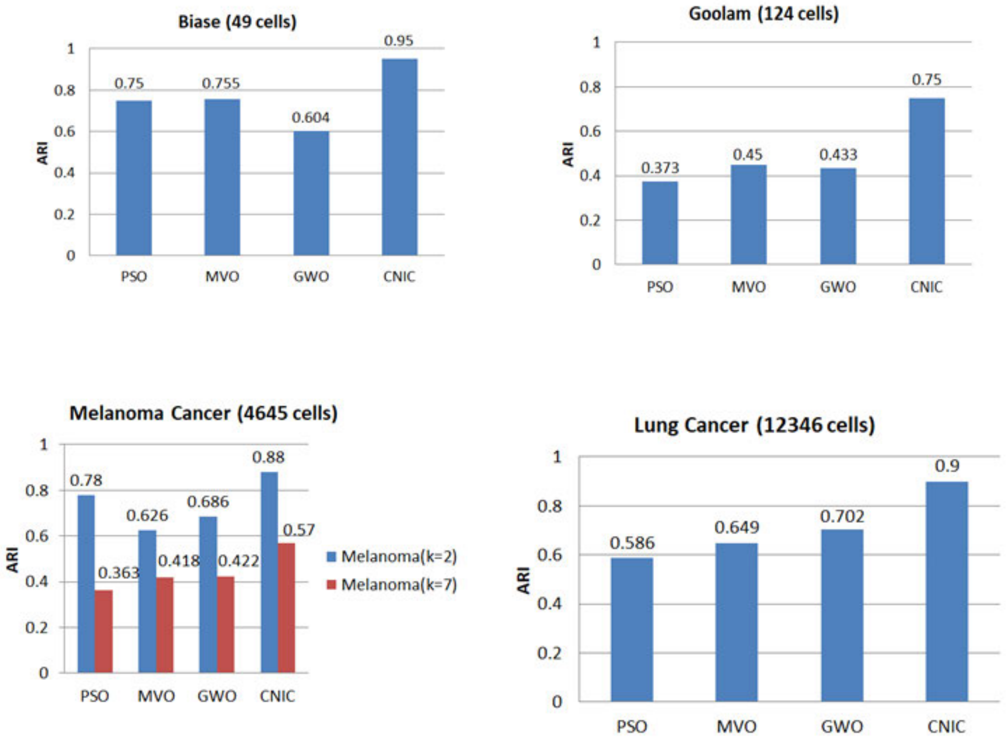

FIGURE 10. Comparison of ARI scores of all implemented algorithms.

1 indicates perfect score. CS is calculated using equation (27):

$$CS = 1 - \frac{H(P \mid T)}{H(P)} \tag{27}$$

c: V-MEASURE (VM)

A harmonic mean that makes no assumption on the cluster structure. It is also used to qualitatively interpret the

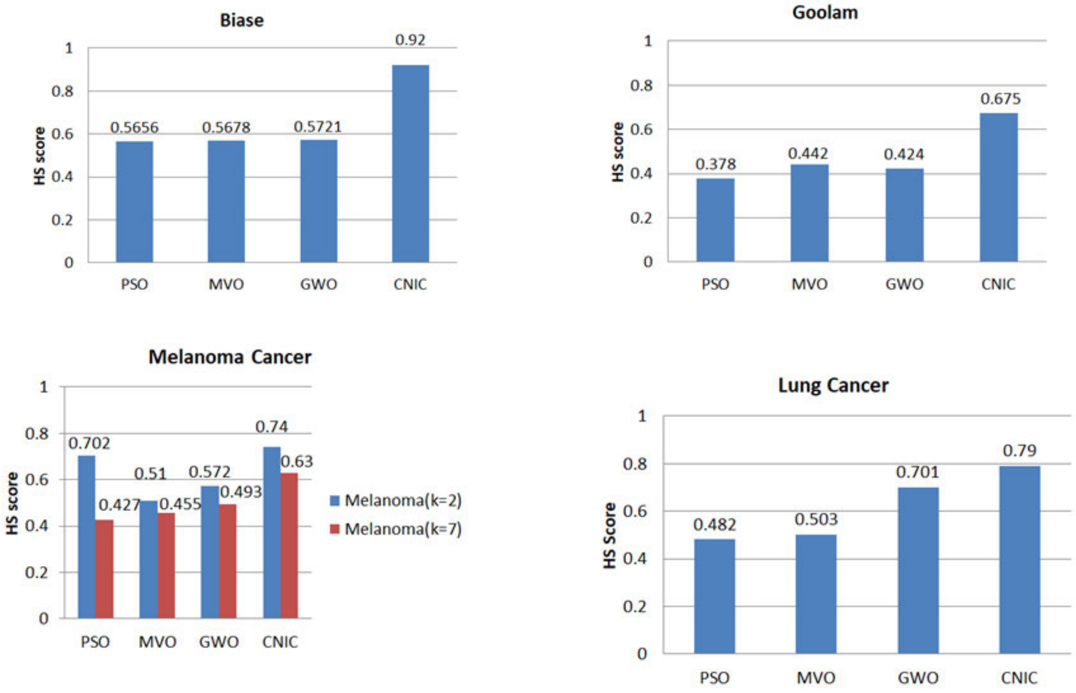

FIGURE 11. Comparison of HS scores of all implemented algorithms.

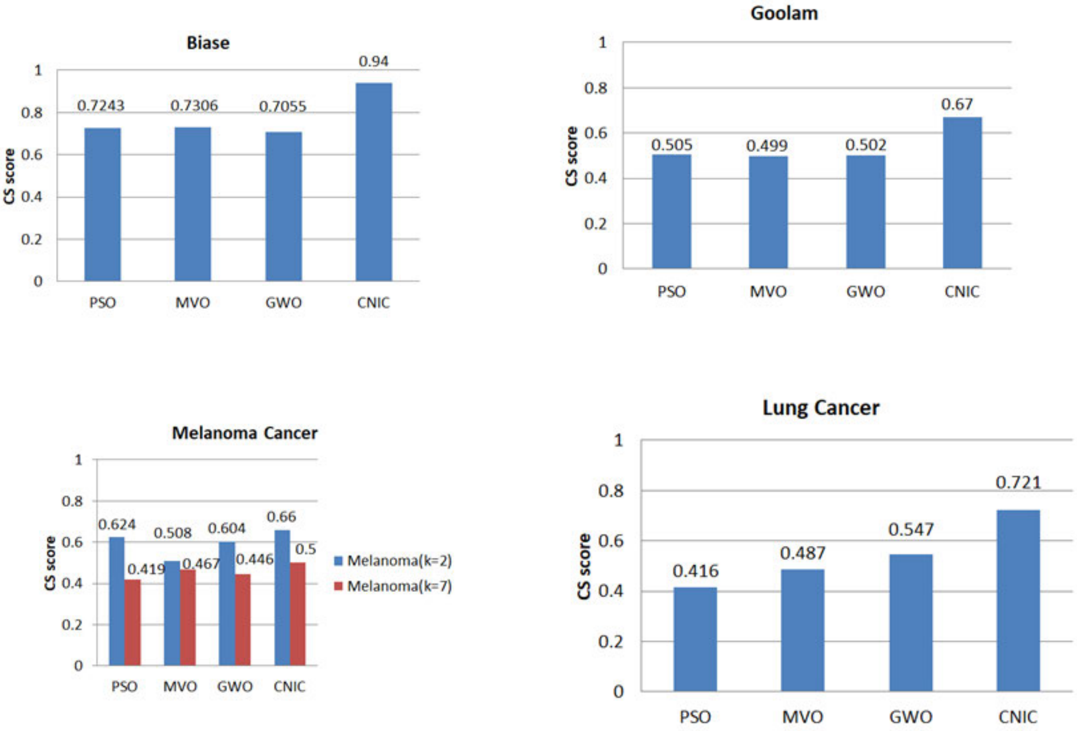

FIGURE 12. Comparison of CS scores of all implemented algorithms.

clustering results. VM is calculated as mentioned in equation (28):

$$\mathrm{VM} = 2.\frac{\mathrm{HS.CS}}{\mathrm{HS} + \mathrm{CS}} \tag{28}$$

Such that $H(P)$ is the cluster entropy, $H(P|T)$ is the clusters conditional entropy, $H(T)$ is the ground truth entropy and $H(T|P)$ is the ground truth conditional entropy.

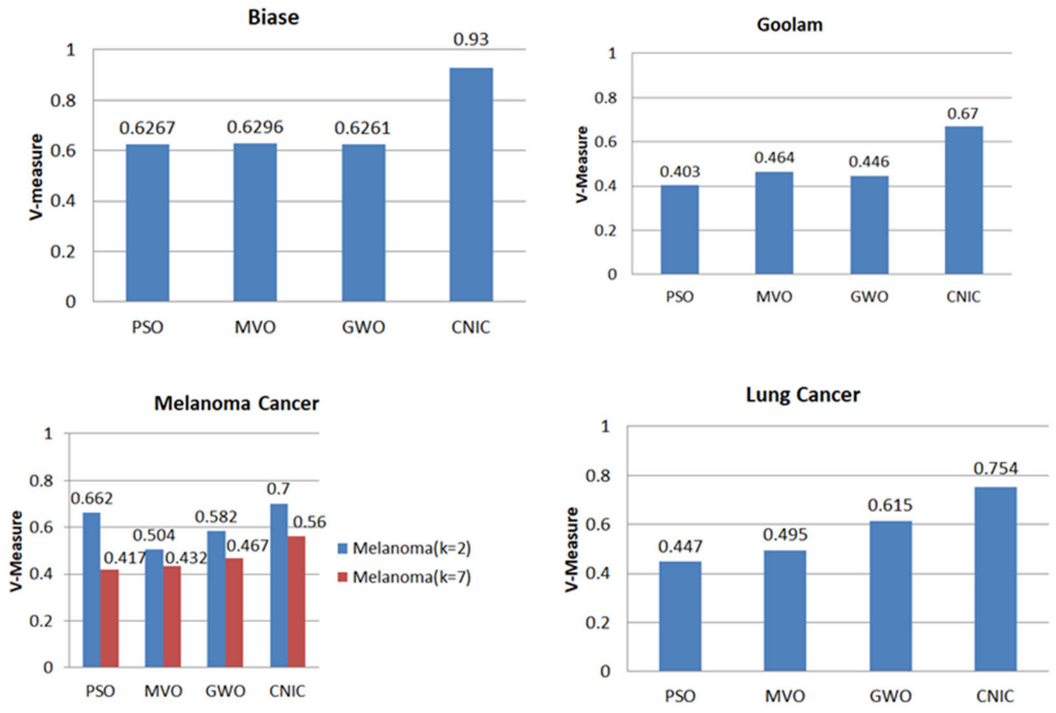

FIGURE 13. Comparison of V-measure scores of all implemented algorithms.

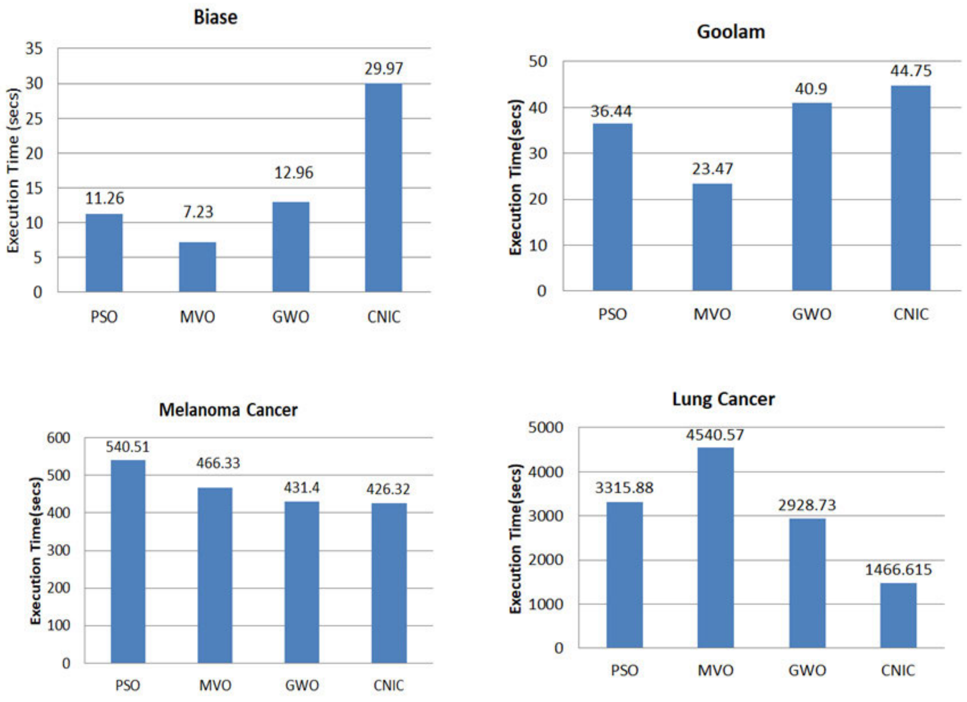

FIGURE 14. Execution Time of all implemented algorithms.

IV. EXPERIMENTS AND RESULTS

In this section, the experimental configuration and parameter settings to evaluate the performance of the proposed approach for the task of clustering single cell sequencing data are discussed. Description of the datasets used for validation is given in this section, as well as the discussion and comparison of the results of the proposed approach with results from literature.

A. DATASET

Two benchmark datasets with golden standards (Biase [50] and Goolam [51]) are used for the experimentation and evaluation of the CNIC clustering approach. The melanoma cancer dataset is used to decipher the cellular composition of the heterogeneous complex ecosystem of the tumor [52]. The lung cancer dataset [53] discusses the complexity of T cells in non-small-cell-lung cancer (NSCLC) considered the main reason of cancer mortality accounting for 85% of lung cancers [54].

All datasets are free to access using NCBI (National Center for Biotechnology Information) and Array Express. Table 1 summarizes the datasets used, their dimensions and the number of clusters according to the authors of the datasets.

B. SYSTEM CONFIGURATION AND PARAMETER SETTINGS

Experiments were executed using a 2.00 GHz Intel(R) Core (TM) i7-3537U- processor with 8 GB memory on Windows 10 operating system. The entire workflow was implemented in Python using Spyder, Google colab and Kaggle Notebooks.

C. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed approach is applied to the four public mentioned datasets, and results are compared to recent used tools for clustering Single cell data. Table 2 shows the number of clusters computed by the proposed approach and other tools against the ground truth mentioned by the original authors on benchmark datasets with golden standards.

1) VAE TRAINING AND TESTING

To train and evaluate the VAE model, the data is split into a training set and a testing set with percentages (70%-30%) respectively. The 70%-30% was chosen in order to have enough data for the training to result in better performance. Also, the error estimation is more accurate with enough data for testing. Other train-test split criteria were used (80%-20%,90%-10%) but the results were not accurate. The loss value is evaluated on both the training set and the testing set. For the hyper parameters such as learning rate, number of epochs and batch size, many experiments were conducted with different values for each parameter. Eventually, Adam optimizer is used with a learning rate of.001 chosen from values (.01,.005,.0001,.1) since it achieved better results. A batch size of 50 and 100 epochs are used out of many suggested values for both parameters based on better performance.

Fig. (5-8) shows the reconstruction loss and validation loss calculated by mean square error as the loss function per dataset.

D. CLUSTERING PHASE

In the experiments, PSO, GWO, MVO and consensus clustering had a maximum of 100 runs, a maximum iteration of 100 and a population size of 50. Experiments were replicated for all datasets used with the same settings. Four datasets were used to assess the performance of the proposed approach against competing approaches.

TABLE 3. Comparison of CNIC with PSO, GWO and MVO over 100 independent runs.

Dataset	Algorithm	Fitness Function (Silhouette Coefficient)	
		Mean	StdDev
Biase	PSO	0.3633	0.004726
	MVO	0.3634	0.004969
	GWO	0.3625	0.004578
	Consensus	**0.03**	**7.31×10^{-18}**
Goolam	PSO	0.3763	0.013000
	MVO	0.3833	0.013108
	GWO	0.3771	0.014654
	Consensus	**0.11**	**1.463 ×10^{-17}**
Melanoma Cancer	PSO	0.354	0.005477
	MVO	0.352	0.0083666
	GWO	0.354	0.015166
	Consensus	**0.0608**	**1.59 ×10^{-17}**
Lung Cancer	PSO	0.468	0.008367
	MVO	0.484	0.008944272
	GWO	0.474	0.011402
	Consensus	**0.33**	**0.000113**

Standard deviation is used as a descriptive statistic to detail the computed solutions obtained by the proposed clustering approach. The average objective function for all 100 runs of the algorithm is calculated and reported as well as the average execution time needed to find the clustering solutions by all algorithms.

Table 3 shows the mean and standard deviation of the fitness function (Silhouette Coefficient) of each algorithm across all runs on all datasets.

1) BENCHMARKING

To evaluate the clustering accuracy, the performance of the proposed CNIC is compared to currently most used methods in clustering Single cell data using their default parameters as mentioned by the authors. These methods are SNN-Cliq, SC3, CIDR, PcaReduce, SEURAT and t-SNE+k-means. All approaches were applied to Biase, Goolam and melanoma cancer datasets. The results were evaluated by ARI.

Fig.9 demonstrates the ARI values obtained by the proposed clustering approach against the other methods. It is shown that the proposed CNIC approach achieves better

TABLE 4. Top 20 Signature genes of predicted cluster for lung cancer T cells dataset.

Cluster number	Signature genes
Cluster 1	CCR7, RPS6, SELL, RPL3, RPL13, EEF1G, TXK, RPL32, RPL5, RPS14, RPL11, TCF7, RPS13,RPS4X, LEF1, RPL19, PRKCQ-AS1, RPL18, RPL31, RPS12
Cluster 2	LTB, KLF2, AES, BIRC3, GPR183, RPLP0, GIMAP7, RPSAP58, GSTK1, ICAM2, TRADD, RPL17, SERINC5, SORL1, NOP53, PTGER2, ADD3, RNASET2, FLT3LG, RASA3
Cluster 3	NKG7, GZMH, GNLY, PRF1, FGFBP2, GZMB, CCL5, TARP, CX3CR1, KLRG1, PLEK, CTSW, A2M, ADGRG1, FCRL6, ZEB2, HLA-DPB1, LITAF, PLAC8, MYOM2
Cluster 4	FOS, CD69, DUSP1, FOSB, CCR6, SLC2A3, OCIAD2, LOC100130476, PDCD4, NABP1, GZMK, GZMA, CST7, CCL4, CCL4L1, CD27, CCL3L3, DTHD1, F2R, CD27-AS1
Cluster 5	CXCR6, HLA-DRB5, HLA-DRB1, HLA-DRB6, ALOX5AP, CAPG, LGALS3, ADAM19, JAML, CXCR3, ITGAE, HLA-DRA, FAM129A, CD2, ANXA2, HLA-DQB1, SH3BGRL3, CKLF, CKLF-CMTM1, ANXA2P2
Cluster 6	CXCL13, SRGN, RGS1, GAPDH, CTLA4, TNFRSF18, DUSP4, RBPJ, NR3C1, RNF19A, PDCD1, TIGIT, BHLHE40-AS1, TNFRSF4, TOX, ITM2A, SNX9, CDK6, SLA, DNPH
Cluster 7	NAP1L4, OAS3, TNFAIP3, ENTPD1, SRGAP3, APOBEC3C, BATF, CD7, SH2D2A, GNG5, PDE4D, SUSD6, SEM1, MAF, VMP1, AHR, GALNT2, SLC1A5, ANXA5, MIR497HG
Cluster 8	TXNIP, GIMAP7, S1PR1, RPL13, CD52, RPS3, UBXN11, LINC00861, RPS6, RPS18, RPS14, GIMAP4, AES, RPL3, GIMAP1-GIMAP5, RPLP2, RPL32, RPL13A, RPS14P3, RPL19
Cluster 9	TNFRSF18, CCR8, CXCR6, CD7, IL1R2, CTSC, TNFRSF9, DUSP4, SH2D2A, GAPDH, TNFRSF4, BATF, CREM , RGS1, ID2, ICOS, TNFAIP3, IL2RB, CTLA4, PHTF2
Cluster 10	CCR7, SELL, LEF1, TCF7, RPL13, RPS6, TXK, RPL4, LDHB, NOSIP, RPL3, EEF1G, RPL32, EEF1A1, RPL5, LDLRAP1, RPL19, SERINC5, LRRC75A-AS1, RPS4X
Cluster 11	IL7R, CD28, LYAR, TNFSF8, MCUB, GPR183, DUSP2, SESN1, MALAT1, EPB41, ATP2B1, CYB561, ZFP36L2, IL10RA, HELB, PATJ, UPP1, JUNB, PTGER2, MYBL1
Cluster 12	FGFBP2, CX3CR1, FCGR3A, ADGRG1, PLEK, FCGR3B, KLRD1, S1PR1, LITAF, GZMH, FCRL6, GNLY, KLRG1, NKG7, S1PR5, PRF1, PLAC8, A2M, ZEB2, FGR
Cluster 13	GZMK, CCL4L1, ITM2C, CD74, CCL4, AOAH, CXCR4, DTHD1, CCL3L3, CCL3L1, CLDND1, CD44, SH2D1A, TRAT1, EOMES, CCL5, F2R, TC2N, FAM102A, PVRIG,
Cluster 14	ZNF683, CAPG, ITGA1, STK17B, JUN, ADAM19, CKLF-CMTM1, CKLF, CXCR3, XCL1, PELO, CTSA, PLTP, VIM, PLP2, SUSD3, CD69, GLUL, S100A11, ZYX
Cluster 15	HAVCR2, SIRPG, GAPDH, ITGAE, GZMB, CCL3, TIGIT, ENTPD1, PDCD1, RBPJ, RGS1, CXCR6, CD63, SAMSN1, CD82, CCND2, ENTPD1-AS1, HLA-DRA, FKBP1A-SDCBP2, COTL1
Cluster 16	KLRB1, SLC4A10, NCR3, LST1, LTB, CCR6, DPP4, SPOCK2, SLAMF1, JAML, DUSP1, RORA, TNFRSF25, CTSH, ERN1, IFNGR1, MAF, IL18RAP, MPZL3, ZBTB16

performance than most of the mentioned tools for all datasets. For Biase dataset, CNIC achieved better ARI value of .95 than SNN-Cliq, CIDR, PcaReduce, SEURAT and t-SNE+k-means. For Goolam dataset, CNIC achieved an ARI value of .75 higher than all methods. For Melanoma cancer dataset, CNIC got an ARI value of .88 higher than all mentioned methods.

It is also shown that the proposed CNIC reaches high ARI values regardless the size of the dataset samples (the number of cells).

2) CNIC CLUSTERING STABILITY

A comparison between solutions of the used metaheuristic algorithms and the consensus solution found in terms of ARI is performed. The ARI values shown are the average of the 100 runs of each algorithm.

Fig. 10 shows a comparison between the ARI values of the best solutions resulting per algorithm and the consensus solution. As shown, the CNIC approach achieves .95 for Biase dataset,.75 for Goolam dataset, .88 for melanoma cancer and.9 for lung cancer respectively. The results indicate the efficiency of the proposed CNIC approach.

For assessing the clustering stability, all implemented algorithms as well as the consensus solution were evaluated in terms of homogeneity score, completeness score and V-measure.

Fig.11, Fig.12 and Fig.13 show the homogeneity score, completeness score and V-measure score of the proposed approach and other implemented algorithms respectively. It is noticed that CNIC achieves better scores indicating clustering stability even in cases of large datasets.

Fig.14 shows the average running time for 100 runs for each algorithm. Results show that it takes seconds to cluster small datasets and a reasonable time in case of large datasets.

As shown above, all experimental results show that the proposed CNIC approach performs better s in terms of ARI, CS, HS and VS indicating clustering stability. Also, performance of the proposed CNIC approach is not affected by the number of samples (cells) as it can perform stably whether the number of samples (cells) is small or large in a feasible running time.

FINDING MARKER GENES

For further analysis of the lung cancer T cells; ANOVA [55] test is performed to identify the signature genes of each cluster. ANOVA test is known as an analysis of variance used to determine the differentially expressed genes. Table 4 shows only top 20 signature genes of each cluster in case it has more than 20 signature genes of the main 16 predicted clusters according to the adjusted p-values.

V. CONCLUSION

In this paper, a new unsupervised consensus clustering approach based on swarm intelligence optimization algorithms is proposed to cluster scRNA-seq data. The proposed approach automatically and accurately computes the number of clusters, k, overcoming the shortcomings of other methods that require that k must be known. The proposed CNIC approach takes VAE as a dimensionality reduction method to project the original feature space into a lower dimension space, yet the created latent feature space is biologically relevant. For clustering, proposed CNIC utilizes metaheuristic algorithms PSO, GWO and MVO to cluster single cell data and returns best solutions found in the search space. Best solutions are concatenated into a binary matrix, and consensus clustering is performed fusing the solutions of the PSO, GWO and MVO into one consensus solution. The proposed CNIC approach achieves higher ARI values compared to other widely used methods indicating better clustering accuracy. The results of evaluation measures CS score, HS score and V-measure score verify the stability of the clustering results.

REFERENCES

[1] A. A. Kolodziejczyk, J. K. Kim, V. Svensson, J. C. Marioni, and S. A. Teichmann, “The technology and biology of single-cell RNA sequencing,” *Mol. Cell*, vol. 58, pp. 610–620, May 2015.

[2] F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, and K. Lao, “mRNA-seq whole-transcriptome analysis of a single cell,” *Nature Methods*, vol. 6, no. 5, pp. 377–382, 2009.

[3] T. Kalisky, S. Oriel, T. H. Bar-Lev, N. Ben-Haim, A. Trink, Y. Wineberg, I. Kanter, S. Gilad, and S. Pyne, “A brief review of single-cell transcriptomic technologies,” *Briefings Funct. Genomics*, vol. 17, no. 1, pp. 64–76, 2018.

[4] D. Tsoucas and G. C. Yuan, “Recent progress in single-cell cancer genomics,” *Current Opinion Genet. Develop.*, vol. 42, pp. 22–32, Feb. 2017.

[5] R. Petegrosso, Z. Li, and R. Kuang, “Machine learning and statistical methods for clustering single-cell RNA-sequencing data,” *Briefings Bioinf.*, vol. 21, no. 4, pp. 1209–1223, 2020.

[6] O. Stegle, S. A. Teichmann, and J. C. Marioni, “Computational and analytical challenges in single-cell transcriptomics,” *Nature Rev. Genet.*, vol. 16, pp. 133–145, Jan. 2015.

[7] P. V. Kharchenko, L. Silberstein, and D. T. Scadden, “Bayesian approach to single-cell differential expression analysis,” *Nature Methods*, vol. 11, pp. 740–742, May 2014.

[8] R. Qi, A. Ma, Q. Ma, and Q. Zou, “Clustering and classification methods for single-cell RNA-sequencing data,” *Briefings Bioinf.*, vol. 21, no. 4, pp. 1196–1208, 2020.

[9] Z. Ji and H. Ji, “TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis,” *Nucleic Acids Res.*, vol. 44, no. 13, p. e117 2016.

[10] M. W. E. J. Fiers, L. Minnoye, S. Aibar, C. B. González-Blas, Z. K. Atak, and S. Aerts, “Mapping gene regulatory networks from single-cell omics data,” *Briefings Funct. Genomics*, vol. 17, no. 4, pp. 246–254, 2018.

[11] T. Kim, I. R. Chen, Y. Lin, A. Y.-Y. Wang, J. Y. H. Yang, and P. Yang, “Impact of similarity metrics on single-cell RNA-seq data clustering,” *Briefings Bioinf.*, vol. 20, no. 6, pp. 2316–2326, 2019.

[12] E. Diday and J. C. Simon, “Clustering analysis,” in *Digital Pattern Recognition*. Berlin, Germany: Springer, 1976, pp. 47–94.

[13] P. Arabie, L. Hubert, and G. De Soete, *Clustering and Classification*. Singapore: World Scientific, 1996.

[14] F. Murtagh and P. Contreras, “Algorithms for hierarchical clustering: An overview,” *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 2, no. 1, pp. 86–97, 2012.

[15] M. G. Omran, A. P. Engelbrecht, and A. Salman, “An overview of clustering methods,” *Intell. Data Anal.*, vol. 11, no. 6, pp. 583–605, 2007.

[16] P. Novak, P. Neumann, and J. Macas, “Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data,” *BMC Bioinf.*, vol. 11, pp. 1–12, Jul. 2010.

[17] S. Vega-Pons and J. Ruiz-Shulcloper, “A survey of clustering ensemble algorithms,” *Int. J. Pattern Recognit. Artif. Intell.*, vol. 25, no. 3, pp. 337–372, May 2011.

[18] A. E. Ezugwu, “Nature-inspired metaheuristic techniques for automatic clustering: A survey and performance study,” *Social Netw. Appl. Sci.*, vol. 2, pp. 1–57, Jan. 2020.

[19] A. José-Garcia and W. Gómez-Flores, “Automatic clustering using nature-inspired metaheuristics: A survey,” *Appl. Soft Comput.*, vol. 41, pp. 192–213, Apr. 2016.

[20] X.-S. Yang, *Nature-Inspired Metaheuristic Algorithms*. Bristol, U.K.: Luniver Press, 2010.

[21] S. Mirjalili, J. S. Dong, and A. Lewis, *Nature-Inspired Optimizers*. Cham, Switzerland: Springer, 2020.

[22] N. G. Blas and O. L. Tolic, “Clustering using particle swarm optimization,” *Int. J. Inf. Theories Appl.*, vol. 23, no. 1, pp. 24–33, 2016.

[23] S. Mirjalili, S. M. Mirjalili, and A. Lewis, “Grey wolf optimizer,” *Adv. Eng. Softw.*, vol. 69, pp. 46–61, Mar. 2014.

[24] S. Mirjalili, S. M. Mirjalili, and A. Hatamlou, “Multi-verse optimizer: A nature-inspired algorithm for global optimization,” *Neural Comput. Appl.*, vol. 27, no. 2, pp. 495–513, 2016.

[25] C. Xu and Z. Su, “Identification of cell types from single-cell transcriptomes using a novel clustering method,” *Bioinformatics*, vol. 31, no. 12, pp. 1974–1980, 2015.

[26] M. Guo, H. Wang, S. S. Potter, J. A. Whitsett, and Y. Xu, “SINCERA: A pipeline for single-cell RNA-seq profiling analysis,” *PLoS Comput. Biol.*, vol. 11, Nov. 2015, Art. no. e1004575.

[27] R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, and A. Regev, “Spatial reconstruction of single-cell gene expression data,” *Nature Biotechnol.*, vol. 33, pp. 495–502, Apr. 2015.

[28] J. Žurauskiene and C. Yau, “pcaReduce: Hierarchical clustering of single cell transcriptional profiles,” *BMC Bioinf.*, vol. 17, pp. 1–11, Mar. 2016.

[29] V. Y. Kiselev, K. Kirschner, M. T. Schaub, T. Andrews, A. Yiu, T. Chandra, K. N. Natarajan, W. Reik, M. Barahona, and A. R. Green, “SC3: Consensus clustering of single-cell RNA-seq data,” *Nature Methods*, vol. 14, no. 5, pp. 483–486, 2017.

[30] J. A. Hartigan and M. A. Wong, “Algorithm AS 136: A k-means clustering algorithm,” *J. Roy. Stat. Soc. C, Appl. Statist.*, vol. 28, no. 1, pp. 100–108, 1979.

[31] A. Strehl and J. Ghosh, “Cluster ensembles—A knowledge reuse framework for combining multiple partitions,” *J. Mach. Learn. Res.*, vol. 3, pp. 583–617, Dec. 2002.

[32] P. Lin, M. Troup, and J. W. Ho, “CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data,” *Genome Biol.*, vol. 18, pp. 1–11, Mar. 2017.

[33] T. Caliński and J. Harabasz, “A dendrite method for cluster analysis,” *Commun. Statist., Theory Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[34] B. Wang, J. Zhu, E. Pierson, D. Ramazzotti, and S. Batzoglou, “Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning,” *Nature Methods*, vol. 14, pp. 414–416, Mar. 2017.

[35] Y. Gan, N. Li, G. Zou, Y. Xin, and J. Guan, “Identification of cancer subtypes from single-cell RNA-seq data using a consensus clustering method,” *BMC Med. Genomics*, vol. 11, pp. 65–72, Dec. 2018.

[36] Y. Yang, R. Huh, H. W. Culpepper, Y. Lin, M. I. Love, and Y. Li, “SAFE-clustering: Single-cell aggregated (from ensemble) clustering for single-cell RNA-seq data,” *Bioinformatics*, vol. 35, no. 8, pp. 1269–1277, 2019.

[37] H. Nguyen, S. J. Louis, and T. Nguyen, “MGKA: A genetic algorithm-based clustering technique for genomic data,” in *Proc. IEEE Congr. Evol. Comput. (CEC)*, Jun. 2019, pp. 103–110.

[38] T. A. Geddes, T. Kim, L. Nan, J. G. Burchfield, J. Y. Yang, D. Tao, and P. Yang, “Autoencoder-based cluster ensembles for single-cell RNA-seq data analysis,” *BMC Bioinf.*, vol. 20, pp. 1–11, Dec. 2019.

[39] J. Hua, H. Liu, B. Zhang, and S. Jin, "LAK: Lasso and K-means based single-cell RNA-seq data clustering analysis," *IEEE Access*, vol. 8, pp. 129679–129688, 2020.

[40] S. H. Yip, P. Wang, J.-P. A. Kocher, P. C. Sham, and J. Wang, "Linnorm: Improved statistical analysis for single cell RNA-seq expression data," *Nucleic Acids Res.*, vol. 45, no. 22, p. e179, 2017.

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

[42] D. M. Witten and R. Tibshirani, "A framework for feature selection in clustering," *J. Amer. Stat. Assoc.*, vol. 105, no. 490, pp. 713–726, 2010.

[43] E. Vans, A. Patil, and A. Sharma, "FEATS: Feature selection-based clustering of single-cell RNA-seq data," *Briefings Bioinf.*, vol. 22, no. 4, 2021, Art. no. bbaa306.

[44] Y. Cui, S. Zhang, Y. Liang, X. Wang, T. N. Ferraro, and Y. Chen, "Consensus clustering of single-cell RNA-seq data by enhancing network affinity," *Briefings Bioinf.*, vol. 22, no. 6, 2021, Art. no. bbab236.

[45] G. Chen, B. Ning, and T. Shi, "Single-cell RNA-seq technologies and related computational data analysis," *Frontiers Genet.*, vol. 10, p. 317, Apr. 2019.

[46] K. Chowdhury, D. Chaudhuri, and A. K. Pal, "A novel objective function based clustering with optimal number of clusters," in *Methodologies and Application Issues of Contemporary Computing Framework*. Singapore: Springer, 2018, pp. 23–32.

[47] S. S. Chen and P. S. Gopalakrishnan, "Clustering via the Bayesian information criterion with applications in speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 1998, pp. 645–648.

[48] X. Wang and Y. Xu, "An improved index for clustering validation based on Silhouette index and Calinski–Harabasz index," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 569, no. 5, 2019, Art. no. 052024.

[49] L. Hubert and P. Arabie, "Comparing partitions," *J. Classification*, vol. 2, no. 1, pp. 193–218, 1985.

[50] F. H. Biase, X. Cao, and S. Zhong, "Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing," *Genome Res.*, vol. 24, no. 11, pp. 1787–1796, 2014.

[51] M. Goolam, A. Scialdone, S. J. Graham, I. C. Macaulay, A. Jedrusik, A. Hupalowska, T. Voet, J. C. Marioni, and M. Zernicka-Goetz, "Heterogeneity in Oct4 and Sox2 targets biases cell fate in 4-cell mouse embryos," *Cell*, vol. 165, pp. 61–74, Mar. 2016.

[52] I. Tirosh *et al.*, "Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq," *Science*, vol. 352, no. 6282, pp. 189–196, 2016.

[53] X. Guo, Y. Zhang, L. Zheng, C. Zheng, J. Song, Q. Zhang, B. Kang, Z. Liu, L. Jin, R. Xing, R. Gao, L. Zhang, M. Dong, X. Hu, X. Ren, D. Kirchhoff, H. G. Roider, T. Yan, and Z. Zhang, "Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing," *Nature Med.*, vol. 24, pp. 978–985, Jun. 2018.

[54] L. T. Tanoue, R. S. Herbst, J. V. Heymach, and S. M. Lippman, "Molecular origins of cancer: Lung cancer," *New England J. Med.*, vol. 359, no. 13, pp. 1367–1380, 2008.

[55] A. Cuevas, M. Febrero, and R. Fraiman, "An ANOVA test for functional data," *Comput. Statist. Data Anal.*, vol. 47, no. 1, pp. 111–122, 2004.

AMANY H. ABOU EL-NAGA received the B.Sc. degree from the Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt, in 2016. After graduation, she applied for her master's degree. She is currently working as a Teaching Assistant at the Department of Computer Science, Future University in Egypt. Her research interests include evolutionary algorithms, metaheuristic techniques, and bioinformatics.

SABAH SAYED received the Ph.D. degree in computer science "A Computational Framework for Colorectal Cancer", in 2019. She is currently working as a Teacher at the Faculty of Computers and Artificial intelligence, Cairo University, Egypt. She has many scientific research articles published in international journals in the topics of bioinformatics, artificial intelligence, machine learning. Her research interests include bioinformatics and biomedical, cloud computing, soft computing, image processing, artificial intelligence, data mining, high performance computing, optimization, and meta-heuristics techniques.

AKRAM SALAH graduated in mechanical engineering. He received the Ph.D. degree in computer and information sciences from the University of Alabama at Birmingham, USA, in 1986. He has worked in computer programming for seven years, before he got his Ph.D. degree at the University of Alabama at Birmingham. He has taught at The American University in Cairo, Michigan State University, and Cairo University, before he joined North Dakota State University, where he designed and started a graduate program that offers Ph.D. and M.Sc. degrees in software engineering. He is currently a Professor at the Faculty of Computers and Artificial Intelligence, Cairo University. He has published more than 100 articles. His research interests include data, knowledge, and software engineering. His current research is in semantics and semantic web.

HEBA MOHSEN received the B.Sc. degree in computer science from Ain Shams University and the M.Sc. and Ph.D. degrees in artificial intelligence and machine learning from Ain Shams University. She joined Future University in Egypt, in 2006, where she is currently working as an Assistant Professor at the Computer Science Department, Faculty of Computers and Information Technology. During her scientific research journey, she published several papers that have been highly cited and recognized in local and international journals and conferences. Her research interests include artificial intelligence, machine learning, bioinformatics, image processing, pattern recognition, biometrics, and medical data mining.