JOURNAL OF SOFT COMPUTING AND DATA MINING
e-ISSN: 2716-621X

JSCDM

Vol. 5 No. 2 (2024) 223-244
https://publisher.uthm.edu.my/ojs/index.php/jscdm

Optimized Cancer Subtype Classification and Clustering Using Cat Swarm Optimization and Support Vector Machine Approach for Multi-Omics Data

Ali Mahmoud Ali[1], Mazin Abed Mohammed[2]*

[1] *Informatics Institute for Postgraduate Studies,*
Iraqi Commission for Computers & Informatics, Baghdad, IRAQ

[2] *Department of Artificial Intelligence, College of Computer Science and Information Technology,*
University of Anbar, Anbar 31001, IRAQ

*Corresponding Author: mazinalshujeary@uoanbar.edu.iq

DOI: https://doi.org/10.30880/jscdm.2024.05.02.017

Article Info

Received: 27 March 2024
Accepted: 11 August 2024
Available online: 18 December 2024

Keywords

Cancer, omics, multi-omics, cat swarm optimization, cancer subtype, K-Means

Abstract

There is no standard approach has been established to define cancer subtypes, making it a challenging task due to the high dimensionality of the data and the limited sample sizes. The addition of multiple levels of data increases the dimensionality, and interpreting the predictions of the machine learning (ML) model introduces an additional layer of complexity. Some prior studies have failed to explain aspects of certain characteristics that affect the classification results. The aim of this work is to improve the feature selection method, thereby increasing the significance and accuracy of characterizing cancer subtypes by using the powerful framework of clustering multi-omics data. With regard to feature selection, we propose a rigorous cat swarm optimization feature selection to isolate relevant features for prediction, K-means for clustering the dataset, and a nonlinear Support Vector Machine (SVM) for multi-classification. The performance is evaluated using the ML model's known measures of accuracy, F1-score, precision, and recall. The silhouette metric is then used to quantify the quality of clusters generated by the shortlisted features. The initial use of the proposed Cat Swarm Optimization (CSO) and SVM model achieved an accuracy of approximately 81%. After integrating the feature selection algorithm, the accuracy significantly improved to 100%, clearly outperforming existing models. The silhouette metric highlighted the effectiveness of our feature selection strategy, demonstrating a distinct and significant improvement in classification accuracy. In addition to improving classification accuracy, this method enhances interpretability, one of the initial steps toward understanding the molecular mechanisms that govern cancer subtypes. Therefore, this work establishes a solid foundation for further biological studies, contributing to the generation of new data and advancements in cancer research.

1. Introduction

Cancer is a broad term encompassing various diseases, characterized by the uncontrolled growth of cells, posing significant risks to health [1]. Majority of doctors and scientists worldwide concur that cancer is a complex adversary that demands considerable sacrifice and effort to control, given its profound effect on people's lives. Cancer is a major health concern identified by the WHO as the leading cause of increased mortality rates. Hence, the various causes of cancer and the ways through which it can be classified must be extensively understood [2].

Omics technologies have greatly aided cancer research by providing a systemic perspective on cancer through quantifiable insights into cancer genes and proteins and their associated pathways. Genomics allows for the distinction of the mutations and variations. Transcriptomics provides information on changes in gene expression disrupted in cancer cells, while proteomics examines alterations in protein expression, interactions, and functions vital to cancer cells. Metabolomics explains changes in metabolic pathways that facilitate rapid cell proliferation, while epigenomics reveals changes in DNA and histones that regulate gene expression without altering the sequence. Researchers can discover new biomarkers, unique therapeutic targets, and effective treatment plans for cancer by integrating these multi-omics approaches, thereby enhancing our understanding of its diagnosis, prognosis, and treatment [3].

Multi-omics refers to the analysis of several omics levels, including genomics, transcriptomics, proteomics, and metabolomics, to gain a comprehensive understanding of a biological system. Unlike specific omics approaches that focus on studying one type of molecular data in isolation, multi-omics integrates diverse layers of data to provide a holistic view of biological processes. In cancer research, adoption of this integrated approach can aid in understanding the molecular intricacies of complex cancer signaling pathways and their regulatory circuits, which contribute to cancer progression and heterogeneity. AI can significantly enhance the application of multi-omics by processing vast and complicated data to pinpoint unique characteristics in different cancer types [4]. Moreover, AI can identify new biomarkers that were undetected previously, estimate the efficiency of various treatments, and even increase the accuracy of cancer differentiation and specifying of treatment strategies.

Modern trends in cancer research include the utilization of omics technologies, which play a significant role in identifying and understanding the genomic, proteomic, and metabolomic differences within malignant cells. Omics must be integrated into the analysis of cancer subtypes because it enables researchers to combine distinct models of how cancer can affect the body. Handling omics data from cancer patients has become feasible due to technological advancements, enabling the management of vast amounts of information [5]. Researchers have used machine learning (ML) tools to analyze these data by utilizing unsupervised learning to identify subtypes based on multi-omics data [6]. Existing methods face several issues, which are as follows: the dimensionality issue, where datasets contain numerous variables and a limited number of cases for classification; and data integration, which exacerbates dimensionality problems when combining different data levels. Another problem related to model interpretability is the difficulty in understanding the predictions made using ML models. Previous studies have failed in presenting the influence of the selected features on the classification performance measures and efficiently work with relatively small and multi-class datasets. This research aims to enhance the reliability and clarity of cancer subtype categorization through a strong feature selection framework using multi-dimensional data.

To address these gaps, this research explicitly focusses on developing a structured approach to enhance cancer subtype classification by systematically integrating multi-omics data. The main goal is to eliminate the issues related to the implementation of classifiers when working with high-dimensional data and the introduction of multi-class classification. Apart from focusing on the performance of the model for better accuracy through sophisticated feature selection and clustering, there is more focus on the simplicity of the model for easy interpretation. Thus, as a result, we are able to give a more explicit description of how particular features affect classification results, addressing the gaps that other researchers have pointed out. The main contributions that can be derived from this work are as follows:

- Handling multi-omics dataset requires clarity and suitability for clustering algorithms. The multi-omics data present numerous challenges (i.e., outlier, noise, high dimensions, and complex relationships).
- A new feature selection method using cat swarm optimization (CSO) was proposed and applied in this study. The proposed method improves the classification accuracy up to 100% after feature selection. This model also improved the model interpretability in a biological context, offering comprehensive insights into underlying pathways. Validation was conducted using widely accepted performance criteria.
- This study also incorporated robust preprocessing of the multi-omics dataset to ensure the reliability and consistency of the integrated data used for analysis.

This paper is organized as follows: Section 2 focuses on a literature review, outlining previous work and describing limitations. Section 3 covers the methodology, detailing the dataset and the proposed feature selection and classification approach. Section 4 provides the results and discussion, detailing the key research points, such as the experimental setup, performance criteria, and a persuasive presentation of the results of the experimental study. Section 5 concludes the study and discusses potential future research directions.

2. Related Work

The application of ML and deep learning (DL) in cancer diagnosis using genomic data faces several significant challenges. The size and complexity of the multi-genomic data pose a major challenge. Training ML and DL models require a significant amount of data and computational resources, making them expensive and time-consuming to implement. The lack of strict regulation in the data collection and analysis process limits the comparability of results from different studies [7]. Considering these obstacles, ML and DL still have the potential to revolutionize cancer detection and management. These algorithms can be used to develop novel, non-invasive cancer diagnosis techniques and individualized treatment plans for patients. Numerous computational methodologies have been developed due to the exponential increase in computing capacity. Traditional medical procedures are expected to undergo significant changes with the introduction of big data and AI technologies. Advancements in technology and tools, such as DL and high-speed computing, have significantly enhanced the role of ML in computational biology. This fact can be explained by the successful results achieved in the field of biology, including its specific subdivisions. These advancements have significantly contributed to other fields, such as the identification of genetic variants, DNA methylation, and image analysis. The application of DL to diverse omics data continues to yield promising results [8].

However, despite the potential of current approaches, they are not without limitations. Numerous multi-omics techniques focus on a narrow range of data types, typically two or three, and often have limited applicability, targeting specific samples or patient groups. Furthermore, in unsupervised learning, interpreting results or classifier labels can be particularly challenging. The opportunities and threats posed by the influx of highly complex and unstructured omics data in healthcare research must be recognized. In light of these considerations, we have reviewed existing literature to map the current landscape [1]. Our study aims to highlight key contributions and unique features within this field, providing insights into the potential and limitations of current methodologies in optimizing the classification and clustering of cancer subtypes using specific approaches, such as CSO and support vector machines (SVMs).

Several studies have advanced cancer subtype classification and clustering using various ML and DL approaches. The work of [9] introduced DeepMO, which leverages encoded mRNA, DNA methylation, and CNV data from TCGA, showing the benefits of feature selection in classification. However, the aforementioned method lacks comparison with cutting-edge algorithms and clear analysis of the influence of feature selection, and the role of TCGA and potential data biases are not discussed. The work of [10] proposed a deep forest model with tiered data analysis, achieving high accuracy and efficiently handling asymmetric labeled data using the METABRIC dataset to address overfitting and ensemble diversity. Nevertheless, this model does not detail explicit strategies for overcoming overfitting and challenges related to high data dimensionality. The work of [7] improved lung cancer classification by integrating multi-omics data from liquid biopsies with ML techniques, achieving high AUC values using AdaBoost and showing enhanced performance. However, individual ML analyses showed limited discriminative performance, heavily relying on multi-omics data integration for improvements.

The work of [11] introduced a multi-stage feature selection system, tested four ML models, and applied the SHAP framework for interpretability, demonstrating improved performance and detailed feature impact explanations. Nevertheless, the small dataset size limited the results, especially in multiclass classification. In 2023, the BioSurv system [12] used ML and DL to identify biomarkers and predict cancer survival with high accuracy for BRCA and LUAD, utilizing statistical tests and the RSLBCSO algorithm for feature selection; however, it did not detail potential data biases and specific challenges in integrating multi-omics data. The work of [13] proposed MOCSS (shared and specific representation learning) for clustering and subtyping cancer using multi-omics data, emphasizing the importance of molecular subtyping and the necessity of sophisticated computational methods, although it was limited in discussing specific challenges and strategies in multi-omics data integration and handling data diversity. In their work, [14] demonstrated the modification of the Fruit Fly Optimization (FO) approach by incorporating a Levy Flight (LF) approach and achieved noteworthy outcomes with 93%. Nonetheless, the results presented in the study are uncertain and address overfitting and underfitting issues when split data is used for a training and testing data set that could affect the model's performance on new data.

Likewise, [15] presents the classification of hybrid DL models using LSTM and GRU classifiers with decision fusion. Despite achieving high reliability—98.0% for the decision fusion, 97.50% for GRU, and 97.0% for LSTM—the generality of the findings is in question due to the use of a small test sample that could influence the stability of the models when applied to other cases. The summary of these studies is provided in Table 1.

Table 1 *Summary of related works*

Author	Summary	Advantages	Limitations
Lin et al. [9]	DeepMO uses DNN and multi-omics datasets to classify breast cancer subtypes with higher accuracy and area under the curve (AUC) than previous approaches.	Greater prediction accuracy in multi-classification when utilizing multi-omics data as opposed to other techniques.	Understanding DeepMO's prediction process is challenging because of its closed-loop nature.
El-Nabawy et al. [10]	The deep forest model utilizes a cascade effect, combining the capabilities of DNN and ensemble models. Cascading deep forests learn class attributes.	The findings showed accuracy rates of 83.45% and 77.55% for 5 and 10 subtypes, respectively.	Difficulties with overfitting and ensemble variety arise due to the short sample size.
Kwon et al. [7]	The Investigation used three ML algorithms, namely, AdaBoost, MLP, and LR, to increase the accuracy of lung cancer classification.	The result revealed that incorporating multi-omics data considerably increased the accuracy of lung cancer categorization and diagnosis accuracy.	The study used a limited sample size of 92 lung cancer patients and 80 healthy individuals, which may restrict the generalizability of the findings.
4. Meshoul et al. [11]	Introduces a multi-stage feature selection system with two data integration strategies and looks into four ML models.	HYBRID raw experimental result accuracy: random forest (RF), extra trees, SVM, and XGBoost = 77.577, 78.066, 56.590, and 79.885, respectively, with AVG = 71.35063752. The model also obtained a high ROC_AUC using XGBoost = 92.438.	The small dataset size influenced deep learning model performance.
Dhillon et al. [12]	The BioSurv framework identifies cancer biomarkers and predicts survival rates.	A comparison of single and multi-omics is provided. BioSurv achieved high AUC values of 90.0% for BRCA and 87.0% for LUAD.	TNBC patients have an unfavorable predictive indicator.
Chen et al. [13].	MOCSS is a new technique for clustering multi-omics data and subtyping cancers.	MOCSS outperforms existing recent multi-omics clustering algorithms in terms of clustering performance, according to the experimental data.	The gene expression profiling criteria may complicate the categorization of the luminal A and B subtypes in BRCA, making it difficult to distinguish between the two subtypes.

Author	Summary	Advantages	Limitations
Anita Desiani. et. al. [14]	The main goal of the work is to improve the fruit fly optimization (FO) method by using a levy flight (LF) strategy.	Achieving 93.83% accuracy, 91.22% recall, and 96.53% specificity	The models may not generalize as well, especially when the data is divided into training and testing sets, meaning that eventual overfitting or underfitting might occur.
Othman, N.A. et al. [15]	Using two separate classifiers, a hybrid DL model is defined and developed to enable decision-making based on data from several sources, by the combination of multiple omics data sets.	LSTM and GRU are both used as classifiers. The decision fusion (LSTM + GRU) has a 98.0 % reliability, 97.50 % reliability from GRU, and 97.0 percentage reliability from LSTM.	The findings are not generalizable because tests were done on a limited dataset.

Despite significant advancements in the application of ML and DL for cancer subtype classification and clustering, several gaps remain in current research [1]. These gaps are identified to provide the areas for future study and enhancement, that is, standardization of data collection and analysis of omics data. The standardization of data collection, analysis, and interpretation of omics data is quite limited, resulting in variations in how the data are collected, analyzed, and interpreted. Consequently, the comparison of omics data across different studies provides challenges in replicating and generalizing the results. The applicability and results of these analyses are largely dependent on how databases, such as TCGA, are designed and on potential biases in data acquisition, which are often not comprehensively discussed. Feature selection has significant consequences for the model's performance. However, the effects of this mechanism might not be evident in analysis or documentation. This situation arises because feature selection is crucial with multi-omics data, and utilizing highly sophisticated methods becomes essential as dimensionality increases.

Moreover, numerous studies do not present a thorough analysis of performance on the newer versions of the new and state of the art algorithms, making it challenging to assess the advantages and disadvantages of these new models. However, overfitting remains a significant concern due to the high dimensionality of the genomic data and often small sample sizes, and the strategies to address these problems are not reciprocated. Current methods in multi-omics integration frequently encounter challenges in managing intricate multi-layered data and in preserving the interdependency of the various omics levels. Unsupervised learning methods may lead to the emergence of weak and non-operational conclusions because of the low use of label data. Unsupervised/label-supervised learning methods can be ineffective when used to make full use of data multi-omics integration due to simple data fusion and possible biases [1]. Accordingly, additional robust strategies are required to manage these problems. Future research can address these gaps and improve the effectiveness and application of ML and DL techniques for cancer subtype classification and clustering, which will enhance treatment and prognosis options in the long term.

3. Methodology

A noticeable increase in the incidence of cancer among patients has been observed at present. Numerous cancer cases have recently been reported across various clinical hospitals. The literature has proposed several ML algorithms for predicting cancer diseases of the same class types using trained and test data[16]. However, this area can still be further researched. This study focuses on investigating different types of cancer by analyzing, classifying, and processing multi-omics datasets within a CSO algorithm. This study developed new hybrid cancer detection schemes by utilizing multi-omics workload learning through an SVM. These schemes involve multiple layers, including the collection of clinical data through laboratory processes and tools (such as mammography, colonoscopy, and biopsy) at distributed omics-based clinics within the network. One of the most common methods in bioinformatics and computational biology to identify patterns, and group similar samples on the basis of multiple types of omics data (e.g., genomics, transcriptomics, and proteomics) is to cluster the samples through the use of K-means. To handle the dataset of multi-omics before doing the clustering, we worked pre-processing the data. This step is important to enhance the clustering results.

This work proposes a model based on K-means, CSO for feature selection, and support vector classifier (SVC) to classify cancer subtypes based on biomarker identification using a multi-omics dataset of patients. First, after pre-processing the dataset, the K-means algorithm extracted the features. Second, CSO selected the most significant features of columns for clustering by K-means. Third, each column is a subtype of cancer in the data frame and has a label given by K-means. The significant features are clustered into 10 clusters. Fourth, features are passed to SVC to classify patients as healthy or subtypes of cancer. Furthermore, the optimized feature markers are passed to K-means to cluster the data frame based on the WCSS fitness function of CSO. Finally, our model is evaluated using five performance parameters comprising accuracy, sensitivity, specificity, precision, and silhouette score for clustering. The complete workflow of the model is shown in Figure 1, which illustrates the process by which data are collected from molecular biology laboratories to obtain multi-omics data, which can be downloaded online. After the data is collected, the necessary steps are taken to process it to ensure the required quality. In this step, the data is unlabeled, and unsupervised learning algorithms apply clustering using the K-means algorithm to provide the label, as shown in the figure. Next, the algorithms are adapted to work with the CSO algorithm to select the most significant feature. Finally, the data is divided into training and test sets, and the SVM algorithm is used to obtain the results and is described as follows:

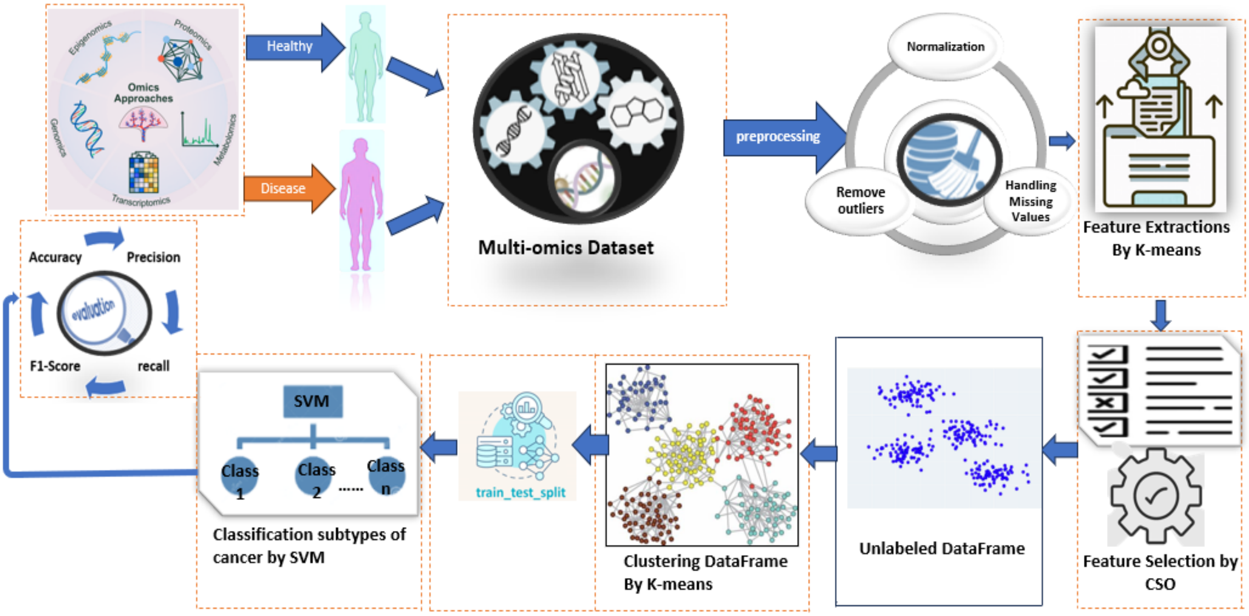


Fig. 1 *Methodology of the proposed model*

3.1 Dataset

The dataset can be downloaded from the Internet and used for academic purpose. Our gathered dataset offers a plethora of information about various types of cancer. This dataset can be considered a massive library, with each row representing a unique patient sample and every column providing valuable knowledge on the molecular origins of cancer. A thorough overview of the malignancies has been provided, including brain cancer (GBM), pancreatic cancer (PaCa), breast cancer (BrCa), colon cancer (CoCa), gastric cancer (GaCa), lung cancer, and hepatocellular carcinoma (HCC). It feels as if creating a detailed roadmap of such diseases' molecular landscape, directed at the cancer cells themselves. Means that there is an extensive data set that engulfs all aspects and details of the intricate molecular structure of different types of cancer. Such data allows for detailed information about the nature and development of cancer at the molecular level, which is so important to contemporary researchers. It is similar to making a detailed map that helps consider certain perks, such as special markers within the cancer cells, their workings, origins, and flaws. This kind of elaborated plan is very handy in identifying the specific proteins and other molecules that take part in the diseases and, hence, assisting in research, diagnosis, and treatment measures to be applied. The dataset size is 20244*69. Additional information about the dataset used in this study can be found in Table 2 and downloaded from (https://cfomics.ncRNAlab.org).

Table 2 *Summary details of the dataset*

Cancer type/healthy	Number of features	Number of gene name
1. Healthy	8	20244
2. Lung cancer	15	20244
3. Hepatocellular carcinoma	10	20244
4. Pancreatic cancer	7	20244
5. Breast cancer	4	20244
6. Colon cancer	4	20244
7. Gastric cancer	5	20244
8. Brain cancer	4	20244
9. Hepatitis B (HBV)	7	20244
10. Blood	4	20244
Total	68	Size = 20244*69

3.2 Preprocessing

The multi-omics dataset must be clearly preprocessed and adequately handled to be suitable for clustering algorithms. Multi-omics data are known to present several challenges, such as outliers, noise, high dimensionality, and complex relationships. Adding more information on data preprocessing and handling missing values would enhance the clarity and effectiveness of the data preparation process.

3.2.1 Remove Outliers

Outliers are a common occurrence in any procedure or field of study. Moreover, outliers can reduce statistical power by increasing data variability. Therefore, removing outliers may enhance the statistical significance of our findings. Several stages are involved in addressing outliers:

Removal of the outliers' steps

Step 1: Calculate Q1j and Q3j for each numeric column: Q1j = 25th percentile (x: j), Q3j = 75th percentile (x: j).

Step 2: Calculate the IQR for each numeric column:

$$\boldsymbol{IQR = Q3j - Q1j}. \tag{1}$$

Step 3: Determine the outlier condition for each numeric column:

a. Determine the lower threshold for outliers.

$$\boldsymbol{low\ hreshold = Q1j - 1.5 * IQRj}\ . \tag{2}$$

b. Identify and mark outliers in column j.

$$\boldsymbol{outliers_j = Xi,j \mid Xi,j < low_t hreshold}. \tag{3}$$

c. Remove outliers from column j, and update the dataset.

$$\boldsymbol{data[j] = data[j] - outliers_j}. \tag{4}$$

Step 4: Return the cleaned dataset without outliers.

3.2.2 Handling Missing Values

Real-world datasets frequently contain missing values for a number of reasons, including data gathering failures, equipment malfunctions, and simply because the information was not captured. Missing values are handled using specific procedures, such as imputation (replacing missing values with estimated ones), deletion, and considering missing data as a different category. Missing values can be replaced by the mean, median, mode, or sophisticated approaches, such as regression or K-nearest neighbor imputation. The goal of addressing missing values is to guarantee that the dataset is full and appropriate for analysis or modeling. In this study, we handled missing values by imputing the mean value of each column. We avoided missing data by calculating the mean value of each column by following these procedures. Handling missing values involves several steps:

Steps for handling missing values

Step 1: Recognize missing values. We used a program that became a dataset analyzer to find missing values in each column.

Step 2: Calculate the mean value. The aggregate mean value for each numeric column is obtained, excluding the null or missing values.

$$Meam_j = \frac{\sum_{i=1}^{m} x_{i,j}}{m}. \tag{5}$$

Step 3: Imputation. We replaced the missing values in each column with the mean value calculated for that column.

Step 4: Update the dataset. The dataset should be updated with imputed values to augment it.

3.2.3 Normalize

Each omics data type should undergo normalization as necessary (e.g., z-score normalization for gene expression data). Normalization converts the values of each feature in the dataset to a mean of zero and a standard deviation of one. In this study, we conducted normalization, which entails computing the data's mean (μ) and standard deviation (σ). This method is also known as z-score normalization or standardization. The aforementioned method normalizes the data to ensure that the mean is zero, and the standard deviation is one. This method maintains the distribution structure of the original data while centering it around zero and modifying the scale. This task is accomplished using the following formula: $X_{normalized.}$

$$\chi_{normalized} = \frac{\chi - \mu}{\sigma}, \tag{6}$$

where x is the original value of the feature, μ is the mean of the feature, σ is the standard deviation of the feature, and x is normalized.

3.3 Feature Extraction Using K-means Clustering

K-means clustering is a common technique that divides data into discrete groups, or clusters, based on feature similarities [17]. K-means may actually be used for feature extraction from numeric datasets [18] However, this approach is more typically utilized for clustering than explicit feature extraction. They distinguished the dataset into k different clusters inherent in the data through unsupervised learning. The centroids of the clusters may also be used as features to discover hidden features of the data. These attributes can be further used for further analysis or can be fed to another round of ML algorithms. The following is a representation of the formula for K-means feature extraction from the numerical datasets:

The dataset, denoted as X, contains n samples and m features, represented as X = {1, 2....}X = {X1, X2, Xn}, where Xi is an m-dimensional vector. When using K-means, the goal is to divide the dataset into k clusters in such a way that the WCSS is reduced to its lowest possible value. Let C = {c1, c2, ..., ck} be the centroids of the clusters. The objective function of K-means can be defined as follows:

$$Minimize\ C \sum_{k}^{i=1} \sum_{x \in si} \left\| x - ci \right\|^2. \tag{7}$$

Once the K-means method converged and effectively clustered all the data items present in a given dataset, the final centroids c1, c2,....., ck may be used in addition to the transformed dataset for further data analysis or fed into other ML algorithms. Feature extraction reveals the underlying patterns in the data that may not have been observed beforehand. Accordingly, we conducted feature extraction by K-means with the help of these procedures. Feature extraction using K-means requires the following steps:

Algorithm: Feature extraction by K-means clustering

```
Input: Multi-omics dataset, number of clusters K
Output: Cluster centroids with extracted features
Step 1: Apply K-means clustering.
    a. To begin the K-means clustering, determine the number of clusters.
    b. Apply the K-means algorithm to the multi-omics dataset.
    c. Acquire the means of centroids and cluster assignments.
Step 2: Feature extraction.
    a. The cluster centroid for each omics feature's average level is calculated for each omics feature
    using every sample for which it has been assigned.
    b. Take the average expression levels as the extracted features for each group.
Step 3: Output the centroids of the cluster groups, where the features are extracted, and all the samples
are assigned to a particular cluster.
End algorithm
```

3.4 Feature Selection Based on the CSO Algorithm

Feature selection is a fundamental data clustering technique accredited for decreasing feature dimensionality and enhancing the effectiveness of learning algorithms in practical applications [19]. In this study, we used the CSO to select the most significant features. The CSO is produced by observing cat activities and is made up of two sub-models: tracing and seeking modes, both of which simulate cat behaviors [20]. However, input data frequently contain duplicate or irrelevant characteristics, increasing processing time, lowering cluster quality, and potentially jeopardizing classification accuracy. Feature selection before classification is an efficient way to improve the efficiency and accuracy of a classifier [21]. We used CSO to choose the most significant features from the 68 columns. The CSO selected 60 attributes (Table 2) to facilitate clustering using the K-means approach. Thereafter, we applied the fitness function. The within-cluster sum of squares (WCSS) measures the cluster compactness. This metric represents the sum of the squared distances between each data point and the centroid of the cluster in which it belongs. WCSS reduction produces tight and coherent clusters.
WCSS calculation: After convergence, the WCSS is computed as follows:

a. Calculate each data point's squared Euclidean distance from the centroid of its allocated cluster.
b. Add the squared distances for all data points in each cluster.
c. Add the sums from all clusters to calculate the overall WCSS value.

$$WCSS = \sum C_i \in C \sum P_j \in C_i \ \|P_j - O_i\|^2. \tag{8}$$

After this step, we selected features from the multi-omics dataset.

Feature selection using CSO

```
Step 1. Initialization:
    *Determine the cardinality, denoted as N, representing the count of cats within the collective.
    *Specify the search scope for each feature dimension, delineating the spatial extent within which the
    exploration will be conducted.
    *Establish the number of clusters, denoted as K, pertinent to the K-means clustering algorithm.
    *Define the coefficient parameters pertinent to CSO, namely, inertial weight (ω), cognitive learning
    factor (c1), and social learning factor (c2).
    *Generate random spatial coordinates for each cat within the designated search range.
Step 2. Fitness function:
    *Evaluate the WCSS metric for each cluster utilizing the prevailing spatial distribution of cats.
    *Compute the global fitness metric, indicative of the collective WCSS, to determine each cat's fitness
    within the swarm.
```

```
Step 3. Cat location update:
    *Velocity computation:
        $V_i^d = ( V_i^d * \omega + c1 * rand( ) * ( P_i^d - X_i^d ) + c2 * rand( ) * ( G_d - X_i^d ) .$    (9)
    *Determine the velocity vector for each cat based on its current location and the collective movement dynamics.
    *Position update:
        $X_i^d = ( X_i^d + V_i^d ) .$    (10)
    *Update the spatial coordinates of each cat based on the computed velocity vector.
    *Boundary enforcement: Ensure that the updated cat positions lie within the predefined search range.
Step 4. Update of the optimal position:
    *Contrast the individual WCSS metric of each cat with its respective global best WCSS.
    *Revise the optimal WCSS value for each cat if the current evaluation yields an improvement.
    *Ascertain the global best position (G) by selecting the cat position that embodies the most optimal performance within the entire swarm.
Step 5. Iterative process:
    *Iterate through steps 2 to 4 until a termination criterion is met, such as reaching a predefined number of iterations or achieving a specified WCSS threshold.
Step 6. Feature selection:
    *Identify cats that exhibit superior WCSS metrics as indicative of proficient cluster representation.
    *Extract the corresponding feature vectors associated with these high-performing cats for utilization as selected features within the optimization framework.
End
```

Interpretation of WCSS: A low WCSS indicates tight clusters with data points near their centroids, indicating high intra-cluster cohesion and clear inter-cluster separation. Meanwhile, an increased WCSS indicates scattered clusters with data points widely distributed from their centroids, indicating weak intra-cluster cohesiveness and the possibility of overlap or near proximity of cluster centroids.

Limitations of WCSS: Ignoring cluster sizes: WCSS ignores the influence of cluster size, potentially biasing toward big clusters, which significantly contribute to the overall sum of squared distances. Disregarding data density: Given that WCSS does not take into account the density distribution within clusters, it may overlook changes in data concentration and cluster compactness.

3.5 Clustering the Data Frame: Through Adaptive K-means and CSO

K-means is one of the most popular algorithms for clustering among all the clustering algorithms. The k-means approach and its extensions are sensitive to initializations specified by a number of clusters based on prior knowledge. K-mean unsupervised clustering is essential for recognizing patterns in ML and data science [22]. This method determines cluster structures in a data collection based on the highest similarity inside a cluster and the most dissimilarity across clusters. The K-means technique is the most popular for clustering analysis in data sciences because it is simple, fast, versatile, and easy to implement [23].

General steps of the K-means algorithm.

```
Step 1: Choose K points as initial centroids.
Step 2: Repeat
Step 3:Form K clusters by assigning each point to its closest centroid.
Step 4: Recompute the centroid of each cluster.
Step 5: Repeat until the centroids do not change.
General CSO algorithm.
Randomly initialize cats.
WHILE (is terminal condition reached)
        Distribute cats to the seeking or tracing mode.
        FOR ($i$ = 0; $i$ < Number_Cat; $i$++)
                Measure fitness for cat$i$.
                        IF (cat$i$ in seeking mode) THEN
                                Search by seeking mode process.
                        ELSE
                                Search by tracing mode process.
                        END
        End FOR
End WHILE
Output optimal solution
```

We change how the initial centroids are selected and updated to incorporate CSO with the fundamental K-means method. CSO may be utilized to optimize centroid selection and refinement during the iterative process. We also defined the fitness function as the WCSS to help in guiding the optimization process. Addressing the incorporation of CSO with the K-means method entails several procedural steps:

Algorithm: Optimize cancer subtype classification using CSO and K-means clustering

```
Input: Multi-omics dataset
Output: Clustered data with optimized centroids
CSO parameters: fitness function: WCSS, number of clusters: determined using CSO and proven by elbow method (10 clusters).
K-means: Number of clusters 10 (determined by CSO).
Step 1: Initialization
        - Randomly initialize the positions of cats, where each cat represents a cluster centroid.
        - Evaluate each cat's fitness using WCSS.
Step 2: Update centroids using CSO
        - Repeat
          - Distribute cats into the seeking or tracing mode.
          - For each cat i ( i = 1 to Number_Cat):
            - Measure fitness for cat i.
            - If cat i is in the seeking mode:
              - Then update the position by the seeking mode process to reduce the WCSS.
            - Else (tracing mode):
              - Update the position by the tracing mode process to reduce the WCSS.
        - Until a terminal condition is reached (e.g., maximum iterations or convergence).
Step 3: K-means Clustering
        - Repeat
          - Assign each data point to the nearest centroid (cluster).
          - Recompute the centroids based on the allocated data points.
        - Until centroids do not change or a maximum number of iterations is reached.
      Output: Optimal cluster centroids and clustered data points
End algorithm.
```

3.6 Classification Subtype of Cancer

Cancer is a complex disease with numerous subtypes classified based on the various criteria, including the affected organ or tissue, histological characteristics, genetic mutations, and clinical behaviors [24]. SVM is an ML algorithm that finds a hyperplane in a high-dimensional space to optimally separate different classes of data points. In multiclass classification problems with more than two classes, SVC internally utilizes a strategy called "one-vs-rest" or "one-vs-one" to efficiently handle multiple classes. This hyperplane maximizes the margin between the classes, allowing for robust classification of new data. This approach uses the results of K-means clustering to uncover potential patterns in a cancer dataset. K-means clustering, a technique that groups data points into 10 distinct clusters, has the potential to uncover hidden patterns related to different forms of cancer. The feature selection process is improved by the CSO algorithm and improves the performance of the K-means clustering algorithm [25]. This process analyzes the entire dataset and identifies the most important features that effectively distinguish between different cancer subtypes. The features selected by the CSO are expected to contain the most important information for classification.

In this study, nonlinear SVMs were used in solution spaces where data could not be properly classified by an optimal hyperplane or in the original feature space. SVM addresses this limitation by using the kernel functions that transform the data into a high dimensional space, making it easy to achieve linear separability. The mathematical computations of the kernel function provide distances between these points, enabling the SVM to distinguish patterns and handle the nonlinearity of attributes. This capability allows the nonlinear SVM to disentangle complex data distributions, such as curved or circular decision surfaces, thus enhancing its generalizability across various problems in ML. These traits can be used to develop an ML model, such as the SVC, which can effectively distinguish between various types of cancer. Accordingly, the model may be able to distinguish between subtypes by utilizing K-means clustering and CSO feature selection, which may lead to the proper diagnosis and further treatment. The SVC equation can be expressed as follows:

$$f\,i(x) = sign\left(\sum_{j=1}^{Nsv} Y_j \alpha_j K(X_j, X) + b_i\right), \tag{11}$$

where: Nsv is the number of support vectors; Yj denotes the labels of the support vectors ($yj \in \{-1,1\}$ y ∈ {−1,1} for binary classification); αj represents the Lagrange multipliers associated with the support vectors, K(Xj, X) is the kernel function, which measures the similarity between training sample xj and input x; and bi is the bias term. After completing this procedure, we classified the cancer subtypes by using the multi-omics dataset subsequent to data clustering using the following steps:

Algorithm: SVM-based cancer subtype classification

```
Input: Clustered data from the CSO-K-means algorithm
Output: Classified cancer subtypes and performance metrics
Parameters SVC: Kernel: nonlinear (RBF), regularization parameter (C): 1.0, Gamma: 0.1
Step 1: Initialize SVC
        - Set up a nonlinear SVC.
Step 2: Train SVC
        - Train the SVC with the training data and cluster labels.
        - Use 80% and 70% of the dataset for training.
Step 3: Prediction
        - Apply the trained SVC to predict the cluster labels for the test data.
        - Use 20% and 30% of the dataset for testing.
Step 4: Evaluate performance
        - Apply relevant measures (e.g., accuracy, precision, recall, and F1-score) to assess the SVC's
      performance.
Step 5: Parameter tuning
        - Fine-tune the SVC's parameters using specific approaches, such as grid-search.
Step 6: Iterate and refine
        - Iterate over steps 1–5, tweaking the model as needed.
Step 7: Interpret results
        - Interpret the classification results, and assess the relevance of the detected clusters.
End algorithm
```

Fig. 1 shows the classification of our model after clustering the dataset using K-means and CSO, it was labeled. Thereafter, the data frame was divided into 70% for training the classifier and 30% for testing the SVC. Finally, the model was evaluated. Class imbalance problem: The issue we confronted, namely, an imbalance in the distribution of categories within the dataset, is a prevalent challenge in data analysis and ML methodologies. This discrepancy arises when the distribution of data samples among distinct classes is unequal, resulting in adverse effects on the efficacy of models, particularly those reliant on classification algorithms. We used the synthetic minority over-sampling technique (SMOTE) to address the problem of an imbalance in the number of classes. SMOTE is a mechanism used to oversample data between different classes by creating synthetic samples of a few classes [26]. The implementation of the SMOTE solution entails the following steps:

Algorithm: SMOTE-based data balancing

```
Input: Multi-omics dataset with imbalanced classes
Output: Balanced dataset and improved classification performance
Step 1: Identification of the underrepresented classes
    - Recognize classes characterized by a scarcity of samples within the dataset.
Step 2: Application of SMOTE
    -Utilize the SMOTE methodology to generate synthetic samples for the underrepresented
    categories by leveraging existing data.
Step 3: Data resampling
    -After implementing SMOTE, obtain a new set of data with a better balance between categories.
Step 4: Subsequent analysis
    -Apply the subsequent algorithms, such as SVC, to the resampled data post-SMOTE
    implementation.
    - Evaluate the performance of these algorithms using relevant metrics (e.g., accuracy, precision,
    recall, and F1-score).
End algorithm.
```

Adhering to these procedural steps enhances data balance, bolstering the performance of the classification-dependent algorithms.

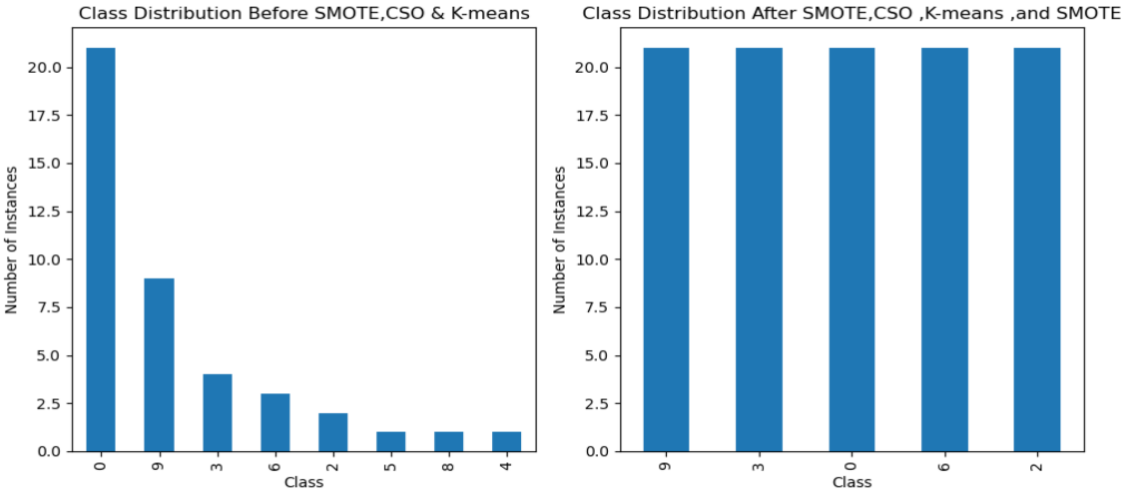


Fig. 2 *Before and after using SMOTE*

3.7 Evaluation of the Model

The model evaluation is a crucial aspect of ML and modeling algorithms, providing a tool to assess model quality and efficiency and support decision-making in different areas. A key statistic for evaluating the effectiveness of clustering is the silhouette score. This score indicates how each data point fits into the given cluster relative to the other clusters. Moreover, this score has an interval of −1 to 1. A high-scoring data point fits well inside its own cluster but not as well with distant clusters. Meanwhile, a low or poor score might suggest that a data item was mislabeled and placed in the wrong cluster. The silhouette score s(i) each sample i is calculated as follows:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}. \tag{12}$$

Accuracy refers to the proportion of correctly categorized positive and negative cases. However, accuracy should not be utilized alone to solve unbalanced situations because a high accuracy score can be achieved by categorizing all cases as the majority class.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}. \tag{13}$$

Precision is measured by dividing the number of properly predicted cases by the total number of positively predicted instances and the number of wrongly predicted cases. This metric is helpful when false positives are more concerning than false negatives. The precision metric is calculated using the following equation:

$$Precision = \frac{TP}{(TP + FP)}. \tag{14}$$

Recall (sensitivity) is calculated by dividing the number of successfully anticipated positive cases by the number of mistakenly predicted negative ones. This metric is helpful when false negatives are more concerning than false positives. The following equation is used to calculate recall:

$$recall = \frac{TP}{(TP + FP)}. \tag{15}$$

F1 score: Increasing the model's accuracy decreases recall and vice versa. This metric is the average precision and recall, providing a comprehensive view of these two measurements. The F1 score ranges from zero (lowest) to one (highest). The score accounts for false positives and negatives. F1 is more beneficial than accuracy in imbalanced class distributions. The F1 score is calculated using the following equation:

$$F1 = \frac{2rp}{(r + p)} = \frac{(2 * TP)}{(2 \times TP + FP + FN)} \quad . \tag{16}$$

4. Results and Discussions

This section thoroughly analyses the research findings, evaluates their efficacy, and presents the conclusion regarding the utilized methodology. The feature selection from the feature extraction, conducted using the K-means technology, resulted in distinct analysis and classification efforts. Nevertheless, utilizing CSO for feature selection might be crucial in enhancing the metrics used in analytical models because of the decrease in dimensions and the resulting improvement in categorical accuracy. Furthermore, this part will examine the outcomes of cancer subclassification and provide a comprehensive study to enhance the performance of our SVM classifier. This work improves the CSO feature selection approach and K-means clustering method to improve the accuracy of the dataset while also strengthening the model. Our technology has clearly demonstrated its exceptional proficiency in distinguishing various cancer subclasses by effectively incorporating diverse methodologies.

4.1 Feature Extraction Results

The K-means algorithm is utilized as a feature extractor to classify cancer subtypes from multi-omics data. This algorithm is an incredible tool to determine the meaningful and significant features of multi-omics data. In our study, K-means is used to identify key features, resulting in the extraction of 69 features from multi-omics cancer data. Using the K-means algorithm for feature extraction from multi-omics data is a crucial step toward extensively understanding and accurately classifying cancer. This approach allows us to efficiently use big data to analyze cancer subtypes, thereby effectively facilitating treatment. These extracted features were distributed as follows: The first feature is named after the genome used, but we excluded it because no computational method is available to process it. Accordingly, we have 68 remaining extracted features distributed among the cancer subtypes and viral hepatitis, as well as features for people with healthy samples free of any disease (healthy). Table 3 provides a summary of these extracted features.

Table 3 *Extracted features*

Cancer type/healthy	Number of extracted features
1. Healthy	8
2. Lung cancer	15
3. Hepatocellular carcinoma	10
4. Pancreatic cancer	7
5. Breast cancer	4
6. Colon cancer	4
7. Gastric cancer	5
8. Brain cancer	4
9.Hepatitis B (HBV)	7
10. Blood	4
Total	68

4.2 Feature Selection Results

The features extracted using K-means encompass the general features found in the dataset and may include features that might potentially confuse the classifier or lack adequate information. Accordingly, we identified and selected the most pertinent features to acquire a cancer subtype. The CSO technique identified and selected 60 more effective and informative features for the classification compared with the preliminary set of features (Table 3). This method may have outperformed other algorithms because it focuses on selecting a subset of the most significant features that influence the clustering, mitigating the risk of overfitting. Overfitting in ML occurs when the model learns to include noise or irrelevant information during training, resulting in worse performance when applied to fresh data. CSO can establish a correlation between the elements proven to be most beneficial, resulting in a universally applicable model.

The model can improve the clustering quality by using CSO in the context of K-means clustering. A high correlation coefficient value indicates improved clustering quality. The improvement observed in clustering quality after using CSO with K-means clustering indicates that the data have better-defined clusters due to using an appropriate basis, reducing dimensions, clarifying differences between values, and enhancing resistance to noise variable factors. Accordingly, this approach provides competitive and interpretable clustering, as reflected in the silhouette score. Table 4 illustrates the selected features.

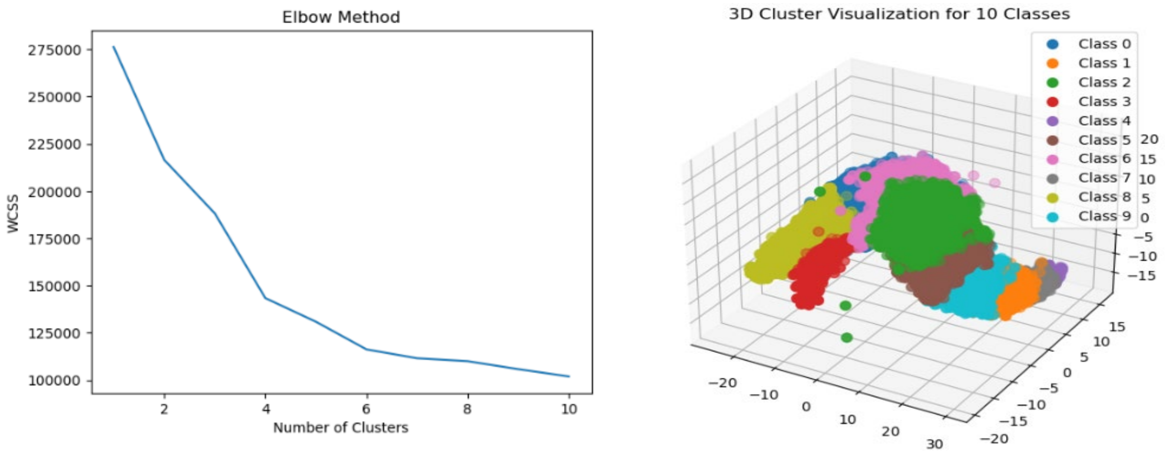
Table 4 *Result of the feature selection by CSO*

Cancer type/healthy	Number of features before CSO	Number of features after CSO
1. Healthy	8	8
2. Lung cancer	15	12
3. Hepatocellular carcinoma	10	8
4. Pancreatic cancer	7	6
5. Breast cancer	4	4
6. Colon cancer	4	3
7. Gastric cancer	5	4
8. Brain cancer	4	4
9. Hepatitis B, (HBV)	7	7
10. Blood	4	4
Total	68	60

The CSO identified the distinctive features that differentiate each group, resulting in clear and tangible borders. Consequently, the end product is enhanced in terms of quality and the capacity to understand the created batches. The CSO (feature selection algorithm) eliminates unnecessary or redundant information to enhance the accuracy of the classifier and optimize the outcomes of the multiple classification. Furthermore, the feature selection approach prioritizes essential features that will reduce the influence of noise on the clusters. K-means clustering with feature selection (CSO) bears joint responsibility for enhancing the overall quality of the model. The clustering procedure is improved by choosing the most pertinent features and segregating the group. This approach results in improved, easily understandable, and reliable clustering outcomes, making it suitable for various applications and facilitating data analysis. CSO's selection of key features resulted in an enhanced quality of clustering performed by the K-means algorithm. This improvement was evident through an increase in the silhouette score index, which serves as a measure of CSO's effectiveness in identifying features that are relevant to the disease.

4.3 Clustering Dataset by K-means

We utilized the K-means algorithm to perform clustering on the dataset, dividing it into 10 clusters and assigning a label to each subtype of cancer determined by K-means. The number of clusters was determined using CSO as the optimal number for these data (10). The elbow method is one of the data sciences and ML methods used for identifying the smallest possible number of clusters in a dataset while clustering it using the k-means algorithm. This method facilitates the choice of a reasonable number of clusters in the model through balancing and trading-off between a factor with a small sum of squared distance inside one cluster and achieving simplicity in model interpretation for visualizing the clustering of the data frame. The elbow method is presented in Fig. 3.



(a) **(b)**

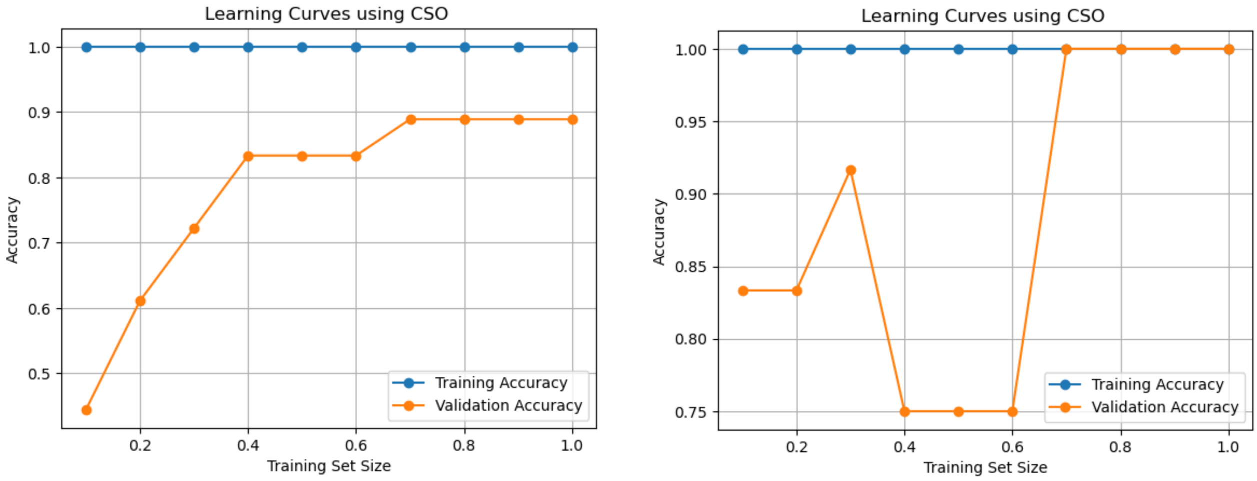
Fig. 3 *Clustering: (a) Elbow method; and (b) 3-D cluster visualization*

4.4 Classification of Cancer Subtype Using SVM

The utilization of a nonlinear SVM for the classification of cancer subtypes and following the assignment of labels to each disease type through the clustering conducted using K-means resulted in an initial accuracy of 81%. However, the accuracy of the SVM classifier significantly increased to 89% after the feature selection algorithm (CSO) has been integrated. This enhancement suggests that CSO effectively selected relevant and informative features pertaining to the disease, thereby reducing noise and improving the classifier's performance.

After utilizing the CSO, classification accuracy increased to 89%, suggesting enhanced classification performance. Our chosen CSO method demonstrated 89% classification accuracy. This result demonstrates that CSO effectively reduced misclassifications, thereby ensuring proper classifications of the different classes of cancer subtypes. The CSO type demonstrated high classification accuracy and minimized misclassifications by selecting the most significant features and improving the clustering quality. Thus, the SVM achieved a high score.

We adopted a data-splitting approach, allocating 70% for training and 30% for testing. When utilizing an alternate data splitting method of 20% for testing and 80% for training, the classifier's accuracy reached 100%. This result indicates the adaptability of our model, particularly when trained on large datasets containing a wide array of cancer types. Nevertheless, the scarcity or unavailability of multi-omics datasets encompassing a broad spectrum of cancer types poses a challenge to obtaining such data at present.



(a) **(b)**

Fig. 4 *Learning curves: (a) 30% testing; (b) 20% testing*

The experimental result (Table 5) shows that the CSO enhances our model better in terms of classification (accuracies, F1 score, precision, and recall) than k-means only when we used an 80% training and 20% testing split. We have high accuracy, equal to 100%.

Table 5 *Proposal result after CSO*

Method	Accuracy	F1-score	Precision	Recall
K-means and SVM	81%	73%	67%	81%
K-means, CSO, and SVM	100%	100%	100%	100%

Our proposed model with K-means, CSO, and SVM classes yielded 100% accuracies in more than one performance metric and outperformed the other models that displayed varied performance. Some models, such as SVM, Naive Bayes (NB), XGBoost, and RF, achieved favorable results but did not reach the level of accuracy attained by the proposed model. By contrast, some models, such as DL-TODA, demonstrated satisfactory performance. Although these models obtained decent scoring accuracy, a slight decline was observed in the other performance metrics. Table 6 provides a summary of these models.

Table 6 *Comparison of the model with other models*

Authors	Method	Accuracy (%)	F1-score (%)	Precision (%)	Recall (%)
Mohammed et al. [27]	SVM and RF	85.29–97.89	94.83–98.67	91.74–99.53	97.82–98.14
Cres et al. [28]	DL-TODA	97	80-98	91–98	76–97
Zhang et al. [29]	OmiEmbed	97.71	96.83	97.05	99.91

Gao et al. [30]	DeepCC	>90	N/A	N/A	>90
Modhukur et al. [31]	SVM, NB, XGBoost, and (RF) algorithms	99	98.3–100	94.4–100	96.4–100
Wang et al. [32]	RF, decision tree, and k-nearest neighbors	AUC 92.1			
Anita Desiani. et. al. [14]	LFFO-SVM	93.83	N/A	96.53	91.22
Othman, N.A. et al. [15]	Decision Fusion (LSTM + GRU)	98.0	N/A	99.0	99.20
Proposed model	K-means, CSO, and SVM	100	100	100	100

4.5 Discussions

The demonstration of the effectiveness of the CSO method in feature selection for tumor subtype classification provides proof that it positively influences the clustering quality and classifier accuracy. Our study utilized a CSO algorithm to identify the most crucial elements based on the fitness function. This approach resulted in a noteworthy enhancement in the classification accuracy. Initially, the accuracy rate was 81%. When the CSO was adopted, the accuracy rate increased to 100%. In this study, the approach was to prioritize key aspects of each type of cancer to enhance the performance of the classifier. This approach significantly enhances the classifier's accuracy levels. The CSO played a crucial role in determining the optimal number of clusters. We verified the efficacy of this number, as determined by the CSO, using the elbow method. This method further validated the efficacy of the CSO by selecting 10 clusters for data clustering.

The silhouette score was enhanced by using CSO, indicating a minimal overlap and strong distinction across the clusters. This research showed that CSO had a crucial role in accurately selecting features, which was the reason for the efficacy of the subgroup classification and type identification tasks in cancer detection. The combination of CSO with K-means and SVM classification was affected by the resolution of the frequent problems associated with imbalanced data clustering that arise after K-means clustering. The two main tasks were taken on which were the balancing the data by applying the SMOTE approach, resulting in increased classifier performance. Moreover, nonlinear SVM is a major and strong classifier in the process of classifying data. This mechanism is particularly inclined toward high-dimensional datasets and is capable of segmenting and analyzing small patterns. This research presented the significance of using CSO, and choosing integrated techniques, such as computer-based algorithms for optimal feature selection and trustworthy classifiers, is crucial in accurately training the classifiers.

The determination of the number of clusters and feature selection were aimed at obtaining highly informative clusters and using these features for the classification of cancer subtypes. One of the first algorithms used by the CSO method was designed to determine the appropriate number of clusters for a given dataset. After clustering the data frames, the shape of the clusters became clear, revealing a total of 10 clusters. The option was based on the fitness function, which aimed to minimize the intra-cluster sum of square (WCSS). This approach facilitated the creation of fairly autonomous and well-defined clusters that reflect the underlying data structures. The elbow method demonstrated that the optimal number of clusters was selected by CSO. Each cluster representing a certain form of cancer had attributes selected to capture unique features of each subtype. The datasets were then divided into separate sections for training and testing. The chosen features encompassed a range of total features, including the number of clusters and the overall count of features. The stringent selection process ensured the inclusion of crucial cancer subtype traits, resulting in improved efficacy and accuracy in the subsequent classification. The process used to define the number of clusters and select features demonstrated a meticulous approach to achieve optimal clustering and relevant features in cancer subtype classification.

An important and potentially controversial issue that needs to be highlighted is the variability in the selection of the number of features by the CSO. Whether these features are selected for clustering or classification concerning the above features, it should be noted that the selected features in this study are for both clustering and classification. First, they employ the features to cluster in order to find numerous subtypes of cancer within the database. Later, these features are used in the classification that enables precise division of new samples into the revealed cancer subtypes. This twofold approach in terms of defining subtypes guarantees that the types which are specified with a help of clustering are as significant as the types which are defined for classification purposes. In classification, a high level of accuracy is achieved when few features are present, which is typical behavior for the classifier. The classifier will distinguish between these samples based on a limited number of features, thereby minimizing confusion. However, if these features are intentionally chosen to be adaptable with a clustering algorithm, then the program will cluster the data and assign a label to each feature, indicating an illness. The greater the number of these features, the more information-dense the results of this cluster become, providing additional data about the disease. This phenomenon forms the basis for addressing the generalization of the results and enables the interpretability and analysis of the classifier. The selection of features is significantly influenced by the fitness function, known as WCSS, which aims to identify the optimal quantity of these traits. The features establish coherence and interconnectedness among the squares within each cluster, indicating their interconnection with one another.

The clusters are distinct from one another, indicating the presence of well-defined borders between them. The findings demonstrated a positive correlation between the number of features and the quality of clusters, as evidenced by the silhouette score. Each feature corresponds to a disease with over 20,000 genomes. In particular, each feature has about 20,000 values derived from the analytical results obtained in these datasets. The significant improvement in accuracy (from 81% to 100%) highlights the effectiveness of the CSO method in feature selection and cluster optimization. The silhouette score, which indicates the cluster separation, was significantly improved by CSO, demonstrating a minimal overlap and strong discrimination between clusters. The combination of CSO, K-means, and SVM effectively addressed the problems of imbalanced data and improved the classification performance. The integration of SMOTE also contributed to data balancing, further improving the performance of the classifier. The ability of the nonlinear SVM to handle high-dimensional data and detect small patterns was critical in achieving high levels of accuracy.

The findings presented in our paper show that cancer subtype classification has been improved at a significant level, which would have practical applications in personalized medicine. SO, through maximizing the accuracy and stability of the classifier models, the proposed method could allow better diagnosis and interventions for the cancer patients due to the more precise classification of the disease. However, there are potential limitations, such as the time and computational resources required to integrate and process large-scale multi-omics data and how well the model generalizes to other datasets. Moreover, although the presented method seems to be effective in a well-organized experimental environment, more validation with real-world patients with different demographics and in larger groups is required.

This section delves deeply into the reasons for the remarkable effectiveness of the CSO algorithm as compared to traditional methods. The principal attributes of CSO are based on nature-inspired processes, and optimization has been enhanced through the methods employed in the given framework. Sophisticated Feature Selection: Traditional optimization procedures also face numerous difficulties, especially when working with a large number of variables since they remain stuck in the local minimum and do not discover other reasonably beneficial regions. These are limitations of the previous methods only. As CSO 's tracing and searching mechanisms are used to broaden the solution space and avoid convergence prematurely, thereby enhancing the feature selection and getting high classification accuracy. Cluster Coherence: Hence, the application of CSO for feature selection before applying K-Means yields better accuracy in forming well-correlated clusters and also forms compact clusters. CSO not only serves the purpose of fine-tuning the feature set in order to obtain more meaningful clusters but also performs a profound enhancement of the clustering procedure and results, which are sometimes virtually unattainable. Handling Complex Multi-Omics Data: Multi-omics datasets involve relations that are complex and many-layered, which makes pattern typology challenging to accomplish. Due to feature management and optimisation in CSO, it is easier to find relationships within the data, hence improving the overall clustering and classification result.

5. Conclusion

This work illustrates the theoretical and practical ramifications of applying cluster optimization and feature selection for cancer subtype classification using the CSO approach. In theory, this study adds to the body of knowledge by illustrating how CSO might improve the precision and quality of tasks involving clustering and classification. Imbalanced data and high-dimensional datasets pose issues that can be addressed by combining CSO with K-means, SMOTE, and SVM classifiers, resulting in significantly improved classification performance. Enhanced classifier efficiency is one of the numerous practical benefits of this work, along with several others. In addition to improving accuracy, the suggested approach effectively shields the classification process from artificial disruptions. Meanwhile, the SMOTE balances datasets and uses nonlinear SVM to handle multiclassification errors. This intelligent classifier can perceive complex data characteristic. Specifically, this mechanism can be applied and further developed for clinical applications to process data related to cancer patients, enhancing possibilities for subtype analysis in diagnosis and therapy strategies. The initial use of the proposed CSO and SVM model achieved an accuracy of approximately 81%. After integrating the feature selection algorithm, the accuracy significantly improved to 100%, clearly outperforming existing models. Nonetheless, this work is not without limitations. Despite the existing advantages, few disadvantages must be noted. Selecting features for CSO requires significant consideration, especially if those features are utilized for either clustering or classification. A significant number of features provides substantial information about the clustering. Nevertheless, few features typically result in accurate classification results. The integration of SVM with SMOTE or CSO and K-means can consume a substantial amount of computational time, which may not be advantageous in technical environments with limited computing power. Before this approach can be potentially applied, it must be validated for different types of cancer and datasets. These areas should be addressed in subsequent research to improve the efficacy and relevance of CSO in cancer subtype classification and other high-dimensional data clustering, such as validation on diverse datasets. Future work should replicate the method on a large sample of datasets across different cancer types to substantiate its effectiveness. Additionally, exploring the use of additional datasets and improving the proposed method through the integration of more advanced or alternative models could further validate and enhance the performance of our approach. Optimization techniques: Another area of interest could be looking at other optimization methodologies or models that can be combined to potentially decrease the computational cost and/or increase the accuracy. Quantum CSO: The literature review on the application of quantum CSO (QCSO) helps in enhancing the understanding of feature selection and clustering performance. QCSO takes advantage of quantum computing techniques that promise super polynomial speedup and improved solution quality. Some ideas for future work would be to conduct an analysis of the feasibility and relevance of the use of QCSO in high-dimensional space data, particularly in the context of carcinoma subtype categorization.

Acknowledgements

The authors would like to show their appreciation to the Department of Artificial Intelligence, College of Computer Science and Information Technology, University of Anbar.

Conflict of interest

The authors declare no conflict of interest regarding the publication of the paper.

Author contribution

The authors' contributions to this work are as follows: Ali Mahmoud Ali and Mazin Abed Mohammed ***designed and conceptualized the study****. Mazin Abed Mohammed conducted* ***data collection****. Mazin Abed Mohammed and Ali Mahmoud Ali* ***analyzed and explained the results****. Ali Mahmoud Ali and Mazin Abed Mohammed* ***write drafted and final versions of the manuscript****. After reviewing the results, all authors approved the final draft of the manuscript.*

References

[1] Ali, A. M., & Mohammed, M. A. (2023). A comprehensive review of artificial intelligence approaches in omics data processing: Evaluating progress and challenges. International Journal of Mathematics, Statistics, and Computer Science, 2, 114–167. https://doi.org/10.59543/ijmscs.v2i.8703

[2] Mohammed, M. A., Lakhan, A., Abdulkareem, K. H., & Garcia-Zapirain, B. (2023). Federated auto-encoder and XGBoost schemes for multi-omics cancer detection in distributed fog computing paradigm. *Chemometrics and Intelligent Laboratory Systems*, *241*, 104932. https://doi.org/10.1016/j.chemolab.2023.104932

[3] Fan, Y., et al. (2022). Integrated multi-omics analysis model to identify biomarkers associated with prognosis of breast cancer. Frontiers in Oncology, 12. https://doi.org/10.3389/fonc.2022.899900

[4] Sarkar, J. P., Saha, I., Sarkar, A., & Maulik, U. (2021). Machine learning integrated ensemble of feature selection methods followed by survival analysis for predicting breast cancer subtype specific miRNA biomarkers. Computers in Biology and Medicine, 131. https://doi.org/10.1016/j.compbiomed.2021.104244

[5] Mohammed, M. A., Lakhan, A., Abdulkareem, K. H., & Garcia-Zapirain, B. (2023). A hybrid cancer prediction based on multi-omics data and reinforcement learning state action reward state action (SARSA). Computers in Biology and Medicine, 154. https://doi.org/10.1016/j.compbiomed.2023.106617.

[6] Momeni, Z., Hassanzadeh, E., Saniee Abadeh, M., & Bellazzi, R. (2020). A survey on single and multi-omics data mining methods in cancer data classification. Journal of Biomedical Informatics. Academic Press Inc. https://doi.org/10.1016/j.jbi.2020.103466

[7] Kwon, H. J., et al. (2023). Enhancing lung cancer classification through integration of liquid biopsy multi-omics data with machine learning techniques. Cancers (Basel), 15(18). https://doi.org/10.3390/cancers15184556

[8] Mohammed, M. A., Abdulkareem, K. H., Dinar, A. M., & Garcia-Zapirain, B. (2023). Rise of deep learning clinical applications and challenges in omics data: A systematic review. Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/diagnostics13040664

[9] Lin, Y., Zhang, W., Cao, H., Li, G., & Du, W. (2020). Classifying breast cancer subtypes using deep neural networks based on multi-omics data. Genes (Basel), 11(8), 1–18. https://doi.org/10.3390/genes11080888

[10] El-Nabawy, A., Belal, N. A., & El-Bendary, N. (2021). A cascade deep forest model for breast cancer subtype classification using multi-omics data. Mathematics, 9(13). https://doi.org/10.3390/math9131574

[11] Meshoul, S., Batouche, M., Shaiba, H., & AlBinali, S. (2022). Explainable multi-class classification based on integrative feature selection for breast cancer subtyping. Mathematics, 10(22). https://doi.org/10.3390/math10224271

[12] Dhillon, A., Singh, A., & Bhalla, V. K. (2023). Biomarker identification and cancer survival prediction using random spatial local best cat swarm and Bayesian optimized DNN. Applied Soft Computing, 146. https://doi.org/10.1016/j.asoc.2023.110649

[13] Chen, Y., et al. (2023). MOCSS: Multi-omics data clustering and cancer subtyping via shared and specific representation learning. iScience, 26(8). https://doi.org/10.1016/j.isci.2023.107378

[14] Desiani, A. A., Lestari, A. T., Al-Ariq, M., Amran, A., & Andriani, Y. (2022). Comparison of support vector machine and k-nearest neighbors in breast cancer classification. Pattimura International Journal of Mathematics (PIJMath), 1(1), 33–42. https://doi.org/10.30598/pijmathvol1iss1pp33-42

[15] Othman, N. A., Abdel-Fattah, M. A., & Ali, A. T. (2023). A hybrid deep learning framework with decision-level fusion for breast cancer survival prediction. Big Data and Cognitive Computing, 7(1). https://doi.org/10.3390/bdcc7010050

[16] Badarudin, P. M., Ghazali, R., Alahdal, A., Alduais, N. A. M., & Mostafa, S. A. (2021). Classification of breast cancer patients using neural network technique. Journal of Soft Computing and Data Mining, 2(1), 13–19. https://doi.org/10.30880/jscdm.2021.02.01.002

[17] Mohammed, A. Z., & George, L. E. (2023). Region of interest extraction using K-means and edge detection for DEXA images. Al-Salam Journal for Engineering and Technology, 2(2), 48–53. https://doi.org/10.55145/ajest.2023.02.02.006

[18] Mandal, S. K., Parida, K., Kumar Mandal, S. S., Das, S. S., & Ranjan Tripathy, A. (2011). Feature extraction using K-means clustering: An approach & implementation. Retrieved from https://www.researchgate.net/publication/334227319

[19] Veysel Aslantaş, P. D., Ahmet Nusret Toprak, D., Khorsheed, F. H., & Khalaf, B. A. (2020). Wrapper feature selection approach based on binary firefly algorithm for spam e-mail filtering. Journal of Soft Computing and Data Mining, 1(2), 44–52. https://doi.org/10.30880/jscdm.2020.01.02.005

[20] Chu, S. C., Tsai, P. W., & Pan, J. S. (2006). Cat swarm optimization. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (pp. 854–858). Springer Verlag. https://doi.org/10.1007/11801603_94

[21] Ahmed, A. M., & Abdulazeez, A. M. (2021). Examining swarm intelligence-based feature selection for multi-label classification. Journal of Soft Computing and Data Mining, 2(2), 63–73. https://doi.org/10.30880/jscdm.2021.02.02.006

[22] Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. IEEE Access, 8, 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796

[23] Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. MDPI AG. https://doi.org/10.3390/electronics9081295

[24] Houssein, E. H., Abdelkareem, D. A., Hu, G., Hameed, M. A., Ibrahim, I. A., & Younan, M. (2024). An effective multiclass skin cancer classification approach based on deep convolutional neural network. Cluster Computing. https://doi.org/10.1007/s10586-024-04540-1

[25] Ali, A. M., & Mohammed, M. A. (2024). Enhanced cancer subclassification using multi-omics clustering and quantum cat swarm optimization. Iraqi Journal for Computer Science and Mathematics, 5(3), 552–582. https://doi.org/10.52866/ijcsm.2024.05.03.035

[26] Kanellopoulos, D., Pintelas, P. E., Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (n.d.). Handling imbalanced datasets: A review. Retrieved from https://www.researchgate.net/publication/228084509

[27] Mohammed, A., Biegert, G., Adamec, J., & Helikar, T. (2017). Identification of potential tissue-specific cancer biomarkers and development of cancer versus normal genomic classifiers. Oncotarget, 8(49), 85692–85715. https://doi.org/10.18632/oncotarget.21127

[28] Cres, C. M., Tritt, A., Bouchard, K. E., & Petrick, J. (2021). Comparative analysis of multiclass support vector machine and deep learning classification of cancer subtypes. Journal of Biomedical Informatics, 117. https://doi.org/10.1016/j.jbi.2021.103708

[29] Zhao, L., Chen, Y., Zhang, X., & Xie, W. (2022). Hybrid feature selection method for cancer data classification based on cat swarm optimization and quantum particle swarm optimization. Journal of Computational Chemistry, 43(11), 1153–1163. https://doi.org/10.1002/jcc.26878

[30] Gao, F., et al. (2019). DeepCC: A novel deep learning-based framework for cancer molecular subtype classification. Oncogenesis, 8(9). https://doi.org/10.1038/s41389-019-0157-8

[31] Modhukur, V., et al. (2021). Machine learning approaches to classify primary and metastatic cancers using tissue of origin-based DNA methylation profiles. Cancers (Basel), 13(15). https://doi.org/10.3390/cancers13153768

[32] Wang, M., et al. (2022). Identification of cancer-associated fibroblast subtype of triple-negative breast cancer. Journal of Oncology, 2022. https://doi.org/10.1155/2022/6452636