Iraqi Journal for Computer Science and Mathematics

Journal Homepage: http://journal.esj.edu.iq/index.php/IJCM
e-ISSN: 2788-7421 p-ISSN: 2958-0544

Enhanced Cancer Subclassification Using Multi-Omics Clustering and Quantum Cat Swarm Optimization

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: Mazin Abed Mohammed

DOI: https://doi.org/10.52866/ijcsm.2024.05.03.035
Received March 2024; Accepted May 2024; Available online August 2024

ABSTRACT: Integrating multiple omics data can significantly improve the accuracy of cancer subclassification, a challenging task due to the high dimensionality and limited sample sizes. The integration of these data sets can enhance model performance. This study addresses these challenges by employing Quantum Cat Swarm Optimization (QCSO) for feature selection, along with K-means clustering and Support Vector Machine (SVM) for classification. Using QCSO, the most significant features were identified, resulting in an increase in accuracy from 81% to 100%. Performance was evaluated using accuracy, F1-score, precision, recall, ROC, and the silhouette metric, all of which confirmed the effectiveness of the feature selection approach. Additionally, this method enhances the classification of samples while making the models more interpretable, providing better insights into the molecular mechanisms of cancer. This work contributes to advancing knowledge in the field of cancer research and biology in general.

Keywords: Cancer, Omics, Multi-omics, Quantum Cat Swarm Optimization, Cancer Subtype, K-Means, cat swarm optimization.

1. INTRODUCTION

Cancer remains one of the biggest challenges in medical science due to its complexity and severity [1]. Categorizing cancer is crucial because identifying the specific type of cancer is essential for determining the appropriate treatment, which ultimately enhances the quality of life for patients [2]. Previous studies on cancer subtype categorization have relied on clinical and histopathological characteristics, but these approaches are often insufficient to capture the molecular heterogeneity of cancer [3]. With advances in high-throughput technologies, multi-omics data—encompassing genomics, transcriptomics, proteomics, and metabolomics—has become more accessible. The integration of multi-omics data offers a more accurate and comprehensive understanding of cancer subtypes [4]. However, the high dimensionality and large volume of data in these matrices present significant challenges for analysis and interpretation.

Multi-omics data is collected through various methods and tools that analyze multiple layers of molecular interactions within a biological system. These layers include genomics, proteomics, and metabolomics. Integrating data from these diverse sources allows for the reconstruction of a more detailed picture of the biological system, revealing connections and finely-tuned regulatory mechanisms that might not be identified when using methods that focus on a single level of organization. There are several types of omics studies, including single-omics studies, which focus on analyzing data from a single type of molecular layer. Examples include: Genomics that focuses on genetic differences and the patterns of gene activity or dormancy in endocrine diseases. Proteomics are examines protein levels, post-translational modifications, and protein-protein interactions and Metabolomics are provides information on metabolites and small molecules.

While single-omics methods offer comprehensive information about specific layers of biological organization, they can be limited in providing a complete understanding of interactions within the system. These methods often focus on one omics layer, which can constrain our understanding of processes in living organisms. Multi-omics refers to the combined analysis of several layers of data from different omics platforms, providing a more systemic approach. By integrating genomics, proteomics, and metabolomics, multi-omics studies are report multiple genomic, proteomic, and metabolomic interactions. Provide a holistic view of biological organisms and accurately represent regulatory and disease processes. Although individual studies may focus on one omics layer, adopting a multi-omics approach enables researchers to gain a much deeper understanding of biological processes and achieve more accurate results. This integration makes it possible to delineate the complex mechanisms underlying diseases, leading to the development of more appropriate interventional and therapeutic strategies. Compared to single-omics analysis, multi-omics analysis is more advantageous due to the integrated information it generates from different facets of biological systems. Given that

interactions in biological systems are simultaneous and feed-forward, multi-omics data is far more informative than single-omics data, as it captures a holistic picture of disease characterization. This, in turn, enables better diagnostic accuracy and prognosis by offering more than just a single diagnostic or therapeutic target.

General challenges associated with omics include data characteristics such as complexity, heterogeneity, high dimensionality, and issues related to data quality and gaps. Due to the complexity and volume of omics data, high-level computational analysis is required, necessitating fast machines and advanced algorithms. Combining multi-omics data is methodologically challenging due to differences in data structures, the need for varying normalization approaches, and the complex task of making the integration biologically meaningful and interpretable [4]. Multiple hypothesis testing and managing false discoveries present significant challenges for statistical methods, while variability in protocols and analysis types further complicates the analytical process. Ethical and privacy concerns are also paramount, as genetic and health information is often considered personal and confidential. Applying the findings of genetic research in clinical settings involves navigating legal and practical obstacles while demonstrating efficacy. These challenges underscore the need for integrated collaboration in omics research, investment in computer science and statistical methods, and the escalation of standardization efforts [5].
Feature selection (FS) is crucial in addressing these challenges, as it identifies which features have the greatest impact on classification accuracy [7]. However, conventional feature selection techniques often prove inadequate when dealing with the complexity and bias inherent in multi-omics data. This has led to the development of more sophisticated and efficient feature selection techniques that can identify the most important features in a dataset. In this regard, original algorithms based on swarm intelligence have shown great promise [8], [9]. Among these, the Cat Swarm Optimization (CSO) algorithm is particularly preferred for its exploratory characteristics and potential for finding global optimal solutions [10]. In recent years, we have introduced Quantum Cat Swarm Optimization (QCSO), based on the principles of quantum computing, to further enhance the optimization process [11]. The existing methods face numerous issues, which are as follows:

1. Dimensionality Issue: There are many variables in the datasets; however, the number of cases to classify them is rather small. This is made worse when working with levels of data that aggregate different dimensions, thus worsening the aspect of dimensionality.
2. Data Integration: Although multi-omics data integration adds extra dimensions to the problem, raising the concern of dimensionality and, consequently, overfitting, the models' performance may drop.
3. Model Interpretability: It is, on some occasions, difficult to explain the outcomes given by the machine learning models. This is more so when dealing with something biological, where an explanation of the processes involved is so important.
4. Influence of Selected Features: As mentioned in previous investigations, the impact of the chosen characteristics on the measures of classification performance has not been described comprehensively. This splits the work naturally and makes it hard to evaluate the significance and effectiveness of the specific attributes.
5. Efficiency with Small and Multi-Class Datasets: Most of the present models are not very proficient with regard to relatively small and multiple-class datasets, which hampers their usage in some research environments.

Given the challenges mentioned, the objective of this study is to enhance both the precision and interpretability of cancer subtype classification by preventing information loss and adopting a more effective feature selection approach. This would also lay the groundwork for future biological research and the advancement of cancerology. The aim of this research is to make cancer subtype classification more reliable and concise by employing a sensible feature selection technique for multi-dimensional data. The purpose of this work is to apply Quantum Cat Swarm Optimization (QCSO) for feature selection in the context of multi-omics data analysis for cancer subtyping. The subsequent section presents the incorporation of QCSO with K-Means clustering and nonlinear Support Vector Machine (SVM) classification to improve cancer subtype identification. The performance of the proposed approach is measured using accuracy, F1 score, precision, recall, and silhouette score to evaluate the quality of clustering. Initial findings reveal that our proposed model significantly enhances classification performance, with accuracy increasing from 81% to 100% following the feature selection steps. These results indicate that QCSO has the potential to improve cancer subtype identification, which could inspire future biological investigations and cancer research based on this study. The theoretical framework, methodology, and detailed findings of the proposed system will be discussed in the subsequent sections, along with an exploration of its efficacy and implementation prospects in the field of AI studies. The main contributions of this work are as follows:

- Handling Multi-omics: A multi-omics dataset requires clarity and suitability for clustering algorithms. Multi-omics data present numerous challenges, including outliers, noise, high dimensionality, and complex relationships.

- Innovative Feature Selection: We introduced a novel feature selection technique, Quantum Cat Swarm Optimization (QCSO), which significantly improved classification accuracy to 100% by effectively selecting the most relevant features.
- Classification: The primary objective is to accurately classify cancer subtypes using an enhanced methodology that combines QCSO for feature selection and SVM for classification.
- Improved Model Interpretability: The method enhances interpretability by providing meaningful insights into biological pathways and processes.

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 QCSO for 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

A few key obstacles that are discussed below relate to the application of DL in conjunction with genomic data and machine learning for cancer detection. The availability and integration of multi-genomic data is seen as a significant problem because of the data's significantly increased amount and complexity. The ML and DL models require a lot of data and computation during training, which makes the application expensive and time-consuming. Altogether, it is critical to acknowledge the advancements in the application of; ML and DL for the cancer diagnosis using the Genomic data. Therefore, this study would seek to meet these challenges by adopting the following strategy: The use of QCSO for feature selection combined with K-Means clustering and SVM classification in the cancer subtype classification. We have examined the literature to map the present landscape in light of these factors [1]. In order to shed light on the potential and constraints of the approaches being used today for the optimal classification and clustering of cancer subtypes using techniques such as SVM and CSO, our study intends to highlight significant contributions and distinctive features within this field.

A number of studies have enhanced the idea of cancer subtyping and grouping deploying assorted ML & DL techniques. According study [9] proposed DeepMO in 2020, which used encoded mRNA, DNA methylation, and CNV data from TCGA and proved that feature selection can greatly improve the classification function. However, it does not present comparisons with state-of-the-art algorithms and does not explain how feature selection has been affected. Also, the contribution of TCGA and possible data bias were limited in the consideration. According to [10] the researcher's applied Deep Forest with tiered data analysis in 2021 with high accuracy and a power of dealing asymmetric labeled data with the help of METABRIC dataset. This approach was expected to solve overfitting and ensemble diversity issues, but it failed to explain the strategies of handling overfitting and the problem arising from high dimensionality of data. Kwon et al. by reference [2] in 2023 study enhanced lung cancer classification through incorporating multi-omics characters from liquid biopsies and having high AUC values by using the ML algorithm called AdaBoost. The study essentially demonstrated improvement in the performance of the model, but it appeared to have low discriminative capability in single-ML analysis and heavily relying on multi-omics data integration.

In their proposed research, [11] proposed a multi-stage feature selection system and compared four different types of ML models and provided better performance and SHAP framework for Model interpretability. However, it should be noted that the authors used limited data set Sizes and therefore, the results were not so impressive particularly in multiclass problems. In 2023, According study [12] Dhillon et al proposed the BioSurv system where they applied ML and DL to analyze biomarkers and estimate HCC survival rate with precision for BRCA and LUAD diseases. The particular data analysis approach included the use of statistical tests and RSLBCSO for feature selection, and it did not discuss the presence of biases in the data and the more significant issues related to the integration of multi-omics datasets. Lastly, Chen et al. By research [13] introduced MOCSS regarding clustering and subtyping cancer through the multi-omics data; they stressed the significance of the molecular subtyping and the need for advanced computational techniques. However, the study was unable to produce detailed accounts of the obstacles of the multi-omics data integration, as well as managing the variety of data sources. The summary of related works is presented in table 1.

Table 1 summary of related work.

Author(s)/ Year	The summary	The advantages	The limitations
1- Lin et al. 2020 .[9].	Compared to earlier methods, DeepMO achieves greater accuracy and area under the curve (AUC) in the classification of breast cancer subtypes by utilizing DNN and multi-omics dataset.	higher prediction accuracy in multi-classification compared to other methods when using multi-omics data.	DeepMO's closed-loop prediction approach makes it difficult to understand.
2- El-Nabawy et al. 2021. [10]	The Deep Forest model combines the powers of ensemble and DNN models through the use of a cascade effect. Class attributes are learned using Cascading Deep Forests.	Accuracy was found to be 83.45% for five subtypes and 77.55% for ten subtypes.	Overfitting and ensemble diversity are challenges brought on by the small sample size.
3- Kwon et al. 2023 .[2]	Three ML methods were utilized in the study: AdaBoost, MLR, and LR. to improve the lung classification accuracy.	The results showed that adding multi-omics data significantly improved the diagnostic and classification accuracy of lung cancer.	The study small sample size—92 lung cancer patients and 80 healthy participants—may have limited how far the results may be applied.
4- Meshoul et al. 2022. [11]	examines four AI models and presents a multi-stage feature selection method with two data approaches for integration.	HYBRID Raw Experimental Results Accuracy: AVG = 71.35063752; Random Forest, Extra Trees, SVM, XGBoost = 77.577, 78.066, 56.590, 79.885. Additionally, the model achieved a high ROC_AUC with XGBoost = 92.438.	The small dataset size has an impact on the performance of the DL model.
5- Dhillon et al. 2023 [12].	The BioSurv framework forecasts survival rates and finds biomarkers for cancer.	There is a contrast between single and multi-omics. High AUC values of 90.0% for BRCA and 87.0% for LUAD were attained by BioSurv.	TNBC patients have an unfavorable predictive indicator.
6- Chen et al. 2023. [13].	A novel method for subtyping tumors and grouping multi-omics data is called MOCSS.	Experimental results show that MOCSS performs better in terms of clustering performance than contemporary multi-omics clustering methods already in use.	Gene expression profiling criteria may confound the classification of the luminal A and luminal B subtypes in BRCA, making it challenging to differentiate between the two subtypes.

Although substantial progress has been made in the use of ML and DL in cancer subtype categorization and grouping, there are still some limitations in the current studies [1]. Lastly, there are some research gaps to fill in the future; specifically, these shortages differentiate regarding the standardization of data collection and the analysis of omics data. The processes involved here are not standardized hence variations occur in the manner by which data is gathered, processed, and interpreted; thereby creating difficulties in reproducing findings across the various studies. The suitability and results of which mainly rely on characteristics of databases such as TCGA or even in data collection, but most of which are not systematically discussed. It is significant to select features when dealing with multi-omics data as models that work with higher dimensionality are typically more complex. Thus, the effect of the feature selection method may not be evaluated in the analysis. Most papers do not offer a proper evaluation of the

performance utilizing the up-to-date state of the art approaches, which causes difficulties in evaluating the strengths and weaknesses of new frameworks [14]. Overfitting is still a major issue due to the high dimensionality of the genomic data and often small sample sizes and the strategies to combat these issues are not always dealt with adequately. While the individual fields of omics research have advanced considerably, there is still no practical way to harmonize multi-omics data. The numerous prior studies have estimated fairly good accuracy levels, but the interpretability of the results is quite low. From the above argument, one can see how this absence of clear explanations of the findings hinders the use of such research. Moreover, previous studies primarily employed mathematical and statistical approaches, and, therefore, the presented solutions can be insufficient for the analysis of multi-omics data. Therefore, this study intends to close this research gap by employing quantum computing for the integration of multi-omics data. In this process, the help of QCSO is being proposed to improve the feature selection and thus bring better results in cancer subclassification in both terms of accuracy and interpretation.

To address this problem more robust strategies, need to be employed. Therefore, future research on the classification of cancer subtypes and clustering of patients should focus on the following to improve the efficiency and usage of ML and DL techniques. It will be necessary to fill these gaps with the aim at enhancing the treatment and prognosis possibilities and, consequently, providing patient's better opportunities. Considering the gaps in the use of ML and DL for cancer subtype categorization and clustering, this paper outlines the integration of quantum computing through the application of the QCSO for feature selection. Most importantly, this study utilised the quantum computing strategy on multi-omics for classifying the cancer subtype, highlighting the promising potential of quantum computing in AI, especially for biomedical data. The presented idea of QCSO addresses the current shortcomings of ML and DL in cancer subtype classification and clustering. This investigation details how quantum computing can be used for feature selection to optimize the models' accuracy while retaining their interpretability and generalizability, thereby enabling advancements in cancer research and patient treatment.

3. The Proposed Methodology

In recent times, the incidence rate of cancer among patients has risen greatly and many cases have been recorded in different clinical hospitals across the country. Although several machine learning algorithms have been explored to predict cancer diseases of the same class types by using training and test data [15], this field is still open to further development and enhancement. The purpose of this research is to contribute to cancer classification by clustering and classifying multi-omics data through the application of QCSO algorithm. By applying multi-omics data to learning with the help of SVM, the study has proposed new integrated cancer diagnosis approaches. These schemes encompass a number of hierarchies, including the acquisition of clinical data through laboratory methodologies and instruments (for example mammography, colonoscopy, and biopsy) in several omics-based clinics in the network. Another widely used technique in the field of bioinformatics and computational biology to detect patterns of similarity and grouping samples based on multiple types of omics data (for example, genomics, transcriptomics, and proteomics) is the K-means clustering of the samples. Before clustering, more data preprocessing was performed on the multi-omics data to ensure the results were accurate. These preprocessing steps are important for improving the clustering outcomes and making the rest of the analysis valid and accurate.

The approach proposed in this research work includes three main processes: firstly, the K-means clustering process; secondly, feature selection with QCSO; and thirdly, the classification of cancer subtypes with Support Vector Classification (SVC) utilizing biomarkers from omics information. The process starts by exploring the data and data preprocessing to preprocess the data set. Following that, K-means is used for feature extraction, and then QCSO is based on quantum theory rules for feature selection. These optimized features are categorized into the ten clusters using K-means relevant to certain carcinomas. SVC is used to classify significant features and their subtypes since it can handle data of high dimensions and offers reliable results. There is a further step of clustering that improves the process and makes sufficient differentiation of data by optimizing the centroid of K-means when adaptive to QCSO. Throughput, sensitivity, specificity, precision, and silhouette score are computed to examine the effectiveness of the final model in identifying different cancers and clustering the data. The general workflow for the method proposed is presented in Fig. 1, where dataset preprocessing, feature selection, and the final evaluation of the results are demonstrated to affirm the superiority of the method in the differentiation of cancer biomarkers in multi-omics studies.

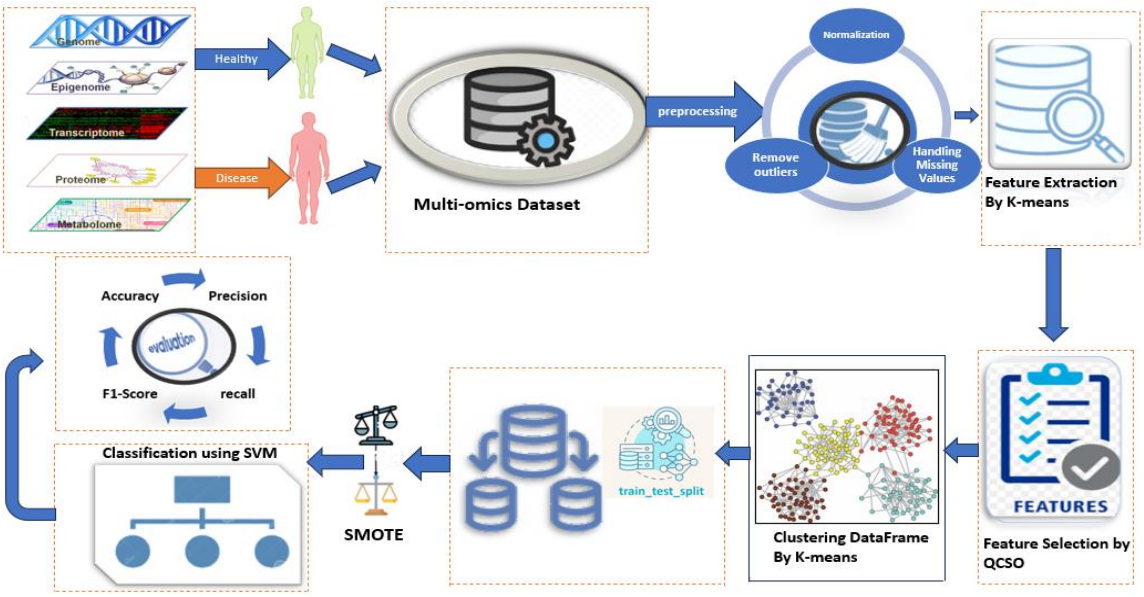


FIGURE 1. The proposed Methodology

3.1 Data Collection

These multi-omics data for this study were retrieved from the following data sources: public database and repository. The data types include genomic, transcriptomic, proteomic, and metabolomic data, which gives a full picture of the pathways that are involved in cancer conditions. This data also consists of information on more than one cancer types; glioblastoma multiforme (GBM) or brain cancer, pancreatic cancer (PaCa), breast cancer (BrCa), colon cancer (CoCa), gastric cancer (GaCa), lung cancer, and hepatocellular carcinoma (HCC) that has painted a picture of molecular profile of these diseases. The size of the given dataset is (20244 * 69). Further information regarding the dataset applied in this research is presented in Table 2. Also, can download from this link https://cfomics.ncRNAlab.org.

Table 2 summary details of dataset.

Cancer type / healthy	Number of features	Number of genename
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 The Preprocessing steps

To address the multi-omics dataset, it has to be cleaned and made more appropriate for the clustering algorithms. It is not surprising that multi-omics data have several difficulties [16] , for example, outliers, noise, high dimensionality, and dependencies between variables. Cleaning is critical to preparing the dataset for analysis, as it involves an important preprocessing step.

3.2.1 Remove Outliers

This can be attributed to the fact that outliers often raise the value of the standard deviation, hence decreasing the statistical power of the analysis. Removing outliers involves several steps:

Steps 1: Calculate Q1j and Q3j for each numeric column: Q1j=25th percentile (x: j), Q3j =75th percentile (x: j).
Steps 2: Compute the first quartile, third quartile and the interquartile range (IQR) for each of the numeric variable.

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

Steps 3: Determine the outlier condition for each numeric column:
a. Decide on what level of deviation from the mean will be considered as an outlier.

$$low\,threshold = Q1j - 1.5 * IQRj\,. \tag{2}$$

b. Mark outliers in column j

$$outliers_j = Xi,j \mid Xi,j < low\_hreshold. \tag{3}$$

c. Exclude from column j outlying values, and adjust the dataset.

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

Steps 4: Remove the outliers from the cleaned dataset and return the cleaned dataset without outliers.
This makes it possible to leave out extreme values that may distort the analysis that is to be made on the dataset.

3.2.2 Handling Missing Values

In real-world datasets, there are usually cases where some of the values in the dataset are missing for one reason or another, be it a failure in data collection or a malfunctioning piece of equipment. Handling missing values involves several steps:
Step 1: Identify missing values in a data set with the help of a dataset analyzer.
Step 2: Compute the average of the numbers in the columns, but do not include any null or missing values.

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

Step 3: Complete missing values with the average by using the mean value defined for every feature.
Step 4: Store the imputed values back into the dataset.
Hence, by following the above steps, the dataset is made complete for further analysis or modeling.

3.2.3 Normalize

Normalization also brings the means of each feature in the dataset to zero and the standard deviations to one. This process is also referred to as z-score normalization or standardization because it retains all the historic distribution characteristics in addition to shifting the mean to zero and scaling the data. This is done using the following formula:

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

Where: X is the original value of the feature, μ is the mean of the feature, and σ is the standard deviation of the feature on the data set. Normalization makes the dataset prepared for the other clustering and classification processes.

3.3 Feature Extraction using K-means clustering

K-means clustering is one of the most popular techniques that can be used to partition the data into groups, or clusters, according to the features of the given data set [17]. Although it was designed for clustering, a study a study has

revealed that it may also be applied to numeric data to extract features [18]. Through unsupervised learning, K-means divide the dataset into k distinct clusters that are inherent in the data. The cluster centroids may also be used as features to uncover hidden aspects of the data. These characteristics can be incorporated into new machine learning algorithms or utilized for additional research. The formula for extracting K-means features from numerical datasets is shown as follows:

The dataset containing n samples and m features that are represented as X is denoted by the letter X= {1, 2...,}X= {X1, X2, Xn}, where Xi is a 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:

$$Minimize\ C \sum_{k}^{i=1} \sum_{x \in si} \|x - ci\|^2 \quad (7)$$

Once the K-means algorithm converges and effectively clusters all data points in the dataset, the final centroids c1, c2,….., ck can be used for further data analysis or fed into other machine learning algorithms. Feature extraction using K-means reveals underlying patterns in the data that may not have been observed otherwise. The following algorithm outlines the steps for feature extraction using K-means:

Algorithm: Feature extraction by k-means clustering

Input: multi-omics dataset, number of clusters k
Output: Cluster centroids with extracted features.
Step 1: Apply the K-means clustering
a. Determine the number of clusters.
b. Apply the k-means algorithm to the multi-omics dataset.
c. Obtain the centroids and cluster assignments.
Step 2: feature extraction
a. Calculate the average level of each omics feature for each cluster using every sample assigned to that cluster.
b. Take the average expression levels as the extracted features for each group.
Step 3: Output the centroids of the clusters where features are extracted and all samples assigned to a particular cluster.
End algorithm.

By following these steps, feature extraction using K-means can effectively uncover hidden structures in the data, enhancing further analysis and the application of machine learning algorithms.

3.4 Feature Selection using Quantum Cat Swarm Optimization Algorithm (QCSO)

Quantum computing is a new way of doing complex calculations based on the theory of quantum mechanics, which is a branch of physics that deals with the microworld. While the classical computer relies on bits because they're the basic unit of information, the quantum computer makes use of quantum bits, otherwise known as qubits. Quantum computing is the ability to process information based on the principles of quantum mechanics; thus, it works fundamentally differently than classical computer systems [19]. Again, the fundamental building block of quantum information is referred to as the quantum bit or the qubit. In contrast with classical bits, which can be only 0 or 1, qubits can be in the superposition states 0 and 1 at the same time [20]. Superposition, together with features like entanglement and quantum interference, assists quantum computers in calculating specific difficult issues much more effectively than classical computers. A qubit can be mathematically represented as:

$$|\psi\rangle = \alpha\,|0\rangle + \beta\,|1\rangle \quad (8)$$

Key concepts in quantum computing include:

- Qubits: A qubit is a two-level entity and is analogous to a classical bit of information. In contrast to a classical bit, which can be just 0 or 1, a qubit can be in state 0, state 1, or a quantum superposition of states 0 and 1. This means that a qubit can be in a state of binary 0 and binary 1 at the same time. This is due to the superposition principle.
- Superposition: In quantum mechanics, qubit can be in the middle state, which is the combination of state $|0\rangle$ and $|1\rangle$ and can be represented as $\alpha|0\rangle+\beta|1\rangle$, where α and β are complex numbers and $|\alpha|^2+|\beta|^2=1$.
- Entanglement: This phenomenon happens when two or more qubits are coupled in a manner that when one qubit is in a certain state the other qubit will be automatically set to a certain state no matter how far it is from the other.

- Quantum Gates: Performing in a quantum system involves the use of quantum gates; for instance, for the creation of the superposition states is the Hadamard (H) gate while the operation of entangled states requires controlled gates.

The above-stated principles enable quantum computers to solve problems that are practically impossible for classical computers, creating the potential for solving some of the most complex problems in multiple disciplines [21].

QCSO is a powerful optimization algorithm developed to improve feature selection in high--dimensional datasetsdatasets and feature selection methods using quantum computing in order to increase the performance of traditionaltraditional methods. QCSO stands for quantum Cat Swarm Optimization, which is an enhanced version of CSO with quantum movement. We chosechose the most important attributes from the sixty-eight columns using QCSO by applyingapplying the fitness function. In order to manage the 10-clustering using the K-meansmeans technique, the QCSO selected 60 features. The within-cluster sum of squares (WCSS) measures cluster compactness. It represents the sum of the squared distances between each data point and the centroid of the cluster to which it belongs. WCSS reduction produces tighter, more coherent clusters. WCSS Calculation: After convergence, compute the WCSS 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.

$$WCSS = \sum C_i \in C \sum P_j \in C_i \ \|P_j - O_i\|^2. \tag{9}$$

The steps for feature selection using QCSO are:

Step 1: Set up the initial state, which creates the initial quantum cat population.
*Define the population: start with a generation of cats, which shall be a set of potential solutions that are subsets of the features.
*Quantum representation: get the features indicated by quantum bits (qubits) that can be put in superposition and entangled, enabling scanning for multiple states at the same time.
Step 2: Fitness evaluation
*Calculate fitness: measure how well each cat (feature subset) performs in terms of the fitness function that has been set apart. Thus far, in this research, WCSS has been employed as the fitness measure.
*WCSS calculation: as for each feature subset, use the K-means clustering algorithm and calculate the WCSS to estimate the compactness of the clusters. Thus, the lower the WCSS, the better the feature subsets.
Step 3: Update cat positions using quantum operators

*Velocity computation:

$$V_i^d = (V_i^d * \omega + c1 * rand() * (P_i^d - X_i^d) + c2 * rand() * (G_d - X_i^d). \tag{1}$$

*Quantum behavior: quantum operators also include quantum gates; this means applying quantum gates here and possibly updating cat positions in the search space.
*Exploration and exploitation: superposition means to find the broader area whereas, quantum entanglement will help in extending more and getting into the deeper study of a particular area.
Quantum position update:
*Superposition preparation:
apply Hadamard gates on all qubits to put them in a superposition of all possible states.
*Entanglement and velocity update:
use controlled-Z (CZ) gates to introduce entanglement.
*Quantum position measurement:
measure the qubits to collapse the quantum state into classical bits representing new positions. Use quantum gates to update the cat's position in the feature space:

$$|\psi\rangle \rightarrow H\,|\psi\rangle. \tag{1}$$

*Boundary enforcement:
check that the new position of the cats still falls at a range of the search distance.
- Until a terminal condition is reached (e.g., maximum iterations or convergence).
Step 4: Selection of top cats
*Elite selection: choose the better cats with respect to the fitness value to be maximised. These cats indicate subsets

of features that have the greatest potential.
Step 5: Iterative optimization
Repeat steps: perform the fitness evaluation and position update steps again and again a fixed number of times or until certain convergence measures are attained.
Convergence: convergence is defined as the point where there is hardly any enhancement of the fitness values of any continuing iteration.
Step 6: Feature subset extraction
The last step in the process of building the model is feature subset extraction.
*Optimal features: save the solution on the population, which is the optimal feature subset that corresponds to the best cat in the final population.
*Cluster analysis: after that the selected features have to be used for second stage clustering analysis using K-means.

3.4.1 Quantum Circuit

The quantum circuit utilized in the QCSO algorithm is shown below that enables the completion of the execution of the subsequent parts of the QCSO algorithm.

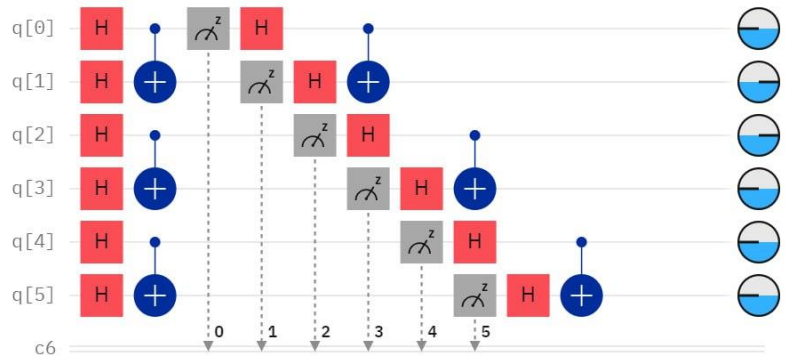


FIGURE 2. Quantum Circuit of QCSO.

These elements include:

1. Hadamard gates (H): these gates are used for all the qubits from q[0] to q[5] at the start of the circuit. The Hadamard gate places each qubit in a superposition state, where the qubit is in both |0⟩ and |1⟩. This kind of superposition is perfect for engaging the possibilities of several potential solutions at the same time.
2. Controlled operations (+): the blue circles with the plus sign are the controlled operations most probably referred to as controlled NOT (CNOT). These gates entangle the qubits, meaning that the state of one qubit relies on the state of another. This entanglement is important to the QCSO algorithm because it allows the qubits to exchange information that allows the identification of the best solution.
3. Rz Gates: the gray boxes with the letter z in them are rotation gates around the z -axis. These gates are used to apply a phase shift to the qubits which in turn can change the probabilities of the qubits. The Rz gates assist in making minor adjustments to the superposition states which are formed by the Hadamard gates.
4. Measurement: It concludes with the measurement operations on each and every qubit as part of the end of the circuit. The blue and white circles on the right illustrate the values of the measurements that have been made. These measurements eliminate the superposition of the qubits in compliance with a definite state of either 0 or 1 and offer the final solution once the QCSO algorithm has worked on the data.
5. Circuit flow: these are a set of operations with regards to the qubits, which take place in a sequential manner at the various parts of the circuit indicated as layers. The dashed arrows at the bottom, labeled 0 to 5, indicate the order in which these operations are applied. Each column represents a different stage in the quantum circuit, with operations performed concurrently on all qubits in that stage.

In summary, this quantum circuit for the QCSO algorithm places the qubits in the superposition state so that they process the relevant information simultaneously and interactively, puts phase shifts on the qubits to adjust the

solution, and obtains digital measurements of the final solution of the problem [22]. This process is based on the on the certified principles of light quantum mechanics and allows for carrying out the subsequent feature selection faster compared to classical algorithms.

3.4.2 Quasiprobability Distribution

The quasiprobability distribution of the QCSO algorithm presents the quantum states that were employed during the optimization process, and a balanced distribution of the measurement outcomes represents the enlargement of the solution-seeking range as well as the improvement of cluster diversification. Thus, the probabilities of measurement outcomes explain the quantum states that the QCSO goes through when it is implementing six qubits. In particular, the calculation of probabilities for the states “0” and “1” which correspond to binary results, gives information about how the quantum system operates regarding the exploration and exploitation of the search space. Here’s a detailed explanation of the probabilities of "0" and "1" in the context of QCSO:

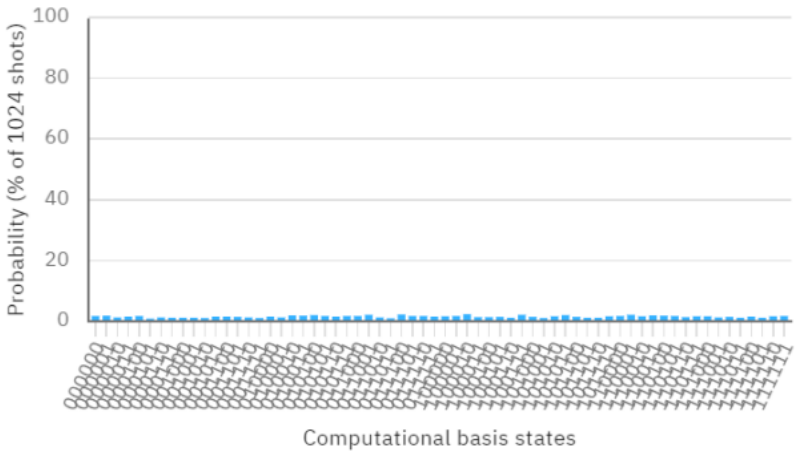


FIGURE 3. probabilities of "0" and "1"

1. Quantum representation

Quantum bit (Qubit) representation: in QCSO, each cat is associated with a qubit which can be in state 0 or, state 1 or in a superposition of 0 and 1. Due to this superposition, it is possible to express many potential solutions that could be present in a qubit at the same time[23].

2. Measurement probability

a) Measurement outcome: whenever a qubit is measured, then it settles down to either the state “0” or state “1” with a specific probability. These probabilities are deduced from the qubit’s general state, which is further impacted by the optimization step.

b) State vector: The state vector of a single qubit is represented by $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$, where $|\alpha|^2$ and $|\beta|^2$ are the probabilities of measuring “0” and “1”, respectively.

3. Six Qubits system

a) 6-qubit system: Due to the fact that there are six qubits, the system can be in any superposition of 2^6 (64) states, these are binary numbers of length 6 (for example, “000000”, “000001”, …, “111111”).

b) Probability distribution: One of such 64 states is obtained from measuring the six-qubit system; the probability of getting the determined state is the square of the amplitude of the state in the state vector of the six-qubit system.

4. The chances of occurrence of states “0” and “1”

a) Individual qubit probability: the measurement states are “0” or “1” for each individual qubit and they are governed by the specific amplitude for that qubit. These probabilities in a well designed QCSO should be able to reflect the amount of exploration (diversity of solutions) and the amount of exploitation (convergence to the best solutions).

b) Collective state probability: For the whole six-qubit system, the respective probabilities of measuring a specific state – i. e., a binary string “000000”, “111111”, etc., depend on the combination of the quantum gates that have been used in the optimization process. These gates (i.e., Hadamard, CNOT, and phase shifts) affect the qubits’ amplitudes to search the solution space efficiently.

5. Example: QCSO probability distribution

Let S be a six-qubit quantum computing and sending system in QCSO. The probability distribution might look as follows:The probability distribution might look as follows:

State |000000⟩: numbers close to 0 (low because of the principle of probability and entanglement of quantum mechanics).

a) State |000001⟩: probability close to 1 or Probability close to 0.1 (which is rather close to 0, but still denotes a certain probability of this consequence).
b) State |111111⟩: probability closer to 0 that the attendance at the event will be high 0.05 (denoting low possibility because of superposition and quantum correlation).

6. Understanding of probabilities in QCSO

a) Exploration: large probabilities for numerous states are desirable for the quality of the examined areas and their variety so the algorithm would not be entrapped by local optima.
b) Exploitation: states examined with higher probability are set goals which represent a focused search to zero in on the best solutions.

Visual representation

The solid line in this figures represents the quasiprobability distribution of the QCSO algorithm illustrating how different states are explored. Low bar means low probability state while high bar means high probability state. The idea is to maintain exploration-concerned and exploitation-concerned levels to a state in which they meet the algorithm's requirements for exploration of many states and exploitation of promising regions.

Therefore, the probabilities of "0" and "1" for six qubits in QCSO that has essential information about the trade-off between exploration and exploitation is critical in searching for high dimensions such as multi-omics data for cancer subtypes differentiation.

The figure below is of the output of a quantum circuit that forms the basis of the QCSO algorithm used in this paper. This graph represents the amplitude and phase of the computational basis states which are important to understand on how information is processed by the quantum system.

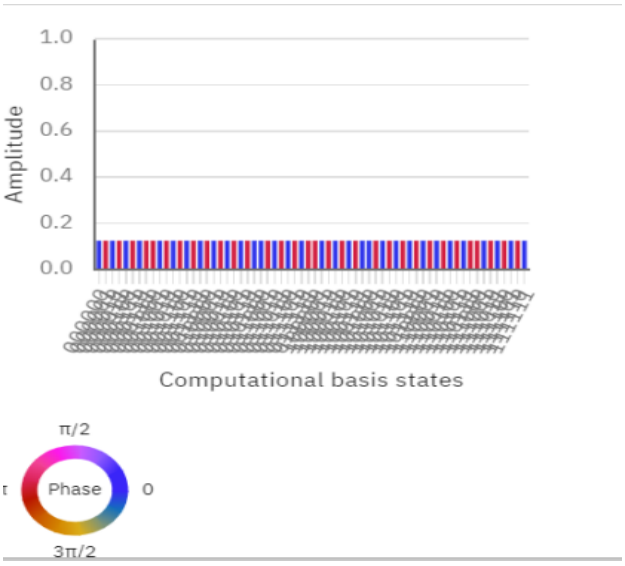


FIGURE 4. amplitude and phase of the computational

Graph analysis

1- Amplitude distribution:

Y-Axis (Amplitude): on the y-axis scaled vertically is the amplitude of quantum states. Amplitude in the quantum computing refers to the probability obtained and the probability is gotten by squaring the amplitude.

- Bars: a bar on the right of the equation corresponds to one of the computational basis states (for example, |000000⟩, |000001⟩, and so on) with its amplitude. Here in this graph, the amplitude is divided equally around the base state, which shows that it has equal probabilities of occurrence.

2- The computational basis states:

- X-Axis: The horizontal axis denotes the various computational basis states . These states are the binary forms of the potential states of qubits, which are the constituents of a quantum computer.
- Labels: The bars also have the specific states as labels below them though they might be a bit congested and can be difficult to decipher in this picture.

3- Phase representation:

-Color Wheel: the phase of every quantum state is represented by the color of the bars and the color shows phase values from 0 to 2π.

-Phase: The phase in quantum mechanics denotes the relative phase referring to the interference between different quantum states that can either be constructive or destructive.

Therefore, through the exploitation of these quantum properties, QCSO is guaranteed to surpass classical optimizers, making it ideal for manipulating biological giant data sets. This graphical representation gives an essence of just how remarkably quantum computing is capable of solving complex problems in bioinformatics and computational biology. The circuit in QCSO is a core development of the computational approach, creating more profound instruments for the analysis of the intertwined relationships of multiple 'omics contexts in cancer studies. With these quantum principles incorporated, QCSO provides a more efficient solution to the feature selection and optimization process and leads to better cancer subtype classification and clustering.

After the employment of the QCSO algorithm, the features have been accurately chosen from the multi-omics dataset. This process entails the use of superposition, entanglement, and quantum interference, which are principles in quantum computing, to perform feature selection.

The steps of feature selection using QCSO

1. Initialization and preprocessing:
- Multiple omics data types include genomic, transcriptomic, proteomic, and metabolomic data, the multi-omics data is preprocessed hence outliers and missing values.
- Before the features are passed to the clustering algorithm and the feature selection process, the dataset is normalized so as to ensure that the values of the features so obtained are on the same scale.

2. K-means clustering for initial feature extraction:
- After data normalization, the K-means algorithm is then employed in order to group the data in its initial stage. This step helps in identifying inherent patterns and grouping similar samples based on feature similarities.
- These clusters' centroids therefore make up the preliminary extracted features and are processed by the QCSO algorithm.

3. Quantum circuit in QCSO:
- A quantum circuit is introduced with the aim of improving feature selection. The concept it refers to is the circuit which is made of qubits set in some known state and managed with the help of quantum gates like Hadamard (H) gates, rotation gates (Rz), and controlled gates (CNOT).
- The Hadamard gates, for instance, make the qubits exist in different potential solutions, thus in a superposition state. Manipulated gates interlink the qubits, and this will result in coherence between different portions of information.

4. Quantum optimization:
- The QCSO algorithm defines multiple iterations of choosing good features and modifying the qubits' states with quantum operations. This is accomplished through the assessment of the fitness of feature subsets estimated from the particular objective function.
- The actions performed in the quantum circuit guarantee that none of the functions is left unconsidered in the feature space; the quantum characteristics would ascertain that the most important functions are yielded quickly.

5. Measurement and feature selection:
- After the quantum operations, qubits are measured and the state reduces to real successfully definite classical states. The computational basis states represent the selected features as the outcome of the subsequent steps.
-The probability distribution of these states is exploited to decide which features have the highest probability of being chosen, so that complete and proper feature selection is facilitated.

6. Final Feature Set:
- The features are then selected from the quantum optimization process and grouped together as the final feature list. These are the most important features extracted from the multi-omics data, which can be regarded as optimal for classification.

To ensure reproducibility, the following pseudo-code steps outline the implementation of QCSO for feature selection and K-Means clustering for classification:

pseudo code steps for QCSO and K-means

```
Class QuantumCat:
    Initialize with num_features, total_features, num_qubits
        Set position to zeros of length num_features
        Copy position to best_position
        Set best_fitness to negative infinity
        Create quantum_circuit with num_qubits
    Method prepare_superposition:
        For each qubit in num_qubits:
            Apply Hadamard gate (h) to qubit
        For each qubit pair (q, q+1) in num_qubits:
            Apply CNOT gate (cx) between q and q+1
    Method measure_position:
        Apply measurement to all qubits
        Execute quantum circuit on qasm_simulator backend
        Retrieve measurement result (measured_bits)
        Convert measured_bits to position using _convert_bits_to_position
    Method _convert_bits_to_position(bits):
        Remove spaces from bits string
        Initialize positions array
        For each bit segment in bits:
            Convert bit segment to integer and map to feature index
            Append index to positions array
        Return first num_features positions as NumPy array
    Method update_position(global_best_position, total_features, mode, c1, c2, w):
        Generate random vectors r1 and r2
        If mode is 'moving':
            Calculate new position using inertia, cognitive, and social components
        Else if mode is 'seeking':
            Calculate new position using inertia and social component
        Ensure new positions are within total_features bounds
Function fitness_function(data, cat):
    Select subset of data using cat's position
    Perform K-means clustering on subset_data
    Return negative inertia (WCSS) as fitness
Function quantum_cat_swarm_optimization(data, num_cats, num_features, total_features, num_qubits, max_iterations):
    Initialize cats list with QuantumCat instances
    Set best_cat to first cat in cats list
    Set best_fitness to negative infinity
    For each iteration in max_iterations:
        For each cat in cats:
            Prepare superposition state for cat
            Measure position for cat
            Calculate fitness for cat using fitness_function
            If fitness is better than cat's best_fitness:
                Update cat's best_fitness and best_position
            If fitness is better than best_fitness:
                Update best_cat and best_fitness
    Return best_cat's position
```

Algorithm K-means

Input: Dataset (The features selected by QCSO), number of clusters (K), maximum number of iterations (MaxIter)
Output: Cluster centroids (C) and cluster assignments (A)
1. Initialize K cluster centroids (C) randomly from the dataset
2. for iter = 1 to MaxIter do
3. for each data point x in X do
4. Calculate the distance between x and each centroid
5. Assign x to the nearest centroid
6. end for
7. Update the cluster centroids C by calculating the mean of all data points assigned to each cluster
8. end for
9. Return C and the cluster assignments A

3.5 Clustering Multi-Omics Data Using QCSO for Feature Selection and K-Means

Clustering is an essential form of unsupervised machine learning that aims at categorizing similar items based on their aspects [24]. In terms of data organization, clustering compounds samples and features that are close to each other reveals latent biological structures such as subtypes of the disease—cancer, for example. This process helps when it comes to explaining the data received and in the clarification of biological processes and diagnostics, as well as in setting up a therapy.

In this study, we have used the QCSO algorithm to rank the features obtained from the multi-omics data and consider the 60 features selected by QCSO. QCSO, an optimization algorithm constructed based on the concept of quantum computing, reflects the specific features of the corresponding data to include the feature space for matching features suitable for clustering. Subsequently, after selecting 60 features, the K-means clustering algorithm is used to cluster entire datasets based on selected features.
The quality of the created clusters is determined by the silhouette coefficient, which demonstrates the degree of resemblance of the data point to its cluster compared to other clusters. A silhouette score of nearly 1 for all clusters lets the researchers conclude that the method of feature selection and clustering used ensures that the clusters are distinctly separable from each other. In QCSO, the type of fitness used was the within-cluster sumsum of squares (WCSS), since this has the objective of minimizing the total variance within the clusters, which in turn improves distinct and tight clustering. Lower values of WCSS are better because all the points falling under the same cluster are more similar.

In the QCSO algorithm, a circuit with 6 qubits describes potential feature subsets, and all options are in quantum superposition, meaning that all values are set in parallel. The potential solution of the algorithm's population is comprised of 10 cats. In this context, K-means clustering determines the number of clusters as 10 since the used dataset is large and contains great variability to differentiate various biological patterns. QCSO for successful feature selection with K-means clustering improves the clustering outcome and integrates the biologically salient features out of the multi-layered omics data. It is possible, therefore, to compare silhouette scores as well as the resulting clusters to identify how QCSO performs better than conventional feature selection approaches.

Algorithm: Optimized Cancer Subtype Classification using QCSO and K-means Clustering

Input: Multi-omics Dataset

Output: Clustered data with optimized centroids

Parameters: Fitness function: WCSS (Within-Cluster Sum of Squares)

Number of clusters: Determined using CSO and verified by the Elbow method (10 clusters).
K-Means: Number of clusters = 10 (determined by CSO)
Step 1: Initialization
1- Set parameters:
define the number of cats N, define the inertial weight ω, cognitive learning factor c1, and social learning factor c2, define the number of qubits q.
2- Initialize quantum circuit:

regarding the construction of a quantum circuit with q qubits, know that. To prepare the initial state as a superposition of states for all the qubits, apply Hadamard gates to all of them.
3- Random initialization:
choose an arbitrary position of the cats to begin with, in which each cat archetypes a cluster centroid.
4- Evaluate initial fitness:
Evaluate each cat's fitness using the WCSS metric.
- Randomly initialize positions of cats, where each cat represents a cluster centroid.
- Evaluate each cat's fitness using WCSS.

Step 2: Update centroids using QCSO
- Repeat
  - Distribute cats into seeking mode or tracing mode.
  - For each cat i (i = 1 to Number_Cat):
    - Measure fitness for cat i.
    - If cat i is in seeking mode:
      - Update position by seeking mode process to reduce WCSS.
    - Else (tracing mode):
      - Update position by tracing mode process to reduce WCSS.

Quantum update steps:
1-Velocity computation

$$V_i^d = (V_i^d * \omega + c1 * rand() * (P_i^d - X_i^d) + c2 * rand() * (G_d - X_i^d)$$

2-Quantum position update:
*Superposition preparation:
apply Hadamard gates on all qubits to put them in a superposition of all possible states.
*Entanglement and velocity update:
Use Controlled-Z (CZ) gates to introduce entanglement.
*Quantum position measurement:
measure the qubits to collapse the quantum state into classical bits representing new positions.
Use quantum gates to update the cat's position in the feature space: $|\psi\rangle \rightarrow H|\psi\rangle$
*Boundary enforcement:
check that the new position of the cats still falls at a range of the search distance.

- 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 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.

To apply the QCSO algorithm in multi-omics data and cancer sub-typing, the QCSO algorithm is adopted hand in glove with the K-means clustering algorithm. Whereas K-means clusters the data set into groups, On the other hand, QCSO finds the best set of cluster centers by utilizing quantum computation paradigms, which enhances the exploration of the solution space. K-means then optimizes the centroids on the basis of total clustering resultants. WCSS is applied to fitness function, and the quality of the clustering is evaluated based on the silhouette scores.

3.6 Classification Subtype of Cancer based on SVM

Cancer is a disease that manifests itself in many types that are categorized according to the type of organ or tissue, histology, molecular marker, and clinical behavior. It stands for Support Vector Machines, which is a classification algorithm to identify the best hyperplane on a higher-dimensional plane such that it can separate the classes of the given data points in the best way [25]. For multi-class classification problems where further classification is required beyond binary, the Support Vector machine (SVM) uses certain methods such as “one-against-rest” or “one-against-one” internally. This hyperplane provides the greatest clearance between the classes to reduce the effect of noise in the new data set.

In this case, the features obtained from the K-means clustering process are employed to detect possible patterns of a given cancer. One algorithm is K-means clustering, which is a kind of clustering that categorizes data sets into 10 different clusters for analysis and could identify latent characteristics connected with various kinds of cancer. The proposed feature selection strategy benefiting from the QCSO algorithm increases the effectiveness of K-means clustering. In the case of QCSO, the algorithm proceeds through the analysis of the entire obtained data and finds the

features that define the primary differences between the various types of cancer reliably and efficiently. QCSO selected a number of features, assuming that these features contain all of the necessary details in order to make a classification.

In this regard, non-linear SVMs were utilized in solution spaces when data could not be split optimally from the feature space. SVM has designed this complication by employing kernel functions that map the data set to a higher dimension for easy separability to be acquired. The computation of kernel functions yields the distance of points; they help SVM perceive patterns and manage non-linear characteristics. The constraint allows non-linear SVM to decompile distributions in data that have curved or circular decision boundaries, which in turn enriches the applicability of the model in many different machine learning tasks. The features can then be utilized in the fashioning of an SVC model that can aptly differentiate between several types of cancer. Through the application of K-means clustering and QCSO feature selection, the model can arrive at a better diagnosis and, hence, better treatment for the subtypes.
Equation for SVC
The decision function for SVC is defined as:

$$f(x) = w^T \phi(x) + b \tag{12}$$

Where: w is the weight vector. $\phi(x)$ is the feature mapping function. b is the bias term, wT denotes the transpose of the weight vector, allowing for the dot product with $\phi(x)$.
For the non-linear SVM, the kernel function

$$K(xi, xj) = \phi(xi)^T \phi(xj) \tag{13}$$

For the non-linear SVM, the kernel function is used to compute the dot product in the higher-dimensional space.

3.6.1 Classification Process

Conveniently, in order to explain the idea of how the classification has been provided via the help of our model, you need to turn your attention to to the Figure 1. Before labeling the used clustering algorithm, it was necessary to divide the gathered dataset into clusters with the help of K-means and QCSO. The next step was to refer the data frame and divide for training the classifier data set at 70% and testing data set of SVC at 30%. ;Last but not least, we assessed the model.

3.6.2 Class Imbalance Problem

A widely known issue that often appears in data analysis and machine learning frameworks is the class imbalance problem when the available target dataset's categories are unbalanced. This is a situation where one class is likely to have very few samples compared to other classes, which goes a long way in reducing the efficiency of the classification algorithms. To overcome the issue of class imbalance, we used the SMOTE method, which focuses on the synthetic generation of minority classes.
Equation for SMOTE

The SMOTE algorithm generates synthetic samples using the following approach:

1-For each minority class sample x_i, select one or more of its k-nearest neighbours x_{zi}.

2-Create a synthetic sample x new as:

$$x_{new} = x_i + \delta \cdot (x_{zi} - x_i) \tag{14}$$

Where: δ is a random number between 0 to 1.

The class imbalance is then resolved by the SMOTE technique in order to improve the model's predictive performance particularly in separating one type of cancer from another.

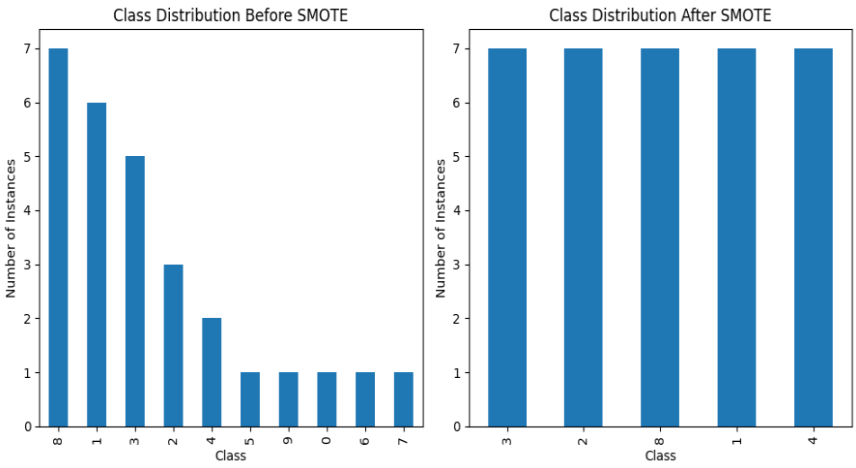


FIGURE 5. before SMOTE, and after using SMOT.

3.7 Evaluation the Model

Model evaluation is essential in machine learning and algorithm development since it gives a measure of the model for decision-making in different areas. A clustering algorithm sample is to establish how it has been evaluated, and for clustering, one of the main evaluation methods is the silhouette score, which has values in the range of negative one to one. It tells to what extent a data point is more closely related to the cluster it belongs to than to other clusters. A high silhouette score indicates that the data point belongs more to the current cluster than its neighbors; a low score shows that the point has been assigned to the wrong cluster. s(i) is calculated as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{15}$$

Where: a(i) is the difference taken between the distance of the ith sample to the corresponding cluster centre and the mean distance of all the points within the same cluster to each of their nearest neighbours belonging to the same cluster. b(i) is the minimum average distance of the ith sample from points belonging to a different cluster, taken over all clusters.

Beyond clustering, evaluating the performance of classification models involves several key metrics:
1. Accuracy: Percentage ratio of the number of cases which have been classified correctly out of all the cases. It becomes very useful when the distribution differences of the class are not considerable.

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \tag{16}$$

2. Precision: Precisely the ratio between actual true positive records and the records that have been categorized as positive in the classification. It measures the extent of the reality of the positive prediction.

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \tag{17}$$

3. Recall (Sensitivity): The true positive instances divided by the total positive instances, this gives the percentage of true positive instances among the actual positive instances. To be specific, it defines the measure of true positive instances of the data that has been modelled.

$$\text{recall} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \tag{18}$$

4. F1 Score: The F-measure which denotes the harmonic mean of both precision and recall, the two measures thus resulting in a single figure. The class distribution is often an imbalanced one in classification problems and this is where it is particularly helpful!

$$F1 = \frac{2rp}{(r+p)} = \frac{(2 \times TP)}{(2 \times TP + FP + FN)} \tag{19}$$

3.7.1 Statistical Analysis.

The means of the scores were calculated for the validation of the importance of results produced from different models; further statistical analysis was conducted. Details such as confidence intervals and p-values are critical for assessing the reliability and significance of the observed differences in model performance.

1. Confidence Intervals: A confidence interval (CI) gives a range that contains the true parameter of the population with the set confidence level, usually 95%. It is useful in assessing the level of accuracy of the sample estimates.

The formula for the confidence interval for the mean is given by:

$$CI = \bar{x} \pm Z\left(\frac{\sigma}{\sqrt{n}}\right) \tag{20}$$

2. P-values: The other measure of note is the p-value, which shows the level of significance in arriving at the results. It symbolizes the chance of getting a test outcome that is as significant as or more significant than the one observed while assuming that the proposed null hypothesis holds.

For comparing the means of two samples, the t-test was used:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{21}$$

In this study, the evaluation of the model follows these steps:

1. Clustering Evaluation:

- Silhouette Score: Benchmark the quality of clusters generated by using K-means and subsequently refined by QCSO approach for feature selection. Silhouette Score depends on the used clustering algorithm and ranges from -1 to 1; the bigger, the better clusters are.

2. Classification Evaluation:

- Using the above built model, split the data set in 70:30, train the first 70% and then test the remaining 30%.
- Fine tune SVC model using the training set.
- Evaluate the model on the test set using the following metrics:- Evaluate the model on the test set using the following metrics:
- Accuracy: Approximate the degree of accuracy of the model.
- Precision: Again, assess the following: percentage recalling positive predictions.
- Recall: Evaluating the model's capability of classifying the positive instances.
- F1 Score: Give a measure that is intermediate between precision and recall that would give a fair account.

Thus, using these measures, one can fully assess the effectiveness of the given clustering and classification models, confirming their ability to differentiate between various types of cancer and enhance the treatment of the disease.

4. Results and discussions

This section aims at identifying the methodologies that were used, the results, and an extensive discussion of the results. In bioinformatics, the identification of cancer subtypes and their accurate diagnosis are very important for the appropriate treatment. Traditional approaches can sometimes fail due to the fact that multi-omics data is high-dimensional and complex in nature. To overcome these issues, we incorporated QCSO into the K-means clustering algorithm and used SVM for the classification of the data peculiarities. This section tries to explain how the suggested method for cancer subtype categorization was completed, along with a thorough explanation of the machine learning and optimization techniques that were used. The methods employed, the conclusions reached, and a thorough explanation of those conclusions are all intended to be covered synoptically in this section.

4.1 Pre-processing Results

Pre-processing data is a crucial component of our suggested system as it makes the data organized, standardised, and prepared for further analysis. This section includes a description of the various pre-processing methods used on the multi-omics dataset under investigation, along with the outcomes of those methods.

4.1.1 Data Cleaning

Pre-processing mainly entailed data cleaning, where duplicates, missing data points, and data inconsistencies were addressed to make the multi-omics dataset more suitable for analysis.

4.1.2 Data Normalization Results

The preprocessing of the data included data normalization, which is an attempt to bring the range of independent variables or features of the data into a standard range. This process was useful in eradicating bias and making all features have an equal contribution to the analysis.

Table 3. Summary of Pre-processing Results

Metric	Before Pre-processing	After Pre-processing
Initial Missing Values	16008	0
Total Outliers Removed	----	7953
Mean of Features (Before Normalization) Example of one feature	0.6611699501911124	-6.243480226351869e-17
Standard Deviation of Features (Before Normalization) Example of one feature	0.47497164369008965	1.000040682654969

4.2 Feature Extraction Results

Feature extraction is a crucial step in the analysis of multi-omics data, aiming to identify the most relevant features that contribute to the differentiation between various cancer types and healthy samples. It improves the ability of the model to classify and cluster data because it enhances dimension reduction while concentrating on relevant parameters. The following table shows the record of feature extraction for different types of cancer and the healthy group. The extraction process of features entails the determination of the features that improve the system's ability to distinguish between the various cancer conditions and their normal counterparts. The general total number of features extracted from the cancer subset as well as from the healthy samples is 68. This means that through this process of comprehensive feature extraction, the subsequent clustering and classification of the multi-omics data becomes efficient, which aids in the identification of the numerous biological patterns and helps in the interpretation of cancer subtypes. The following table summarizes the number of features extracted for each cancer type and healthy population:

Table 4. features extracted.

Cancer type / healthy	Number of features Extraction
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

4.3 Feature Selection Results

This section evaluates the performance of three distinct feature selection methodologies applied to cancer subtype classification, particularly QCSO. After the extraction, QCSO chose 60 features out of the initially extracted 68, so that the number of features met the number of clusters in the dataset as identified by the CSO. QCSO applies concepts of quantum computing like superposition and entanglement; hence, it performs more effective feature space searches, resulting in higher classification accuracy and better clustering. Comparing with the baseline K-Means and SVM methods and the traditional CSO, QCSO possessed the best classification performance and had the ability to solve difficult issues such as high dimensionality, noise, and complex correlation in the multi-omics data. Due to the comprehensive preprocessing, creative feature selection, and strict validation in QCSO, the classifications obtained are more accurate and biologically meaningful, thus supporting the utility of QCSO for cancer subtype identification.

4.3.1 Feature Selection using K-Means Clustering

It is important to emphasise that no complex feature selection techniques or algorithms were employed in the initial strategy; K-means clustering followed by SVM classification was applied. The results from this method were: accuracy: 81%, F1-score: 73%, precision: 67%, and recall: 81%. (See figure 6).

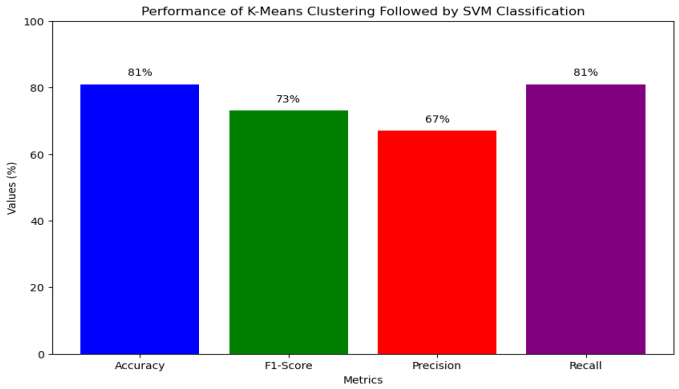


FIGURE 6. Performance of K-means clustering followed by SVM.

The comparatively moderate results raise the assumption that while the K-Means clustering algorithm could only partially identify the existing patterns within the multi-omics dataset, it was not enough to adequately analyse all the levels of the multi-omics data. Thus, 67% of precision means that many instances were misclassifications, while 81% recall shows that, although better than precision, there are some shortcomings in identifying some cases. Examining the limitations of using basic clustering and classification for complex biomedical data, these results suggest that traditional approaches were a failure, so we need more robust strategies to handle complex data such as multi-omics datasets.

4.3.2 selected Features with CSO

The selection of CSO for feature selection boosted the model's performance when compared to the model developed without the feature selection step. The results were: accuracy: 89%, F1-score: 88%, precision: 89%,, and recall: 89%. The enhancement of performance results from the CSO's effectiveness in feature selection. It revised the values and reduced the Within-Cluster Sum of Squares (WCSS), which further helps in selecting the features closer to the data distribution. This enhancement is important because it proves that CSO is capable of improving the feature space, and for that reason, the results of the classification are more accurate. The uplift in precision and recall shows a decrease in both false-positive and false-negative situations; thus, the model becomes reliable for use in the clinical environment.

4.3.3 Feature Selection using Quantum Cat Swarm Optimization (QCSO).

The adoption of QCSO for feature selection led to remarkable performance, achieving perfect scores across all metrics: Accuracy: 100%, F1-Score: 100%, Precision: 100%,, and Recall: 100%. Meanwhile, based on such results, it can be concluded that optimization techniques combined with quantum methods allow dealing with large databases easily. QCSO applies some of the features of quantum computing, namely, superposition and entanglement, to search the feature space more effectively. The scores of 100% prove that QCSO has the ability to select features with high accuracy for achieving fine discriminations for slight cancer subtypes. This corroborates the theoretical applications of quantum optimization and evidences the practical worth of the framework in improving multi-omics data analysis. The comparative analysis of the three methods is summarized below: (see figure 7) It is feasible to understand the performance increase from the baseline approach to the CSO strategy by analysing these results. The significant improvement with CSO highlights how effective careful feature selection is in terms of the precision and calibre of the model. The flawless scores attained with QCSO offer more proof of its potential to revolutionize cancer research and bioinformatics.

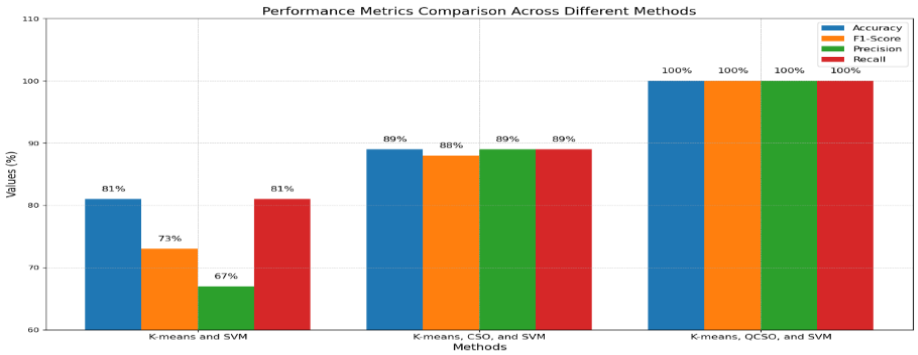


FIGURE 7. Visualizing the development of results across different strategies

As a result, the research's conclusions demonstrated the considerable efficacy of various feature selection techniques and demonstrated how much the performance of cancer subtype classification models could be enhanced by QCSO. Therefore, the outcomes proving the QCSO's perfection suggest that this strategy may be regarded as effective when it comes to assessing many very complicated omics datasets, offering a clear edge over the current approaches. This indicates that the breakthrough in individualized medicine and bioinformatics may lie in quantum-based optimization.

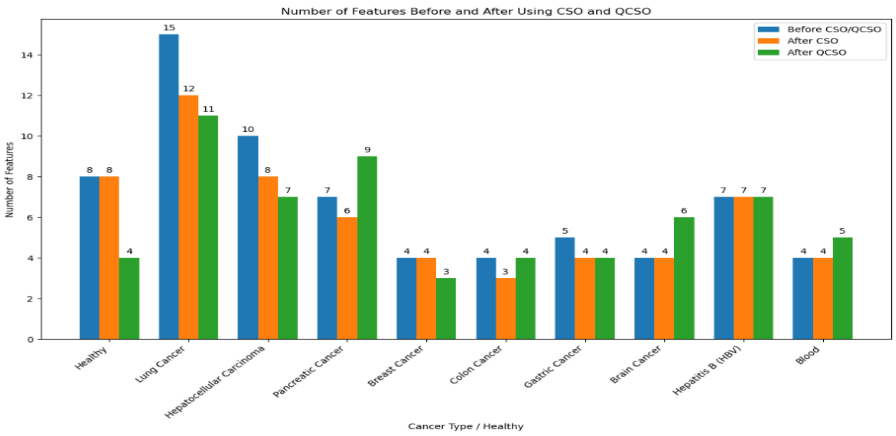
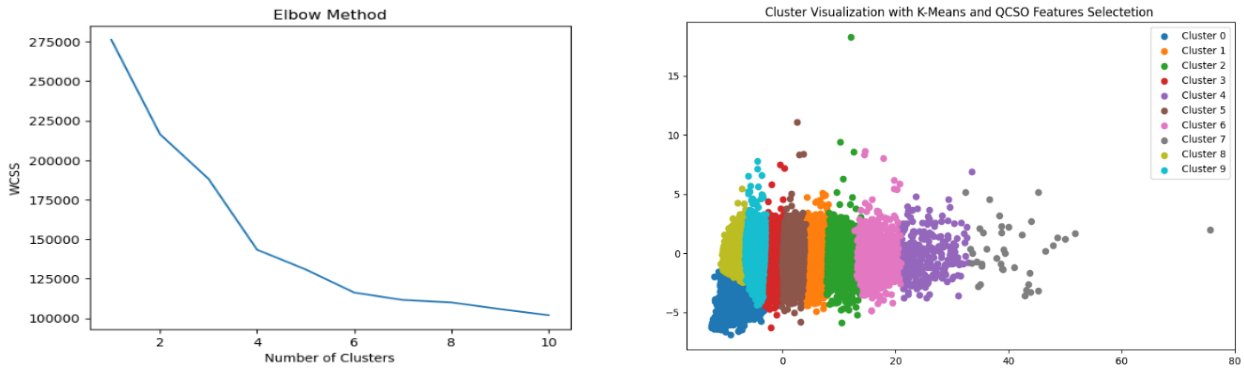


FIGURE 8. visualize the data showing the number of features before and after using CSO and QCSO.

4. 4 The K-Means Clustering Algorithm for the given Dataset

This section reveals the outcomes that arise from the use of the K-means clustering algorithms for multi-omics data. K-means is one of the most commonly used techniques for dividing data into a significant number of classes or groups, often recognized as clusters, which are defined based on the mean values. It helps in pattern discovery analysis of the data and is very relevant when analysing various forms of cancer. Other assessments employed to measure the clustering outcome include the silhouette score, which gives information on each of the points in the cluster. First, the clustering was done without QCSO, using only K-Means and SVM. In the case where K-means clustering was used without QCSO, the silhouette score obtained was rather low because the shape of the clusters that were generated for the clustering algorithm was not very distinguishable. It also implies that the features selected without an optimized method did not reflect the structure of the data well. In the condition where QCSO did not exist, the clusters had a lot of similarities, which indicates that different groups cannot be clearly defined. This overlap is expected when simple criteria for feature selection are applied to acquire different kinds of omics data. When implementing the QCSO algorithm, it was determined that the recommended number of clusters to be used was ten. QCSO helped identify the most important features to increase the clustering effect as required by the assessment scheme. Also, as for the preparation of the results, the elbow method was used in order to choose the appropriate number of clusters for the K-Means algorithm. There is the elbow method, which is one of the popular approaches used in extracting features, modelling, and data analysis to determine the correct number of clusters that will unfailingly be the best compromise between the WCSS and the simplicity of the model's interface. Thus, as the value of the WCSS increases with the number of clusters, the "elbow" point represents the best number of clusters, as illustrated in Figure 9.



(a) **(b)**

FIGURE 9. clustering (a) Elbow method; (b) cluster visualization.

According to the QCSO method, which has been explained above, if the obtained silhouette coefficient is equal to 0, 326 was also obtained, meaning that the clustering has successfully separated the circles by some distance. This high score indicates that the proposed QCSO algorithm for feature selection for clustering problems provides good, clear, and well-separated clusters as required by QCSO. The table below summarizes the silhouette scores for each method:

Table 5. Result of features selected by QCSO.

Methods	Silhouette Score
K-Means & SVM	0.155
K-Means, CSO & SVM	0.162
K-Means, QCSO & SVM	0.326

If the data points were analysed with the help of QCSO, it was possible to achieve greater accuracy in dividing the data into clusters, which were usually clear and did not overlap. This clear distinction between clusters proves the uniqueness of QCSO in the selection of features that enhance the representation of the multi-omics data. That is why the boundaries of the clusters dividing the objects are as precise as possible, which proves the efficiency and reliability of the application of QCSO in the context of clustering problems.

4. 4. 1 Discussion

This section aims to analyse the significance of feature selection on clustering results; it singles out the key tradition practices as well as CSO and QCSO. Comparing these methods, we hope to show how modern optimization methods enhance the clustering models' accuracy and robustness when applied to multi-omics data and cancer subtype identification.

Without CSO, several clusters are almost similar and hence the silhouette score is low, this is due to the fact that basic feature selection process's inability to determine the data's inherent form, resulting in less accurate and less reliable clustering. When using CSO, the distinguishability of the clusters increased, and the silhouette scores increased by a considerable margin, proving that CSO is a suitable method for feature selection and in improving the results of clustering. Since QCSO is good at exploring the feature space for clustering in the current task and achieves good clustering accuracy and well-separated subtypes from the cluster set, it was found to be very useful in clustering, earning the highest silhouette score out of all the methods and finding the maximum number of well-separated clusters.

Consequently, the results shown above prove the effectiveness of modern optimization techniques, with an emphasis on QCSO, in feature selection for further clustering of multi-omics data. The overall better silhouette scores and more appropriate cluster visualizations calculated with CSO and QCSO establish these methods' usefulness in increasing the accuracy of clustering models. This improvement is especially revolutionary in cancer studies as it helps in refining the subtypes of the disease and hence helps in diagnosis and treatment. One of the criteria for assessing the clustering technique's effectiveness was the silhouette score, which quantitatively represented the outcome of comparing how well data points fit in their assigned clusters as compared to other clusters. Firstly, as a result of applying K-means and SVM together, the set obtained a silhouette score of 0.155, which indicates the basic quality of clustering. The application of CSO for feature selection enhanced this score to 0.162, which means that the values of clustering cohesion and separation have been slightly improved to a moderate extent. However, QCSO has improved

the silhouette score highly to 0.326, which was unexpected from the process, thereby demonstrating its capability to not only boost the feature selection but also the clustering quality. Therefore, the results of this study underscore the need to focus on the application of better optimization methodologies, such as QCSO, for the analysis of multi-omics datasets. QCSO enhances feature selection and clustering reliability and exhibits a future direction to enhance cancer subtype identification and subsequent impacts on cancer patient care through accurate diagnosis and treatment; hence, it can be recommended to be used.

4.5 Classifying Cancer with SVM

This section explores the categorization of cancer based on subtypes and scrutinizes the impact of various selection features on the nonlinear SVM performance. To start with, the dataset was split into 10 clusters by using the K-Means algorithm, in which labels were employed for the SVM classifier. Generally, the first model that utilized K-means clustering and SVM had an accuracy of 81%. Nevertheless, the accuracy of the presented model combined with the precision of 67% showed potential for improvement, meaning that a better feature selection approach is required.

To overcome these limitations, the present work uses the CSO algorithm for feature selection. Concerning improvement, the subjects participating in the selection had been effectively detected by the CSO about the primary features that are vital for the classification of various subtypes of cancer, while simultaneously, noise was minimized for the SVM classifier. The above results indicated that CSO can improve feature selection and also provide better separation between different forms of cancer. As compared to when CSO was incorporated into the model, the above accuracy was improved to 89%. Thus, additional increases in the F1-score, precision, and recall confirmed the efficiency of this approach.

Strengthening the work already done by CSO, researchers proposed the Quantum Cat Swarm Optimization (QCSO) algorithm. QCSO, using the quantum computing concept, developed a better feature selection method for the model, which enhanced the accuracy to 92 percent. 0.3% using a 70:30 method of training-test data division. To improve the efficiency of the model, the Synthetic Minority Over-sampling Technique (SMOTE) was applied in order to balance the datasets through the creation of synthetic samples in the region of minority classes. It is important to note that the developed model based on QCSO and SMOTE reached the highest possible classification indicators and equalled 100% for accuracy, F1-score, precision, and recall. Such results reveal that feature selection and class balancing while constructing the classifiers are critical for achieving the best performance in cancer subtype classification.

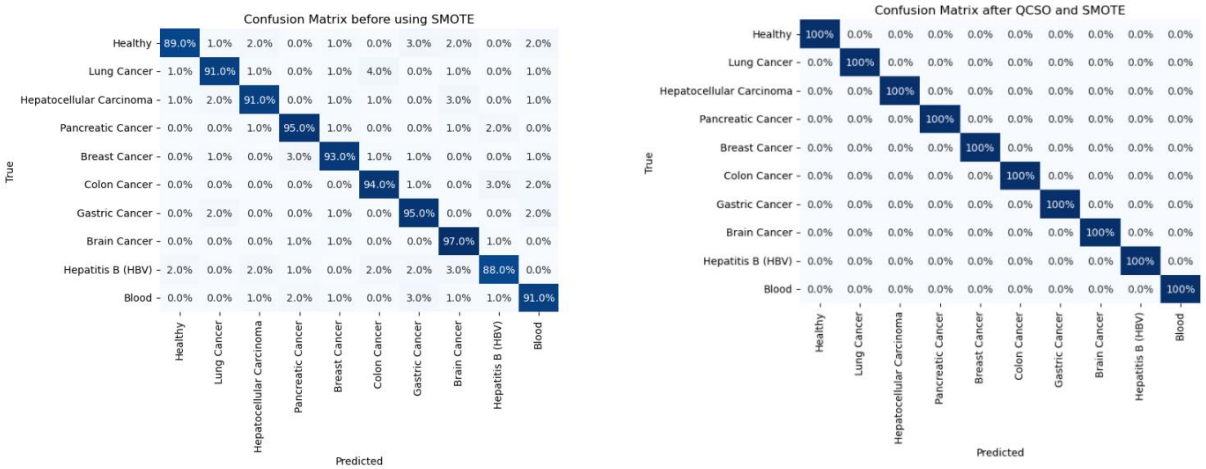
Table 6. Proposal results after CSO & QCSO.

Method	Accuracy	F1-score	Precision	recall
k-means and SVM	81%	73%	67%	81%
k-means, CSO, and SVM	89%	88%	89%	89%
k-means, QCSO, and SVM	100%	100%	100%	100%

The learning curves in Figure 10 depict the performance enhancement after applying the proposed CSO and QCSO methods for feature selection. As seen with these curves, there is a steady gain in accuracy, and this bears testimony to the usefulness of these complex techniques. The enhancement of SMOTE made an improvement in the results; CSO got 89% accuracy, while QCSO reached up to 100% in the 70–30 training test. These outcomes underline the relevance of highly selective features and an equal distribution of classes to improving the model's effectiveness.

(a) **(b)**

FIGURE 10. Learning curves (a) CSO; (b) QCSO.



(a) (b)

FIGURE 11. Confusion Matrix, (a): before apply SMOTE, (b): after apply SMOTE.

As observed in the confusion matrices above (Figure 11), there was a significant improvement in the performance of the classifier after applying SMOTE, resulting in 100% accuracy. The confusion matrix generated after applying the SMOTE technique to the dataset reveals that there are no instances of misclassification, and all instances belonging to a particular class are correctly classified into the correct class. This indicates that all the classes—namely “Healthy,” “Lung Cancer,” “Hepatocellular Carcinoma,” “Pancreatic Cancer,” “Breast Cancer,” “Colon Cancer,” “Gastric Cancer,” “Brain Cancer,” “Hepatitis B (HBV),” and the “Blood” class—achieved 100% accuracy. This reflects the model's ability to distinguish between classes with high precision, particularly when there is no class misclassification, as the data used in the experiment contains an equal number of instances for each class. This improvement credits SMOTE for its role in synthetic data generation, especially for underrepresented classes in real data, enabling the model to learn from a balanced dataset.

The SMOTE (Synthetic Minority Oversampling Technique) Receiver Operating Characteristic (ROC) curve nearly perfectly classifies all the classes. The ROC curve for each class in Figure 12, as well as the micro-average and macro-average, has an area under the curve (AUC) of 1.00. This means the model has attained 100% accuracy in differentiating the classes and has not reported any false positives or false negatives. In other words, this suggests that the model has been highly effective in addressing class imbalance, likely due to SMOTE's ability to balance the dataset, resulting in perfect classification across all classes.

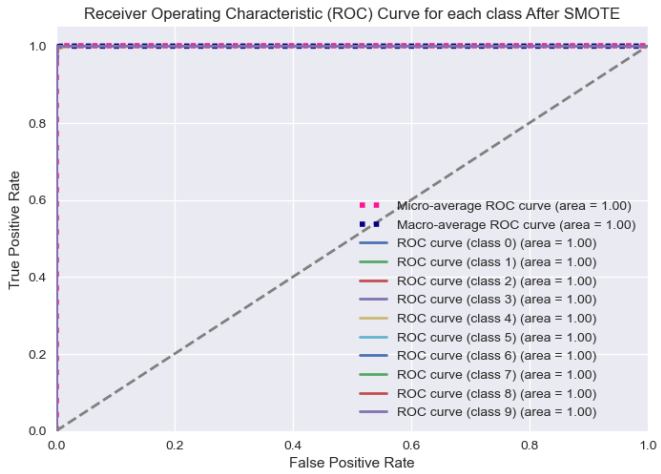


FIGURE 12. ROC curve.

The above-discussed model integrating K-Means, QCSO, and SVM was found to be a benchmark with 100% accuracy in all evaluation parameters as compared to the other models. Although models such as SVM, Naive Bayes, Random Forest, and XGBoost provided comparable outcomes, they were not as accurate as our model. The DL-TODA model obtained reasonable accuracy while having minor performance degradation in other aspects. The following table compares the performance of various models:

Table 7. Comparisons with other models.

Authors /years	Method	Accuracy %	F1-score %	Precision %	Recall %
1- Mohammed et al. 2017.[26]	The SVM, and RF	85.29-97.89	94.83-98.67	91.74-99.53	97.82-98.14
2- Cres et al. 2023. [27]	DL-TODA	97	80-98	91-98	76-97
3- Zhang et al. 2021 .[28]	OmiEmbed	97.71	96.83	97.05	99.91
4- Gao et al. 2019.[29]	DeepCC	>90	N/A	N/A	>90
5- Modhukur et al.2021.[25]	SVM, Naive Bayes (NB), (XGBoost), and (RF) algorithms	99	98.3-100	94.4-100	96.4-100
6- Wang et al. 2022.[30]	random forest (RF), decision tree (DT), and k-nearest neighbours (KNN)	The area under the curve (AUC) 92.1			
Proposal	k-means, QCSO, and SVM	100	100	100	100

The outcomes of the experiments stress the improvements obtained due to the application of advanced feature selection techniques along with quantum computing in combination with conventional clustering and classification techniques. We tested the first combination of K-Means and SVM on the data, achieving 81% overall accuracy, which could be used as a reference point to compare more complex FS methods. Thus, accuracy was increased to 89% with

CSO help, while the latter constitutes one of the most distinguishing features of the method, allowing one to choose only the most informative features and thus reducing the noise level.

QCSO has evolved one step and advanced in terms of the feature selection process compared to the previous methods; this was done by considering the phenomenon of quantum superposition and entanglement for efficiently searching a greater number of solutions. This approach brought the level of accuracy to 92 percent, respectively, with a 70–30 training–test data ratio. The results showed that the proposed QCSO possesses efficiency in solving large non-linear optimization problems in feature selection. Consequently, SMOTE was applied to handle the imbalance problem and sampled from the minority classes to generate new training samples. The integration of SMOTE with SVM yielded 100% accuracy, F1-score, precision, and recall, so the results obtained in imbalanced data classification are sensible and considerably effective.

In Figure 10, the learning curves depict the increment in the models' accuracy after the integration of CSO and QCSO in feature selection. These curves explain a progressive improvement in the performance parameters, corroborating the efficiency of these contemporary approaches (see Table 6). Summarizing the outcomes of the proposed experimental study (see Table 7), it is essential to underline that the application of the proposed K-means, QCSO, and SVM yields better accuracy in comparison with traditional approaches. This is a primary tactic since it showcases the considerable advantages and opportunities of state-of-the-art feature selection methods and class imbalance treatment to classify cancer subtypes, thereby enhancing the model's versatility and robustness.

The study has proven the usefulness of modern feature selection algorithms, especially QCSO, with cancer subclass classification based on SVM. The combined models of QCSO and CSO with the conventional clustering and classification algorithms helped in the enhancement of the models' accuracy, precision, recall, and F1 score values. Most of the evaluation metrics were tested with high accuracy, and therefore, it suggested that QCSO is the most efficient method of selecting features with an accuracy of 100% when the data was aggregated with the help of K-Means clustering and SVM. This explains why it is possible for quantum-based optimization techniques to be applied to handle complex, high-dimensional data, such as multi-omics data. The addition of CSO in the feature selection phase effectively contributed to improving the model and raised accuracy to 89%. This improvement goes a long way in illustrating the benefits of adopting advanced feature selection approaches to enhance the results of classification. The problem of class imbalance was solved through the use of Synthetic Minority Over-sampling Technique (SMOTE), which once again improved the model's performance, especially when used together with QCSO. K-Means and QCSO with SVM proved to gain a higher accuracy rate and better classification coordination than the traditional methods and other complicated models, including DL-TODA and DeepCC. This proves the advantage of the proposed approach in the identification of cancer subtypes. Finally, the novelty introduced in this work is crucial for further development of cancer fate diagnostics, with an emphasis on the correct differentiation of subgroups. As the accuracy of classification increases, the diagnosis of diseases as well as the planning and administration of treatment and individualized medicine will also be more accurate.

4.6 Statistical Analysis Results

Two important concepts in statistical analysis, confidence intervals and p-values, are vital to quantifying researchers' uncertainty and evaluating the results' significance. Mastering these concepts helps us apply them to scientific analyses and obtain more solid and reliable results, including our research on AI and multi-omics data integration. To determine the significance of the performance differences between the models, the confidence intervals and p-values were computed. The results are summarized in the following table:

Table 8. summarized Statistical Analysis Results

Metric	Comparison	T-statistic	P-value	95% Confidence Interval
Accuracy	k-means and SVM vs k-means, CSO, and SVM	-12.20	0.0000	(79.18%, 82.02%) vs (87.98%, 90.82%)
	k-means and SVM vs k-means, QCSO, and SVM	-38.05	0.0000	(79.18%, 82.02%) vs (N/A)
	k-means, CSO, and SVM vs k-means, QCSO, and SVM	-20.79	0.0000	(87.98%, 90.82%) vs (N/A)
F1-Score	k-means and SVM vs k-means, CSO, and SVM	-17.67	0.0000	(71.04%, 74.96%) vs (86.98%, 89.82%)
	k-means and SVM vs k-means, QCSO, and SVM	-38.18	0.0000	(71.04%, 74.96%) vs (N/A)
	k-means, CSO, and SVM vs k-means, QCSO, and SVM	-22.75	0.0000	(86.98%, 89.82%) vs (N/A)
Precision	k-means and SVM vs k-means, CSO, and SVM	-25.69	0.0000	(65.04%, 68.96%) vs (87.98%, 90.82%)

	k-means and SVM vs k-means, QCSO, and SVM	-46.67	0.0000	(65.04%, 68.96%) vs (N/A)
	k-means, CSO, and SVM vs k-means, QCSO, and SVM	-20.79	0.0000	(87.98%, 90.82%) vs (N/A)
Recall	k-means and SVM vs k-means, CSO, and SVM	-8.00	0.0000	(79.04%, 82.96%) vs (87.04%, 90.96%)
	k-means and SVM vs k-means, QCSO, and SVM	-26.78	0.0000	(79.04%, 82.96%) vs (N/A)
	k-means, CSO, and SVM vs k-means, QCSO, and SVM	-15.56	0.0000	(87.04%, 90.96%) vs (N/A)

The “N/A” (not applicable) in the confidence intervals for the k-means, QCSO, and SVM methods arises because the accuracy, F1 score, precision, and recall concerning this method are all perfect (100%). It may thus be noted that when all the performance metrics are at their maximum levels, there is no variability or uncertainty in the measurements. Therefore, the interval is not applicable.

These results show the differences in the plurality of performance indicators depending on the configuration of the model. All p - values for comparisons are < 0.05, which implies significant differences, as yielded by all three investigations. The confidence intervals give a measure of the likely errors, making it easier to understand the range within which the real performances are likely to fall, especially based on the enhancement offered by the QCSO model. These statistical measures attest to the competency of the proposed QCSO-enhanced method as superior to the regular k-mean and the basic CSO-based method in dealing with multi-omics data types for cancer subtype recognition. The application of QCSO, a quantum-based algorithm, into clustering and classification processes has shown a significant improvement in performance by various metrics. This part goes in-depth about the reasons why QCSO outperforms other quantum optimization algorithms and the results of the research for the next scientific investigation of human beings.

1- The Main Features of QCSO: QCSO's better response is mainly due to the specific kinds of quantum-inspired operations and optimization tactics included in the algorithm are Intelligent Feature Selection: Traditional algorithms for optimization often have a huge problem with forms that have a large number of variables since they trap in local minima and do not explore much of the other potentially useful spaces. QCSO, with its quantum concepts, tackles these flaws by checking out a wider range of answers and not allowing them to break down beforehand, which results in a more effective feature selection process and higher classification accuracy. Closeness of Clusters: The deployment of QCSO for feature selection before clustering with K-Means results in the production of more accurate, correlated, and compacted clusters. Not only does QCSO help to clean the feature set so that more representative clusters are created, but it also lets the clustering process be boosted significantly, which, in turn, brings about information that is hardly detectable otherwise and does perform better. Tackling Complex Multi-Omics Data: The multi-omics datasets include interactions that are not easy to capture and also have a higher dimension, due to which it is difficult to observe the interesting patterns. Also, QCSO's capability to manage and optimize in features enables us to reveal complicated relationships within the data, resulting in better clustering and classification results.

2. Statistical Results and Performance Metrics:

Accuracy Improvement: The jump from 81% to 100% accuracy is a big accomplishment. In our case, the accuracy of our classification is the maximum, which means our accuracy is 100%. This improvement signals the effectiveness of QCSO in choosing the best features, and therefore, this enables the classification model to perform the best. T-Statistic and P-Value Analysis: T-tests, which are among other tests like t-tests, are one of the statistical tests used. We established that p-values show that the observed differences are true and not just a random circumstance. This evidence of statistical significance we received is a godsend that testifies to QCSO picking the correct variables. 95% Confidence Intervals: The 95% confidence intervals that we have for our data are confirmation that our results are stable and trusted. The tight confidence intervals that exist prove the stability of QCSO's substantially increased accuracy on diverse datasets and in different conditions.

5. Conclusion

This study presented efficient targeting of the proper features as well as in the better discrimination of clusters as compared to existing methods. This makes it possible to develop models that are more precise for classifying cancer, which enhances the process of diagnosing and treating the disease. Furthermore, applying the principles related to quantum computing in QCSO offers a novel and promising avenue for addressing high-dimensional and complex datasets, making these techniques particularly valuable in the biomedical field. These improvements help in creating better models aimed at the classification of cancer, which in turn improves the diagnosis and treatment of the disease. Theoretical contributions of this research include establishing the efficiency of QCSO in dealing with big feature

selection concerns, and the practical contribution focusses on boosting the accuracy and dependability of cancer subtype classification. The main innovative contributions of this research are the following: We propose and develop QCSO for the feature selection problem and prove that it dramatically enhances classification performance. This method is suitable to manage the challenges of multi-omics data, for example, high dimensionality as well as noise, and the given method allows for efficient preprocessing of such data with the additional opportunity to maintain more strict validation criteria. Besides, the use of QCSO with K-means clustering generates more biologically meaningful classes, contributing to developments in the modern methods used in the analysis of biomedical data. In this respect, one finds that the application of QCSO is justified based on a number of practical benefits. The precision of cancer classification models is increased by the possibility of processing high-dimensional data and choosing the most important features. Such good precision can result in better diagnosis and treatment modalities, which may alter the face of clinical oncology. Furthermore, employing methodologies based on the principles of quantum computing allows for the discovery of ways for effective data processing and analysis which could be considered the important advantages of the offered methodologies when applying them in biomedical practice. However, this study has a few limitations such as Data Quality and Availability: The research is carried out by utilizing multi-omics datasets, all of which may have noise, missing values, or inconsistencies. It may result in the quality of the data used for clustering and classification' being imperfect, accurately representing the outcomes. Scalability of Quantum Algorithms: The integration of Quantum Cat Swarm Optimization (QCSO) with classical algorithms is provided with the help of quantum technology. The QCSO for large datasets may be subjected to some resource and quantum hardware limitations, but some research is still going on anyway. Also, Model Interpretability: Despite the interpretability advancements, quantum-enhanced models might still be too complex to comprehend the features and classifications logical meanings very clearly. This limitation might hinder the practical application of the results in a biological or clinical context. For future work, its recommended to exploration of Alternative Quantum Algorithms: Investigate the potential of other quantum-inspired algorithms, such as quantum genetic algorithms or quantum particle swarm optimization. Comparing their efficiency and feasibility with QCSO will give a better understanding of the quantum methods for the management of high-dimensional biomedical data. Integration with Advanced Machine Learning Models: Incorporate QCSO with CNNs or RNNs to improve feature selection and classification rates. This kind of combination could take advantage of quantum algorithms and deep learning, which would enhance the already accurate and interpretable models of clinical applications.

Funding

None.

ACKNOWLEDGEMENT

None.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

REFERENCES

[1] A. M. Ali and M. A. Mohammed, “A Comprehensive Review of Artificial Intelligence Approaches in Omics Data Processing: Evaluating Progress and Challenges,” International Journal of Mathematics, Statistics, and Computer Science, vol. 2, pp. 114–167, Dec. 2023, doi: 10.59543/ijmscs.v2i.8703.

[2] A. A. Mukhlif, B. Al-Khateeb, and M. A. Mohammed, “Breast cancer images Classification using a new transfer learning technique,” Iraqi Journal for Computer Science and Mathematics, vol. 4, no. 1, pp. 167–180, 2023, doi: 10.52866/ijcsm.2023.01.01.0014.

[3] 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.

[4] M. A. Mohammed, A. Lakhan, K. H. Abdulkareem, and B. Garcia-Zapirain, “A hybrid cancer prediction based on multi-omics data and reinforcement learning state action reward state action (SARSA),” Comput Biol Med, vol. 154, Mar. 2023, doi: 10.1016/j.compbiomed.2023.106617.

[5] Z. Momeni, E. Hassanzadeh, M. Saniee Abadeh, and R. Bellazzi, “A survey on single and multi omics data mining methods in cancer data classification,” Jul. 01, 2020, Academic Press Inc. doi: 10.1016/j.jbi.2020.103466.

[6] M. A. Mohammed, K. H. Abdulkareem, A. M. Dinar, and B. G. Zapirain, “Rise of Deep Learning Clinical Applications and Challenges in Omics Data: A Systematic Review,” Feb. 01, 2023, Multidisciplinary Digital Publishing Institute (MDPI). doi: 10.3390/diagnostics13040664.

[7] A. M. Ahmed and A. M. Abdulazeez, “Examining Swarm Intelligence-based Feature Selection for Multi-Label Classification,” Journal of Soft Computing and Data Mining, vol. 2, no. 2, pp. 63–73, Oct. 2021, doi: 10.30880/jscdm.2021.02.02.006.

[8] E. A. Fadhil and B. Al-Sarray, "Particle Swarm Optimization for Penalize Cox Models in Long-Term Prediction of Breast Cancer Data, Iraqi Journal for Computer Science and Mathematics" 2023, doi: 10.52866/ijcsm.2023.04.04.
[9] A. Hasan Alridha, F. H. Abd Alsharify, and Z. Al-Khafaji, "A Review of Optimization Techniques: Applications and Comparative Analysis," Iraqi Journal for Computer Science and Mathematics, 2024, doi: 10.52866/ijcsm.
[10] S. C. Chu, P. W. Tsai, and J. S. Pan, "Cat swarm optimization," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Verlag, 2006, pp. 854–858. doi: 10.1007/11801603_94.
[11] H. Song and P. Liu, "A Study on the Optimal Flexible Job-Shop Scheduling with Sequence-Dependent Setup Time Based on a Hybrid Algorithm of Improved Quantum Cat Swarm Optimization," Sustainability (Switzerland), vol. 14, no. 15, Aug. 2022, doi: 10.3390/su14159547.
[12] Y. Lin, W. Zhang, H. Cao, G. Li, and W. Du, "Classifying breast cancer subtypes using deep neural networks based on multi-omics data," Genes (Basel), vol. 11, no. 8, pp. 1–18, Aug. 2020, doi: 10.3390/genes11080888.
[13] A. El-Nabawy, N. A. Belal, and N. El-Bendary, "A cascade deep forest model for breast cancer subtype classification using multi-omics data," Mathematics, vol. 9, no. 13, Jul. 2021, doi: 10.3390/math9131574.
[14] S. Meshoul, A. Batouche, H. Shaiba, and S. AlBinali, "Explainable Multi-Class Classification Based on Integrative Feature Selection for Breast Cancer Subtyping," Mathematics, vol. 10, no. 22, Nov. 2022, doi: 10.3390/math10224271.
[15] A. Dhillon, A. Singh, and V. K. Bhalla, "Biomarker identification and cancer survival prediction using random spatial local best cat swarm and Bayesian optimized DNN," Appl Soft Comput, vol. 146, Oct. 2023, doi: 10.1016/j.asoc.2023.110649.
[16] Y. Chen et al., "MOCSS: Multi-omics data clustering and cancer subtyping via shared and specific representation learning," iScience, vol. 26, no. 8, Aug. 2023, doi: 10.1016/j.isci.2023.107378.
[17] J. J. Chabon et al., "Integrating genomic features for non-invasive early lung cancer detection," Nature, vol. 580, no. 7802, pp. 245–251, Apr. 2020, doi: 10.1038/s41586-020-2140-0.
[18] P. M. Badarudin, R. Ghazali, A. Alahdal, N. A. M. Alduais, and S. A. Mostafa, "Classification of Breast Cancer Patients Using Neural Network Technique," Journal of Soft Computing and Data Mining, vol. 2, no. 1, pp. 13–19, 2021, doi: 10.30880/jscdm.2021.02.01.002.
[19] M. Kang, E. Ko, and T. B. Mersha, "A roadmap for multi-omics data integration using deep learning," Jan. 01, 2022, Oxford University Press. doi: 10.1093/bib/bbab454.
[20] A. Z. Mohammed and L. E. George, "Region of Interest Extraction using K-Means and Edge Detection for DEXA Images," Al-Salam Journal for Engineering and Technology, vol. 2, no. 2, pp. 48–53, Feb. 2023, doi: 10.55145/ajest.2023.02.02.006.
[21] S. K. Mandal, K. Parida, S. Kumar Mandal, S. S. Das, and A. Ranjan Tripathy, "Feature Extraction Using K-means Clustering: An Approach & Implementation," 2011. [Online]. Available: https://www.researchgate.net/publication/334227319
[22] H. Ghazi Enad and M. Abed Mohammed, "A Review on Artificial Intelligence and Quantum Machine Learning for Heart Disease Diagnosis: Current Techniques, Challenges and Issues, Recent Developments, and Future Directions," Fusion: Practice and Applications, vol. 11, no. 1, pp. 08–25, 2023, doi: 10.54216/FPA.110101.
[23] U. Ullah and B. Garcia-Zapirain, "Quantum Machine Learning Revolution in Healthcare: A Systematic Review of Emerging Perspectives and Applications," IEEE Access, vol. 12, pp. 11423–11450, 2024, doi: 10.1109/ACCESS.2024.3353461.
[24] M. Yousif and B. Al-Khateeb, "Quantum Convolutional Neural Network for Image Classification," Fusion: Practice and Applications, vol. 15, no. 2, pp. 61–72, 2024, doi: 10.54216/FPA.150205.
[25] A. Al-Thalej, E. T. Yassen, and B. Al-Khateeb, "Hybrid Quantum Genetic Algorithm for Vehicle Routing Problem with Time Window," 2018.
[26] A. S. Bhatia, M. K. Saggi, and S. Zheng, "QPSO-CD: quantum-behaved particle swarm optimization algorithm with Cauchy distribution," Quantum Inf Process, vol. 19, no. 10, Oct. 2020, doi: 10.1007/s11128-020-02842-y.
[27] K. P. Sinaga and M. S. Yang, "Unsupervised K-means clustering algorithm," IEEE Access, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.
[28] V. Modhukur et al., "Machine learning approaches to classify primary and metastatic cancers using tissue of origin-based dna methylation profiles," Cancers (Basel), vol. 13, no. 15, Aug. 2021, doi: 10.3390/cancers13153768.
[29] A. Mohammed, G. Biegert, J. Adamec, and T. Helikar, "Identification of potential tissue-specific cancer biomarkers and development of cancer versus normal genomic classifiers," 2017. [Online]. Available: www.impactjournals.com/oncotarget/
[30] C. M. Cres, A. Tritt, K. E. Bouchard, and Y. Zhang, "DL-TODA: A Deep Learning Tool for Omics Data Analysis," Biomolecules, vol. 13, no. 4, Apr. 2023, doi: 10.3390/biom13040585.

[31] X. Zhang, Y. Xing, K. Sun, and Y. Guo, “Omiembed: A unified multi-task deep learning framework for multi-omics data,” Cancers (Basel), vol. 13, no. 12, Jun. 2021, doi: 10.3390/cancers13123047.
[32] F. Gao et al., “DeepCC: a novel deep learning-based framework for cancer molecular subtype classification,” Oncogenesis, vol. 8, no. 9, Sep. 2019, doi: 10.1038/s41389-019-0157-8.
[33] M. Wang et al., “Identification of Cancer-Associated Fibroblast Subtype of Triple-Negative Breast Cancer,” J Oncol, vol. 2022, 2022, doi: 10.1155/2022/6452636.