Journal of Biomedical Informatics 107 (2020) 103466

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

A survey on single and multi omics data mining methods in cancer data classification

Zahra Momeni[a], Esmail Hassanzadeh[a], Mohammad Saniee Abadeh[a,b,*], Riccardo Bellazzi[c,d]

[a] Faculty of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran
[b] School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran
[c] Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Italy
[d] IRCCS ICS Maugeri, Pavia, Italy

ARTICLE INFO

Keywords:
Cancer classification
Single and multi omics data
Gene selection
High dimensional datasets
Data integration

ABSTRACT

Data analytics is routinely used to support biomedical research in all areas, with particular focus on the most relevant clinical conditions, such as cancer. Bioinformatics approaches, in particular, have been used to characterize the molecular aspects of diseases. In recent years, numerous studies have been performed on cancer based upon single and multi-omics data. For example, Single-omics-based studies have employed a diverse set of data, such as gene expression, DNA methylation, or miRNA, to name only a few instances. Despite that, a significant part of literature reports studies on gene expression with microarray datasets. Single-omics data have high numbers of attributes and very low sample counts. This characteristic makes them paradigmatic of an under-sampled, small-n large-p machine learning problem. An important goal of single-omics data analysis is to find the most relevant genes, in terms of their potential use in clinics and research, in the batch of available data. This problem has been addressed in gene selection as one of the pre-processing steps in data mining. An analysis that use only one type of data (single-omics) often miss the complexity of the landscape of molecular phenomena underlying the disease. As a result, they provide limited and sometimes poorly reliable information about the disease mechanisms. Therefore, in recent years, researchers have been eager to build models that are more complex, obtaining more reliable results using multi-omics data. However, to achieve this, the most important challenge is data integration. In this paper, we provide a comprehensive overview of the challenges in single and multi-omics data analysis of cancer data, focusing on gene selection and data integration methods.

1. Introduction

Cancer is one of the main causes of death in the world, and nearly 9.6 million people died from January to September 2018 [1]. In general, one out of every six deaths in the world is due to Cancer. Recent advances in molecular medicine have both allowed providing more insights in the basic mechanism of the disease and designing new drugs to complement standard therapeutic strategies, such as surgery and chemotherapy. Therefore, analysing molecular data for prediction of the cancer incidence, survival and the patient's response to drugs, are the major areas of research in bioinformatics, biology and computer science. Over the last two decades, machine learning techniques have been used in omics data analysis. Experiments conducted on omics datasets aim to investigate: (1) the genetic mechanisms of cancer, (2) the presence of different cancer subgroups, (3) the molecular signatures of cancerous and non-cancerous tissues (4) the time course of gene expression in rapidly evolving cancer types. All these research activities mainly focus on providing an interpretation of the results that is able to increase the knowledge about the underlying mechanism of cancer [2].

Omics data usually consists of a large number of features with a small number of samples. Analysis of a data set containing a small number of samples and a large number of features is a challenging task. Such data sets are prone to overfitting, generating non-reproducible results and are affected by high variance. This problem also occurs due to various sources of noise in the underlying omics data. A well-known method of handling these types of data is to select a few features that promise to convey stable information across the experimental conditions. Finding the most effective features among thousands of other ones in a feature selection process is a fundamental challenge in the field of single-omics data analysis.

To reduce data dimensions, there are two common techniques: (1) Feature selection and (2) Feature extraction. Feature extraction

* Corresponding author.
E-mail address: saniee@modares.ac.ir (M. Saniee Abadeh).

https://doi.org/10.1016/j.jbi.2020.103466
Received 11 December 2019; Received in revised form 1 May 2020; Accepted 31 May 2020
Available online 07 June 2020

techniques produce new features with lower dimensions than the main features. These newly created features are a linear or non-linear combination of the main features. Some example of extraction techniques are Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA), Canonical Correlation Analysis (CCA) and autoencoder deep learning feature reduction. Compared to the feature extraction techniques, feature selection methods select a subset of features (without any change in the meaning of the main features) with the least redundancy and the most relevance to the target class in order to obtain the highest classification accuracy. Tibshirani [3], Guyon [4], Diaz [5] and Peng [6] are among the pioneers of gene selection in the cancer classification. The main goal of both techniques is to improve classification accuracy with minimum computational cost. In feature extraction, the original feature space is mapped into a new feature space with lower dimensions. As a consequence, in most cases the resulted classification model from the aforementioned new feature space suffers from lack of interpretability. Feature selection methods are often preferred over feature extraction techniques for dimensionality reduction in single-omics data pre-processing.

The correlation between genotypes and phenotypes (e.g., cancer) of a living organism is not a simple one-to-one relationship, but rather they participate in a complex set of interactions between different types of molecular mechanism [7]. As such, in order to achieve accurate results along with a better understanding of the disease process, the direction of research in this area has been switched from single-omics data analysis to simultaneous use of multiple omics datasets. A fundamental challenge of multi-omics data analysis is data integration. A proper data integration method in multi-omics data analysis is a method that considers the set of complex interactions between genetic products, i.e., it achieves stable and accurate results when analyzing the relationships between various omics data. Gevaert [8] and Meng [9] are among the pioneers in this area of research.

Fig. 1 depicts the relationship between single- and multi-omics data summarizing most important challenges of their analysis. Despite the fact that existing literature reviews [2,10–13] provide a clear description of data integration methods for multi-omics data analysis, unfortunately none of them addressed the importance of the feature selection process in multi-omics data analysis, the effect of this process on integration methods as well as the choice of method. To the best of our knowledge, we are the first to review single- and multi-omics data analysis challenges simultaneously.

The main reason behind discussing single- and multi-omics data in a single review paper is that there are important challenges in genomic data analysis. Specifically, the choice of an integration approach often depends on two contributing factors: (1) the used feature selection method and (2) the integration stage after the feature selection process is performed. Choosing a specific integration process (e.g., early integration) sometimes makes it very challenging to choose a powerful feature selection method (e.g., wrapper-based). In other words, selection of an integration technique enforces the use of a specific feature selection technique. Since in multi-omics data, features are often genes, we use “gene selection” instead of “feature selection” in rest of this paper.

This paper is organized as follows. Section 2 presents an overview of gene selection methods for single-omics data analysis. Section 3 overviews data integration techniques for multi-omics data analysis. In Section 4, we summarize the challenges found in single- and multi-omics data analysis. Section 5 reviews datasets exploited so far in the field. Finally, Section 6 concludes this paper.

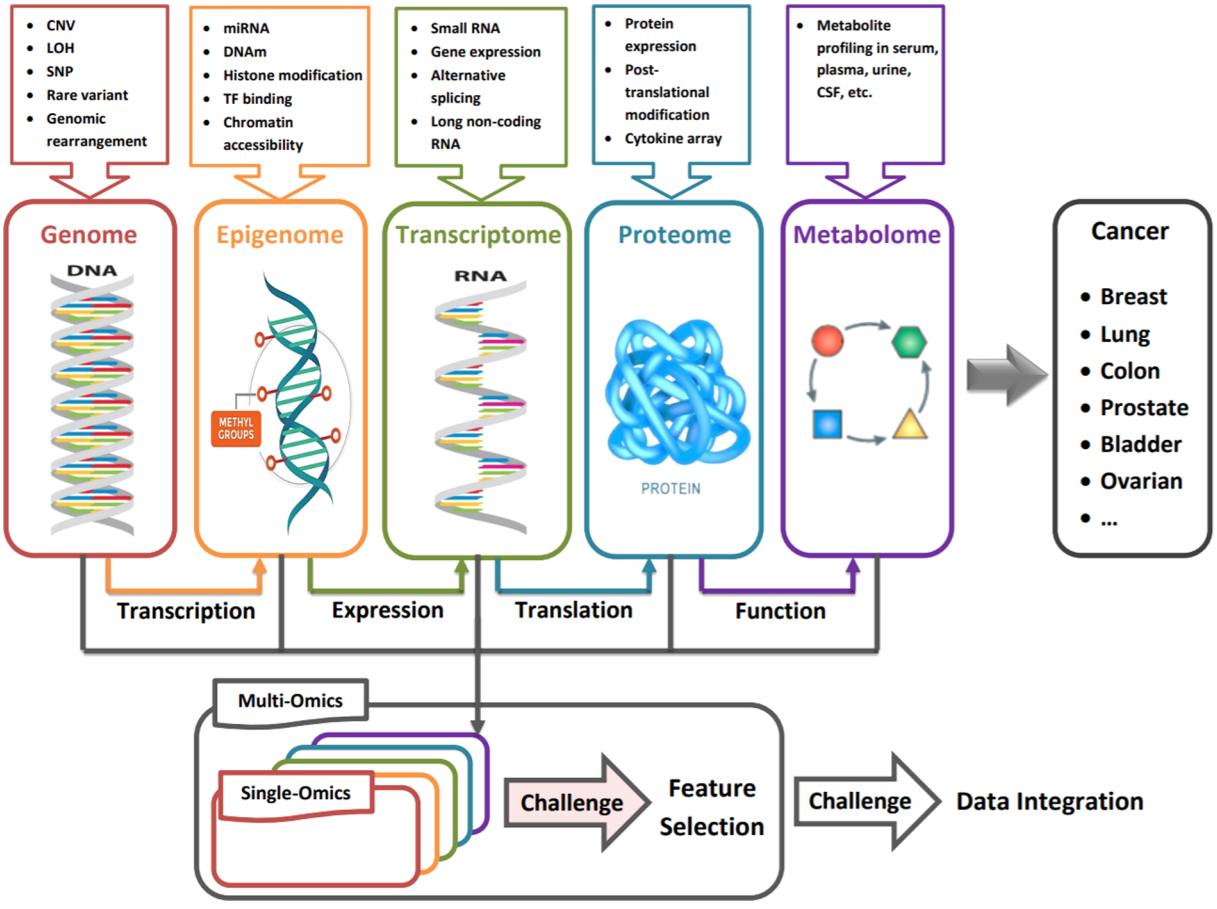


Fig. 1. The overall diagram of relationship between single and multi omics data analysis challenges.

2. Review on gene selection methods

There are ~ 22,000 protein coding genes and many regulator elements in the human body, not all of which are in the causal pathways of cancers, and only a few are involved in the management of the disease and its effects on human body. Thus, identifying genes involved in a specific cancer (such as breast cancer) may help in analysing the cancer disease.

In the gene selection process, we look for a subset of informative genes in the main dataset that can effectively describe the input data, providing a reasonable predictive performance. Gene selection techniques can provide a better understanding of the problem, reduce the computational cost, and increase the accuracy of prediction. Omics data, such as gene expression, may include hundreds of attributes that can be correlated with each other. Due to the fact that the relevant features do not have any additional information about classes, they are considered as noise for predictive models. In fact, the information can be obtained with a smaller number of unique features that contain more information about the distinction between classes. After removing the features, the volume of data is reduced, possibly resulting in improvement of the performance of classification.

During the selection of best feature sets, various subsets may be obtained. An optimal subset contains all the features that are not redundant and are completely interrelated. The presence of fully relevant features is necessary to increase the detection capability and accuracy of prediction [14,15].

2.1. Gene selection approaches

In general, gene selection has three approaches: supervised, unsupervised and semi-supervised. Supervised gene selection is the most commonly used approach. This approach uses labelled data in the process of gene selection. In this approach, considering the class feature, the most important and most separable features are selected. In cancer data, the record class can indicate whether or not the sample is cancerous, or may indicate a subtype of cancer. Therefore, the goal is to find the best genes that are effective in separation and identifying this class. However, the challenge is to label data by using external knowledge sources. The labelling process is costly and may not be completely reliable. The unreliability of labels due to the unwanted deletion of related features, or the selection of unrelated features increases the risk of over-fitting of learning process.

There is no information about the data label in unsupervised gene selection. Therefore, in order to select the best subset of features, additional information such as distribution, variance, and separability of data can be used. Without labeling, the data does not require the aid of an expert or an external knowledge and can still work well even when no prior knowledge is available. The term semi-supervised gene selection is used when some data is labelled and some unlabelled. Labelled data is usually used to maximize the margin between data points of different classes, and unlabelled data is used to explore the geometric structure of feature space.

2.2. Gene selection methods

In this section, we investigate ca. 300 papers from Scopus in the field of cancer classification and gene selection published in recent years. We select 100 of those according to the relevance of the subject of this research, and disregard others. Fig. 2 shows the exact query as well as the criteria for searching and filtering the papers in Scopus.

The selected papers are categorized according to the methods of gene selection. Similar to prior studies, the fraction of papers we found in the field of semi-supervised approaches were noticeably small. As such, we did not consider papers that applied semi-supervised approach in this review paper. Fig. 3 depicts the hierarchy of papers considered.

As discussed earlier, gene selection is the most challenging task in single- and multi-omics data analysis. It is also one of the most important steps in data pre-processing that is done before modeling. Gene selection methods are classified into five categories, as will be discussed next.

2.2.1. Filter approaches for gene selection

Filtering methods often use statistical methods to select features with a low computational cost. These methods do not use any kind of classifier or learning algorithm. A filter algorithm mainly measures the characteristics of features based on four types of evaluation criteria: Dependency, Information, Distance, and Compatibility. These methods are known to be more efficient and faster than wrapper methods (see below). Therefore, using filter approaches for gene selection is a wise choice particularly for huge datasets. Filtering methods are divided into uni-variate and multi-variate methods. Multi-variate methods have the ability to find relationships between several features, while uni-variate methods consider each feature individually.

Unsupervised gene selection is a more difficult task given the lack of class labels. Tabakhi et al. [91] presented an unsupervised gene selection method based on ant colony optimization, called UFSACO. This method seeks to find a subset of optimal features through several iterations. In addition, the relevance of features is calculated based on the similarity between the features, which leads to minimization of redundancy. Therefore, it can be categorized as a multi-variate filter-based method. The proposed method has a low computational complexity, which makes it a perfect choice for high dimensional datasets.

Most gene selection methods are supervised and use class labels as a guide. For example the method presented by Mohammadi et al. [25] uses a Maximum–Minimum Corr-entropy Criterion Approach (MMCC) to select useful genes from the microarray dataset. This method is stable, acts fast and it is robust to noisy data as well as the high diversity problem in data and outliers. There are also methods that first reduce the dimensions of feature space and then apply filter methods. For example, Xu et al. [23] presented a method that reduces feature space by leveraging a conventional method to reduce data dimensions, called PCA, at its first step. In the second step, a correlation-based filtering method along with a threshold is employed to exclude improper features from the subset of features. It should be noted PCA changes the meaning of features, which can be a problem when analyzing microarray data.

New forms of filtering techniques are methods used as the fitness function of a meta-heuristic algorithm. As an example, Zheng and Wang [31] presented a new method for selecting features called FS-JIME. In this paper, a meta-heuristic algorithm, namely BPSO, is used to find the optimal subset of features.

2.2.2. Wrapper approaches for gene selection

Wrapper methods use a predictor in a black box setting. The efficiency of predictor is used as a target function to evaluate the subset of selected features. This method selects a subset of discriminative features by minimizing the prediction error of a particular classifier. Wrapper methods are not as popular as filter methods due to their higher computational cost. As the number of features increases, the space of existing subsets of features expands exponentially. This becomes a critical aspect when the number of features reaches the order of ten thousand. Furthermore, these methods have the potential to over-fit on data with a small number of samples. In general, wrapper methods are divided into two groups: Sequential selection methods and Heuristic search algorithms. Examples follow.

- Sequential selection methods begin with an empty (or full) feature subset. Then, before reaching the maximum value of the target function, they add (or remove) some features, so that the value cannot be increased anymore. A criterion is chosen to speed up the selection process. The number of selected features increases gradually so that, with the least number of features, the algorithm reaches

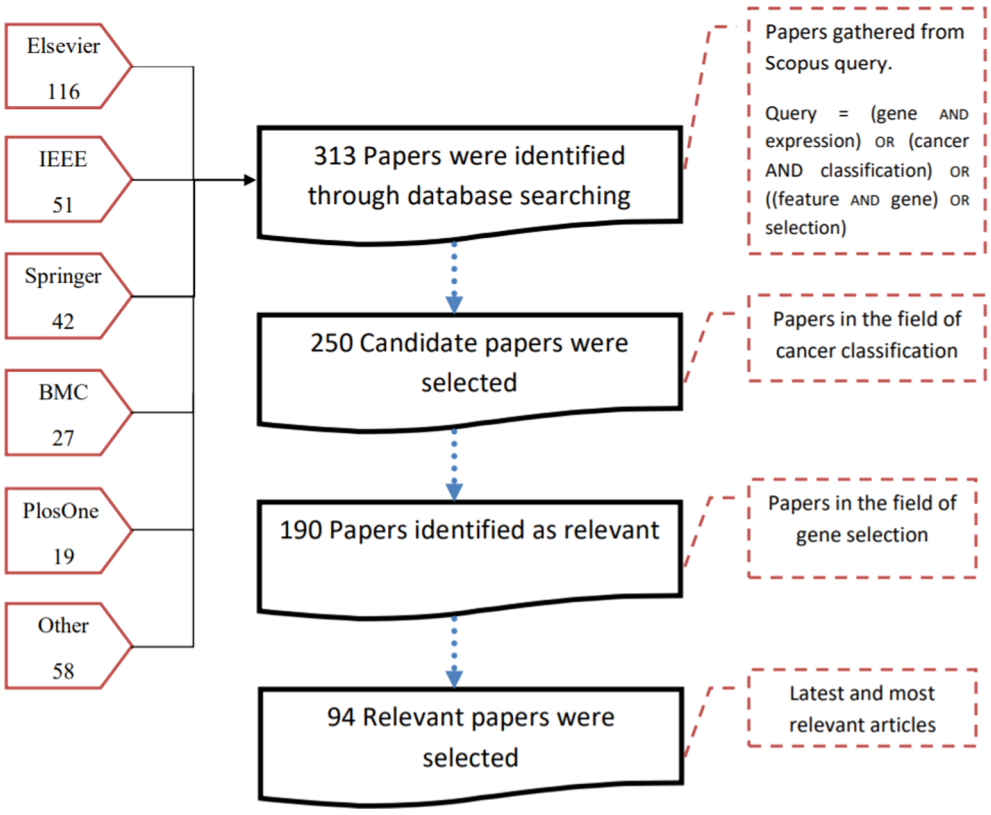


Fig. 2. The paper selection process in gene selection.

to the maximum value of the objective function. Wang et al. [34] proposed a method to accelerate the sequential wrapper methods. In this method, a classifier distance matrix is used to reduce the computational complexity of sequential wrapper methods.

- Heuristic search algorithms evaluate different subsets to optimize the target function. Different subsets are created either through searching in a search space or by creating a solution to an optimization problem. As an example Moradi and Gholampour [41] presented a new algorithm based on the particle swarm optimization algorithm, so-called HPSO-LS. In this method, a local search strategy that optimizes a particle swarm is used to select a subset of effective features that have less correlation. The purpose of the local search technique is to guide the search process for particle swarm optimization to select distinctive features according to their correlation information. Additionally, it uses the KNN classifier to evaluate a subset of candidate features. Wang et al. [39] proposed a weighted gene selection strategy, which specifies features based on their classification performance as well as the frequency of occurrence in the population, based on two matrices. The goals of this approach are to minimize the number of features, maximize performance, and minimize computational costs. In addition, in this method, a gene selection algorithm based on bacterial colony optimization is proposed to reduce computational complexity and also improve search capability even in discrete optimization problems. The best combination of features can considerably enhance the functionality of the classification. In another example, Pati et al. [36] proposed a method based on improved genetic algorithm. The fitness function of this algorithm is based on the selection of the least number of genes having the highest classification accuracy, which is measured by the performance of a SVM classifier.

2.2.3. Embedded approaches for gene selection

The embedded method is a gene selection mechanism in which gene selection is performed during the execution of the learning algorithm. The embedded method is more efficient and computationally less complicated than wrapper method, while retaining a similar performance. This is because of the fact that the embedded method avoids duplicate execution and examines each feature subset in the learning process of the algorithm. The widely used LASSO[1] method [3] places a limitation on the sum of model parameter values in that their sum should be smaller than a constant value. To do this, LASSO method enters a setup process in which some coefficients of the regression variables are set to zero. During the feature selection process, variables that still have non-zero coefficients are selected as final attributes for the modelling process. Although LASSO is a very efficient method, its great disadvantage is that it tends to over-regularize the model. As such, ELASTIC NETs [98] address this problem by establishing a balance between LASSO and RIDGE penalties.

The most popular embedded method is SVM-RFE, introduced by Guyon et al. [4]. This method uses a SVM as a classifier. In the SVM-RFE, the model is a linear equation that separates the samples of the two classes. Thanks to an iterative backward selection approach, the least-weighted features of the SVM equation are removed each time, because they are supposed to have really less effect on the separation of the two classes. It should be noted that the number of features that are eliminated in each round and the number of features that are eventually remain are predetermined.

In recent years, Guo et al. [20] presented a method that used a class separability criterion (suitable for multi-class data) to select the features. Their proposed approach is a general gene selection method in which each feature is scored based on a criterion. Subsequently, features with a higher score are selected.

2.2.4. Ensemble approaches for gene selection

An ensemble method searches for a group of best feature subsets by

[1] Least Absolute Shrinkage and Selection Operator

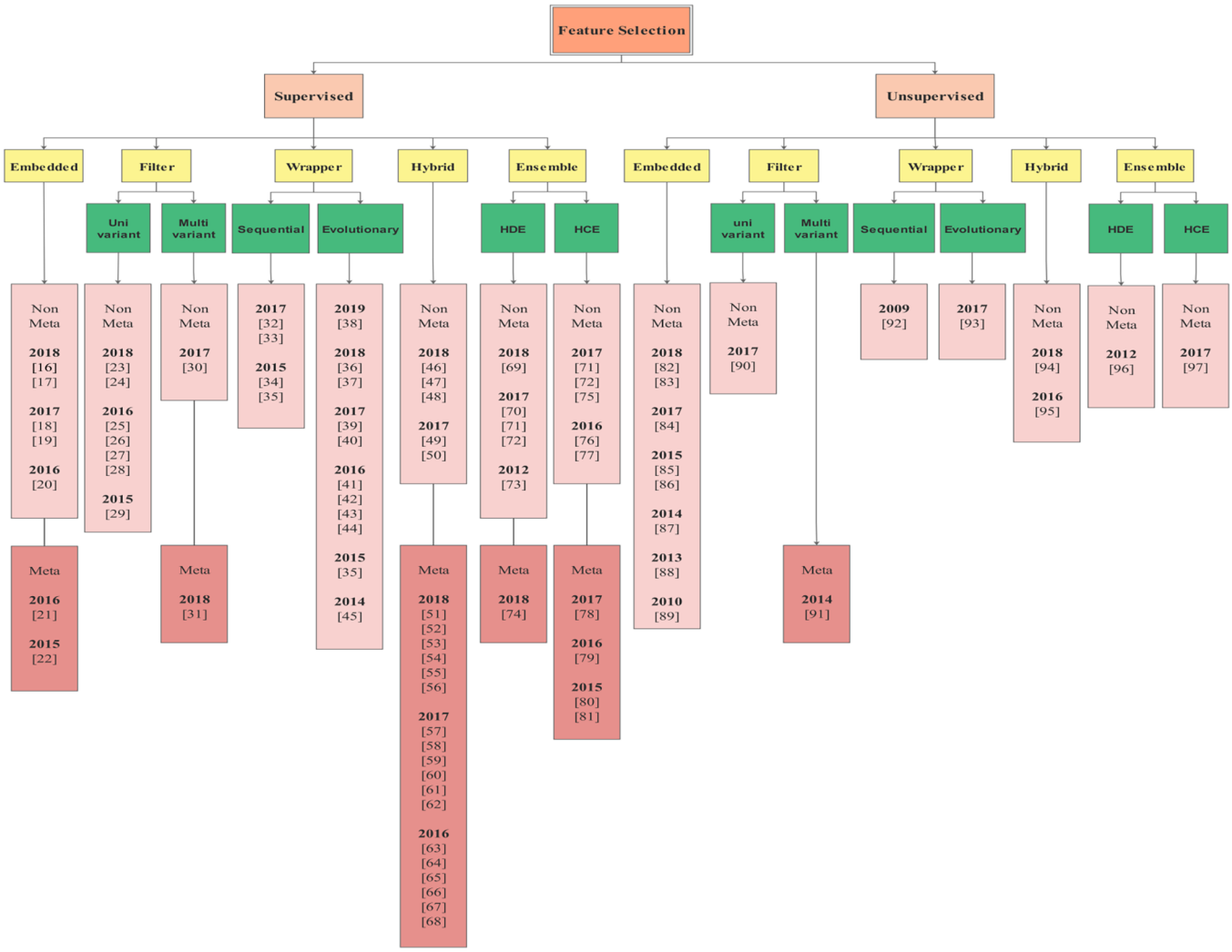


Fig. 3. Tree of papers in gene selection. (See above-mentioned references for further information.)

relying on different feature selection strategies, and then produces a merged result from these groups. The idea is to derive more stable feature sets, less dependent on the data as well as the design parameters of algorithms. Two common approaches [72] of ensemble methods can be found in the literature:

- **Homogeneous distribution:** According to Fig. 4, in this approach, the dataset is divided into n parts and distributed between n nodes to run in parallel. In all nodes, a single gene selection method is executed. Finally, the rankings derived from each node are combined using ensemble methods. For example, Pes et al. [70] initially reduce the dataset into a number of smaller subsets with placement sampling. Then, on each data sample, a ranking algorithm is executed to obtain a subset of features that have higher ratings. Finally, using one of the integration methods, the final best feature subset is obtained. In another example, Ebrahimpour et al. [74] proposed a

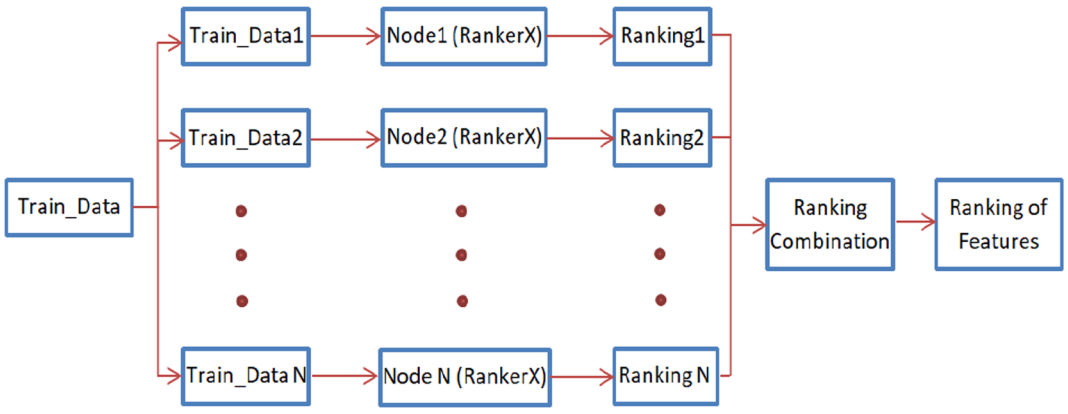


Fig. 4. Homogeneous distribution.

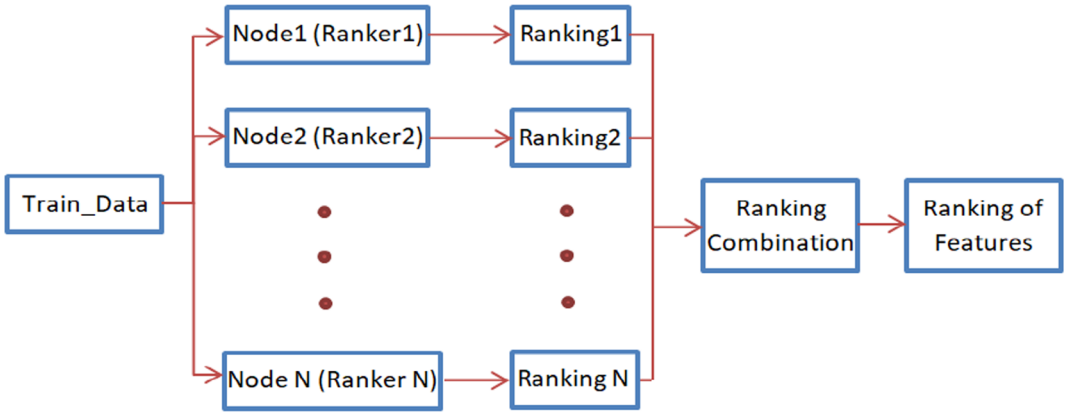


Fig. 5. Heterogeneous Centralized.

feature selection algorithm to handle the high dimensional space of the gene selection problem. The CCFS method first divides the dataset into a number of subsets using a random method (on the features). For each of the generated subsets, an evolutionary algorithm runs with a filter criterion as a fitness function to optimize each of the subsets. The calculation of the fitness function of the evolutionary algorithm is based on the entire problem (all features).

- **Heterogeneous Centralized:** In this approach, depicted in Fig. 5, an identical gene selection algorithm with different parameters or several gene selection algorithms are applied to the original dataset, each of which finally generates a subset of the best genes as output. The results generated from each of the gene selection algorithms are combined in the next step with the help of aggregation methods. In this way, the strengths and weaknesses of the unique gene selection methods can be exploited. For example, the method presented by Mohapatra et al. [99] normalizes the dataset as its first step using the max–min normalization method. Then, using the modified cat swarm optimization (MCSO) algorithm, a subset of optimal set of features is obtained from the normalized dataset. Then the MCSO extracts ten subsets of 10–100 genes in a ten-fold interval (first time 10 genes, second time 20 genes … and tenths of 100 genes from each subset). To obtain the most desirable candidate features in each of these subsets, the KNN classifier is used to find the classification accuracy. Finally, the subset with the least number of features and the highest classification accuracy is selected as the optimal subset of the candidates.

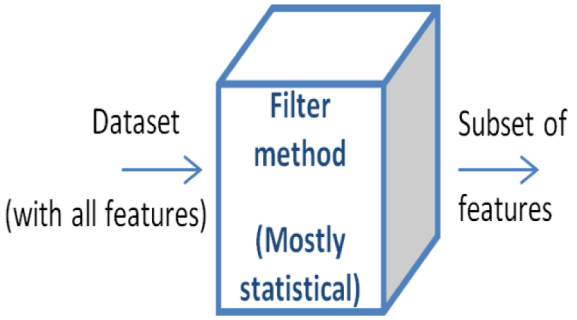
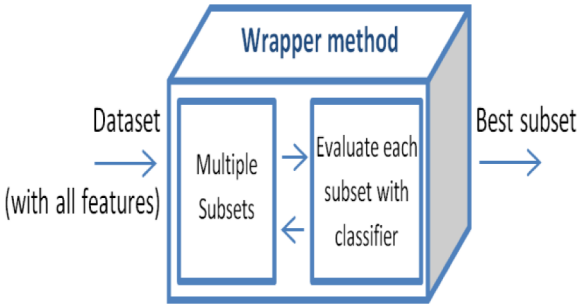
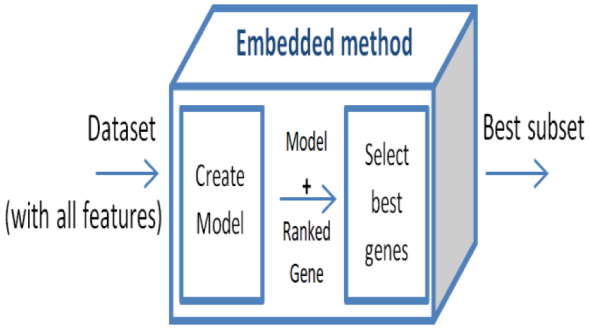
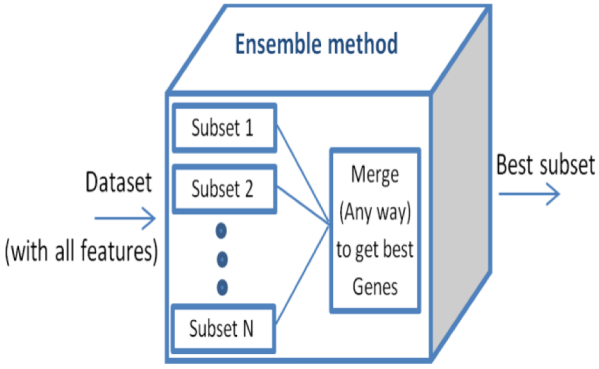
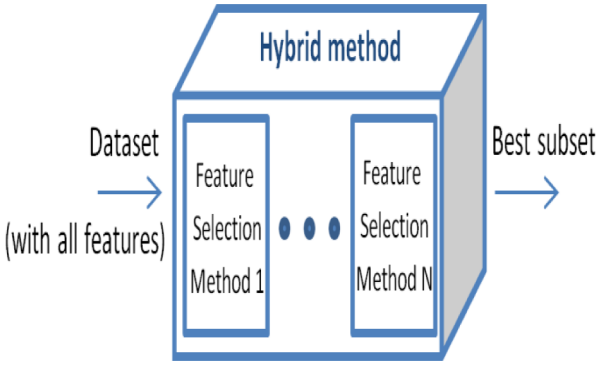
Elyasigomari et al. [81] developed a new hybrid optimization algorithm, namely COA-GA, leveraging the recently discovered cuckoo optimization algorithm as well as a traditional genetic algorithm trend, which is used for data clustering and selection of popular genes. First, the clustering algorithm is executed six times using the combined evolutionary algorithm, the COA-GA, and each time the number of clusters increases from one to six in 100 iterations. The fitness function of this evolutionary algorithm is defined by reducing the distance within the examples of the same cluster and increasing the distance between the examples not in the same cluster. After using the COA-GA algorithm in each run, 20 subtypes of the attribute are obtained; at the end of the implementation of the sixth, 120 subsets are obtained. To select 25 of these features, 120 of these subsets are used to score points based on the number of times a gene has been presented in the last 6 subsets. Finally, genes are sorted based on their score and the 25 best genes that have the highest scores are selected.

2.2.5. *Hybrid approaches for gene selection*

In hybrid methods, two or more gene selection algorithms (filter, wrapper, embedded or ensemble) are combined in a certain order. The hybrid method inherits the benefits of each individual method combined. A hybrid method uses different evaluation criteria at different stages of the search to improve the efficiency with better computing performance.

The most common hybrid method is a combination of filter and wrapper methods. For example, in [63], the advanced mRMR filter algorithm uses pre-selection features. Lv et al. [63], defined two criteria: increasing the accuracy and reducing the number of features. As the first criterion is relatively more important, they design a multi-objective model called MOEDA. MOEDA is a type of distribution estimation algorithm. The search process guides to find the best candidate. Elyasigomari et al. [57] presents an approach where genes are selected using mRMR in the second stage. In this approach, a combination of a Cuckoo optimization and Harmony Search algorithms are implemented as a wrapper part. Finally, both papers use a support vector machine as the base classifier. As another example, Lu et al. [58], combines mutual information maximization (MIM) as a filter component and genetic algorithm as a wrapper component to take advantage of both. Jain et al. [51] used a hybrid method called CSF-iBPSO to select features. The filter method used in this paper is the CFS method, which is a multivariants filter algorithm. Then, the subset of features selected by this algorithm is given by the improved BPSO-NB algorithm to construct a model. Dashtban et al. [52] initially used a single-variable filter method called Fisher score to rank genes, and subsequently select the 500 genes with the highest score. It should be noted that the Fisher score is actually an assigned number to each gene, representing how much the gene is capable of differentiating between different classes. The subset of selected features from the first stage is given to the wrapper algorithm to select the final subset of classes with the highest degree of separability. The wrapper method used in [48] is a variant of the traditional BAT algorithm. Venkataraman et al. [53] proposed a two-step method. First, it selects a few related features by computing the Symmetric Uncertainty (SU) values between the features and the class. Next, a genetic algorithm is used to find the optimal subset of features. The SU between features and classes is used to obtain the best features for classification. Features that have a larger SU value will gain more weight and are more likely to be selected. In [54], a combination of ensemble and wrapper methods are used for gene selection. Lai [54] firstly used an ensemble of chi-squared, correlation, gain ratio, information gain and relief methods to select the most informative genes. Then in the second step, the genes selected from the previous step are given to the multi-objective simplified swarm optimization (MOSSO). Agarwalla and Mukhopadhyay [55] also proposed a bi-stage hierarchical swarm based gene selection technique which combines two methods. In the first stage, a multi-fitness discrete particle swarm optimization (MFDPSO) is applied. This stage uses multi-filtering based gene selection method. In the second stage a new blended Laplacian artificial bee colony algorithm (BLABC) is proposed and it is used for automatic clustering of the selected genes obtained from the first stage.

Table 1
Comparison between gene selection methods.

Method	Shape	Advantage	Disadvantage
Filter		Fast and scalable Independent of classifier Faster than wrapper methods Better computational complexity than wrapper methods	Less accuracy due to non-consideration of the classifier Ignores the relationship between features / variables. May involve redundancy.
Wrapper		Interact with classifier Interaction between features Better accuracy than the filter methods	Overfitting High computational complexity Expensive operation Tends to local optimization
Embedded		Higher Accuracy and efficiency than filter methods Less computational complexity than wrapper methods More focus on relationship between features	Dependent to classifier
Ensemble		Less tendency to overfitting More scalability for high-dimensional data sets Stability	Understanding the combination of classifiers is difficult
Hybrid		More efficient than filter methods Less tendency to overfittingLower computing cost	Dependent to classifier Depending on the combination of different gene selection algorithms

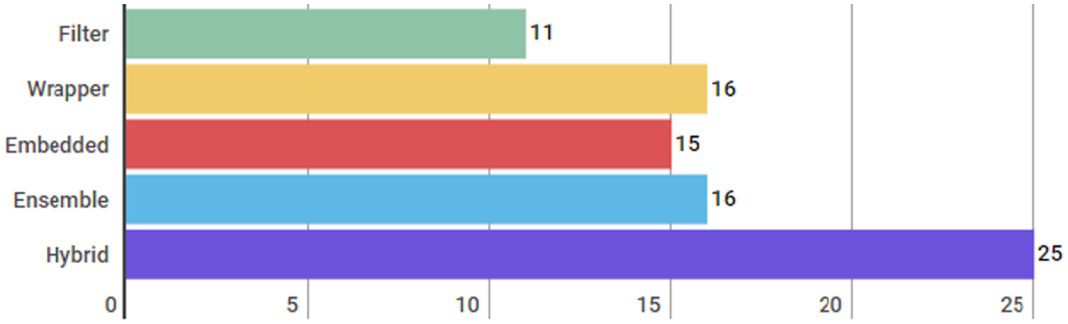


Fig. 6. The number of papers examined in this paper in field of gene selection.

Table 1 provides an overall comparison of different gene selection methods previously studied.

Fig. 6 summarizes the papers examined in this section in accordance to the gene selection methods used. According to the figure, the most popular methods for gene selection are hybrid approaches. This popularity is because hybrid gene selection approaches benefit from the advantages of both filtering and wrapper methods as well as their high implementation speed.

2.3. Stability

As discussed in the previous section, there are a variety of gene selection methods that can be applied to a dataset. If the underlying problem is related to classification, on the one hand, the best method is the classification algorithm that achieves the highest classification accuracy by selecting the lowest number of features. On the other hand, one of the problems that is usually neglected is the stability of results in the gene selection algorithms, i.e., the selected gene should not be changed by changing the training data set or by minor changes in the design parameters of the gene selection algorithms. The stability of a gene selection algorithm can be proved by observing if the algorithm chooses almost the same subset of features when different subset of records are selected in the training dataset. Among the methods mentioned in the previous section, and according to their implementation technique, ensemble methods naturally rank first in terms of stability. For example, Castellanos-Garzón et al. [46] used a five-step method to provide a stable gene selection. Each step filters and passes the data into the next one. In the first two steps, the data is preprocessed to eliminate outliers and noisy data from the dataset. In the third step, several filtering methods are applied in parallel (ensemble of filters) to the dataset, each of which acquires a subset of features, and at the end of this step all of these features sets that come from different filtering methods are gathered and a union set is obtained. In the fourth step, the selected features in the previous step are given as input to forward gene selection and backward gene selection wrapper algorithm. Several classifiers are also used as evaluation functions in these feature selection algorithms. In this step, by combining each selection algorithm and classifier as evaluation function for feature selection algorithm, a subset of features is obtained that maximizes the classification accuracy. Finally, all feature subsets that were selected at this step are merged and entered as the input of the fifth step to the next step. In the fifth step, the goal is to find a subset of stable genes that results the highest accuracy for all classes. Note that this algorithm is clearly a costly method.

3. From gene expression to multi-omics: a review of data integration methods

In this section, we downloaded ca. 103 papers published in recent years from Scopus. Each of these papers has been carefully examined. We select 61 papers out of those 103 according to the relevance of the topic of papers. Fig. 7 shows the exact query used in Scopus for searching and filtering the papers.

Fig. 8 depicts the hierarchy of the papers considered. Multi-omics data integration methods are divided into three categories, which will be reviewed in the following sections.

As we reported in the introduction, analysis of multi-Omics data takes place in order to better exploit the variety of data that can be collected to represent the interactions between genotypes and phenotypes of a living organism and give more and accurate facts as outputs. The biggest challenge for analyzing multi-omics data is data integration.

As discussed earlier the data integration is most challenging problem in multi omics data analysis. First we should decide about integration mechanism. Then the selected integration method is used for multi-omics data analysis. We classify integration approaches in three major groups: concatenation-based integration, transformation-based integration and model-based integration, as discussed next.

3.1. Concatenation-based integration (early integration)

In the concatenation-based integration methods, known as “early integration”, the matrixes of omics data are combined and form a large data matrix before the model is constructed. An advantage of an early integration is that after determining how to combine all variables in a matrix, it is easy to use a variety of machine learning methods for analyzing continuous or discrete data. As such, these methods are theoretically powerful because the selected machine learning algorithm can find any relationship between the features of the concatenated multi-omics data. Early integration methods have various computational frameworks, such as Bayesian networks, i-cluster methods, etc.

Manzalawy et al. [100] used CNA, methylation data and RNA-Seq to predict survival in patients with ovarian cancer. They have proposed a two-stage hybrid method for feature selection process. In the first stage, several feature selection methods (filters such as chi2 and embedded such as lasso) are defined that one of them is activated by a signal and separately applied on each omics data. In the second stage, a MRMR filtering method (which is one of the feature selection methods in the filter feature selection category, and is explained in section 2) is applied to the selected features in the previous stage. Finally, the features selected from each of these data were merged into a final consolidated matrix for the learning process.

3.2. Transform-based integration (intermediate integration)

In this method, each dataset is first converted to an intermediate form, such as a graph or a core matrix. The transformed core or graph is subsequently used for modeling. One of the advantages of this method is that if data has a single variable as an identifier for linking a patient's data types, it can integrate either continuous or discrete data types. Transform-based methods include statistical frameworks such as core-based integration and graph-based semi-supervised for building a model. Gade et al. [101] is one of the first papers in this category of

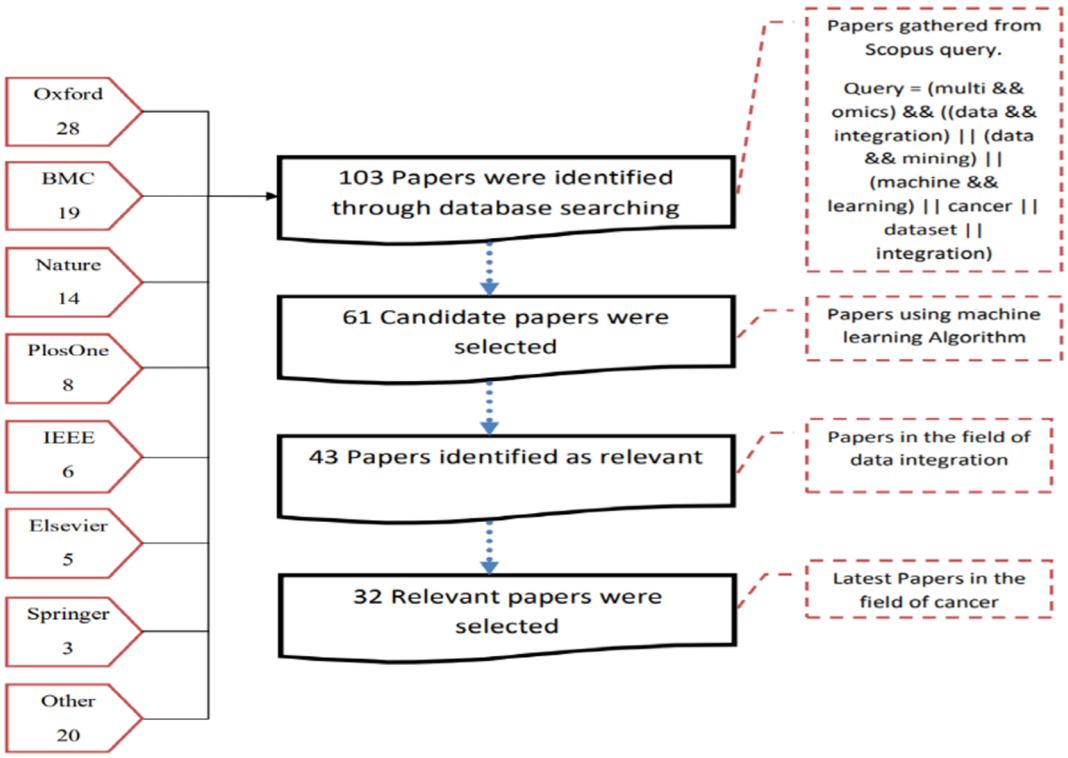


Fig. 7. The paper selection process in multi-omics data integration.

integration methods. This study used mRNA and miRNA expression changes to improve clinical outcome prediction in prostate cancer. Since miRNAs are known regulators of mRNAs they used the correlations between them as well as the target prediction information to build a bipartite graph representing the relations between miRNAs and mRNAs. This graph is used to guide the feature selection process in order to improve final prediction accuracy. Using the above-mentioned method enabled Gade et al. [101] to significantly improve the prediction performance as well as the stability of the feature selection process.

By searching for the most authoritative papers, Peng et al. [102] gathered well-known biomarkers known in bladder cancer as seeds in different omics data types. They considered these biomarkers as a positive class (true positive) in classification of genetic factors in different omics data, and attempted to find other effective biomarkers in bladder cancer using binary classification. They used CNV, methylation, gene expression and miRNA omics data in their study. Initially, the authors calculated the Pearson correlation of features in each omics data with the collected seeds. The mostly correlated features are selected while disregarding the remaining features. For each given omics data, a graph is constructed. The nodes of the graph express the features of omics data and its edges expresses the correlation between the two nodes. Then regulatory relationships between gene expression and the other three omics are further analyzed by linear regression model. The corresponding coefficients are then utilized to weight the edges connecting the networks of different omics. In this way, the heterogeneous network model of genes is constructed. Finally, a modified propagation algorithm is implemented on the model to identify BC-related genes. In this process, the information flow propagates from seeds to candidate genes iteratively and a score is obtained for each candidate gene by the end of propagation process. They finally considered top-k gene as the most effective bladder cancer genes. To evaluate the performance of the proposed method, they performed leave-one-out cross-validation (LOOCV) in the test process. In each round, the authors take one seed as test data and all other seeds as training data and validated their study.

There are also examples of papers that used deep learning. For example, Zhang et al. [103] used CNA and gene expression omics data to predict the type of cancer in Neuroblastoma. They adopt an auto-encoder deep learning algorithm to integrate multi-omics data. The so-called deep learning algorithm is then combined with K-means clustering to identify two cancer subtypes with significant survival differences. They validated the final classification algorithm using two independent datasets. Sharifi-Noghabi et al. [104] used copy number aberration, somatic mutation, and gene expression data for multi-omics integration. They used deep learning for each input omics data to learn features for each omics type. The learned features integrated into one representation network by concatenation. The concatenated network serves as input for final classification task.

As another example, Rakshit et al. [105] used data from mRNA, methylation and miRNA to detect breast cancer subtypes. They used an autoencoder to transform the features into a smaller feature space. Finally, the final matrix is given to the five classifiers for learning process.

Another set of papers such as [106–108] used the matrix factorization method to integrate multi-omics data. This method is shown in Fig. 9. It focuses on reflecting the variation in dataset by mapping it to a reduced dimensional space (a core matrix). In fact, this method looks for low-dimensional representations of high-dimensional data and attempts to find hidden common factors in the data.

3.3. Model-based integration (late integration)

Model-based integration is not limited to the type of data. In this method, integration is performed to identify interactions between different levels of omics data associated with a specific disease or phenotype. More specifically, a number of models are first constructed using several omics data as training datasets. The final model is then built using the constructed models from the previous stage and a reasoning strategy. The model-based approach has several computational frameworks for constructing the final model, which includes, among others, majority voting, ensemble approaches, or probabilistic causal networks. One of the benefits of model-based methods is that these methods can integrate predictive models of different data types.

Kim et al. [109] used CNA, methylation, gene expression and

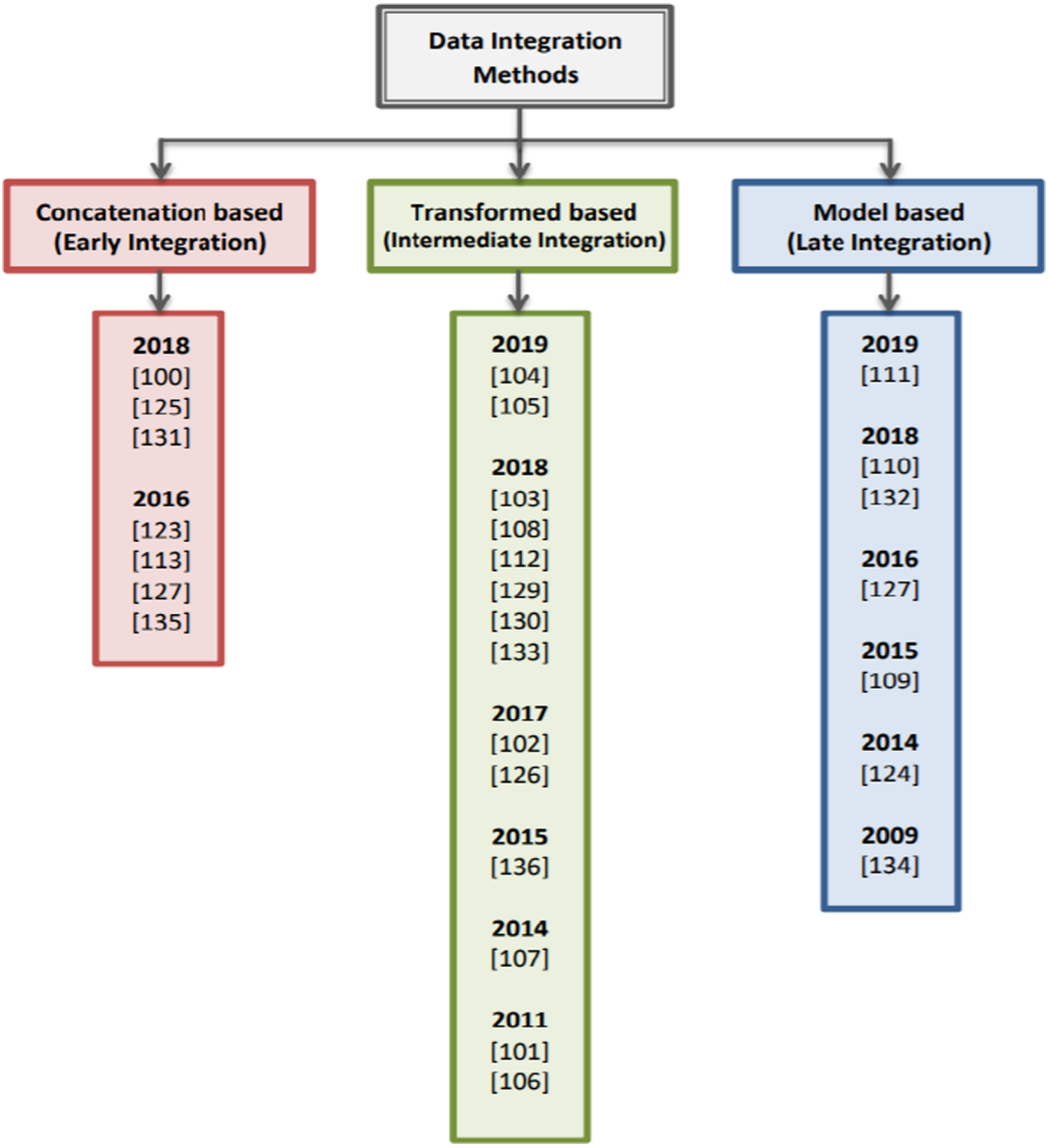


Fig. 8. Tree of papers in data integration.

miRNA omics datasets for classification task in ovarian cancer. They used the clinical information from ovarian cancer patients to formulate a binary classification problem. Three sets of classifications were defined as follows based on the clinical outcomes of ovarian cancer: (i) early stage or late stage, (ii) low grade or high grade, and (iii) short-term (less than 3 years) or long-term (at least 3 years) survival. They built a graph model for each omics data type using a similarity matrix calculation. Each patient is a node in this graph, and it has a value of 1 or -1 corresponding to each class, respectively. Nodes are connected together with labelled edges. The label is a number indicting the similarity of the two patients based on their genetic values. For each patient data, its similarity metric is calculated and its class is predicted according to the score obtained in the model. As predicting clinical outcomes by only leveraging a unidimensional genomic data set is not accurate enough, Kim et al. [109] used a method to integrate models via finding optimum combination of classifiers coefficients.

As another example, Yang et al. [110] used kernel fusion with a genetic algorithm to tune the kernel parameters so as to integrate CNV, mRNA and miRNA omics data for the diagnosis of breast cancer subtypes. Tao et al. [111] used multi-omics data for breast cancer subtypes prediction. The authors employed different kernels (Linear, Gaussian, and Polynomial) to generate kernels of several SVMs using multiple kernel learning (MKL) algorithm on CNV, mRNA and methylation omics data separately. MKL is a multi-omics fusion algorithm that use l-norm regularization can lead to a sparse solution.

In this paper, we classify the data integration approaches into three general categories, together with their pros and cons, as summarized in Table 2.

Fig. 10 presents the papers considered in accordance to the data integration methods used in this review paper. The most popular method of data integration is transform-based methods that are in turn based on the creation of intermediate graphs and the consideration of the relationships between different levels of genetic products.

Multi-omics data analysis is one of the recent and challenging problems in bioinformatics. Although the most important challenge in multi-omics data analysis is data integration, gene selection is still a challenging problem. This means that selecting an integration technique largely depends on the dimensions of the problem which needs to be addressed. Table 3 represents some papers in which gene selection plays an important role in improving the final classification performance.

As discussed above, one of the important issues in multi-omics data

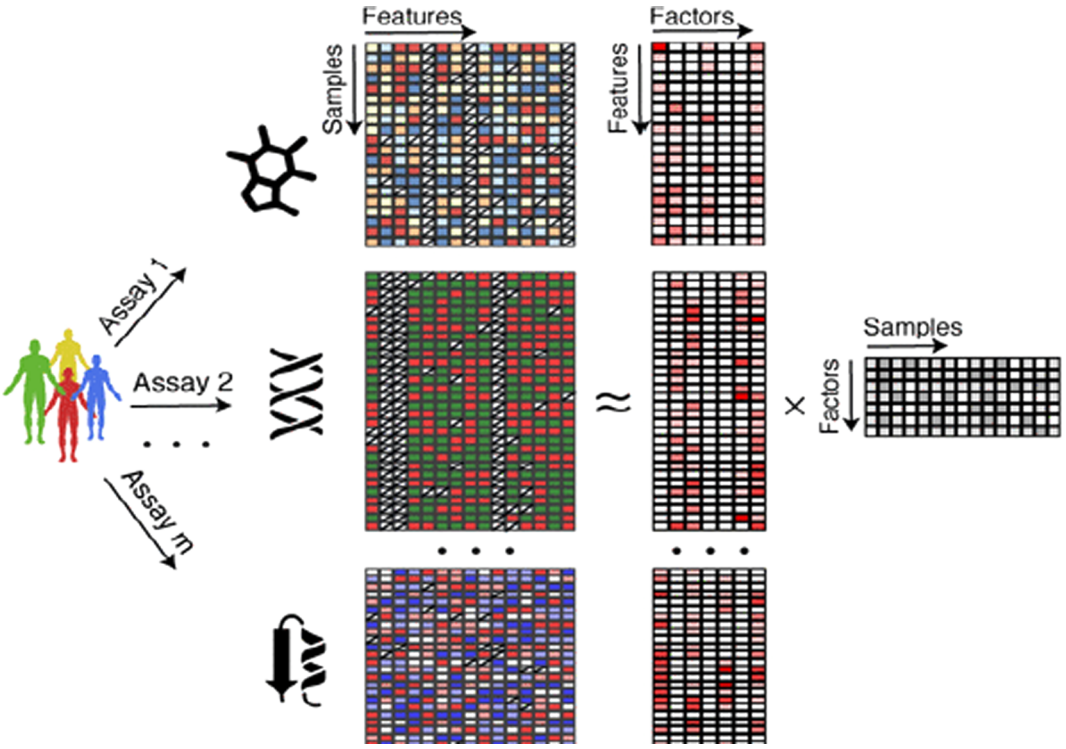


Fig. 9. Matrix factorization method [108].

integration methods is the problem of feature selection. The extent of this problem depends on the type of integration method. In early integration methods, the number of features will be increased at once if one does not use any dimensionality reduction method. Thus, the best and most efficient approach is to first reduce the size of each single-omics by using one of the filter gene selection methods and then combine the reduced single-omics data. Hybrid feature selection methods can also be used alongside early integration. However, one have to remove many of the features at once using this approach. In the intermediate integration methods, the feature extraction methods are used if the meaning of features changes. In contrast, if the data transformation is done to an intermediate type that does not change the meaning of features (e.g., graph, or network), one has to utilize a feature selection technique that work on those spaces. As mentioned in the late integration methods, a modeling algorithm is created on each omics data individually, and finally leveraging a voting method among models, the outcome is produced. Therefore, any of gene selection methods can be applied independently to single omics data.

4. Challenges in single and multi omics datasets

Single- and multi-omics datasets have been widely used in biomedical research and studies, for the purpose of finding biomarker signatures of a specific disease, discovering disease subtypes, predicting response to therapy or survival time and for functional omics studies. Study design may vary, from case-control studies to time series. As discussed earlier, the main challenge for single-omics data analysis is the small number of samples compared to the number of variables and the main challenge for multi-omics data analysis is the high variety of multiple omics data. In addition, there are a lot of empirical complexities that have collectively transformed the analysis of single- and multi-omics data into an attractive area. Here are some of the key challenges that we believe are quite general and extends to other application contexts.

4.1. Lack of samples

The first problem encountered in dealing with omics data is the challenge of having few records (usually less than 100 samples). In this setting, it is crucial to define a proper procedure to estimate classification error. Small sample size is an important problem in omics data analysis since some well-known research did not pan out (e.g., [114]).

4.2. Class imbalance

A common problem in omics data is class imbalance. This problem occurs when a dataset is dominated by a class or a few primary classes and the number of instances of these classes is much larger than other classes. To solve this problem, over-sampling or under-sampling techniques are commonly used.

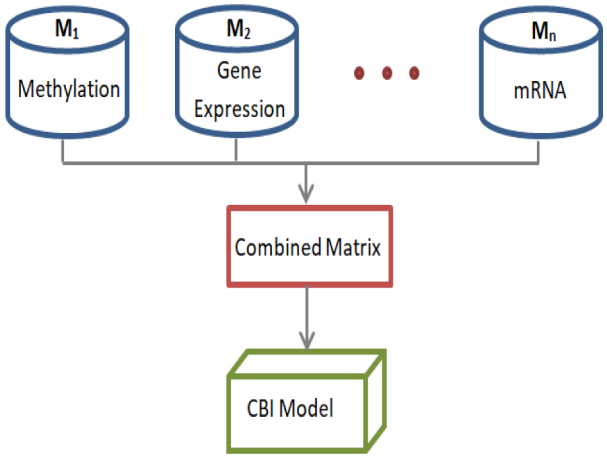
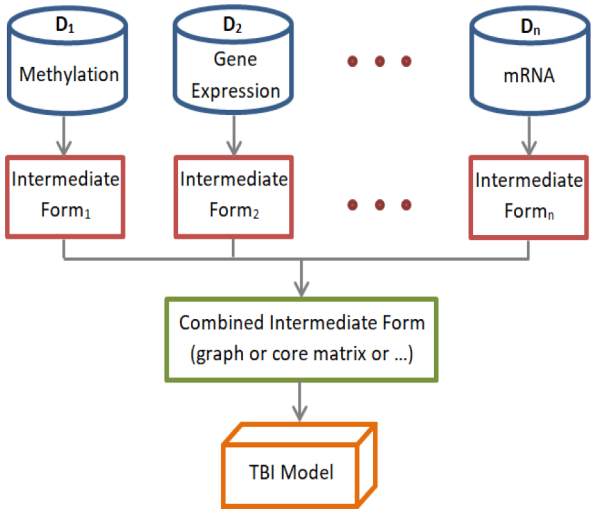
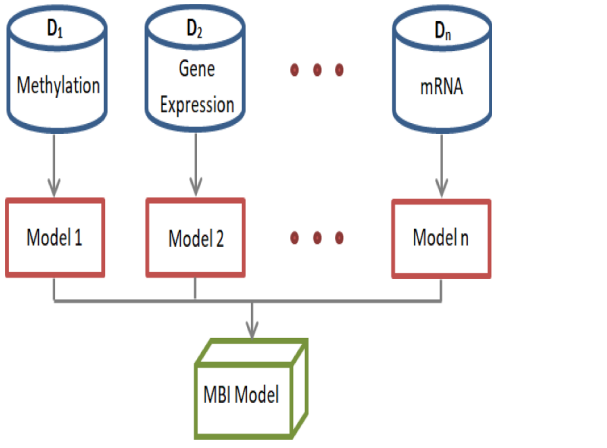
4.3. Data complexity

Data Complexity metrics are criteria that represent data specification. This metric measures the resolution, the linearity of the decision boundary and the overlap between classes to achieve a better classification. For example, the criterion F1 measures the overlap of classes, which focuses on the class discrimination capability of one feature individually [115].

4.4. Dataset shift

Another common problem that is caused by dividing the dataset into training and test data is called the dataset shift. This phenomenon occurs when the distribution of test data in a feature or combination of features or boundaries of classes differs with the distribution of training data. Consequently, the assumption that the training and test data have the same distribution, both in applications and in the real world, is violated and leads to volatility in the process of gene selection and classification.

Table 2
Advantage and disadvantage of data integration methods.

Integration Method	Shape	Advantage	Disadvantage
Concatenation-based (Early)		After a large input matrix is formed, it is easy to use different machine learning methods to analyze continuous or discrete data. Correlation between different omics data is considered	The combination of small matrices may produce a super matrix. The lack of some data typesFor a patient that exists in other data. It results in the appearance of a large number of Nan cells.
Transform-based (Intermediate)		Unique variables such as patient IDs can be used to connect multi-omics data types and integrate multiple values of continuous or discrete data.	The challenge is to turn into an intermediate form.
Model-based (Late)		Build a model independent of data type. Do not need to have different data from the same people.	Overfitting Correlation between different omics data is not considered

4.5. Outliers

Another important aspect that has been neglected in omics data is the discovery of pertinent data. In some omics data, there are records that are mistakenly tagged or identified, which should be identified and removed from the dataset, because they can have a negative influence on the selection of informative biomarkers and sample classification.

4.6. Missing value

In some papers, especially those that use DNA Methylation datasets, one of the biggest challenges is the existence of missing data values, simply because this issue may have a significant effect on the conclusions that can be drawn from the data. In recent years, several papers have been published in this area that used the imputation methods to solve this problem in genomic datasets [116–118].

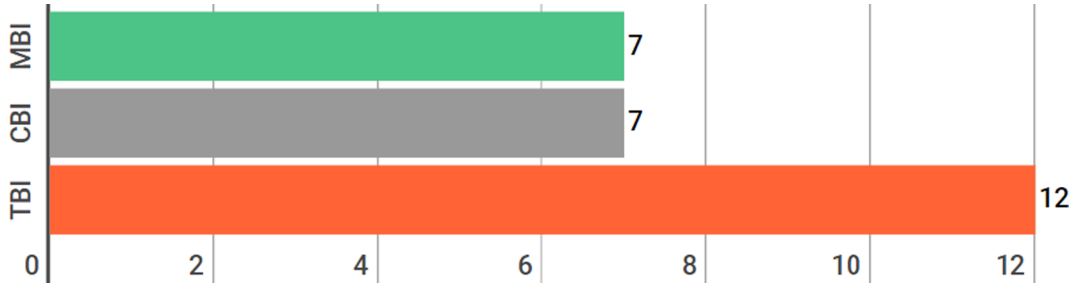


Fig. 10. The number of papers examined in this paper in field of data integration in multi-omics data analysis.

Table 3
Classification of papers based used integration and gene selection methods.

Integration Method	Reference	Learning Model	Gene Selection Method
Concatenation based (Early)	[100]	Results based on comparison of three models (Random Forest, eXtreme Gradient Boosting, Logistic Regression)	Hybrid (two stage)
Transform based (Intermediate)	[112]	Logistic Regression analysis	Filter (DESeq2, *t*-test)
Model based (Late)	[111]	Multiple Kernel Learning	Filter (Wilcoxon rank-sum test)
	[110]	kernel partial least squares	Embedded (Penalized logistic regression)
	[113]	Multiple Kernel Learning	Filter (Wilcoxon rank-sum test)

Table 4
Highest prediction accuracies and the number of selected genes for cancer datasets.

Cancer dataset	Ref.	Gene selection method	# Genes	# Samples	#class	Accuracy	Classifier / evaluation method	# Selected Genes
Leukemia	[23]	Filter	7129	72	2	99.7	SVM	5
	[36]	Wrapper	7129	38	2	100	SVM	5
	[19]	Ensemble	22,283	64	2	95.23	Random Forest	4
	[57]	Hybrid	7129	73	2	100	SVM	6
	[50]	Hybrid	7129	72	2	100	SVM	2
Lung	[23]	Filter	12,600	203	2	94.8	SVM	3
	[36]	Wrapper	12,533	32	2	100	SVM	4
	[20]	Embedded	12,600	203	5	94.5	5-fold CV	11
	[71]	Ensemble	12,533	181	2	98.36	KNN	N/A
	[51]	Hybrid	12,533	181	2	100	NB[1]	11
	[18]	Hybrid	7129	96	2	100	LIBSVM	3
DLBCL	[38]	Wrapper	7070	77	2	100	MLP[2]	3
	[20]	Embedded	7129	77	2	96.2	5-fold CV	41
	[71]	Ensemble	7129	77	2	94.21	KNN	N/A
	[18]	Hybrid	4026	47	2	100	LIBSVM	3
Colon	[23]	Filter	2000	62	2	95.4	SVM	4
	[38]	Wrapper	7464	36	2	100	KNN	2
	[20]	Embedded	2000	62	2	90	5-fold CV	N/A
	[19]	Ensemble	22,277	111	2	87.38	Random Forest	4
	[50]	Hybrid	2000	62	2	99.46	SVM	8
	[57]	Hybrid	7457	62	2	100	SVM	5
SRBCT	[38]	Wrapper	2308	83	4	100	SVM	5
	[20]	Ensemble	2308	83	4	99.5	5-fold CV	36
	[51]	Hybrid	2308	83	2	100	NB	34
	[120]	Hybrid	2308	83	4	100	ELM	9
Breast	[28]	Filter	7129	72	2	89.30	5-fold CV	99
	[71]	Ensemble	24,482	97	2	65.93	KNN	N/A
	[51]	Hybrid	24,481	97	2	92.75	NB	32
	[68]	Hybrid	24,481	97	2	94.74	SVM	37
Prostate	[23]	Filter	12,600	102	2	97.3	SVM	5
	[38]	Wrapper	12,533	102	2	100	MLP	3
	[48]	Embedded	54,675	21	2	85.8 (Avg)	N/A	28.5 (Avg)
	[71]	Ensemble	12,600	102	2	91.09	KNN	N/A
	[57]	Hybrid	12,600	102	2	100	SVM	5
	[58]	Hybrid	12,600	34	2	96.54	ELM	3
	[68]	Hybrid	12,600	136	2	100	SVM	11

[1] Naive Bayes.
[2] Multi-Layer Perceptron.

4.7. Data variety

Given the fact that data variety is one of the main features of the Big Data, it can be mentioned that the multi-omics data is a big data in terms of complexity. Each of the omics data has a specific computational complexity. This makes it extremely difficult to integrate this kind of data at the data level. It is also hard to figure out the interactions between these different genetic products. In other words, given

Table 5
Datasets used in recent papers in the field of data integration in cancer classification.

Cancer Type	Ref.	Omics Data	Method	Subject	# Class	Accuracy	AUC
Breast	[111]	CNV, mRNA, DNAm[1]	MBI	Subtypes	2	79.8	91.6
	[110]	CNV, mRNA, miRNA	MBI	Subtypes	2	91	N/A
	[105]	mRNA, miRNA, DNAm	TBI	Subtypes	5	93.95	N/A
	[123]	mRNA, miRNA	CBI	Subtypes	4	84.6	N/A
	[124]	CNV, GE[2]	MBI	Survival	2	74	N/A
	[125]	GE, DNAm	CBI	Survival Risk Group	3	N/A	79
	[126]	CNV, GE	TBI	Survival	2	N/A	81
	[113]	GE, DNAm	CBI	Grade	2	82.06	83.5
	[127]	GE, Proteome, CNVPhosphoproteome	CBI	Survival	2	60.7	N/A
			MBI	Survival	2	56.1	N/A
	[128]	GE, CNA	TBI	Subtypes	5	79.2	85
	[112]	GE, DNAm	TBI	Survival	2	74	81
Ovarian	[129]	CAN, GE, DNAm	TBI	Stage	2	N/A	85
				Grade			87
				Survival			75
	[109]	CAN, miRNA, DNAm, GE	MBI	Stage	2	N/A	84
				Grade			84
				Survival			79
	[100]	CAN, GE, DNAm, RNA-Seq	CBI	Survival	2	N/A	70
	[130]	GE, DNAm	TBI	Survival	2	N/A	79.98
	[131]	mRNA, miRNA, DNAm, Somatic mutations	CBI	Survival	2	N/A	87
Prostate	[132]	DNAm, miRNA Transcript1, miRNA Transcript2, N-Glycan serum, Serum protein	MBI	Growing speed	3	N/A	91
	[133]	GE, Metabolomics	TBI	Prognosis	2	N/A	95.36
	[134]	CNV, GE	MBI	Grade	2	N/A	90.06
				Stage			85.28
				Metastasis			98.68
				Recurrence			78.57
Bladder	[102]	CNV, GE, DNAm, miRNA	TBI	Genes identification	1	N/A	95.9
Neuroblastoma	[103]	CAN, GE	TBI	Subtypes	2	75.53	99.62
Rectal	[134]	GE, Proteome	MBI	Wheeler	3	N/A	92.69
				PN-Stage			96.30
				CRM			98.70
Kidney	[135]	GE, DNAm	CBI	Survival Risk Group	3	N/A	79.2
Glioblastoma Multiform	[136]	SNP, DNAm, CNV, GE	TBI	Survival Time	2	N/A	72

[1] DNA methylation.
[2] Gene expression.

the variety of features produced by high-performance devices, each with different sensitivity, error rate and data structure, it is quite challenging to combine and integrate them [119].

4.8. Unavailability of some data types

Some types of omics data may not be available for some patients in the learning process of multi-omics data. Suppose that the multi-omics data model is based on five types of omics data, but there are only three available types of omics data for each patient, which may vary from patient to patient. This also raises the complexity of building a single model for this type of dataset, since many statistical methods cannot be applied directly to an incomplete dataset.

5. Review on single- and multi-omics cancer datasets

In this section, we review the single and multi-omics cancer datasets that are commonly used in the literature.

5.1. Single omics datasets

The purpose of this section is to introduce and review the datasets used in gene selection research. According to Table 4, it can be seen that the accuracy of results are high in many datasets (in some datasets, 100% accuracy is reported), and researchers attempts to develop feature selection algorithms capable of finding lower number of biomarkers while preserving higher classification accuracies which is obviously much more valuable since it is more interpretable.

The highlighted records in Table 4 illustrate the research whose details of the proposed method are presented in Section 3.

According to Table 4, it can be concluded that the most challenging type of cancer, among the single-omics data from a variety of cancer cases, is breast cancer. Another result that can be drawn is that the hybrid method has the best performance among all methods. It must be noted that achieving 100% accuracy in realistic dataset poses a question about reproducibility of the presented results. Sometimes, hidden overfitting can be present, and only fully reproducibility of experiments can lead to really assess the performance of the methods.

Some papers also use a large methylation data that recognizes several types of cancers [121,122]. For example, Liu et al. assembled data from methylation of 32 types of cancers from the TCGA and constructed a dataset with 32 classes, 9223 CpG-sites and 10,000 samples and scored 83.5% accuracy in detecting multiple types of cancers in this dataset [122].

5.2. Multi omics datasets

Most papers published in the field of multi-omics data analysis and cancer classification have used the datasets that exist in the TCGA datasets portal. In some papers that use graph-based or network-based methods, they use protein–protein-interaction and pathway datasets from their specific database along with these TCGA omics datasets. In Table 5, several datasets are reported all of which are related to some papers in the field of multi-omics data analysis and cancer classification. The highlighted records in this table belong to the research that details of their proposed method were described in Section 4.

According to Table 5, the most important subject in breast cancer classification is the subtype prediction. The best reported accuracy

among these researches is 93.95 using TBI method. Also, the most important subject in ovarian cancer classification is survival prediction. The best reported area under curve in these papers is 79.98 using TBI method. Another result that can be drawn is that subjects in prostate cancer classification are diverse, so they are not comparable.

6. Conclusion

In this paper, we highlighted three points. The first point relates to methods of gene selection as one of the important sectors in single- and multi-omics data analysis for cancer classification. Gene selection methods are divided into five groups: filter, wrapper, embedded, ensemble and hybrid. Filter methods are fast but the quality of their selected features is poor in contrast to other methods. On the other hand, wrapper methods select useful subsets of features, but they come with a huge computational cost, and as such, are rarely used alone. Hybrid methods are the most popular methods in the field of gene selection as they can combine the capabilities of other feature selection methods, and thus perform better. The underlying goals of developing gene selection methods are improving accuracy, increasing classifier efficiency and reducing computational complexity. In most of the reviewed papers, another important objective is to reduce the number of features that are effective in predicting the disease.

The second point relates to data integration methods which is an important challenge in analyzing multi-omics datasets. By data integration, relationships between various levels of genetic products can be considered in analysis, providing better, more reliable and more stable results. The transform-based integration method, known as intermediate integration, is the most popular method of integration. Intermediate graph is also more frequently used among the intermediate integration methods since it considers the relationship between different levels of genetic products.

The third point is that although data integration is the most important challenge in multi-omics data analysis, the use of feature selection techniques developed in single-omics data analysis research can significantly improve the final performance.

Last but not least, analysis of data on breast cancer in both single- and multi-omics data is the most problematic issue among diseases.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research was in part supported by a grant from the Institute for Research in Fundamental Sciences, Iran (No. CS1398-4-98).

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre, A. Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA. Cancer J. Clin. 68 (6) (2018) 394–424.
[2] V.N. Kristensen, O.C. Lingjærde, H.G. Russnes, H.K.M. Vollan, A. Frigessi, A.L. Børresen-Dale, Principles and methods of integrative genomic analyses in cancer, Nat. Rev. Cancer 14 (5) (2014) 299–313.
[3] R. Tibshirani, Regression shrinkage and selection via the Lasso, J. Royal Stat. Soc. Ser. B (Methodol.), vol. 58. WileyRoyal Statistical Society, pp. 267–288, 1996.
[4] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (1–3) (2002) 389–422.
[5] R. Díaz-Uriarte, S. Alvarez de Andrés, Gene selection and classification of microarray data using random forest, BMC Bioinf. 7 (1) (Jan. 2006) 1–13.
[6] H. Peng, F. Long, C. Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (Aug. 2005) 1226–1238.
[7] B. Lehner, Modelling genotype-phenotype relationships and human disease with genetic interaction networks, J. Exp. Biol. 210 (9) (2007) 1559–1566.
[8] O. Gevaert, V. Villalobos, B.I. Sikic, S.K. Plevritis, Identification of ovarian cancer driver genes by using module network integration of multi-omics data, Interface Focus 3 (4) (Aug. 2013).
[9] C. Meng, B. Kuster, A.C. Culhane, A.M. Gholami, A multivariate approach to the integration of multi-omics datasets, BMC Bioinf. 15 (1) (May 2014) 1–13.
[10] M.D. Ritchie, E.R. Holzinger, R. Li, S.A. Pendergrass, D. Kim, Methods of integrating data to uncover genotype-phenotype interactions, Nat. Rev. Genet. 16 (2) (2015) 85–97.
[11] S. Huang, K. Chaudhary, L.X. Garmire, More is better: Recent progress in multi-omics data integration methods, Front. Genet., vol. 8, no. JUN, 2017, pp. 1–12.
[12] E. Lin, H.Y. Lane, Machine learning and systems genomics approaches for multi-omics data, Biomark. Res. 5 (1) (2017) 1–6.
[13] I.S.L. Zeng, T. Lumley, Review of statistical learning methods in integrated omics studies (An integrated information science), Bioinform. Biol. Insights 12 (2018).
[14] I.A. Gheyas, L.S. Smith, Feature subset selection in large dimensionality domains, Pattern Recognit. 43 (1) (Jan. 2010) 5–13.
[15] J.C. Ang, A. Mirzal, H. Haron, H.N.A. Hamed, Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection, IEEE/ACM Trans. Comput. Biol. Bioinforma. 13 (5) (2016) 971–989.
[16] J. López, S. Maldonado, M. Carrasco, Double regularization methods for robust feature selection and SVM classification via DC programming, Inf. Sci. (Ny) 429 (2018) 377–389.
[17] S.B. Chen, Y. Zhang, C.H.Q. Ding, Z.L. Zhou, B. Luo, A discriminative multi-class feature selection method via weighted l2,1-norm and extended elastic net, Neurocomputing 275 (2018) 1140–1149.
[18] L. Gao, M. Ye, X. Lu, D. Huang, Hybrid method based on information gain and support vector machine for gene selection in cancer classification, Genomics. Proteomics Bioinformatics 15 (6) (Dec. 2017) 389–395.
[19] M. Ram, A. Najafi, M.T. Shakeri, Classification and biomarker genes selection for cancer gene expression data using random forest, Iran. J. Pathol. 12 (4) (2017) 339–347.
[20] S. Guo, D. Guo, L. Chen, Q. Jiang, A centroid-based gene selection method for microarray data classification, J. Theor. Biol. 400 (2016) 32–41.
[21] B. Tran, B. Xue, M. Zhang, Genetic programming for feature construction and selection in classification on high-dimensional data, Memetic Comput. 8 (1) (2016) 3–15.
[22] K.H. Chen, K.J. Wang, K.M. Wang, M.A. Angelia, Applying particle swarm optimization-based decision tree classifier for cancer classification on gene expression data, Appl. Soft Comput. J. 24 (2014) 773–780.
[23] J. Xu, H. Mu, Y. Wang, F. Huang, Feature genes selection using supervised locally linear embedding and correlation coefficient for microarray classification, Comput. Math. Methods Med. 2018 (Jan. 2018) 1–11.
[24] J.R. Ummadi, B.V.R. Reddy, A novel statistical feature selection measure for decision tree models on microarray cancer detection, in: Proceedings of International Conference on Computational Intelligence and Data Engineering, 2018, pp. 229–245.
[25] M. Mohammadi, H. Sharifi Noghabi, G. Abed Hodtani, H. Rajabi Mashhadi, Robust and stable gene selection via maximum-minimum correntropy criterion, Genomics 107 (2–3) (2016) 83–87.
[26] H. Chen, Y. Zhang, I. Gutman, A kernel-based clustering method for gene selection with gene expression data, J. Biomed. Inform. 62 (2016) 12–20.
[27] M.S. Raza, U. Qamar, An incremental dependency calculation technique for feature selection using rough sets, Inf. Sci. (Ny) 343–344 (2016) 41–65.
[28] P.A. Mundra, J.C. Rajapakse, Gene and sample selection using T-score with sample selection, J. Biomed. Inform. 59 (2016) 31–41.
[29] S. Begum, D. Chakraborty, R. Sarkar, Data classification using feature selection and kNN machine learning approach, in: 2015 International Conference on Computational Intelligence and Communication Networks (CICN), 2015, pp. 811–814.
[30] Y. Chen, Z. Zhang, J. Zheng, Y. Ma, Y. Xue, Gene selection for tumor classification using neighborhood rough sets and entropy measures, J. Biomed. Inform. 67 (Mar. 2017) 59–68.
[31] K. Zheng, X. Wang, Feature selection method with joint maximal information entropy between features and class, Pattern Recognit. 77 (2018) 20–29.
[32] C. Liu, W. Wang, Q. Zhao, X. Shen, M. Konan, A new feature selection method based on a validity index of feature subset, Pattern Recognit. Lett. 92 (2017) 1–8.
[33] A. Wang, N. An, J. Yang, G. Chen, L. Li, G. Alterovitz, Wrapper-based gene selection with Markov blanket, Comput. Biol. Med. 81 (December) (2016, 2017) 11–23.
[34] A. Wang, N. An, G. Chen, L. Li, G. Alterovitz, Accelerating wrapper-based feature selection with K-nearest-neighbor, Knowledge-Based Syst. 83 (1) (2015) 81–91.
[35] R. Panthong, A. Srivihok, Wrapper feature subset selection for dimension reduction based on ensemble learning algorithm, Procedia Comput. Sci. 72 (2015) 162–169.
[36] S.K. Pati, S. Sengupta, A.K. Das, Improved genetic algorithm for selecting significant genes in cancer diagnosis, Prog. Adv. Comput. Intell. Eng. 564 (2018) 395–405.
[37] N.Y. Moteghaed, K. Maghooli, M. Garshasbi, Improving classification of cancer and mining biomarkers from gene expression profiles using hybrid optimization algorithms and fuzzy support vector machine, J. Med. Signals Sens. 8 (1) (2018) 1–11.
[38] M. Ghosh, S. Begum, R. Sarkar, D. Chakraborty, U. Maulik, Recursive memetic algorithm for gene selection in microarray data, Expert Syst. Appl. 116 (2019) 172–185.
[39] H. Wang, X. Jing, B. Niu, A discrete bacterial algorithm for feature selection in classification of microarray gene expression cancer data, Knowledge-Based Syst.

126 (2017) 8–19.
[40] E. Aličković, A. Subasi, Breast cancer diagnosis using GA feature selection and Rotation Forest, Neural Comput. Appl. 28 (4) (2017) 753–763.
[41] P. Moradi, M. Gholampour, A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy, Appl. Soft Comput. J. 43 (2016) 117–130.
[42] M. García-Torres, F. Gómez-Vela, B. Melián-Batista, J.M. Moreno-Vega, High-dimensional feature selection via feature grouping: A Variable Neighborhood Search approach, Inf. Sci. (Ny) 326 (2016) 102–118.
[43] S. Gunasundari, S. Janakiraman, S. Meenambal, Velocity bounded boolean particle swarm optimization for improved feature selection in liver and kidney disease diagnosis, Expert Syst. Appl. 56 (2016) 28–47.
[44] B. Xue, M. Zhang, W.N. Browne, X. Yao, A survey on evolutionary computation approaches to feature selection, IEEE Trans. Evol. Comput. 20 (4) (2016) 606–626.
[45] X. Han, et al., Feature subset selection by gravitational search algorithm optimization, Inf. Sci. (Ny) 281 (2014) 128–146.
[46] J.A. Castellanos-Garzón, J. Ramos, D. López-Sánchez, J.F. de Paz, J.M. Corchado, An ensemble framework coping with instability in the gene selection process, Interdiscip. Sci. Comput. Life Sci. 10 (1) (2018) 12–23.
[47] H. Güney, H. Öztoprak, Microarray-based cancer diagnosis: repeated cross-validation-based ensemble feature selection, Electron. Lett. 54 (5) (2018) 272–274.
[48] J. Li, W. Dong, D. Meng, Grouped gene selection of cancer via adaptive sparse group lasso based on conditional mutual information, IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 5963, no. c, 2017, pp. 1–11.
[49] R.E. Naftchali, M.S. Abadeh, A multi-layered incremental feature selection algorithm for adjuvant chemotherapy effectiveness/futileness assessment in non-small cell lung cancer, Biocybern. Biomed. Eng. 37 (3) (2017) 477–488.
[50] H. Motieghader, A. Najafi, B. Sadeghi, A. Masoudi-Nejad, A hybrid gene selection algorithm for microarray cancer classification using genetic algorithm and learning automata, Informatics Med. Unlocked 9 (Jan. 2017) 246–254.
[51] I. Jain, V.K. Jain, R. Jain, Correlation feature selection based improved-Binary Particle Swarm Optimization for gene selection and cancer classification, Appl. Soft Comput. J. 62 (2018) 203–215.
[52] M. Dashtban, M. Balafar, P. Suravajhala, Gene selection for tumor classification using a novel bio-inspired multi-objective approach, Genomics 110 (1) (2018) 10–17.
[53] S. Venkataraman, Rajalakshmi Selvaraj, Optimal and novel hybrid feature selection framework for effective data classification, 2018, pp. 499–514.
[54] C.M. Lai, Multi-objective simplified swarm optimization with weighting scheme for gene selection, Appl. Soft Comput. J. 65 (2018) 58–68.
[55] P. Agarwalla, S. Mukhopadhyay, Bi-stage hierarchical selection of pathway genes for cancer progression using a swarm based computational approach, Appl. Soft Comput. J. 62 (2018) 230–250.
[56] Manosij Ghosh, Sukdev Adhikary, Kushal Kanti Ghosh, Aritra Sardar, Shemim Begum, Ram Sarkar, Genetic algorithm based cancerous gene identification from microarray data using ensemble of filter methods, Med. Biol. Eng. Comput. 57 (1) (2019) 159–176, https://doi.org/10.1007/s11517-018-1874-4.
[57] V. Elyasigomari, D.A. Lee, H.R.C. Screen, M.H. Shaheed, Development of a two-stage gene selection method that incorporates a novel hybrid approach using the cuckoo optimization algorithm and harmony search for cancer classification, J. Biomed. Inform. 67 (2017) 11–20.
[58] H. Lu, J. Chen, K. Yan, Q. Jin, Y. Xue, Z. Gao, A hybrid feature selection algorithm for gene expression data classification, Neurocomputing 256 (2017) 56–62.
[59] P. Shunmugapriya, S. Kanmani, A hybrid algorithm using ant and bee colony optimization for feature selection and classification (AC-ABC Hybrid), Swarm Evol. Comput. 36 (Oct. 2017) 27–36.
[60] H. Salem, G. Attiya, N. El-Fishawy, Classification of human cancer diseases by gene expression profiles, Appl. Soft Comput. J. 50 (2017) 124–134.
[61] F. Han, et al., A gene selection method for microarray data based on binary PSO encoding gene-to-class sensitivity information, IEEE/ACM Trans. Comput. Biol. Bioinforma. 14 (1) (Jan. 2017) 85–96.
[62] C. Arunkumar, S. Ramakrishnan, Attribute selection using fuzzy roughset based customized similarity measure for lung cancer microarray gene expression data, Futur. Comput. Informatics J. 3 (1) (2018) 131–142.
[63] J. Lv, Q. Peng, X. Chen, Z. Sun, A multi-objective heuristic algorithm for gene expression microarray data classification, Expert Syst. Appl. 59 (2016) 13–19.
[64] M. Xi, J. Sun, L. Liu, F. Fan, X. Wu, Cancer feature selection and classification using a binary quantum-behaved particle swarm optimization and support vector machine, Comput. Math. Methods Med. 2016 (2016).
[65] J. Apolloni, G. Leguizamón, E. Alba, Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments, Appl. Soft Comput. J. 38 (2016) 922–932.
[66] E. Bonilla-Huerta, A. Hernández-Montiel, R. Morales-Caporal, M. Arjona-López, Hybrid framework using multiple-filters and an embedded approach for an efficient selection and classification of microarray data, IEEE/ACM Trans. Comput. Biol. Bioinforma. 13 (1) (2016) 12–26.
[67] L. Chuang, C. Ke, C. Yang, A hybrid both filter and wrapper feature selection method for microarray classification, vol. I, 2008, pp. 19–21.
[68] X. Huang, L. Zhang, B. Wang, F. Li, Z. Zhang, Feature clustering based support vector machine recursive feature elimination for gene selection, Appl. Intell. 48 (3) (2018) 594–607.
[69] M.K. Ebrahimpour, M. Eftekhari, Distributed feature selection: A hesitant fuzzy correlation concept for microarray high-dimensional datasets, Chemom. Intell. Lab. Syst. 173 (January) (2018) 51–64.
[70] B. Pes, N. Dessì, M. Angioni, Exploiting the ensemble paradigm for stable feature selection: A case study on high-dimensional genomic data, Inf. Fusion 35 (2017) 132–147.
[71] A. Ben Brahim, M. Limam, Ensemble feature selection for high dimensional data: a new method and a comparative study, Adv. Data Anal. Classif., 2017, pp. 1–16.
[72] B. Seijo-Pardo, I. Porto-Díaz, V. Bolón-Canedo, A. Alonso-Betanzos, Ensemble feature selection: Homogeneous and heterogeneous approaches, Knowledge-Based Syst. 118 (2017) 124–139.
[73] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, An ensemble of filters and classifiers for microarray data classification, Pattern Recognit. 45 (1) (2012) 531–539.
[74] M.K. Ebrahimpour, H. Nezamabadi-pour, M. Eftekhari, CCFS: A cooperating coevolution technique for large scale feature selection on microarray datasets, Comput. Biol. Chem. 73 (2018) 171–178.
[75] M.K. Ebrahimpour, M. Eftekhari, Ensemble of feature selection methods: A hesitant fuzzy sets approach, Appl. Soft Comput. J. 50 (2017) 300–312.
[76] T. Nguyen, S. Nahavandi, Modified AHP for gene selection and cancer classification using type-2 fuzzy logic, IEEE Trans. Fuzzy Syst. 24 (2) (2016) 273–287.
[77] K.H. Liu, Z.H. Zeng, V.T.Y. Ng, A Hierarchical Ensemble of ECOC for cancer classification based on multi-class microarray data, Inf. Sci. (Ny) 349–350 (2016) 102–118.
[78] A.K. Das, S. Das, A. Ghosh, Ensemble feature selection using bi-objective genetic algorithm, Knowledge-Based Syst. 123 (2017) 116–127.
[79] M. Mollaee, M.H. Moattar, A novel feature extraction approach based on ensemble feature selection and modified discriminant independent component analysis for microarray data classification, Biocybern. Biomed. Eng. 36 (3) (Jan. 2016) 521–529.
[80] P. Mohapatra, S. Chakravarty, P.K. Dash, Microarray medical data classification using kernel ridge regression and modified cat swarm optimization based gene selection system, Swarm Evol. Comput. 28 (2016) 144–160.
[81] V. Elyasigomari, M.S. Mirjafari, H.R.C. Screen, M.H. Shaheed, Cancer classification using a novel gene selection approach by means of shuffling based on data clustering with optimization, Appl. Soft Comput. J. 35 (2015) 43–51.
[82] J. Sun, A. Zhou, S. Keates, S. Liao, Simultaneous Bayesian clustering and feature selection through student's t mixtures model, IEEE Trans. Neural Networks Learn. Syst. 29 (4) (2017) 1187–1199.
[83] M. Luo, F. Nie, X. Chang, Y. Yang, A.G. Hauptmann, Q. Zheng, Adaptive unsupervised feature selection with structure regularization, IEEE Trans. Neural Networks Learn. Syst. 29 (4) (2018) 944–956.
[84] X. Zhu, X. Li, S. Zhang, C. Ju, X. Wu, Robust joint graph sparse coding for unsupervised spectral feature selection, IEEE Trans. Neural Networks Learn. Syst. 28 (6) (2017) 1263–1275.
[85] S. Wang, J. Tang, H. Liu, Embedded unsupervised feature selection, in Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015, p. 7.
[86] L. Du, Y.-D. Shen, "Unsupervised Feature Selection with Adaptive Structure Learning", Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15, 2015, pp. 209–218.
[87] Z. Li, J. Liu, Y. Yang, X. Zhou, H. Lu, Clustering-guided sparse structural learning for unsupervised feature selection, IEEE Trans. Knowl. Data Eng. 26 (9) (2014) 2138–2150.
[88] M. Qian, C. Zhai, "Robust unsupervised feature selection", Twenty-Third international joint conference on Artificial Intelligence, 2013, pp. 1621–1627.
[89] Z. Zhao, L. Wang, H. Liu, Efficient spectral feature selection with minimum redundancy, in: Twenty-Fourth AAAI Conference on Artificial Intelligence, 2010.
[90] N. Nidheesh, K.A. Abdul Nazeer, P.M. Ameer, An enhanced deterministic K-Means clustering algorithm for cancer subtype prediction from gene expression data, Comput. Biol. Med. 91 (2017) 213–221.
[91] S. Tabakhi, P. Moradi, F. Akhlaghian, An unsupervised feature selection algorithm based on ant colony optimization, Eng. Appl. Artif. Intell. 32 (Jun. 2014) 112–123.
[92] S. Maldonado, R. Weber, A wrapper method for feature selection using Support Vector Machines, Inf. Sci. (Ny) 179 (13) (2009) 2208–2217.
[93] X. Chen, J. Z. Huang, Q. Wu, M. Yang, Subspace weighting co-clustering of gene expression data, IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 5963, no. c, 2017, pp. 1–1.
[94] L. Sun, X. Zhang, J. Xu, W. Wang, R. Liu, A Gene selection approach based on the fisher linear discriminant and the neighborhood rough set, Bioengineered 9 (1) (2018) 144–151.
[95] S. Solorio-Fernández, J.A. Carrasco-Ochoa, J.F. Martínez-Trinidad, A new hybrid filter–wrapper feature selection method for clustering based on ranking, Neurocomputing 214 (2016) 866–880.
[96] S. Zhang, H.S. Wong, Y. Shen, D. Xie, A new unsupervised feature ranking method for gene expression data based on consensus affinity, IEEE/ACM Trans. Comput. Biol. Bioinforma. 9 (4) (2012) 1257–1263.
[97] X. Yu, G. Yu, J. Wang, Clustering cancer gene expression data by projective clustering ensemble, PLoS ONE 12 (2) (2017) 1–21.
[98] Hui Zou, Trevor Hastie, Regularization and variable selection via the elastic net, J. Royal Statistical Soc. B 67 (2) (2005) 301–320, https://doi.org/10.1111/rssb:2005.67.issue-210.1111/j.1467-9868.2005.00503.x.
[99] P. Mohapatra, S. Chakravarty, P.K. Dash, Microarray medical data classification using kernel ridge regression and modified cat swarm optimization based gene selection system, Swarm Evol. Comput. 28 (2015) 144–160.
[100] Yasser EL-Manzalawy, Tsung-Yu Hsieh, Manu Shivakumar, Dokyoon Kim, Vasant Honavar, Min-redundancy and max-relevance multi-view feature selection for predicting ovarian cancer survival using multi-omics data, BMC Med. Genomics 11 (S3) (2018), https://doi.org/10.1186/s12920-018-0388-0.
[101] Stephan Gade, Christine Porzelius, Maria Fälth, Jan C Brase, Daniela Wuttig, Ruprecht Kuner, Harald Binder, Holger Sültmann, Tim Beißbarth, Graph based fusion of miRNA and mRNA expression data improves clinical outcome prediction

in prostate cancer, BMC Bioinf. 12 (1) (2011), https://doi.org/10.1186/1471-2105-12-488.
[102] C. Peng, A. Li, M. Wang, Discovery of bladder cancer-related genes using integrative heterogeneous network modeling of multi-omics data, Sci. Rep. 7 (1) (2017) 1–11.
[103] L. Zhang, et al., Deep learning-based multi-omics data integration reveals two prognostic subtypes in high-risk neuroblastoma, Front. Genet., vol. 9, no. OCT, 2018, pp. 1–9.
[104] H. Sharifi-Noghabi, O. Zolotareva, C.C. Collins, M. Ester, MOLI: Multi-omics late integration with deep neural networks for drug response prediction, bioRxiv, p. 531327, 2019.
[105] S. Rakshit, I. Saha, S.S. Chakraborty, D. Plewczyski, Deep learning for integrated analysis of breast cancer subtype specific multi-omics data, IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON, vol. 2018-Octob, no. October, pp. 1917–1922, 2019.
[106] S. Zhang, Q. Li, J. Liu, X.J. Zhou, A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules, Bioinformatics, vol. 27, no. ii, 2011, pp. 401–409.
[107] M. Zitnik, B. Zupan, Data fusion by matrix factorization, IEEE Trans. Pattern Anal. Mach. Intell. 37 (1) (2014) 41–53.
[108] R. Argelaguet, et al., Multi-omics factor analysis — a framework for unsupervised integration of multi-omics data sets, Mol. Syst. Biol. 14 (6) (2018) 1–13.
[109] D. Kim, et al., Knowledge boosting: A graph-based integration approach with multi-omics data and genomic knowledge for cancer clinical outcome prediction, J. Am. Med. Informatics Assoc. 22 (1) (2015) 109–120.
[110] H. Yang, H. Cao, T. He, T. Wang, Y. Cui, Multilevel heterogeneous omics data integration with kernel fusion, Brief. Bioinform. 00 (April) (2018) 1–15.
[111] M. Tao, et al., Classifying breast cancer subtypes using multiple kernel learning based on omics data, Genes (Basel) 10 (3) (2019) 200.
[112] S.Y. Kim, T.R. Kim, H.-H. Jeong, K.-A. Sohn, Integrative pathway-based survival prediction utilizing the interaction between gene expression and DNA methylation in breast cancer, BMC Med. Genomics 11 (S3) (Sep. 2018) 68.
[113] T. Song, Y. Wang, W. Du, S. Cao, Y. Tian, Y. Liang, The method for breast cancer grade prediction and pathway analysis based on improved multiple kernel learning, J. Bioinform. Comput. Biol. 15 (1) (2017) 1–21.
[114] L.J. Van't Veer, et al., Gene expression profiling predicts clinical outcome of breast cancer, Nature 415 (6871) (2002) 530–536.
[115] L. Mor, Data complexity measures for analyzing the effect of SMOTE over microarrays, no. April, pp. 27–29, 2016.
[116] L.S. Zou, et al., BoostMe accurately predicts DNA methylation values in whole-genome bisulfite sequencing of multiple human tissues, BMC Genomics 19 (1) (2018) 1–15.
[117] E.B. Wijaya, E. Lim, D. Agustriawan, C. Huang, J.J.P. Tsai, K. Ng, Algorithms for Computational Biology, vol. 10849, Springer International Publishing, 2018.
[118] P. Di Lena, C. Sala, A. Prodi, C. Nardini, Missing value estimation methods for DNA methylation data, Bioinformatics (2019).
[119] P.J. Fabres, C. Collins, T.R. Cavagnaro, C.M. Rodríguez López, A concise review on multi-omics data integration for terroir analysis in Vitis vinifera, Front. Plant Sci., vol. 8, no. June, 2017, pp. 1–8.
[120] A. Chinnaswamy, R. Srinivasan, Hybrid information gain based fuzzy roughset feature selection in cancer microarray data, 2017 Innov. Power Adv. Comput. Technol. i-PACT 2017, vol. 2017-Janua, 2018, pp. 1–6.
[121] F. Celli, F. Cumbo, E. Weitschek, Classification of large DNA methylation datasets for identifying cancer drivers, Big Data Res. 13 (2018) 21–28.
[122] L. Liu, et al., Targeted methylation sequencing of plasma cell-free DNA for cancer detection and classification, Ann. Oncol. 29 (6) (2018) 1445–1453.
[123] Y. Li, X.Q. Tang, Z. Bai, X. Dai, Exploring the intrinsic differences among breast tumor subtypes defined using immunohistochemistry markers based on the decision tree, Sci. Rep. 6 (October) (2016) 1–13.
[124] J.A. Seoane, I.N.M. Day, T.R. Gaunt, C. Campbell, A pathway-based data integration framework for prediction of disease progression, Bioinformatics 30 (6) (2014) 838–845.
[125] J.A. Thompson, B.C. Christensen, C.J. Marsit, Methylation-to-expression feature models of breast cancer accurately predict overall survival, distant-recurrence free survival, and pathologic complete response in multiple cohorts, Sci. Rep. 8 (1) (2018) 1–10.
[126] A. González-Reymúndez, G. De Los Campos, L. Gutiérrez, S.Y. Lunt, A.I. Vazquez, Prediction of years of life after diagnosis of breast cancer using omics and omic-by-treatment interactions, Eur. J. Hum. Genet. 25 (5) (2017) 538–544.
[127] S. Ma, J. Ren, D. Fenyö, Breast Cancer Prognostics Using Multi-Omics Data, AMIA Jt. Summits Transl. Sci. proceedings. AMIA Jt. Summits Transl. Sci. (2016, 2016) 52–59.
[128] Y.W., Md. Mohaiminul Islam, P. Hu, Deep learning models for predicting phenotypic traits and diseases from omics data, Artif. Intell. Emerg. Trends Appl., vol. i, no. Artificial Intelligence, 2018, p. 13.
[129] D. Kim, R. Li, A. Lucas, S.S. Verma, S.M. Dudek, M.D. Ritchie, Using knowledge-driven genomic interactions for multi-omics data analysis: Metadimensional models for predicting clinical outcomes in ovarian carcinoma, J. Am. Med. Informatics Assoc. 24 (3) (2017) 577–587.
[130] A.D. Torshizi, L.R. Petzold, Graph-based semi-supervised learning with genomic data integration using condition-responsive genes applied to phenotype classification, J. Am. Med. Informatics Assoc. 25 (1) (2018) 99–108.
[131] A. Fu, H.R. Chang, Z.F. Zhang, Integrated multiomic predictors for ovarian cancer survival, Carcinogenesis 39 (7) (2018) 860–868.
[132] K. Murphy, et al., Integrating biomarkers across omic platforms: an approach to improve stratification of patients with indolent and aggressive prostate cancer, Mol. Oncol. 12 (9) (2018) 1513–1525.
[133] W. Liu, et al., Topologically inferring pathway activity toward precise cancer classification via integrating genomic and metabolomic data: Prostate cancer as a case, Sci. Rep. 5 (June) (2015) 1–15.
[134] A. Daemen, et al., A kernel-based integration of genome-wide data for clinical decision support, Genome Med. 1 (4) (2009) 1–17.
[135] J.A. Thompson, C.J. Marsit, A methylation-to-expression feature model for generating accurate prognostic risk scores and identifying disease targets in clear cell kidney cancer, Biocomput. 2017 (2016) 509–520.
[136] Y.L. Bernal Rubio, et al., Whole-genome multi-omic study of survival in patients with glioblastoma multiforme, G3;#58; Genes|Genomes|Genetics, vol. 8, no. 11, 2019, pp. 3627–3636.