# A Review on Cancer Dataset Classification using Data Mining Methods

**G.Jeya kumar[a] and Dr.T.Kamala kannan[b]**
[a] Research Scholar, Department of Information Technology. Computer Sciences.VISTAS, Chennai
[b]Associate Professor, Department of Information Technology. Computer Sciences.VISTAS, Chennai

**Article History**: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 28 April 2021

**Abstract:** Data analytics is observed as routine biomedical research domain focused on relevant clinical condition of cancer. Bioinformatics is characterized with consideration of several disease aspects. At present, several research focused on data processing of medical data. The medical data incorporates vast range of data such as miRNA, DNA with consideration of instances. The analysis of medical data expressed the micro array expression of gene with consideration of data attributes for minimal sample data count. For clinical and research process by using available data, medical data analysis extract equivalent data about genes. Data mining concentrated on analysis of cancer data with complexity analysis of underlying molecular diseases. The data mining of medical data is provided for processing reliable information and mechanism. However, construction of building model for medical data is complex for achieving reliable data. In this review, presented about contribution of data analysis for cancer data. The research concentrated on gene selection and data integration. Further, the analysis is based on the examination of data mining method analysis with consideration of gene data.

**Keywords:** Cancer, Data Mining, Data Integration, High Dimensional Dataset, Leukemia

## 1. Introduction

Cancer is one of the significant causes of death worldwide and from January to September 2018, almost 9.6 million people died. In total, cancer is responsible for one out of every six deaths in world. In 2010, roughly 325,000 individuals in the UK were diagnosed with cancer (roughly 890 individuals every day)[1]. During their lifespan, more than 1 in 3 individuals in UK will experience a form of cancer. About 12.7 million new cancer cases were estimated in 2008 worldwide. Half of those afflicted with cancer currently live for at least 5 years. With continued progress in cancer research as well as healthcare provision, cancer survival rates in UK have doubled in last 40 years. Today, cancer is most frequent cause of death (29%) followed by circulatory disorders (28%) such as heart failure as well as stroke[2]. Further technologies are crucially based on reliable as well as equivalent data collection.

Recently, advancement in medical processing provides clear insight about design and disease for standard therapeutic strategies, such as surgery as well as chemotherapy. Therefore, analysis of medical data provides clear cancer prediction. Medical data analysis provides significant advantage for computer science, bioinformatics, biology and survival rate of patient. Over past years, data analysis exhibits promising performance characteristics with incorporation of machine learning. Medical dataset studies are planned to investigate: (1) genetic causes of cancer; (2) existence of various subgroups of cancer; (3) molecular signatures of cancerous and non-cancer tissues; (4) timing of gene expression in rapidly changing forms of cancer. Many of these study efforts concentrate on offering an understanding of the findings that would improve awareness of the underlying cancer mechanism [3].

In medical field, Data Mining (DM) provides considerable advancement in scientific and technological medical data processing [4]. The data mining process is considered as competitive and powerful mechanism for construction of algorithm, which involved in data processing based on data pattern or relationships [5]. The incorporation of Machine Learning (ML) provides statistical approach for computation for extraction of generalized data. The DM evaluated the features of medical application with increase in research processing. The characteristics includes different characteristics with consideration of cellular and molecular levels like bioinformatics; tissue and organ level as imaging informatics; single patient information about clinical; based on public information health society population level [6].

In this regard, it is by data mining that recent work has been committed to the early identification of cancer [7]. Other pathologies, such as coronary and respiratory disorders, diabetes, asthma, meningitis, form a major part of the study into more correct diagnosis [8]. Several psychiatric disorders, such as Attention Deficit Hyperactivity Disorder (ADHD), schizophrenia, depression and Parkinson are also object of extensive investigation [9]. This paper presented as review about cancer medical data processing with data mining. The analysis is based on consideration of data mining methods. Further, the cancer medical data features are presented in this paper. Paper organisation is as follows as follows: In section 2 dataset of cancer is presented with consideration of features of medical data. In section 3 data mining methods are presented and presented overall conclusion and future perspective of medical data mining process.

## 2. Cancer Dataset

The idea of clinical proof is for most part bantered by clinical DM audits according to subject of demonstrative help. In particular,[10] presented overviews that basically corroborate each other, with the exception of the concepts protected by the variability of medical details. Indeed,[10] accept that fluctuation is clarified by changed nature, amount and incorrectness of demonstrative information, absence of a solitary clinician 's depiction, nonappearance of a solitary sickness jargon, just as helpless consistency with numerical thinking. Heterogeneity of clinical confirmation, on other hand[11], is seen by its different roots and by presence of various systems to acquire a similar estimation esteem. In table 1 attributes of data is presented for analysis of cancer data.

TABLE I. FEATURES OF MEDICAL DATA

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Imprecision | | | | | | | |
| High Dimensionality | | | | | | | |
| Different Data Types and Measurement Methods | | | | | | | |
| Incompleteness | | | | | | | |
| Non-Single Disease Terminology | | | | | | | |
| Inconsistency | | | | | | | |
| Non-Single Interpretation | | | | | | | |
| Sensitive Property and Use issues | | | | | | | |
| Temporal Components | | | | | | | |
| Poor Compatibility with the Logic of Mathematics | | | | | | | |

We connect the heterogeneity of data with the characteristics in the remainder of the present survey that preclude their consistent processing, or between records or attributes. This leads one to extend definition of heterogeneity with additional characteristics. In second part of this section, in open datasets published by the data sharing community, we outline the causes that enhanced this inherent heterogeneity of medical data.

*A. Cancer Dataset Features*

The precise existence of medical data demands specific handling, thus requiring further (implicit) attention for creation of diagnostic aid models. We outline these traits in the following paragraphs; others are explained in the particular diagnosis sense.

***Heterogeneity:*** In medical records, multiple sources of variability occur. Health records, such as biological data, photographs, signals, discrete values, interviews, are of multiple forms.

- To evaluate any discrete metrics, multiple scales exist. For starters, different measures used to calculate the academic quotient are Stanford-binet, Raven's matrices, Wechsler [12].
- Diseases can be distinguished by a form of heterogeneity; forms of cancer are incidentally checked from time to time[13].

***High dimensionality***. There may be several forms of data available for a particular patient. In addition, imaging and signal acquisition systems produce cumbersome archives and raw data. At a given time , for example, a brain MRI consists of several 2D images obtained during a full rotation of the MRI apparatus as slices of whole brain volume. As a result , multiple brain volumes, i.e. collections of 2D slices, are collected per patient to measure brain function over a given amount of time[14]

***Imprecision.*** It is subject to test findings, assumptions and experiments. It is determined by various measures which don't accomplish flawless. Further they are separated by sensitivity and accuracy [15].

***Incompleteness.*** Missing information in patient datasets is widely obtained. There could be an issue of missing values due to economic, ethical, medical or technological reasons[16].

***Inconsistency***: Accuracy of medical knowledge is not assured. Indeed, equipment for imaging as well as signal acquisition is prone to noise, which can lead to inaccurate results. While preprocessing pipelines exist, it is not assured that noise has been extracted properly.

**3. Data mining for data selection**

In data mining process data selection is observed as essential factor for regulation of elements in human body. The medical data consists of ~ 22,000 with identification of cancer pathway management and analysis of disease in human body. Thus, processing of medical data based on specific cancer type improves data mining performance. In data processing, medical data subset provides effective description about input and offers predictive performance. The appropriate selection of medical data provides clear understanding about medical data with computational cost reduction and increased prediction accuracy. Medical data includes several attributes and features for processing medical data. Based on this, relevant features provides information about additional classes in predictive models. This ensures that data is derived even from a small number of features based on different knowledge classes. Elimination of features decreases the amount of data resulting in better classification performance[17].

Various subsets can be obtained during the collection of the best feature sets. Many of the functions that are not overlapping and are fully interrelated are used in an ideal subset. In order to enhance detection capacity and accuracy of prediction, existence of completely applicable features is necessary[18]. There are three methods to gene selection in general: supervised, unsupervised and semi-supervised. The most widely used approach is

supervised gene selection. In the gene selection process, this method uses labelled data. The most important as well as most separable characteristics are chosen in this approach, given class functionality. The record class may mean whether the sample is cancerous or not, or can mean a subtype of cancer, in cancer records. The aim, therefore, is to find the right genes that are successful in distinguishing this class and defining it. However, when using external sources of information, the task is to mark data. The method of labelling is expensive and may not be absolutely accurate. Owing to the unintended omission of similar features, or the choice of irrelevant features, the unreliability of labels raises the possibility of over-fitting the learning process.

In unsupervised gene collection, there is no knowledge about the data mark. Therefore, additional details such as distribution, volatility, and data separability may be used in order to pick the right subset of functions. The data does not need the assistance of an expert or additional expertise without marking and can still work well even though no previous available knowledge. Word semi-directed quality determination is utilized where some information is marked and some is unlabeled. Labeled information is normally used to enhance edge between information focuses in different classifications, and unlabeled information is utilized to examine mathematical state of component space.

*A. Data Mining Methods for data selection*

In this section presented about existing literature conducted for analysis of methods involved in data mining. The analysis is examined based on research subject for processing medical data. The data mining methods examined in this review for cancer medical data processing are filtering, Wrapper, embedded, ensemble and hybrid methods.

*B. Filtering*

To pick features with less computational cost, filtering methods use statistical methods. A classifier or learning algorithm of some type is not used in these approaches. Based on four sorts of filter methods, capacity qualities are decided. They are Compatibility, Expertise, Dependency and Distance. Owing to the lack of class lables, unsupervised gene selection is a more difficult task. Tabakhi et al.[19] researched the unsupervised approach integrated with ant colony optimization. The recommended solution is known as UFSACO. This approach tries to find a subset of ideal functions through multiple iterations. Often gene selection procedures are supervised as well as use class labels as a guide. Method proposed by Mohammadi et al.[20], for example , involves the Maximum-Minimum Corr-entropy Criterion (MMCC) method to choose the maximal point in the dataset. This technique is stable, works efficiently and is resilient for noisy data and for issue of high diversity of data as well as outliers. In comparison, with the implementation of the filter system, MMCC showed decreased feature space. Xu et al.[21] developed a technique that, at its first level, decreases feature space by using a traditional technique called PCA to minimize data measurements. In the second step , in order to remove inappropriate features from subset of features, a correlation-based filtering approach along with a threshold is used. It should be remembered that PCA affects sense of functions, which may be a concern when analyzing data from microarrays. New approaches to filter methods are used as fitness function of a meta-heuristic method. In addition, Zheng and Wang have suggested a technique known as FS-JIME for data function selection[22]. Meta-heuristics is also used for the compilation of optimal medical dataset functions.

*C. Wrapper*

In a black box setting, wrapper methods use predictor. To calculate selected features subset, predictors efficiency is used as target function.

Wang et al.[23] suggested a strategy for accelerating sequential wrapper processes. A distance matrix classifier is utilized to decrease numerical complexity of sequential wrapper techniques in this method. Heuristic search algorithms test various subsets in order to refine the goal function. Moradi and Gholampour [24] proposed a optimization based meta-heuristic algorithm. The proposed model incorporates particle swarm optimization for identification of effective dataset with minimal correlation between the feature variables. The aim of local search method is to direct PSO search method to pick distinctive characteristics according to their correlation data. Also, for selection of optimal features KNN classifier is applied for classification of medical dataset. Wang et al. suggested weighted gene selection technique to find characteristics based on utility in classification and frequency occurrence in population based on 2 matrices [25]. Objective of this method are to reduce number of functions, optimize efficiency as well as minimize computing costs. The dataset features are selected based on the consideration of optimal parameters such as bacterial colony optimization for computational complexity reduction and increased discrete optimization features for classification. The right mix of features will increase the usefulness of classification considerably. A further analysis review by Pati et al. [26] proposed a tool for enhancing dataset usability. The suggested model implements a genetic algorithm for predicting the fitness function assessment based on the least number selection with an improvement in classification precision.

*D. Embedded Approach*

The embedded technique is a mechanism for gene selection in which gene selection is done while learning technique is executed. Guyon et al. Developed most popular embedded technique called SVM-RFE [27]. The SVMRFE integrates several SVM classifier with integration of linear equation based on separation of sample classes. The iterative backward selection includes least-weight features for reducing computation time for classification of data classes. The proposed model eliminates the features in the computation and predetermined

values. Recently, Guo et al. [28] developed a method for separability of class in multi-class data for feature selection, The data features are computed based on consideration of each criteria with evaluation of high selected features.

*E. Ensemble Approach*

By depending on various feature selection methods, an ensemble approach looks for a group of best feature sub-sets and then generates a merged outcome from these categories.

Two common approaches [29] can be found in the literature on ensemble methods:

1. Homogeneous distribution:
2. Heterogeneous Centralized:

*F. Homogeneous distribution:*

In this step, dataset is divided into n parts and distributed in parallel between n nodes. A single method of gene selection is executed in all nodes. Finally, using ensemble techniques, the rankings obtained from each node are merged. Pes et al. [30] focused on reduction of dataset with subset estimation with sampling placement. Based on the ranking algorithm each data is sampled for subset of features extraction for calculating higher rating. Finally, the optimal feature subsets are computed through integration methods. Similarly, Ebrahimpour et al. [31] developed a feature selection technique for handling high dimensional space medical data. In figure 1 homogeneous method adopted in data mining is presented.

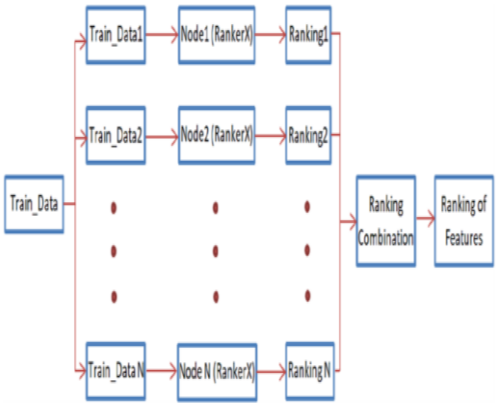


Fig.1. Homogeneous Method

*G. Heterogeneous Centralized:*

Obtained data collection is trailed by equal choice estimation with different boundaries or quality estimation of genes, in which at end of this process, every one gives output with best qualities. Mohapatra et al. [32] normalized the medical dataset with max–min normalization method. Then, based on analysis MCSO (Modified Cat Swarm Optimization) method, with optimal feature subset for normalized dataset. Elyasigomari et al. [33] proposed a new hybrid optimization method, defined as COA-GA. Proposed model involved in leveraging discovered cuckoo optimization method instead of traditional genetic method trend, proposed model incorporates clustering of data and gene selection. At first, the combination of evolutionary process cluster size is increases based on consideration of 100 iterations. In figure 2 process involved in heterogeneous methods is illustrated.

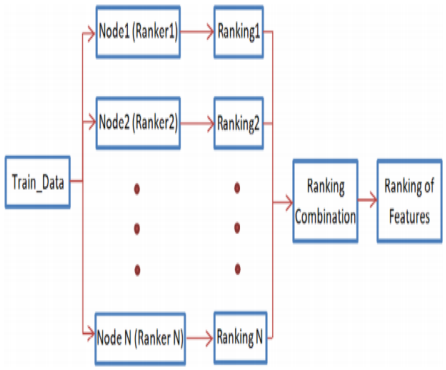


Fig.2. Heterogeneous Method

*F. Hybrid Approach*

Two or more algorithms for gene selection (filter, wrapper, embedded or ensemble) are mixed in a certain order in hybrid approaches. Composite approach inherits advantages of each blended human approach. In order to increase productivity with improved computing performance, a hybrid approach utilises various measurement parameters at different stages of the quest. A mixture of filter and wrapper methods is the most prevalent hybrid process. Lv et al.[34] used pre-selection features in the advanced mRMR filter algorithm, specifying two criteria:

increasing precision and decreasing number of features. They develop a multi-objective model known as MOEDA as the first criterion is comparatively more relevant. MOEDA is a type of algorithm for distribution calculation.

Elyasigomari et al. [35] presented a meta-heuristics optimization approach for medical data feature selection. The proposed model consists of two optimization model such as Cuckoo optimization as well as Harmony Search algorithms integrated with wrapper technique. Lu et al. [36], integrates MIM (Mutual Information Maximization) with combination of filter as well as wrapper components with wrapper technique. Jain et al. [37] utilized medical feature selection stated as CSF-iBPSO. The medical features incorporates multivariant filtering technique with consideration of feature subset for improving model. Initially,

Dashtban et al. uses Fisher score which is a single variable filter technique to rate and select 500 genes with high score [38].It should be recalled that Fisher score is merely a number given to each gene, providing degree to which gene will discriminate between different types. In chosen features, 1st subset is given to wrapper technique to choose final subset classes with high separability degree. A variation of the standard BAT algorithm is the wrapper approach used in [39]. Venkataraman et al. [40] developed technique for medical data feature subset classification and class. Analysis is based on the computation of relevant feature with estimation of Symmetric Uncertainty (SU) values. In next section, optimal medical feature subset are estimated based on the genetic algorithm. The classification is based on the estimation of SU medical data classes and feature. The large value of SU exhibits large features with high weight are selected for processing.

TABLE II. OVERALL SUMMARY OF DATA MINING METHODS

| Data Mining Method | Advantage | Disadvantage |
|---|---|---|
| Filter | Scalable and quick<br>Classifier-independent<br>Speedier than wrapper strategies<br>Better complexity of computing | Less precision attributable to a classifier's lack of attention<br>Ignores relationship between features/ variables<br>May involve redundancy.<br>Overfitting |
| Wrapper | Classifier relationship<br>Interaction between characteristics<br>Better precision | Huge complexity of computing<br>Costly working<br>Local optimisation tendencies |
| Embedded | Better accuracy and performance compared to filters<br>Less computational complexity than wrapper approaches More emphasis is placed on function relationships | Dependent to classifier |
| Ensemble | Less overfitting propensity<br>For high-dimensional data, greater scalability<br>Stability | It is difficult to grasp the combination of classifiers. |
| Hybrid | More efficient than filter methods<br>Less tendency to overfitting<br>Lower computing cost | Depends on the classifier<br>Based on the mixture of various algorithms for gene selection. |

In table 3 shows analysis of leukemia dataset for data mining is presented.

Also, in [41] combined wrapper and ensemble classifier for estimation of medical data features. Further, the medical features are computed with utilization of correlation, chi-square, information gain, relief methods and gain ratio. Then, in the second step, MCSO (Multi-objective Simplified Swarm Optimization) is given to genes selected from previous step. Agarwalla and Mukhopadhyay developed bi stage hierarchical swarm based method for combination of two methods [42]. MFDPSO (Medical data features defined as multi-fitness discrete PSO) are used in the proposed model estimation. In table 2 data mining methods are overall performance summary is presented.

TABLE III. ANALYSIS OF LEUKEMIA DATASET

| Data | Ref | Method | Samples | Class | Accuracy | Classifier |
|---|---|---|---|---|---|---|
| Leukemia | 21 | Filter | 72 | 2 | 99.7 | SVM |
| | 26 | Wrapper | 38 | 2 | 100 | SVM |
| | 43 | Ensemble | 64 | 2 | 95.23 | Random Forest |
| | 33 | Hybrid | 73 | 2 | 100 | SVM |
| | 44 | Hybrid | 72 | 2 | 100 | SVM |

**4. Conclusion**

The review paper is based on the classification of cancer data analysis for classification. Data analysis is based on classification of several methods such as filter, wrapper, embedded, ensemble and hybrid. Filter techniques are fast, but in comparison to other techniques, the consistency of their selected characteristics is poor. Wrapper methods, on other hand , prefer useful subsets of functions, but they come at a large expense of computation and are barely seen on their own. Hybrid techniques are most common in gene selection field, as they can combine strengths of other methods of selection and thus perform good. Improving precision, increasing classifier performance and reducing statistical complexity are underlying aims of improving gene selection methods. Another essential goal of most of reviewed articles is to minimize number of features which are significant in disease prediction. Second argument involves methods of data integration that are an significant issue in study of multi-omics datasets. Data convergence may consider relationships between different levels of genetic products in science, generating better, more precise and more predictable results. Most popular integration method call transform based integration also known as intermediate integration. Intermediate integration approaches, intermediate graph is often more commonly used when it recognizes interaction between various levels of genetic products. Data mining method subjected to challenge related to selection of features for classification. In future, data mining method can be developed based on consideration of various features of cancer dataset.

**References**

1. Bray F, Ferlay J, Soerjomataram I, Siegel R. L, Torre L A and Jemal A, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries", *CA: a cancer journal for clinicians*, vol.68, no.6, pp.394-424, 2018.
2. Statistics O. N, "Deaths registered in England and Wales (series DR)", *Newport: Office for National Statistics*, 2012.
3. Kristensen V. N, Lingjærde O. C, Russnes H. G, Vollan H. K. M, Frigessi A and Børresen-Dale A. L, "Principles and methods of integrative genomic analyses in cancer", *Nature Reviews Cancer*, vol.14, no.5, pp.299-313, 2014.
4. PhridviRaj M. S. B and GuruRao C. V, "Data mining–past, present and future–a typical survey on data streams", *Procedia Technology*, vol.12, pp.255-263, 2014.
5. Bellazzi R, Ferrazzi F and Sacchi L, "Predictive data mining in clinical medicine: a focus on selected methods and applications", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol.1, no.5, pp.416-430, 2011.
6. Giudici P and Figini S, "Applied data mining for business and industry", *Chichester: wiley*, pp.147-162, 2009.
7. Lyu B and Haque A, "Deep learning based tumor type classification using gene expression data", *Proceedings of the ACM international conference on bioinformatics, computational biology, and health informatics*, pp. 89-96, 2018.
8. Woo C. W, Chang L. J, Lindquist M. A and Wager T. D, "Building better biomarkers: brain models in translational neuroimaging", *Nature neuroscience*, vol.20, no.3, 2017.
9. Abraham A, Milham M. P, Di Martino A, Craddock R. C, Samaras D, Thirion B and Varoquaux G, "Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example", *NeuroImage*, vol.147, pp.736-745, 2017.
10. Cios K. J and Moore G. W, "Uniqueness of medical data mining", *Artificial intelligence in medicine*, vol.26, no.1, pp.1-24, 2002.
11. Harrison Jr, J. H, "Introduction to the mining of clinical data", *Clinics in laboratory medicine*, vol.28, no.1, pp.1-7, 2008.
12. Wagholikar K. B, Sundararajan V and Deshpande A. W, "Modeling paradigms for medical diagnostic decision support: a survey and future directions", *Journal of medical systems*, vol.36, no.5, pp.3029-3049, 2012.
13. Miotto R, Wang F, Wang S, Jiang X and Dudley J. T, "Deep learning for healthcare: review, opportunities and challenges", *Briefings in bioinformatics*, vol.19, no.6, pp.1236-1246, 2018.
14. Retico A, Gori I, Giuliano A, Muratori F and Calderoni S, "One-class support vector machines identify the language and default mode regions as common patterns of structural alterations in young children with autism spectrum disorders", *Frontiers in neuroscience*, vol.10, 2016.
15. Majid A, Ali S, Iqbal M and Kausar N, "Prediction of human breast and colon cancers from imbalanced data using nearest neighbor and support vector machines", *Computer methods and programs in biomedicine*, vol.113, no.3, pp.792-808, 2014.
16. Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang J. F and Hua L, "Data mining in healthcare and biomedicine: a survey of the literature", *Journal of medical systems*, vol.36, no.4, pp.2431-2448, 2012.
17. Gheyas I. A and Smith L. S, "Feature subset selection in large dimensionality domains", *Pattern recognition*, vol.43, no.1, pp.5-13, 2010.

18. Ang J. C, Mirzal A, Haron H and Hamed H. N. A, "Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection", *IEEE/ACM transactions on computational biology and bioinformatics*, vol.13, no.5, pp.971-989, 2015.
19. Tabakhi S, Moradi P and Akhlaghian F, "An unsupervised feature selection algorithm based on ant colony optimization", *Engineering Applications of Artificial Intelligence*, vol.32, pp.112-123, 2014.
20. Mohammadi M, Noghabi H. S, Hodtani G. A and Mashhadi H. R, "Robust and stable gene selection via maximum–minimum correntropy criterion", *Genomics*, vol.107, no.2, pp.83-87, 2016.
21. Xu J, Mu H, Wang Y and Huang F, "Feature genes selection using supervised locally linear embedding and correlation coefficient for microarray classification", *Computational and mathematical methods in medicine*, 2018.
22. Zheng K and Wang X, "Feature selection method with joint maximal information entropy between features and class", *Pattern Recognition*, vol.77, pp. 20-29, 2018.
23. Wang A, An N, Chen G, Li L and Alterovitz G, "Accelerating wrapper-based feature selection with K-nearest-neighbor", *Knowledge-Based Systems*, vol.83, pp.81-91, 2015.
24. Moradi P and Gholampour M, "A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy", *Applied Soft Computing*, vol.43, pp.117-130, 2016.
25. Wang H, Jing X and Niu B, "A discrete bacterial algorithm for feature selection in classification of microarray gene expression cancer data", *Knowledge-Based Systems*, vol.126, pp. 8-19, 2017.
26. Pati S. K, Sengupta S and Das A. K, "Improved Genetic Algorithm for Selecting Significant Genes in Cancer Diagnosis", *Progress in Advanced Computing and Intelligent Engineering*, pp. 395-405, 2018.
27. Guyon I, Weston J, Barnhill S and Vapnik V, "Gene selection for cancer classification using support vector machines", *Machine learning*, vol.46, no.1, pp.389-422, 2002.
28. Guo S, Guo D, Chen L and Jiang Q, "A centroid-based gene selection method for microarray data classification", *Journal of theoretical biology*, vol.400,pp.32-41, 2016.
29. Seijo-Pardo B, Porto-Díaz I, Bolón-Canedo V and Alonso-Betanzos A, "Ensemble feature selection: homogeneous and heterogeneous approaches", *Knowledge-Based Systems*, vol.118, pp.124-139, 2017.
30. Pes B, Dessì N and Angioni M, "Exploiting the ensemble paradigm for stable feature selection: a case study on high-dimensional genomic data", *Information Fusion*, vol.35, pp.132-147, 2017.
31. Ebrahimpour M. K, Nezamabadi-Pour H and Eftekhari M, "CCFS: A cooperating coevolution technique for large scale feature selection on microarray datasets", *Computational biology and chemistry*, vol.73, pp.171-178, 2018.
32. Ebrahimpour M. K, Nezamabadi-Pour H and Eftekhari M, "CCFS: A cooperating coevolution technique for large scale feature selection on microarray datasets", *Computational biology and chemistry*, vol.73, pp.171-178, 2018.
33. Elyasigomari V, Mirjafari M. S, Screen H. R and Shaheed M. H, "Cancer classification using a novel gene selection approach by means of shuffling based on data clustering with optimization", *Applied Soft Computing*, vol.35, pp. 43-51, 2015.
34. Lv J, Peng Q, Chen X and Sun Z, "A multi-objective heuristic algorithm for gene expression microarray data classification", *Expert Systems with Applications*, vol.59, pp.13-19, 2016.
35. Elyasigomari V, Lee D. A, Screen H. R and Shaheed M. H, "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", *Journal of biomedical informatics*, vol.67, pp.11-20, 2017.
36. Lu H, Chen J, Yan K, Jin Q, Xue Y and Gao Z, "A hybrid feature selection algorithm for gene expression data classification", *Neurocomputing*, vol.256, pp.56-62, 2017.
37. Jain I, Jain V. K and Jain R, "Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification", *Applied Soft Computing*, vol.62, pp.203-215, 2018.
38. Dashtban M, Balafar M and Suravajhala P, "Gene selection for tumor classification using a novel bio-inspired multi-objective approach", *Genomics*, vol.110, no.1, pp.10-17, 2018.
39. Li J, Dong W and Meng D, "Grouped gene selection of cancer via adaptive sparse group lasso based on conditional mutual information", *IEEE/ACM transactions on computational biology and bioinformatics*, vol.15, no.6,pp.2028-2038, 2017.
40. Venkataraman S and Selvaraj R, "Optimal and novel hybrid feature selection framework for effective data classification", *Advances in Systems, Control and Automation*,pp.499-514, 2018.
41. Lai C. M, "Multi-objective simplified swarm optimization with weighting scheme for gene selection", *Applied Soft Computing*, vol.65, pp.58-68, 2018.

42. Agarwalla P and Mukhopadhyay S, "Bi-stage hierarchical selection of pathway genes for cancer progression using a swarm based computational approach", *Applied Soft Computing*, vol.62, pp.230-250, 2018.
43. Ram M, Najafi A and Shakeri M. T, "Classification and biomarker genes selection for cancer gene expression data using random forest", *Iranian journal of pathology*, vol.12, pp.4, 2017.
44. Motieghader H, Najafi A, Sadeghi B and Masoudi-Nejad A, "A hybrid gene selection algorithm for microarray cancer classification using genetic algorithm and learning automata", *Informatics in Medicine Unlocked*, vol.9, pp.246-254, 2017.