FinJeHeW Finnish Journal of eHealth and eWelfare

SCIENTIFIC PAPERS

VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

Cancer prediction using graph-based gene selection and explainable classifier

Mehrdad Rostami[1], Mourad Oussalah[1,2]

[1] Center of Machine Vision and Signal Processing (CMVS), Faculty of Information Technology, University of Oulu, Oulu, Finland; [2] Research Unit of Medical Imaging, Physics, and Technology, Faculty of Medicine, University of Oulu, Oulu, Finland

Mehrdad Rostami, MHSc, Center of Machine Vision and Signal Processing (CMVS), Faculty of Information Technology, University of Oulu, P.O. Box 4500, FI-90014 University of Oulu, FINLAND. Email: Mehrdad.Rostami@oulu.fi

Abstract

Several Artificial Intelligence-based models have been developed for cancer prediction. In spite of the promise of artificial intelligence, there are very few models which bridge the gap between traditional human-centered prediction and the potential future of machine-centered cancer prediction. In this study, an efficient and effective model is developed for gene selection and cancer prediction. Moreover, this study proposes an artificial intelligence decision system to provide physicians with a simple and human-interpretable set of rules for cancer prediction. In contrast to previous deep learning-based cancer prediction models, which are difficult to explain to physicians due to their black-box nature, the proposed prediction model is based on a transparent and explainable decision forest model. The performance of the developed approach is compared to three state-of-the-art cancer prediction including TAGA, HPSO and LL. The reported results on five cancer datasets indicate that the developed model can improve the accuracy of cancer prediction and reduce the execution time.

Keywords: artificial intelligence, supervised machine learning, medical informatics applications, classification, decision trees

Introduction

Worldwide, cancer remains the leading cause of death for both men and women. It is estimated that one out of every six deaths worldwide are caused by cancer, which makes it the leading cause of death globally [1], and around 19.3 million of new cancer cases and 10 million cancer deaths are recorded worldwide in 2020 alone. Therefore, improving cancer prediction becomes crucial for increasing survival chances by providing opportunity for early diagnosed patients to receive appropriate treatment [2-4].

Microarray technology has already been employed in several healthcare applications to advance in cancer prediction. Especially, Deoxyribonucleic Acid (DNA) microarray data [5], biomarker analysis [6], Ribonucleic Acid (RNA) sequencing [7] and Polymerase Chain Reaction (PCR) [8,9] to classify input samples into predefined groups (e.g. benign or cancerous). These groups could be a specific diagnosis (e.g. melanoma) or a diagnostic category (e.g. malignant versus benign) [10]. On the other hand, the availability of large volume of DNA microarray data enabled the development of tailored Machine Learning (ML) algorithms [11,12].

Nevertheless, there are at least two significant challenges that constraint the large-scale clinical deployment of ML techniques for cancer DNA microarray data classification/prediction tasks. The first one is associated with the high-dimensionality of microarray data where the number of genes is much greater than the number of patterns [13,14], and many genes may be irrelevant or redundant to cancer prediction or classification task. The second one is associated with the explainability issue where clinicians and health authorities are reticent to rely on hardly explainable / transparent results raised by the employed complex ML and black-box like systems [15-17]. Indeed, in many healthcare applications, it is necessary to know how the prediction model made a specific prediction, allowing the healthcare stakeholders (e.g., physicians, specialists, patients, researchers and public) to trust the model. Explainability here refers to machine learning approaches that can provide human-understandable explanation for their models' behavior.

High-dimensional cancer microarray dataset with small number of patterns leads to the well-known problem of “curse of dimensionality”. Gene selection is a powerful and efficient technique in DNA microarray data analysis to deal with such a challenge and it is among popular techniques that enable eliminating irrelevant and/or redundant genes as well as enhancing computational complexity, learning efficiency and generalization capabilities [18-20]. Whereas various techniques have been suggested to handle the lack of explainability in ML in a way to earn clinician trusts and to achieve positive clinical impact, although with a limited success. This includes diagnostic methods that aim to provide model understanding of ML system such as linear approximation models, gene importance visualization, saliency map, sensitivity analysis, LIME, Anchors, instance-based explanation methods (e.g., data instances as [23]), among others. See, [21,22] for an overview. The handling of these two aspects (high dimensionality and explainability) in DNA microarray data analysis for cancer prediction / classification tasks showed mixed results in health literature. See Table 1 for a discussion of some relevant works in this field, and no satisfactory solution has been universally accepted yet. This calls for further research in this issue. For instance, it emerged from our literature survey that models that achieved state-of-the-art results in terms of cancer prediction are rather based on deep learning approaches [24-28] without any explainability issues. As a result, developing new models for cancer prediction and classification using DNA microarray data that are both explainable and interpretable, as well as highly accurate, is considered of paramount importance in the eHealth community.

Table 1. Outlining the reviewed cancer prediction models and used techniques.

Paper	Technique	Task	Dataset	Accuracy	Explainability
Al-Betar et al. [29]	Gene selection and SVM classifier	Cancer classification	DNA microarray dataset (Breast, Colon, Lymphoma and etc.)	Medium	Low
Rostami et al. [30]	PSO and feature selection	Cancer classification	DNA microarray dataset (Colon, Lung Cancer and etc.)	High	Low
Gu et al. [17]	Case-based ensemble learning	Breast cancer prediction	Breast cancer recurrence cases	Low	High
Nayak et al. [31]	Artificial Neural Network and PSO	Cancer prediction	WBC Breast Cancer, Lung Cancer and Cervical cancer	Medium	Low
Ghiasi et al. [32]	DT	Breast cancer classification	Breast Cancer Database	Low	High
Lai and Huang [3]	Multi-filter ensemble technique and simplified swarm optimization	Cancer classification	Microarray gene expression datasets (Brain Tumor, Lung Cancer, and etc.)	Medium	Low
Maleki et al. [2]	KNN and genetic algorithm	Lung cancer prognosis	lung cancer dataset	Medium	Medium
Babu et al. [4]	Cellular learning automata with SVM, Naive Bayes and KNN	Microarray data classification	DNA microarray datasets (Prostate tumor, SRBCT and etc.)	Medium	Medium
Hamid et al. [33]	Gene Selection, PSO and SVM	Cancer Classification	Breast Cancer and Lymphography datasets	Medium	Low
Xiao et al. [28]	Deep Learning	Cancer diagnosis	RNA-seq datasets (LUAD, STAD and BRCA)	High	Low
Koh et al. [27]	Deep Learning	Breast cancers detection	Chest CT scans	High	Low
Zheng et al. [34]	Dual latent representation learning	Microarray data classification	DNA microarray Datasets (Breast, Lung Cancer and etc.)	Medium	Low
Doppalapudi et al. [26]	Deep Learning	Lung cancer survival period prediction	Surveillance, Epidemiology, and End Results Dataset	High	Low
Alomari et al. [35]	Gray Wolf Optimizer and SVM	Cancer classification	DNA microarray Datasets (Colon, CNS, Lung Cancer and etc.)	Medium	Low
Chai et al. [25]	Deep learning	Cancer prognosis prediction	The Cancer Genome	High	Low
Liu et al. [24]	Deep learning	Breast cancer prediction	Breast cancer CT images	High	Low

The major aim of this work is to introduce a novel graph-based gene selection method integrated with an explainable classifier that satisfies the explainability requirement. In the developed model a prediction technique based on Decision Tree (DT) is developed to improve the explainability of the learning algorithm. Transparency of DT makes it widely accepted in healthcare application that requires a comprehension of both the classifier structure and its prediction. This developed cancer prediction model has a number of innovations compared with previously surveyed methods:

1. Unlike black-box deep learning-based cancer prediction models [24-28], the use of DT in our scheme enforces explainability due to its acknowledged transparency.

2. The proposed method uses a novel graph clustering-based technique to identify similar genes. This, unlike other clustering algorithms such as k-means [36] and fuzzy clustering [37], automatically identifies the number of clusters, which does not need to be specified in advance.

3. The proposed method uses a graph-based approach for gene selection that is faster and more accurate than nature-inspired methods such as MOSSO [38], AutoGeneS [39], C-HMOSHSSA [40].

4. The developed gene selection satisfies both objectives of genes selection Minimum redundancy and Maximum relevance [41-43] in its search strategy.

5. In comparison with previous wrapper-based gene selection approaches [44-46], the developed method does not employ any learning model in its gene selection process.

Material and methods

In this section, an innovative explainable prediction model for cancer prediction is proposed by incorporating the concept of Gene Selection with Explainable Classifier (GSEC). In the developed GSEC, in order to handle the high occurrence of irrelevant and redundant genes in DNA cancer microarray data analysis, which decreases the prediction accuracy [47-49], a gene selection phase is added to the main prediction phase with the aim of removing irrelevant and redundant genes.

Since their ability to encode similarity relationships among data, graph-based models such as graph embedding [20], graph clustering [50], and semi-supervised learning [51] have played an important role in machine learning tasks. Through the use of graph-based models for cancer prediction, a universal and versatile framework can be created that reflects the complex relationships and structure of the gene space. for this purpose, a novel-graph-based and explainable cancer prediction model is developed in this paper. In overall, the developed model consists of four main steps. In the first step, the primary genes are represented as a graph. In the second step, a graph clustering algorithm is utilized to find gene clusters. Next, high score genes are selected from each cluster to generate the final gene set. In the fourth step, a DT-based prediction technique is developed to improve the explainability of the learning algorithm. Figure 1 provides a high-level graphical illustration of the developed cancer prediction model.

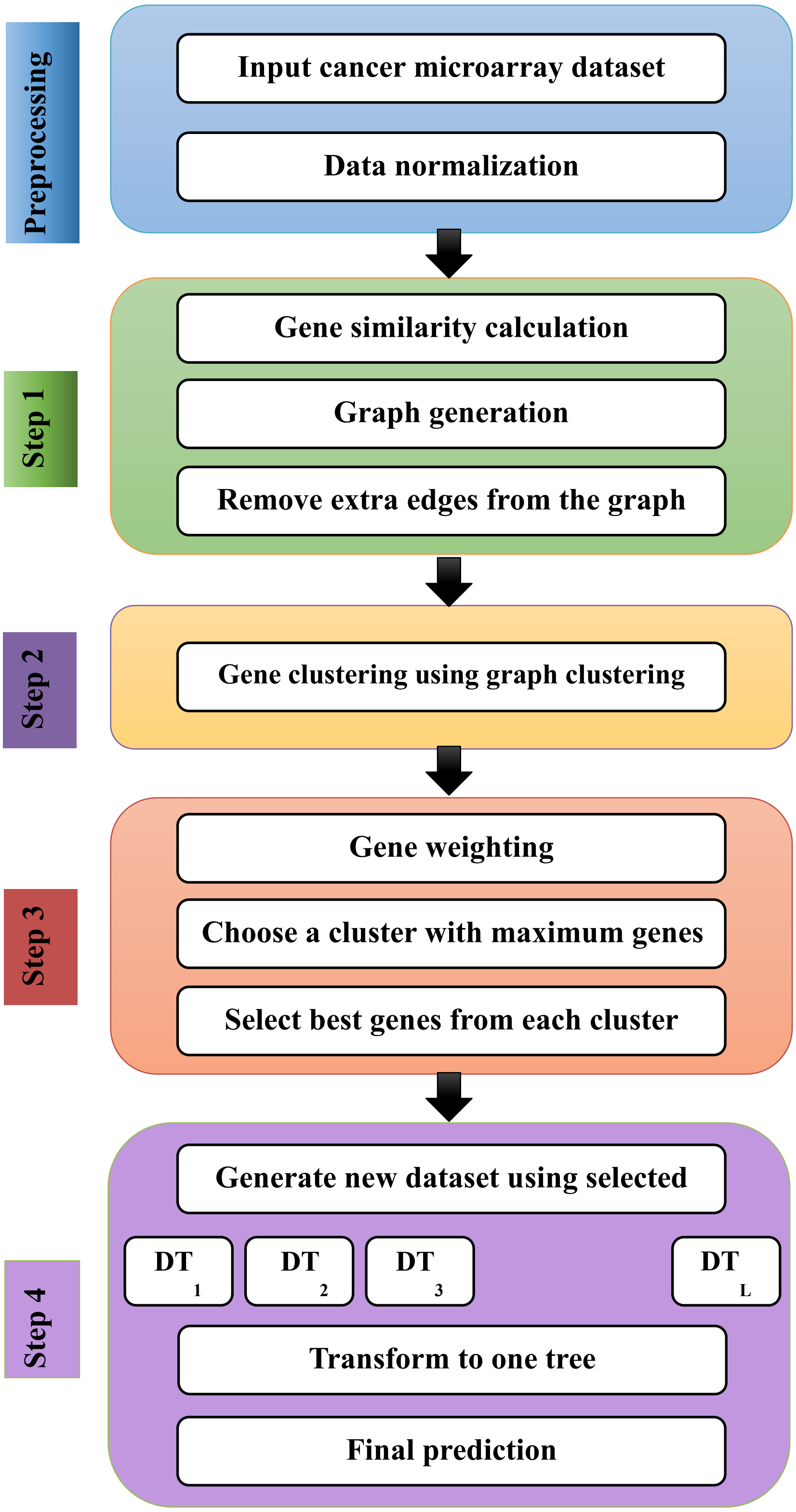

Preprocessing
Input cancer microarray dataset
Data normalization
Step 1
Gene similarity calculation
Graph generation
Remove extra edges from the graph
Step 2
Gene clustering using graph clustering
Step 3
Gene weighting
Choose a cluster with maximum genes
Select best genes from each cluster
Step 4
Generate new dataset using selected
DT_1
DT_2
DT_3
DT_L
Transform to one tree
Final prediction

Figure 1. The overall schema of the developed method for cancer prediction.

Gene similarity calculation

The aim of the first step is to represent the gene space of the DNA microarray dataset to an undirected and weighted graph. In this representation, each gene is demonstrated using a node and the weight of each edge is the similarity between their corresponding genes. In this graph representation, the Pearson correlation coefficient measure [52] is employed to calculate the similarity values between different genes. Using this similarity measure, the similarity values are in the range of [0 1], where identical gene expressions take value 1 and two completely dissimilar genes take value 0.

This similarity measure maps the gene space of a microarray dataset into a fully weighted and connected graph. To make the graph sparser, the edges with associated weights lower than some threshold value θ are removed. θ is an adjustable parameter that takes values in the unit interval [0 1]. When θ value is small (resp. large), more (resp. fewer) edges will be considered in the next steps. In our experiments θ is empirically set to 0.6, which is found to work well.

Gene clustering

One of the important goals of the developed method is to select subset of genes that have least similarity to each other and the highest similarity to the target class. In order to achieve the first objective, genes are grouped into similar clusters. For this purpose, the fast graph clustering algorithm [53] is applied to cluster the genes. This graph clustering algorithm can quickly detect gene communities in a high-dimensional cancer microarray dataset, due to the use of fast parallel model for community detection. Since the generated gene graph is sparse enough, this algorithm is faster than previous methods for gene clustering such as [54] and [55].

Final gene selection

In order to achieve the second objective (similarity with the target class), a selection strategy based on Pearson similarity measures and gene scoring is developed. Specifically, at each iteration, the cluster with the highest number of genes is identified and then among the existing genes in this maximum cluster, the most representative genes are selected using an approach that seeks to maximize Pearson similarity scores while ensuring gene relevance. Finally, the remaining genes available in this cluster are removed from the set of genes and the process is repeated for the remaining genes in this graph. In short, using the concept of gene relevance, the genes of each cluster are ranked and then using Pearson similarity measure, non-redundant genes are considered to represent the initial genes of this cluster. In the proposed method, the search process is guided in such a way that at least one gene is selected per cluster. As a result, the selected genes satisfy both conditions: maximum relevancy and minimum redundancy.

More specifically, in the proposed method, the incremental gene weighting mechanism [56], which yields a score in the unit interval, is utilized for gene scoring. The purpose of gene scoring is to select a representative gene that is most relevant to the target class. Therefore, a gene with the highest score in the cluster is added to the selected gene set. After eliminating this gene from the cluster, the next gene with the highest score is considered as the candidate gene and the average similarity of this gene with the previously selected genes is calculated using Pearson similarity measure. If the average similarity value of the candidate gene and the previously selected genes is lower than some predefined threshold δ, this gene is added to the selected gene set and the next relevant gene from this cluster is considered as a candidate gene. Otherwise, if the average Pearson similarity between the candidate gene and the previously selected genes was higher than δ, all remaining genes in this cluster are removed from the gene graph. This process is repeated again for all remaining genes after deleting the genes present in the extracted cluster. Then, as mentioned earlier, for each cluster, relevant gene is selected by performing the incremental gene scoring technique.

Explainable cancer prediction

In this step, a novel decision tree-based predication technique is developed to improve the explainability of cancer prediction. To address the limited prediction performance of DT model to capture complex interactions between input genes, which result in important biases, due to the nearsightedness characteristic of their induction model, Decision Forest (DF), or ensemble of DTs, were employed. This Ensemble model is a powerful technique to combine the results of several prediction models into a single decision. This typically yields more accurate prediction capabilities due to the diversity of the constituent individual classifiers and their performances. Moreover, in this paper, a new technique for converting DF into a single DT is developed. Based on the original DF, the final decision tree approximates its prediction accuracy, while providing explainable and faster classification. As compared to previous prediction models, the developed model can be applied to all sizes of forests and does not need complicated hyperparameter setting. In the developed prediction model, first, a conjunction set that represents the original DF is created and then a DT that forms the conjunction set in a tree structure is built.

More formally, suppose a dataset with n samples, m genes, and c classes (D={(x_i,y_i)}|D|n,x_i∈R^m,y_i∈1,…,c). In DF a m-dimensional gene vector is mapped into a c-dimensional probability vector by collecting different progressively increasing functions as below:

$$\varphi(x_i) = \frac{\sum_{l=1}^{|L|} t_l(x_i)}{|L|}, t_l \epsilon R^c \tag{1}$$

where, L represents the set of DTs contained within the DF. The main goal of the developed model is to build a new explainable tree that approximates the prediction function of a given DF. This new explainable tree t┴^ is calculated as follows:

$$\forall x_i, \hat{t}(x_i) \approx \varphi(x_i) \tag{2}$$

This method relies on the idea that both DF and DT can be demonstrated as sets of disjoint rules. As an organization tree structure is constructed, conjunctions can be organized to provide accurate and precise classifications. Generating an explainable DT that approximates the prediction function of a given DF is the main aim of this formula. It should be noted that the developed prediction model does not measure any interdependencies between different trees. For this reason, it is more appropriate to employ this technique for autonomous DFs.

Afterwards, the DF is partitioned into a set of rule conjunctions that each is associated with an appropriate ultimate outcome of DF. In the Hierarchical form of a DT, t_i that is part of the DF of T, is ignored and the tree is considered as a series of rule conjunctions CS_i. Conjunction in CS_i is a set of rules c_ij mapped to y┴^_(c_ij), a K-dimensional vector that K demonstrates the number of classes and each cell donates the probability of the respective label. It is possible to combine two conjunctions CS_1 and CS_2 using a Cartesian product where conjunction c_1 is combined with conjunction c_2 to generate a new conjunction (c_1j∧c_2j,y┴^_(c_1j)+y┴^_(c_2j)).

Figure 2 indicates the pseudo-code of the developed cancer prediction model.

```
Algorithm: Cancer prediction model based on Gene Selection with Explainable Classifier (GSEC)

Input    $D_T$: Input Cancer microarray data
         $\theta$: Threshold for edge removing
         $\delta$: Threshold for final gene selection
Output   Prediction model and explanations

Begin algorithm
Data Normalization
Gene similarity calculation using Pearson measure
$\boldsymbol{Graph}$ = Generate a primary graph of genes using calculated gene similarities
$\boldsymbol{Sparser\_Graph}$ = Remove edges which their associated weights are less than threshold $\theta$
$\boldsymbol{Clutstered\_Graph}$= $\boldsymbol{Fast\_Graph\_Clusrering}(Sparser\_Graph)$
$\boldsymbol{Gene\_Weight} = \boldsymbol{Feature\_Weighting}(D_T)$
$\boldsymbol{G'} = \{\,\}$
Do
    $\boldsymbol{Candidate\_Cluster} = \boldsymbol{Max\_Cluster}(Clutstered\_Graph)$
    $\boldsymbol{Candidate\_Gene} = \boldsymbol{Max\_Gene\_in\_Candidate\_Cluster}(Gene\_Weight)$
    Do
        $\boldsymbol{G'}$=$\boldsymbol{G'}$ +$\boldsymbol{Candidate\_Gene}$
    $\boldsymbol{Candidate\_Cluster} = \boldsymbol{Candidate\_Cluster} - \boldsymbol{Candidate\_Gene}$
    $\boldsymbol{Candidate\_Gene} = \boldsymbol{Max\_Gene\_in\_Candidate\_Cluster}(Gene\_Weight)$
    While ($\boldsymbol{Average\_Pearson\_Similarity}(Candidate\_Gene, G') < \delta$) Then
    $\boldsymbol{Clutstered\_Graph} = \boldsymbol{Clutstered\_Graph} - \boldsymbol{Candidate\_Cluster}$
While ($\boldsymbol{Clutstered\_Graph} \neq \emptyset$)
Report $\boldsymbol{G'}$ as a final gene set
Generate new Microarray Dataset using the selected gene set ($\boldsymbol{G'}$)
Evaluating different decision trees
Transform different Trees to a single tree for final explanation
Final Cancer prediction and explanations
End algorithm
```

Figure 2. Pseudo-code of the proposed explainable cancer prediction model.

Results

To measure the efficiency of the developed cancer prediction model, various experiments are performed. The performance of the developed model is compared with three very recent methods of filter-based cancer microarray data classification: Tabu Asexual Genetic Algorithm (TAGA) [57], Harmonize Particle Swarm Optimization (HPSO) [33], Least Loss (LL) [58].

Moreover, the experiments in this study use a variety of datasets with different properties to demonstrate the effectiveness and robustness of the developed approach. These microarray data consist of Colon, Leukemia, SRBCT, Prostate Tumor, and Lung Cancer. The primary characteristics of these datasets are detailed in Table 2. Both Colon and Leukemia datasets are publicly available from the Universidad Pablo de Olavide's Bioinformatics Research Group [59], while the SRBCT dataset, Prostate Tumor dataset, and the Lung Cancer dataset are available at Vanderbilt University's Gene Expression Model Selector [60]. Leukemia, Colon, and Prostate Tumor datasets are binary classification problems and SRBCT and Lung Cancer datasets represent multi-class problems whose task it is to classify different tumor types. In Table 2, #Genes refers to the number of initial genes used to build the prediction model, #Patterns refers to the number samples (i.e. patients) and #Class indicates to the number groups (e.g. benign or cancerous and etc.). In order to achieve more precise and acceptable results, the results are obtained over ten separate and autonomous runs. Microarray datasets are randomly divided into train data (66% of the initial data) and test data (34% of the initial data) for each run. Train part is utilized to model generation, while test data is used to measure the model. The comparative methods are evaluated on the same training and testing sets in order to ensure fairness. The efficiency of the developed cancer prediction model is evaluated in terms of classification accuracy, number of selected genes and execution time. The classification accuracy of a prediction model can be summarized as the number of correct predictions divided by the total predictions. Furthermore, the number of selected genes refers to the number of genes used by each model for prediction task.

Table 3 shows the average accuracy over ten independent runs of the different cancer prediction models. The reported results show that in all cases the developed model performs better than the recently state-of-the-art approaches. For example, for the Lung Cancer dataset, the proposed method yields a 91.82% classification accuracy while TAGA [57], HPSO [33], and LL [58] reported 90.19%, 90.83%, and 89.61%, correspondingly.

Furthermore, Figure 3 records the number of selected genes of the different gene selection approaches for each microarray dataset. It can be seen that, in general, all compared models have succeeded in significantly reducing the number of initial genes by choosing only a small number of the original genes. Especially, for Colon, SRBCT, Leukemia, Prostate Tumor, Lung Cancer, the developed model achieved the best result in terms of size of the selection, among the three other alternative approaches, resulting averagely in a selection size of 13.5, 17.4, 20.3, 22.8 and 28.3, respectively.

In the next experiment, different gene selection models are compared in term of execution times. In these experiments, corresponding execution times (in second) for each gene selection method are shown in Figure 4. The reported results revealed that among the state-of-the-art gene selection models, the proposed model has the lowest average execution time, which testifies of its computational efficiency.

Sensitivity analysis

Like many other prediction models, in the developed model, two parameters should be determined: edge weight threshold θ and threshold for gene selection δ. For this purpose, a sensitivity analysis is performed. Figure 5 exhibits the variation of the accuracy evaluation with respect to various choices of θ and δ. The reported results in this figure demonstrate that in most cases when the parameter θ is adjusted to 0.6 and δ parameter to 0.7, the developed cancer prediction model achieves the best accuracy.

Table 2. Characteristics of the used microarray datasets.

Dataset	# Genes	# Classes	# Patterns
Colon	2000	2	62
SRBCT	2328	4	83
Leukemia	7129	2	72
Prostate Tumor	10509	2	102
Lung Cancer	12600	5	203

Table 3. Average prediction accuracy of different models.

Dataset	Proposed Model	TAGA	HPSO	LL
Colon	88.99	81.54	84.81	87.24
SRBCT	83.39	80.82	78.31	77.81
Leukemia	92.16	89.63	87.23	88.13
Prostate Tumor	83.87	80.83	82.81	79.15
Lung Cancer	91.82	90.19	90.83	89.61

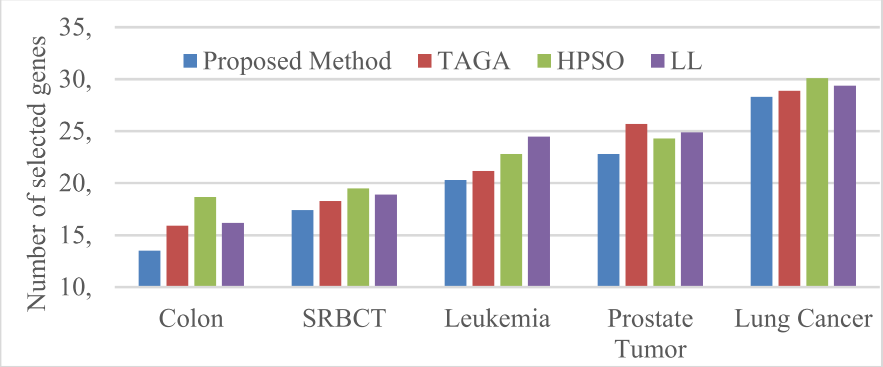


Figure 3. Average number of selected genes of the different methods in ten independent runs.

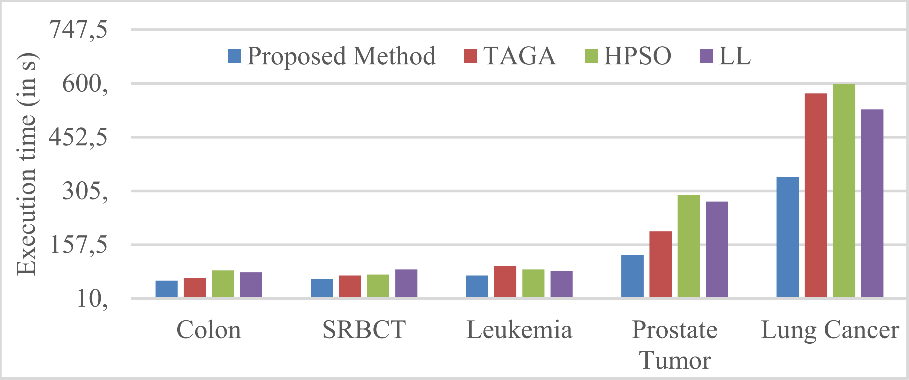
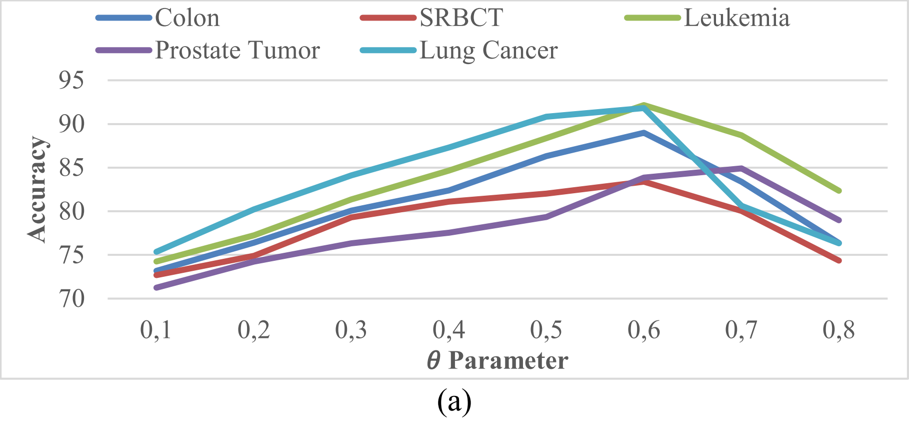


Figure 4. Average execution time (in second) of different gene selection approaches.



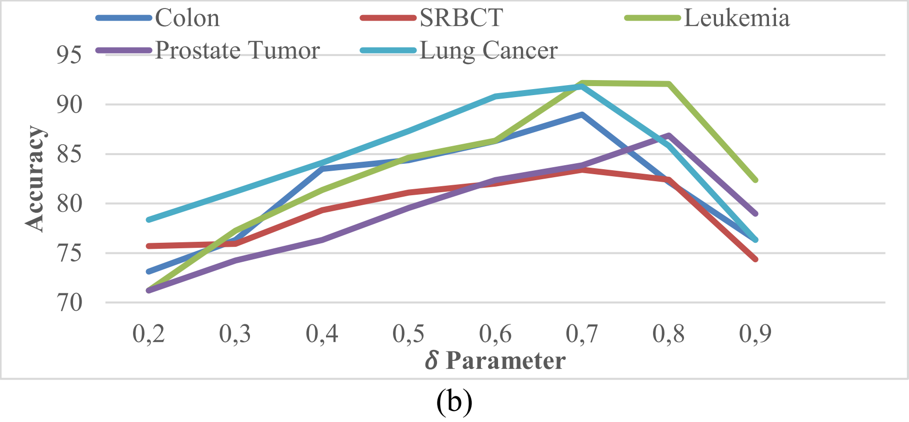


Figure 5. Sensitivity analysis of (a) θ parameter and (b) δ parameter.

Discussion and conclusion

In this section first the performance of the proposed model is analysed and then the novelties and future plan are highlighted.

In this paper an explainable cancer prediction model is developed by integrating gene selection and explainable decision forests. With the developed gene selection strategy, not only will the most redundant genes be selected, but also their relevance to DNA microarray cancer data will be maximized. Moreover, a DT-based prediction technique is developed to improve the explainability of the learning algorithm. A DT's transparency makes it a powerful tool when understanding the model structure and its predictions is necessary. The performance of the developed cancer prediction model is compared to state-of-the-art models in terms of accuracy, number of selected genes, and execution time. Compared to other prediction models, the developed method had higher accuracy, and the number of selected genes was lower than other models. Moreover, the reported results demonstrated that generally the developed model was not significantly faster than the other models. Because of the higher accuracy and explainability of the developed model, this result is particularly justifiable for medical problems, for which accuracy is more important than execution time. Furthermore, our DT-based model suffers from instability and high standard deviation in different executions. This model is inherently unstable, in that a small change in train data can result in a large change in the structure of the final DT. As a result, our model has a high probability of being overfit. In order to reduce the probability of overfitting and instability of prediction models, it would be better to increase the sample size since. However, the amount of train data available in the healthcare and DNA microarray data might be limited, so we can only use machine learning methods for train data augmentation.

Error analysis

Scrutinizing the results of the developed approach indicates that the best accuracy that can be achieved by our model is around 92-93%. Although, this is already outperforming several state-of-the-art methods as detailed in the result section of this paper, it is still legitimate to question whether we can improve further this performance. An examination of this process reveals several factors that halt further enhancement. First, the quality and size of training data are subject to inherent limitations due to complex annotation and clinical protocols employed in generating the various dataset. Second, from methodological perspective, the abrupt thresholding employed in graph construction and gene selection, although it yields significant reduction of algorithmic complexity and improves efficiency in overall, it also bears information reduction that can discard useful patterns that would have positively impacted the overall accuracy. Third, the use of decision forest model, although efficient and with good transparency/explainability capabilities, bears the inherent limitation of not predicting beyond the range of the training data and it can overfit in case of noisy data.

Novelties and future works

The developed model has four major novelties that have made it perform better than other cancer prediction models:

1. Many of state-of-the art machine learning-based models in cancer prediction [24-28] used black-box like deep learning techniques with no explanation for their prediction. In this paper, a simple and easily interpreted decision system for

cancer prediction is developed using decision forest model. Moreover, a graph representation model is utilized to ensure transparency of the gene selection model in cancer prediction that clinicians can use to explain the process of cancer prediction.

2. The majority of existing cancer prediction methods use a single classifier. Due to the specific characteristics of each prediction model, single classifier-based models are less generalizable than ensembles classifier-based models. In contrast to previous single classier-based cancer prediction models [33,57,58], this study developed a model based on ensemble DF. This yields an increase in prediction accuracy and reduced the likelihood of overfitting.

3. The efficiency of a prediction model and the result of prediction are strongly decreased by irrelevant and redundant genes. Many of previously univariate gene selection algorithms [61-66] do not consider possible gene-to-gene dependencies, therefore, they fail to remove redundant genes accurately. Moreover, several gene selection techniques, including UFSACO [67], select only a subset of dissimilar genes and do not distinguish irrelevant genes. To consider these two objectives simultaneously, an efficient graph-based gene selection search strategy is developed that can efficiently and effectively discards irrelevant and redundant genes.

4. The developed gene selection model utilizes social network analysis-based approach to develop an accurate mechanism to select the final gene set, without any classification method, which yields a low computational complexity, and its complexity is efficient for high-dimensional cancer datasets.

In high-dimensional datasets, the proposed method has the disadvantage of being a multi-phase method, which may slightly increase computational complexity. Therefore, future research can focus on developing a model that combines the different phases into one overall phase. Furthermore, we aim to integrate our proposed gene selection approach with other alternative powerful explainable deep learning techniques, e.g., LIME, in our future work for accurate and transparent cancer prediction and comparison purpose.

Acknowledgements

This project is supported by the Academy of Finland Profi5 DigiHealth project, which is gratefully acknowledged.

Conflicting interests

The authors declare no conflicts of interest.

References

[1] Mazlan AU, Sahabudin NA, Remli MA, Ismail NSN, Adenuga KI. Chapter 16 - An enhanced feature selection and cancer classification for microarray data using relaxed Lasso and support vector machine. In: Raza K, Dey N (Editors). Translational Bioinformatics in Healthcare and Medicine. Volume 13 in Advances in ubiquitous sensing applications for healthcare. Academic Press; 2021. p. 193-200. https://doi.org/10.1016/B978-0-323-89824-9.00016-1

[2] Maleki N, Zeinali Y, Niaki STA. A k-NN method for lung cancer prognosis with the use of a genetic

algorithm for feature selection. Expert Systems with Applications. 2021;164:113981. https://doi.org/10.1016/j.eswa.2020.113981

[3] Lai CM, Huang HP. A gene selection algorithm using simplified swarm optimization with multi-filter ensemble technique. Applied Soft Computing. 2021;100:106994. https://doi.org/10.1016/j.asoc.2020.106994

[4] Babu P SA, Annavarapu CSR, Dara S. Clustering-based hybrid feature selection approach for high dimensional microarray data. Chemometrics and Intelligent Laboratory Systems. 2021;213:104305. https://doi.org/10.1016/j.chemolab.2021.104305

[5] Rostami M, Forouzandeh S, Berahmand K, Soltani M, Shahsavari M, Oussalah M. Gene selection for microarray data classification via multi-objective graph theoretic-based method. Artif Intell Med. 2022 Jan;123:102228. https://doi.org/10.1016/j.artmed.2021.102228

[6] Böttger F, Schaaij-Visser TB, de Reus I, Piersma SR, Pham TV, Nagel R, Brakenhoff RH, Thunnissen E, Smit EF, Jimenez CR. Proteome analysis of non-small cell lung cancer cell line secretomes and patient sputum reveals biofluid biomarker candidates for cisplatin response prediction. J Proteomics. 2019 Mar 30;196:106-119. https://doi.org/10.1016/j.jprot.2019.01.018

[7] Rosati D, Giordano A. Single-cell RNA sequencing and bioinformatics as tools to decipher cancer heterogenicity and mechanisms of drug resistance. Biochem Pharmacol. 2022 Jan;195:114811. https://doi.org/10.1016/j.bcp.2021.114811

[8] Nikolaeva A, Sigin V, Kalinkin AI, Litviakov NV, Slonimskaya E, Tsyganov M, Kharitonova A, Vinogradov I, Vinogradov M, Strelnikov VV, Tanas AS, Vinogradov I. 70P A DNA methylation markers panel for prediction of luminal B breast cancer neoadjuvant chemotherapy response by quantitative PCR. Annals of Oncology. 2021;32(Suppl 5):S386. https://doi.org/10.1016/j.annonc.2021.08.350

[9] Kalinkin AI, Sigin V, Ignatova E, Tanas AS, Strelnikov VV. 1143P Virtual amplicons for methylation-sensitive restriction enzyme quantitative PCR (MSRE-qPCR) derived from genome-wide DNA methylation sequencing: Application to prediction of breast cancer neoadjuvant chemotherapy response. Annals of Oncology. 2021;32(Suppl 5):S927-S928. https://doi.org/10.1016/j.annonc.2021.08.784

[10] Daneshjou R, He B, Ouyang D, Zou JY. How to evaluate deep learning for cancer diagnostics – factors and recommendations. Biochim Biophys Acta Rev Cancer. 2021 Apr;1875(2):188515. https://doi.org/10.1016/j.bbcan.2021.188515

[11] Wahid A, Khan DM, Iqbal N, Khan SA, Ali A, Khan M, Khan Z. Feature selection and classification for gene expression data using novel correlation based overlapping score method via Chou's 5-steps rule. Chemometrics and Intelligent Laboratory Systems. 2020;199:103958. https://doi.org/10.1016/j.chemolab.2020.103958

[12] Shukla AK, Muhuri PK. Big-data clustering with interval type-2 fuzzy uncertainty modeling in gene expression datasets. Engineering Applications of Artificial Intelligence. 2019;77:268-282. https://doi.org/10.1016/j.engappai.2018.09.002

[13] Chen H, Li T, Fan X, Luo C. Feature selection for imbalanced data based on neighborhood rough sets. Information Sciences. 2019;483:1-20. https://doi.org/10.1016/j.ins.2019.01.041

[14] Maniruzzaman M, Jahanur Rahman M, Ahammed B, Abedin MM, Suri HS, Biswas M, El-Baz A, Bangeas P, Tsoulfas G, Suri JS. Statistical characterization and classification of colon microarray gene expression data using multiple machine

learning paradigms. Comput Methods Programs Biomed. 2019 Jul;176:173-193. https://doi.org/10.1016/j.cmpb.2019.04.008

[15] Petch J, Di S, Nelson W. Opening the black box: the promise and limitations of explainable machine learning in cardiology. Can J Cardiol. 2022 Feb;38(2):204-213. https://doi.org/10.1016/j.cjca.2021.09.004

[16] Magesh PR, Myloth RD, Tom RJ. An Explainable Machine Learning Model for Early Detection of Parkinson's Disease using LIME on DaTSCAN Imagery. Comput Biol Med. 2020 Nov;126:104041. https://doi.org/10.1016/j.compbiomed.2020.104041

[17] Gu D, Su K, Zhao H, A case-based ensemble learning system for explainable breast cancer recurrence prediction. Artif Intell Med. 2020 Jul;107:101858. https://doi.org/10.1016/j.artmed.2020.101858

[18] Wang H, Zhang Y, Zhang J, Li T, Peng L. A factor graph model for unsupervised feature selection. Information Sciences. 2019;480:144-159. https://doi.org/10.1016/j.ins.2018.12.034

[19] Tang X, Dai Y, Xiang Y. Feature selection based on feature interactions with application to text categorization. Expert Systems with Applications. 2019;120:207-216. https://doi.org/10.1016/j.eswa.2018.11.018

[20] Forouzandeh S, Berahmand K, Rostami M. Presentation of a recommender system with ensemble learning and graph embedding: a case on MovieLens. Multimedia Tools and Applications. 2020;80: 7805-7832. https://doi.org/10.1007/s11042-020-09949-5

[21] Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM computing surveys. 2018;51(5):1-42. https://doi.org/10.1145/3236009

[22] Oussalah M. AI Explainability. A Bridge Between Machine Vision and Natural Language Processing. In: ICPR 2021: Pattern Recognition. ICPR International Workshops and Challenges. Springer; 2021. p. 257-273. https://doi.org/10.1007/978-3-030-68796-0_19

[23] Kim B, Khanna R, Koyejo OO. Examples are not enough, learn to criticize! criticism for interpretability. In: NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems. December 2016. p. 2288-2296.

[24] Liu Z, Ni S, Yang C, Sun W, Huang D, Su H, Shu J, Qin N. Axillary lymph node metastasis prediction by contrast-enhanced computed tomography images for breast cancer patients based on deep learning. Comput Biol Med. 2021 Sep;136:104715. https://doi.org/10.1016/j.compbiomed.2021.104715

[25] Chai H, Zhou X, Zhang Z, Rao J, Zhao H, Yang Y. Integrating multi-omics data through deep learning for accurate cancer prognosis prediction. Comput Biol Med. 2021 Jul;134:104481. https://doi.org/10.1016/j.compbiomed.2021.104481

[26] Doppalapudi S, Qiu RG, Badr Y. Lung cancer survival period prediction and understanding: Deep learning approaches. Int J Med Inform. 2021 Apr;148:104371. https://doi.org/10.1016/j.ijmedinf.2020.104371

[27] Koh J, Yoon Y, Kim S, Han K, Kim EK. Deep Learning for the Detection of Breast Cancers on Chest Computed Tomography. Clin Breast Cancer. 2022 Jan;22(1):26-31. https://doi.org/10.1016/j.clbc.2021.04.015

[28] Xiao Y, Wu J, Lin Z. Cancer diagnosis using generative adversarial networks based on deep

learning from imbalanced data. Comput Biol Med. 2021 Aug;135:104540. https://doi.org/10.1016/j.compbiomed.2021.104540

[29] Al-Betar MA, Alomari OA, Abu-Romman SM. A TRIZ-inspired bat algorithm for gene selection in cancer classification. Genomics. 2020 Jan;112(1):114-126. https://doi.org/10.1016/j.ygeno.2019.09.015

[30] Rostami M, Forouzandeh S, Berahmand K, Soltani M. Integration of multi-objective PSO based feature selection and node centrality for medical datasets. Genomics. 2020 Nov;112(6):4370-4384. https://doi.org/10.1016/j.ygeno.2020.07.027

[31] Nayak M, Das S, Bhanja U, Senapati MR. Elephant herding optimization technique based neural network for cancer prediction. Informatics in Medicine Unlocked. 2020;21:100445. https://doi.org/10.1016/j.imu.2020.100445

[32] Ghiasi MM, Zendehboudi S. Application of decision tree-based ensemble learning in the classification of breast cancer. Comput Biol Med. 2021 Jan;128:104089. https://doi.org/10.1016/j.compbiomed.2020.104089

[33] Hamid TMTA, Sallehuddin R, Yunos ZM, Ali A. Ensemble Based Filter Feature Selection with Harmonize Particle Swarm Optimization and Support Vector Machine for Optimal Cancer Classification. Machine Learning with Applications. 2021;5:100054. https://doi.org/10.1016/j.mlwa.2021.100054

[34] Zheng X, Zhang C. Gene selection for microarray data classification via dual latent representation learning. Neurocomputing. 2021;461:266-280. https://doi.org/10.1016/j.neucom.2021.07.047

[35] Alomari OA, Makhadmeh SN, Al-Betar MA, Alyasseri ZAA, Doush IA, Ammar KA, Awadallah MA, Zitar RA. Gene selection for microarray data classification based on Gray Wolf Optimizer enhanced with TRIZ-inspired operators. Knowledge-Based Systems. 2021;223:107034. https://doi.org/10.1016/j.knosys.2021.107034

[36] Anusha M, Sathiaseelan JGR. Feature Selection Using K-Means Genetic Algorithm for Multi-objective Optimization. Procedia Computer Science. 2015;57:1074-1080. https://doi.org/10.1016/j.procs.2015.07.387

[37] Marcelloni F. Feature selection based on a modified fuzzy C-means algorithm with supervision. Information Sciences. 2003;151:201-226. https://doi.org/10.1016/S0020-0255(02)00402-4

[38] Lai CM. Multi-objective simplified swarm optimization with weighting scheme for gene selection. Applied Soft Computing. 2018;65:58-68. https://doi.org/10.1016/j.asoc.2017.12.049

[39] Aliee H, Theis FJ. AutoGeneS: Automatic gene selection using multi-objective optimization for RNA-seq deconvolution. Cell Syst. 2021 Jul 21;12(7):706-715.e4. https://doi.org/10.1016/j.cels.2021.05.006

[40] Sharma A, Rani R. C-HMOSHSSA: Gene selection for cancer classification using multi-objective meta-heuristic and machine learning methods. Comput Methods Programs Biomed. 2019 Sep;178:219-235. https://doi.org/10.1016/j.cmpb.2019.06.029

[41] González J, Ortega J, Damas M, Martin-Smith P, Gan JQ. A new multi-objective wrapper method for feature selection – Accuracy and stability analysis for BCI. Neurocomputing. 2019;333:407-418. https://doi.org/10.1016/j.neucom.2019.01.017

[42] Xue B, Zhang M, Browne WN. Particle Swarm Optimization for Feature Selection in Classifica-

tion: A Multi-Objective Approach. IEEE Transactions on Cybernetics. 2013;43(6):1656-1671. https://doi.org/10.1109/TSMCB.2012.2227469

[43] Moradi P, Rostami M. Integration of graph clustering with ant colony optimization for feature selection. Knowledge-Based Systems. 2015;84:144-161. https://doi.org/10.1016/j.knosys.2015.04.007

[44] Liu W, Wang J. Recursive elimination - election algorithms for wrapper feature selection. Applied Soft Computing. 2021;113(Part B):107956. https://doi.org/10.1016/j.asoc.2021.107956

[45] Singh N, Singh P. A hybrid ensemble-filter wrapper feature selection approach for medical data classification. Chemometrics and Intelligent Laboratory Systems. 2021;217:104396. https://doi.org/10.1016/j.chemolab.2021.104396

[46] Nouri-Moghaddam B, Ghazanfari M, Fathian M. A novel multi-objective forest optimization algorithm for wrapper feature selection. Expert Systems with Applications. 2021;175:114737. https://doi.org/10.1016/j.eswa.2021.114737

[47] Moradi P, Rostami M. A graph theoretic approach for unsupervised feature selection. Engineering Applications of Artificial Intelligence. 2015; 44:33-45. https://doi.org/10.1016/j.engappai.2015.05.005

[48] Rostami M, Berahmand K, Forouzandeh S. A novel method of constrained feature selection by the measurement of pairwise constraints uncertainty. Journal of Big Data. 2020;7(1):83. https://doi.org/10.1186/s40537-020-00352-3

[49] Rostami M, Berahmand K, Forouzandeh S. A novel community detection based genetic algorithm for feature selection. Journal of Big Data. 2021;8(1):2. https://doi.org/10.1186/s40537-020-00398-3

[50] Wu Y, Hu Q, Wang S, Liu C, Shan Y, Guo W, Jiang R, Wang X, Gu J. Highly Regional Genes: graph-based gene selection for single cell RNA-seq data. J Genet Genomics. 2022 Feb 7;S1673-8527(22)00033-9. https://doi.org/10.1016/j.jgg.2022.01.004

[51] Jia X, Wen T, Ding W, Li H, Li W. Semi-supervised label distribution learning via projection graph embedding. Information Sciences. 2021;581:840-855. https://doi.org/10.1016/j.ins.2021.10.009

[52] Kabir MM, Shahjahan M, Murase K. A new local search based hybrid genetic algorithm for feature selection. Neurocomputing. 2011;74(17):2914–2928. https://doi.org/10.1016/j.neucom.2011.03.034

[53] Bai L, Cheng X, Liang J, Guo Y. Fast graph clustering with a new description model for community detection. Information Sciences. 2017;388-389:37-47. https://doi.org/10.1016/j.ins.2017.01.026

[54] Chen Z, Chen Q, Zhang Y, Zhou L. Clustering-based feature subset selection with analysis on the redundancy–complementarity dimension. Computer Communications. 2021;168:65-74. https://doi.org/10.1016/j.comcom.2021.01.005

[55] Yuan T, Deng W, Hu J, An Z, Tang Y. Unsupervised adaptive hashing based on feature clustering. Neurocomputing. 2019;323:373-382. https://doi.org/10.1016/j.neucom.2018.10.015

[56] Wang L, Meng J, Huang R, Zhu H, Peng K. Incremental feature weighting for fuzzy feature selection. Fuzzy Sets and Systems. 2019;368:1-19. https://doi.org/10.1016/j.fss.2018.10.021

[57] Salesi S, Cosma G, Mavrovouniotis M. TAGA: Tabu Asexual Genetic Algorithm embedded in a filter/filter feature selection approach for high-dimensional data. Information Sciences.

2021;565:105-127. https://doi.org/10.1016/j.ins.2021.01.020

[58] Thabtah F, Kamalov F, Hammoud S, Shahamiri SR. Least Loss: A simplified filter method for feature selection. Information Sciences. 2020;534:1-15. https://doi.org/10.1016/j.ins.2020.05.017

[59] Dataset Repository. UPOBioinfo Group, 2020-2021. Centro Andaluz de Biología del Desarrollo (CABD), CSIC-UPO; 2022 [cited 1 September 2021]. Available from: http://www.bioinfocabd.upo.es/

[60] Statnikov A, Tsamardinos I, Dosbayev Y, Aliferis CF. GEMS: a system for automated cancer diagnosis and biomarker discovery from microarray gene expression data. Int J Med Inform. 2005 Aug;74(7-8):491-503. https://doi.org/10.1016/j.ijmedinf.2005.05.002

[61] Raileanu LE, Stoffel K. Theoretical comparison between the Gini index and information gain criteria. Ann Math Artif Intell. 2004;41:77-93. https://doi.org/10.1023/B:AMAI.0000018580.96245.c6

[62] Mitchell TM. Machine Learning. New York: McGraw-Hill; 1997.

[63] Theodoridis S, Koutroumbas C. Pattern Recognition. 4th Edition. Elsevier Inc; 2009.

[64] Xu Y, Jones G, Li JT, Wang B, Sun CM. A study on mutual information-based feature selection for text categorization. Journal of Computational Information Systems. 2007;3(3):1007-1012.

[65] He X, Cai D, Niyogi P. Laplacian Score for Feature Selection. Adv Neural Inf Process Syst. 2005;18:507-514.

[66] Gu Q, Li Z, Han J. Generalized Fisher Score for Feature Selection. In: Proceedings of the International Conference on Uncertainty in Artificial Intelligence. July 2011. p. 266-273.

[67] Tabakhi S, Moradi P, Akhlaghian F. An unsupervised feature selection algorithm based on ant colony optimization. Engineering Applications of Artificial Intelligence. 2014;32:112-123. https://doi.org/10.1016/j.engappai.2014.03.007