PLOS ONE

RESEARCH ARTICLE

HDG-select: A novel GUI based application for gene selection and classification in high dimensional datasets

Shilan S. Hameed[1,2]*, **Rohayanti Hassan**[3], **Wan Haslina Hassan**[1], **Fahmi F. Muhammadsharif**[4], **Liza Abdul Latiff**[5]

1 Computer Systems and Networks (CSN), Malaysia-Japan International Institute of Technology (MJIIT), Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia, **2** Directorate of Information Technology, Koya University, Koya, Kurdistan Region-F.R., Iraq, **3** School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia, Johor Bahru, Johor, Malaysia, **4** Department of Physics, Faculty of Science and Health, Koya University, Koya, Kurdistan Region-F.R., Iraq, **5** U-BAN Research Group, Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia

* shilan.sameen@koyauniversity.org

Citation: Hameed SS, Hassan R, Hassan WH, Muhammadsharif FF, Latiff LA (2021) HDG-select: A novel GUI based application for gene selection and classification in high dimensional datasets. PLoS ONE 16(1): e0246039. https://doi.org/10.1371/journal.pone.0246039

Editor: Bryan C. Daniels, Arizona State University & Santa Fe Institute, UNITED STATES

Received: April 18, 2020

Accepted: January 12, 2021

Published: January 28, 2021

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0246039

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Abstract

The selection and classification of genes is essential for the identification of related genes to a specific disease. Developing a user-friendly application with combined statistical rigor and machine learning functionality to help the biomedical researchers and end users is of great importance. In this work, a novel stand-alone application, which is based on graphical user interface (GUI), is developed to perform the full functionality of gene selection and classification in high dimensional datasets. The so-called HDG-select application is validated on eleven high dimensional datasets of the format CSV and GEO soft. The proposed tool uses the efficient algorithm of combined filter-GBPSO-SVM and it was made freely available to users. It was found that the proposed HDG-select outperformed other tools reported in literature and presented a competitive performance, accessibility, and functionality.

Introduction

The microarray is a tool used to estimate whether mutations in specific genes are present in a particular individual. The most common type of microarray is utilized to measure gene expression, where the expression values of thousands of genes are calculated from the microarray sample [1]. The identification of the most attributed genes to a specific disease can be carried out by means of gene selection and classification of the microarray datasets, wherein various statistical and optimization algorithms are involved. The outcome of accurate selection of attributed genes would ultimately lead to establishing a cost-effective and useful studies on the altered genes [2]. Furthermore, the identified genes help in classifying the clinical samples to normal and disease samples. Gene selection methods are classified into two main types: filter-based methods and wrapper based ones [3, 4]. Filter based methods work separately without using any connected classifier, so they provide the results faster. They are better applied in analyzing high dimensional data of microarray datasets with thousands of genes and hundreds of

Funding: RH received a financial support from the Fundamental Research Grant Scheme (FRGS), Ministry of Education and Universiti Teknologi Malaysia under Vote No: RJ130000.7851.5F037.

Competing interests: NO authors have competing interests.

samples [5, 6]. The weakness of filter methods is that most of them are unable to establish a useful correlation among the genes and hence there would be the possibility of selecting redundant genes. This drawback acts to reduce the final classifier accuracy if only filter is applied to select the discriminative genes [4]. Hence, the best approach is to use filters in the preliminary selection steps [6]. Wrappers perform better in selecting discriminative genes since they depend on the model hypothesis to train and test in the gene space [4]. However, wrapper-based techniques are heavy and could be a worst choice if they are directly applied on high dimensional datasets without any preprocessing [7].

Many computational methods are failed to extract a small subset of attributed genes in high dimensional datasets because of the presence of various correlations and redundancy among the genes. Interestingly, studies in the field of cancer informatics have shown a splendid contribution of data mining and machine learning to find the attributed genes [8–11]. It is however proved that machine learning can perform well in cancer classification, it is yet required further improvement and robustness in terms of efficiency and computational cost, especially when high dimensional datasets are investigated. This is because high dimensional datasets contain several redundant and variant genes expression, which in turn acts upon reducing the accuracy and efficiency of the computational techniques used to mine the most attributed genes [12]. In general, the noise in gene expression level is occurred due to biological variations associated with the experiments or the existence of alterations in the genes [4, 13]. Therefore, it is not an easy or a straightforward task to find the attributed genes in high dimensional datasets unless a careful analysis and selection rule is carried out.

Along this line, a binary variant of Harris hawk's optimizer (HHO) was proposed to boost the efficacy of wrapper-based gene selection in high dimensional dataset [14]. Besides, a two-stage sparse logistic regression was reported aiming at obtaining efficient subset of genes with high classification capabilities [15]. That is by combining the screening approach as filter method and adaptive lasso with a new weight as wrapper method. Gene selection in high-dimensional colon cancer microarray dataset was seen to be enhanced by using an ensemble of gene selection technique based on *t*-test and GA [16]. After preprocessing the data using *t*-test, a Nested-GA was employed to get the optimal subset of genes. As such, various approaches were reported in literature in order to increase the gene selection efficacy in high dimensional datasets such as hybrid binary coral reefs optimization algorithm with simulated annealing [17], ensembles of regularized regression models with resampling-based lasso [18], variable-size cooperative coevolutionary particle swarm optimization [19], hybrid dimensionality reduction forest with pruning [20], hybrid feature selection based on reliefF and binary dragonfly [21] as well as hybrid rough set theory and hypergraph [22]. It is observed that the effective approach for gene selection in microarray dataset can be a combination of filter and wrapper algorithms. Obviously, there exist numerous techniques used to select attributed genes in high dimensional datasets, however the complexity of the algorithms and computational cost are limiting their reproducibility with rapid selection of discriminated genes in massive datasets. Nevertheless, particle swarm optimization (PSO), as a searching strategy for genes selection, is proved to be more efficient and easy to implement compared to other methods [23, 24]. This is because few parameters are needed to perform its adjustment and therefore it saves memory. The modified geometric binary of PSO (GBPSO) was effectively utilized for gene selection in autism dataset [12]. Details on PSO and its GBPSO variant can be found in literature [12, 24–27]. GBPSO can be used as a wrapper feature selection method with a support vector machine (SVM). SVMs represent a group of supervised machine-learning methods which were developed by Vapnik [28]. The various forms of this algorithm are widely used [9, 24, 29], particularly for medical related data classification [30–34]. Moreover, SVM can perform both linear and nonlinear separable data classification. When using SVM, it is essential

that the number of coefficients to be determined are primarily based on the number of samples not on the number of genes. In the case of gene classification, SVM utilizes kernel functions to get an orthogonal hyperplane to separate the genes in a specific dimension. Different types of kernels can be applied [24, 35, 36], whereas each kernel type is appropriate for different data. In the current work, a polynomial kernel was utilized for the SVM due to its highest classification accuracy when it is applied for high dimensional datasets.

It is well-known that the process of gene selection and classification is becoming tedious and time consuming when the datasets are not curated such as soft GEO datasets. A review of literature showed that there are various tools created for sequence and genomic data analysis [37–40], while there has been few applications established for gene selection and classification [41–43]. For instance, a java GUI application was developed for microarray data classification using SVM classifier [43]. The researchers concluded that the application performs well when a radial basis SVM kernel is used. However, their tool is not accessible now and it is created only for classification. The varSelRF package and GeneSrF tool were developed for gene selection given the associated error of classification using R language and python [41]. The package of varSelRF can be only used on Linux and Unix OS, while GeneSrF is a web-based tool and is not currently accessible. In another study [42], ArrayMining.net, which is a web-based tool, was constructed for gene selection and class identification using supervised and un-supervised techniques. In the current work, a novel user-friendly and stand-alone (non-web based) application is proposed for a simple and efficient gene selection and classification in the high dimensional datasets. The software program was developed with the help of interfacing MATLAB with Weka tool, combining their benefits in one package. The proposed application is named as HDG-select, referring to its capability of high dimensional gene selection. It can be used by researchers and students to reduce the burden of hard-working steps of dataset curation, gene selection and classification on a one-platform scheme. The main advantages of the proposed application include dataset curation, user-defined gene filtration, handling both numerical and categorical samples and combining the functionality of MATLAB and Weka in a single tool. If someone wants to perform a complete gene selection including dataset curation and filtration, a comprehensive coding is first required in MATLAB and then the results need to be transferred to Weka tool in order to run the GBPSO-SVM algorithm. Interestingly, the proposed HDG-select is the collection of all necessary operations within a single user-friendly graphical interface, which helps the users to practice simplicity, accuracy and reduced computational cost. Noticeably, the tool uses a combination of filters and GBPSO wrapper for gene selection, while SVM is used for classification. Furthermore, the reported tools in literature accept CSV files as input datasets. However, our developed tool can handle both CSV and.Soft file formats, which is specifically useful for analysing the non-curated genomic data available in the GEO database.

Materials and methods

Implementation procedure

The process of selection and classification of genes in high dimensional microarrays using the developed HDG-select tool and the built-in structure of the application are shown in Figs 1 and 2, respectively. The first step was to reduce the dimensionality of the datasets by removing the redundant/irrelevant genes whose expression values are close amid the control and non-control classes. For this purpose, the values of mean and median ratio were calculated based on the variance of the genes expression, which is discussed later in detail. This process is performed so that the next steps become more efficient and easier. Later on, two different filters, namely *t*-test (TT) and Wilcoxon rank sum (WRS), are used to filter the desired number of top

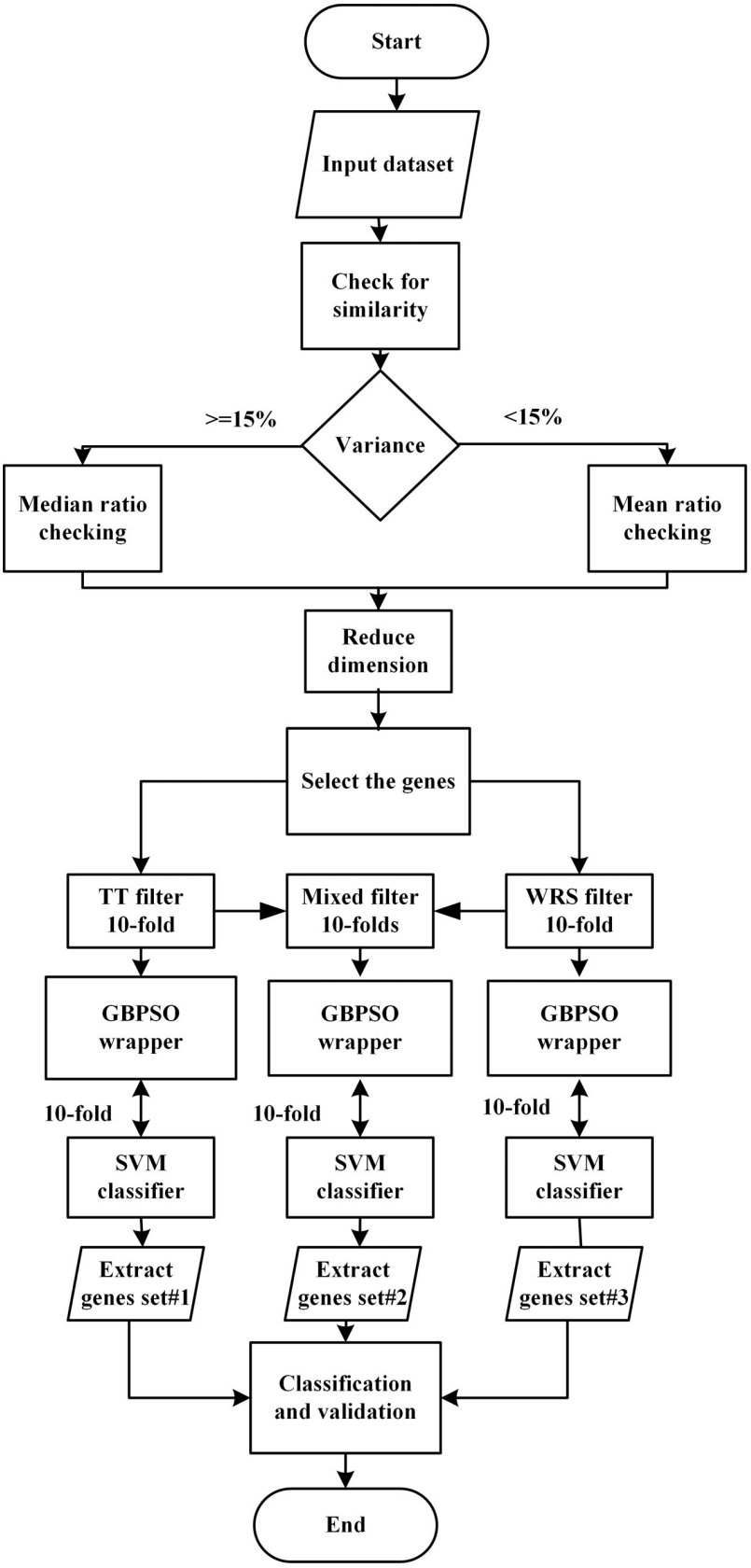

Start
Input dataset
Check for similarity
Variance
>=15%
<15%
Median ratio checking
Mean ratio checking
Reduce dimension
Select the genes
TT filter 10-fold
Mixed filter 10-folds
WRS filter 10-fold
GBPSO wrapper
GBPSO wrapper
GBPSO wrapper
10-fold
10-fold
10-fold
SVM classifier
SVM classifier
SVM classifier
Extract genes set#1
Extract genes set#2
Extract genes set#3
Classification and validation
End

Fig 1. The flowchart of selecting genes and classification in high dimensional datasets using the HDG-select application. The highly irrelevant genes are fist removed (dataset curation) by considering the values of mean and median ratio, followed by the use of different filters in combination with the GBPSO algorithm.

https://doi.org/10.1371/journal.pone.0246039.g001

relevant genes. These filters and their combination process are elaborated in the next sections. As shown in Fig 2, the curation and filtration steps are implemented by MATLAB coding, while the use of wrapper based GBPSO-SVM algorithm is realized by Java programming, that is by interfacing the GBPSO-SVM algorithm from Weka with MATLAB.

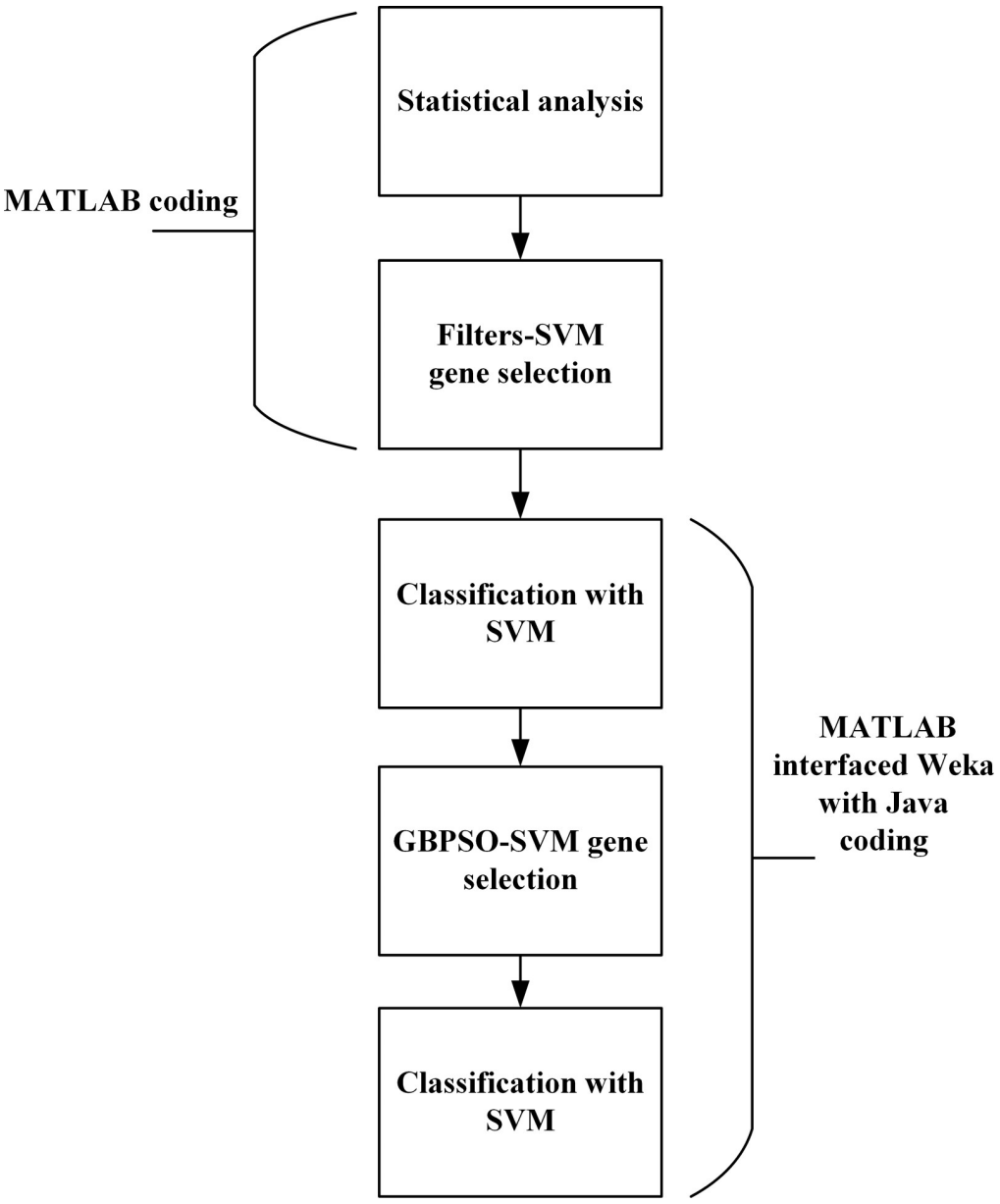


Fig 2. The built-in structure of the developed HDG-select application. The curation and filtration steps are implemented by MATLAB coding, while the use of wrapper based GBPSO-SVM algorithm is realized by MATLAB interfaced Weka through Java coding.

https://doi.org/10.1371/journal.pone.0246039.g002

Microarray datasets

Validation and assessment of the developed HDG-select application was carried out by testing 11 high dimensional datasets of different types of diseases. The characteristics of the datasets are given in Tables 1 and 2. The first six datasets in Table 1 are in csv format, which were pre-processed and previously used for gene expression analysis [11]. The Leukemia cancer dataset was achieved from [44]. The dataset of leukemia cancer is given as S1 Dataset. The colon cancer microarray dataset was originally analyzed by Alon *et al* [45], it is given as S2 Dataset. The prostate cancer dataset is based on oligonucleotide microarray, which was obtained from [46]. The dataset of prostate cancer is given as S3 Dataset. The rest of the datasets (Breast, CNS and Ovarian) were achieved from [47] whose datasets are given as S4–S6 Datasets, respectively.

The second batch of investigated datasets are in soft format with originally non-curated condition, which can be downloaded from the well-known public repository of GEO (NCBI) [48] under GDS file name. The main characteristics of these datasets are given in Table 2 with some descriptions as follows:

Inflammatory breast cancer (GDS3097). Tumor epithelium and underlying stromal cells were extracted using laser capture microdissection of human breast cancer to study gene expression variations based on inflammatory and non-inflammatory breast cancer tissue types.

Breast cancer (GDS3716). With Affymetrix HU133A microarrays, 42 total laser capture micro dissected histologically normal samples of breast tissue were analyzed in this dataset.

Brain metastatic breast cancer (GDS5306). Gene expression of 19 HER2+ breast cancer brain metastases were comparably examined with HER2+ nonmetastatic primary tumors.

Autism disorder (GDS4431). Total RNA was extracted for microarray experiments with Affymetrix Human U133 Plus 2.0 39 Expression Arrays. The autistic samples were diagnosed by medical professionals of developmental pediatrician and psychologist according to the DSM-IV criteria and the diagnosis was confirmed on the basis of ADOS and ADI-R criteria [49].

Influenza A (GDS6063). More than 2600 genes were expressed differently in pDCs exposed to influenza A compared to controls (no viruses) blood pDCs.

Dimensionality reduction of soft format datasets using mean and median ratio

Because of the presence of variance among the genes expression in high dimensional datasets and the non-curated nature of GEO soft datasets [48, 50], it is imperative to perform a pre-processing mechanism in order to reduce the dimensionality of the datasets, thereby removing the redundant and highly irrelevant genes. For this purpose, the overall similarity of genes expression was assessed through the estimation of their mean and median values among the two

Table 1. The main characteristics of the pre-reduced high dimensional datasets in csv format.

Dataset	#Genes	#Samples	#Class (class1:class2)
Leukemia	3051	72	2(25:47)
Colon	2000	62	2(22:40)
Prostate	6033	102	2(50:52)
Breast	24481	97	2(46:51)
CNS[a]	7129	72	2(21:39)
Ovarian	15154	253	2(162:91)

[a] Central Nervous system.

https://doi.org/10.1371/journal.pone.0246039.t001

Table 2. The main characteristics of the original/non-curated GEO datasets in soft format.

Dataset	#Genes	#Samples	#Class (class1:class2)
Inflammatory Breast Cancer (GDS3097)	22283	48	2(35 NIBC:13 IBC)
Breast Cancer (GDS3716)	22283	42	2(24 control:18 breast cancer)
Brain Metastatic Breast Cancer (GDS5306)	61359	38	2(19(BMBC) tumor:19(NBC) tumor)
Autism (GDS4431)	54675	146	2(69 control:77 autism)
Influenza A (GDS6063)	48107	10	2(5 postive:5negative)

https://doi.org/10.1371/journal.pone.0246039.t002

classes. When a mean criterion is applied to identify the redundant/irrelevant genes, the mean of genes expression is determined. Similarly, when the median criterion is considered, the median value of genes expressions is calculated.

It was seen that when there is a high variance in the genes expression (variance $\geq$ 15%), the application of median criterion to reduce the dataset dimensionality is performed better in comparison to the application of mean criterion [12]. This is because the mean value of genes expression is affected by the high variance. Therefore, in this work, median criterion is applied on the genes of variance $\geq$ 15%, while mean criterion is used for those with variance $<$ 15%. Consequently, genes whose median and mean ratio of their expression are between 0.95 and 1/0.95 are removed from the dataset. This threshold range is chosen intentionally in order to remove the redundant and less significant genes from the whole dataset, and hence making the next steps of the gene selection simple and cost-effective without compromising the selection accuracy.

Gene selection using statistical filters

The second step of gene selection in the GEO soft datasets, after the dimensionality reduction, was performed by using two different statistical filters and their combination, namely two-sample t-test (TT), Wilcoxon rank sum test (WRS) and combined TT-WRS. However, this selection by means of the filters can be the first step of gene selection for the pre-curated datasets of CSV format because the proposed HDG-select tool allows users to bypass the dimensionality reduction step for the pre-curated CSV datasets. The choice of these filters is based on the findings that the TT and WRS filters performed well when they used for gene selection in the high dimensional dataset of autism [12]. This is because each filter is based on different assumptions related to the mean, median and variance which can be found in the high dimensional datasets. Because the filtration power is different for each filter, the combination of them might yield a better selection performance [11, 51, 52]. The TT filter was applied to microarray genes [53] and it was seen that the filter shows a strong scalability when the number of genes is high [54]. Hence, some researchers have used the TT filter as the only step of gene selection [55, 56]. Also, WRS filter was effectively used for the pre-selection of genes [57, 58], especially when the data are associated with high variance [59].

In this work, the statistical filtrations are applied on the datasets in a 10-fold run in order to avoid overfitting. As such, the genes are ranked among the 10-fold from the most significant to the least significant ones. Hence, based on their ranking position (weight), the desired number of most highly ranked genes can be extracted. The equation used to weigh the genes and ranking their positions based on their significancy is a formula of global weight that is given by:

$$w(f) = \sum_{i=1}^{K} w_i(f) \quad (1)$$

where each i in K = the number of current fold iterations in the entire 10-fold run.

The *t*-test (TT) filter is a univariate filter that is commonly used for binary classes [53, 54]. The general assumption of the *t*-test is that the values are uniformly distributed with a bell-shaped distribution curve among the two classes. The *t*-test null hypothesis supposes equal means and equal variances, and this assertion is rejected by the alternative hypothesis. The *t*—test formula is [60]:

$$t = \frac{c_1 - c_2}{\sqrt{\frac{{\sigma_1}^2}{n} + \frac{{\sigma_2}^2}{m}}} \tag{2}$$

where n and m are the first- and second-class population size, respectively. The result of the evaluation calls t, its value is ranged from 0 to 1 based on the significancy level. The value of 1 refers the abandonment of the null hypothesis at the 5 percent and 0 refers to the acceptance of the null hypothesis at the same level of significance. The test also returns the probability value of t. The lower p-value implies a noticeable difference between the compared samples. The parametric form of TT filter assumes equal variance and normal distribution, while non-parametric one assumes unequal variance and random distribution. In this work, the non-parametric TT filter is used because most of the data distribution in high dimensional datasets follow unequal variance and random distribution due to the presence of high noise and various expression values.

The second filter is Wilcoxon rank sum (WRS) test, which is a non-parametric filter method [61]. Hence, it is not essential for the gene values in the classes to have a normal distribution such as seen in the high dimensional datasets. This method is also known as the Mann-Whitney test [62, 63]. It uses a median based criterion to distinguish between the two classes. The test compares the samples medians and provides results on a ranking manner rather than in numerical values [64]. The index value and rank for each element in the result can be determined by arranging them in an ascending order. The null hypothesis considered by WRS test is that all genes originate from one class. The statistical formula of the Wilcoxon rank sum is as follows [57]:

$$s(g) = \sum_{i \in N_0} \sum_{j \in N_1} I\left((\mathbf{x}_\mathbf{j}^{(\mathbf{g})} - \mathbf{x}_\mathbf{i}^{(\mathbf{g})}) \right) \leq 0 \tag{3}$$

where I is the function used to distinguish the classes. If the logical expression $(\mathbf{x}_\mathbf{j}^{(\mathbf{g})} - \mathbf{x}_\mathbf{i}^{(\mathbf{g})}) \leq 0$ is true, I is 1; otherwise, it is 0. $\mathbf{x}_\mathbf{i}^{(\mathbf{g})}$ is the expression value of gene g in sample I, N_0 and N_1 represent the number of observations in each of the two classes, respectively, and $s(g)$ denotes the difference in the expression of the gene in the two classes. Based on whether $s(g)$ becomes 0 or reaches the maximum of $N_0 \times N_1$, the considered gene is ranked in importance in the classification process. The following equation is used to calculate the gene's importance:

$$q(g) = \max\left(s(g), N_o \times N_1 - s(g)\right) \tag{4}$$

Gene selection using GBPSO-SVM algorithm

In the final step of gene selection, the wrapper based GBPSO-SVM algorithm is applied. The GBPSO uses SVM's accuracy prediction to select the best subset of genes. SVM algorithm was used with GBPSO due to its sufficient ability in giving sensible classification accuracy for microarray data regardless of the number of samples. This is a useful feature of SVM for microarray data due to the low sample-to-gene ratio in this dataset. GBPSO starts with a number of randomly selected genes, then in each iteration it searches for the optimum subset of genes. The SVM classifier assesses the performance of each candidate subset using 10-fold cross validation. Hence, every current candidate subset of genes is commonly better than the

previous subset. The GBPSO original package of the algorithm can be retrieved from [65]. In the current work, a polynomial kernel was utilized for the SVM due to its highest classification accuracy when it is applied for high dimensional datasets.

Development of HDG-select application

The so-called HDG-select application was created using graphical user interface (GUI) in MATLAB. The developed application has a user-friendly interface which is easy to understand and implement for gene selection and classification in both of high dimensional and normal datasets. The first two steps of gene selection were written in MATLAB, while the third step was written in Java, taking the advantage of Weka packages and Java interfacing [66]. Meanwhile, we used Java coding for interfacing the Weka functionality with MATLAB. It was designed in a way that it can handle errors and control the user's inputs to perform each step correctly. This was achieved by using message handlers during the application process. The HDG-select application was made freely available to users, which can be downloaded from GitHub (https://github.com/Shilan-Jaff/HDG_select). Fig 3 shows the interface of the tool, which is composed of four major sections described as below:

a. *Input (Dataset import)*: The user is able to import two formats of datasets, namely soft and csv, as shown in part (a) of Fig 7. Most of the curated datasets available online are in the form of csv format, which is the common format for machine learning applications. The other dataset type is soft file, which is the format of the gene expression profiling datasets made available to public at GEO NCBI database [48]. This type of dataset is usually in the form of a non-curated and high dimensional structure. Before importing the csv files, users have to make sure that the last column contains the class label, while for the soft file the range of the sample class should be manually given to the application. This is because information regarding the sample class is inherently not presented in the soft files, so it must be obtained from the dataset description given in the NCBI database.

b. *Preprocessing*: In this step of analysis, a reduction process can be made upon the high dimensional datasets, or upon the datasets that have not yet been reduced/curated by researchers. As such, the dataset will be easily handled for the next steps of analysis. It can be seen from part (b) of Fig 7 that this section has two options, which allows the user to choose between reducing the dataset or leaving it as it is. However, for the soft format

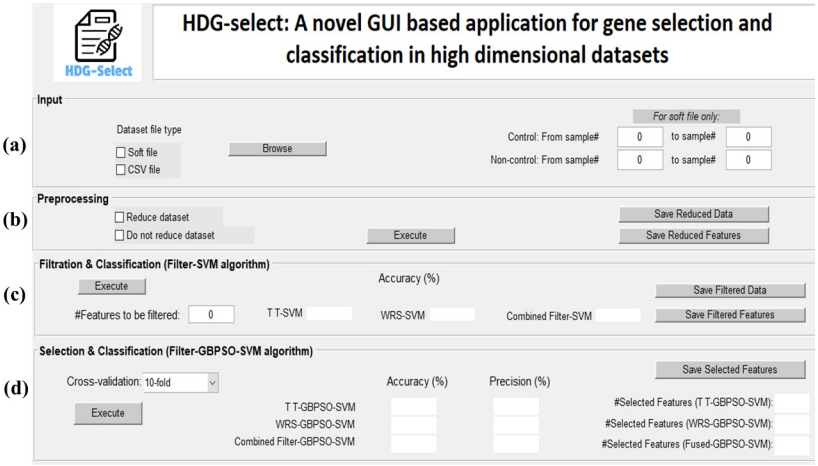


Fig 3. The main interface of the developed HDG-select application used for gene selection and classification in high dimensional datasets.

https://doi.org/10.1371/journal.pone.0246039.g003

datasets, gene reduction is obligatory, otherwise the process would be computationally costly and memory overload is resulted. At the end of this process, users can save the reduced dataset for their future use if required.

c. *Filtration and classification*: Once the user finalized the preprocessing step, the dataset is proceeded to the next stage of filtering the most significant genes, classification assessment and saving the filtered dataset, as shown in part (c) of Fig 7. It is worth to mention that with the help of this application, the user can get accessibility to decide on the number of genes to be filtered. Hence, one can choose the optimum filtered genes based on the preference and understanding of the dataset. Nevertheless, the default number of gene filtration was set to be 200.

d. *Selection and classification using Filter-GBPSO-SVM algorithm*: The last and most important action is to select the genes and to apply the SVM classifier on the selected genes, as shown in part (d) of Fig 7. This is applied on the results achieved from previous steps and is performed on each dataset generated from the filtration step. Here, the user can see how many genes are selected by each approach and has access to save them.

The HDG-select application uses the following equations to determine the accuracy and precision of the gene selection and classification, respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

Where *TP*, *TN*, *FN*, *FP* are the true positive, true negative, false negative and false positive detected samples, respectively.

Results and discussions

In order to show the functionality and robustness of the proposed HDG-select tool, a step by step analysis is presented. As we mentioned earlier, users can choose to not perform the dimensionality reduction on the pre-curated CSV datasets. However, the preprocessing step (see Fig 3B) for the soft GEO datasets is a must since these datasets are not pre-curated. It leads to a heavy computational burden and low classification accuracy if they are directly applied for gene selection. Consequently, the soft GEO files were preprocessed and the dataset dimensionality was interestingly reduced. For example, the genes in Autism and influenza A datasets were reduced from 54613 to 14530 and from 22283 to 17939, respectively upon the application of the preprocessing step. As such, with the help of HDG-select, the impact of filtered genes on the SVM classification accuracy in the soft GEO datasets was investigated, as shown in Fig 4. Results showed that limiting the filtered genes to below 150 genes has negatively affected the classification accuracy, except for the influenza A dataset which showed a stable performance regardless of the change in the genes number. Concludingly, the dataset with a steady curve indicating the presence of good correlation between the genes and hence extra filtration does not further improve the classification accuracy. Nevertheless, filtering the genes to a low possible number can save memory and speed up the execution time in the final stage of gene selection and classification.

Fig 5A and 5B show the SVM accuracy of gene selection in CSV datasets and soft GEO datasets after the application of TT filter and TT filter-GBPSO algorithm with the help of the

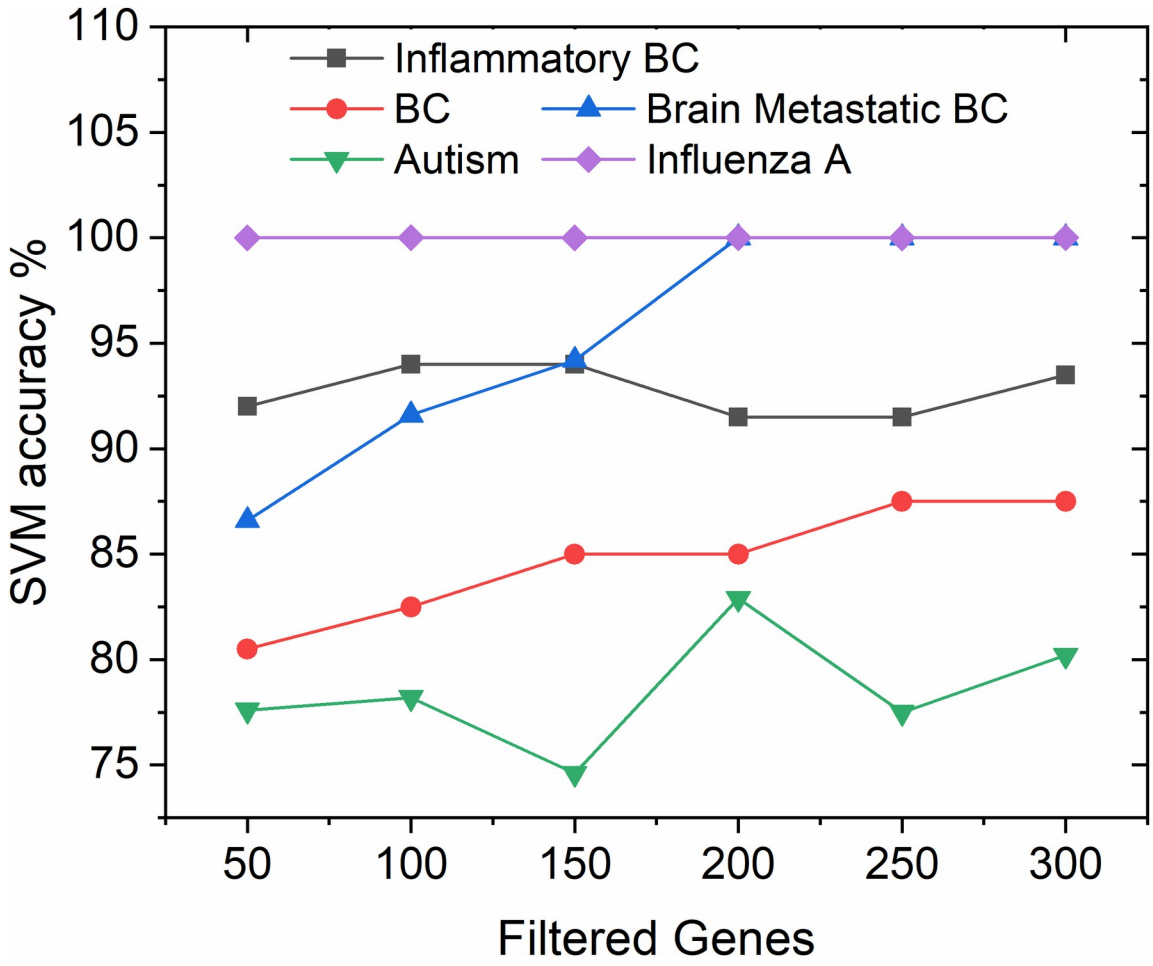
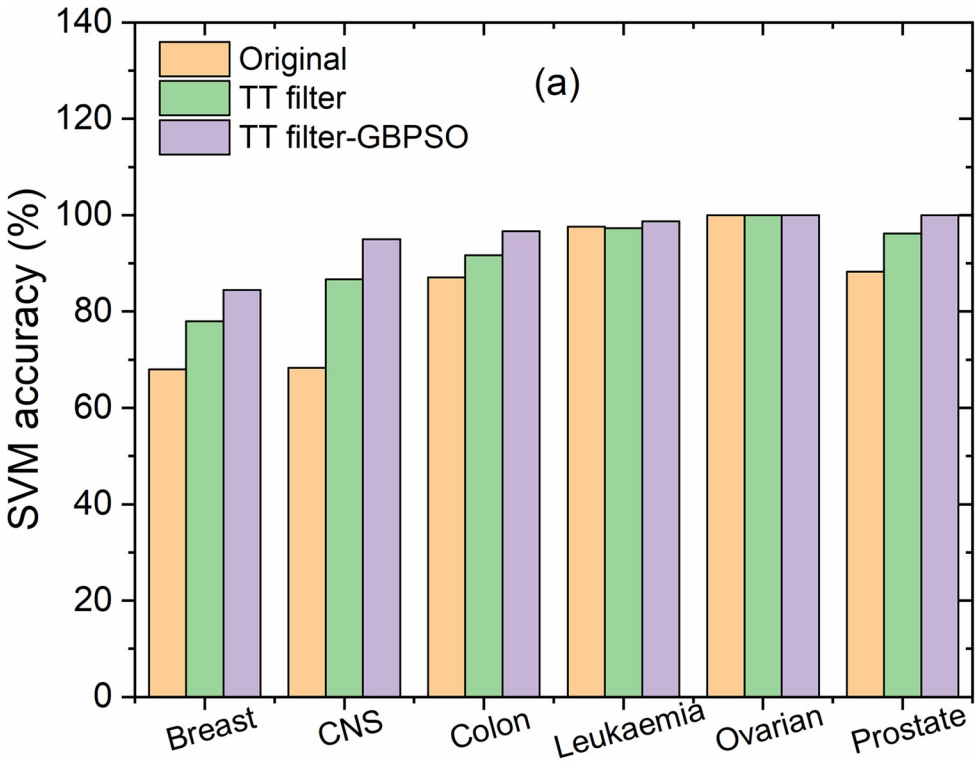


Fig 4. Effect of the number of filtered genes on the SVM classification accuracy in the first step of gene selection in high dimensional datasets.

https://doi.org/10.1371/journal.pone.0246039.g004

proposed HDG-select, respectively. More data results from the application of different filters and GBPSO are given as S1 and S2 Tables. One can notice from the results that the accuracy of SVM is largely improved when the statistical filters are used in combination with GBPSO. The application of filters has improved the classification accuracy when it is compared to the results of the original dataset. Comparably, the use of GBPSO algorithm in combination with the filters has led to improved performance. For instance, the classification accuracy in leukemia and colon cancer has reached 100% when a combined TT-WRS filter with GBPSO was utilized, surpassing the results obtained by Nested-GA algorithm [16]. Noteworthy, the SVM accuracy for brain metastatic BC and Influenza A datasets remained 100% in both steps of genes selection by filters and GBPSO-SVM, indicating that the HDG-select tool has a strong power on the dimensionality reduction of the datasets to maintain the most important genes that are quite useful for the subsequent steps of gene selection process.

Results showed that the best approach to increase the accuracy of classification of the gene selection in high dimensional datasets is to utilize different filters in combination with the GBPSO-SVM algorithm, as shown in Fig 6. It was observed that the combination of TT-WRS filter with GBPSO has led to improve the SVM accuracy in eight datasets out of eleven datasets.


Original
TT filter
TT filter-GBPSO
(a)
SVM accuracy (%)
0
20
40
60
80
100
120
140
Breast
CNS
Colon
Leukaemia
Ovarian
Prostate

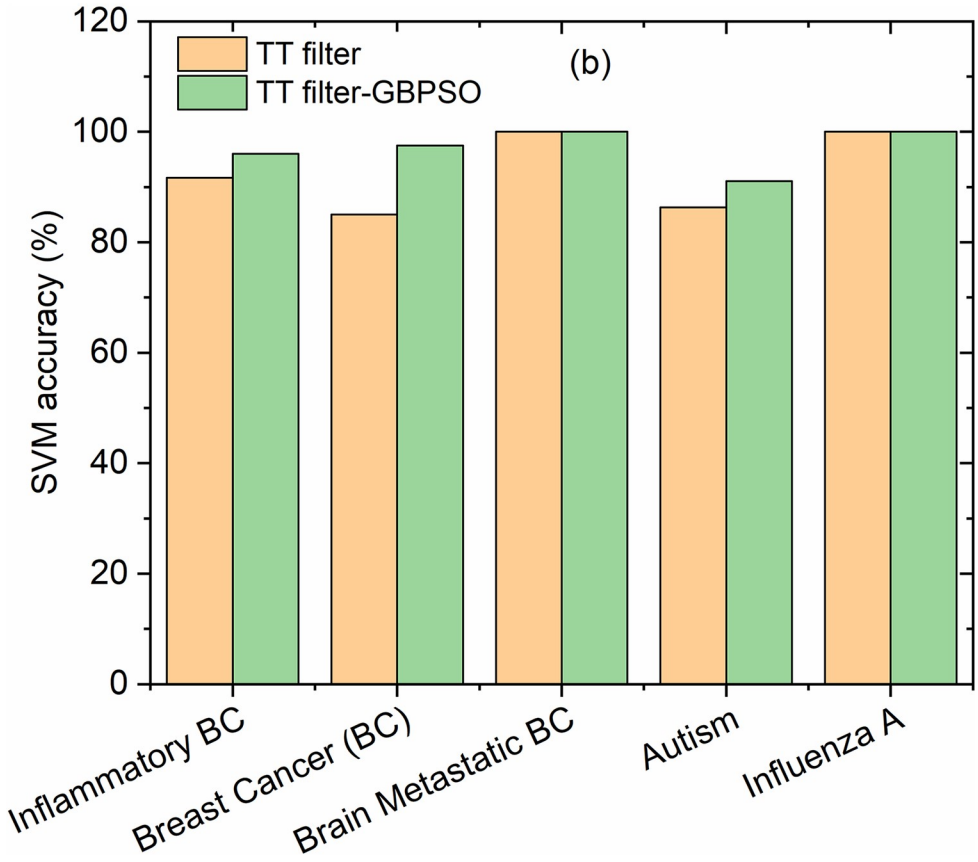

TT filter
TT filter-GBPSO
(b)
SVM accuracy (%)
0
20
40
60
80
100
120
Inflammatory BC
Breast Cancer (BC)
Brain Metastatic BC
Autism
Influenza A

Fig 5. A representative comparison of SVM accuracy of gene selection in the CSV datasets (a) and soft GEO datasets (b) after the application of TT filter and TT filter-GBPSO algorithm using the HDG-select application.

https://doi.org/10.1371/journal.pone.0246039.g005

It is worth to mention that the use of HDG-select tool is also important even if the accuracy is not much improved after the selection process because the HDG-select toll can help in selecting a small subset of the attributed genes while maintaining the original accuracy but reducing the computational burden.

Fig 7 shows the achieved accuracy and precision of gene selection from various high dimensional datasets using the proposed HDG-select application. It was seen from the results that the values of accuracy and precision are in the range from 90% to 100% for different datasets. For instance, the precision of gene selection in the Prostate, Ovarian, Inflammatory BC and Influenza A datasets has reached 100% which is close enough to their classification accuracy. Hence, it can be concluded from the coincidence of the accuracy and precision data that the

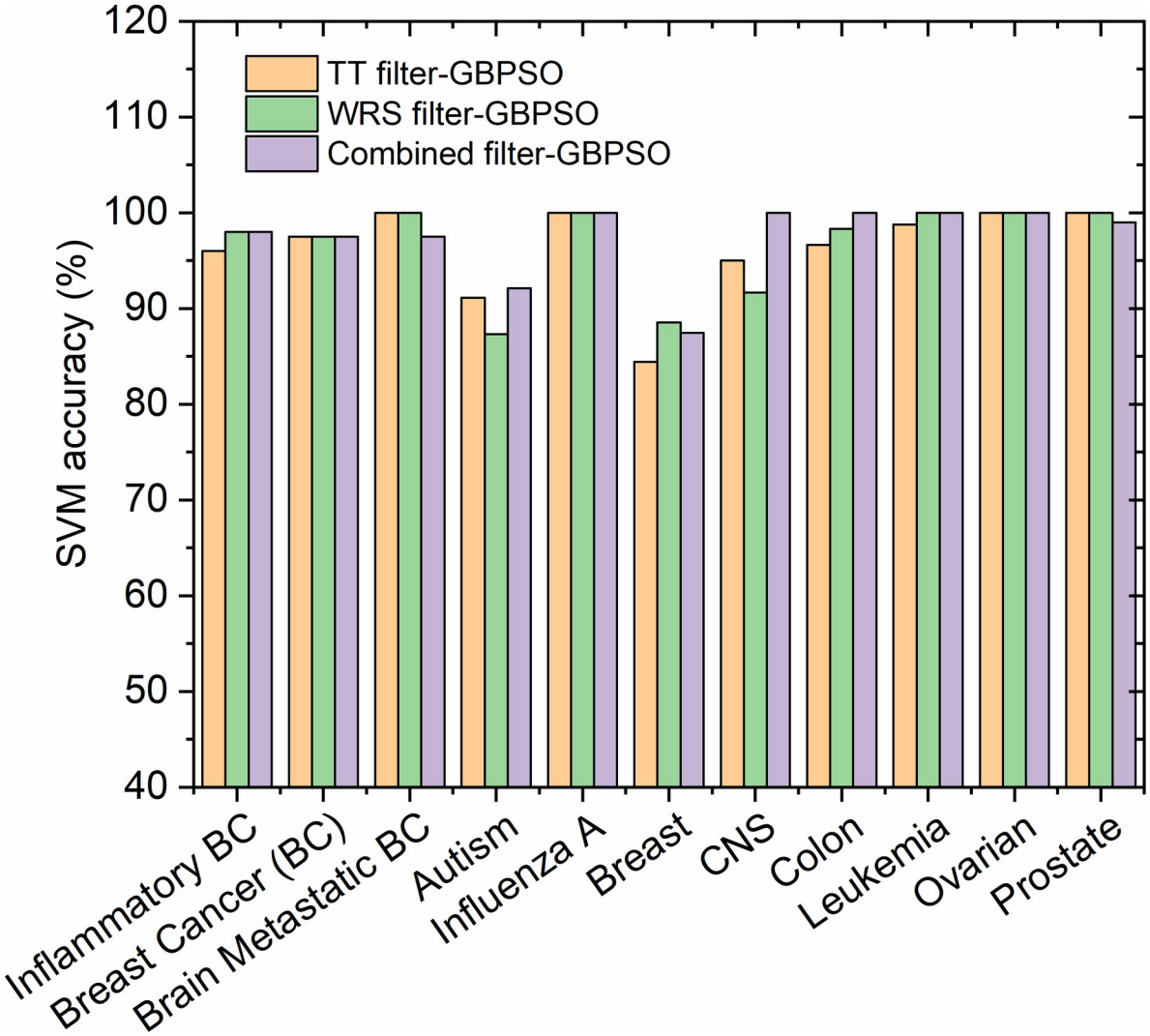


Fig 6. Comparison of the SVM accuracy in selected genes by different filters-GBPSO-SVM approach using the proposed HDG-select application.

https://doi.org/10.1371/journal.pone.0246039.g006

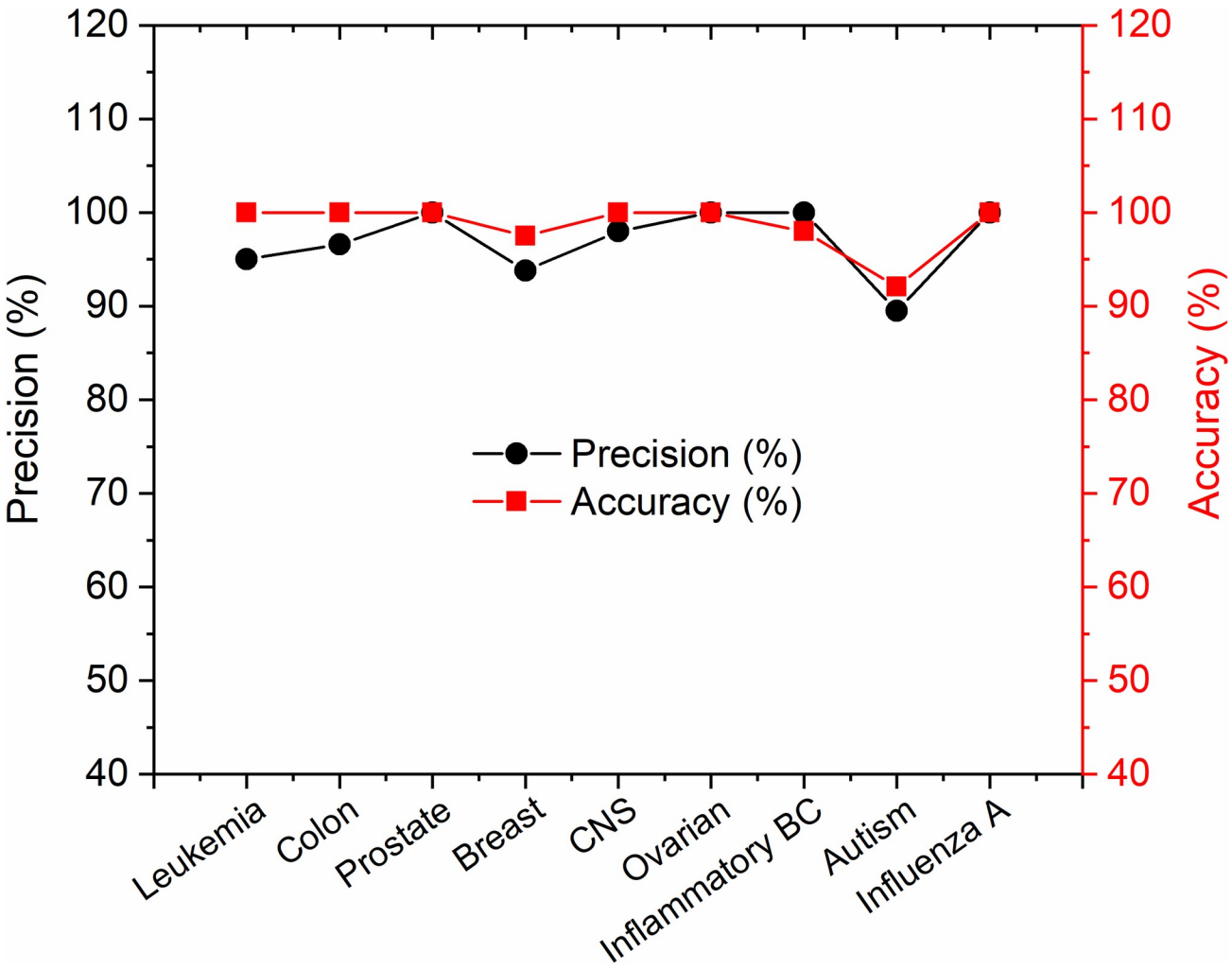


Fig 7. Accuracy and precision of the classification of datasets after final gene selection using the proposed HDG-select application.

https://doi.org/10.1371/journal.pone.0246039.g007

proposed HDG-select has performed very well on various datasets when it was used to select the most attributed genes efficiently, thereby providing a competitive classification accuracy.

Furthermore, a comparison of the results obtained by the HDG-select application to those reported in literature suggesting that the filter-GBPSO-SVM algorithm is more efficient than PCC-BPSO-SVM and PCC-GA-SVM algorithm [11] when it comes to the selection and classification of attributed genes in high dimensional datasets, as shown in Fig 8 and S3 Table.

In order to show the effectiveness of the proposed HDG-select application in the final step of selecting the biomarker/attributed genes, heatmap graphical analysis was plotted. Figs 9 and 10 show the produced heatmap of seven biomarker genes versus the samples for breast cancer and influenza A datasets. It can be seen from the heatmap that the biomarker genes show a high discrimination ability between the control and non-control samples. For instance, gene number 1 and 3 have the highest discrimination ability among the breast cancer and influenza A datasets, respectively.

Table 3 shows a detailed comparison of our proposed application with those reported in literature. It was concluded that the proposed HDG-select outperformed the other tools in terms of overall performance, accessibility and functionality. Noticeably, the most competitive tool

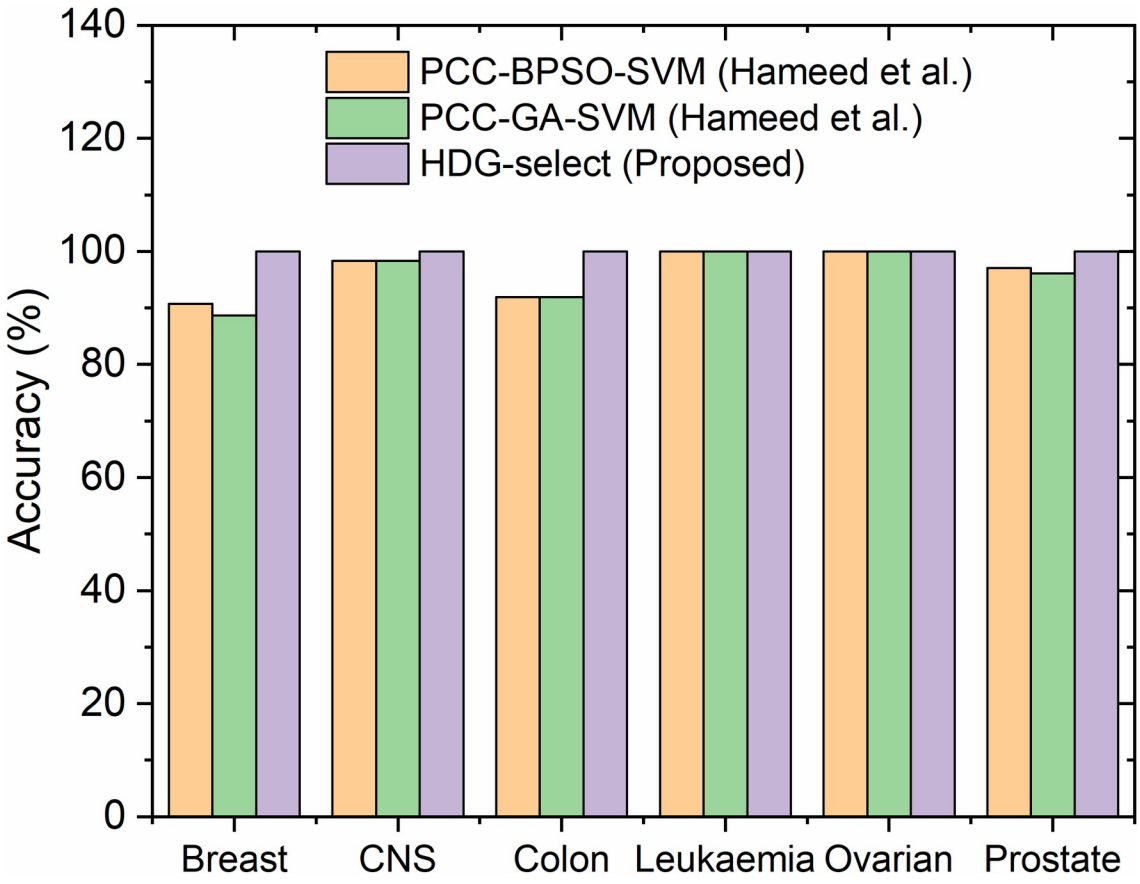


Fig 8. Comparison of the classification accuracy in genes selection by different methods.

https://doi.org/10.1371/journal.pone.0246039.g008

to the proposed HDG-select can be ArrayMining [42]. However, this tool accepts dataset files of CSV format only, while with the help of HDG-select one can also perform dataset curation and dimensionality reduction on the soft GEO datasets. Furthermore, with the proposed HDG-select a multiple gene selection and classification can be performed simultaneously and the selected genes with their expression can be downloaded in CSV format, while ArrayMining [42] can perform one task at each time and the selected genes is downloadable in text format. Table 4 shows a comparison between HDG-select and ArrayMining tool for two representative CSV datasets that were previously analyzed by ArrayMining in terms of accuracy and precision.

Conclusions

A novel GUI based stand-alone application, named as HDG-select, was developed to select and classify the attributed genes in high dimensional datasets effectively. The application was validated on 11 datasets and it was found to perform well on most of high dimensional datasets, including CSV and GEO soft file formats. The proposed HDG-select tool uses efficient algorithm of combined filter-GBPSO-SVM. It was observed that the best approach of increasing gene selection efficiency in high dimensional data is to utilize a mixed filter-GBPSO-SVM algorithm. It was concluded that the proposed HDG-select outperformed the other tools in terms of overall performance, accessibility, and functionality.

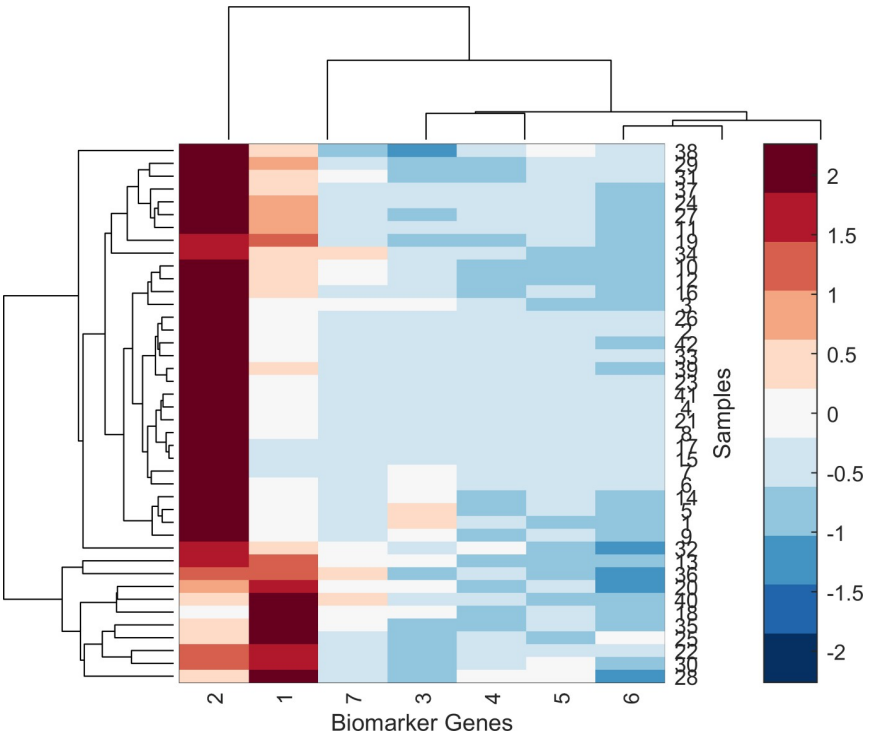


Fig 9. The heatmap of seven selected biomarker genes of breast cancer using the proposed HDG-select application.

https://doi.org/10.1371/journal.pone.0246039.g009

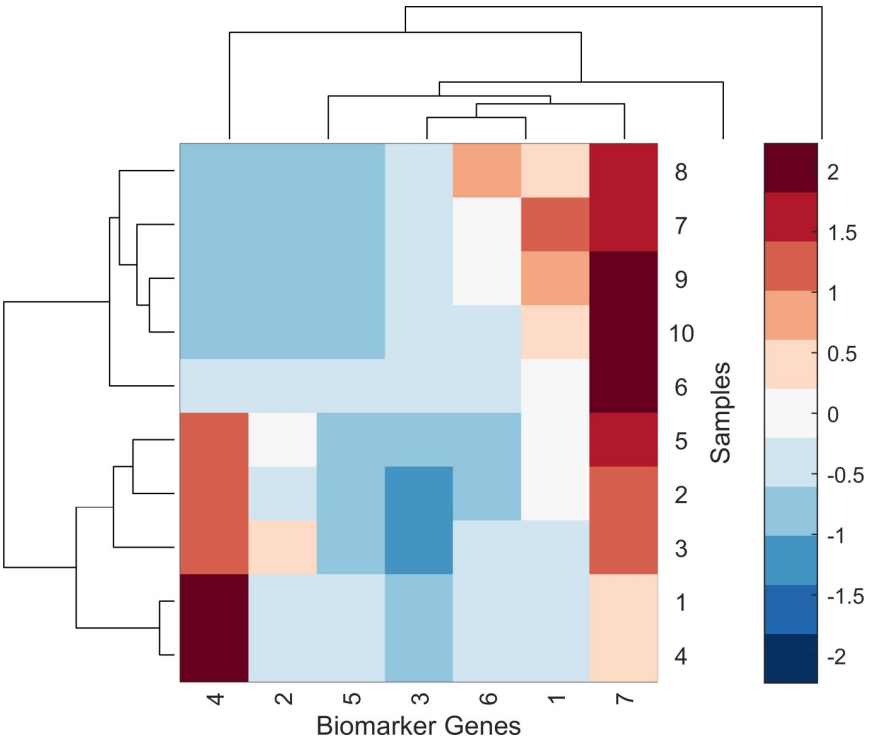


Fig 10. The heatmap of seven selected biomarker genes of influenza A using the proposed HDG-select application.

https://doi.org/10.1371/journal.pone.0246039.g010

Table 3. Comparison of the proposed HDG-select tool with those reported in literature for gene selection and classification.

Application name	language/ package	Gene selection/ classification	Algorithm	Accessibility (online/offline)	GUI Interface	Operating system	Dataset format	User-friendliness
SVM Classifier [43]	Java	No/Yes	SVM	No (online)	Yes	No-restriction	Not known	Low
R. GeneSrF and varSelRF [41]	R, Python	Yes/No	Random forest	No (online)	No	Linux, Unix for R package	CSV	Low
ArrayMining [42]	R, C++ and a PHP-interface	Yes/Yes	Filter Classification clustering	Yes (online)	Yes	No restriction	CSV	Medium
HDG-select (proposed)	Weka, Java and MATLAB	Yes/Yes	Two filters, their combination, GBPSO wrapper and SVM classifier	Yes (offline-no internet required)	Yes	No restriction	CSV and (. soft) GEO dataset	High

https://doi.org/10.1371/journal.pone.0246039.t003

Table 4. Comparison of the proposed HDG-select tool with ArrayMining tool in terms of accuracy and precision of gene selection and classification in high dimensional datasets.

Dataset	Application name	Algorithm (# selected genes)	SVM classification accuracy (%)	Precision (%)
Colon	ArrayMining [42]	Filter (80)	80.7±13	86.6
	HDG-select (Proposed)	Filter-wrapper (30)	98.3±1.7	96.6
Prostate	ArrayMining [42]	Filter (50)	82.2±13	86.6
	HDG-select (Proposed)	Filter-wrapper (30)	100±0	100

https://doi.org/10.1371/journal.pone.0246039.t004

Supporting information

S1 Dataset. The microarray dataset of leukemia cancer in csv format.
(CSV)

S2 Dataset. The microarray dataset of colon cancer in csv format.
(CSV)

S3 Dataset. The microarray dataset of prostate cancer in csv format.
(CSV)

S4 Dataset. The microarray dataset of breast cancer in csv format.
(CSV)

S5 Dataset. The microarray dataset of central nervous system cancer in csv format.
(CSV)

S6 Dataset. The microarray dataset of ovarian cancer in csv format.
(CSV)

S1 Table. The accuracy percentage of selecting attributed genes using Filter-SVM and Filter-GBPSO-SVM approach compared to that of the original dataset.
(DOCX)

S2 Table. The accuracy percentage of Filter-SVM and Filter-GBPSO-SVM algorithm upon various soft files of high dimensional datasets.
(DOCX)

S3 Table. Comparison of the accuracy result of the proposed HDG-select application to those reported in literature.
(DOCX)

Author Contributions

Conceptualization: Shilan S. Hameed.

Data curation: Shilan S. Hameed.

Formal analysis: Shilan S. Hameed, Fahmi F. Muhammadsharif.

Funding acquisition: Rohayanti Hassan.

Investigation: Shilan S. Hameed, Liza Abdul Latiff.

Methodology: Shilan S. Hameed.

Resources: Fahmi F. Muhammadsharif.

Software: Shilan S. Hameed.

Supervision: Rohayanti Hassan, Wan Haslina Hassan, Liza Abdul Latiff.

Validation: Shilan S. Hameed, Fahmi F. Muhammadsharif.

Writing – original draft: Shilan S. Hameed.

Writing – review & editing: Rohayanti Hassan, Wan Haslina Hassan, Fahmi F. Muhammadsharif, Liza Abdul Latiff.

References

1. Govindarajan R, Duraiyan J, Kaliyappan K, Palanisamy M. Microarray and its applications. Journal of Pharmacy & Bioallied Sciences. 2012; 4(Suppl 2):S310–S2. https://doi.org/10.4103/0975-7406.100283 PMC3467903. PMID: 23066278
2. Cosma G, Brown D, Archer M, Khan M, Pockley AG. A survey on computational intelligence approaches for predictive modeling in prostate cancer. Expert Systems with Applications. 2017; 70:1–19. https://doi.org/10.1016/j.eswa.2016.11.006.
3. Singh RK, Sivabalakrishnan M. Feature selection of gene expression data for cancer classification: a review. Procedia Computer Science. 2015; 50:52–7. https://doi.org/10.1016/j.procs.2015.04.060
4. Hira ZM, Gillies DF. A review of feature selection and feature extraction methods applied on microarray data. Advances in bioinformatics. 2015;2015. https://doi.org/10.1155/2015/198363 PMID: 26170834
5. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. bioinformatics. 2007; 23(19):2507–17. https://doi.org/10.1093/bioinformatics/btm344 PMID: 17720704
6. Bommert A, Sun X, Bischl B, Rahnenführer J, Lang M. Benchmark for filter methods for feature selection in high-dimensional classification data. Computational Statistics & Data Analysis. 2020; 143:106839.
7. Bolón-Canedo V, Sánchez-Marono N, Alonso-Betanzos A, Benítez JM, Herrera F. A review of microarray datasets and applied feature selection methods. Information Sciences. 2014; 282:111–35. https://doi.org/10.1016/j.ins.2014.05.042.
8. Chandra Sekhara Rao Annavarapu SD, Banka H. Cancer microarray data feature selection using multi-objective binary particle swarm optimization algorithm. EXCLI journal. 2016; 15:460. https://doi.org/10.17179/excli2016-481 PMID: 27822174
9. Rejani Y, Selvi ST. Early detection of breast cancer using SVM classifier technique. arXiv preprint arXiv:09122314. 2009.
10. Guyon I, Weston J, Barnhill S, Vapnik V. Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning. 2002; 46(1):389–422. https://doi.org/10.1023/A:1012487302797
11. Hameed SS, Muhammad FF, Hassan R, Saeed F. Gene Selection and Classification in Microarray Datasets using a Hybrid Approach of PCC-BPSO/GA with Multi Classifiers. JCS. 2018; 14(6):868–80.
12. Hameed SS, Hassan R, Muhammad FF. Selection and classification of gene expression in autism disorder: Use of a combination of statistical filters and a GBPSO-SVM algorithm. PloS one. 2017; 12(11). https://doi.org/10.1371/journal.pone.0187371 PMID: 29095904
13. Han J, Pei J, Kamber M. Data mining: concepts and techniques: Elsevier; 2011.

14. Thaher T, Heidari AA, Mafarja M, Dong JS, Mirjalili S. Binary Harris Hawks Optimizer for High-Dimensional, Low Sample Size Feature Selection. Evolutionary Machine Learning Techniques: Springer; 2020. p. 251–72.

15. Algamal ZY, Lee MH. A two-stage sparse logistic regression for optimal gene selection in high-dimensional microarray data classification. Advances in data analysis and classification. 2019; 13(3):753–71.

16. Sayed S, Nassef M, Badr A, Farag I. A nested genetic algorithm for feature selection in high-dimensional cancer microarray datasets. Expert Systems with Applications. 2019; 121:233–43.

17. Yan C, Ma J, Luo H, Patel A. Hybrid binary coral reefs optimization algorithm with simulated annealing for feature selection in high-dimensional biomedical datasets. Chemometrics and Intelligent Laboratory Systems. 2019; 184:102–11.

18. Kim ARP. Combination of Ensembles of Regularized Regression Models with Resampling-Based Lasso Feature Selection in High Dimensional Data. Mathematics. 2020; 8(1):110.

19. Song X-f, Zhang Y, Guo Y-n, Sun X-y, Wang Y-l. Variable-size Cooperative Coevolutionary Particle Swarm Optimization for Feature Selection on High-dimensional Data. IEEE Transactions on Evolutionary Computation. 2020.

20. Chen W, Xu Y, Yu Z, Cao W, Chen CP, Han G. Hybrid Dimensionality Reduction Forest With Pruning for High-Dimensional Data Classification. IEEE Access. 2020; 8:40138–50.

21. Karizaki AA, Tavassoli M, editors. A novel hybrid feature selection based on ReliefF and binary dragonfly for high dimensional datasets. 2019 9th International Conference on Computer and Knowledge Engineering (ICCKE); 2019: IEEE.

22. Raman MG, Nivethitha S, Kannan K, Sriram VS. A hybrid approach using rough set theory and hypergraph for feature selection on high-dimensional medical datasets. Soft Computing. 2019; 23 (23):12655–72.

23. Chen L-F, Su C-T, Chen K-H, Wang P-C. Particle swarm optimization for feature selection with application in obstructive sleep apnea diagnosis. Neural Computing and Applications. 2012; 21(8):2087–96. https://doi.org/10.1007/s00521-011-0632-4

24. Alba E, Garcia-Nieto J, Jourdan L, Talbi E-G, editors. Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms. Evolutionary Computation, 2007 CEC 2007 IEEE Congress on; 2007: IEEE.

25. Kennedy J, Eberhart RC, editors. A discrete binary version of the particle swarm algorithm. Systems, Man, and Cybernetics, 1997 Computational Cybernetics and Simulation, 1997 IEEE International Conference on; 1997: IEEE.

26. Zhang Y, Wang S, Phillips P, Ji G. Binary PSO with mutation operator for feature selection using decision tree applied to spam detection. Knowledge-Based Systems. 2014; 64:22–31.

27. Moraglio A, Di Chio C, Togelius J, Poli R. Geometric particle swarm optimization. Journal of Artificial Evolution and Applications. 2008;2008.

28. Cortes C, Vapnik V. Support-vector networks. Machine learning. 1995; 20(3):273–97.

29. Ardjani F, Sadouni K, Benyettou M, editors. Optimization of SVM MultiClass by Particle Swarm (PSO-SVM). 2010 2nd International Workshop on Database Technology and Applications; 2010 27–28 Nov. 2010.

30. Jirapech-Umpai T, Aitken S. Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes. BMC bioinformatics. 2005; 6(1):148. https://doi.org/10.1186/1471-2105-6-148 PMID: 15958165

31. Hassanien AE, Al-Shammari ET, Ghali NI. Computational intelligence techniques in bioinformatics. Computational biology and chemistry. 2013; 47:37–47. https://doi.org/10.1016/j.compbiolchem.2013.04.007 PMID: 23891719

32. Huerta EB, Duval B, Hao J-K, editors. A hybrid GA/SVM approach for gene selection and classification of microarray data. Workshops on Applications of Evolutionary Computation; 2006: Springer.

33. Qian R, Wu Y, Duan X, Kong G, Long H. SVM Multi-Classification Optimization Research based on Multi-Chromosome Genetic Algorithm. International Journal of Performability Engineering. 2018;14(4).

34. Barash E, Sal-Man N, Sabato S, Ziv-Ukelson M. BacPaCS—Bacterial Pathogenicity Classification via Sparse-SVM. Bioinformatics. 2018.

35. Latkowski T, Osowski S, editors. Developing Gene Classifier System for Autism Recognition. International Work-Conference on Artificial Neural Networks; 2015: Springer.

36. García-Nieto J, Alba E, Jourdan L, Talbi E. Sensitivity and specificity based multiobjective approach for feature selection: Application to cancer diagnosis. Information Processing Letters. 2009; 109(16):887–96.

37. Duez M, Giraud M, Herbert R, Rocher T, Salson M, Thonier F. Vidjil: a web platform for analysis of high-throughput repertoire sequencing. PLoS One. 2016; 11(11):e0166126. https://doi.org/10.1371/journal.pone.0166126 PMID: 27835690

38. Kaya H, Hasman H, Larsen J, Stegger M, Johannesen TB, Allesøe RL, et al. SCCmecFinder, a web-based tool for typing of staphylococcal cassette chromosome mec in Staphylococcus aureus using whole-genome sequence data. Msphere. 2018; 3(1). https://doi.org/10.1128/mSphere.00612-17 PMID: 29468193

39. Bruyneel AA, Colas AR, Karakikes I, Mercola M. AlleleProfileR: A versatile tool to identify and profile sequence variants in edited genomes. Plos one. 2019; 14(12):e0226694. https://doi.org/10.1371/journal.pone.0226694 PMID: 31877162

40. Tamazian G, Dobrynin P, Krasheninnikova K, Komissarov A, Koepfli K-P, O'brien SJ. Chromosomer: a reference-based genome arrangement tool for producing draft chromosome sequences. GigaScience. 2016; 5(1):s13742-016–0141-6. https://doi.org/10.1186/s13742-016-0141-6 PMID: 27549770

41. Diaz-Uriarte R. GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest. BMC bioinformatics. 2007; 8(1):328. https://doi.org/10.1186/1471-2105-8-328 PMID: 17767709

42. Glaab E, Garibaldi JM, Krasnogor N. ArrayMining: a modular web-application for microarray analysis combining ensemble and consensus methods with cross-study normalization. BMC bioinformatics. 2009; 10(1):1–7. https://doi.org/10.1186/1471-2105-10-358 PMID: 19863798

43. Pirooznia M, Deng Y. SVM Classifier–a comprehensive java interface for support vector machine classification of microarray data. BMC bioinformatics. 2006; 7(S4):S25.

44. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. science. 1999; 286 (5439):531–7. https://doi.org/10.1126/science.286.5439.531 PMID: 10521349

45. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences. 1999; 96(12):6745–50. https://doi.org/10.1073/pnas.96.12.6745 PMID: 10359783

46. Díaz-Uriarte R, De Andres SA. Gene selection and classification of microarray data using random forest. BMC bioinformatics. 2006; 7(1):3. https://doi.org/10.1186/1471-2105-7-3 PMID: 16398926

47. Z Z., Ong YS, Dash M. Markov blanketembedded genetic algorithm for gene selection. Pattern Recognition. 2007; 40:3236–48. https://doi.org/10.1016/j.patcog.2007.02.007.

48. Autistic children and their father's age: peripheral blood lymphocytes [Internet]. from www.ncbi.nlm.nih.gov. 2011. Available from: http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4431.

49. Alter MD, Kharkar R, Ramsey KE, Craig DW, Melmed RD, Grebe TA, et al. Autism and increased paternal age related changes in global levels of gene expression regulation. PloS one. 2011; 6(2):e16715. https://doi.org/10.1371/journal.pone.0016715 PMID: 21379579

50. El-Fishawy P, State MW. The genetics of autism: key issues, recent findings, and clinical implications. Psychiatric Clinics of North America. 2010; 33(1):83–105.

51. Latkowski T, Osowski S. Computerized system for recognition of autism on the basis of gene expression microarray data. Computers in biology and medicine. 2015; 56:82–8. https://doi.org/10.1016/j.compbiomed.2014.11.004 PMID: 25464350

52. Latkowski T, Osowski S. Data mining for feature selection in gene expression autism data. Expert Systems with Applications. 2015; 42(2):864–72. https://doi.org/10.1016/j.eswa.2014.08.043

53. Lai C, Reinders MJ, van't Veer LJ, Wessels LF. A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. BMC bioinformatics. 2006; 7(1):235. https://doi.org/10.1186/1471-2105-7-235 PMID: 16670007

54. Huertas C, Juárez-Ramírez R, editors. Filter feature selection performance comparison in high-dimensional data: A theoretical and empirical analysis of most popular algorithms. Information Fusion (FUSION), 2014 17th International Conference on; 2014: IEEE.

55. Haury A-C, Gestraud P, Vert J-P. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. PloS one. 2011; 6(12):e28210. https://doi.org/10.1371/journal.pone.0028210 PMID: 22205940

56. Lazar C, Taminau J, Meganck S, Steenhoff D, Coletta A, Molter C, et al. A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2012; 9(4):1106–19. https://doi.org/10.1109/TCBB.2012.33 PMID: 22350210

57. Li S, Wu X, Tan M. Gene selection using hybrid particle swarm optimization and genetic algorithm. Soft Computing-A Fusion of Foundations, Methodologies and Applications. 2008; 12(11):1039–48.

58. Saha S, Seal DB, Ghosh A, Dey KN. A novel gene ranking method using Wilcoxon rank sum test and genetic algorithm. International Journal of Bioinformatics Research and Applications. 2016; 12(3):263–79.

59. Bridge PD, Sawilowsky SS. Increasing physicians' awareness of the impact of statistics on research outcomes: comparative power of the t-test and Wilcoxon rank-sum test in small samples applied research. Journal of clinical epidemiology. 1999; 52(3):229–35. https://doi.org/10.1016/s0895-4356(98)00168-1 PMID: 10210240

60. Ruxton GD. The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test. Behavioral Ecology. 2006; 17(4):688–90.

61. Wilcoxon F. Individual comparisons by ranking methods. Biometrics bulletin. 1945; 1(6):80–3.

62. Wild C, Seber G. The Wilcoxon rank-sum test. Chapter; 2011.

63. Khoshgoftaar T, Dittman D, Wald R, Fazelpour A, editors. First order statistics based feature selection: A diverse and powerful family of feature seleciton techniques. Machine Learning and Applications (ICMLA), 2012 11th International Conference on; 2012: IEEE.

64. Sprent P, Smeeton NC. Applied nonparametric statistical methods: CRC Press; 2016.

65. Geometric Particle Swarm Optimisation [Internet]. 2016 [cited 28/20/2020]. Available from: https://github.com/sebastian-luna-valero/PSOSearch/.

66. wekalab [Internet]. 2016 [cited 28/20/2020]. Available from: https://github.com/NicholasMcCarthy/wekalab.