RESEARCH ARTICLE

Use of relevancy and complementary information for discriminatory gene selection from high-dimensional gene expression data

Md Nazmul Haque[1]*, Sadia Sharmin[2], Amin Ahsan Ali[3], Abu Ashfaqur Sajib[4]*, Mohammad Shoyaib[1]

1 Institute of Information Technology, University of Dhaka, Dhaka, Bangladesh, **2** Department of Computer Science & Engineering, Islamic University of Technology, Dhaka, Bangladesh, **3** Department of Computer Science & Engineering, Independent University, Dhaka, Bangladesh, **4** Department of Genetic Engineering & Biotechnology, University of Dhaka, Dhaka, Bangladesh

* bsse0635@iit.du.ac.bd (MNH); abu.sajib@du.ac.bd (AAS)

OPEN ACCESS

Citation: Haque MN, Sharmin S, Ali AA, Sajib AA, Shoyaib M (2021) Use of relevancy and complementary information for discriminatory gene selection from high-dimensional gene expression data. PLoS ONE 16(10): e0230164. https://doi.org/10.1371/journal.pone.0230164

Editor: Y-h. Taguchi, Chuo University, JAPAN

Received: February 22, 2020

Accepted: September 21, 2021

Published: October 6, 2021

Data Availability Statement: GDS6063, GDS5306, GDS4431, GDS3341, GDS3610, GDS4824, GSE106291 datasets are available at the Gene Expression Omnibus (GEO) database at https://www.ncbi.nlm.nih.gov/geo/.

Funding: AAS received research grant from the innovation fund (2019-2020) of the ICT Division, Ministry of Posts, Telecommunications and Information Technology, Bangladesh (http://ims.ictd.gov.bd/). MNH received research fellowship (56.00.0000.028.33.029.19-329) from ICT Division, Ministry of Posts, Telecommunications

Abstract

With the advent of high-throughput technologies, life sciences are generating a huge amount of varied biomolecular data. Global gene expression profiles provide a snapshot of all the genes that are transcribed in a cell or in a tissue under a particular condition. The high-dimensionality of such gene expression data (*i.e.*, very large number of features/genes analyzed with relatively much less number of samples) makes it difficult to identify the key genes (biomarkers) that are truly attributing to a particular phenotype or condition, (such as cancer), *de novo*. For identifying the key genes from gene expression data, among the existing literature, mutual information (MI) is one of the most successful criteria. However, the correction of MI for finite sample is not taken into account in this regard. It is also important to incorporate dynamic discretization of genes for more relevant gene selection, although this is not considered in the available methods. Besides, it is usually suggested in current studies to remove redundant genes which is particularly inappropriate for biological data, as a group of genes may connect to each other for downstreaming proteins. Thus, despite being redundant, it is needed to add the genes which provide additional useful information for the disease. Addressing these issues, we proposed Mutual information based Gene Selection method (*MGS*) for selecting informative genes. Moreover, to rank these selected genes, we extended *MGS* and propose two ranking methods on the selected genes, such as MGS_f—based on frequency and MGS_{rf}—based on Random Forest. The proposed method not only obtained better classification rates on gene expression datasets derived from different gene expression studies compared to recently reported methods but also detected the key genes relevant to pathways with a causal relationship to the disease, which indicate that it will also able to find the responsible genes for an unknown disease data.

and Information Technology, Bangladesh (http://ims.ictd.gov.bd/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Genes are the physical and functional units of hereditary genetic information. The activity and/or expression level of a gene affects the synthesis of downstream protein(s) that dictates specific functionality in a cell. Therefore, the properties as well as the expression levels of a particular set of genes are responsible for a particular phenotype such as disease or tissue morphology. Those genes which are able to differentiate between different states (such as normal vs. diseased, quiescent vs. proliferating, adult vs. stem cells, etc.) of cells are called informative genes or biomarkers (a measurable indicator of a particular state). Identification of these informative genes is very important for elucidating developmental and disease mechanisms, disease diagnosis, drug development, etc. Especially, for the identification of different cancers, these informative genes may provide invaluable information for the improvement of diagnosis, prognosis, and treatment. For a set of known diseases, such informative genes are already identified using wet-lab verification. A computational method that can identify these known informative genes can be considered as a reliable method. Again, for known diseases, there might be few more informative genes (due to ethnicity variation) which are responsible for that disease. More importantly, for a new disease, these informative genes are unknown. Identifying these genes through wet-lab techniques are costly and time consuming. These time and cost can be significantly reduced by a reliable computation based method which is the main objective of this paper.

Usually, studies to generate disease specific gene expression profiles such as cancer comprise of a small number of control and patient samples, but tens of thousands of genes (high dimensional data) in each sample where only a few of the genes are responsible for a disease. Identification of a small subset of differentially expressed genes among thousands in cancerous cells compared to the normal ones is a challenging task and considered as NP (non-deterministic polynomial time) hard or NP-complete [1]. Therefore, the feature/gene selection methods can be a convenient and useful way to find a subset of genes relevant to a particular cancer. In this paper, we use the terms “gene” and “feature” interchangeably.

Till to date, several gene selection methods have been proposed, particularly for cancer data classification [2–4]. These methods can be categorized into three types, such as “Filter”, “Wrapper”and “Hybrid” [5]. Among them, filter based methods are more popular as these can assess the property of features without being dependent on any particular classifier. Filter based methods select a subset of features based on some criteria such as correlation coefficient [6], t-statistics [7], distance [8, 9], Mutual Information(MI) [10–13]. Among these, MI based methods are popular for feature selection due to their ability to capture non-linear dependencies between features. One of the recent works used Minimum Redundancy Maximum Relevance (MRMR) [3] where each gene was selected incrementally to hold the highest discriminatory power (relevancy) with the target class (control/cancer) and the lowest dependency (redundancy) with other selected genes. However, in this method, bias corrections (errors occurred due to finite number of samples) are not considered and there are some genes which add some additional information about the class that are discarded. To solve this issue, a new information theoretic measure such as complementary (additional) information that a gene has about the class (which is not found in the already selected subset of genes) has been proposed in [11, 14]. These methods attempted to estimate the joint mutual information of a feature subset with the class. Another method, modified Discretization and feature Selection based on Mutual information (mDSM) [11] includes bias correction and captures complementary information. Relaxmrmr [14] and DSbM [13] add a higher order term, namely feature-feature interaction in addition to the complementary information. However, all these methods discard those genes considering as redundant which may provide complementary information about a particular disease (class). The exclusion of a gene considering only a pair-wise correlation may hamper of finding informative and distinguishable genes because a group of genes is connected to each other to perform a particular function.

In contrast to filter based methods, wrapper based methods are classifier dependent. Wrapper based methods select the most discriminant subset of features by minimizing the prediction error of a particular classifier [15]. Support Vector Machine based on the Recursive Feature Elimination (SVM-RFE) [2] is considered to be one of the best performing wrapper methods. It ranks the genes using SVM and selects the important genes using recursive feature elimination strategy. Different variants of SVM-RFE have also been proposed [16, 17]. Although the wrapper based feature selection methods provide better performances, these methods become computationally expensive when the feature size grows. Moreover, these methods may not provide the optimal solution for other classifiers [18].

To combine the advantages of wrapper and filter based methods, a hybrid approach was introduced which first selected candidate gene subset from the original gene set via computationally-efficient filter method and then the candidate gene subset was further refined by wrapper method. An example of a hybrid method named Information Guided Interactive Search (IGIS) [19] that selected the best set of genes based joint MI. However, this method selected more genes than the wrapper or the hybrid algorithms. Addressing the limitations of IGIS, improved Interaction information-Guided Incremental Selection (*IGIS*+) [20] was proposed, where the first gene was selected based on the highest accuracy using KNN and CART classifiers and utilized Cohen's *d* test to add a new gene into the selected gene subset. One major limitation of *IGIS*+ is that it uses several handcrafted thresholds. There are several popular bio-inspired algorithms to find out the optimal set of features. Almugren et al. in [21] provided an extensive review of the bio-inspired hybrid methods. Alshamlan et. al. proposed a hybrid artificial bee colony [22] and a genetic bee colony [23] optimization method that uses MRMR criterion. El Akadi et al. [24] proposed a genetic algorithm based on MRMR criterion. In these methods, MRMR criterion is used to filter noise and redundant genes in the high-dimensional microarray data and then the bio-inspired algorithm uses the classifier accuracy as a fitness function to select the highly discriminating genes. Particle swarm optimization is a kind of bio-inspired swarm intelligence optimization method which was used to select informative genes [25]. In this method, informative genes are selected in autism spectrum disorder by utilizing a combination of various statistical filters and a wrapper-based Geometric Binary Particle Swarm Optimization-Support Vector Machine (GBPSO-SVM) algorithm. Another recent hybrid method was introduced by Hameed et al. [26] named HDG-select. It provides a graphical user interface that uses mixed filter-GBPSO-SVM for feature selection, while SVM is used for disease classification. Most bio-inspired algorithms use local searches with random restart or population based methods. However, these algorithms still can get stuck at a local optimum. In order to solve the optimization problems globally, a parallel search strategy was attempted in [27]. It incorporated parallel search strategies based on semi-definite programming or quadratic programming that can find the feature subset in polynomial time.

Recently, deep learning based methods show better accuracy in different classification problems such as image [28, 29], text [30] or audio [31] classification. Deep learning based methods have also been proposed for gene expression data [4] where the authors developed a new model namely Forest Deep Neural Network (*fDNN*) that incorporated deep neural network (DNN) with random forest (RF) to solve the problem of learning from small sample data having a large number of genes. RF was used to reduce the dimension of these datasets by detecting the important genes in a supervised manner. This new feature representation was then fed into DNN to predict the outcomes. However, this method does not make use of the main advantage of deep learning, which is automatic feature extraction in solving classification problems. On the other hand, using a neural network as a black box to extract new features from gene expression data reduces the interpretability of the classifier, which is important in studies such as disease (cancer) classification.

We in this paper choose filter based methods for gene selection instead of deep learning or wrapper/hybrid methods due to the useful properties that filter based methods have, namely interpretability, classifier independence, and superior performance. Moreover, we adopt filter based methods that use selection criteria based on MI for reasons mentioned previously. However, one of the challenges of MI based methods is to reliably estimate the *MI* when the dataset is high-dimensional but contains few samples. Gene expression datasets have this characteristic. There has been a lot of effort to better approximate the *MI*. Among them, one of the recent works is modified Discretization and Selection of feature based on *MI* (*mDSM*) [11], where the authors showed that during the calculation of MI for finite samples, there exist some errors (bias) for all the three terms namely relevancy, redundancy and complementary information. Moreover, for selecting a feature, they proposed to use χ^2 statistics by showing that these terms follow χ^2 distribution. Despite having a few good characteristics, MI based methods might discard informative genes by incorporating the term *redundancy* in gene expression data [20]. Note that, usually most of the existing works improve the classification accuracy whereas it is also important to identify genes that improve classification accuracy and are relevant to a particular disease [5].

To solve the aforementioned problems, we propose a new MI based filter method, namely Mutual information based Gene Selection (*MGS*) that achieves better classification performance as well as captures biological significance with high dimensional data. The main contributions of this study are as follows: first, a gene selection technique is proposed for identifying the discriminating genes relevant to a particular disease based on their relevancy and complementary information. Second, a statistical test is used to select genes without a handcrafted threshold. Third, two ranking techniques are proposed to rank the selected genes and select top η genes as biomarkers for disease classification. Finally, the selected informative genes are validated using already wet-lab tested results with a causal relationship to a particular type of disease (phenotype) with the hope that this method will also work well for unknown disease data.

Materials and methods

Dataset description

To find the informative genes and to assess the performance of the proposed method compared to the existing ones, we choose datasets that have different characteristics, such as balanced and imbalanced datasets, and small and relatively large samples having different diseases. We used seven different gene expression datasets such as GDS3341 [32], GDS3610 [33], GDS4824 [34], GSE106291 [35], GDS4431 [36], GDS5306 [37] and GDS6063 [38] retrieved from the Gene Expression Omnibus (GEO) database [39] at the National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov). These datasets are grouped into two sets based on the distribution of control and disease samples: Balanced datasets (GDS3341, GDS3610, GDS4824 and GSE106291) and Imbalanced datasets (GDS4431, GDS5306 and GDS6063) having small and relatively large samples. The description of datasets is given in Table 1. Expression data of multiple probes for the same gene were merged. At first, we download these datasets (.soft extension) from the NCBI site, and it is transformed into CSV format taking samples and features (genes). Here, features are defined by probe id and gene symbol. There are various cases where probe id is different but same gene symbol name. In this case, we merge these gene symbols by taking their mean. In addition, the genes which

Table 1. Summary of the datasets used in this study.

Dataset type	Sample size	Dataset ID	Description	Total samples	Control samples	Disease samples	Features/ genes
Balanced	Small	GDS6063	Influenza A infected plasma-cytoid dendritic cells(pDC)	10	5	5	36825
		GDS5306	Breast cancer brain metastasis specimens and non-metastatic primary breast tumors	38	19	19	32389
	Relatively large	GDS4431	Peripheral blood lymphocytes of autistic and non-autistic child	146	69	77	30803
Imbalanced	Small	GDS4824	Prostate cancer	21	8	13	30872
		GDS3610	Nasopharyngeal carcinoma	28	3	25	14126
		GDS3341	Nasopharyngeal carcinoma	41	10	31	30865
	Relatively large	GSE106291	Acute myeloid leukemia	235	71	164	21403

https://doi.org/10.1371/journal.pone.0230164.t001

have no expression values over the samples are discarded. All these datasets contained much less number of samples compared to the number of genes. These datasets are publicly available at https://doi.org/10.6084/m9.figshare.16680355.

Gene selection and validation processes

The overall process of the proposed MI based Gene Selection (*MGS*) is shown in Fig 1. We first identified the informative genes using *MGS* and then selected top η genes by ranking them according to their performance (Fig 1A). Finally, we use these η genes for classification (Fig 1B) and validating the biological significance. The following subsections describe our method with further details.

Gene selection

For the identification of a gene subset, we used a filter based gene selection method that approximated the joint *MI* with respect to the class variable. In order to identify an informative gene subset, we first subdivided the given gene expression dataset into *K* subsets and applied K-fold cross validation (KFCV). However, when the number of samples (n) is small ($n < 100$), Leave One Out Cross Validation (LOOCV) is applied where $K = 1$. In *MGS*, we incorporated a variant of the *mDSM* [11] by modifying the selection criteria so that it can identify biologically relevant genes for a disease. The accumulation of all genes identified by *MGS* from *K* different subsets was defined here as selected gene subset (G_S). Finally, to rank these selected gene subset (G_S), two ranking criteria namely *MGS* frequency-based ranking (MGS_f) and *MGS* Random Forest (RF) based ranking (MGS_{rf}) were proposed to select the top η genes as biomarkers.

- **Gene subset (G_S) selection**: To measure how much information a particular gene expression dataset provided for the identification of a disease, we calculated MI between the expression values of a gene g_i and the class variable C. This MI represented the relevancy of a gene that revealed the degree of importance of that gene in disease data classification. Note that, before calculating the MI, the gene expression data was discretized which was necessary for noise reduction and data simplification, and thus resulted in maximizing the relevancy of a gene to the target class C. For calculating the relevance between g_i and C, MI was calculated using Eq 1.

$$J_{rel}(g_i) = I(g_i^{d_i}; C) - \frac{(\mathcal{I} - 1)(\mathcal{K} - 1)}{2N \ln 2} \tag{1}$$

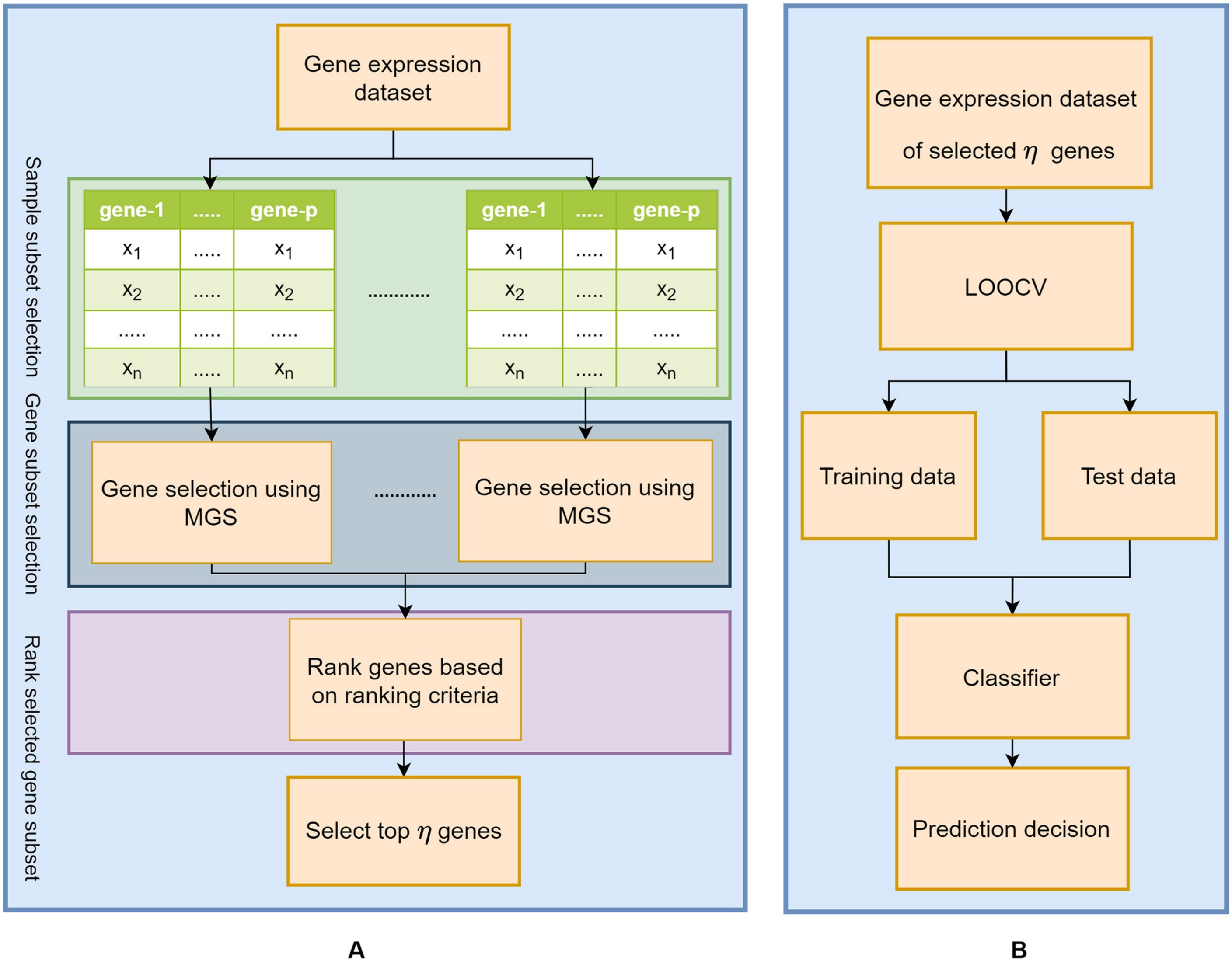

Fig 1. Overall process of the proposed method. (A) Gene selection. (B) Classification.

https://doi.org/10.1371/journal.pone.0230164.g001

where, $g_i^{d_i}$ denotes gene g_i with d_i discretization levels. The second term of the right hand side of Eq (1) is the bias correction term for calculating the relevancy where $\mathcal{I}$, $\mathcal{K}$ and N represent the discretization levels of gene g_i, the total number of classes in C and the total number of samples respectively. For each gene g_i, the minimum discretization levels d_i was chosen for which $J_{rel}(g_i)$ was greater than its χ^2 critical value ($x_C^2(rel)$) and thus helped to determine whether the gene was significantly relevant or not. This test could be done as it could be shown that the relevancy followed χ^2 distribution with $(\mathcal{I}-1)(\mathcal{K}-1)$ degrees of freedom. The genes which satisfied the χ^2 critical value were included in the candidate gene subset, G_c. Then, these candidate genes, G_c were ranked in descending order based on the relevancy. As the top ranked gene was considered to be the most important, we included it to the selected gene subset G_S at first. Now, the second ranked one was evaluated for selection based on its score calculated using Eq 2.

$$\begin{aligned} J_{MGS}(g_i) = \; & I(g_i^{d_j}; C) - \frac{(\mathcal{I}-1)(\mathcal{K}-1)}{2N\ln 2} \\ & + \frac{1}{\mid G_S \mid} \sum_{g_s \in G_S} \left[I(g_i^{d_j}; g_s \mid C) - \frac{(\mathcal{I}-1)(\mathcal{J}-1)\mathcal{K}}{2N\ln 2} \right] \end{aligned} \tag{2}$$

Here, along with relevancy, the complementary information ($I(g_i^{d_j}; g_s \mid C)$) of a new gene was also calculated. The complementary information $I(g_i^{d_j}; g_s \mid C)$ due to g_i for the already selected gene in g_s revealed the dependency among those genes while identifying the class variable C. Here, $\mathcal{J}$ represents the discretization levels of a gene in G_S. The last term in Eq 2 is the bias correction for complementary information. While calculating the value of J_{MGS}, the discretization level (d_i) of the g_i which was fixed using Eq (1) was also shifted by a small amount (±δ) to check whether the value of J_{MGS} is increasing because a small shifting of discretization might increase the value of J_{MGS} and this new discretization value was chosen dynamically considering the dependency among the genes. Now, for a particular gene (g_i), if the value of J_{MGS} was larger than the χ^2 critical value ($\chi^2_C(MGS)$), then it was placed into the selected gene subset. When the relevancy and complementary information of a g_i was significant, it was selected, otherwise discarded. So, identification of genes that maximize J_{MGS} indicated the genes which were strongly relevant with the class C with greater additional information would be adopted to the selected subset throughout this process.

It is noteworthy to mention that a group of genes with similar expression values may exist which will be identified as redundant. However, if these have complementary (additional) information about the class, it is necessary to incorporate that gene into the selected subset even though these are redundant. Inclusion of the redundant genes is sensible because; usually a set of genes contributes mutually for a particular task in our body and these genes may share a similar or correlated expression profile. The biological importance of such inclusions is presented in the Results and discussion section. The whole procedure of selecting gene subset are illustrated in Algorithm **1**.

- **Rank the selected gene subset**: The same subset of genes was not always selected during the selection of genes by *MGS* at each iteration of LOOCV. For example, in a dataset having n number of samples, we used (n − 1) samples for training and the n^{th} sample for testing. After passing the training data to *MGS*, we got an informative selected gene subset. This was repeated n times and aggregated all the selected gene subsets (G_S) and considered the union of these subsets to get G_{SU}. Afterward, these genes in G_{SU} were ranked using one of the following two ranking criteria.
  - ***MGS*$_f$**: This ranking was performed based on the following assumption. *Assumption*: The genes which are selected in every iterations are likely to have more discriminating power and biological significance.

    To quantify the *Assumption*, we computed the relative frequency of every selected gene, S_i in G_{SU} using Eq 3.

$$P(S_i) = \frac{F_{S_i}}{N_{G_{SU}}} \tag{3}$$

Here, $N_{G_{SU}}$, F_{S_i} and $P(S_i)$ are the total number of genes in G_{SU}, frequency of the selected gene S_i and the relative frequency of gene S_i respectively. For example, we had two selected gene subsets, $L_1 = \{g_1, g_3, g_4, g_5, g_6\}$ and $L_2 = \{g_1, g_2, g_4, g_6\}$. Here, the unique genes were G_{SU}

= {g_1, g_2, g_3, g_4, g_5, g_6} and the frequencies of these unique genes were F = 2, 1, 1, 2, 1, 2 respectively. So, the relative frequencies were $P(S_i)$ = 2/6, 1/6, 1/6, 2/6, 1/6, 2/6. Thus, based on the $P(S_i)$, ranked genes were G_{SR} = g_1, g_4, g_6, g_2, g_3, g_5.

- ***MGS**$_{rf}$*: Informative genes have the ability to split the control and disease samples into two groups. To find the more informative genes, it is needed to rank the selected gene subset. In order to rank the genes G_{SU}, it is necessary to measure how much information a gene contains. To measure the information content of a gene, we can use Information Gain (IG) criterion. IG is used in decision trees [40] to select features that reduces the entropy of the data most by splitting data into two groups (called the the left and right child in a decision tree). We used weighted IG given in Eq 4.

$$IG = \frac{N_t}{N}\left[H(node_{Parent}) - \frac{N_L}{N_t} * H(node_{Leftchild}) - \frac{N_R}{N_t} * H(node_{Rightchild})\right] \tag{4}$$

Where, N_t is the number of samples at the current (parent) node, N is the total number of samples, N_L is the number of samples in the left child, and N_R is the number of samples in the right child. $H(node)$ is the entropy at the node. The entropy was calculated using Eq 5.

$$H(node) = -\sum_{i=1}^{C} P_i log P_i \tag{5}$$

Here, P_i is the probability of the outcome/class, i. Each node in a DT contains a gene with its corresponding weighted IG. Besides, to make the weighted IG more robust, we used M number of DTs to construct a Random Forest and took the average of IGs for each gene g_j ∈ G_{SU} using Eq 6.

$$IG_{g_j} = \frac{1}{\sum_{i=1}^{V} \delta(v_i.g, g_j)}\left[\sum_{i=1}^{V} \delta(v_i.g, g_j) * v_i.IG\right] \tag{6}$$

Here, $V = \{v_i, v_{i+1},..,v_k\} = \{(g_i, IG_i), (g_{i+1}, IG_{i+1}),..,(g_k, IG_k)\}$ and k is the total number of nodes in the random forest. That is, for each node of the random forest, we stored the corresponding gene and its weighted IG in V. $\delta(v_i.g, g_j) = 1$ if $v_i.g = g_j$, and 0, otherwise. This average score can be used as the importance score of each gene. In our case, this importance score represented how important a particular gene was to explain the target class. Then, based on the importance score, the genes from G_{SU} were ranked in descending order. And finally, from the ranked genes, top η genes were taken as biomarkers and the performance metrics were calculated.

Algorithm 1: MGS

```
Input: Set of genes $G$, maximum discretization level $max_d$
Output: Selected subset of genes, $G_S$
Initialize candidate gene subset ($G_c$), its discretization level ($D_c$) and relevance ($J_c$) with $\emptyset$.
for each $g_i \in G$ do
  for all $j$ = 2 to $max_d$ do
    Discretize $g_i$ with $j$ intervals
    Calculate $J_{rel}(g_i)$ using Eq (1)
    if $J_{rel}(g_i) > \chi^2_C(rel)$ then
      $D_c \Leftarrow D_c \cup j$;
      $J_c \Leftarrow J_c \cup J_{rel}(g_i)$
      $G_c \Leftarrow G_c \cup g_i$
      break
    end if
  end for
end for
Sort $G_c$ in decreasing order based on their corresponding $J_c$ values
Select $g_1$ and the corresponding $d_1$ from $G_c$ with max $J_c$
$G_S \Leftarrow \{g_1\}$
$D_S \Leftarrow d_1$
$G_c \Leftarrow G_c \setminus g_1$
for each $g_i \in G_c$, $d_i \in D_c$, $j_i \in J_c$ do
  Initialize threshold, $T \Leftarrow 0$
  for all $j = d_i - \delta$ to $d_i + \delta$ do
    Discretize $g_i$ with $j$ intervals
    Calculate $J_{MGS}(g_i)$ using Eq (2)
    if $J_{MGS}(g_i) > \chi^2_C(MGS)$ then
      $d_i \Leftarrow j$;
      $T \Leftarrow \chi^2_C(MGS)$
      $j_i \Leftarrow J_{MGS}(g_i)$
    end if
  end for
  if $j_i > T$ then
    $G_S \Leftarrow G_S \cup g_i$
    $D_S \Leftarrow D_S \cup d_i$
  end if
  $G_c \Leftarrow G_c \setminus g_i$
end for
Return $G_S$
```

Classification

For classification, as shown in Fig 1B, only the selected top η genes from the previous step were used in the train and test data to predict the outcome. To assess the performance of a gene selection method, we considered two performance metrics, *accuracy* and Area Under the Receiver Operating Characteristic Curve (*AUROC*). *Accuracy* is the percentage of samples that are predicted as the true class. *AUROC* represents degree or measure of separability between classes, and it can be used both balanced and imbalanced datasets, specially imbalanced dataset. *ROC* is a probability curve of a classifier at various thresholds. It plots a curve based on the true positive rate (TPR) and false positive rate (FPR) represented in Eqs 7 and 8.

$$TPR = \frac{TP}{TP + FN} \tag{7}$$

$$FPR = \frac{FP}{FP + TN} \tag{8}$$

here, "TP" and "TN" are the numbers of positive and negative samples that are correctly classified. "FP" is the number of negative-class samples misclassified as the positive class, and "FN" is the number of positive-class samples misclassified as the negative class. To compute the points in a *ROC* curve, *AUROC* computes an aggregate measure of various thresholds. For our experiments, the reported results were calculated by taking the average over the KFCV/LOOCV process for these two metrics. Selection of the informative features in the *MGS* step was implemented using MATLAB and ranking these informative features in MGS_f and MGS_{rf} step was implemented using Python with the package scikit-learn [41]. To evaluate the performance of the proposed and existing methods, different classifiers such as SVM, RF classifiers, XGboost [42], PE*k*NN [43] can be used. In this paper, we only use two simple classifiers namely SVM (linear kernel) and Random Forest to compare different methods. These classifiers were implemented using Python with Scikit-learn packages. The source code is available, which can be downloaded from GitHub (https://github.com/Shisir/MGS). All experiments are conducted using a PC with Core-i7, CPU (3.60GHz x 8), and 16GB of RAM.

Biological interpretation of the selected genes

We used NetworkAnalyst [44] to interpret the biological significance of the selected genes. NetworkAnalyst is a bioinformatics platform to interpret gene expression data within the context of protein-protein interaction (PPI) networks which is widely used in many renowned researches such as [45–47]. It uses well-established walktrap algorithm [48]. The general idea of walktrap algorithm is that if we perform random walks on the PPI network, the walks are more likely to stay within the same module (nodes those are closely connected to each other) because there are only a few edges that lead outside a given module. Then, to assess the goodness of these modules, modularity [49] is used. The modularity quality function is based on the comparison with a random graph that is not expected to have a cluster structure. As the input of NetworkAnalyst, we used top η selected genes for each dataset determined by our proposed and the previously described methods [4, 11, 20, 40]. Only the PPI networks that accommodate these genes with False Discovery Rate (FDR) < 0.05 were considered. FDR is more stringent than the p-value and has become invaluable in transcriptional profiling, and large-scale bioinformatics analysis in general. Since the nature of the biological samples in the datasets was known, we assessed the performance of the compared methods based on their abilities to identify the key pathways affected in the corresponding sample types.

An illustrative example

Here, we present a toy example in order to demonstrate the overall mechanism of the proposed method. Let us assume a dataset having 20 genes (g_1, g_2, .., g_{20}) with 10 samples where the distribution of control and disease samples are seven and three, respectively. First, in *MGS*, we calculate the relevance for the expression values of each gene with the minimum discretization level using Eq 1 and the first gene is selected which has the highest relevance. Then, the next gene subset is selected by a small change ($\pm\delta$) of the initial discretization level of each gene to maximize the J_{MGS} criterion mentioned in Eq 2. In this way, the genes are assessed considering its interaction with other genes. In this example, 10 out of 20 genes are selected as a selected gene subset (G_S). After that, to rank the selected gene subset, in MGS_f, the relative frequencies of the selected genes are recorded using Eq 3 over the LOOCV. On the other hand, in MGS_{rf}, an importance score is given to every candidate gene using Eqs 4–6 over the LOOCV. Table 2 represents the relative frequency distribution and importance score of the selected genes after applying MGS_f and MGS_{rf}, respectively. From these ranking, top $\eta(= 5)$ genes are considered as biomarkers. So, the top 5 ranked gene subset of MGS_f and MGS_{rf} are g_2, g_3, g_5, g_9, g_{12} and g_3, g_2, g_{13}, g_9, g_5, respectively. Finally, using these biomarkers, two classifiers (*SVM* and *RF*) are applied to compute *accuracy* and *AUROC*. Besides, these biomarkers are also assessed for biological interpretation.

Table 2. Relative frequency distribution (MGS_f) and importance score (MGS_{rf}) of the selected gene subset.

Gene	g_2	g_3	g_5	g_9	g_{12}	g_{13}	g_{15}	g_{17}	g_{19}	g_{20}
MGS_f score	0.8	0.8	0.7	0.6	0.6	0.5	0.4	0.4	0.2	0.2
MGS_{rf} score	0.92	0.97	0.73	0.76	0.69	0.89	0.44	0.57	0.39	0.21

https://doi.org/10.1371/journal.pone.0230164.t002

Results and discussion

We compared the performances of our proposed filter based methods (*MGS_f* and *MGS_{rf}*) to other already renowned methods such as *RF* (filter) [40], *fDNN* (embedded) [4], *IGIS+* (hybrid) [20], *HDG* (hybrid) [26] and *mDSM* (filter) [11].

In this study, we applied the aforementioned methods on seven gene expression datasets. Here, we first discussed the classification performance of all methods using the top η genes and then compared the performance of all methods for different numbers of top η genes to assess the robustness of our method. In situations where the gene selection methods selected less than 10 genes, we used all the selected ones in further analysis. Finally, we provided a biological interpretation of the top ten (η) selected genes. For a fair comparison, we followed the same training and testing protocol for all the methods. With *RF*, *fDNN* and *MGS_{rf}* (where random forest was used), we applied 300 decision trees.

Classification performance

Tables 3 and 4 summarized the comparative results of the proposed methods along with the existing methods on balanced and imbalanced datasets respectively and the values in boldface represent the classification performance of the best performing method for a particular classifier. For balanced datasets, the average *accuracy* and *AUROC* indicated the superiority of *MGS_f* and *MGS_{rf}* against other five gene selection methods on two different classifiers (Table 3). With GDS6063 and GDS5306 datasets, *MGS_f* and *MGS_{rf}* performed better than the other reported methods. Although *fDNN* performed slightly better than *MGS_f* and *MGS_{rf}* with GDS5306 dataset, the biological significance of the selected genes was not as satisfactory as our methods (discussed later).

For imbalanced datasets, the similar superiority of the *MGS_f* and *MGS_{rf}* could be observed in most of the cases compared to the existing methods (Table 4), which indicated that the proposed methods selected more informative genes. With GDS3341 and GDS4824 datasets, all methods except *RF* were able to perfectly differentiate the control and disease samples for both *SVM* and *RF* classifiers. The small number of samples compared to a large number of genes might be the reason behind the relatively poor performance of *RF*. Even though the other methods performed well for selecting the distinguishable genes between the control and disease samples, all these genes were not biologically informative (discussed later). With the GDS3610 and GSE106291 datasets, *MGS_f* and *MGS_{rf}* methods achieved better performances compared to the other methods except one (fDNN using RF classier with GSE106291 dataset).

Table 3. Classification accuracy and AUROC of different methods for balanced datasets.

Methods	Dataset: GDS6063				Dataset: GDS5306				Dataset: GDS4431			
	Accuracy		*AUROC*		*Accuracy*		*AUROC*		*Accuracy*		*AUROC*	
	SVM	RF	SVM	RF	SVM	RF	SVM	RF	SVM	RF	SVM	RF
RF	0.700	0.700	0.898	0.766	0.447	0.500	0.583	0.710	0.561	0.625	0.632	0.719
fDNN	0.900	0.800	0.900	0.932	**0.690**	**0.888**	0.814	0.926	0.697	0.742	0.750	0.824
IGIS+	0.600	0.400	0.460	0.480	0.605	0.868	0.610	0.890	0.616	0.705	0.560	0.780
HDG	0.900	0.891	0.953	0.972	0.685	0.763	0.798	0.916	0.725	0.740	0.787	0.788
$mDSM_f$	0.900	0.900	0.920	0.940	0.650	0.775	0.758	0.866	0.705	0.767	0.830	0.839
$mDSM_{rf}$	0.900	0.900	0.920	0.940	0.650	0.775	0.758	0.866	0.705	**0.767**	**0.830**	0.839
MGS_f	0.900	0.900	0.960	0.990	0.658	0.868	0.820	0.940	0.733	0.753	0.820	0.850
MGS_{rf}	**0.900**	**0.900**	**0.960**	**0.990**	0.658	0.868	**0.812**	**0.940**	**0.733**	0.753	0.820	**0.850**

https://doi.org/10.1371/journal.pone.0230164.t003

Table 4. Classification accuracy and AUROC of different methods for imbalanced datasets.

Methods	Dataset: GDS3341				Dataset: GDS4824				Dataset: GDS3610				Dataset: GSE106291			
	Accuracy		*AUROC*		*Accuracy*		*AUROC*		*Accuracy*		*AUROC*		*Accuracy*		*AUROC*	
	SVM	**RF**	**SVM**	**RF**	**SVM**	**RF**	**SVM**	**RF**	**SVM**	**RF**	**SVM**	**RF**	**SVM**	**RF**	**SVM**	**RF**
RF	0.878	0.878	0.955	0.940	0.476	0.476	0.289	0.389	0.679	0.893	0.253	0.507	0.698	0.702	0.277	0.622
fDNN	1.00	1.00	1.00	1.00	0.952	1.00	1.00	1.00	0.750	0.893	0.560	0.827	0.766	**0.779**	0.778	0.783
IGIS+	0.976	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.893	0.893	0.853	0.940	0.732	0.762	0.695	0.765
HDG	1.00	1.00	1.00	1.00	0.952	1.00	0.960	1.00	0.964	0.964	1.00	0.980	0.698	0.690	0.600	0.567
$mDSM_f$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.964	0.929	1.00	0.980	0.728	0.694	0.638	0.629
$mDSM_{rf}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.964	0.929	1.00	0.980	0.698	0.689	0.400	0.5419
MGS_f	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.964	0.929	0.960	0.973	0.757	0.762	0.764	0.793
MGS_{rf}	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	**1.00**	**0.964**	**1.00**	**0.987**	**0.770**	0.757	**0.787**	**0.796**

https://doi.org/10.1371/journal.pone.0230164.t004

Besides the superiority of *MGS* in terms of accuracy, it also performed significantly better in many cases in comparison to most of the methods.

It is observed from the Tables 3 and 4 that *mDSM*, *fDNN* and *HDG* performed reasonably and *MGS* outperformed most of its competitive methods. Even though MGS_f performed better in most of the cases compared to the existing methods, MGS_{rf} performs slightly better than MGS_f. It is because of the selection of informative genes based on the RF classifier. As the number of samples is relatively small with respect to the number of genes, MGS_f could easily overfit training samples and thus performed poorly for unseen data. MGS_{rf} solved the overfitting problem by using *RF* classifier [40].

Comparison of performances for different number of genes

We also investigated the performances of the aforementioned methods for a different number of selected genes (η) using two metrics *accuracy* and *AUROC* as shown in Figs 2–7. Except *RF*, all the methods performed well (Figs 2–7).

In the case of balanced dataset (Figs 2 and 3), with GDS5306 and GDS4431 dataset, MGS_f and MGS_{rf} outperformed other methods which indicate that the proposed gene selection methods were able to select those genes which give additional information about the disease. For the small and highly imbalanced dataset GDS3610, our methods showed superior performances with different number of genes (Fig 5). With GDS3341 and GDS4824 datasets, all the gene selection methods classified the samples for different number of genes almost perfectly as shown in Figs 4 and 6. For these two datasets, the expression values of genes are more distinguishable between classes which would be the reason for the almost equal performance of every method. This might be the reason why the performance did not vary with an increase in the number of selected genes. We have also shown the strength of our methods with the GSE106291 dataset, which has a comparatively large number of samples (Fig 7).

Based on the results presented in Figs 2–7 and Tables 3 and 4, it is evident that the performances of MGS_f and MGS_{rf} are clearly better than the existing methods for balanced and imbalanced datasets. *MGS* performed well for all classifiers and thus, it is classifier independent. The datasets used for experimentation had a highly imbalanced distribution of the classes. This indicates that *MGS* is tolerant to imbalanced datasets. However, MGS_{rf} achieved slightly better performance for every value of η, indicating MGS_{rf} could classify more samples accurately than MGS_f.

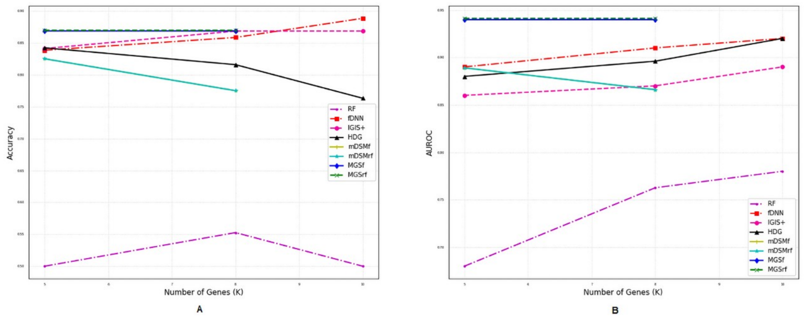

Fig 2. Performance comparison using different number of selected genes for the GDS5306 dataset. (A) Accuracy. (B) AUROC.

https://doi.org/10.1371/journal.pone.0230164.g002

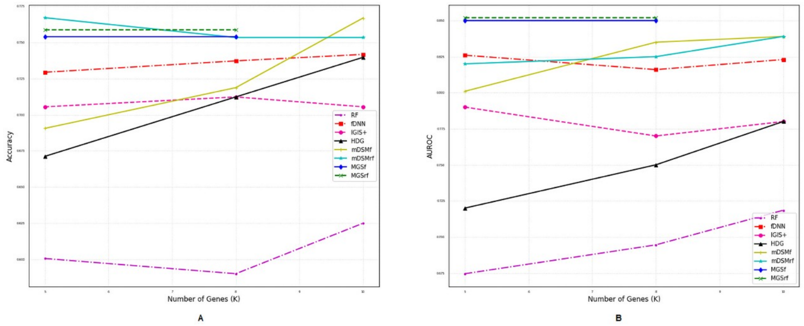

Fig 3. Performance comparison using different number of selected genes for the GDS4431 dataset. (A) Accuracy. (B) AUROC.

https://doi.org/10.1371/journal.pone.0230164.g003

Biological interpretation

It is not always a requisite that the selected genes with better classification ability are also relevant to a particular biological process. Therefore, to assess the performances of MGS_f and MGS_{rf}, we investigated the ability of the top (≤ 10) selected genes to identify the most relevant

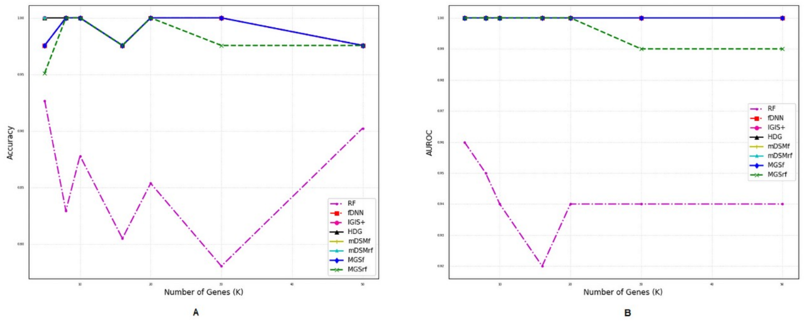

Fig 4. Performance comparison using different number of selected genes for the GDS3341 dataset. (A) Accuracy. (B) AUROC.

https://doi.org/10.1371/journal.pone.0230164.g004

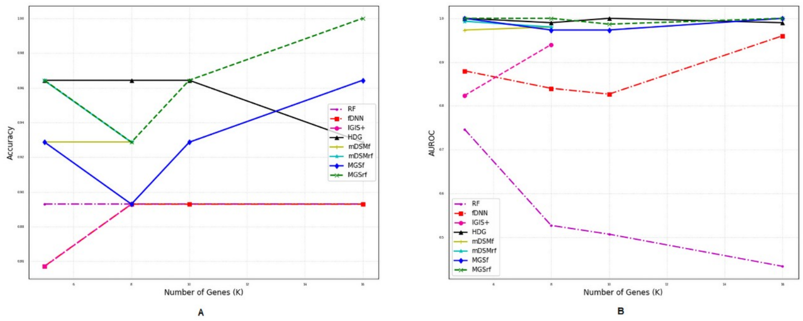

Fig 5. Performance comparison using different number of selected genes for the GDS3610 dataset. (A) Accuracy. (B) AUROC.

https://doi.org/10.1371/journal.pone.0230164.g005

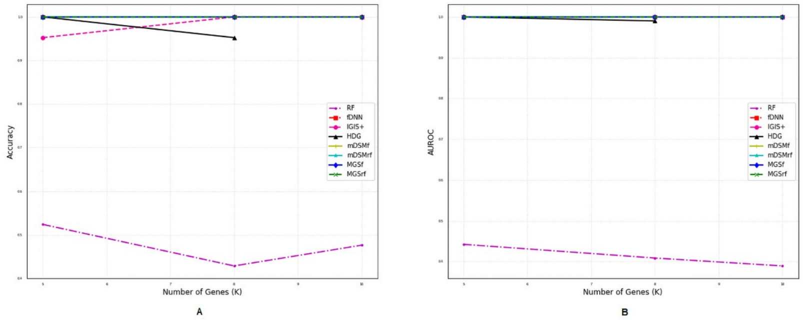

Fig 6. Performance comparison using different number of selected genes for the GDS4824 dataset. (A) Accuracy. (B) AUROC.

https://doi.org/10.1371/journal.pone.0230164.g006

pathways in the cancer types used in different balanced and imbalanced datasets (Tables 5 and 6).

For balanced dataset, as this study primarily focused on disease data classification and the identification of relevant genes, we investigated the performances of all methods for capturing biological significance on various disease datasets derived from varied biological sources, such as cancer metastasis (GDS5306), autism (GDS4431) and viral infection (GDS6063). The results

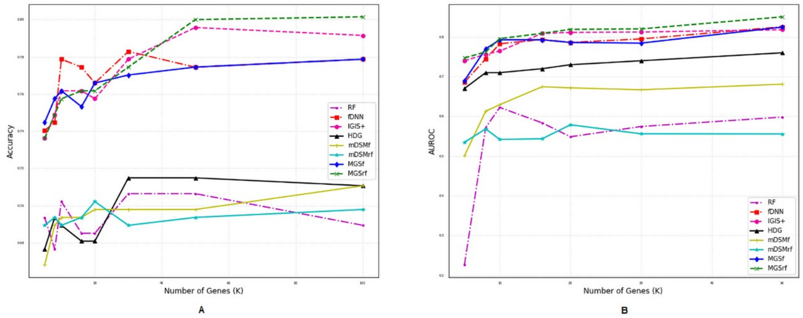

Fig 7. Performance comparison using different number of selected genes for the GSE106291 dataset. (A) Accuracy. (B) AUROC.

https://doi.org/10.1371/journal.pone.0230164.g007

are shown in Table 5. GDS6063 dataset incorporates gene expression profiles of primary plasmacytoid dendritic cells following exposure to influenza A for 8 hours [50]. Human dendritic cells (DCs) are susceptible to infection with various viruses, including human T-lymphotropic virus Type 1 (HTLV-1), human immunodeficiency virus type 1 (HIV-1), measles virus and influenza virus [51]. In fact, the HTLV-1 acts more like the influenza virus in terms of infection to the DC cells and differs significantly with measles virus or HIV-1 [51]. The influenza virus-infected DCs induce a considerably higher proliferative response [51]. Transforming growth factor-beta (TGF-β) is a multifunctional cytokine and its activity increases during influenza virus [52, 53]. Along with *MGS*, $mDSM_f$ and $mDSM_{rf}$ could identify these as the top pathways (Table 5). Besides, *HDG* identified two pathways but other three methods could not identify any of the pathways. TGF-β signaling also plays an important role by stimulating cell invasion during metastasis of breast cancer [54, 55]. GDS5306 dataset contains gene expression data of HER2+ breast cancer brain metastasis specimens and HER2+ nonmetastatic primary breast tumors [50]. As shown in Table 5, MGS_f and MGS_{rf} identified pathways relevant to cancer and metastasis. Compared to *MGS*, *mDSM* and *HDG* performed reasonably well. GDS4431 dataset includes gene expression data of peripheral blood lymphocytes from autistic and non-autistic children [50]. Interestingly, the pathways indentified by majority of the methods overrepresented the pathways associated with cancer. It was recently reported that autism and cancer share risk genes [56]. Mutations in genes encoding the ubiquitin proteasome system (UPS) are associated with an increased risk for the development of autism spectrum disorders [57–59]. Viral carcinogenesis and ubiquitin mediated proteolysis were identified as the top pathways affected in autistic children. Connection between autism and Hepatitis B infection (one of the other top-ranked pathways) is not obvious based on the available information and may be explored in further studies.

For imbalanced datasets, it is evident that MGS_f and MGS_{rf} performed better in capturing the genes more relevant to the cancer type (Table 6). For example, Epstein-Barr virus (EBV) is well known to cause nasopharyngeal carcinoma (NPC), which is a type of epithelial cancer prevalent in Southeast Asia [60–62]. GDS3341 and GDS3610 datasets contain NPC samples [32, 33]. Although GDS3341 and GDS3610 are independent datasets, both MGS_{rf} and MGS_f could detect the genes involved in viral carcinogenesis and Epstein-Barr virus infection (Table 6). We used two different datasets (GDS3341 and GDS3610) on the same cancer type as built-in controls in the study to increase confidence with the experimental results. With both the datasets, MGS_{rf} and MGS_f performed almost equally well, although the genes selected by MGS_{rf} performed somewhat better. The other methods (*RF*, *fDNN*, *IGIS+*, $mDSM_f$ and $mDSM_{rf}$) could detect these pathways only for the GDS3610 dataset whereas *HDG* could detect pathways for both GDS3341 and GDS3610 datasets. In fact, *RF* and *IGIS+* could detect one of these pathways. The GDS4824 dataset contains gene expression data from prostate cancer samples. Both the MGS_{rf} and MGS_f detected genes that are involved in prostate cancer. Although the prostate cancer pathway was ranked 6th in the detected pathways (based on the FDR values) with the genes selected by the MGS_{rf} and MGS_f, the top ranked pathways (FoxO signaling pathway, colorectal cancer, pancreatic cancer and endometrial cancer) are relevant to cancer as well [63–66]. In fact, unlike nasopharyngeal carcinoma, prostate cancer development involves different pathways. Fork head box O transcription factors (FoxO) regulates multiple cellular processes, including cell cycle arrest, cell death, DNA damage repair, stress resistance, and metabolism [67]. Inactivation of FoxO protein is linked to multiple tumorigenesis including prostate cancer [67–69]. Among the other methods, *fDNN*, *HDG* and *mDSM* could detect the genes associated with prostate cancer, although the rank of the pathway and associated FDR values were less significant.

Table 5. Comparative performance of different methods in identification of relevant biological pathways for balanced datasets.

Dataset ID	Cancer type	Method	No. of genes	Pathway	Output rank	FDR
GDS6063	Influenza A infected plasmacytoid dendritic cells(pDC)	*RF*	10	Cell cycle/HTLV-I infec./TGF-beta sig. path.	ND	-
		fDNN				
		IGIS+	1			
		HDG	8	Cell cycle	23	6.00E-01
				HTLV-I infection	16	2.39E-01
		mDSM$_{f}$	2	Cell cycle	1	1.41E-06
				HTLV-I infection	2	7.63E-06
				TGF-beta signaling pathway	3	2.18E-05
		mDSM$_{rf}$	2	Cell cycle	1	1.41E-06
				HTLV-I infection	2	7.63E-06
				TGF-beta signaling pathway	3	2.18E-05
		MGS$_{f}$	2	Cell cycle	**1**	**1.41E-06**
				HTLV-I infection	**2**	**7.63E-06**
				TGF-beta signaling pathway	**3**	**2.18E-05**
		MGS$_{rf}$	2	Cell cycle	**1**	**1.41E-06**
				HTLV-I infection	**2**	**7.63E-06**
				TGF-beta signaling pathway	**3**	**2.18E-05**
GDS5306	Breast cancer brain metastasis specimens and nonmetastatic primary breast tumors	*RF*	10	Cell cycle/Path. in cancer/TGF-beta sig. path	ND	-
		fDNN	10	Cell cycle	ND	-
				Pathways in cancer	6	1.96E-08
				TGF-beta signaling pathway	ND	-
		IGIS+	10	Cell cycle	ND	-
				Pathways in cancer	1	9.97E-08
				TGF-beta signaling pathway	ND	-
		HDG	4	Cell cycle	10	0.533
				Pathways in cancer	9	0.526
				TGF-beta signaling pathway	4	0.0525
		mDSM$_{f}$	2	Cell cycle	2	3.64E-10
				Pathways in cancer	1	3.64E-10
				TGF-beta signaling pathway	3	1.71E-07
		mDSM$_{rf}$	2	Cell cycle	2	3.64E-10
				Pathways in cancer	1	3.64E-10
				TGF-beta signaling pathway	3	1.71E-07
		MGS$_{f}$	2	Cell cycle	**1**	**1.10E-12**
				Pathways in cancer	**2**	**3.92E-11**
				TGF-beta signaling pathway	**4**	**5.15E-08**
		MGS$_{rf}$	2	Cell cycle	**1**	**1.10E-12**
				Pathways in cancer	**2**	**3.92E-11**
				TGF-beta signaling pathway	**4**	**5.15E-08**

(*Continued*)

Table 5. (Continued)

Dataset ID	Cancer type	Method	No. of genes	Pathway	Output rank	FDR
GDS4431	Peripheral blood lymphocytes of autistic and non-autistic children	*RF*	10	Viral carcinogenesis	ND	-
				Hepatitis B	ND	-
				Ubiquitin mediated proteolysis	1	1.13E-04
		fDNN	10	Viral carcinogenesis	5	4.72E-04
				Hepatitis B	ND	-
				Ubiquitin mediated proteolysis	ND	-
		IGIS+	10	Viral carcinogenesis	2	6.72E-07
				Hepatitis B	ND	-
				Ubiquitin mediated proteolysis	ND	-
		HDG	10	Viral carcinogenesis	ND	-
				Hepatitis B	ND	-
				Ubiquitin mediated proteolysis	2	1
		$mDSM_f$	10	Viral cycle/Hep. B/Ub. mediat. prote.	ND	-
		$mDSM_{rf}$				
		MGS_f	4	Viral carcinogenesis	**1**	**4.55E-03**
				Hepatitis B	**2**	**5.12E-03**
				Ubiquitin mediated proteolysis	**3**	**1.19E-02**
		MGS_{rf}	4	Viral carcinogenesis	**1**	**4.55E-03**
				Hepatitis B	**2**	**5.12E-03**
				Ubiquitin mediated proteolysis	**3**	**1.19E-02**

ND—Not detected

FDR—False discovery rate

https://doi.org/10.1371/journal.pone.0230164.t005

Although multiple proteins interact in a network inside a cell to attain a particular function, each of these does not play equally important role. Some proteins in a network are more connected and play a pivotal role in the overall biological process. MGS_{rf} and MGS_f selected top genes play important roles in pathways relevant to cancer (Fig 8).

It is noteworthy to mention that the proposed methods (MGS_f and MGS_{rf}) performed better compared to the *mDSM* ($mDSM_f$ and $mDSM_{rf}$) despite sharing a closely similar methodology. These methods differed in the exclusion of redundancy term. *mDSM* discards a gene if it finds another gene with similar expression level. But as mentioned earlier, both genes may be informative despite redundancy and may provide useful information. Avoidance of the redundant genes may not be appropriate as genes working together in a pathway may be regulated in a more coordinated fashion than a random set of genes, and thus share a more coherent expression profile [70]. To understand this issue, let us consider an example of two genes named *MAN*1*C*1 and *ARCN*1 in dataset GDS3610. mDSM discarded *ARCN*1 gene since the redundancy value (0.685461) with *MAN*1*C*1 is greater than χ^2 critical value (0.558168). Both of these genes work in pathways that inhibit cancer cell proliferation [71, 72]. Therefore, we did not consider redundancy in Eq 2 to select genes with MGS_f and MGS_{rf}. Our proposed methods selected both *MAN*1*C*1 and *ARCN*1 as these provide additional information (0.598510).

Conclusion

Here, we present a gene selection method followed by two gene ranking methods for the selection of informative genes from high dimensional low sample size gene expression data. The

Table 6. Comparative performance of different methods in identification of relevant biological pathways for imbalanced datasets.

Dataset ID	Cancer type	Method	No. of genes	Pathway	Output rank	FDR
GDS3341	Nasopharyngeal carcinoma	*RF*	10	Viral carcinogenesis	ND	-
				Epstein-Barr virus infection	ND	-
		fDNN	10	Viral carcinogenesis	ND	-
				Epstein-Barr virus infection	ND	-
		IGIS+	3	Viral carcinogenesis	ND	-
				Epstein-Barr virus infection	ND	-
		HDG	10	Viral carcinogenesis	1	0.000121
				Epstein-Barr virus infection	ND	-
		mDSM_{f}	4	Viral carcinogenesis	ND	-
				Epstein-Barr virus infection	ND	-
		mDSM_{rf}	4	Viral carcinogenesis	ND	-
				Epstein-Barr virus infection	ND	-
		MGS_{f}	10	Viral carcinogenesis	4	0.00259
				Epstein-Barr virus infection	14	0.166
		MGS_{rf}	10	Viral carcinogenesis	**1**	**1.38E-14**
				Epstein-Barr virus infection	**4**	**2.56E-07**
GDS3610	Nasopharyngeal carcinoma	*RF*	10	Viral carcinogenesis	ND	-
				Epstein-Barr virus infection	29	0.53
		fDNN	10	Viral carcinogenesis	79	7.97E-08
				Epstein-Barr virus infection	113	6.85E-05
		IGIS+	7	Viral carcinogenesis	6	0.338
				Epstein-Barr virus infection	ND	-
		HDG	10	Viral carcinogenesis	10	3.82E-08
				Epstein-Barr virus infection	12	0.00000155
		mDSM_{f}	9	Viral carcinogenesis	1	4.83E-13
				Epstein-Barr virus infection	5	0.0165
		mDSM_{rf}	9	Viral carcinogenesis	1	4.83E-13
				Epstein-Barr virus infection	5	0.0165
		MGS_{f}	10	Viral carcinogenesis	**1**	**4.83E-13**
				Epstein-Barr virus infection	**5**	**0.000259**
		MGS_{rf}	10	Viral carcinogenesis	**1**	**4.83E-13**
				Epstein-Barr virus infection	**5**	**0.0165**
GDS4824	Prostate cancer	*RF*	10	ND	ND	-
		fDNN	10	Prostate cancer	28	0.435
		IGIS+	10	ND	ND	-
		HDG	8	Prostate cancer	7	0.632
		mDSM_{f}	6	Prostate cancer	7	1.25E-16
		mDSM_{rf}	6	Prostate cancer	7	1.25E-16
		MGS_{f}	10	Prostate cancer	**6**	**1.29E-22**
		MGS_{rf}	10	Prostate cancer	**6**	**1.29E-22**

(*Continued*)

Table 6. (Continued)

Dataset ID	Cancer type	Method	No. of genes	Pathway	Output rank	FDR
GSE106291	Acute myeloid leukemia	*RF*	10	Chronic myeloid leukemia	ND	-
				Acute myeloid leukemia	ND	-
		fDNN	10	Chronic myeloid leukemia	ND	-
				Acute myeloid leukemia	ND	-
		IGIS+	10	Chronic myeloid leukemia	22	0.000319
				Acute myeloid leukemia	ND	-
		HDG	10	Chronic myeloid leukemia	ND	-
				Acute myeloid leukemia	ND	-
		$mDSM_f$	10	Chronic myeloid leukemia	ND	-
				Acute myeloid leukemia	ND	-
		$mDSM_{rf}$	10	Chronic myeloid leukemia	26	0.448
				Acute myeloid leukemia	ND	-
		MGS_f	10	Chronic myeloid leukemia	ND	-
				Acute myeloid leukemia	ND	-
		MGS_{rf}	10	Chronic myeloid leukemia	**1**	**2.78E-12**
				Acute myeloid leukemia	**8**	**8.74E-08**

ND—Not detected

FDR—False discovery rate

https://doi.org/10.1371/journal.pone.0230164.t006

proposed gene selection method utilizes the maximum relevance and complementary information for selecting informative genes that have biological importance. Experimental results with known disease datasets illustrate that the proposed methods consistently achieve higher classification accuracy and select more biologically relevant genes than the previously reported

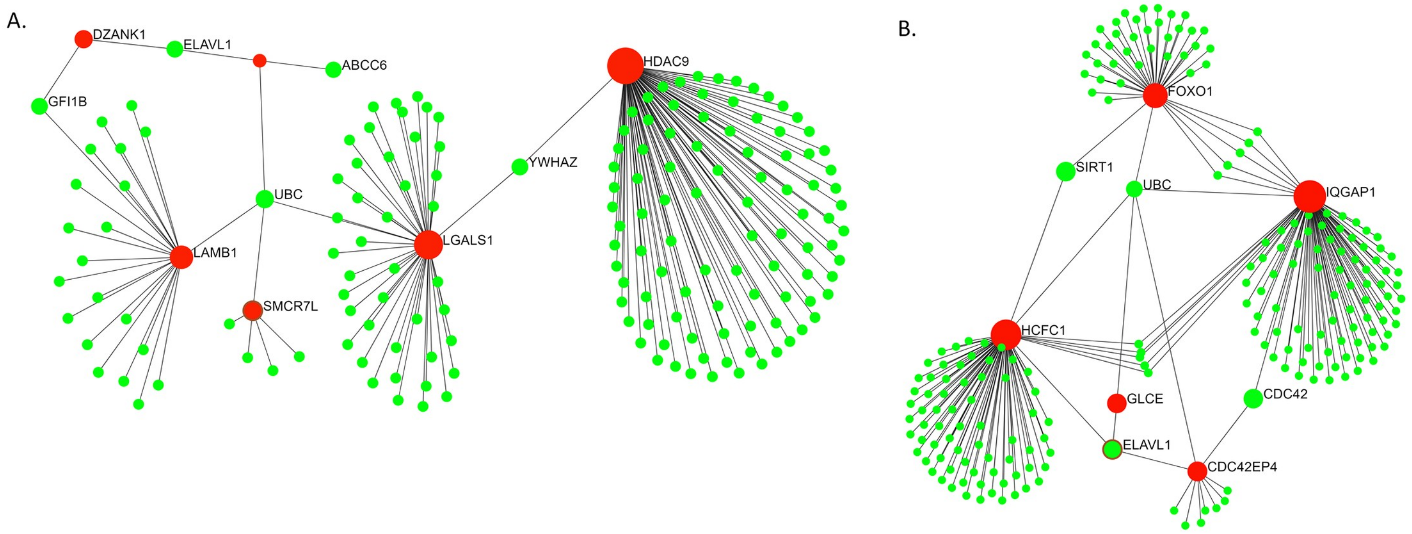

Fig 8. Roles of MGS_{rf} selected top genes in pathways related to cancer. (A) *LGALS*1 and *LAMB*1 were selected among the top 10 genes from GDS3341 dataset by the MGS_{rf}. These (highlighted in red) are part of a sub-network that contains many other proteins (highlighted in green) known to play roles in different cancers [44]. (B) *HCFC*1, *FOXO*1 and *IQGAP*1 were selected among the top 10 genes from GDS4824 dataset by the MGS_{rf}. These (highlighted in red) are part of a sub-network that contains many other proteins (highlighted in green) known to play roles in different cancers [44].

https://doi.org/10.1371/journal.pone.0230164.g008

methods. Moreover, we anticipate that the proposed method will also identify genes responsible for an unknown disease because it identifies effective and responsible genes for known diseases. However, there are a few challenges that need to be addressed in further studies. First, we believe that introducing a higher-order gene interaction may help to reduce the number of selected genes but it may increase the computational complexity. Second, a semi-definite programming based search strategy may help to obtain globally optimum gene subsets.

Supporting information

S1 File.
(ZIP)

Author Contributions

Conceptualization: Md Nazmul Haque.

Data curation: Md Nazmul Haque.

Formal analysis: Md Nazmul Haque, Sadia Sharmin, Amin Ahsan Ali, Abu Ashfaqur Sajib, Mohammad Shoyaib.

Investigation: Md Nazmul Haque, Sadia Sharmin, Amin Ahsan Ali, Abu Ashfaqur Sajib, Mohammad Shoyaib.

Methodology: Md Nazmul Haque, Sadia Sharmin, Amin Ahsan Ali, Abu Ashfaqur Sajib, Mohammad Shoyaib.

Software: Md Nazmul Haque, Sadia Sharmin.

Supervision: Amin Ahsan Ali, Abu Ashfaqur Sajib, Mohammad Shoyaib.

Validation: Md Nazmul Haque, Sadia Sharmin, Amin Ahsan Ali, Abu Ashfaqur Sajib, Mohammad Shoyaib.

Writing – original draft: Md Nazmul Haque, Sadia Sharmin, Amin Ahsan Ali, Mohammad Shoyaib.

Writing – review & editing: Md Nazmul Haque, Sadia Sharmin, Amin Ahsan Ali, Abu Ashfaqur Sajib, Mohammad Shoyaib.

References

1. Narendra PM, Fukunaga K. A branch and bound algorithm for feature subset selection. IEEE Transactions on computers. 1977;(9):917–922. https://doi.org/10.1109/TC.1977.1674939
2. Li Z, Xie W, Liu T. Efficient feature selection and classification for microarray data. PloS one. 2018; 13 (8):e0202167. https://doi.org/10.1371/journal.pone.0202167 PMID: 30125332
3. Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. Journal of bioinformatics and computational biology. 2005; 3(02):185–205. https://doi.org/10.1142/S0219720005001004 PMID: 15852500
4. Kong Y, Yu T. A deep neural network model using random forest to extract feature representation for gene expression data classification. Scientific reports. 2018; 8(1):16477. https://doi.org/10.1038/s41598-018-34833-6 PMID: 30405137
5. 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
6. Hall MA. Correlation-based feature selection for machine learning. 1999.
7. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proceedings of the National Academy of Sciences. 2002; 99(10):6567–6572. https://doi.org/10.1073/pnas.082099299 PMID: 12011421

8. Akhter, Suravi and Sharmin, Sadia and Ahmed, Sumon and Sajib, Abu Ashfaqur and Shoyaib, Mohammad. mRelief: A Reward Penalty Based Feature Subset Selection Considering Data Overlapping Problem. International Conference on Computational Science. 2021;278–292.

9. Urbanowicz Ryan J and Olson Randal S and Schmitt Peter and Meeker Melissa and Moore Jason H. Benchmarking relief-based feature selection methods for bioinformatics data mining. Journal of biomedical informatics. 2018; 85:168–188. https://doi.org/10.1016/j.jbi.2018.07.015 PMID: 30030120

10. Sharmin S, Ali AA, Khan MAH, Shoyaib M. Feature selection and discretization based on mutual information. In: 2017 IEEE icIVPR. IEEE; 2017. p. 1–6.

11. Sharmin S, Shoyaib M, Ali AA, Khan MAH, Chae O. Simultaneous feature selection and discretization based on mutual information. Pattern Recognition. 2019; 91:162–174. https://doi.org/10.1016/j.patcog.2019.02.016

12. Ross BC. Mutual information between discrete and continuous data sets. PloS one. 2014; 9(2). https://doi.org/10.1371/journal.pone.0087357 PMID: 24586270

13. Roy Puloma and Sharmin Sadia and Ali Amin Ahsan and Shoyaib Mohammad. Discretization and feature selection based on bias corrected mutual information considering high-order dependencies. Advances in Knowledge Discovery and Data Mining. 2020; 12084:830. https://doi.org/10.1007/978-3-030-47426-3_64

14. Vinh NX, Zhou S, Chan J, Bailey J. Can high-order dependencies improve mutual information based feature selection? Pattern Recognition. 2016; 53:46–58. https://doi.org/10.1016/j.patcog.2015.11.007

15. 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 (TCBB). 2012; 9(4):1106–1119. https://doi.org/10.1109/TCBB.2012.33 PMID: 22350210

16. Mundra PA, Rajapakse JC. SVM-RFE with MRMR filter for gene selection. IEEE transactions on nanobioscience. 2009; 9(1):31–37. https://doi.org/10.1109/TNB.2009.2035284 PMID: 19884101

17. Yoon S, Kim S. Mutual information-based SVM-RFE for diagnostic classification of digitized mammograms. Pattern Recognition Letters. 2009; 30(16):1489–1495. https://doi.org/10.1016/j.patrec.2009.06.012

18. Karegowda AG, Jayaram M, Manjunath A. Feature subset selection problem using wrapper approach in supervised learning. International journal of Computer applications. 2010; 1(7):13–17. https://doi.org/10.5120/169-295

19. Nakariyakul S. High-dimensional hybrid feature selection using interaction information-guided search. Knowledge-Based Systems. 2018; 145:59–66. https://doi.org/10.1016/j.knosys.2018.01.002

20. Nakariyakul S. A hybrid gene selection algorithm based on interaction information for microarray-based cancer classification. PloS one. 2019; 14(2). https://doi.org/10.1371/journal.pone.0212333 PMID: 30768654

21. Almugren N, Alshamlan H. A survey on hybrid feature selection methods in microarray gene expression data for cancer classification. IEEE Access. 2019; 7:78533–78548. https://doi.org/10.1109/ACCESS.2019.2922987

22. Alshamlan H, Badr G, Alohali Y. mRMR-ABC: a hybrid gene selection algorithm for cancer classification using microarray gene expression profiling. Biomed research international. 2015; 2015. https://doi.org/10.1155/2015/604910 PMID: 25961028

23. Alshamlan HM, Badr GH, Alohali YA. Genetic Bee Colony (GBC) algorithm: A new gene selection method for microarray cancer classification. Computational biology and chemistry. 2015; 56:49–60. https://doi.org/10.1016/j.compbiolchem.2015.03.001 PMID: 25880524

24. El Akadi A, Amine A, El Ouardighi A, Aboutajdine D. A two-stage gene selection scheme utilizing MRMR filter and GA wrapper. Knowledge and Information Systems. 2011; 26(3):487–500. https://doi.org/10.1007/s10115-010-0288-x

25. 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): e0187371. https://doi.org/10.1371/journal.pone.0187371 PMID: 29095904

26. Hameed SS, Hassan R, Hassan WH, Muhammadsharif FF, Latiff LA. HDG-select: A novel GUI based application for gene selection and classification in high dimensional datasets. PloS one. 2021; 16(1): e0246039. https://doi.org/10.1371/journal.pone.0246039 PMID: 33507983

27. Naghibi T, Hoffmann S, Pfister B. A semidefinite programming based search strategy for feature selection with mutual information measure. IEEE Trans Pattern Anal Mach Intell. 2014; 37(8):1529–1541. https://doi.org/10.1109/TPAMI.2014.2372791

28. Araújo T, Aresta G, Castro E, Rouco J, Aguiar P, Eloy C, et al. Classification of breast cancer histology images using convolutional neural networks. PloS one. 2017; 12(6). https://doi.org/10.1371/journal.pone.0177544 PMID: 28570557

29. Islam, SM Sofiqul and Rahman, Shanto and Rahman, Md Mostafijur and Dey, Emon Kumar and Shoyaib, Mohammad. Application of deep learning to computer vision: A comprehensive study. 2016 5th international conference on informatics, electronics and vision (ICIEV). 2016;592–597.

30. Haque, Md Nazmul and Mahbub, Mahir and Tarek, Md Hasan and Lota, Lutfun Nahar and Ali, Amin Ahsan Nurse Care Activity Recognition: A GRU-based approach with attention mechanism Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers

31. Noda K, Yamaguchi Y, Nakadai K, Okuno HG, Ogata T. Audio-visual speech recognition using deep learning. Applied Intelligence. 2015; 42(4):722–737. https://doi.org/10.1007/s10489-014-0629-7

32. Dodd LE, Sengupta, et al. Genes involved in DNA repair and nitrosamine metabolism and those located on chromosome 14q32 are dysregulated in nasopharyngeal carcinoma. Cancer Epidemiology and Prevention Biomarkers. 2006; 15(11):2216–2225. https://doi.org/10.1158/1055-9965.EPI-06-0455 PMID: 17119049

33. Bose S, Yap, et al. The ATM tumour suppressor gene is down-regulated in EBV-associated nasopharyngeal carcinoma. The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland. 2009; 217(3):345–352. https://doi.org/10.1002/path.2487 PMID: 19142888

34. Arredouani MS, et al. Identification of the transcription factor single-minded homologue 2 as a potential biomarker and immunotherapy target in prostate cancer. Clinical cancer research. 2009; 15(18):5794–5802. https://doi.org/10.1158/1078-0432.CCR-09-0911 PMID: 19737960

35. Herold T, Jurinovic V, Batcha AM, Bamopoulos SA, Rothenberg-Thurley M, Ksienzyk B, et al. A 29-gene and cytogenetic score for the prediction of resistance to induction treatment in acute myeloid leukemia. haematologica. 2018; 103(3):456–465. https://doi.org/10.3324/haematol.2017.178442 PMID: 29242298

36. 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

37. McMullin RP, Wittner BS, Yang C, Denton-Schneider BR, Hicks D, Singavarapu R, et al. A BRCA1 deficient-like signature is enriched in breast cancer brain metastases and predicts DNA damage-induced poly (ADP-ribose) polymerase inhibitor sensitivity. Breast Cancer Research. 2014; 16(2):1–10. https://doi.org/10.1186/bcr3625 PMID: 24625110

38. Bajwa G, DeBerardinis RJ, Shao B, Hall B, Farrar JD, Gill MA. Cutting edge: Critical role of glycolysis in human plasmacytoid dendritic cell antiviral responses. The Journal of Immunology. 2016; 196(5):2004–2009. https://doi.org/10.4049/jimmunol.1501557 PMID: 26826244

39. Barrett T, Wilhite, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic acids research. 2012; 41(D1):D991–D995. https://doi.org/10.1093/nar/gks1193 PMID: 23193258

40. Breiman L. Random forests. Machine learning. 2001; 45(1):5–32. https://doi.org/10.1023/A:1010933404324

41. Pedregosa F, Varoquaux, et al. Scikit-learn: Machine learning in Python. Journal of machine learning research. 2011; 12(Oct):2825–2830.

42. Chen, Tianqi and Guestrin, Carlos. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016;785–794.

43. Kadir Md Eusha and Akash Pritom Saha and Sharmin Sadia and Ali Amin Ahsan and Shoyaib Mohammad. A proximity weighted evidential k nearest neighbor classifier for imbalanced data. Journal of biomedical informatics. 2020; 12085:71.

44. Zhou G, Soufan O, Ewald J, Hancock RE, Basu N, Xia J. NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. Nucleic acids research. 2019;. https://doi.org/10.1093/nar/gkz240 PMID: 30931480

45. Liang Q, Dharmat R, Owen L, Shakoor A, Li Y, Kim S, et al. Single-nuclei RNA-seq on human retinal tissue provides improved transcriptome profiling. Nature communications. 2019; 10(1):1–12. https://doi.org/10.1038/s41467-019-12917-9 PMID: 31848347

46. Bayrak CS, Zhang P, Tristani-Firouzi M, Gelb BD, Itan Y. De novo variants in exomes of congenital heart disease patients identify risk genes and pathways. Genome medicine. 2020; 12(1):1–18.

47. Ouyang Y, Yin J, Wang W, Shi H, Shi Y, Xu B, et al. Downregulated Gene Expression Spectrum and Immune Responses Changed During the Disease Progression in Patients With COVID-19. Clinical Infectious Diseases. 2020; 71(16):2052–2060. https://doi.org/10.1093/cid/ciaa462 PMID: 32307550

48. Pons P, Latapy M. Computing communities in large networks using random walks. In: International symposium on computer and information sciences. Springer; 2005. p. 284–293.

49. Newman ME, Girvan M. Finding and evaluating community structure in networks. Physical review E. 2004; 69(2):026113. https://doi.org/10.1103/PhysRevE.69.026113 PMID: 14995526

50. Camarata PJ, McGeachie RE, Haines SJ. Dorsal midbrain encephalitis caused by Propionibacterium acnes: Report of two cases. Journal of neurosurgery. 1990; 72(4):654–659. https://doi.org/10.3171/jns.1990.72.4.0654 PMID: 2319325

51. Makino M, Shimokubo S, Wakamatsu SI, Izumo S, Baba M. The role of human T-lymphotropic virus type 1 (HTLV-1)-infected dendritic cells in the development of HTLV-1-associated myelopathy/tropical spastic paraparesis. Journal of virology. 1999; 73(6):4575–4581. https://doi.org/10.1128/JVI.73.6.4575-4581.1999 PMID: 10233916

52. Carlson CM, Turpin EA, Moser LA, O'Brien KB, Cline TD, Jones JC, et al. Transforming growth factor-β: activation by neuraminidase and role in highly pathogenic H5N1 influenza pathogenesis. PLoS Pathog. 2010; 6(10):e1001136. https://doi.org/10.1371/journal.ppat.1001136 PMID: 20949074

53. Denney L, Branchett W, Gregory LG, Oliver RA, Lloyd CM. Epithelial-derived TGF-β1 acts as a pro-viral factor in the lung during influenza A infection. Mucosal immunology. 2018; 11(2):523–535. https://doi.org/10.1038/mi.2017.77 PMID: 29067998

54. Imamura T, Hikita A, Inoue Y. The roles of TGF-β signaling in carcinogenesis and breast cancer metastasis. Breast cancer. 2012; 19(2):118–124. https://doi.org/10.1007/s12282-011-0321-2 PMID: 22139728

55. Drabsch Y, Ten Dijke P. TGF-β signaling in breast cancer cell invasion and bone metastasis. Journal of mammary gland biology and neoplasia. 2011; 16(2):97–108. https://doi.org/10.1007/s10911-011-9217-1 PMID: 21494783

56. Crawley JN, Heyer WD, LaSalle JM. Autism and cancer share risk genes, pathways, and drug targets. Trends in Genetics. 2016; 32(3):139–146. https://doi.org/10.1016/j.tig.2016.01.001 PMID: 26830258

57. Nakashima M, Kato M, Matsukura M, Kira R, Ngu LH, Lichtenbelt KD, et al. De novo variants in CUL3 are associated with global developmental delays with or without infantile spasms. Journal of human genetics. 2020; p. 1–8.

58. Louros SR, Osterweil EK. Perturbed proteostasis in autism spectrum disorders. Journal of neurochemistry. 2016; 139(6):1081–1092. https://doi.org/10.1111/jnc.13723 PMID: 27365114

59. Kasherman MA, Premarathne S, Burne TH, Wood SA, Piper M. The ubiquitin system: a regulatory hub for intellectual disability and autism spectrum disorder. Molecular neurobiology. 2020; p. 1–15. PMID: 31974941

60. Tsao SW, Tsang CM, Lo KW. Epstein–Barr virus infection and nasopharyngeal carcinoma. Philosophical Transactions of the Royal Society B: Biological Sciences. 2017; 372(1732):20160270. https://doi.org/10.1098/rstb.2016.0270

61. Cao Y. EBV based cancer prevention and therapy in nasopharyngeal carcinoma. NPJ precision oncology. 2017; 1(1):10. https://doi.org/10.1038/s41698-017-0018-x PMID: 29872698

62. Young LS, Dawson CW. Epstein-Barr virus and nasopharyngeal carcinoma. Chinese journal of cancer. 2014; 33(12):581. PMID: 25418193

63. Kagawa Y, Ishizuka M, Saishu T, Nakao S. Stable structure of thermophilic proton ATPase beta subunit. Journal of biochemistry. 1986; 100(4):923–934. https://doi.org/10.1093/oxfordjournals.jbchem.a121805 PMID: 2880841

64. Shukla S, Bhaskaran N, Maclennan GT, Gupta S. Deregulation of FoxO3a accelerates prostate cancer progression in TRAMP mice. The Prostate. 2013; 73(14):1507–1517. https://doi.org/10.1002/pros.22698 PMID: 23765843

65. Hiripi E, Lorenzo Bermejo J, Li X, Sundquist J, Hemminki K. Familial association of pancreatic cancer with other malignancies in Swedish families. British journal of cancer. 2009; 101(10):1792–1797. https://doi.org/10.1038/sj.bjc.6605363 PMID: 19826425

66. O'Neill M, Whelton M, Doyle C, Shorten E, Hennessy T. Endoscopic findings in patients after definitive gastric surgery. Irish medical journal. 1975; 68(1):9–12. PMID: 1110154

67. Shukla S. FOXO3a: A potential target in prostate cancer. Austin journal of urology. 2014; 1(1). PMID: 25584362

68. Liu Y, Ao X, Ding W, Ponnusamy M, Wu W, Hao X, et al. Critical role of FOXO3a in carcinogenesis. Molecular cancer. 2018; 17(1):104. https://doi.org/10.1186/s12943-018-0856-3 PMID: 30045773

69. Shan Z, Li Y, Yu S, Wu J, Zhang C, Ma Y, et al. CTCF regulates the FoxO signaling pathway to affect the progression of prostate cancer. Journal of cellular and molecular medicine. 2019; 23(5):3130–3139. https://doi.org/10.1111/jcmm.14138 PMID: 30873749

70. Huang R, Wallqvist A, Covell DG. Comprehensive analysis of pathway or functionally related gene expression in the National Cancer Institute's anticancer screen. Genomics. 2006; 87(3):315–328. https://doi.org/10.1016/j.ygeno.2005.11.011 PMID: 16386875

71. Legler K, Rosprim R, Karius T, Eylmann K, Rossberg M, Wirtz RM, et al. Reduced mannosidase MAN1A1 expression leads to aberrant N-glycosylation and impaired survival in breast cancer. British journal of cancer. 2018; 118(6):847–856. https://doi.org/10.1038/bjc.2017.472 PMID: 29381688

72. Oliver D, Ji H, Liu P, Gasparian A, Gardiner E, Lee S, et al. Identification of novel cancer therapeutic targets using a designed and pooled shRNA library screen. Scientific reports. 2017; 7(1):1–16. https://doi.org/10.1038/srep43023 PMID: 28223711