Graph-based Hub Gene Selection Technique using Protein Interaction Information: Application to Sample Classification

Pratik Dutta*, *Student Member, IEEE,* Sriparna Saha*, *Senior Member, IEEE*, and Saurabh Gulati†

***Abstract*—Classification of samples of gene expression profile plays a significant role in prediction and diagnosis of diseases. In the task of sample classification, a robust feature selection algorithm is very much essential to identify the important genes from the high dimensional gene expression data. This paper explores the information of protein-protein interaction (PPI) with a graph mining technique for finding a proper subset of features (genes), which further takes part in sample classification. Here, our contribution for feature selection is three-fold: firstly, all the genes are grouped into different clusters based on the integrated information of the gene expression values and their protein interactions using a multi-objective optimization (MOO) based clustering approach. Secondly, the confidence scores of the protein interactions are incorporated in a popular graph mining algorithm namely Goldberg algorithm to find out the relevant features. These features are the topologically and functionally significant genes, named as hub genes. Finally, these hub genes are identified varying the degrees of the nodes, and those are utilized for the sample classification task. Different machine learning classifiers are exploited for this purpose, and the classification performance is measured with respect to various performance metrics namely accuracy, sensitivity, specificity, precision, F-measure and Mathews coefficient correlation (MCC). Comparative analysis with respect to two baselines and several existing approaches proves the efficiency of the proposed approach. Further, the robustness of the identified hub-gene modules is endorsed using some strong biological significance analysis.**

***Index Terms*—Multi-objective optimization, Protein-protein interaction (PPI), Hub gene, Goldberg algorithm, Feature selection.**

Availability of codes and data: https://github.com/sduttap16/GraphPPI

I. Introduction

A. Background

With the technical enhancement in genomics, analysis of gene expression profiles leads to the discovery of some biologically significant genes. In the field of biomedical research, identification of informative genes is carried out in two ways; (i) the genes are clustered into some homogeneous groups and further analysis of these clusters gives the knowledge of the informative genes; (ii) group the genes into different clusters and from each cluster, extract a subset of informative genes (features). This subset of genes can be further used for sample classification.

* Department of Computer Science and Engineering, Indian Institute of Technology, Patna (e-mail: pratik.pcs16@iitp.ac.in, sriparna@iitp.ac.in). Both these authors have equally contributed for this work.

† Department of Chemical and Biochemical Engineering, IIT Patna.

In the current literature, the second method (i.e., informative feature extraction and sample classification) is used for identifying disease related genes. For the grouping of genes, clustering is a very popular unsupervised pattern classification method but that often gets stuck at local optima depending on the initial values of centroids. To get rid of these problems and also to optimize various cluster quality measures simultaneously, multi-objective optimization (MOO) based clustering approach has become popular. Unlike single-objective optimization based techniques, in case of MOO-based approaches, a set of non-dominated solutions are present in the final solution set [1].

However, due to the presence of huge number of genes in the microarray profiles, each cluster contains a large number of similar genes. To reduce the redundancy and the complexity of the high dimensional space, the immediate solution is to select representative features (genes) from each cluster. Extracting the representative genes from a cluster using biological knowledge is an emerging field of research in recent years. For exploring the biological knowledge, protein-protein interaction (PPI) network [2] is one of the enriched sources. Basically, the topology of the PPI network maintains "scale-free" property [3], i.e., a few number of proteins are strongly connected while most of the proteins are loosely connected. These strongly connected nodes are called hub proteins/genes [4]. As the hub proteins possess influential characteristics, it is conceivable that hub proteins can be considered as informative features (genes).

B. Related Works & Motivation

There are several existing works related to the improvements of feature selection algorithms as well as identification of hub genes. In [5], Liu et al. proposed an integrated method for feature (gene) selection that combines statistical similarity measure and supervised learning named as recursive feature addition (RFA). Recently, feature (gene) selection algorithm utilizing different biological knowledge is gaining the interest of the researchers. An unsupervised feature selection technique utilizing biological knowledge extracted from gene ontology (GO) is proposed by Acharya et al. [6].

In the literature, it has been found that the information about disease gene associations obtained from the protein-protein interaction leads to the higher classification accuracy in informative gene selection [7]. In the PPI network, the nodes which are strongly connected, play some significant roles in

identifying drug targets and termed as hub proteins. Generally, hub proteins have special biological properties compared to non-hub proteins and may act as the informative features in feature extraction algorithm. Recently, Xionglei et al.[8] showed why hub proteins are fundamental components in the protein networks. As there is no predefined rule for the cutoff degree of the hub genes, it is very much necessary to calculate the threshold degrees of the hub proteins. In [4], three different methods are described to identify hub proteins from PPINs.

There are several available works on the themes of feature selection and hub gene identification, separately; but utilizing the characteristics of hub genes in feature selection algorithm was not explored so much. Motivated by this fact, we have proposed a graph-based feature (hub gene) selection method utilizing the protein interaction information. Further the reduced set of genes are utilized in a MOO-based clustering framework to group the samples into some reasonable categories.

II. Proposed Methodology

A. Multi-objective Clustering Architecture

In the proposed method, a new MOO-clustering technique [9], where the goodness of clusters is validated by four objective functions $\{f_1, f_2, f_3, f_4\}$, is used. First, all the genes $\{\mathcal{G}_i \mid i \in [1, N]\}$ of $\mathcal{G}$, are normalized across the samples $[\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_\mathcal{S}]$, i.e., the normalized value of $\mathcal{X}^k_{i|i\in[1,\mathcal{S}]}$ is

$$\widehat{\mathcal{X}}^k_i = \frac{\max\limits_{k\in[1,N]}\{\mathcal{X}^k_i\} - \mathcal{X}^k_i}{\max\limits_{k\in[1,N]}\{\mathcal{X}^k_i\} - \min\limits_{k\in[1,N]}\{\mathcal{X}^k_i\}} \tag{1}$$

Now, the new MOO-based clustering technique is applied on this normalized gene expression profile $\widehat{\mathcal{G}} = \{\widehat{\mathcal{G}_1}, \widehat{\mathcal{G}_2}, \ldots, \widehat{\mathcal{G}_N}\}$ and produced a clustering solution (partitioning) π^* aiming that

$$f_t(\pi^*) = \max_{\pi\in\Omega} f_t(\pi) \;\text{ where }\; t \in \{1, 2, 3, 4\} \tag{2}$$

Where Ω is the set of all feasible clustering solutions, π^* is the clustering solution that has optimized all t objective functions. In the proposed approach, the four objective functions are $f_1 :=$ fuzzy partition coefficient (FPC), $f_2 :=$ PBM index, $f_3 :=$ biological homogeneity index (BHI) and $f_4 :=$ protein-protein interaction confidence score (PPICS). Here, $\{f_1, f_2\}$ are traditional cluster validity indices and $\{f_3, f_4\}$ are bio-oriented cluster validity indices. Among $\{f_3, f_4\}$, f_3:= BHI, which measures the biological or functional homogeneity (similarity) of a particular cluster of genes and f_4:= PPICS [9], which is the newly developed cluster quality measure calculated using confidence scores of the protein-protein interactions.

The proposed MOO-based clustering technique utilizes three genetic operators, namely crossover, mutation, and selection.

1) *Crossover*: In the current study, we have used single point crossover (with crossover probability p_c) where the crossover point is stochastically selected from the parent chromosomes. The procedure for selecting the crossover point ensures that the length of the offspring lies over a range $[2, \sqrt{N}]$.
2) *Mutation*: For the mutation purpose, the three types of mutations, e.g., *Normal Mutation*(ρ_n) (changing any cluster center by a small amount), *Insert Mutation*(ρ_i) (inserting a new cluster center) and *Delete Mutation*(ρ_d) (deleting a center from the chromosome), are applied on a chromosome at a particular time.
3) *Selection*: Lastly, to select the best chromosomes for the next generation, a binary tournament selection operator is used. Two individual chromosomes are randomly picked up to play the tournament and winner is chosen after considering their non-domination ranks and crowding distances[1].

For our MOO-based clustering technique, the values of different parameters are tabulated in Table-I of the **Supplementary material**. Finally, this MOO-based clustering technique simultaneously optimizes $\{f_1, f_2, f_3, f_4\}$ and generates a Pareto front $\Pi^\mathcal{P} = \{\pi^\mathcal{P}_1, \pi^\mathcal{P}_2, \ldots, \pi^\mathcal{P}_M\}$, where each of the solutions of Pareto front is non-dominated to each other, i.e., $\{\pi^\mathcal{P}_i \mid \nexists \pi^\mathcal{P}_j \in \Pi^\mathcal{P} : \pi^\mathcal{P}_j \succ \pi^\mathcal{P}_i\}$ where $\succ$ represents the dominance relation. Figure-1 represents the schematic flowchart of our proposed architecture.

B. Creation of Induced Network and Finding the Dense Subgraph (DSG)

From the obtained non-dominated solutions $\Pi^\mathcal{P} = \{\pi^\mathcal{P}_1, \pi^\mathcal{P}_2, \ldots, \pi^\mathcal{P}_M\}$, we picked a single partitioning (clustering solution) $\pi^\mathcal{P}_\mathcal{M}$ on the basis of the Silhouette Index $s(C)$. As the MOO-based clustering technique simultaneously optimizes bio-oriented cluster validity indices$\{f_3, f_4\}$, the genes of each cluster are functionally and biologically similar. Now to identify the representative features from these similar sets of genes, we utilize the characteristics of hub genes. Generally, hub genes possess some special functional and topological significance. The functional properties of genes are taken care of by the MOO-based clustering technique (described in the previous section). Now to understand the topological significance of the genes, we explore the information of the protein-protein interaction (PPI) network.

Let, the partitioning $\pi^\mathcal{P}_\mathcal{M}$ contain a set of clusters $\{C_1, C_2, \ldots, C_K\}$, and consider a PPI network $\mathcal{N}$ that is defined by a triplet $\langle V, E, \Phi\rangle$, where $V = \{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_\mathbb{Z}\}$ is the set of proteins/genes, $E = \{e_1, e_2, \ldots, e_m\} \subseteq V\text{X}V$ is the edges (links) between proteins and $\Phi = \{\phi_1, \phi_1, \ldots, \phi_m\}$ where, $\forall i : \phi_i$ represents the confidence score of e_i. In the proposed feature selection technique, the confidence score (ϕ_i) [10] acts as the edge weight of e_i.

Now, a set of induced subgraphs $\mathcal{I}^{'} = \{\mathcal{N}^\mathcal{I}_1, \mathcal{N}^\mathcal{I}_2, \ldots, \mathcal{N}^\mathcal{I}_K\}$ are generated by mapping each of the clusters $\{C_i : C_i \in \pi^\mathcal{P}_\mathcal{M}\}$ to $\mathcal{N}$. Here $\{\mathcal{N}^\mathcal{I}_i : i \in [1, K]\}$ denoted by $\{(V^\mathcal{I}_i, E^\mathcal{I}_i) \mid \forall i : (V^\mathcal{I}_i \subset V \in C_i) \wedge (E^\mathcal{I}_i \subset E)\}$. This mapping is basically performed to acquire the topological or the structural information of the genes. The generation of the induced network ($\mathcal{N}^\mathcal{I}_i$) is carried out using Cytoscape [11]. Figure-2a shows an induced network ($\mathcal{N}^\mathcal{I}_i$) generated from $\{C_i : C_i \in \pi^\mathcal{P}_\mathcal{M}\}$ of B-CLL dataset.

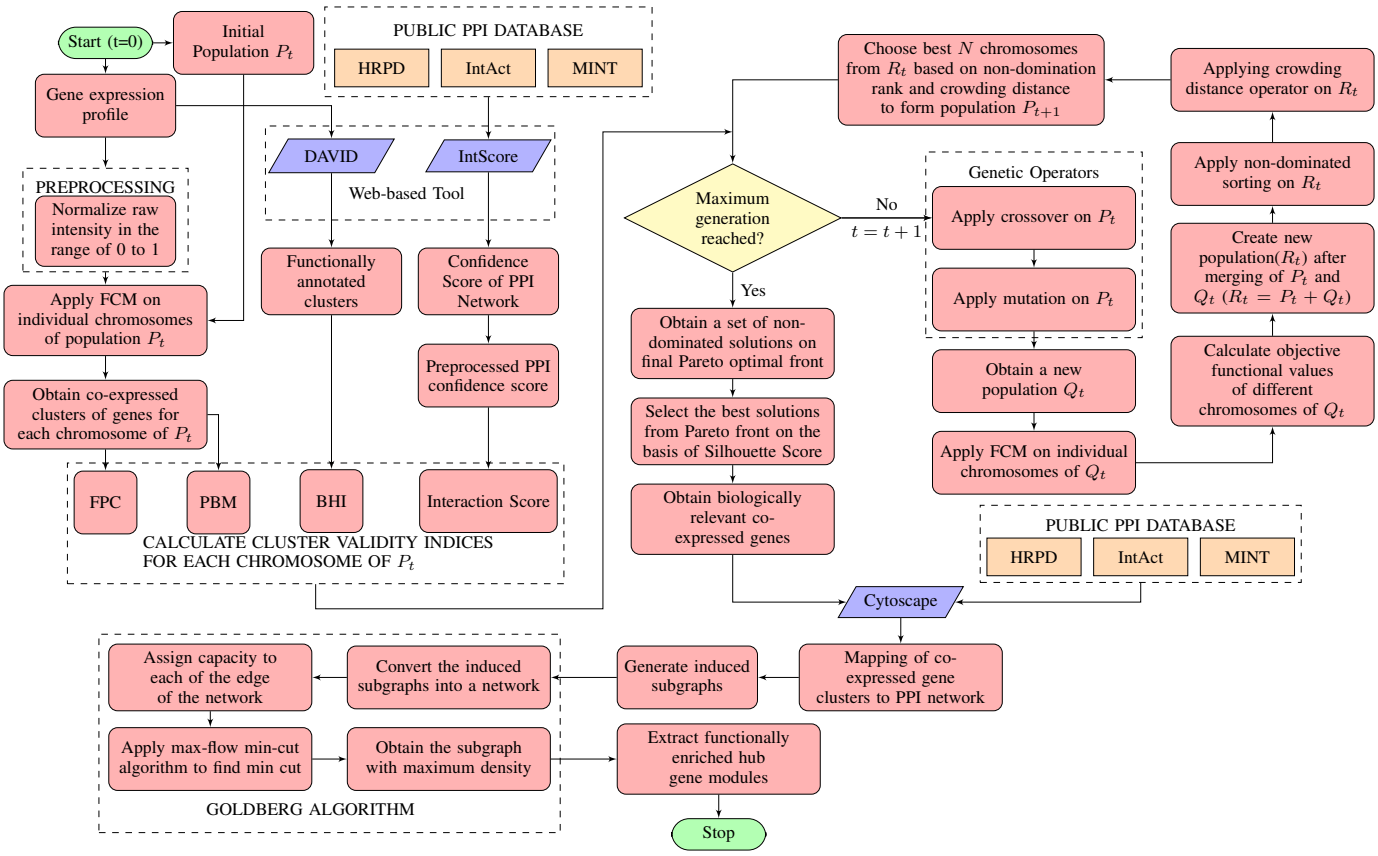

Figure 1: Flowchart of Our Proposed Hub-genes Identification Framework based on the Multi-objective Optimization technique and the Goldberg Algorithm, FPC: Fuzzy partition Coefficient, PBM: Pakhira-Bandyopadhyay-Maulik Index, BHI: Biological Homogeneity Index

Here, all the nodes (genes/proteins) of any $\mathcal{N}_i^{\mathcal{I}}$ are functionally similar, but not all are strongly connected, i.e., topologically significant. To identify those intensely connected nodes, the condensed part (region) of any $\mathcal{N}_i^{\mathcal{I}}$ is found by utilizing a popular graph mining algorithm, named as Goldberg algorithm [12]. In the proposed approach, the confidence score (ϕ_i) of each edge (e_i) is prudently incorporated into the Goldberg algorithm during the search for dense part of each $\mathcal{N}_i^{\mathcal{I}}$. The modified Goldberg Algorithm is illustrated in Algorithm-1

The core part of the Goldberg algorithm is to generate a network $\mathcal{N}_i^*$ from $\mathcal{N}_i^{\mathcal{I}}$, which is described in line-6 to line-19 of Algorithm-1. During the construction of $\mathcal{N}_i^*$, the confidence scores, Φ of the interactions (edges) are prudently utilized (in line number-12 of Algorithm-1). This modified Goldberg algorithm takes $\{\mathcal{N}_i^{\mathcal{I}} : i \in [1, K]\}$ and generates $\mathbb{D}$ = $\{\mathfrak{D}_1, \mathfrak{D}_2, \ldots, \mathfrak{D}_K\}$, where $\mathfrak{D}_{i|i\in[1,K]}$ = $\{\partial_{i1}^{\mathcal{I}}, \partial_{i2}^{\mathcal{I}}, \ldots, \partial_{id}^{\mathcal{I}}\}$ is the set of dense subgraphs (DSG) obtained using the Algorithm-1. The time complexity of this modified Goldberg algorithm is $\mathcal{O}(Knm[\log(n)]^2)$ (refer to section-I of the supplementary material), where n and m are the number of nodes (genes) and edges of any induced network ($\mathcal{N}_i^*$), respectively. Different dense subgraph (DSG) modules obtained from the induced network (illustrated in Figure-2a) are shown in Figure-2b to Figure-2j .

C. Identifying Hub Proteins

Now, we have combined all the DSGs obtained from $\mathcal{N}_i^{\mathcal{I}}$, to find all the genes, which belong to different dense subgraphs of $\mathcal{N}_i^{\mathcal{I}}$. After combining, we get a set of topologically significant genes, i.e., $\mathcal{D}_i^* = \{\{\partial_{i1}^{\mathcal{I}}\} \cup \{\partial_{i2}^{\mathcal{I}}\} \cup \ldots \{\partial_{id}^{\mathcal{I}}\}\} = \{\mathcal{G}_{i1}, \mathcal{G}_{i2}, \ldots, \mathcal{G}_{i\mathcal{F}}\}$. Similarly, from all other induced networks, we obtained topologically significant genes and finally a set of all topologically significant genes is obtained, i.e., $\mathbb{D}^*$ = $\{\mathcal{D}_1^*, \mathcal{D}_2^*, \ldots, \mathcal{D}_K^*\}$

Now, to identify the hub genes from $\mathcal{D}_{i|[1,K]}^*$, the degree ($\delta_{l|l\in[1,P]}$) of the node (gene) $\mathcal{G}_{il}$ is considered. The degree of a node (gene) represents the number of genes interacting to that particular gene. As there is no consensus rule for the threshold degree (δ_{th}) of the hub genes, we heuristically vary δ_{th} for identifying the hub genes. Inspired by the recent literature [4], we extracted the hub genes depending upon the value of $\delta_{th} \in \{3, 5, 17, 33, 85\}$. Hence, the final hub gene set is defined by $HG_{z|z\in\{3,5,17,33,55,85\}} = \{H_z^1, H_z^2, \ldots H_z^K\}$; where $H_z^i = \{\mathcal{G}_{ij|j\in[1,\mathcal{F}]} : (\mathcal{G}_{ij} \in \mathcal{D}_i^*) \wedge (\delta_j > \delta_{th} = z)\}$.

These hub genes (features) are both functionally (taken care by MOO-based clustering technique) and topologically (taken care by modified Goldberg algorithm) significant. Thus $HG_{z|z\in\{3,5,17,33,85\}}$ can act as the informative features of the feature space. Further to analyze the goodness of extracted features, we have shown the utility of the subset of features

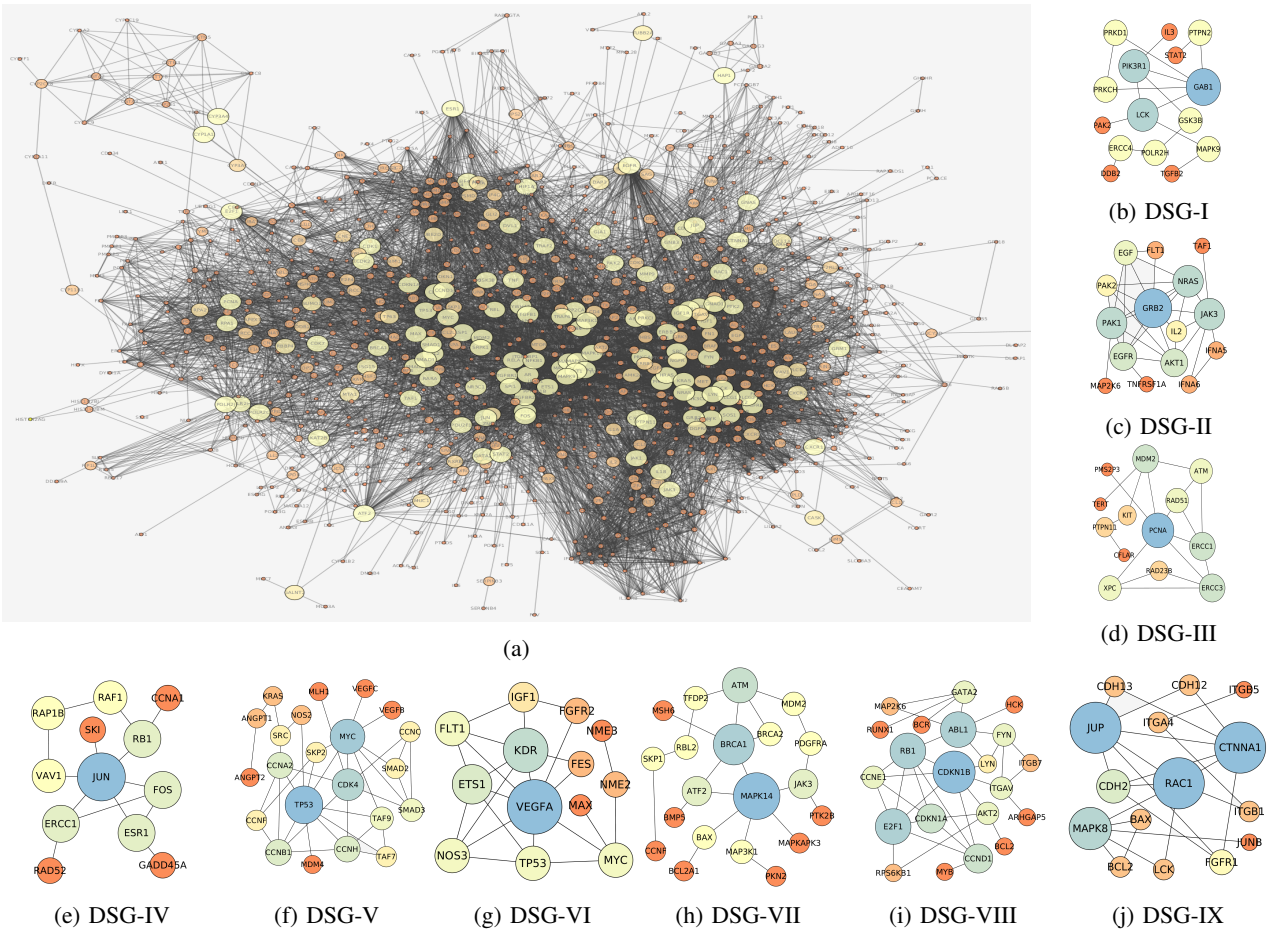

Figure 2: Visualization of generated induced network and dense-subgraphs using the Cytoscape; (a) Induced network ($\mathcal{N}_i^{\mathcal{I}}$) generated from $\{C_i : C_i \in \pi_{\mathcal{M}}^{\mathcal{P}}\}$ of B-CLL dataset. (b)-(j) DSG-I to DSG-IX generated from $\mathcal{N}_i^{\mathcal{I}}$ after application of Algorithm-1; blue nodes have higher degrees and color of the nodes changes to orange if the degree of the node reduces.

on sample classification task. The sample classification task is briefly described in the following section.

III. Experimental Results and Discussion

A. Datasets

In the current experiment, two types of datasets are used: gene expression datasets and PPI datasets. In case of microarray datasets, six NCBI's GEO datasets, i.e., B-CLL, ILD, Prostate dataset, GDS3268, GDS3795 and GDS4206 are used. Also, we have used three simulated datasets with different noise levels and sample sizes. The brief description of all these datasets is provided in Table-IV of the supplementary material. The protein-protein interaction(PPI) network is obtained after merging three PPI databases, i.e., Human Protein Reference Database(HRPD), Molecular Interaction database(MINT) and IntAct database.

B. Selection of Evaluation Metrics

The choice of cluster validity indices plays a crucial role in multi-objective clustering technique. In this regard, we have applied an interactive approach [13] to find out the best set of cluster validity indices. After applying this interactive approach, we found that $\{FPC, PBM, BHI, PPICS\}$ is one of the best set of objective functions for gene clustering by utilizing their molecular functionalities. The Silhouette score is used as a prioritize criteria among the non-dominated solutions because it has been widely used to understand the goodness of gene clustering.

C. Analysis of DSG as the Extracted Features

To analyze the goodness of the extracted features (hub genes), sample classification is done by utilizing four well-known classifiers, namely k-NN(k=3), random forest, support vector machine(SVM) and artificial neural network (ANN). As all the datasets contain two types of samples (tumour and normal), we perform the binary classification. The parameter values of these four binary classifiers are tabulated in Table-II of the supplementary material. Here, the extracted features are termed as hub gene modules ($HG_{z|z\in\{3,5,17,33,55,85\}}$). We have obtained the average results after conducting 10 times 10-fold cross-validation and the results are evaluated using five performance metrics, namely sensitivity [14], specificity [14], precision [14], F-measure[14] and Mathews coefficient correlation (MCC) [15]. The performance metrics of the extracted features and the dense subgraphs of three real-life datasets

Algorithm 1 Modified Goldberg algorithm

linenosize=

Input: $\mathcal{N}_i^{\mathcal{I}} = (V_i^{\mathcal{I}}, E_i^{\mathcal{I}})$
Output: $\mathfrak{D}_i = \{\partial_{i1}^{\mathcal{I}}, \partial_{i2}^{\mathcal{I}}, \ldots, \partial_{id}^{\mathcal{I}}\}$ ▷ A set of dense subgraphs (DSG)

procedure GOLDBERG($\mathcal{N}_i^{\mathcal{I}}$)
  $temp \leftarrow 0; counter = 0$ ▷ Initialization
  $temp_1 \leftarrow m = \left|E_i^{\mathcal{I}}\right|; \quad temp_2 \leftarrow n := \left|V_i^{\mathcal{I}}\right|$
  while $\left((temp_1 - temp) \geq \frac{1}{temp_2(temp_2-1)}\right)$ **do**
    $g \leftarrow \frac{temp+temp_1}{2}$
    Construct a network $\mathcal{N}_i^* = (V_i^*, E_i^*)$ from $\mathcal{N}_i^{\mathcal{I}}$
    $V_i^* \leftarrow V_i^{\mathcal{I}} \cup \{s, t\}$ ▷ Add two vertices source (s) and sink (t)
    $E_u^* = \{\}$
    for $k \in [1, 2, \ldots, temp_1]$ **do**
      if $e_k \in (\mathcal{G}_\alpha, \mathcal{G}_\beta)$ **then**
        $E_u^* = E_u^* \cup \{\vec{e}_{\alpha,\beta}, \vec{e}_{\beta,\alpha}\}$ ▷ Replace an undirected edge by two directed edges
        $w(\vec{e}_{\alpha,\beta}) = w(\vec{e}_{\beta,\alpha}) = \phi_k$ ▷ $w(\vec{e}_{\alpha,\beta})$ is the weight or capacity of $\vec{e}_{\alpha,\beta}$
    $E_s^* = \{\}; E_t^* = \{\}$
    for $\forall \mathcal{G}_k \in \{V_i^{\mathcal{I}}\}$ **do**
      Generate $\vec{e}_{s,\mathcal{G}_k}$ and $\vec{e}_{\mathcal{G}_k,t}$
      $w(\vec{e}_{s,\mathcal{G}_k}) = m; E_s^* = E_s^* \cup (\vec{e}_{s,\mathcal{G}_k})$
      $w(\vec{e}_{\mathcal{G}_k}, t) = m + 2g - d_k; E_t^* = E_t^* \cup (\vec{e}_{\mathcal{G}_k,t})$ ▷ d_k is the degree of $\mathcal{G}_k$
    $E_i^* = \{E_u^* \cup E_s^* \cup E_t^*\}$
    $V_1 = \{\}$ ▷ Make an empty list that will contain the nodes (genes) of the DSG obtained from $\mathcal{N}_i^*$
    FIND MIN-CUT(S,T)
    if $S = \{s\}$ **then**
      $temp_1 \leftarrow g$
    else
      $temp \leftarrow g$
      $V_1 \leftarrow S - \{s\}$
    $temp_2 = temp_2 - |V_1|$
    $\partial_{i.counter}^{\mathcal{I}} = \{V_1\}$
    $counter+ = 1$
  end while

Table I: Comparison of Accuracy of Different Existing Algorithms with Our Proposed Algorithm

Accuracy							
		B-CLL	ILD	Prostate	GDS3268	GDS3795	GDS4206
Falt et al. (2005)	WV	0.71	-	-	-	-	-
	LDA	0.90	-	-	-	-	-
Chuang et al. (2007)	kNN	-	-	0.81	-	-	-
	RF	-	-	0.82	-	-	-
	SMO	-	-	0.87	-	-	-
Taylor et al. (2009)	-	-	-	0.68	-	-	-
Ahn et al. (2011)	-	-	-	0.82	-	-	-
	-	-	-	0.84	-	-	-
Cho at al. (2011)	RbFS	-	1.00	-	-	-	-
Swarnakar et al. (2015)	3NN	0.81	0.83	0.85	-	-	-
	RF	0.57	0.79	0.84	-	-	-
	SVM	0.81	0.86	0.87	-	-	-
Ge at al. (2016)	3NN	0.77	0.75	0.80	0.49	0.88	0.71
	RF	0.76	0.86	0.84	0.51	0.91	0.77
	SVM	0.64	0.82	0.83	0.63	0.87	0.66
	ANN	0.64	0.73	0.90	0.43	0.75	0.70
Riverol et al. (2017)	3NN	0.65	0.86	0.85	0.65	0.85	0.72
	RF	0.72	0.86	0.80	0.63	0.91	0.75
	SVM	0.71	0.86	0.82	0.53	0.88	0.78
	ANN	0.50	0.67	0.88	0.65	0.91	0.76
Kang et al. (2017)	3NN	0.64	0.61	0.85	0.51	0.91	0.78
	RF	0.61	0.76	0.87	0.65	0.88	0.77
	SVM	0.71	0.79	0.79	0.63	0.83	0.69
	ANN	0.67	0.79	0.90	0.67	0.88	0.73
Proposed Method	3NN	0.71 ± 0.10	0.86 ± 0.09	0.88 ± 0.06	0.71 ± 0.04	0.92 ± 0.02	0.78 ± 0.07
	RF	0.67 ± 0.12	0.89 ± 0.03	0.87 ± 0.02	0.67 ± 0.03	0.95 ± 0.02	0.82 ± 0.06
	SVM	0.76 ± 0.10	0.86 ± 0.06	0.91 ± 0.02	0.67 ± 0.01	0.91 ± 0.02	0.78 ± 0.06
	ANN	0.75 ± 0.15	0.88 ± 0.06	0.93 ± 0.03	0.70 ± 0.02	0.95 ± 0.02	0.85 ± 0.10

are reported in Table-V, VII, and VIII of the supplementary material. Also, we analyze the performance of the extracted features with respect to two baseline approaches.

- Baseline 1: Here, we consider all the preprocessed genes of the dataset during sample classification.
- Baseline-2: Here, we assume that the centroids, $\{C_i \mid C_i \in \pi_{\mathcal{M}}^{\mathcal{P}}\}$, of the MOO-based clustering technique are used as the representative features and further utilized for sample classification using the above mentioned three classifiers.

In Table-V of the supplementary material, we have shown the performance metrics for the B-CLL dataset with respect to different DSGs (DSG-I to DSG-IX), the hub genes modules $(HG_{z|z\in\{3,5,17,33,55\}})$ and two baselines. It is shown that the performance of $HG_{z|z\in\{3,5,17,33,85\}}$ identified by the proposed approach for B-CLL data outperforms the baselines and most of the dense subgraphs. In Table-VII and Table-VIII of the supplementary material, a comparative analysis of the performance metrics for baselines and different hub gene modules of ILD and prostate datasets are presented, respectively. For both datasets, the hub gene modules identified by the proposed method outperform two baselines in sample classification with respect to all the performance metrics irrespective of classification model used.

The obtained comparative results confirm that the proposed graph-based hub gene (as informative feature) identification method utilizing the power of MOO-based clustering technique is a productive approach. Incorporation of the confidence scores (ϕ_i) of the PPI in the Goldberg algorithm enables to identify the biologically and topologically significant gene modules. These gene modules are further used as the informative features for sample classification.

D. Comparative Analysis of the Proposed Method with the Existing Methods

To validate the superiority of the proposed method over other state-of-art algorithms, we have compared the performance of our approach with several existing methods in terms of different performance metrics. All the performance metrics are calculated as a two-class classification problem. In Table-I, a comparative analysis with nine state-of-art techniques in terms of accuracy is tabulated. Among these methods, Fält et al.[16] used different supervised and unsupervised clustering methods to identify the important genes. Chuang et al. [17] and Taylor et al. [18] exploit protein interaction networks for predicting the gene markers. Similarly, Ahn et al. [19] and Swarnakar et al. [20] identified important genes by combining different genetic information with protein interaction network. System biology approaches, with global expression data sets, were used in Cho et al. [21]. Ge et al. (2016)[22] developed a feature selection technique which is based on a recent correlation measurement, Maximal Information Coefficient (MIC). In Riverol et al. (2017)[23], an integrated feature selection technique is developed by combining different popular feature selection methods with Correlation Matrix(CM) and Principal Component Analysis(PCA). Kang et al. (2017) proposed a new feature selection method by utilizing the significance analysis of microarrays (SAM). In Table-I, we have reported the comparative study of the accuracies for all six NCBI datasets. For our proposed technique, we run ten times the 10-fold cross-validation and report the variance of the accuracies in Table-I. Also, it is evident from Table-I that the overall performance of the proposed method outperforms other existing methods.

Table II: Comparison of Different Performance Metrics (Sensitivity, Specificity, F-measure and Matthews correlation coefficient(MCC)) of Different Existing Algorithms with Our Proposed Algorithm for Prostate Dataset

	Chuang et al. (2007)			Taylor et al.(2009)	Ahn et al. (2011)		Swarnakar et. al (2015)			Ge et. al (2016)				Riverol et. al (2017)				Kang et. al (2017)				Proposed Method			
	kNN	RF	SMO	-	-	-	3NN	RF	SVM	3NN	RF	SVM	ANN	3NN	RF	SVM	ANN	3NN	RF	SVM	ANN	3NN	RF	SVM	ANN
Sensitivity	0.88	0.82	0.90	0.75	0.86	0.90	0.84	0.71	0.74	0.63	0.63	0.66	0.65	0.67	0.64	0.67	0.63	0.76	0.86	0.82	0.67	**0.87**	**0.89**	**0.92**	**0.95**
Specificity	0.74	0.82	0.84	0.62	0.78	0.78	0.86	0.90	0.93	0.85	0.88	0.85	0.93	0.86	0.79	0.81	0.93	0.62	0.74	0.83	0.79	**0.86**	**0.79**	**0.83**	**0.91**
F-measure	-	-	-	-	-	-	0.78	0.74	0.78	0.66	0.66	0.68	0.67	0.68	0.61	0.68	0.66	0.78	0.82	0.78	0.79	**0.90**	**0.91**	**0.94**	**0.96**
MCC	0.63	0.65	0.75	0.49	0.65	0.69	0.60	0.62	0.69	0.65	0.61	0.71	0.82	0.62	0.62	0.69	0.80	0.65	0.66	0.62	0.65	**0.69**	**0.67**	**0.70**	**0.83**

Table III: Comparison of Different Performance Metrics for B-CLL, ILD, GDS3268, GDS3795 and GDS4206 datasets

		Swarnakar et. al (2015)			Ge et. al (2016)				Riverol et. al (2017)				Kang et. al (2017)				Proposed Method			
Dataset	Performance metrics	3NN	RF	SVM	3NN	RF	SVM	ANN	3NN	RF	SVM	ANN	3NN	RF	SVM	ANN	3NN	RF	SVM	ANN
B-CLL	Precision	0.56	0.58	0.58	0.50	0.50	0.40	0.20	0.50	0.50	0.50	0.20	0.55	0.50	0.50	0.55	**0.80**	**0.57**	**0.88**	**0.73**
	F-measure	0.70	0.61	0.61	0.39	0.37	0.33	0.36	0.36	0.36	0.43	0.30	0.36	0.33	0.43	0.39	**0.72**	**0.64**	**0.61**	**0.73**
	MCC	0.16	0.14	0.14	0.22	0.10	0.14	0.13	0.14	0.24	0.14	0.14	0.44	0.14	0.14	0.29	**0.44**	**0.15**	**0.17**	**0.62**
ILD	Precision	0.50	1.00	0.67	0.80	0.80	0.80	0.70	0.80	0.80	0.80	0.80	0.67	0.80	0.67	0.69	**0.95**	**0.95**	**0.82**	**0.79**
	F-measure	0.25	0.50	0.44	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.81	0.81	0.80	0.89	**0.90**	**0.93**	**0.85**	**0.76**
	MCC	0.20	0.53	0.39	0.75	0.61	0.65	-0.09	0.50	0.37	0.58	0.66	0.61	0.65	0.65	0.71	**0.67**	**0.75**	**0.67**	**0.83**
GDS3268	Precision	-	-	-	0.63	0.64	0.64	0.64	0.65	0.63	0.65	0.63	0.65	0.63	0.63	0.65	**0.75**	**0.81**	**0.81**	**0.81**
	F-measure	-	-	-	0.49	0.53	0.65	0.48	0.55	0.55	0.58	0.55	0.53	0.49	0.49	0.61	**0.65**	**0.58**	**0.65**	**0.74**
	MCC	-	-	-	0.12	0.13	0.09	-0.31	0.23	0.22	-0.26	0.23	0.19	0.23	0.22	0.22	**0.50**	**0.45**	**0.46**	**0.56**
GDS3795	Precision	-	-	-	0.05	0.05	0.02	0.05	0.02	0.02	0.05	0.01	0.02	0.02	0.02	0.01	**0.24**	**0.25**	**0.32**	**0.31**
	F-measure	-	-	-	0.01	0.01	0.01	0.01	0.01	0	0.01	0.01	0	0	0.01	0.03	**0.13**	**0.13**	**0.15**	**0.09**
	MCC	-	-	-	-0.05	-0.02	-0.03	-0.10	-0.06	-0.02	-0.02	-0.08	-0.03	-0.03	-0.02	-0.03	**0.10**	**0.12**	**0.12**	**0.12**
GDS4206	Precision	-	-	-	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.67	0.78	0.78	0.80	**0.83**	**0.80**	**0.82**	**0.92**
	F-measure	-	-	-	0.75	0.77	0.70	0.79	0.75	0.76	0.78	0.79	0.75	0.69	0.72	0.79	**0.79**	**0.83**	**0.85**	**0.90**
	MCC	-	-	-	-0.12	-0.09	-0.04	-0.10	-0.05	-0.13	-0.01	0.14	0.03	0.03	0.09	0.21	**0.21**	**0.19**	**0.17**	**0.41**

In Table-II, a comparative study of different performance metrics for the Prostrate dataset is tabulated. Similarly, Table-III shows the comparative analysis of different performance metrics for B-CLL, ILD, GDS3268, GDS3795 and GDS4206 datasets against other comparative methods. But results of many of the existing methods are not available for six NCBI datasets. Therefore, the unavailable information is represented by the dash (-) in Table-I, II and III. Fig-1 of the supplementary material shows the performance evaluation of the proposed feature selection method against other existing methods for all six datasets. Only for ILD dataset, the error measure of Chuang et al. is less than the proposed method. But from the Fig-1 of the supplementary material, it is clearly evident that overall performance of proposed method is better than all other existing methods. From this comparative study, it is clearly evident that the proposed feature extraction technique surpasses the existing methods in terms of overall performance for all six datasets.

To prove the effectiveness of the proposed method, we have also considered three simulated datasets with different noise levels and sample sizes. These datasets are generated in a way so that they possess similar characteristics [24] of real microarray datasets. The comparative study of performance for the simulated datasets is reported in Table-VI of the supplementary material which proves the effectiveness of the proposed method for simulated datasets. The results of the comparative study (for both NCBI datasets and simulated datasets) have demonstrated that our proposed method is effective for detecting significant genes and classification tasks. Finally, we perform a statistical hypothesis, Welch's t-test [25], for all the performance metrics to prove that the performance improvements attained by the proposed method are statistically significant. Here the null hypothesis presumes that there is an insignificant difference between the mean values of two groups and alternative hypothesis states that there are significant differences in the mean values of two groups. The p-values of Welch's t-test on six gene expression profiles are reported in Table-IX of the supplementary material and it is seen that all the p-values are less than 0.05. Thus the obtained p-values evidently support that the improvements attained by the proposed method are statistically significant and are not occurred by chance.

E. Biological Analysis

To analyze the biological significance of the extracted features, we have followed two different procedures. In the first procedure, Gene Ontology Consortium (http://www.geneontology.org/) is used to find out the significant gene ontology (GO) terms corresponding to extracted gene sets. Also it gives the enrichment of the GO category in terms of p-value, percentage (HG%) of genes of $HG_{z|z\in\{3,5,17,33,55\}}$ or DSG related to a particular GO term, and percentage (Genome%) of genes of Gene Ontology Consortium related to a particular GO term. In Table-X of the supplementary material, we have reported the major biological and cellular processes of selected genes of the dense and hub gene modules for the B-CLL dataset. For example, in DSG-I of B-CLL dataset, 87.09% genes are related to GO term GO:0065007, which is responsible for the biological regulation of the cell. It is also shown that only 57.88% genes of the Gene Ontology Consortium are related to the particular GO term(GO:0065007). Hence the gene modules

or gene set of DSG-I are more enriched than the existing gene ontology database with respect to biological regulation process(GO:0065007).

In the second procedure, we have found out the correlations between the selected genes with their respective diseases by validating our result using a disease-gene association database namely DisGeNET [26]. In Table-III of the supplementary material, we have reported those genes (genes selected by our proposed method) which belong to the top 10 disease-related genes of the DisGeNET database. Results clearly show that 80%-100% of the top 10 disease-related genes of DisGeNET belong to our selected gene (feature) set. The obtained results of these two tables are strong evidence to justify that gene modules obtained by the proposed graph-based technique are strongly correlated with the particular disease and very much responsible for the biochemical process of the living cell.

IV. Conclusion

The current paper reports about the development of a method for hub-gene selection utilizing the concepts of multi-objective based clustering and a modified version of Goldberg algorithm. The MOO-based clustering technique utilizes the integrated information of PPI and expression profiles for simultaneously optimizing four objective functions, which help in identifying both biologically relevant and functionally similar genes. Finally, incorporation of protein interactions with Goldberg algorithm helps in identifying the densest part of the gene clusters obtained from the aforementioned MOO-based clustering technique. The performance of the proposed hub-gene selection approach is evaluated for sample classification. The results are compared with different existing methods with respect to five performance metrics. From the comparative study, described in section-III, it is easily inferred that the proposed method is superior to all the existing methods with respect to the overall performance. In future, we aim to develop a deep learning based feature selection algorithm integrated with multi-objective optimization to identify the informative genes.

Acknowledgment

PD would like to acknowledge the support from Visvesvaraya PhD Scheme of Government of India and SS would like to acknowledge the support from SERB Women in Excellence Award-SB/WEA-08/2017 for conducting this research.

References

[1] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: Nsga-ii," *IEEE transactions on evolutionary computation*, vol. 6, no. 2, pp. 182–197, 2002.

[2] P. G. Sun, "The human drug–disease–gene network," *Information Sciences*, vol. 306, pp. 70–80, 2015.

[3] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.

[4] R. R. Vallabhajosyula, D. Chakravarti, S. Lutfeali, A. Ray, and A. Raval, "Identifying hubs in protein interaction networks," *PloS one*, vol. 4, no. 4, p. e5344, 2009.

[5] Q. Liu, A. H. Sung, Z. Chen, J. Liu, X. Huang, and Y. Deng, "Feature selection and classification of maqc-ii breast cancer and multiple myeloma microarray gene expression data," *PloS one*, vol. 4, no. 12, p. e8250, 2009.

[6] S. Acharya, S. Saha, and N. Nikhil, "Unsupervised gene selection using biological knowledge: application in sample clustering," *BMC bioinformatics*, vol. 18, no. 1, p. 513, 2017.

[7] M. T. Dittrich, G. W. Klau, A. Rosenwald, T. Dandekar, and T. Müller, "Identifying functional modules in protein–protein interaction networks: an integrated exact approach," *Bioinformatics*, vol. 24, no. 13, pp. i223–i231, 2008.

[8] X. He and J. Zhang, "Why do hubs tend to be essential in protein networks?" *PLoS genetics*, vol. 2, no. 6, p. e88, 2006.

[9] P. Dutta and S. Saha, "Fusion of expression values and protein interaction information using multi-objective optimization for improving gene clustering," *Computers in Biology and Medicine*, vol. 89, pp. 31–43, 2017.

[10] J. Yu, T. Murali, and R. L. Finley, "Assigning confidence scores to protein–protein interactions," *Two Hybrid Technologies: Methods and Protocols*, pp. 161–174, 2012.

[11] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome research*, vol. 13, no. 11, pp. 2498–2504, 2003.

[12] A. V. Goldberg, *Finding a maximum density subgraph*. University of California Berkeley, CA, 1984.

[13] A. Mukhopadhyay, U. Maulik, and S. Bandyopadhyay, "An interactive approach to multiobjective clustering of gene expression patterns," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 1, pp. 35–41, 2013.

[14] J. Ji, A. Zhang, C. Liu, X. Quan, and Z. Liu, "Survey: Functional module detection from protein-protein interaction networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 2, pp. 261–277, 2014.

[15] L. He, Y. Wang, Y. Yang, L. Huang, and Z. Wen, "Identifying the gene signatures from gene-pathway bipartite network guarantees the robust model performance on predicting the cancer prognosis," *BioMed research international*, vol. 2014, 2014.

[16] S. Fält, M. Merup, G. Gahrton, B. Lambert, and A. Wennborg, "Identification of progression markers in b-cll by gene expression profiling," *Experimental hematology*, vol. 33, no. 8, pp. 883–893, 2005.

[17] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis," *Molecular systems biology*, vol. 3, no. 1, p. 140, 2007.

[18] I. W. Taylor, R. Linding, D. Warde-Farley, Y. Liu, C. Pesquita, D. Faria, S. Bull, T. Pawson, Q. Morris, and J. L. Wrana, "Dynamic modularity in protein interaction networks predicts breast cancer outcome," *Nature biotechnology*, vol. 27, no. 2, pp. 199–204, 2009.

[19] J. Ahn, Y. Yoon, C. Park, E. Shin, and S. Park, "Integrative gene network construction for predicting a set of complementary prostate cancer genes," *Bioinformatics*, vol. 27, no. 13, pp. 1846–1853, 2011.

[20] T. Swarnkar, S. N. Simões, A. Anura, H. Brentani, J. Chatterjee, R. F. Hashimoto, D. C. Martins, and P. Mitra, "Identifying dense subgraphs in protein–protein interaction network for gene selection from microarray data," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 4, no. 1, p. 33, 2015.

[21] J.-H. Cho, R. Gelinas, K. Wang, A. Etheridge, M. G. Piper, K. Batte, D. Dakhlallah, J. Price, D. Bornman, S. Zhang *et al.*, "Systems biology of interstitial lung diseases: integration of mrna and microrna expression changes," *BMC medical genomics*, vol. 4, no. 1, p. 8, 2011.

[22] R. Ge, M. Zhou, Y. Luo, Q. Meng, G. Mai, D. Ma, G. Wang, and F. Zhou, "Mctwo: a two-step feature selection algorithm based on maximal information coefficient," *BMC bioinformatics*, vol. 17, no. 1, p. 142, 2016.

[23] Y. Perez-Riverol, M. Kuhn, J. A. Vizcaíno, M.-P. Hitz, and E. Audain, "Accurate and fast feature selection workflow for high-dimensional omics data," *PloS one*, vol. 12, no. 12, p. e0189875, 2017.

[24] S. Kang and J. Song, "Robust gene selection methods using weighting schemes for microarray data analysis," *BMC bioinformatics*, vol. 18, no. 1, p. 389, 2017.

[25] B. L. Welch, "The generalization of 'student's' problem when several different population variances are involved," *Biometrika*, vol. 34, no. 1/2, pp. 28–35, 1947. [Online]. Available: http://www.jstor.org/stable/2332510

[26] J. Piñero, À. Bravo, N. Queralt-Rosinach, A. Gutiérrez-Sacristán, J. Deu-Pons, E. Centeno, J. García-García, F. Sanz, and L. I. Furlong, "Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants," *Nucleic acids research*, p. gkw943, 2016.