BMC Medical Genomics

RESEARCH

Open Access

Disease candidate genes prediction using positive labeled and unlabeled instances

Sepideh Molaei[1] and Saeed Jalili[1*]

Abstract

Identifying disease genes and understanding their performance is critical in producing drugs for genetic diseases. Nowadays, laboratory approaches are not only used for disease gene identification but also using computational approaches like machine learning are becoming considerable for this purpose. In machine learning methods, researchers can only use two data types (disease genes and unknown genes) to predict disease candidate genes. Notably, there is no source for the negative data set. The proposed method is a two-step process: The first step is the extraction of reliable negative genes from a set of unlabeled genes by one-class learning and a filter based on distance indicators from known disease genes; this step is performed separately for each disease. The second step is the learning of a binary model using causing genes of each disease as a positive learning set and the reliable negative genes extracted from that disease. Each gene in the unlabeled gene's production and ranking step is assigned a normalized score using two filters and a learned model. Consequently, disease genes are predicted and ranked. The proposed method evaluation of various six diseases and Cancer class indicates better results than other studies.

Keywords Disease gene prediction, Positive-unlabeled learning, Gene expression profile, Score relevance, Support vector machine

Introduction

Genes are the factors of inherited and genetic disorders that can path through into future generations. Also, they can be hidden and may be revealed in the future. Hence, genetic disease treatment or prevention has been challenging for physicians and health researchers from the past to now. Thereby, predicting disease genes and understanding their mechanisms is the first critical step in pharmacology and medicine for treatment and prevention. Today, new studies have significantly enhanced for finding the disease's molecular basis to prevent, diagnose, and treat genetic diseases.

The utilization of machine learning methods to solve various problems has shown promising performance compared to traditional and experimental methods. In particular, machine learning techniques in medicine have attracted significant attention. Experimental and laboratory-based methods for solving medical problems are often cost-intensive and time-consuming, which has led to a growing interest in computational methods, including machine learning. Furthermore, while some genes are classified as non-disease genes, they may be identified as disease-related in different contexts. This complexity has made it difficult to definitively classify non-disease genes, as knowledge in this area remains limited. However, recent studies have shown that some human genes play a role in diseases and can be valuable for predicting disease-related genes using machine learning methods.

In predicting and ranking disease genes using machine learning, the disease-known genes are considered a positive data set, and unknown genes are considered unlabeled genes. Prediction and labeling the genes causing a disease (based on ranking) among the unknown genes using that disease's known genes is the purpose of this

*Correspondence:
Saeed Jalili
sjalili@modares.ac.ir
[1] Computer Engineering Department, Tarbiat Modares University, Tehran, Iran

issue. Due to the data nature, one of the most proper solving methods of this issue (which consists of the data's nature) is the Positive Unlabeled Learning (PU-Learning) approach [1]. The PU-Learning method is semi-supervised; this method is used for binary classes with positive labeled and unlabeled samples. This type of learning has no negative labeled samples, and it distinguishes it from other learning types. The available data in this type of learning is as two following types: (*i*) data set including positive labeled samples; (*ii*) data set without label that potentially can be the cause of the disease (positive) or non-disease (negative). The studies regarding solving this issue with the PU-Learning approach are classified into two general approaches: 1) Identifying negative samples approach; 2) not identifying negative samples approach.

The negative genes (non-disease) are initially selected among the unlabeled genes in the identifying negative samples approach. Next, binary models are learned separately for each disease using data set containing genes causing that disease (with positive label) and non-disease genes (with negative genes). Selecting reliable negative genes is the main challenge in this strategy. The more reliable they are the learning will be accurate in the next step. In the not identifying negative samples approach, learning of one class is only carried out using positive samples. This method will be useful if the number of positive samples is ample and sufficient.

Moreover, the efficiency of this method is very low if the number of positive samples is insufficient [2] or the entire unlabeled genes consider negative samples. Consequently, the problem will be changed to an unbalanced binary classification, and then binary models will be learned. Since the dataset of unlabeled genes is included potential negative and positive samples, utilization of this method have high error. Recently, the use of this method has been reduced [3].

The extraction of reliable negative genes in the proposed method is as follows: negative genes extraction is carried out separately for each disease in the one-class learning step. Then, the most distant negative genes from known disease genes are selected. Indeed, designing reliable negative gene extraction in such a way will enhance the trust in extracted negative genes. Disease genes will be selected separately for each disease in the binary model learning step based on the proposed method's designed scoring system. The score-relevance indicator is used for this purpose. The score of each disease gene is normalized using a scoring system. Then, it is decided whether or not to select any disease gene as positive educational data based on the score of each gene. Eventually, a binary disease model is learned using the Support Vector Machine (SVM) algorithm. The other two filters are used in the unlabeled genes' prediction and ranking step after determining the sample's label using the learned binary model. These two filters are based on: 1) each gene's distance from the support vector; 2) the closeness of the gene to disease genes. A normalized score is laid out for each gene using the designed scoring system in the proposed method and the distance of every unlabeled gene from the disease binary model's support vectors. Next, another score is laid out for each gene using a designed scoring system and score relevance related to every unlabeled gene. Eventually, a single score for each gene is obtained by formulating scores. Then, the decision is made for the unlabeled gene (in other words, whether the gene is a candidate for the disease or not). Besides, the rank of the gene is determined if it is a candidate for the disease. The outcomes of evaluating the proposed method compared to the best previous available proposed method are as follows:

The recall measure of Adrenal, Colon, Lung, Prostate, and Heart Failure diseases and Cancer disease class are increased by 0.53%, 5.32%, 1.29%, 3.33%, 4.04%, and 3.11%, respectively. Moreover, the increase of precision measure is 2.64%, 2.14%, 1.75%, 3.14%, 3.13%, and 2.38%, respectively. The increase of AUC measure for Neurological disease is 8.82% compared to other studies.

Basic concepts

Gene expression profile (GEP)

Gene expression data provides valuable information regarding cellular situations, biological networks, and understanding of genes' performance. Indeed, the genetic codes have been stored in DNA strands. Furthermore, they will interpret by gene expression. Determining how genes are expressed in non-disease and diseased cells is one of the purposes of gene expression interpretation. Scientists utilize DNA microarray (biochips) to measure gene expression amount. A set of gene expression samples is the result of determining the gene expression amount's experiment. Every row in the gene expression matrix indicates the related gene expression profile. Time series of gene expression profiles (which state the gene expression level in determining periods) are used in this study.

Similarity-based communication principle

Similarity-based communication principle is used in most disease candidate gene prediction problem-solving methods. The mentioned principle declares that the greater the physical and performance similarity of genes, the greater the probability of their role in developing the same diseases. The closeness amount to the disease genes can be used as a rank.

Score relevance

The scores for each gene based on Score-Relevance can be considered a score for the effectiveness of that gene in the specific disease formation. Indeed, the mentioned scores are based on the simultaneous presence of two elements in the Medline[1] document. This score is based on a formula (the base of this formula is the Boleyn model) and is calculated for finding coincident documents and their conformity amount. Overall, the mentioned formula has used the concepts of Term Frequency-Inverse Document Frequency (TF-IDF), Vector Space, Coordination factor, and field length normalization [4].

Comparing the number of documents in which two elements are present next to each other and the number of documents in which elements independently appear with the expected amount is carried out based on the hypergeometric distribution. The greater the simultaneous presence of elements (compared to the expected amount) will reduce the random occurrence of this happen. Consequently, the scores will enhance [5]. Unfortunately, these scores are not significant absolutely and only are sequentially significant in the related genes list of each disease and have particular importance. Moreover, the absolute amounts of scores may vary from one version to another version.

Research history

The previous studies regarding disease candidate gene prediction are introduced in two groups.

Identifying negative samples approach

Yousef and Moghadam [6] used proteins' amino acid sequences for predicting and ranking the diseases' genes. They construct four various characteristic vectors using amino acid sequences. Moreover, they use cosine distance for extracting reliable negative genes. Eventually, the characteristics of a model are learned separately for each vector. The results of every category are integrated, and the final result will be announced.

VasighiZaker and Jalili [7] presented the C-PUGP method. In this method, the clustering of positive samples is considered initially. Next, a one-class model with an OCSVM learning algorithm is carried out for every cluster. Labeling of unlabeled samples is performed using learned models. Then, the unlabeled gene, which gives a negative label based on the entire one-class models, is considered a reliable negative sample. Finally, the SVM binary model is learned using the obtained negative samples and initial positive samples. Many initial studies considered the entire unlabeled genes as negative samples and learned a binary model. Since the dataset of unlabeled genes is included negative and positive samples, utilization of this method have high error. Smalter et al. [8] predicted disease candidate genes using the protein–protein interaction dataset and SVM binary model. Radivojac et al. [9] used three various datasets and learned an SVM binary model for every dataset. They identify disease candidate genes using these three disease binary models' results. The used datasets were protein sequences, protein performance information, and the PPI network.

Not identifying negative samples approach

Learning is carried out only with positive samples in this method. The efficiency of this method is very low if the number of positive samples is insufficient [2]. Yousef and Moghadam [10] identified disease genes using the SVDD one-class model (only by using the sequences of disease genes). This method generates the characteristic vector by converting protein consequences to numerical vectors using their physicochemical properties translation. Then, they reduced the characteristics sizes to find the critical characteristics using Principal Component Analysis (PCA). The disease genes (positive samples) are learned using SVDD one-class model in the next step. The unlabeled samples will predict using the learned model. The entire disease genes are initially considered a positive set in the method of VasighiZaker and Jalili [11]. This set will normalize by the Min–Max method. Then, the number of the characteristic will reduce using the PCA method. Next, the learning is performed by OCSVM one-class model. The unlabeled genes are labeled after finding the optimal parameters. Nikdel and Jalili [12] studied the clustering of disease genes based on a constructed matrix by measuring semantic similarity among the disease types; this is carried out based on the gene ontology. Next, the Hidden Markov Model (HMM) is learned for each cluster; a threshold is calculated for each cluster separately. The unlabeled genes are given to the entire learned hidden Markov models of that disease. The label of that gene will identify given the probability obtained from each hidden Markov model and calculated threshold for each cluster. In other words, if at least one of the hidden Markov models (among the entire learned hidden Markov models of that disease) considers an unlabeled gene as a disease candidate, the positive label is attributed to that gene. After normalizing gene expression data, Vasighizaker et al. [13] used a one-class support vector machine model with a linear kernel for predicting disease genes in Acute Myeloid Leukemia (AML) cancer.

The proposed method

The scoring-based method using the SVM binary model is introduced to solve the prediction and ranking problem of disease candidate genes; this method scores

[1] *It is one of the most famous free databases worldwide and includes bibliographic research information for the entire medical and biology fields.*

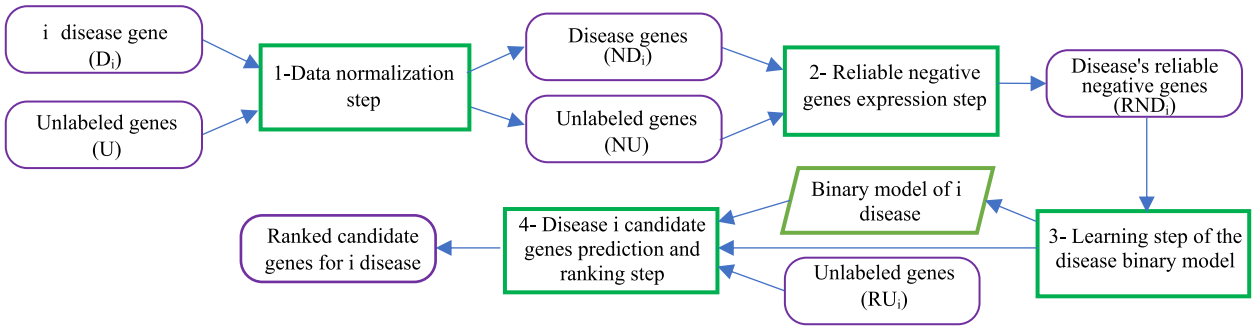

Fig. 1 S-PUL proposed method process

effective factors in predicting and ranking disease candidate genes. The main aim of this method is disease candidate genes prediction and ranking from an unlabeled gene set. The higher priority belongs to the gene more likely to belong to the disease candidate genes group. Unlabeled genes are human genome that does not belong to disease genes. Notably, determining gene expression is performed in various laboratories. Therefore, a gene may have more than one gene profile. Consequently, the entire calculation is carried out separately for a gene's profile.

The S-PUL[2] proposed method has four following steps: 1) data normalization; 2) reliable negative genes extraction; 3) disease binary model learning; 4) disease candidate genes prediction and ranking (see Fig. 1). The gene expression data is normalized in the first step. In the second step, reliable negative genes are extracted from unlabeled samples separately for every disease. The binary model is learned separately for every disease with positive samples (disease genes) in the third step. In the fourth step, reliable negative genes are eliminated from unlabeled genes (U). Then, the remaining unlabeled gene set (Rui) is given to the disease binary model for label prediction.

The term "S-PUL" stands for Scored-Positive Unlabeled Learning. It is a combination of two used methods: Positive Unlabeled Learning (PUL) and a Scoring system. The scoring aspect refers to the integration of a scoring system within the Support Vector Machine (SVM) algorithm. This hybrid approach leverages the strengths of both techniques to enhance the learning process.

Data normalization step

Each gene's time expression range is different, and their difference is high. The entire data is normalized separately for two datasets (disease and unlabeled genes). The normalization is carried out based on Eq. 1. The highest and lowest amounts of every gene's time expression are indicated by X_{max} and X_{min} in Eq. 1, respectively.

$$X_{normalized} = \frac{(X_{max} - X)}{(X_{max} - X_{min})} \quad (1)$$

Reliable negative genes selection step

Learning disease binary model, in addition to disease genes set (as positive samples), requires reliable negative genes set (as negative samples). It is evident that the accuracy of predicting unlabeled genes by the disease binary model (as disease genes) increases with enhancing the trust degree in the identified negative genes (among unlabeled genes). Figure 2 illustrates the reliable negative gene extraction process related to each disease class.

In the first action (i.e., Action 1 Algorithm), the Robust Gaussian, KNN, Parzen window, and SVDD one-class classification algorithms are used for learning the positive samples model separately for each disease class. Moreover, other disease classes' genes (after eliminating common genes) are used as test data. After learning a disease model, other diseases genes are expected to appear in the negative data role. Hence, the evaluation indicator to select the best learning algorithm is the percentage of considered accurate negative samples. Eventually, the learned one-class algorithm that has the highest percentage of accurate negative samples is selected as the best one-class model of i-th disease. In the Action2, unlabeled genes are given to the best one-class model as input, and unlabeled genes are labeled. The outcome of this step is a set of negative genes. Finally, Reliable negative genes are selected from the set of negative genes in the third step (i.e., Filter1 algorihtm.). The shortest Euclidean distance of every negative gene is calculated from its correspondent disease genes. If a disease gene expression profile (Ne) from the ND_i set is shown by $Ne = \{d_1, d_2, d_3, \ldots, d_m\}$ and the negative gene expression profile (Ng_i) is shown by $Ng_i = \{n_1, n_2, n_3, \ldots, n_m\}$, the Euclidean distance is calculated using Eq. 2. Thus, the minimum distance of every negative gene from its correspondent disease genes is calculated based on Eq. 3. Eventually, the farthest genes from disease genes are selected as reliable negative genes for every disease i (RND_i).

[2] Scored-Positive Unlabeled Learning.

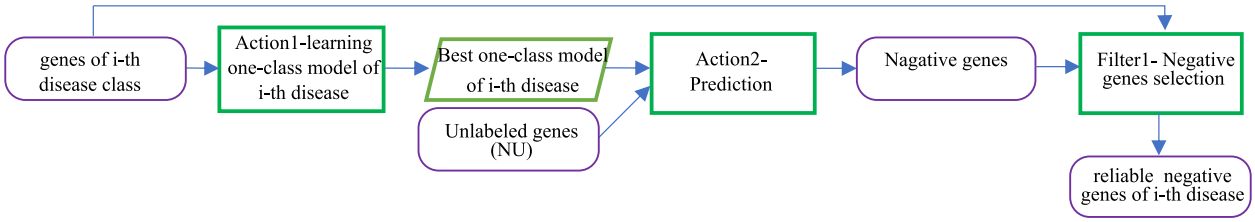

Fig. 2 The reliable negative genes extraction process

$$Dis_{Eu(Ne,Ng_i)} = \sqrt{\sum_{k=1}^{m} (d_k - n_k)^2} \quad (2)$$

$$Ne_k = \min_{\forall Ne \in NDi} \{Dis\_Eu(Ne, Ng_i)\} \quad (3)$$

It is worth noting that the role of genes in the arising of disease has different degrees. The reliability of learning results will enhance using genes (as training data) that have higher correspondent S-R[3] values in the learning process. The selection of disease genes is performed using S-R for a positive training set in this study. The value of S-R related to every disease gene (separately for each disease) is available in [4].

Action1 Algorithm (Learning one-class model of i-th Diseases)

Input: $D = \{g_1, g_2, .., g_n\}$ // *a disease class where g_i are genes.*
$D\text{-set}=\{D_2, D_2.., D_k\}$ // *$D_i=\{g_{i,1}, g_{i,2}, ..., g_{i,|Di|}\}$ is the gene set of i-th test disease classes, K is the no. of test disease classes, $g_{i,j}$ shows the j-th gene of D_i and $|P|$ shows the no. of elements of P set*
$A\text{-set} = \{A_1, A_2, A_3, A_4\}$ // *A_1= "Robust Gaussian", A_2= "KNN", A_3= "Parzen Window," A_4="SVDD" learning algorithms*

Output: *One-class model* // *best one-class model of disease class D*

//Learning models using all learning algorithms
1. *Initialize parameters of A_i learning algorithms*
2. *For each $A_i \in$ A-set*
3. *Train a one-class model M_i using algorithm A_i and save M_i in M-set where M-set = $\{M_1, M_2, M_3, M_4\}$*
End of For loop (step 2)
//Testing all learned models Mi on genes from other disease classes, D-set
// Exclude common genes of disease class D and all disease classes of D-set
4. *Compute $E = (\cup_{i=1}^{k} Di \in D\_set)$, $F = (E \cap D)$, $G = (E - F)$*
5. *For i = 1 to 4:*
6. *Test $M_i \in$ M-set on genes of G set (derived in step 4) and save their labels in $L_i \in$ L-set*
// L-set = $\{L_1, L_2, L_3, L_4\}$ where $L_i = \{l\text{-}g_{i,1}, l\text{-}g_{i,2}, .., l\text{-}g_{i,|G|}\}$ and $l\text{-}g_{ij}$ is the label of $g_j \in G$ given by $M_i \in$ M-set and labels are 0 means (positive label) or 1 (means negative label)
End of For loop (Step 5).
// Evaluating the performance of one-class classification methods of A-set
7. *A = |G|, Max = 0 , I = 0 // Variable A shows the number of genes of G set*
8. *For i = 1 to 4:*
9. $B = \sum_{r=1}^{|G|} l\_gi_r$ *Where* $l_{g_i, r} \in Li$
10. $TNG_i = \left(\frac{A-B}{A}\right) * 100$ // *TNG-set = $\{TNG_1, TNG_2, TNG_3, TNG_4\}$*
11. *If TNG_i > Max then {I = i, Max = TNG_i}*
End of For loop (Step 8)
12. *One-class model = MI // best learned model that is learned by A_i algorithm $\in$ A-set*

Learning step of the disease binary model

The prediction and ranking problem of disease candidate genes are solved based on binary model learning. Figure 3 indicates the learning process of the disease binary model. Selecting the positive training data from the disease genes set of every disease is another challenge of this study.

Positive genes selection (Filter 2)

Positive genes of each disease are selected in four steps. This process is described step by step in the following and presented formally by "Filter 2 algorithm".

In the first step, disease class genes are categorized based on their S-R values (separately for each disease). The gene

[3] Score Relevance (explained in "Score relevance" section).

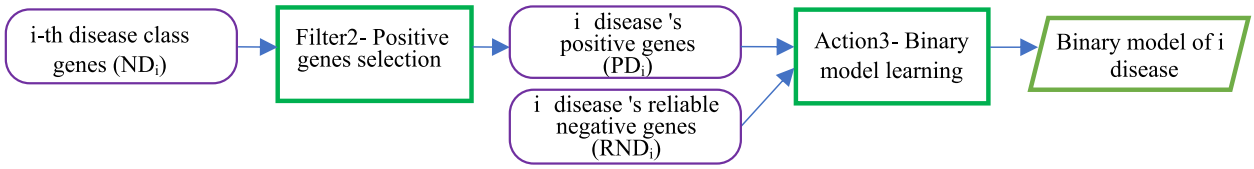

Fig. 3 Learning process of disease binary model

will belong to a higher category by enhancing its S-R value, thus obtaining a higher score. The categories with equal intervals of ten units will create for categorizing genes based on their S-R value. Therefore, the first category is related to the genes whose range is [0,10). In other words, the first category has the lowest value. Accordingly, each gene will belong to a category (the length of these categories is 10). One of the challenges of this study is determining the value of these categories' range. The distribution of disease genes number based on the S-R values is not uniform. Each category's range should be determined so that it does not lead to the over-elimination of genes. The length of 10 for categories is a logical number for the entire disease. The mentioned length has been obtained by trial and error in this study. Moreover, this number can be calculated more accurately in future studies.

In the second step, every category gets a portion of 100 points according to its obtained score. In other words, the highest percentage will obtain by the highest category. The category score related to the i-th category is shown by NGr_i; this reaches the base of 100. The NGr_i can be calculated by Eq. 4.

$$NGr_i = \left(\left\lfloor\left(\frac{SR_i}{10}\right)\right\rfloor + 1\right) \times \frac{200}{\text{Max}\{|Gr|\}(\text{Max}\{|Gr|\}+1)} \tag{4}$$

The category score of the entire genes belonging to the disease is saved in Gr set. In Eq. 4, Max(|Gr|) is the highest category score of a gene belong to a disease class; $S\text{-}R_i$ indicates the S-R value related to the i-th gene of Gr set.

In the third step, the final score of the i-th gene F_Score_i is calculated by Eq. 5.

$$F\_Score_i = NGr_i \times S - R_i \tag{5}$$

The mean scores range ($\overline{IL}$) is calculated based on the Eq. 6 (separately to every disease). Moreover, some genes are selected as positive training data (their final score is over the mean of the scores range). In Eq. 6, the final score of the entire genes is in the {F_Score} set.

$$\overline{IL} = \frac{Max\{F\_Score\} + Min\{F\_Score\}}{2} \tag{6}$$

Filter2 Algorithm (Positive genes selection)

Input: $ND_i = \{(g_1, s_1), (g_2, s_2), \ldots, (g_n, s_n)\}$ // ND_i is i-th disease class where in pair (g_i, s_i), g_i are genes and s_i are correspondent S-R value of g_i genes
Output: $PD_i = \{g_1, \ldots, g_k\}$ // Selected positive genes, initially $PD_i = \emptyset$

//Categorize genes based on their S-R values
1. For each $(g_i, s_i) \in ND_i$
2. $j = int\ (S\text{-}R_i / 10) + 1$ // j shows gene g_i belongs to j-th category
3. Append $((g_i, g_i\text{-category}), categories)$ // categories = $\{(g_1, g_1\text{-category}), \ldots, (g_n, g_n\text{-category})\}$
End of For loop (Step 1)
// Calculate category scores of genes
4. For each $(g_i, J) \in categories$ // J shows g_i-category
5. Calculate NGr_j based on Equation 4
6. Append $((g_i, NGr_j), Gr)$ // Gr set contains the category score of all genes of ND_i set and initially $Gr = \emptyset$; $Gr = \{(g_1, g_1\text{-NGr}), \ldots, (g_n, g_n\text{-NGr})\}$
End of For loop (Step 4)
// Calculate final Score for each gene in ND_i set
7. For each $(g_i, s_i) \in ND_i$
8. Find $(g_i, k) \in Gr$ // k shows NGr_j
9. Calculate $F\_Score_i = k * S_i$
10. append $((g_i, F\_Score_i), FS)$ //FS set contains F-Score of all genes of ND_i set and initially $FS = \emptyset$
End of For loop (Step 7)
//Select Positive genes of ND_i disease
11. Compute IL based on Equation 6
12. For each $(g_i, F\_Score_i) \in FS$
13. If $F\text{-}Score_i > IL$ then append (g_i, PD_i)
End of For loop (Step 12)
14. return PD_i

Positive genes selection (Filter3)

Filter 3 is an optimization step in the proposed method designed to eliminate low-significance genes and reduce noise in the data. Specifically, this filter removes genes that received a negative label from the SVM binary model during the learning phase, and have S-R values in the lowest scoring range ([0, 10)).

The primary goal of this filter is to focus the learning process on genes that are more likely associated with the disease, while excluding genes that have the least impact on disease formation. By doing so, the learning process is refined, and it is expected that the prediction accuracy for disease candidate genes will improve.

Binary model learning

In Action3 of Fig. 3, the binary learning using binary learning algorithms is performed using selected positive training genes (PD_i) from i-th disease genes and reliable negative genes (RND_i) from unlabeled genes. Eventually, the algorithm that obtains the highest recall evaluation value for all diseases is selected.

Disease candidate genes prediction and ranking step

The remaining unlabeled gene sets (i.e., the unlabeled genes set that the extracted negative genes are eliminated in that set in Reliable negative genes selection step) are given to the disease binary model as test data after learning and selecting the best binary learning algorithm (SVM) with having the best learning parameters. A scoring algorithm is also used in the disease candidate prediction and ranking step, as illustrated in Fig. 4. There are two critical factors in the scoring algorithm: 1) The distance of every unlabeled gene from the disease gene; 2) The distance of every unlabeled gene from the support-vector of the i disease model. genes give a score based on each mentioned factor. The final score of the gene will obtain by multiplying these two scores. Eventually, the prediction and ranking are carried out according to the final score.

Action4- Identifying the valuable genes

The unlabeled genes given to the i disease (i.e.; the extracted reliable negative genes of i-th disease eliminated from the unlabeled genes set; RUi indicates this set) are labeled and stored in the DS1 set using the i-th disease learning model. Suppose that the expression profile of disease gen (Ne) from the ND_i set is $Ne = \{d_1, d_2, \ldots, d_m\}$, and the expression profile of an unlabeled gene (Ru) from the RU_i set is $Ru = \{u_1, u_2, \ldots, u_m\}$. The closet i disease gene (Ne) to each Ru studied expression profile from the RU_i dataset is identified using Eqs. 2 and 3 (in terms of Euclidean distance). Moreover, it is stored in the DS2 set. Negatively labeled genes are eliminated from the DS1 dataset to preserve valuable genes (separately to each profile). Their correspondent S-R values are settled in the DS2 dataset of the first category (the least valuable category). The remaining genes are stored in the VRU_i dataset. These remaining genes are negatively labeled genes with high S-R and positive labeled genes that are valuable genes.

Action5- The prediction and ranking of disease candidate genes

The F_Score value of the nearest disease gene is attributed to Ru studied gene profile from the VRU_i dataset. It is worth noting that the nearest disease gene to each studied gene profile is identified in the "Reliable negative genes selection step" section and maintained in the DS2 dataset. In this method, the given label to each Ru studied gene is maintained from the VRU_i dataset. Each gene has many gene profiles. Thus, the final score of a gene is the algebraic summation of scores of that gene's profiles. The output of this step is the DS3 dataset, which contains the entire valuable genes input from the VRU_i to this step, along with the second score of each gene (DP_Score_i). It is worth noting that the reliability of the sample belonging to the i disease class increases with enhancing the distance of the tested sample from the support vectors of the i disease model. In contrast, the reliability of the sample belonging to the i disease class reduces by reducing

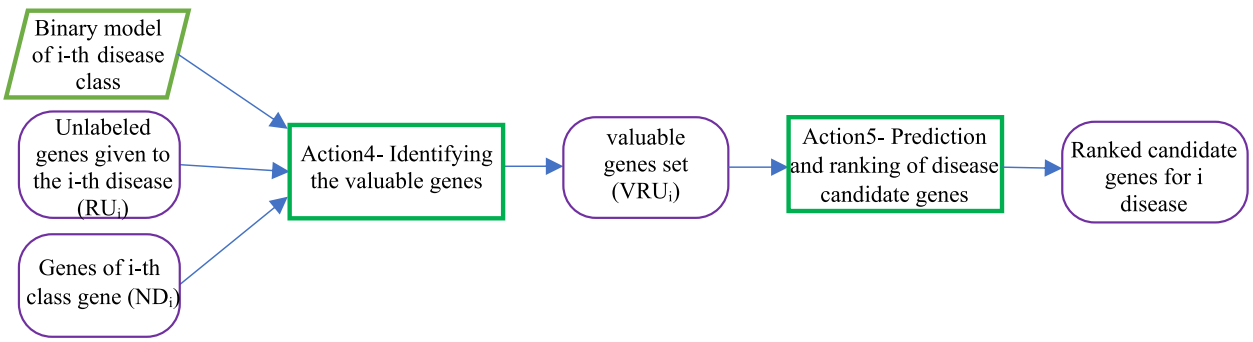

Fig. 4 The prediction and ranking process of disease candidate genes

the distance of the tested sample from the support vectors of the i disease model. Consequently, the gene score will increase by distancing the studied gene (VRU_i) from the support vectors of i disease in the calculation of the second score of each gene (DS_Score_i). The calculation of the third score is carried out in three steps.

In the first step, the value of Gr_{sv} parameter for i-th gene from positive and negative labeled genes considered $\lfloor DS_i \rfloor + 1$ value and $\lfloor DS_i \rfloor$ value, respectively. Gr_{sv} is the category's score, including the ith gene, and DS_i is the distance of the i-th gene from the Support Vector.

In the second step, the Eq. 7 is used to calculate the value of NGr_{svi}; it is the category's score of the i-th gene (Gr_{svi}). The mentioned score has reached the base of 100. The category's score of all genes belonging to the disease is in the $\{Gr_{sv}\}$ set.

$$NGr_{svi} = Gr_{svi} \times \frac{200}{\text{Max}\{|Gr_{sv}|\}(\text{Max}\{|Gr_{sv}|\}+1)} \quad (7)$$

In the third step, the final correspondent score with the i-th gene DS_Score_i is calculated by Eq. 8.

$$DS\_Score_i = NGr_{svi} \times |DS_i| \quad (8)$$

The second and third scores are simultaneously used for predicting and ranking disease candidate genes. Each gene may have several profiles in the unlabeled genes dataset. Thus, each gene obtains a score based on its profile number. The final score for that gene is obtained from the algebraic summation of gene profiles.

The final score of the studied gene ($Final\_Score_i$) is calculated based on the Eq. 9 with the algebraic summation of gene profiles' scores (DP_Score_i and DS_Score_i for each profile of that gene). The prediction of disease candidate genes is carried out based on the score of each gene.

$$Final\_Score_i = \sum_{i=1}^{m} (DS\_Score_i \times |DP\_Score_i|) \quad (9)$$

The number of gene profiles is indicated by m in Eq. 9.

Finally, genes whose *Final_Score* values are negative will eliminate; other genes are predicted as disease candidate genes. The obtained final score of each disease candidate gene is used for ranking.

Results

The efficiency of the S-PUL method is evaluated in six versions, namely S-PUL_Vn in this section. The number of S-PUL versions and used filters in that version are reported in Table 1. It is worth noting that the version of S-PUL_V5 is the proposed S-PUL method, which uses all filters.

Table 1 The used filters in the versions of the S-PUL proposed method

S-PUL_ Version	Filter 1	Filter 2	Filter 3	Second Score	Third Score
S-PUL_V0	✓				
S-PUL_V1	✓	✓			
S-PUL_V2	✓	✓	✓		
S-PUL_V3	✓	✓	✓	✓	
S-PUL_V4	✓	✓	✓		✓
S-PUL_V5	✓	✓	✓	✓	✓

The efficiency of the S-PUL method results (separately for each version) is compared with the previous studies. Finally, this method's efficiency is evaluated separately in 2016 and 2020 using the newly identified disease genes. The MATLAB Software (2019 version) is used for learning binary classification and calculation. Moreover, the dd-tools library is used for learning one-class models. The entire evaluation is carried out on a computer with an Intel Core TM i5 processor and main memory of 32 GB in Windows 10 pro.

Dataset

The used genes in the learning and testing phases are extracted from the dataset of Yang et al. (2014) [14] (the second row of Table 2). The dataset for the Cancer disease class has 210 genes, these 210 genes are common among the three diseases: colon, prostate, and lung (the number of disease genes is provided in Table 8), and the dataset of unlabeled genes has 12,001 genes. GeneCards [4] (the third and fourth rows of Table 2) are used for the dataset of disease genes from 2015 to 2020. The characteristics of disease genes are represented in Table 2 separately for each disease and period. Notably, each disease gene may be the cause of several diseases.

Evaluation measures

The accuracy, recall F_1, and AUC measures (area under the ROC curve, which is the changes of TPR to FPR) are used to evaluate the S-PUL method). The mentioned measures are defined in Table 3. In these equations: the TP parameter is the number of positive samples that are categorized correctly; the TN parameter is the number of negative samples that are categorized correctly; the FP is the number of negative samples that are categorized as positive incorrectly (in other words, the number of spurious positive samples); the FN is the number of positive samples that are categorized as negative incorrectly (in other words, the number of spurious negative samples); the TPR is the correct positive rate; FPR is the spurious positive rate. This study has considered disease genes as positive samples and extracted negative samples from

Table 2 Characteristics of genes expression profile datasets

The name of the disease class	Cancer			Endocrine	Cardiovascular	Neurological	
The name of the disease	Colon	Lung	Prostate	Adrenal	Heart Failure	Neurological	Row number
Sequence length of gene expression profiles	18	18	5	37	9	42	1
Biologists [14] Number of disease genes by 2014	342	245	325	81	107	219	2
Biologists [4] Number of new disease genes from 2015 to 2016	240	-	191	9	-	-	3
Biologists [4] Number of new disease genes from 2017 to 2020	56	27	67	29	58	16	4
Total known disease genes	638	272	583	119	165	235	5

unlabeled genes. All evaluations are performed with K-fold C.V and k = 10.

The evaluation of extracted reliable negative genes

The extraction of reliable negative genes is carried out in two steps (extraction of negative genes using a one-class learning algorithm and selecting reliable negative genes using distance measure). The quality of reliable negative genes extraction is evaluated at each step.

Selecting the one-class learning algorithm

Negative genes are initially extracted separately for each disease to select the one-class learning algorithm and each one-class classification learning algorithm of SVDD, Robust Gaussian, KNN, and Parzen Window (the first and second steps in Reliable negative genes selection step). Each algorithm's parameters and a brief introduction are provided below, with references for detailed explanations.

Table 3 The relations of evaluation measures

Equation number	Equation	Equation number	Equation
18	$Precision = \frac{TP}{TP+FP} * 100$	**19**	$Recall = \frac{TP}{TP+FN} * 100$
20	$F_1 = \frac{2*P*R}{P+R}$	**21**	$TNG = \frac{A-B}{A} * 100$
22	$TPR = \frac{TP}{TP+FN}$	**23**	$FPR = \frac{FP}{FP+TN}$

A. Precision measure – this measure indicates the percentage of positive predictions that are performed correctly. Moreover, this measure is calculated from Eq. 18 in Table 3

B. Recall measure – this measure indicates the percentage of positive samples that are categorized correctly. Moreover, this measure is calculated from Eq. 19 in Table 3

C. F_1 measure – it is a compatible mean between precision and recall. Moreover, this measure is calculated from Eq. 20 in Table 3. Additionally, R symbol refers to recall and P symbol refers to precision

D. Trust in Negative Genes (TNG) measure – this measure is used for trust value in extracted negative genes measurement in unlabeled genes. Indeed, it compares the extracted negative genes and disease-known genes from 2014 to 2020. The trust value in negative genes is calculated from Eq. 21 in Table 3. In Eq. 21, A is the number of extracted negative genes, the B parameter is the number of available common genes in the disease genes list from 2014 to 2020 and extracted negative genes

Support Vector Data Description (SVDD) is a machine learning algorithm used for anomaly detection and classification. It constructs a sphere in the feature space that encompasses the training data. The parameter used in this algorithm is the width parameter in the RBF kernel [15].

Robust Gaussian is an algorithm that models data distribution as a Gaussian distribution and employs robust statistics to handle outliers. The parameter for this algorithm is the error tolerance on the mean and covariance matrix [16].

K-Nearest Neighbors (KNN) is an instance-based algorithm that classifies data based on the distances to the k-nearest neighbors. The parameter for this algorithm is the number of neighbors [17].

Parzen Window is a non-parametric method for estimating the probability density function of a dataset using kernel functions. The parameter for this algorithm is the width parameter [18].

Then, each one-class learning algorithm's efficiency is examined through two evaluation methods.

The first evaluation method: The percentage of correct negative samples (%TN) is considered the evaluation method in the first method. The selected parameters of each one-class learning algorithm are presented in Table 4. The error value on the target class (Fracrej) parameter considers 0.1 for all one-class learning algorithms. The efficiency results of each one-class learning algorithm are reported in Table 5.

The first evaluation method

The percentage of correct negative samples (%TN) is considered the evaluation method in the first method. The selected parameters of each one-class learning algorithm are presented in Table 4. The error value on the target class (Fracrej) parameter considers 0.1 for all one-class learning algorithms. The efficiency results of each one-class learning algorithm are reported in Table 5.

Table 4 The selected parameters for each one-class learning algorithm separately for each disease

The name of the disease ↓	Algorithm →	SVDD		KNN	Robust Gaussian	ParzenWindow
	Parameter →	γ^4	kernel	^{3}K	^{2}Tol	^{1}h
Adrenal		0.14	RBF	2	e-3	1
Colon		0.14	RBF	1	0.015	1.2
Lung		0.3	RBF	2	e-3	1
Prostate		0.2	RBF	1	e-3	1
Heart Failure		0.14	RBF	2	e-3	0.9
Neurological		0.14	RBF	2	0.005	1

[1] Width parameter

[2] Error tolerance on mean and cov. Matrix

[3] Number of neighbors

[4] Width parameter in the RBF kernel

Table 5 The results of one-class learning algorithms efficiency evaluation in the extraction of reliable negative genes based on the percentage of correct negative samples measure (%TN)

The name of the disease ↓	Algorithm →	SVDD	KNN	Robust Gaussian	Parzen Window
Adrenal		**89.5%**	73%	74.8%	52.7%
Colon		**82.1%**	75.8%	12.6%	35.1%
Lung		**84%**	80%	19.4%	64%
Prostate		**82.6%**	69.3%	71%	72%
Heart Failure		**78.4%**	75.7%	37%	42.8%
Neurological		**71%**	42.62%	49.5%	68.7%

The highest efficiency of the one-class learning algorithm is related to SVDD. SVDD labeled the most percentage of negative samples for the entire types of diseases.

The second evaluation method

In this evaluation method, the efficiency of the S-PUL_V5 learning method is learned using considered positive disease genes separately for each extracted reliable negative genes set for each one-class learning algorithm (Reliable negative genes selection step). The results of this evaluation are illustrated in Fig. (5a to f) for each disease separately. It is worth noting that the number of selected reliable negative genes for each disease is equal to the number of disease genes; this prevents the unbalanced problem of positive and negative data).

According to Fig. 5, the highest efficiency is related to the S-PUL_V5 method if reliable negative genes are extracted using the SVDD method (compared to the other three one-class learning algorithms).

Measuring the trust degree in the extracted negative genes

According to Eq. 21, the trust degree in negative genes is extracted by the SVDD algorithm for each disease separately (see Table 6); it demonstrates that the reliability of extracted negative genes by the SVDD algorithm is high.

Evaluation of the binary classification algorithms performance and selection of the disease genes

Table 7 reports the parametrization for learning algorithm and disease separately. Moreover, Table 8 presents used disease genes information in the binary models learning for each disease individually.

The efficiency evaluation of five binary classification algorithm results is illustrated in Fig. 6 in the filtered/unfiltered status of disease genes.

Figure 6 indicates the evaluation results. The recall measure increased using S-PUL-V1 compared to the efficiency evaluation of S-PUL_V0 classification algorithms. The precision and, subsequently, the F_1 measures increase if it does not affect the recall measure. Hence, the filtering disease genes method in the S-PUL method will be used. Furthermore, the efficiency of the SVM binary model learning algorithm is more than other algorithms (see Fig. 6). Hence, the SVM binary classification method in the S-PUL method will be used. Table 9 indicates the value of the used parameter in the SVM learning algorithm. If the kernel is a

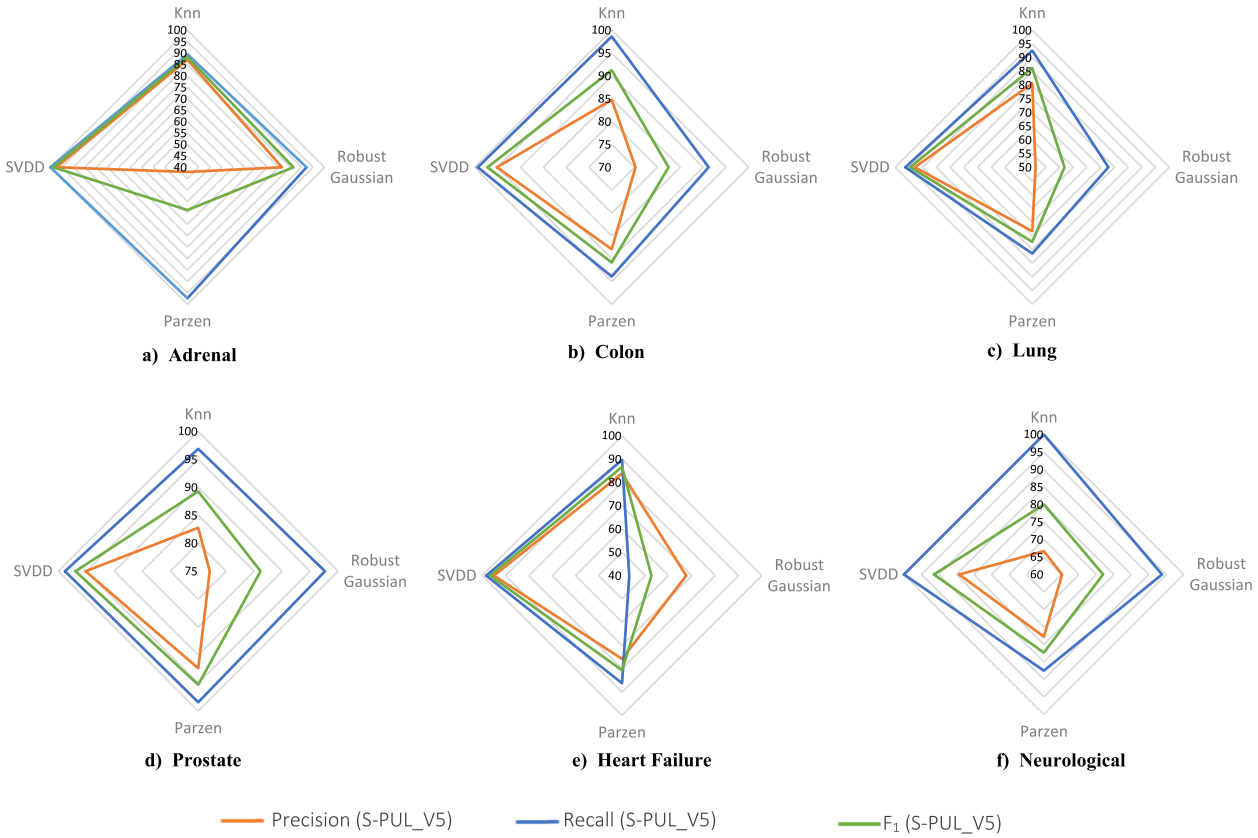

Fig. 5 The results of one-class learning algorithm efficiency evaluation in the extraction of reliable negative genes based on the efficiency of the S-PUL_V5 method

Table 6 The trust degree in extracted negative genes (TNG) using the SVDD method

The name of the disease	Parameter B[2]	Parameter A[1]	TNG
Adrenal	0	77	100%
Colon	2	323	99.38%
Lung	12	191	93.71%
Prostate	16	268	94.02%
Heart Failure	2	101	98.01%
Neurological	7	151	95.36%

[1] Parameter A represents the number of extracted negative genes

[2] Parameter B represents the number of common samples between the two sets of disease genes from the years 2014 to 2020 and the set of extracted negative genes

quadratic function, γ parameter is set to the one divided by the number of features (1/ number of features).

The evaluation of disease candidate genes prediction and ranking

The efficiency of disease candidate genes prediction and ranking is examined in this section for implementing filter 3, using the second and third scores separately.

The evaluation of selecting valuable genes efficiency (filter 3)

This section assesses the elimination of genes given a negative label by the SVM binary learning algorithm, and their S-R is in the first category ([0,10) range). The statistics of eliminated genes based on their related S-R range are presented in Table 10 for each disease.

Figure 7 illustrates the evaluation results of the S-PUL_V2 version (by implementing filter 3) compared to the S-PUL_V1 version (without implementing filter 3) to evaluate the filter 3 implementation value.

According to Fig. 7, by implementing filter 3, the recall measure of S-PUL_V2 increases in all diseases. Contrary, without implementing filter 3, the recall measure reduces in all diseases. The highest and lowest increase in recall measures in S-PUL_V2 is related to Lung disease (7.40%) and Colon disease (0.67%), respectively. Therefore, filter 3 will be used in the S-PUL method.

The efficiency evaluation of utilizing the second genes of the VRU set

Every gene has several gene expression profiles. Thus, in the S-PUL_V1, the gene will be considered a disease candidate gene if at least one of its gene expression profiles

Table 7 The values of parameters for learning algorithm and disease separately

The name of the disease ↓	Algorithm →	Logistic Regression		KNN	Decision Tree	Discriminative	
	Parameter →	Distribution	Size	K	ɵ	Type	δ
Adrenal		Binomial	1	1	0.25	Quadratic	0
Colon		Binomial	1	1	0.25	Linear	0.11
Lung		Binomial	1	2	0.32	Quadratic	0
Prostate		Binomial	1	1	0.25	Linear	0.14
Heart Failure		Binomial	1	2	0.24	Quadratic	0
Neurological		Binomial	1	2	0.30	Quadratic	0

Table 8 Disease genes information for each disease

The name of the disease	Number of disease genes before filtering	S-R range	S-R rating range	S-R score threshold	Number of disease genes after filtering
Adrenal	81	[0.08 , 108.21]	[0.12 , 1803.5]	901.68	77
Colon	342	[0.14 , 225.43]	[0.05 , 1878.583]	939.26	323
Lung	245	[0.1 , 216.63]	[0.03 , 1883.739]	941.84	191
Prostate	325	[0.1 , 251.1]	[0.02 , 1860]	929.98	268
Heart Failure	107	[0.12 , 109.44]	[0.18 , 1824]	911.90	101
Neurological	219	[0.07 , 42.06]	[0.46 , 1402]	700.76	151

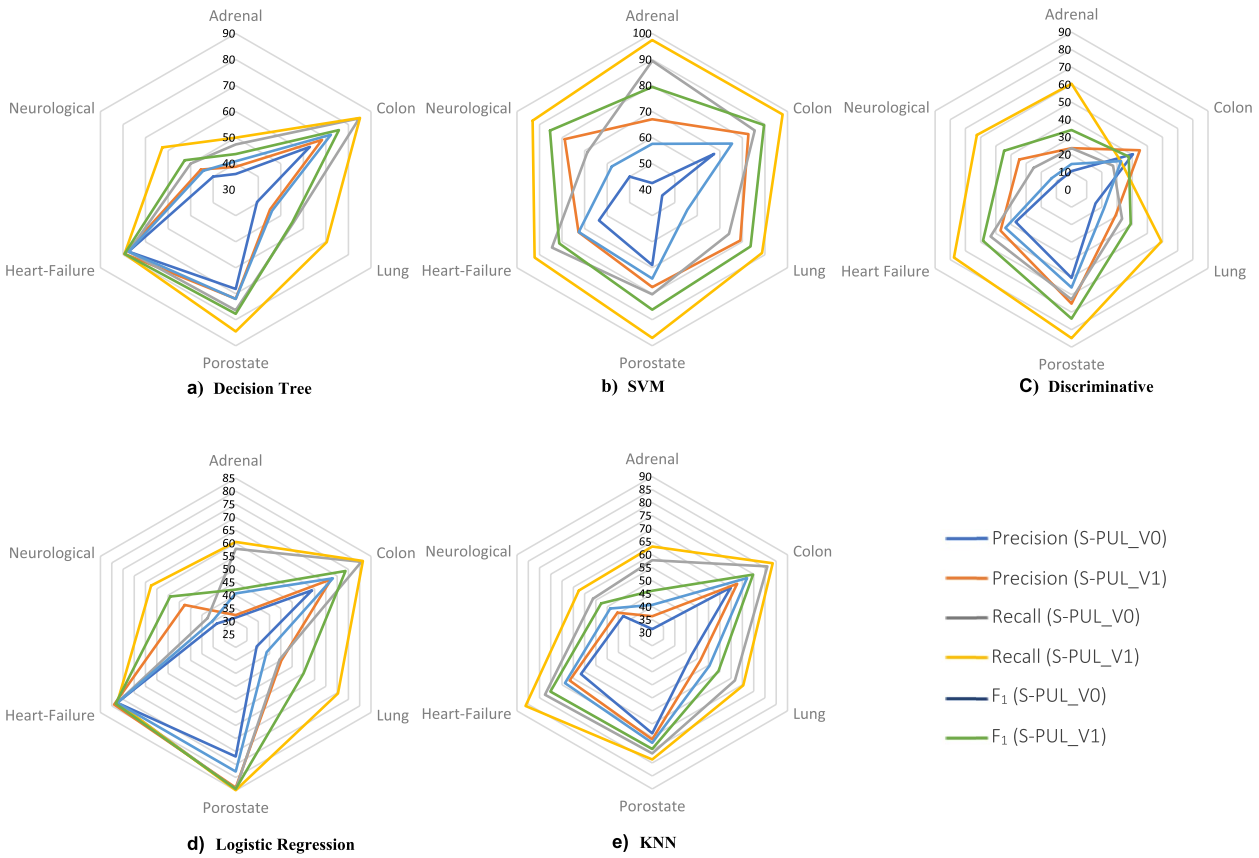

Fig. 6 The results of efficiency evaluation of binary classification algorithms for disease genes filtering (S-PUL_V1) and non-filtering (S-PUL-V0)

Table 9 Setting parameters of the SVM learning algorithm with polynomial kernel for each disease separately

The name of the disease	Parameter C	Parameter γ
Colon	15.33	0.14
Lung	16.43	0.14
Prostate	9.20	0.2
Adrenal	8.75	0.14
Heart Failure	13.61	0.14
Neurological	11.52	0.14

Table 10 Statistics of eliminated genes (genes having negative labels and being in the first S-R category)

The name of the disease	S-R range	Number of deletion genes
Adrenal	[0.08 , 10)	68
Colon	[0.06 , 10)	51
Lung	[0.1 , 10)	21
Prostate	[0.1 , 10)	37
Heart Failure	[0.12 , 10)	48
Neurological	[0.07 , 10)	28

obtains a positive label. Consequently, the number of spurious positive samples is very high. Therefore, the method for reducing the number of spurious positive samples is presented in "Action5- The prediction and ranking of disease candidate genes" section.

Figure 8 illustrates the results of S-PUL_V3 version efficiency in disease candidate genes ranking using the second score ("Action5- The prediction and ranking of disease candidate genes" section). According to these figures, the precision measure value enhances in the V3 version while recall is maintained. Moreover, Table 11 reports statistical information of the second score implementation in "Action5- The prediction and ranking of disease candidate genes" section for each disease and unlabeled gene number (which is introduced as a disease gene in this step).

The efficiency evaluation of using the third score of VRU set genes

Another measure (the third score) is used in "Action5- The prediction and ranking of disease candidate genes" section to reduce the number of spurious positive samples; this measure is calculated from the distance of the unlabeled gene for the support vector.

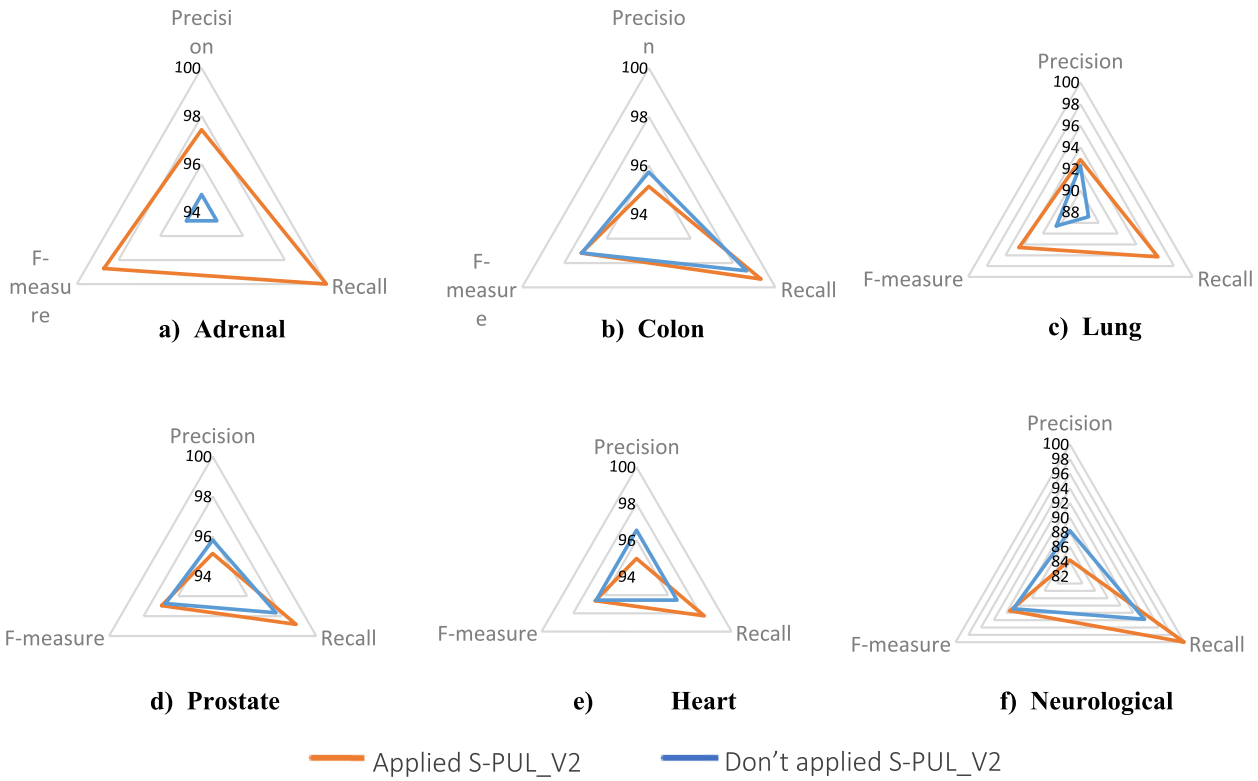

Fig. 7 The results of the S-PUL_V2 version evaluation (by implementing filter 3) compared to the S-PUL_V1 version (without implementing filter 3) for each disease

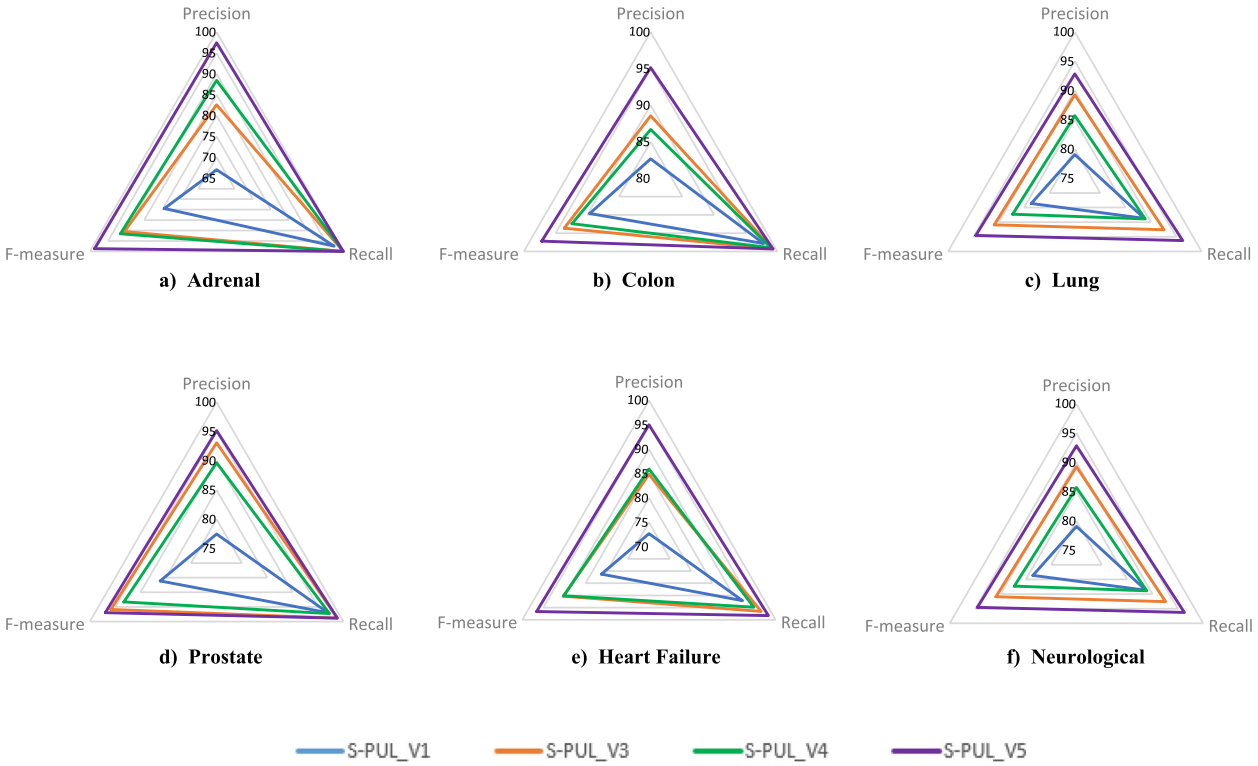

Fig. 8 The evaluation and comparing results of S-PUL of the V2 version (with implementing filter 3), V3 version (using the second score), V4 version (using the third score), and V5 version (using both second and third scores) in the prediction and ranking of disease genes

Figure 8 demonstrates the results of the S-PUL_V4 efficiency evaluation in disease candidate genes ranking using the third score. According to the figures, the entire evaluation measures increased in the V4 version. The highest and lowest increase value of recall measures is in Adrenal disease (2.52%) and Lung disease (0.19%), respectively. Further, the highest and lowest increase value of the precision measure is in the Adrenal disease (17%) and Colon disease (4.02%), respectively. Based on the results, using the third score in the V4 version (compared to the V2 version) dramatically increases the precision measure with maintaining the recall measure. Additionally, Table 12 reports the statistical information of the third score implementation in "Action5- The prediction and ranking of disease candidate genes" section for each disease and unlabeled gene number (introduced as the disease gene in this step).

The evaluation of the S-PUL method efficiency

Figure 8 illustrates the results of the efficiency evaluation of the S-PUL method (introduced in Table 1 with the S-PUL_V5 version) in disease candidate genes ranking using 1 and 2 filters in the learning step and using filter

Table 11 Statistical information of the second score implementation in "Action5- The prediction and ranking of disease candidate genes" section for each disease

The name of the disease	S-R range	Score range	Number of candidate disease genes	The second score range of disease genes
Adrenal	[2.98 , 108.21]	[4.51 , 1803.5]	46	[4.52 , 116.15]
Colon	[2.53 , 204.15]	[3.83 , 6495.6]	332	[1.11 , 2244.57]
Lung	[9.57 , 164.87]	[3.78 , 1107.8]	28	[6.47 , 2285.45]
Prostate	[10.09 , 102.672]	[6.20 , 347.4]	274	[36.09 , 4246.79]
Heart Failure	[4.45 , 72.04]	[6.74 , 873.21]	66	[13.52 , 1600.88]
Neurological	[1.79 , 29.05]	[11.93 , 581]	20	[30 , 1452.5]

Table 12 The statistical information of the third score implementation in "Action5- The prediction and ranking of disease candidate genes" section for each disease

The name of the disease	Distance interval from Support Vector	Score range	Number of candidate disease genes	The third score range of disease genes
Adrenal	[−3.28 , 1.12]	[−131.2 , 74.66]	45	[0.16 , 14.32]
Colon	[−6.21 , 14.78]	[−155.25 , 184.75]	338	[1.71 , 295]
Lung	[−9.25 , 3.78]	[−168.18 , 151.2]	28	[0.05 , 153.7]
Prostate	[−8.4 , 12.6]	[−168 , 180]	280	[0.43 , 302.5]
Heart Failure	[−7.1 , 3.03]	[−157.7 , 121.2]	44	[1.9 , 123.5]
Neurological	[−6.93 , 5.66]	[−173 , 161.71]	18	[0.03 , 301.71]

3 and both second and third scores compared to V2, V3, and V4 versions. According to the figures, all of the evaluation measures are enhanced in the V5 version compared to V2, V3, and V4 versions. The precision measure in Adrenal, Colon, Lung, Prostate, Heart Failure and Neurological diseases is enhanced by 12.48%, 13.78%, 17.68%, 22.31%, and 5.38%, respectively; besides, the recall measure for these diseases is increased by 2.52%, 1.47%, 7.6%, 1.84%, 6.11%, and 6.94%, respectively.

Comparing the efficiency of the S-PUL proposed method with other methods

The efficiency of the S-PUL proposed method is compared with previous methods in this section.

Table 13 reports the efficiency results of the proposed method compared to the [12] study. The recall measure is increased in Colon and Prostate diseases for all versions of the S-PUL method and Lung disease in the V5 version of the S-PUL method. The reason for comparing the results of S_PUL only with Reference [12] in Table 13 is due to the fact that these particular diseases were exclusively studied in that reference. However, from Tables 14, 15, 16 and 17, the diseases are common among various studies, allowing for comparisons across multiple references. The values of precision and recall are enhanced in the V5 version of the S-PUL method of each illness. The recall measure's highest and lowest increase values are in Colon disease (5.32%) and Lung disease (1.29%), respectively. The precision measure's highest and lowest increase values are in Prostate disease (3.14%) and Lung disease (1.75%).

According to Table 14, the precision and recall measures values for Cardiovascular disease class, including Heart Failure disease in the V5 version of the S-PUL method, are increased by 4.04% and 3.13%, respectively, compared to the [12] study. Recall measure in both V3 and V4 versions of the S-PUL method is increased by 0.59% and 2.32%, respectively, compared to the [12] study. The highest value of the recall measure is reported in the ProDige [19] study among the previous methods. The mentioned measure is increased by 0.25% and 1.97% in the V3 and V5 versions of the S-PUL method, respectively. The F_1 measure in all S-PUL versions is increased compared to previous studies, except for the [12] and EPU [14] studies.

According to Table 15, the recall value for Endocrine disease class (including Adrenal disease) increased by 0.53% in all V3, V4, and V5 versions of the S-PUL method than the [12] study (which had the best efficiency). Notably, this study's recall value for the Endocrine disease class reached 100%. In addition to the recall measure, the precision measure is increased by 2.46% in the V5 version of the S-PUL method. The F_1 measure is increased n V3, V4, and V5 versions of the S-PUL method compared to the previous studies, except for the [12] study.

Table 13 Comparing the performance of the S-PUL method (for four versions separately) with the [12] study

The name of the disease	Method	Precision	Recall	F_1
Colon	Nikdel et al. [12]	93%	94%	93%
	S-PUL_V1	82.66%	97.84%	89.61%
	S-PUL_V3	88.55%	**99.32%**	93.69%
	S-PUL_V4	86.68%	**98.98%**	92.42%
	S-PUL_V5	**95.14%**	**99.32%**	**97.19%**
Lung	Nikdel et al. [12]	91.1%	95%	93.3%
	S-PUL_V1	79.06%	88.69%	83.6%
	S-PUL_V3	89.28%	92.59%	90.9%
	S-PUL_V4	85.71%	88.88%	87.27%
	S-PUL_V5	**92.85%**	**96.29%**	**94.54%**
Prostate	Nikdel et al. [12]	92%	95%	93%
	S-PUL_V1	77.46%	96.98%	86.13%
	S-PUL_V3	93.06%	98.83%	95.86%
	S-PUL_V4	89.64%	97.28%	93.3%
	S-PUL_V5	**95.14%**	**98.83%**	**96.95%**

Table 14 Comparing the performance of the S-PUL method with previous methods in the prediction of unlabeled genes of Cardiovascular disease class (including Heart Failure disease)

Method	Precision	Recall	F_1
PUDI [20]	83.6%	75.3%	79.2%
ProDiGe [19]	57.3%	87.7%	69.3%
Smalter et al. [8]	76.4%	58.8%	66.5%
Xu et al. [21]	75.4%	62%	68%
EPU [14]	88.1%	87.7%	87.9%
Nikdel et al. [12]	94.79%	99.47%	97.07%
S-PUL_V1	67.05%	97.47%	79.45%
S-PUL_V3	82.60%	100%	90.47%
S-PUL_V4	84.44%	100%	91.56%
S-PUL_V5	**97.43%**	**100%**	**98.7%**

The efficiency results of the proposed method are reported in Table 16 compared to the efficiency of the previous methods for predicting the Cancer disease class candidate genes, including Colon, Prostate, and Lung diseases for V3, V4, and V5 versions of the S-PUL method. Based on the evaluation results, the efficiency of all three versions of the S-PUL method is enhanced compared to the [12] study (which had the best efficiency). The best results relate to the V5 version of the S-PUL method compared to the [12] study; its precision, recall, and F_1 measures are improved by 2.38%, 3.11%, and 4.75%.

The AUC value in the V5 version of the S-PUL method for Neurological disease class (including Neurological disease) is increased by 8.82%, compared to the SFM method [6] (the best previous method), according to Table 17. The recall measure in all versions of the S-PUL method is more than in previous methods. The best precision, recall, and F1 measures' values are related to EPU [14] by 78.2%, 80.4%, and 78.6%. These measures' values reached 84.21%, 100%, and 91.42% in the V5 version of the S-PUL method.

Table 15 Comparing the efficiency of the S-PUL method with previous methods in the prediction of unlabeled genes of Endocrine disease class (including Adrenal disease)

Method	Precision	Recall	F_1
PUDI [20]	82%	80.3%	80.4%
ProDiGe [19]	54.3%	96.3%	69.3%
Smalter et al. [8]	75.4%	67.6%	70.6%
Xu et al. [21]	72.1%	60%	65.4%
EPU [14]	85.2%	81%	84.1%
Nikdel et al. [12]	91.87%	94.23%	93.03%
S-PUL_V3	84.84%	96.55%	90.32%
S-PUL_V4	85.93%	94.82%	90.16%
S-PUL_V5	**95%**	**98.27%**	**96.61%**

Table 16 Comparing the efficiency of the S-PUL method with the previous methods in the prediction of Cancer class unlabeled genes, including Colon, Prostate, and Lung diseases

Method	Precision	Recall	F_1
PUDI [20]	76.3%	80%	78%
ProDiGe [15]	71.1%	79.8%	75.3%
Smalter et al. [8]	73.8%	79%	76.3%
Xu et al. [21]	71%	79.7%	75.1%
EPU [14]	81.2%	84.5%	82.6%
SFM [6]	76.9%	79.8%	78.3%
Nikdel et al. [11]	96.73%	95.83%	94.28%
S-PUL_V3	98.94%	**98.94%**	98.94%
S-PUL_V4	98.76%	**98.94%**	98.85%
S-PUL_V5	**99.11%**	**98.94%**	**99.03%**

Comparing the efficiency of the S-PUL proposed method with biologists' efficiency

It is worth noting that the biological researchers identified other unlabeled genes as disease genes (six diseases introduced in Table 2) from 2015 to 2020 through laboratory methods. Then, they are introduced in the [4] dataset. The predicted disease genes by the S-PUL method and [12] study with the disease genes set (introduced for the 2015–2016 period and 2017–2020 period by biological researchers) are compared in Tables 18 and 19, respectively, to determine the efficiency. Notably, 2015 to 2016 and 2017 to 2020 sets are reported in the third and fourth rows of Table 2, respectively.

According to Table 18, the efficiency of the V5 version of the S-PUL method compared to the [12] study is as follows:

The prediction in Adrenal disease is the same; in Colon disease has ten more disease genes; in Prostate disease has 11 more disease genes. On the other hand, the V5 version of the S-PUL method has predicted two disease

Table 17 Comparing the efficiency of the S-PUL method with previous methods in the prediction of the unlabeled genes of Neurological disease class (including Neurological disease)

Method	Precision	Recall	F_1	AUC
PUDI [20]	70.3%	80.1%	74.9%	85.4%
ProDiGe [19]	63.1%	74%	68.1%	64.6%
Smalter et al. [8]	60.6%	65.6%	63.1%	73.9%
SFM [6]	-	-	-	88.2%
Xu et al. [21]	59.7%	66.7%	63%	-
EPU [14]	78.2%	80.4%	78.6%	-
S-PUL_V1	78.82%	93.05%	85.35%	-
S-PUL_V3	80%	100%	88.88%	-
S-PUL_V4	83.33%	93.75%	88.23%	-
S-PUL_V5	**84.21%**	**100%**	**91.42%**	**97.02%**

Table 18 Comparing the S-PUL method efficiency and the [12] study with biologists in the prediction of disease genes from 2015 to 2016

The name of the disease ↓	Method→	Biologists [4]	Nikdel et al. [12]	SPUL_V3	SPUL_V4	SPUL_V5
Adrenal	Number of genes	9	9	8	8	9
	Recall		100%	88.88%	88.88%	100%
Colon	Number of genes	240	229	233	238	239
	Recall		95.41%	97.08%	99.16%	99.58%
Prostate	Number of genes	191	178	182	188	189
	Recall		93.61%	95.28%	98.42%	98.95%

genes and only one disease gene lesser than biologists [4] in Prostate and Colon diseases, respectively.

According to Table 19, compared to biologists, the V4 and V5 versions of the S-PUL method are predicted all genes in Adrenal and Neurological diseases. Moreover, the V5 version of the S-PUL method only predicted one disease gene lesser than biologists in Colon, Per, Lung, and Heart Failure diseases. Hence, according to the learned models, the V5 version efficiency of the S-PUL method in predicting disease genes is very proper based on the 2014 dataset (introduced in the second row of Table 2).

Conclusion

In two steps, the reliable negative genes are extracted in this study to reduce available noise in extracted negative genes from unlabeled genes. These two steps are (i) one-class learning and (ii) filtering based on the distance measure. The proposed method initially filters positive educational genes in the disease binary model learning step. Then, the SVM binary model is learned using selected positive samples and extracted reliable negative samples for each disease separately. In the prediction step, the binary model is learned to predict unlabeled samples' labels (labeling) and rank them. Moreover, two filters of (i) nearness of gene to disease genes and (ii) distance of each gene from the support vector are used.

Using influential factors to predict and rank disease candidate genes and properly use them in the S-PUL method leads to the strong performance of this method compared with previous methods. In this line, the mentioned claim is proved by 99.51% average correspondence of predicted disease genes with introduced disease genes from 2015 to 2016 and 98.54% from 2017 to 2020. Moreover, 96.74% lack of average of considered negative genes in evaluating disease genes during the mentioned periods proves this claim.

The following propositions are presented for future studies in this regard based on the performed implementation and advantages and disadvantages of the presented method:

Table 19 Comparing the S-PUL method efficiency with biologists in predicting disease genes from 2017 to 2020

The name of the disease ↓	Method→	Biologists [4]	SPUL_V3	SPUL_V4	SPUL_V5
Adrenal	Number of genes	29	28	29	29
	Recall		96.55%	100%	100%
Colon	Number of genes	56	54	55	55
	Recall		96.42%	98.21%	98.21%
Prostate	Number of genes	67	62	66	66
	Recall		92.53%	98.50%	98.50%
Lung	Number of genes	27	23	25	26
	Recall		85.18%	92.59%	96.29%
Heart Failure	Number of genes	58	50	56	57
	Recall		86.20%	96.55%	98.27%
Neurological	Number of genes	16	13	16	16
	Recall		81.25%	100%	100%

A) The delimitation of distances and scoring to genes are carried out discretely and integrity units in the S-PUL method. More or less of the genes located at borders (even fractional) can lead to changes in category and score in such a way that eliminates or maintain the gene. This method should be improved.
B) Two steps are used in this study to find reliable negative genes. It is proposed to use other information sources (such as the PPI network) to increase trust in extracted negative genes.
C) Two filtering methods based on statistical measures are used in this study to reduce errors in identifying and ranking disease candidate genes. Meanwhile, other genetic factors that are effective in the formation of a disease can consider and introduce in the final score.
D) A deep learning approach in PU learning is proposed to improve the results of identifying and predicting disease candidate genes.

Abbreviations

AML	Acute Myeloid Leukemia
AUC	Area Under the Curve
DNA	Deoxyribonucleic Acid
DP_Score	Disease Prediction Score
DS1	Dataset 1
DS2	Dataset 2
DS3	Dataset 3
DS_Score	Disease Score
F1	F1 Score (Harmonic Mean of Precision and Recall)
F_Score	Final Score
FN	False Negative
FP	False Positive
FPR	False Positive Rate
Fracrej	Fraction Rejected
GEP	Gene Expression Profile
HMM	Hidden Markov Model
IL	Interval Length
KNN	K-Nearest Neighbors
NGr	Normalized Gene Relevance
OCSVM	One-Class Support Vector Machine
PCA	Principal Component Analysis
PPI	Protein–Protein Interaction
Precision	Proportion of correctly predicted positive cases
PU-Learning	Positive-Unlabeled Learning
RBF	Radial Basis Function
Recall	Percentage of correctly predicted disease genes
Recall (TPR)	True Positive Rate
RU_i	Remaining Unlabeled Gene Set
S-PUL	Scored-Positive Unlabeled Learning
S-R	Score Relevance
SVM	Support Vector Machine
SVDD	Support Vector Data Description
TF-IDF	Term Frequency-Inverse Document Frequency
TNG	Trust in Negative Genes
TN	True Negative
TP	True Positive
TPR	True Positive Rate
VRU_i	Valuable Remaining Unlabeled Gene Set

Acknowledgements
Yang, Peng, et al. "Ensemble positive unlabeled learning for disease gene identification." PloS one 9.5 (2014): e97079. https://www.genecards.org/,29/04/2020.

Authors' contributions
S.M. wrote the main manuscript text. S.J. supervised the study and edited and improved the manuscript. The proposed method was presented and evaluated by S.M., and reviewed and enhanced by S.J.

Funding
The authors have no funding to declare that are relevant to the content of this article.

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
Not applicable. The gene expression data used in this study were obtained from the publicly available GeneCards database.

Consent for publication
Not applicable. This study does not involve any individual data requiring consent for publication.

Competing interests
The authors declare no competing interests.

Received: 23 October 2024 Accepted: 18 February 2025
Published online: 16 April 2025

References

1. Fusilier DH, et al. Detecting positive and negative deceptive opinions using PU-learning. Inform Process Manage. 2015;51(4):433–43.
2. Shao YH, et al. Laplacian unit-hyperplane learning from positive and unlabeled examples. Inform Sci. 2015;314:152–68.
3. Zhang Z, et al. Biased p-norm support vector machine for PU learning. Neurocomputing. 2014;136:256–61.
4. Genecards, the human gene database, Weizman Institute of Science. https://www.genecards.org. Accessed 24 Apr 2020.
5. Scoring theory. https://www.elastic.co. Accessed 11 May 2020.
6. Yousef A, Charkari NM. SFM: a novel sequence-based fusion method for disease genes identification and prioritization. J Theor Biol. 2015;383:12–9.
7. Vasighizaker A, Jalili S. C-PUGP: A cluster-based positive unlabeled learning method for disease gene prediction and prioritization. Comput Biol Chem. 2018;76:23–31.
8. Smalter A, Lei SF, Chen XW. Human disease-gene classification with integrative sequence-based and topological features of protein-protein interaction networks. 2007 IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2007).
9. Radivojac P, et al. An integrated approach to inferring gene–disease associations in humans. Proteins: Structure, Function, and Bioinformatics. 2008;72(3):1030–7.
10. Yousef A, Charkari NM. A novel method based on physicochemical properties of amino acids and one class classification algorithm for disease gene identification. J Biomed Inform. 2015;56:300–6.
11. Vasighi Zaker A, Saeed J. Candidate disease gene prediction using one-class classification. Soft Computing J. 2016;4(1):74–83.
12. Nikdelfaz O, Jalili S. Disease genes prediction by HMM based PU-learning using gene expression profiles. J Biomed Inform. 2018;81:102–11.
13. Vasighizaker A, Sharma A, Dehzangi A. A novel one-class classification approach to accurately predict disease-gene association in acute myeloid leukemia cancer. PLoS ONE. 2019;14(12):e0226115.
14. Yang P, et al. Ensemble positive unlabeled learning for disease gene identification. PloS one. 2014;9(5):e97079.
15. Tax DMJ, Duin RPW. Support vector data description. Mach Learn. 2004;54(1):45–66.
16. Huber PJ. Robust Statistics. New York: John Wiley & Sons; 1981.
17. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theory. 1967;13(1):21–7.

18. Parzen E. On Estimation of a Probability Density Function and Mode. Ann Math Stat. 1962;33(3):1065–76.
19. Mordelet F, Vert J-P. ProDiGe: Prioritization Of Disease Genes with multitask machine learning from positive and unlabeled examples. BMC Bioinformatics. 2011;12(1):1–15.
20. Yang P, et al. Positive-unlabeled learning for disease gene identification. Bioinformatics. 2012;28(20):2640–7.
21. Xu J, Li Y. Discovering disease-genes by topological features in human protein–protein interaction network. Bioinformatics. 2006;22(22):2800–5.

Publisher's Note