scientific reports

Check for updates

OPEN

Providing an optimized model to detect driver genes from heterogeneous cancer samples using restriction in subspace learning

Ali Reza Ebadi[1], Ali Soleimani[2]✉ & Abdulbaghi Ghaderzadeh[1]

Extracting the drivers from genes with mutation, and segregation of driver and passenger genes are known as the most controversial issues in cancer studies. According to the heterogeneity of cancer, it is not possible to identify indicators under a group of associated drivers, in order to identify a group of patients with diseases related to these subgroups. Therefore, the precise identification of the related driver genes using artificial intelligence techniques is still considered as a challenge for researchers. In this research, a new method has been developed using the subspace learning method, unsupervised learning, and with more constraints. Accordingly, it has been attempted to extract the driver genes with more precision and accurate results. The obtained results show that the proposed method is more to predict the driver genes and subgroups of driver genes which have the highest degree of overlap due to *p*-value with known driver genes in valid databases. Driver genes are the benchmark of MsigDB which have more overlap compared to them as selected driver genes. In this article, in addition to including the driver genes defined in previous work, introduce newer driver genes. The minister will define newer groups of driver genes compared to other methods the *p*-value of the proposed method was 9.21e-7 better than previous methods for 200 genes. Due to the overlap and newer driver genes and driver gene group and subgroups. The results show that the *p* value of the proposed method is about 2.7 times less than the driver sub method due to overlap, indicating that the proposed method can identify driver genes in cancerous tumors with greater accuracy and reliability.

Cancer is one of the deadliest diseases, and according to the estimation of the American Cancer Association in 2019, about 1,762,450 people has cancer worldwide, of them about 606, 880, individuals have died. Cancer is the second leading cause of death among all diseases[1]. One of the reasons for the abnormal tumor growth, is the rate of DNA mutation in the driver genes, which consequently causes mess in the function of the cancerous cell of a tumor. Due to this reason, having integrated information on this field helps establishing cancer detection and treatment strategies[2]. The large genome changes is one of the causes of cancer, using the second-generation technology of DNA sequencing and analysis, which would significantly contribute to the biological understanding of diagnosis and treatment of cancer. This insight helps us examining each type of change in the somatic genome, and also facilitates the detection of mutant genes in cancer samples[3]. In this study, although we have been able to identify all mutant genes in the tumor, many of these mutant genes have no effect on the tumor development, which are known as passenger genes. Accurate and direct identification of whether this passenger gene has an impact on the development of the tumor or not, still remains a challenge. So, one of the major works in the field of cancer research is identifying the passenger gene from the driver's gene in cases with cancer[4,5]. One common way to deduce driver genes, is a hypothesis that "the driver's mutated genes are primarily among the large groups of sample mutated tumor genes". Therefore, based on this hypothesis, many scientific studies have been driven using computational methods of identifying driver genes among the mutated gene groups in this field[6]. In OncodriveCLUST, specific genes that tend to cluster mutations throughout the protein sequence, were identified, which indicated that these genes have a particular bias toward their dependent gene sequences. Moreover,

[1]Department Computer Engineering, Sanandaj Branch, Islamic Azad University, Sanandaj, Iran. [2]Department of Computer Engineering, College of Technical and Engineering, Malard Branch, Islamic Azad University, Tehran, Iran. ✉email: a.soleimani.uni@iaumalard.ac.ir

based on this hypothesis, in this method, a number of genes that had high mutation frequencies were the driving candidate genes, which were later found to have no significant effect on tumor growth[7]. The MutSigCV method solves one of the challenges in identifying driver genes. Previous methods have identified a list of driver genes, but because of mutation heterogeneity, some of them have not been identified properly. Therefore, by the use of this method, this problem has been solved[8]. Because cancer is a heterogeneous disease, there are many different subtypes for one type of cancer, and the driver genes of each subgroup may be different from the other genes. If a mutated gene acts as a driver gene for several specimens in a subgroup, it can be identified as the driver gene of the subgroup and also can be used as a criterion for separating subgroups[9]. Considering the genomic diversity and heterogeneity of subgroups of specific genes in a group, which their driver is small part of samples, so they are rarely changing among all the samples[10]. Other methods are also used to identify a rare mutation except for mutation frequency, such as modifying the amino acid of the flanking sequence Another method based on optimizing SpeMDP and the maximum matrix weight, is used to identify the driver genes. In this method, the genome data of twelve different types of heterogeneous cancer are used to form a common biological path, and finally the genes in this common path, are used as candidate genes[11,12]. All the results of the previous methods are encountered the problem that methods are suitable for the idea mode. It is appropriate when all subgroup information are available, which are not mostly available in many cases. The accurate extraction of the driver genes, without providing the subgroup information to find the exact treatment of cancer and personal medicine, still remains as a challenge[10]. To solve the problem of inaccessibility of information, the margin writing of the subgroups, as Driversub method was proposed. Correspondingly, in this method, an unsupervised learning method was used[13], which needs no information about subgroup[14]. One of the challenges in analyzing the results of this method is that the available data has noise and there is still discarded data, which consequently affects the accuracy of the results. Therefore, in this study, we have tried to overcome this problem by developing this method. In this article, we achieved better accurate for this method via developing the driversub method, and by applying more restrictions on the data. To achieve this goal, robust adaptive graph regularized non-negative matrix factorization method has been used, and by applying less weight to noisy as well as the discarded data, and giving more weight to clean data, we have tried to improve the accuracy of the results. This article used the Cancer Genome Atlas (TCGA) program and Cancer Gene Gensus (CGC).

Method

Subspace learning. Due to the lack of subgroup information, we have used an unsupervised learning method. To do this, a subspace learning framework has been used[15]. Afterward, we displayed the marginal writing information of a gene, as a vector, so that the mutation data of the gene with high dimensions was converted into a small subspace with smaller dimensions. Gene mutation input data was converted to a binary matrix. The mutation vector of each gene is $X = [x_1, x_2, x_3, \ldots x_p]$, where p is equal to the total number of genes. The input matrix contains p-genes and n-samples, and each entry of this matrix indicates whether the ith gene has been mutated in the j sample or not[16]. The output matrix was $Z = [z1, z_2, z_3, \ldots z_i \ldots z_p]$, which was compressed space with less dimensions, so that $k << n$, where k is the dimension of the output matrix Z[17]. Low-dimensional output matrices in vector space can be better suited for the computational analysis. Although the output matrix can well represent the mutation index of the input matrix, the main challenge is that there is no indication to show that the investigated gene is from which one of the subgroups. In fact, there is no general criterion for matching a gene with a subgroup. Due to the fact that the sub-space dimensions can determine the hidden features related to each gene, and the sub-space output dimensions are almost able to determine the indicators related to each subgroup. However, there is no guarantee that the dimensions of the subspecies matrix represent those indicators related to that subgroup, so it can be used to determine whether the special checked genes is relevant to that subgroup[18]. Based on the two hypotheses proposed in the driversub method, the values of the output vector indices can be used as criteria for evaluating the driver's genes. Also, in the second hypothesis, the values of the output vectors can be used indicators for determining whether a gene belongs to a specific subgroup. However, to increase the guarantee of the first hypothesis in the driversub method, the regularization of L1 norm was used to ensure that the output vectors are sparse[19]. Because the values of the espresso of output vector index are large, the output vectors will more tend to be inclined to match the coordinates of the dimensions subspace. The axes of dimensions of subspace can be used as an indicator to recognize if a gene belongs to a particular subgroups. Hence, at the beginning of the first step, we use the objective function defined in the GNMF method [20]as follows:

$$\min_{U\geq 0, V\geq 0} \left\| X - UV^T \right\|_F^2 + \lambda T_r(V^T L_S V) \tag{1}$$

where LS = DS -S is called graph Laplacian. S is the data similarity matrix, and DS is the degree matrix which the used attribute function in this method is as follows:

$$\begin{gathered} \min \sum_{\substack{i=1 \\ w,z}}^{p} \|x_i - wz_i\|_2^2 + \lambda_z \sum_{i=1}^{p} \| z_i \|_1 \\ \text{s.t } W \geq 0 \text{ and } Z_i \geq 0, \ \forall_i = 1, \ldots p \end{gathered} \tag{2}$$

Which λ_z controls the distance between the output vectors Z and the coordinate axes and the coefficient of the regulator of sparse value. Also, one of the problems of the space learning method is an overflowing problem[21]. To overcome this problem, Frobenius norm regularization has been used in the driversub method, which changed the attribute function as follows.

$$\min_{W,z_i} \sum_{i=1}^{p} \| x_i - Wz_i \|_2^2 + \lambda_z \sum_{i=1}^{p} \| z_i \|_1 + \lambda_w \| W \|_F^2$$
$$\text{s.t } W \geq 0 \text{ and } Z_i \geq 0. \ \forall_i = 1, \ldots, p \tag{3}$$

Here, our parameters are the weight matrix W which can reverse the relationship between a subset of samples and subspace dimensions to calculate the real values of Matrix W and Z, we used the basic method of matrix factorization, and each time we repeated the initial W, Z we obtained more accurate values… What has been forgotten is that in calculating the similarity between the Samples, the Gaussian kernel function can also be used, which is as follows:

$$s_{i,j} = e^{\frac{\|x_i - x_j\|_2^2}{2\sigma^2}} \tag{4}$$

$s_{i,j}$ is the similarity between i and j samples. While the Euclidean distance is used to compute the difference between two different samples, real trait space of sample including noise and large amounts of unrelated features, which play no role in similarity, but they can be used for similarity. To enhance the accuracy and precision, a number of irrelevant and disconnected attributes should be eliminated. It was shown that the attributes that have an impact, will have more weight in the distance calculation. Therefore, in order to achieve this goal, it is necessary to learn an M matrix to obtain the exact distance, so we have used M matrix in this article where M is a diagonal matrix. Herein, we get the distance as follows:

$$\| x_i - x_j \|_M^2 = (x_i - x_j)^T M (x_i - x_j) \tag{5}$$

In this article, we have attempted to reduce noise by combining driversub. The methods as well as applying more restrictions on the obtained samples. Robust adaptive graph regularized NMF (RAGNMF) was also used, which is as follow:

$$\min_{w,z,W,M} Tr[M(x - wz_i)W(x - wz_i)^T] + \lambda Tr(z^T l_s z) + \alpha \| W \|_F^2 + \beta \| M \|_F^2$$
$$\text{s.t } w \geq 0, \quad z \geq 0, \quad W \geq 0, \quad M \geq 0 \tag{6}$$

Optimization

To solve the desired method, a duplicate updating method was used, which is as follows.

By keeping W, M constant, the values of w and z were calculated as follows:

$$w_{ir} = w_{ir} \frac{(MxWz)_{ir}}{(MwzWz^t)_{ir}} \tag{7}$$

$$z_{jr} = z_{jr} \frac{(Wx^t Mw)_{jr}}{(Wz^t w^t Mw + \lambda l_s z^t)_{jr}} \tag{8}$$

In order to update the W value, by keeping values of M, w, and z constant, the following relationships were obtained.

$$\min_w = Tr[E^M W] + \alpha \| W \|_F^2$$
$$\text{s.t } W_i \geq 0, \tag{9}$$

where E^M is as follows:

$$E^M = (x - wz)M(x - wz)^T$$

$$\sum_{i=1}^{n} W_i = C_i \tag{10}$$

Function (10) can be converted to the following equation:

$$\min \sum_{\substack{i=1 \\ M}}^{n} \left(W + \frac{E_i^M}{2\sigma} \right)_i^2$$
$$\text{s.t } W_i \geq 0, \ \sum_{i=1}^{n} W_i = C_i \tag{11}$$

Also, by keeping the values of W, z, and w constant, the value of M was calculated as follows:

$$\min_{M} Tr[E^{W}M] + \beta \parallel M \parallel_{F}^{2}$$
$$\text{s.t} \;\; M_i \geq 0, \sum_{i=1}^{m} M_i = C_i \tag{12}$$

We have used the matrix factorization method here Another point is that here λ, α, β are the control and regulating parameters and the correctness of the method depends on these parameters when the accuracy of the results is reduced when α is too large or too small, Here we have created a filter on the weights between the input and output vectors by imposing a constraint on the softness of the filter and low weights or high impact weights that are abnormal and the use of control parameters. We have achieved better results.

In this study, to solve this equation, we used the Accelerated Gradient Method, which was earlier used in[22]. The steps of performing this work are shown in the algorithm 1.1:

Algorithm 1.1

Start

Input

Regularized Parameters: λ, α, β

$X = [x_1, x_2, x_3, ...x_p]$ # Mutation vector of each gene

W, M initialize is Identity matrix # is similarity between (i, j) genes and (i, j) sample in ordering Output

$Z = [z_1, z_2, z_3, ...z_i.....z_p]$, W

Begin

1. The calculated Z according to the formula (6)
2. The calculated W according to the formula (5)
3. The Updated W and M according to the formula (7),(10)
4. The Updated Attribute function according to the formula (4)

End

Results

In this study, we used breast cancer data (Cancer Genome Atlas Network and others, 2012), which included somatic mutations of 507 samples and 12,233 genes that can be downloaded from the cBioPortal database[23]. By default, we considered the dimensions of k subspace as 4. In the present study, we have used Python 3.7 to implement this method. Moreover, we used Gsea Msigdb web-based software[24] to analyze the results. Firstly, we calculated the mutation score of each gene from the output vector obtained from the learning subspace, and then arranged it in descending order. Thereafter, we separated the top 500 genes with the highest mutation scores, and then selected them as the candidate for driver genes. Finally, we compared the results with the Benchmarks on Msigdb show in Figs. 1, 2 and 3. Thus, we have taken from the 200 candidate driver genes obtained from this method about 13 genes which had the lowest *p*-value and highest mutation score, for example, and used the outputs obtained by the Msigdb web software and curated gene sets as a benchmark. Driver genes results have a very good overlap with the defined driver genes. They are also very similar to the previous methods in terms of defined driver genes. Besides, the new genes defined in this method are very similar due to *p*-value compared to previous works. They have better overlap with breast cancer benchmarks. The results of our simulation in the model presented in Table 1, Figs. 2 and 3. In Fig. 1, the details of candidate driver genes obtained from this method for breast cancer show 13 genes with characteristics which is based on the amount of *p*-value overlap with the driver genes in the specific subgroup of benchmark, As we can see in Fig. 2, from the 13 proposed driver genes in specific subgroups of the benchmark, we see the different numbers of genes of that overlap. Hence, the more driver genes overlap with the benchmark, and the lower the amount *p*-value, the better the outcome. In the Fig. 3, and as it can be seen, the more black cells there are, the more they overlap in one gene. In 200 driver genes obtained from top to bottom ranking, the number of genes proposed by our method overlapped better than previous methods so that we have achieved *p*-value = 9.21e-07. In Fig. 4 and Table 1, Comparing the proposed

Collections	# Overlaps Shown	# Gene Sets in Collections	# Genes in Comparison (n)	# Genes in Universe (N)
C2	10	6226	13	40071

Conversion Details

Original Member	NCBI (Entrez) Gene Id	Gene Symbol	Gene Description
BRCA1	672	BRCA1	BRCA1 DNA repair associated [Source:HGNC Symbol;Acc:HGNC:1100]
BRCA2	675	BRCA2	BRCA2 DNA repair associated [Source:HGNC Symbol;Acc:HGNC:1101]
EPHA4	2043	EPHA4	EPH receptor A4 [Source:HGNC Symbol;Acc:HGNC:3388]
ERBB2	2064	ERBB2	erb-b2 receptor tyrosine kinase 2 [Source:HGNC Symbol;Acc:HGNC:3430]
ISTN1	no mapping	no mapping	
MAP2K4	6416	MAP2K4	mitogen-activated protein kinase kinase 4 [Source:HGNC Symbol;Acc:HGNC:6844]
MAP3K1	4214	MAP3K1	mitogen-activated protein kinase kinase kinase 1 [Source:HGNC Symbol;Acc:HGNC:6848]
MTOR	2475	MTOR	mechanistic target of rapamycin kinase [Source:HGNC Symbol;Acc:HGNC:3942]
MYO10	4651	MYO10	myosin X [Source:HGNC Symbol;Acc:HGNC:7593]
NCOA2	10499	NCOA2	nuclear receptor coactivator 2 [Source:HGNC Symbol;Acc:HGNC:7669]
PIK3CA	5290	PIK3CA	phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha [Source:HGNC Symbol;Acc:HGNC:8975]
SLIT2	9353	SLIT2	slit guidance ligand 2 [Source:HGNC Symbol;Acc:HGNC:11086]
TP53	7157	TP53	tumor protein p53 [Source:HGNC Symbol;Acc:HGNC:11998]
WRN	7486	WRN	WRN RecQ like helicase [Source:HGNC Symbol;Acc:HGNC:12791]

Figure 1. Details of 13 top driver genes in the proposed method.

method with the previous methods due to the amount of p-value for 200 driver genes due to the averages can be seen, which the proposed method of this paper performs better. We compared the proposed specific subgroups of 100, 200, and 500 members of driver genes with the driver genes in the benchmark due to the degree of overlap. The results indicate a good degree of overlap. In details of Fig. 1, you can see 13 driver genes in the proposed method as showed in Table 1 and Fig. 4. Comparing p-values between the previous and the proposed methods for an average of a subset of 200 driver genes that the lowest p-value and highest mutation score which were compared by different methods in the Table 1. So that the proposed method has a significantly lower average p-value. Due to Fig. 5, the overlap of the number of suggested genes with curated gene sets (Misgdb) is observed. Considering the subgroups of 100, 200 and 500 members of driver genes are obtained and their overlap with the driver gene database is seen. The results indicate a good degree of overlap. to further analyze the results on breast cancer data, we compared the superior driver genes selected by existing method with previous methods so that 40 genes are shared between the proposed method and MutsigCV, and also between MutsigCV and Driversub. Under similar conditions, there are 21 common genes, and also between OncodriveCLUST and Driversub, there are about 81 common driver genes, while between the method proposed in this article and OncodriveCLUST, we had about 108 common genes which indicates that the proposed method is better. BRCA2, ERBB2, and PIK3CA are common, which are genes with high mutations, and also in the overlap between the proposed methods of OncodriveCLUST driver genes Which AKAP9, MTOR, TP53 are high mutant driver. The results indicate that

Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p-value	FDR q-value
GRESHOCK_CANCER_COPY_NUMBER_UP [323]	Genes from common genomic gains observed in a meta analyis of copy number alterations across a panel of different cancer cell lines and tumor samples.	8		$2.03\ e^{-14}$	$1.26\ e^{-10}$
WP_SIGNALING_PATHWAYS_IN_GLIOBLASTOMA [83]	Signaling Pathways in Glioblastoma	6		$1.11\ e^{-13}$	$3.47\ e^{-10}$
WP_BREAST_CANCER_PATHWAY [156]	Breast cancer pathway	6		$5.3\ e^{-12}$	$1.1\ e^{-8}$
DACOSTA_UV_RESPONSE_VIA_ERCC3_DN [856]	Genes down-regulated in fibroblasts expressing mutant forms of ERCC3 [GeneID=2071] after UV irradiation.	8		$4.92\ e^{-11}$	$7.11\ e^{-8}$
WP_PANCREATIC_ADENOCARCINOMA_PATHWAY [89]	Pancreatic adenocarcinoma pathway	5		$6.12\ e^{-11}$	$7.11\ e^{-8}$
WP_ERBB_SIGNALING_PATHWAY [91]	ErbB Signaling Pathway	5		$6.85\ e^{-11}$	$7.11\ e^{-8}$
TCGA_GLIOBLASTOMA_MUTATED [8]	Genes significantly mutated in 91 glioblastoma samples.	3		$1.49\ e^{-9}$	$1.33\ e^{-6}$
WP_PATHWAYS_AFFECTED_IN_ADENOID_CYSTIC TIC_CARCINOMA [66]	Pathways Affected in Adenoid Cystic Carcinoma	4		$4.74\ e^{-9}$	$3.69\ e^{-6}$
KEGG_PANCREATIC_CANCER [70]	Pancreatic cancer	4		$6.03\ e^{-9}$	$3.76\ e^{-6}$
PID_CDC42_PATHWAY [70]	CDC42 signaling events	4		$6.03\ e^{-9}$	$3.76\ e^{-6}$

Figure 2. The number of driver genes overlapping with the benchmark dataset in a subset of the top 13 driver genes.

the used method in this article with a high ability to predict and deduce driver genes was shown to be better than previous methods.

Table 1 and Fig. 4 Comparing *p*-values between the previous and the proposed methods for an average of a subset of 200 driver genes the lowest *p* value and highest mutation score which were compared by different methods in the Table 1, so that the proposed method has a significantly lower average *p* value.

In this method, BRCA1, BRCA2, ERBB2, PIK3CA, TP53, and KDM6A genes were introduced as driver candidate genes, which were also common in previous methods. The genes introduced by the proposed method, had a good overlap. In addition, the genes MYO10, ISTN1, EPHA4, SLIT2, WRN, DOP1B PLXNA2, and TCHH were introduced using the proposed method. Due to the elimination of overflow and suspension in the proposed method, the predicted genes were significantly different from the previous methods. Figure 6 shows the heat map diagram of seven genes with the highest score subspace (z) with k = 4 in the proposed method, which showed the heterogeneity of the mutation of specific genes in each one of the subgroups.

Figure 5 overlap of the number of suggested genes with curated gene sets (Misgdb) is observed. Considering the subgroups of 100, 200 and 500 members of the driver genes are obtained and their overlap with the driver gene database is seen. The results show that the candidate driver genes overlap well.

In Fig. 6, we see the overlap and distribution of the driver genes of new top candidate defined for 200 genes which were randomly selected with four specific subgroups of driver genes defined in the color bar, which the adjustment margin increases *p*-value from bottom to top. Black color indicates the lowest *p*-value and highest mutation score, for example, the gene GH2 has more overlap.

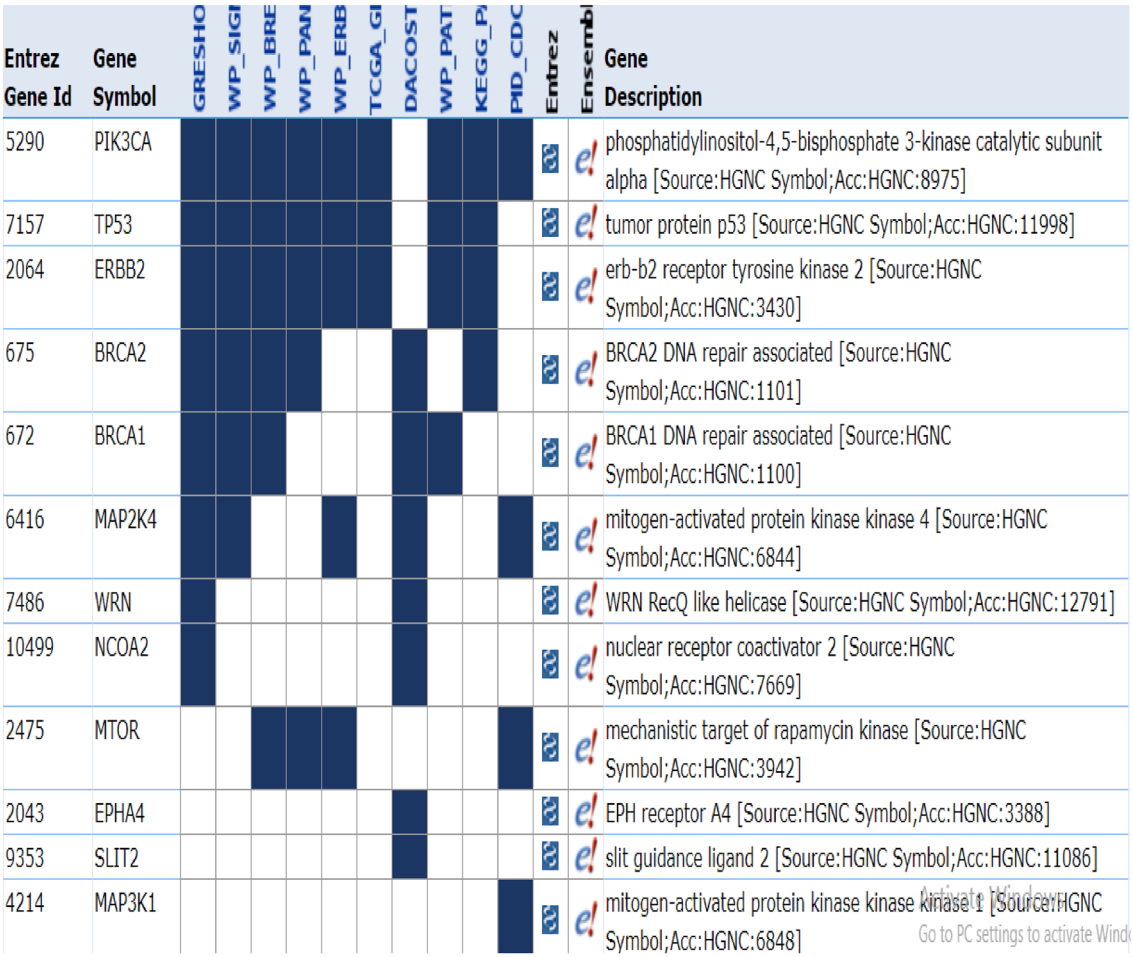

Figure 3. Proposed driver genes are compared to several benchmarks at the same time.

Method	Average (*p* value)
MutSigCV	8.35e−02
OncodriveCLUST	1.23e−02
DriverSub	1.46e−06
proposed method	9.21e−07

Table 1. Comparing *p*-values between the previous and the proposed methods for an average of a subset of 200 driver genes the results of comparison between different methods in the method proposed in this paper have an average of *p* value less for 200 driver genes.

Discussion

Extraction of subgroups of driver genes is one of the most important cases in personal medicine and heterogeneity in cancer. One of the problems in this regard is the lack of subspace margin information due to the fact that annotation of the subtypes of cancer samples is not available in many cases and previous methods cannot correctly determine the driver genes of each subgroup; hence, we predict the subtypes of driver genes in the heterogeneous cancers. A very important point which was forgotten in the past is that in calculating Z where our output is less than the input X under the confined space, the weight of input samples which are less important to influence the output of Z are not removed and cause the accuracy of Z matrix. In this work, we have achieved better results by creating constraints. In this method, we have used the subspace learning method and the unsupervised learning method. Due to the used method in this paper, more restrictions were applied on the distance between the input vector (X) and the output vector (z) in the subgroups, which was done by applying more weight to the samples that were more effective, and giving less weight to those that had no effect, and then applying it to the Euclidean distance between the two input and output vectors' subspace. Herein, we attempted to extract the subgroups of the driver's genes more accurately. The results show that the proposed method can extract the driver genes more accurately and realistically compared to the previous methods. here There are many

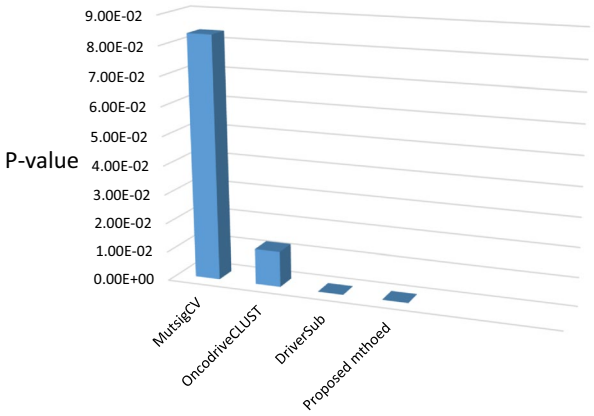

Figure 4. Comparing *p*-values between the previous and the proposed methods for an average of a subset of 200 driver genes.

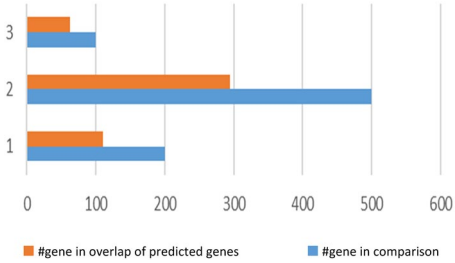

Figure 5. Overlap of the number of suggested genes with specific subgroups of 100, 200 and 500 genes of driver genes.

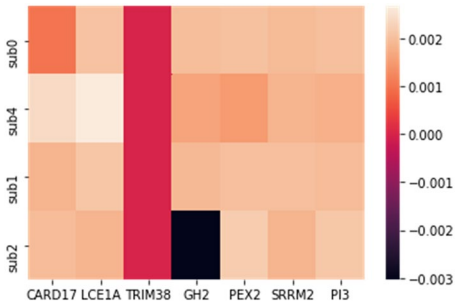

Figure 6. The amount of distribution *p* value of driver genes in the proposed method among specific subgroups.

ideas for researchers to work with in the future, for example in[25,26], the extraction of the characteristics of normal cancer cells through image processing using CNN and deep learning methods[27,28] to isolate healthy cells from cancer, which can be done to identify the driver genes. Due to the openness of article subject, the researchers can achieve more accurate predictions from other methods such as deep learning and combining it with the method in this article. However, using CNN networks with computational complexity and high memory consumption due to the number FLOPS could be due to the volume of input data, and in this case, it should be improved by speeding up CNN through pruning methods. In fact, one of the advantages of using the method used in this article is the low computational complexity and low memory consumption compared to CNN which suffer from memory and computational complexity. Furthermore, in my future work, I decide to use deep learning and convolutional neural network (CNN) with the addition of other information genes in[29,30] such as KEGG pathway and gene transcriptionally changes to more accurately predict specific subgroups of driver genes, Furthermore, using Weighted Gene Co-expression Network Analysis methods in[31] for using in the body of the method of this

paper to calculate the weight between input vector x and output vector z achieved better results. In future work, better results can be obtained for more accurate extraction to further analyze the results on breast cancer data.

Received: 15 December 2020; Accepted: 13 April 2021
Published online: 28 April 2021

References

1. Siegel, R. L. *et al.* Cancer statistics, 2019. *CA Cancer J. Clin.* **69**(1), 7 (2019).
2. Bailey, M. H. *et al.* Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**(2), 371–385 (2018).
3. Meyerson, M. *et al.* Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.* **11**(10), 685–696 (2010).
4. De, S. & Ganesan, S. Looking beyond drivers and passengers in cancer genome sequencing data. *Ann. Oncol.* **28**(5), 938–945 (2017).
5. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**(6127), 1546–1558 (2013).
6. Tokheim, C. J. *et al.* Evaluating the evaluation of cancer driver genes. *Proc. Natl. Acad. Sci.* **113**(50), 14330–14335 (2016).
7. Tamborero, D. *et al.* OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* **29**(18), 2238–2244 (2013).
8. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**(7457), 214–218 (2013).
9. Alizadeh, A. A. *et al.* Toward understanding and exploiting tumor heterogeneity. *Nat. Med.* **21**(8), 846 (2015).
10. Cyll, K. *et al.* Tumour heterogeneity poses a significant challenge to cancer biomarker research. *Br. J. Cancer* **117**(3), 367–375 (2017).
11. Carter, H. *et al.* Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Can. Res.* **69**(16), 6660–6667 (2009).
12. Zhang, J. & Zhang, S. Discovery of cancer common and specific driver gene sets. *Nucleic Acids Res.* **45**(10), e86–e86 (2017).
13. Xi, J. *et al.* Inferring subgroup-specific driver genes from heterogeneous cancer samples via subspace learning with subgroup indication.*Bioinformatics* **36**(6), 1855-1863 (2020).
14. Xi, J. *et al.* Discovering mutated driver genes through a robust and sparse co-regularized matrix factorization framework with prior information from mRNA expression patterns and interaction network. *BMC Bioinform.* **19**(1), 214 (2018).
15. Zheng, R. *et al.* SinNLRR: a robust subspace clustering method for cell type detection by non-negative and low-rank representation. *Bioinformatics* **35**(19), 3642–3650 (2019).
16. Hofree, M. *et al.* Network-based stratification of tumor mutations. *Nat. Methods* **10**(11), 1108–1115 (2013).
17. Wang, K. *et al.* Joint feature selection and subspace learning for cross-modal retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* **38**(10), 2010–2023 (2015).
18. Jolliffe, I. T. & Cadima, J. Principal component analysis: a review and recent developments. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **374**(2065), 20150202 (2016).
19. Ramirez, C. *et al.* Why ℓ1 is a good approximation to ℓ0: a geometric explanation. *J. Uncertain Syst.* **7**(3), 203–207 (2013).
20. He, X. *et al.* Robust adaptive graph regularized non-negative matrix factorization. *IEEE Access* **7**, 83101–83110 (2019).
21. Li, Z. *et al.* Robust structured subspace learning for data representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(10), 2085–2098 (2015).
22. Huang, J., et al. A new simplex sparse learning model to measure data similarity for clustering. In: *Twenty-Fourth International Joint Conference on Artificial Intelligence* (2015)
23. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**(269), l1 (2013).
24. Gsea Msigdb web-based software available in http://www.gsea-msigdb.org/gsea/msigdb/annotate.jsp
25. Iqbal, M. S., Ahmad, I., Bin, L., Khan, S., & Rodrigues, J. J. (2020). Deep learning recognition of diseased and normal cell representation. Trans. Emerg. Telecommun. Technol. e4017.
26. Iqbal, M. S. *et al.* Efficient cell classification of mitochondrial images by using deep learning. *J. Opt.* **48**(1), 113–122 (2019).
27. Iqbal, M. S., Ahmad, I., Asif, M., Kim, S. H. & Mehmood, R. M. Drug investigation tool: identifying the effect of drug on cell image by using improved correlation. *Softw. Pract. Exp.* **51**(2), 260–270 (2021).
28. Liu, M., Li, F., Yan, H., Wang, K., Ma, Y., Shen, L., & Alzheimer's Disease Neuroimaging Initiative. A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease. *NeuroImage*, 208, 116459 (2020).
29. Yu, H. *et al.* LEPR hypomethylation is significantly associated with gastric cancer in males. *Exp. Mol. Pathol.* **116**, 104493 (2020).
30. Chen, J. *et al.* Genetic regulatory subnetworks and key regulating genes in rat hippocampus perturbed by prenatal malnutrition: implications for major brain disorders. *Aging (Albany NY)* **12**(9), 8434 (2020).
31. Li, H. *et al.* Co-expression network analysis identified hub genes critical to triglyceride and free fatty acid metabolism as key regulators of age-related vascular dysfunction in mice. *Aging (Albany NY)* **11**(18), 7620 (2019).

Author contributions

A.R.E., who wrote the main manuscript. His research interests include data mining, bioinformatics, machine learning, Computational Biology and Artificial Intelligence in Medical Applications A.S., who supervised the manuscript and is the corresponding author: Dr. Soleimani's researches blend Computer Science, Virtual Worlds, Virtual Education, Data mining, and Machine Learning. A.G., who reviewed the manuscript: His research focuses on the design, analysis and control of telecommunication networks and embedding distributed intelligence in pure P2P systems, cloud computing and Internet of Things. His corrent interests include: Reliability in IoT, distributed P2P computing networks, modeling and performance evaluation; live streaming systems; and distributed intelligence.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.S.

Reprints and permissions information is available at www.nature.com/reprints.