IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 6, 2022, accepted May 9, 2022, date of publication May 18, 2022, date of current version May 26, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3176376

A Novel Machine Learning Model for Identifying Patient-Specific Cancer Driver Genes

HEEWON JUNG[1], JONGHWAN CHOI[2], (Graduate Student Member, IEEE), JIWOO PARK[1], AND JAEGYOON AHN[1]

[1]Department of Computer Science and Engineering, Incheon National University, Incheon 22012, Republic of Korea
[2]Department of Computer Science, Yonsei University, Seoul 03722, Republic of Korea

Corresponding author: Jaegyoon Ahn (jgahn@inu.ac.kr)

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) through the Ministry of Science and Information and Communication Technologies (ICT), under Grant NRF-2019R1A2C3005212.

ABSTRACT The identification of patient-specific cancer driver genes plays a crucial role in the development of personalized cancer treatment and drug development. Several computational methods have been proposed for identifying patient-specific cancer driver genes, most of which rank driver genes ac-cording to scores calculated from various gene or protein network information. In this paper, we propose a machine learning model for more accurate identification of patient-specific cancer driver genes. The training data for the proposed model is composed of the gene vectors, which indicate the impacts that one gene can have on or receive from all the genes. The gene vector is patient-specific, in other words, one gene can have many gene vectors from many cancer patients. To make gene vectors, first a patient-specific gene network is built using the gene expression data of each cancer patient and gene regulatory network, then modified PageRank is applied to the patient-specific gene network to make the impact matrix, from which gene vectors can be extracted. We used the Random Forest model to train gene vectors to find and discriminate patterns that show how known driver genes affect, or are affected by, other genes. The proposed model was tested through cross validations and independent tests using different sets of known cancer driver genes and six cancer types from The Cancer Genome Atlas (TCGA) data, and showed higher F-scores than existing patient-specific driver gene identification algorithms. The majority of predicted driver genes were rare, and F-scores calculated with these rare genes are higher than or comparable to those of frequently identified driver genes.

INDEX TERMS Patient-specific driver gene prediction, pagerank, machine learning, patient-specific gene network.

I. INTRODUCTION

Identification of cancer driver genes is important because it enables us to have a deeper understanding of cancer, leading to development of superior anti-cancer drugs or therapies. With accumulation of high-throughput genomic and transcriptomic data, many computational methods have been proposed to identify cancer driver genes or mutations, which can be roughly divided into three categories. The first group of methods identifies driver genes or mutations by their frequency [1], [2]. The main disadvantage of these methods is that they cannot find rare driver genes or mutations unless massive amount of data is provided. The second group is based on machine learning models, which learn genetic or transcriptomic patterns of known driver mutations or genes [3]–[6]. While machine learning based approaches have shown to exhibit high accuracy from recent studies, they are limited by their small number of available training data due to a limited number of known cancer driver mutations or genes. The third group adopts various network searching algorithms to the network of genes, such as gene regulatory network or protein-protein interaction network, to identify cancer driver genes [7], [8].

A majority of the above mentioned methods are focused on identifying drivers from cancer cohort studies. However, it is highly likely that individual cancer patients with the same cancer type have heterogeneous cancer drivers [9], [10]. A small fraction of these heterogeneous drivers have high them are well studied, however, most of them are rare and hard to identify [11], [12]. Several methods have been

The associate editor coordinating the review of this manuscript and approving it for publication was Xinyu Du.

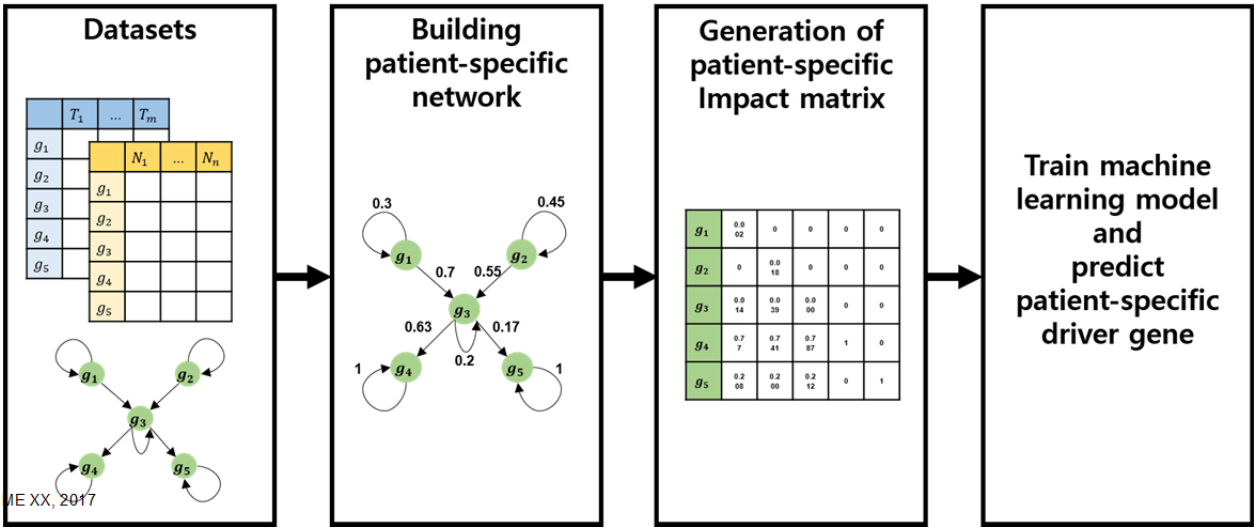

FIGURE 1. Workflow of MPD.

developed to find rare and patient-specific drivers from an individual cancer patient data, and most of the methods are based on network search (third group). DawnRank [13] modified PageRank [14] to apply variable damping factor which is calculated based on the number of incoming edges. Genes are scored using modified PageRank and individual somatic mutation information, and highly ranked genes with somatic mutations are selected as driver genes. Personalized Network Control (PNC) [15] applies Maximum Matching Set algorithm to a patient-specific gene network to find the minimum number of driver genes that affect whole network. The paired Single Sample Network (SSN) constructs a patient-specific gene network whose edges show significant changes in correlation of gene expression between normal and tumor states. Single-sample Controller Strategy (SCS) [16] identifies the minimum number of mutated genes needed to control differentially expressed genes using a method named CTC (Constrained Target Control), which results in several gene modules composed of one driver gene and other downregulated genes. PRODIGY [17] ranks driver genes by calculating the impact of mutated genes on deregulated pathways. These mutated genes are significantly enriched in differentially expressed genes when performing a hypergeometric test.

In this paper, we propose a novel machine learning model named MPD (Machine learning model for Patient-specific Driver gene identification) for identifying patient-specific cancer driver genes. The training data for MPD is composed of the gene vectors. The gene vector of gene G represents the impacts that G can have on, or receives from, all the genes. Gene vectors are patient-specific, which means that the vectors for each gene are made for all cancer patients, and gene vectors of the same gene can be different.

To make the gene vectors, we first construct a gene network for each cancer patient, i.e., a patient-specific gene network. Weights of the patient-specific gene network are calculated using DNA mutations and gene expressions. Then we apply modified PageRank for each gene of the network to make an impact matrix, of which element e_{ij} implies an impact from $gene_j$ to $gene_i$. A gene vector is a concatenation of a column and transpose of a row of a gene in the impact matrix.

Gene vectors for all the patients comprise the training and test data. Because gene vectors are patient-specific, gene vectors of the same gene can be divided into training and test data. This means that known driver genes can also be predicted as cancer driver genes, even if they were already used for training.

Gene vectors of known cancer driver genes with somatic mutation (SM), copy number alteration (CNA) or DNA methylation (DNAm) are assumed to be patient-specific driver genes, and labeled as T. The same number of gene vectors not known to be cancer drivers that do not feature SM, CNA, or DNAm are randomly selected and labeled as F. We used CNA and DNAm in addition to SM data, because they are often associated with cancer driver genes [18], [19]. The machine learning models are trained to classify the gene vectors into T and F. The unlabeled gene vector of gene G is classified by the trained model, and G is predicted to be a patient-specific cancer driver gene if it is classified as T. Because a gene vector is the impacts that a gene gives to and receives from all the genes, classification modeling can find and discriminate patterns that show how a known patient-specific driver affects other genes.

We performed five-fold cross validation and independent tests using known driver genes from Intogen [20], Cancer Gene Census (CGC) [21] and The Network of Cancer Genes (NCG) [22], and SM, CNA and DNAm data of six cancer types (breast, colon, liver, pancreatic, and stomach cancer) in TCGA [23]. Through cross validation, we found that Random Forest (RF) [24] was effective for our purpose, and

that all omics data types showed best results for colon and liver cancer, while the rest of the cancer types were shown best by SM and DNAm data. Through independent tests, we confirmed that MPD shows higher F1 and F0.5 scores than existing methods for identifying patient-specific cancer driver genes. The majority of predicted driver genes are found to be rare, and F1 and F0.5 scores calculated with these rare genes are higher than or comparable to those of frequently identified driver genes.

II. METHODS

A. OVERVIEW

To describe how the proposed model, MPD, works, we first explain how to obtain gene vectors in Section 2.2, and explain how to train gene vectors using the machine learning methods and how to predict patient-specific cancer driver genes in Section 2.3. Fig.1 shows the workflow of MPD.

B. GENERATION GENE VECTORS

1) BUILDING THE PATIENT-SPECIFIC GENE NETWORK

The first step to make a gene vector is to construct the patient-specific gene network. It is built using gene expression data and integrated gene networks made with directed edges of FI networks from Reactom [25] and the gene regulatory networks from the RegNetwork [26] and TRRUST [27]. The patient specific gene network can be represented as a weighted adjacency matrix W for each patient. $W$is calculated using the product of two matrices, Ψ and Φ as in (1).

$$W = (I - \Psi)\,\Phi + \Psi \tag{1}$$

In (1), I is an identity matrix, and Ψ is a diagonal matrix. Each element of Ψ represents differences in gene expression between a tumor sample and a group of normal samples, which is calculated as t-statistics using one sample t-test. T-statistics are then normalized to a range from 0.1 to 0.9 by the min-max scaling method. The value should neither be 0 nor 1 because a value of 0 means that the node had no effect on itself while a value of 1 means that the node was not affected by the gene network. The elements of Ψ are weights of self-loops of a patient-specific gene network.

The matrix Φ is used for calculate weights of non-self-loops, and defined as the elementwise multiplication ($\otimes$) of four matrices, an adjacency matrix A, W_C, W_D and W_P, as in (2).

$$\Phi = A \otimes W_C \otimes W_D \otimes W_P \tag{2}$$

A is an adjacency matrix where $\mathbf{A_{ij}}$**=2 or 1** if i-th and j-th genes are connected in the integrated gene networks, and $A_{ij} = 0$ otherwise. $A_{ij} = 2$ if i-th gene has SM, CNA or DNAm, and $A_{ij} = 1$ otherwise.

W_C and W_D are matrices of which elements are calculated using Pearson's Correlation Coefficient (PCC) as in (3) and (4),respectively.

$$W_C\,[i,j] = |PCC\,(X_c\,[i]\,,X_c\,[j])| \tag{3}$$

$$W_D\,[i,j] = 0.5 \times |PCC\,(X_c\,[i]\,,X_c\,[j]) - PCC\,(X_n\,[i]\,,X_n\,[j])| \tag{4}$$

where X_c, and X_n are matrices of gene expression data of cancer and normal samples, respectively. The value of $W_D[i,j]$ is close to zero if there are similar patterns between cancer and normal samples, and otherwise its value increases up to the maximum of 1.

W_P shows which interactions are particularly crucial in a patient, compared to the other patients. A value of $W_P[i,j]$ is derived from calculation of PCC and gets close to 1 if a patient has similar linear correlation pattern between the i-th and j-th genes compared to PCC of the cancer sample group. (5), as shown at the bottom of the next page. In (5), $X_C[i,k]$ is the value of the i-th gene expression level of the k-th patient; μ_i and σ_i are the mean and standard deviation of the expression level of the i-th gene in the cancer samples, respectively; sgn is a function that returns the sign of an input value, 1 or -1. The sigmoid function is used to give weight of zero to edge (i,j) (i.e. to remove edge), if i-th and j-th genes are not correlated in a specific sample, compared to cancer sample group.

2) GENERATION OF IMPACT MATRIX

Once the patient-specific gene networks are made, we apply modified PageRank to the network to make the impact matrix for each patient. The impact matrix, or IM, is composed of n feature vectors that correspond to the dimension and the number of genes. A feature vector of a specific gene implies the impacts that it has on all the genes, and element e_{ij} of IM implies an impact from $gene_j$ to $gene_i$. Feature vectors calculated by the modified PageRank are used to make gene vectors, which act as the input to the machine learning methods.

To apply the modified PageRank algorithm, a stochastic matrix $\tilde{W}$ of the patient-specific gene weight matrix W is first calculated as (6).

$$\tilde{W} = W \times D^{-1} \tag{6}$$

In (6), D is a diagonal matrix where the i-th diagonal entry is equal to the sum of elements on the i-th column in W, and D^{-1} is the inverse matrix of D. One of the properties of stochastic matrix $\tilde{W}$ is that the sum of elements on each column is 1 and an element $\tilde{W}\,[i,j]$ can be interpreted as the probability with which we proceed to j-th gene from i-th gene.

The modified PageRank algorithm computes feature vectors iteratively using the patient-specific stochastic matrices, as illustrated in Fig.2. An initial feature vector of the i-th gene $IM_0\,[i]$ is a one-hot vector of which the dimension is equal to the number of genes. At the initial time, the i-th entry of the feature vector has value 1 and the other entries have zero, which means that the initial impact matrix IM_0 is an identity matrix. By iteratively multiplying the patient-specific stochastic matrix by the feature vector (7), the positive value of i-th entry spreads to other entries and this process repeats

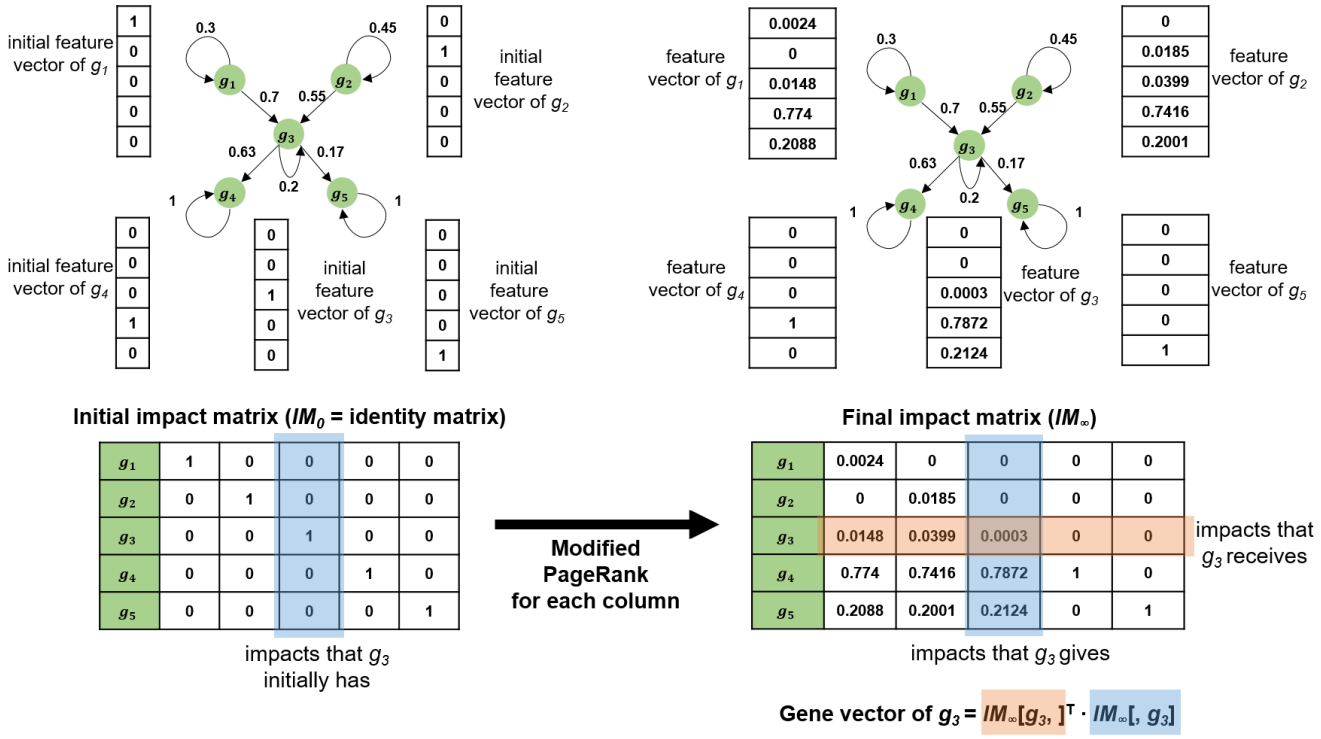

FIGURE 2. Example for generation.

until the feature vector reaches the steady state.

$$IM_{t+1}[i] = \tilde{W} \times IM_t[i], \quad \forall t \geq 0 \tag{7}$$

Equation (7) takes a different approach compared to standard PageRank. First, while the damping factor of PageRank has a fixed value, MPD has a dynamic damping factor – weights of self-loops (Ψ in (1)) can be seen as (1-damping factor). The weights of self-loops are correlated with gene expression differences between cancer samples and normal samples, assuming that genes which have a greater expression difference tend to be affected more by other genes. This suggests that nodes with greater self-loop weights should have smaller impacts on neighboring genes, corresponding to a smaller damping factor.

Second, while initial node values are multiplied by (1-damping factor) for every time point in PageRank, node values at time point t are multiplied by (1-damping factor) to calculate node values of time point $t+1$ in MPD. The reason for this modification is that the initial node values are the same at time 0 in MPD, and we assumed that node's impact from other nodes in the network is accumulated as time flows.

In the steady state, the feature vector $IM_\infty[i]$ shows how much each gene is affected by the i-th gene. The final impact matrix is composed of these feature vectors, and $IM_\infty[i,]$ and $IM_\infty[, i]$ can be interpreted as impacts that the i-th gene receives and the i-th gene gives, respectively.

The gene vector of the i-th gene is defined as the concatenation of $IM_\infty[i,]^T$ and $IM_\infty[, i]$. The reason why we used both row and column of the impact matrix is because driver genes are not necessarily associated with upstream genes in the whole gene network. The example for generation of a gene vector is illustrated in Fig.2.

C. PREDICTION OF PATIENT-SPECIFIC CANCER DRIVER GENE

The gene vector of gene G indicates the impacts that G can have on, or receive from, all the genes, and all the cancer patients have a different gene vector for gene G. In this section, we explain how a machine learning model learns latent information about cancer drivers from gene vectors and determines which gene acts as a cancer driver.

We first get positive and negative gene vectors. The gene vectors are labeled as positive if a gene is known driver in Intogen, CGC, and NCG, and has SM, CNA, or DNAm for each patient. A positive set comprises the positively labeled gene vectors for all the patients.

$$W_P[i,j] = sigmoid\left(\frac{(X_c[i,k]-\mu_i)(X_c[j,k]-\mu_j)}{\sigma_i\sigma_j} \times sgn\left(PCC\left(X_c[i], X_c[j]\right)\right)\right) \tag{5}$$

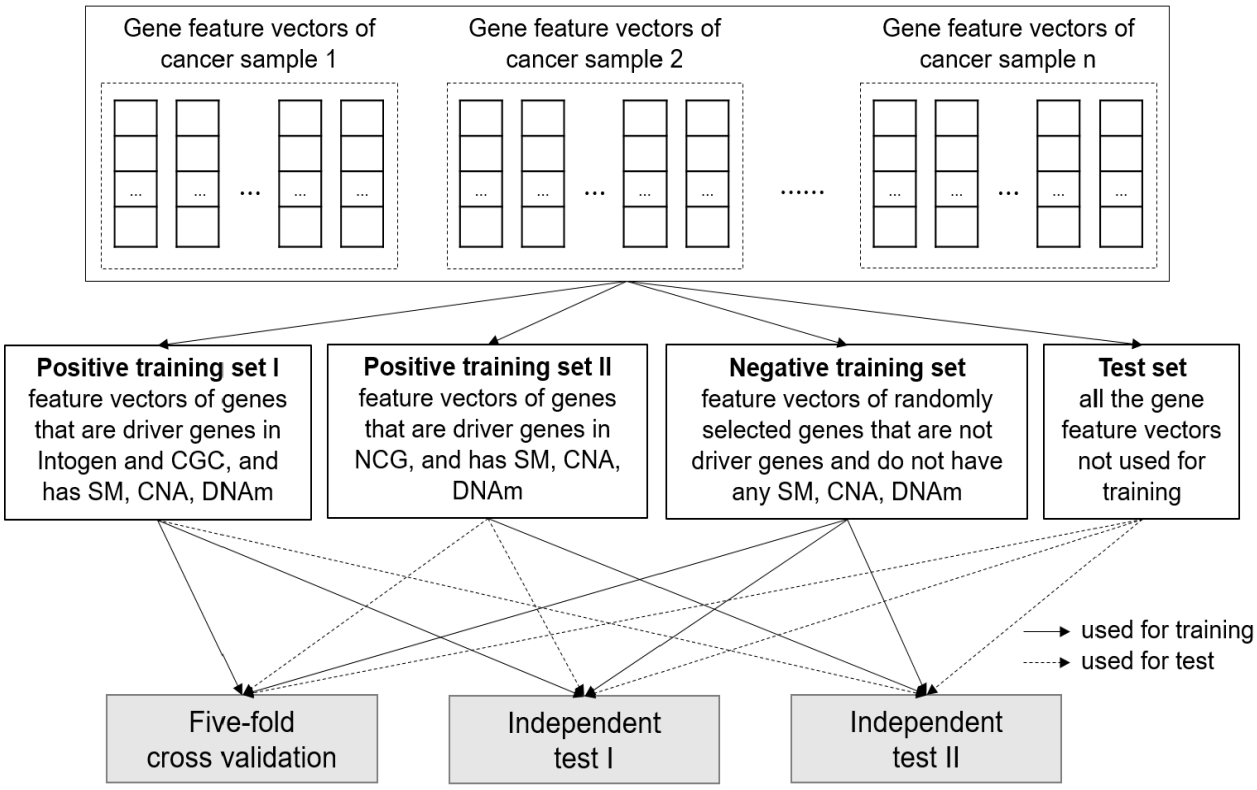

FIGURE 3. Training and test data used for the experiments.

TABLE 1. Detailed description of data downloaded from TCGA and number of driver gene by cancer type.

Cancer type	Number of samples (tumor / normal)	Number of genes	Average number of positive labeled gene vectors[a]	GCG	Intogen	NCG
BRCA	972 / 113	15597	97.1086	54	99	711
COAD	278 / 41	15362	83.9710	58	72	711
LIHC	360 / 50	15002	110.3025	29	31	711
LUAD	510 / 59	15454	98.4370	27	42	711
PAAD	171 / 4	16123	88.0764	32	52	711
STAD	407 / 35	15661	105.9459	35	35	711

* Average number of known driver genes with SM, CNA, or DNAm

For each patient, we randomly select gene vectors from genes that are not known to be driver gene and do not have SM, CNA, and DNAm, and label them as negative. Negatively labeled gene vectors for all the patients compose the negative set.

To build a classification model, we performed five-fold cross validation and two independent tests as in Fig.3. The goal of cross validation is to select a machine learning algorithm, patient-specific gene network construction method, and type of omics data, that will give the best results. The goal of independent tests is to compare the performance of patient-specific driver gene identification with those of existing methods.

For the cross validation, a positive set of each fold consists of gene vectors of one-fifth of the known driver genes. A negative set of each fold consists of gene vectors that are randomly selected genes not known to be drivers. Note that we used only positive set I for cross validation, because driver genes in NCG are not cancer type specific.

For independent tests, all the positive gene vectors not used for training are tested – if positive set I is used for training, positive set II is used for test, and vice versa. Because randomly selected negative set can affect training performance, we made five negative sets and averaged recall and precision.

For both cross validation and independent tests, the same number of gene vectors for the negative and positive set were

TABLE 2. Detailed description of the network data.

	Number of nodes	Number of edges
FI network	14071	110721
Regnetwork	23336	372774
TRRUST	2852	9383
Intergrated network	25167	490200

selected for training. We used RF, Naïve Bayesian classifier (NB) [28] and Deep Neural Networks (DNNs) to train a classification model.

Note that known driver genes can also be predicted as cancer driver genes, even if they were already used for training. This is because gene vectors of known driver genes are patient-specific, which means gene vectors of the same gene are different for different cancer patients. So, even if gene vectors of same genes are used for both training and test, they do not overlap as long as they are not from the same cancer patients.

III. RESULT

A. DATASETS

Four types of omics data (mRNA expression, SM, CNA, and DNAm data) from six cancer types (BRCA, COAD, LIHC, LUAD, PAAD, and STAD) were downloaded from the TCGA data portal. For DNAm data, the top 5% methylation levels of each sample were replaced by 1 and the rest by 0. For gene expression data, genes with zero FPKM value in more than 80% of samples were excluded. The known driver gene information was downloaded from Intogen, CGC, and NCG. Driver genes provided by CGC were divided into two tiers, and only Tier 1 genes were used because driver genes corresponding to Tier 1 had more evidence for cancer occurrence than those in Tier 2. The described datasets are summarized in the Table 1 and the IDs of the omics data that was downloaded from the TCGA portal is provided in supplementary Table 1 by cancer type.

We also downloaded the only directed edges of FI network provided by the Reactom and the gene regulatory networks from the RegNetwork and TRRUST, and integrated them. The details about the integrated network are given in the Table 2.

B. FIVE-FOLD CROSS VALIDATION RESULTS

To choose the best 1) machine learning method, 2) patient-specific gene network construction method, and 3) combination of omics data, we performed a five-fold cross validation as shown in Fig.3.

1) COMPARISON ON DIFFERENT MACHINE LEARNING METHODS

We compared three machine learning models, RF, NB, and DNN in order to find the machine learning method that can best learn the gene vectors we created. We found optimal

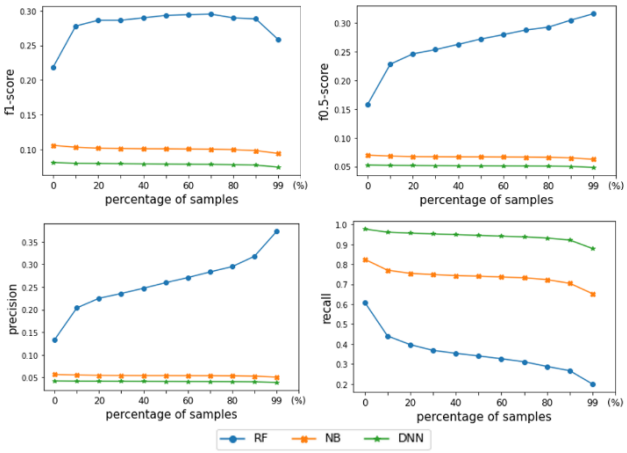

FIGURE 4. Comparison of different machine learning models.

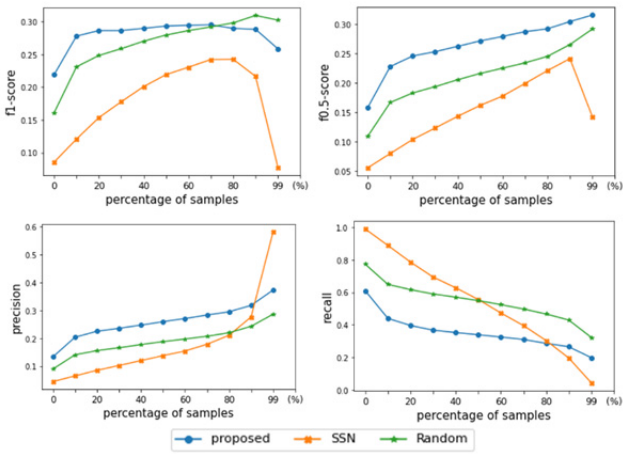

FIGURE 5. Comparison on methods of patient-specific gene network construction methods.

parameters for each method through iterative experiments, and used the parameters with *n_estimator* of 50 for RF, and four hidden layers with size 5,000, 1,000 and 100 were used for DNN.

We calculated recall, precision, F1 score and F0.5 score using the genes selected as driver genes in 0% to 99% of all patient samples for each of six cancer type and averaged them as shown in Fig.4. S1 Fig show F1 score, F0.5 score, precision and recall for each cancer type. Note that 0% means that we use the union of selected driver genes of all the samples, and 100% is not shown because no genes were selected as driver gene in all the samples. We can see that RF shows much

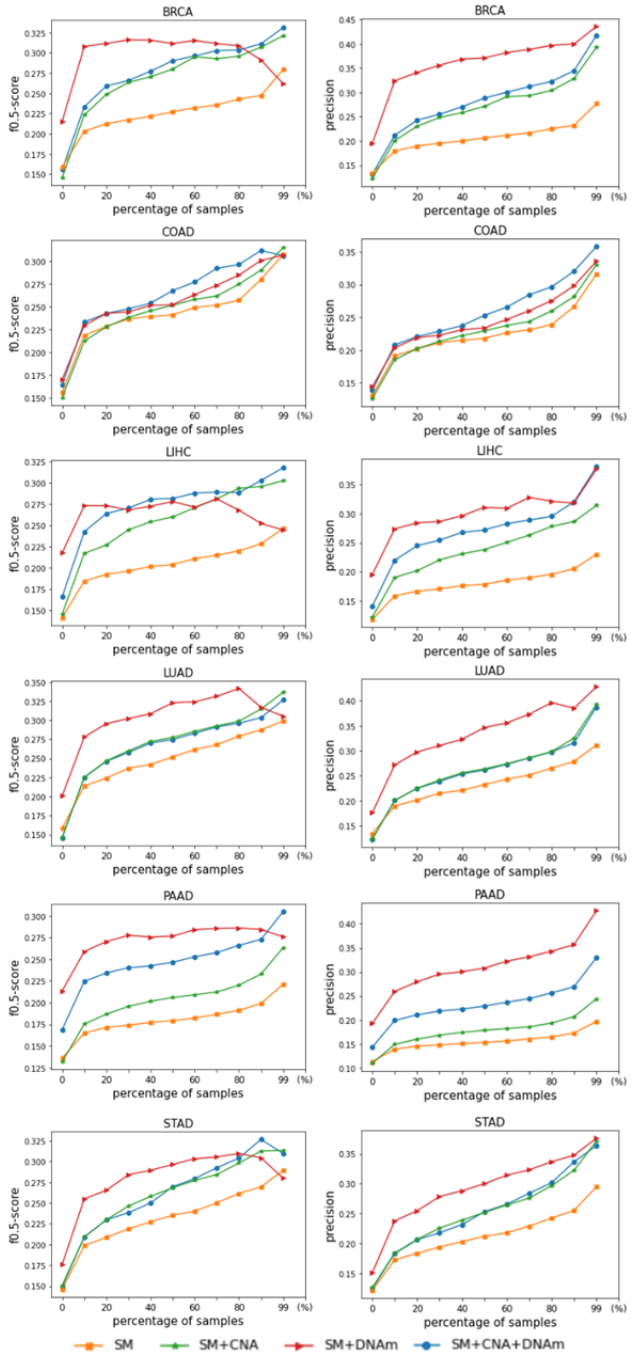

FIGURE 6. Comparison on different omics data types.

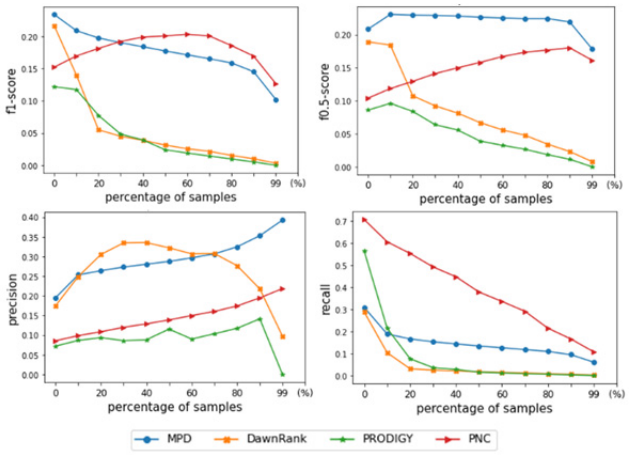

FIGURE 7. Results of independent test I.

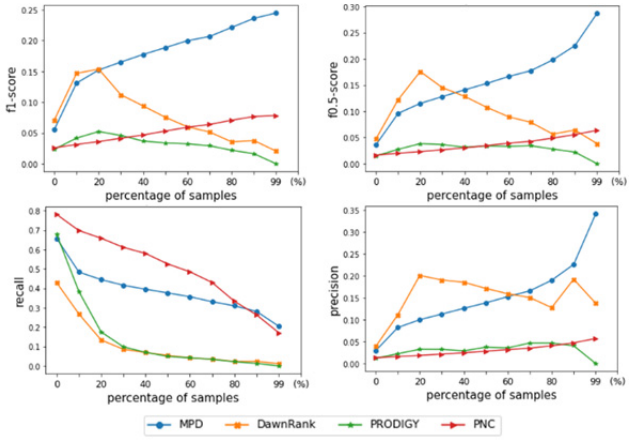

FIGURE 8. Results of independent test II.

higher precision which leads to high F1 and F0.5 scores, so RF was used for the rest of the experiments. The proposed patient-specific gene network construction method was used to build gene networks and all the omics data were used for Fig.4 and S1 Fig.

2) COMPARISON ON METHODS FOR PATIENT-SPECIFIC NETWORK CONSTRUCTION

The differences between patients are represented by patient-specific networks, so accurately constructed patient-specific net-works are very important for accurate identification of the patient-specific driver genes. To show the performance of the proposed patient-specific network construction method, driver genes were found using the proposed network Single Sample Network (SSN) [16], and a random weight network, and F1 score, F0.5 score, precision and recall of each case were compared. Fig.4 shows averaged results, and S2 Fig shows F1 score, F0.5 score, precision and recall for each cancer type. We can see in Fig.4 that the proposed patient-specific network had higher precision than others except 99% of samples, which lead to higher F0.5 score and higher F1 score for less than 70% of samples. All the omics data were used for Fig.4 and S2 Fig.

3) COMPARISON ON METHODS OF DIFFERENT OMICS DATA

We compared the F1 score, F0.5 score, precision and recall when used 1) only SM data, 2) SM and CNV data, 3) SM and DNAm data, and 4) used all data types. F0.5 score and precision for each cancer type are shown in Fig.5, and F1 score, F0.5 score, precision and recall for each cancer type are provided in S3 Fig. But for COAD and LIHC, because using all the data together shows slightly better F0.5 score,

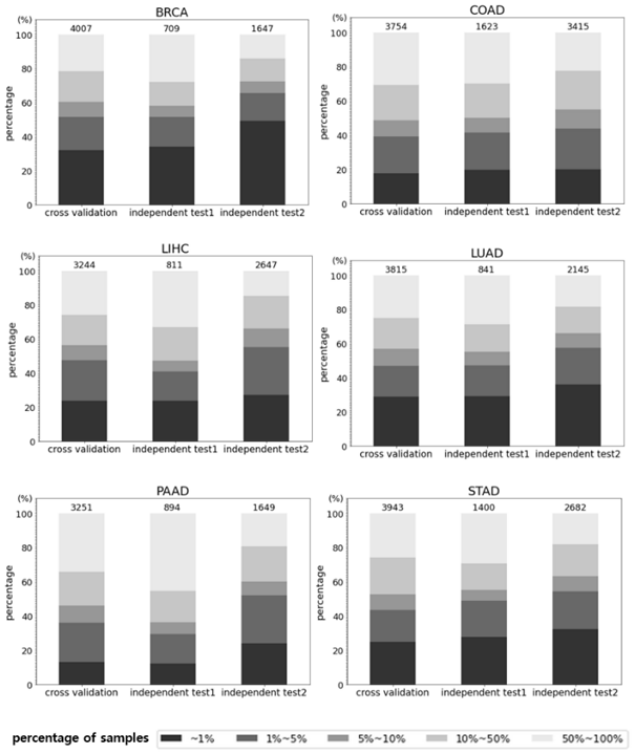

FIGURE 9. Frequency of genes identified as drivers. Graphs show fractions of genes identified as driver genes in less than 1%, 1-5%, 5-10%, 10-50%, and 50-100% of samples. The numbers at the top of each graph indicate total number of genes identified as drivers.

we use all types of omics data for independent tests for COAD and LIHC.

C. INDEPENDENT TEST RESULTS

We performed two independent tests as explained in Fig.2. In independent tests, we compare the F1 score, F0.5 score, precision and recall of MPD, and those of Dawnrank, PNC and PRODIGY. Note that in one other recently published method, SCS was not used in the comparison because we were not able to get its code or executable file. For PNC and PRODIGY, only paired samples were used, because PNC showed better performance for paired samples, and PRODIGY required too much time to get through all the samples. As mentioned in section B, we used all the omics data types for COAD and LIHC, and used SM and DNAm for the rest of the cancer types.

Fig.7 show averaged results for independent test I, and S4 Fig shows F1 score, F0.5 score, precision and recall for each cancer type. We can see in Fig.7 that precision of MPD is higher to or comparable to DawnRank but recall is higher, which leads to highest F1 score when less than 20% of samples are used, and always highest F0.5 score. PNC has highest recall but low precision.

Fig.8 show averaged results for independent test II, and S5 Fig shows F1 score, F0.5 score, precision and recall for each cancer type. Fig.8 also shows that precision of MPD is higher than or comparable to DawnRank, but recall is higher. MPD showed the highest average F1 and F0.5 score when percentage of samples exceeded 20 and 30, respectively. In S5 Fig, we can see that MPD showed good performance for BRCA, LIHC, LUAD and STAD, but has low F1 score, F0.5 score and precision when smaller samples were used. This was especially true for COAD and PAAD, which was likely due to the small number of COAD and PAAD samples, resulting in small number of training gene vectors.

Next, we counted the number of driver genes found for each cancer type to calculate the ratio of rare driver genes. Fig.9 shows the ratio of genes identified as driver genes in less than 1%, 1% to 5%, 5% to 10%, 10% to 50%, and 50% to 100% of samples. We can see that about 20∼40% of genes were identified in ∼1% of samples and 40-50% of genes were identified in ∼5% of samples, which means the majority of genes identified as drivers were rare. In Figu.6 we can see that MPD shows higher F1 and F0.5 scores as less samples were used. In Fig.7, unlike Fig.6, F1 and F0.5 scores increased as more samples were used, and MPD continued to increase relative to others when more than 10% of samples are used. Collectively, these results tell us that MPD successfully identifies rare driver genes with strong accuracy.

IV. DISCUSSION

While existing methods of identifying patient-specific cancer driver genes are usually based on various kinds of network searching algorithms, we proposed a machine learning based method named MPD to reveal patient-specific cancer driver genes. We showed that MPD generally produced higher F1 and F0.5 scores in comparison with existing methods. Compared to DawnRank which frequently shows good performance among existing methods, MPD showed higher or comparable precision but with higher recall.

Expected reasons for good performance of MPD are 1) accurately constructed patient-specific networks, 2) ability of gene vectors to characterize the latent roles of genes in cancer genome, and 3) intrinsic ability of machine learning techniques to find hidden patterns. Machine learning based search can be expected to show high performance in many cases due to recent advances in machine learning methods, but often times, the number of samples is too small. Our work solves this problem of having too few samples by creating a sufficient number of gene vectors for each known driver gene.

Despite its strong results, MPD has some limitations and additional work remains. The main disadvantage of MPD is that it still requires a number of tumor and normal samples to create accurate patient-specific gene networks. Research to overcome this barrier could be the subject of a future study. MPD has another limitation: it does not tell us why a gene is identified as a driver gene because most machine learning models (including Random Forest) are black box models. Interpreting the trained model could be the subject of another future study. In addition, as a classification model, MPD requires negatively labeled samples that are hard to optimize. Because good negative samples can be important

for better prediction performance, we are planning to develop a better way to select negative samples.

REFERENCES

[1] D. Tamborero, A. Gonzalez-Perez, and N. Lopez-Bigas, "OncodriveCLUST: Exploiting the positional clustering of somatic mutations to identify cancer genes," *Bioinformatics*, vol. 29, no. 18, pp. 2238–2244, Sep. 2013.

[2] M. S. Lawrence *et al.*, "Mutational heterogeneity in cancer and the search for new cancer-associated genes," *Nature*, vol. 499, no. 7457, pp. 214–218, 2013.

[3] P. Luo, Y. Ding, X. Lei, and F.-X. Wu, "DeepDriver: Predicting cancer driver genes based on somatic mutations using deep convolutional neural networks," *Frontiers Genet.*, vol. 10, p. 13, Jan. 2019.

[4] J. Nulsen, H. Misetic, C. Yau, and F. D. Ciccarelli, "Pan-cancer detection of driver genes at the single-patient resolution," *Genome Med.*, vol. 13, no. 1, pp. 1–14, Dec. 2021.

[5] C. Arnedo-Pac, L. Mularoni, F. Muiños, A. Gonzalez-Perez, and N. Lopez-Bigas, "OncodriveCLUSTL: A sequence-based clustering method to identify cancer drivers," *Bioinformatics*, vol. 35, no. 22, pp. 4788–4790, Nov. 2019.

[6] H. Yang, Q. Wei, X. Zhong, H. Yang, and B. Li, "Cancer driver gene discovery through an integrative genomics approach in a non-parametric Bayesian framework," *Bioinformatics*, vol. 33, no. 4, pp. 483–490, 2017.

[7] V. V. H. Pham, L. Liu, C. P. Bracken, G. J. Goodall, J. Li, and T. D. Le, "DriverGroup: A novel method for identifying driver gene groups," *Bioinformatics*, vol. 36, no. 2, pp. i583–i591, Dec. 2020.

[8] D. Bertrand, K. R. Chng, F. G. Sherbaf, A. Kiesel, B. K. H. Chia, Y. Y. Sia, S. K. Huang, D. S. B. Hoon, E. T. Liu, A. Hillmer, and N. Nagarajan, "Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles," *Nucleic Acids Res.*, vol. 43, no. 7, p. e44, Apr. 2015.

[9] D. Pe'er and N. Hacohen, "Principles and strategies for developing network models in cancer," *Cell*, vol. 144, no. 6, pp. 864–873, 2011.

[10] M. R. Stratton, "Journeys into the genome of cancer cells," *EMBO Mol. Med.*, vol. 5, no. 2, pp. 169–172, Feb. 2013.

[11] L. Ding, M. C. Wendl, D. C. Koboldt, and E. R. Mardis, "Analysis of next-generation genomic data in cancer: Accomplishments and challenges," *Hum. Mol. Genet.*, vol. 19, no. R2, pp. R188–R196, Oct. 2010.

[12] J. Reimand and G. D. Bader, "Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers," *Mol. Syst. Biol.*, vol. 9, no. 1, p. 637, Jan. 2013.

[13] J. P. Hou and J. Ma, "DawnRank: Discovering personalized driver genes in cancer," *Genome Med.*, vol. 6, no. 7, pp. 1–16, Jul. 2014.

[14] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web," Stanford InfoLab, Stanford, CA, USA, Tech. Rep., 1999.

[15] W.-F. Guo, S.-W. Zhang, T. Zeng, Y. Li, J. Gao, and L. Chen, "A novel network control model for identifying personalized driver genes in cancer," *PLOS Comput. Biol.*, vol. 15, no. 11, Nov. 2019, Art. no. e1007520.

[16] W.-F. Guo, S.-W. Zhang, L.-L. Liu, F. Liu, Q.-Q. Shi, L. Zhang, Y. Tang, T. Zeng, and L. Chen, "Discovering personalized driver mutation profiles of single samples in cancer by network control strategy," *Bioinformatics*, vol. 34, no. 11, pp. 1893–1903, Jun. 2018.

[17] G. Dinstag and R. Shamir, "PRODIGY: Personalized prioritization of driver genes," *Bioinformatics*, vol. 36, no. 6, pp. 1831–1839, Nov. 2019.

[18] E. van Dyk, M. Hoogstraat, J. T. Hoeve, M. J. T. Reinders, and L. F. A. Wessels, "RUBIC identifies driver genes by detecting recurrent DNA copy number breaks," *Nature Commun.*, vol. 7, no. 1, pp. 1–10, Nov. 2016.

[19] Y.-C. Chen, V. Gotea, G. Margolin, and L. Elnitski, "Significant associations between driver gene mutations and DNA methylation alterations across many cancer types," *PLOS Comput. Biol.*, vol. 13, no. 11, Nov. 2017, Art. no. e1005840.

[20] G. Gundem, C. Perez-Llamas, A. Jene-Sanz, A. Kedzierska, A. Islam, J. Deu-Pons, S. J. Furney, and N. Lopez-Bigas, "IntOGen: Integration and data mining of multidimensional oncogenomic data," *Nature Methods*, vol. 7, no. 2, pp. 92–93, Feb. 2010.

[21] Z. Sondka, S. Bamford, C. G. Cole, S. A. Ward, I. Dunham, and S. A. Forbes, "The COSMIC cancer gene census: Describing genetic dysfunction across all human cancers," *Nature Rev. Cancer*, vol. 18, no. 11, pp. 696–705, Nov. 2018.

[22] D. Repana, J. Nulsen, L. Dressler, M. Bortolomeazzi, S. K. Venkata, A. Tourna, A. Yakovleva, T. Palmieri, and F. D. Ciccarelli, "The network of cancer genes (NCG): A comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens," *Genome Biol.*, vol. 20, no. 1, pp. 1–12, Dec. 2019.

[23] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "The cancer genome atlas (TCGA): An immeasurable source of knowledge," *Contemp. Oncol.*, vol. 19, no. 1A, p. A68, 2015.

[24] M. Pal, "Random forest classifier for remote sensing classification," *Int. J. Remote Sens.*, vol. 26, no. 1, pp. 217–222, 2007.

[25] D. Croft *et al.*, "Reactome: A database of reactions, pathways and biological processes," *Nucleic Acids Res.*, vol. 39, no. 1, pp. D691–D697, 2010.

[26] Z.-P. Liu, C. Wu, H. Miao, and H. Wu, "RegNetwork: An integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse," *Database*, vol. 2015, Jan. 2015, Art. no. bav095.

[27] H. Han, J.-W. Cho, S. Lee, A. Yun, H. Kim, D. Bae, S. Yang, C. Y. Kim, M. Lee, E. Kim, S. Lee, B. Kang, D. Jeong, Y. Kim, H.-N. Jeon, H. Jung, S. Nam, M. Chung, J.-H. Kim, and I. Lee, "TRRUST v2: An expanded reference database of human and mouse transcriptional regulatory interactions," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D380–D386, Jan. 2018.

[28] I. Rish, "An empirical study of the naive Bayes classifier," in *Proc. IJCAI Workshop Empirical Methods Artif. Intell.*, 2001, vol. 3, no. 22, pp. 41–46.