Information Technology and Management Science
ISSN 2255-9094 (online)
2023, vol. 26, pp. 10–16
https://doi.org/10.7250/itms-2023-0002
https://itms-journals.rtu.lv

1862
RIGA TECHNICAL UNIVERSITY

Prediction of Cancer Driver Genes Using a Deep Convolutional Network

Natalia Novoselova[1*], Igor Tom[2]
[1,2]*United Institute of Informatics Problems, Minsk, Belarus*

Abstract – The paper describes a method for predicting genes associated with the development of cancer. The method applies the convolutional neural network for the purpose of predicting disease driver genes. Distinctive features of the method are the use of gene expression data to determine the topological structure of the network, the efficiency of prediction with limited information about genes associated with the disease, and the possibility of jointly including information on mutations and similarity of gene expression profiles to improve the accuracy of prediction.

***Keywords* – Driver genes, gene expression, gene mutation, neural network, prediction.**

I. Introduction

In the last decade, many computational methods were proposed to identify cancer driver genes based on biogenetic data. Generally, these methods can be divided into three groups: frequency-based methods, network methods and machine learning based methods. Frequency-based methods identify genes that are significantly hypermutated compared to the background mutation frequency distribution [1]–[2] as driver genes. For example, MutSigCV [1] calculates the statistical significance of mutation rates among all samples to identify driver genes. OncodriveCLUST [2] finds positions with mutation rates higher than the background mutation rates and predicts driver genes using clusters created from these seed positions. However, due to tumour heterogeneity, it is difficult to construct a reliable background mutation model. In addition, these methods cannot be used to detect disease driver genes with low mutation rates. This is explained by the fact that some driver genes mutate at high frequencies (> 20 %), while most cancer driver genes mutate at intermediate (2 %–20 %) or even lower frequencies.

Cancer is a type of complex disease that typically involves multiple gene mutations and dysregulation of various cellular pathways. Therefore, a more effective approach to identifying driver mutations is an approach based on biological networks, including both known regulatory pathways, networks of protein-protein interactions, and networks built on various sources of genetic data. Typically, genes that are located close to each other in biological networks have similar functions and, therefore, the same functional disorders can be a consequence of mutational changes in different genes. Methods based on biological networks can be broadly divided into cluster methods aimed at finding modules of associated genes that cluster together in a network and methods using network diffusion or network propagation [3] to detect altered sub-networks. Both types of methods are based on the fact that genes influencing the same phenotype interact within a network. For example, DawnRank method [4] predicts driver genes by distributing expression information throughout a biological network using the Page Rank algorithm. SCS [5] uses a network strategy to search for driver mutations that can modify the normal functioning of the regulatory network. Although network-based methods have been successfully used to discover cancer driver genes, they are still limited by unreliable and incomplete information about interactions in biological networks. Developing an integrated approach by incorporating multi-omics data (somatic mutations, structural variations, gene expression and methylation) and using hybrid approaches will improve the prediction of disease-causing genes.

As the number of experimentally tested driver genes increases, researchers are using machine learning techniques to predict new candidate genes. Machine learning (ML) methods can include additional information about known driver genes to improve prediction performance.

The main component of the machine learning methods is training data that characterises the samples from different perspectives, and much of the model success depends on these data. A given prediction problem may have its concept represented in many different ways, which is especially true in the genomics domain. A large variety of information is used as model input features for cancer driver prediction. We can classify all the existing data sources of genomic features into five categories based on the properties evaluated as predictors: genomic variation, functional impact, functional genomics, network-based, and ontology-based. Genomic variation category includes mutations and copy number alteration (CNA). The wide use of mutations is based on the hypothesis that driver genes are mutated more frequently than expected as compared to a background mutation rate (BMR) estimated from cancer samples for a given cancer type. Functional impact category includes functional impact (FI) of SNVs on protein function and properties of protein sequence and structure. Functional genomics category includes large-scale gene expression profiles or statistics derived from differential gene

[*] Corresponding author. E-mail: novosel@gmail.com
Article received 29.09.2023; accepted 06.11.2023

expression analyses, epigenomic changes, and protein expression data. The reasons for including these data are based on the fact that mutation frequency across the genome is strongly correlated with transcriptional activity and DNA replication timing, and that driver gene mutations are tightly tied to DNA methylation landscape in multiple types of cancer. Thus, integrating these types of predictors in the ML models may be helpful in differentiating cancer genes from the rest of human genes. The network-based category relates to features extracted from molecular networks, and ontology-based category takes into account the annotations, characterising biological processes such as those provided by Gene Ontology (GO) or the more specific categorisation of gene role in organism functioning and phenotype.

According to literature review [6], genomic variation is the most common feature category, followed by functional impact and functional genomics. When using only one source of data many cancer driver genes will not be discovered because of their high heterogeneity in population. Therefore, the efficient use of additional data can improve the prediction of cancer-driving genes. Based on features integrating protein-protein interactions (PPIs) at the genomic and mutational level, it is possible to identify whether a driver or a passenger is a somatic mutation.

Methods for predicting candidate genes based on machine learning usually use a classification model (Random Forest (RF), Support Vector Machine (SVM), etc.) with features characterising the functional impact of mutations. For example, 20/20+ [7] is a random forest machine learning algorithm for extracting driver genes from somatic mutations. 20/20+ uses features capturing mutational clustering, evolutionary conservation, predicted functional impact of variants, mutation consequence types, gene interaction network connectivity, and other relevant covariates. In [8] SVM and RF machine learning algorithms were selected to induce predictive models to classify supposedly driver genes as real drivers or false-positive drivers based on both mutation data and gene network interactions. Using different data sources to train a classifier can improve prediction accuracy, but simply combining features and increasing the feature space may not be the best approach for integrating different types of data. It is of interest to embed the genomic information into the structure of classification model.

In recent years, deep learning models on graphs (for example, graph neural networks) have been actively used in the field of machine learning to achieve high performance in the field of image processing and computer vision. Deep convolutional models have a neural network architecture that takes advantage of the graph structure and aggregates information about the neighbourhood of network nodes in a convolutional manner [9]. The use of such networks to predict cancer driver genes allows for the integration of multiple data sources, taking into account topological information about the functional similarity of genes available using a network approach. The disease driver gene prediction method proposed in this paper uses a convolutional neural network to simultaneously analyse mutation information and similarity networks, which improves the prediction. A convolutional neural network is trained on a feature matrix based on mutations in genes, built taking into account the similarity of gene expression profiles. The method allows combining two types of biogenetic information due to the two-dimensional organisation of the feature space. It takes into account both topological and functional similarity of the analysed set of genes to identify genes associated with the development of cancer.

II. Method for Predicting Driver Genes Based on a Deep Convolutional Network

A. Deep Convolutional Network Model

Convolutional neural networks (CNN) are multi-layer neural networks. The idea behind convolutional neural networks is to use a “moving filter” or kernel that runs through the image. This moving filter, or convolution, is applied to a specific neighbourhood of nodes. CNN consists of different types of layers: convolutional layers, subsampling layers and layers of a “regular” neural network – a perceptron, in accordance with Fig. 1. The first two types of layers (convolutional, subsampling) alternate with each other, form the input feature vector for a multilayer perceptron.

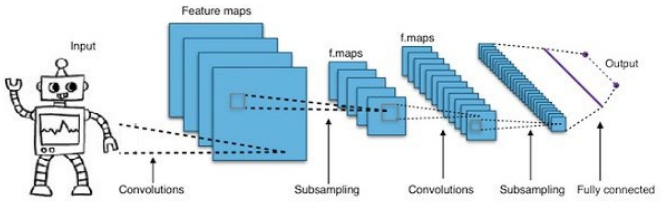

Fig. 1. General architecture of deep convolutional network.

CNN is successful in many areas such as image classification and speech recognition. The CNN model significantly reduces the number of tuning parameters compared to a multilayer perceptron, which can effectively cope with the high dimensionality of raw images. The key component of CNN is the convolutional layer, which helps the model analyse local and global structures of the input data.

In the task of predicting disease driver genes, traditional input data contain features characterising various properties of genes that cannot be directly applied to a CNN. If we provide traditional input data to construct similarity networks in a CNN architecture, we can apply this model for classification and for predicting genes associated with cancer. Figure 2 shows the structure of the one-dimensional CNN that is used in our study. The CNN model consists of five types of layers: input layer, convolution (CONV) layers, pooling layers, Fully-Connected (FC) layers, and output layer. The elements of the input layer are a feature matrix $\phi_i \in R^{2k \times n_f}$, consisting of the gene vector g_i and its k neighbours, where n_f is the dimension of the feature vector g_i. The outputs of the convolutional layer CONV correspond to the input signals ϕ_i when applying the filter ω_i and are calculated as follows:

$$A(i,j) = f\left(\omega_j \phi_i + b_j\right), \quad (1)$$

where b_j represents the bias corresponding to the filter ω_j, f is the activation function, $\omega_j\phi_i$ is the dot product, which is calculated locally. Each convolutional layer is followed by a subsampling layer, and the CONV-POOL pair is repeated several times. The final structure of the CNN model for predicting cancer driver genes is determined using grid search.

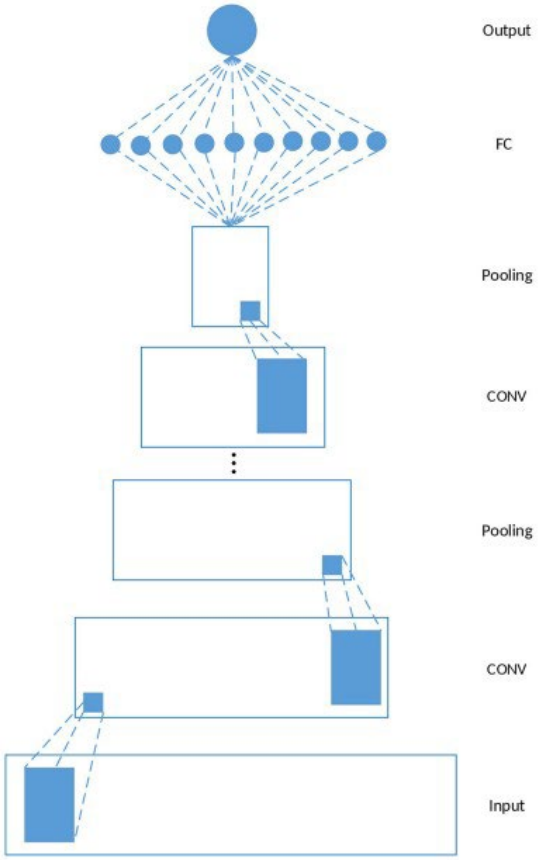

Fig. 2. General scheme of 1D CNN used in the study.

B. Construction of a Feature Space for a Convolutional Network

According to the proposed CNN convolution is performed by combining features based on gene mutation data and gene similarity network. Various approaches can be used to calculate gene similarity. In our study, the Pearson correlation coefficient is used to calculate a similarity measure

$$pcc(g_i, g_j) = \frac{\sum_{q=1}^{v}(e_{iq}-\bar{e}_i)(e_{jq}-\bar{e}_j)}{\sqrt{\sum_{q=1}^{v}(e_{iq}-\bar{e}_i)^2}\sqrt{\sum_{q=1}^{v}(e_{jq}-\bar{e}_j)^2}}, \quad (2)$$

where $e_i = (e_{i1}, e_{i2}, \ldots, e_{iv})$ defines the gene g_i expression values for v disease cases, and $\bar{e}_i$ is the average value of the vector e_i. An undirected network N is constructed using the k nearest neighbours (kNN) algorithm [10], in which each gene is connected to the k genes that have the highest values of *pcc* coefficients.

After defining the network N the construction of the input feature matrix ϕ_i used in convolution is shown in Fig. 3. Assuming that x_i is a feature vector for gene g_i, the combination of genes $g_{s1}, g_{s2}, \ldots, g_{sk}$ determines the k nearest neighbours for gene g_i in network N, where $pcc(g_i, g_{s1}) > pcc(g_i, g_{s2}) > \cdots > pcc(g_i, g_{sk})$. According to Fig. 3 for given values of the gene feature vector g_i, and its k nearest neighbours $g_{s1}, g_{s2}, \ldots, g_{sk}$, the feature matrix ϕ_i is constructed by aggregating $2k$ vectors into a matrix of dimension $2k \times n_f$, which is used as an input to the CNN model.

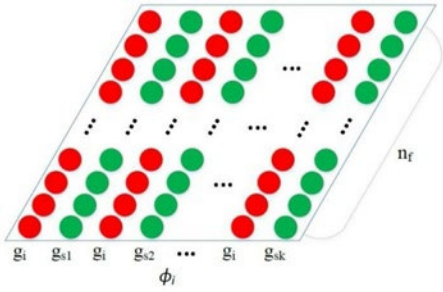

Fig. 3. Scheme of the feature matrix ϕ_i of the CNN input layer.

The feature vector for each gene is based on mutation information and consists of twelve features that are extracted from gene mutation dataset. A number of features are associated with the functional impact of mutations (FIS), i.e., how mutations affect protein function and therefore potentially change phenotype [11]. The SIFT algorithm is used to estimate the FIS mutation coefficient [11]. Table I provides the names and descriptions of all features.

TABLE I
DESCRIPTION OF INPUT FEATURES

№	DESCRIPTION
1	Proportion of silent mutations
2	Proportion of nonsense mutations
3	Proportion of splicing mutations
4	Proportion of missense mutations
5	Proportion of repeated missense mutations
6	Fraction of frameshift mutations
7	Proportion of mutations such as nucleotide insertion and deletion
8	Proportion of start and stop mutations
9	Ratio of missense mutations to silent mutations
10	Ratio of non-silent mutations to silent mutations
11	Normalized entropy of missense mutations
12	Normalized entropy of all mutations

All of the gene features in Table I were previously considered in a number of studies to build machine learning models [7]. These features are used in our study to conduct a comparative analysis of the proposed method with other alternative machine learning methods.

C. Description of a Method for Predicting Genes Associated with Cancer

The proposed driveNet method is implemented based on a deep convolutional network architecture and includes the following steps:

1. Pre-processing of gene expression and mutation data, characterising the cases with malignant oncological diseases.

2. Construction of a gene similarity network. This network determines the topological similarity of genes and allows one to take into account spatial information when constructing a classification and prediction model.

3. Construction of an architecture of CNN convolutional network, which is a model for classifying genes into two classes, namely the class of driver genes that affect the disease and the class of passenger genes, mutations in which are random and not associated with the disease.

4. Construction of the input feature matrix for the CNN convolutional network using the gene similarity network and mutation data.

5. Training the parameters of the CNN convolutional network and assessing the classification efficiency based on the cross-validation and bagging procedure, which allows obtaining an unbiased assessment of the performance criteria and evaluating the generalizing ability of the constructed classification model. Tuning of model parameters such as number of convolutional and subsampling layers, number of elements in layers, number of nearest neighbors for each gene.

6. Assessing the biological significance of driver gene predictions based on cancer databases and literature sources.

D. Testing the Proposed Method

The proposed method for predicting disease genes was tested on two data sets: invasive breast carcinoma (BRCA) and colon adenocarcinoma (COAD). Mutation and gene expression data for the two diseases were obtained from the TCGA cancer genome database [12]. Pre-processing of mutation data included the filtering, namely the exclusion of hypermutated cases (more than 1000 intragenic somatic variants). Gene expression data for BRCA and COAD consisted of 1102 BRCA, 478 COAD primary tumour samples measured using RNA-Seq technology. Three-step filtering of genes that were weakly expressed in tumour samples was performed. In the first step expression values smaller than one were determined to be unreliable and set to zero, then all expression values were log-transformed. In the final step, genes expressed in less than 10 % of samples were removed from consideration. Genes for which data were available in both the mutation dataset and gene expression dataset were selected for analysis. As a result, after quality control 13 777 BRCA genes and 11 282 COAD genes were selected.

Typically, training a supervised machine learning model requires access to labelled data. To obtain gene labels, we used Cancer Gene Census (CGC) data from the COSMIC database [13]. A list of 723 genes that have a causal relationship with cancer was downloaded and this set was selected as the gold standard (driver genes). In our study, both oncogene and tumour suppressor gene are considered as a driver gene. A total of 37 driver genes for BRCA and 42 driver genes for COAD were selected from the CGC. The effectiveness of the driveNet method was assessed in two stages. At the first stage, the proposed method was compared with the 20/20+ driver gene identification method [14] and with the SVM-based method. 20/20+ method was run using default parameters, as described in [14]. For comparison, the classification model efficiency indicator AUC (Area under the ROC curve) was used. The AUC score for each method was obtained using a 10-fold cross-validation procedure. The ROC curve plots the false positive rate (FPR) versus the true positive rate (TPR) at different thresholds. FPR and TPR are defined as

$$FPR = \frac{FP}{FP+TN}$$
$$TPR = \frac{TP}{TP+FN,} \quad (3)$$

where *TP*, *FP*, *TN* and *FN* determine the number of true positive, false positive, true negative and false negative prediction results, respectively.

Due to the fact that the number of randomly mutated genes is much greater than the number of driver genes, it is necessary to solve the problem of imbalance in the data. Our study uses a data-level approach, namely undersampling, which consists of forming a subsample from a larger class equal in size to the small class. According to the proposed approach, a subset of passenger genes is randomly selected from all those available for analysis in such a way that the number of positive marks (driver genes) and negative marks is the same. This approach is applied five times, generating five data sets. In the cross-validation process, for each of the five datasets, all positive and negative samples are randomly divided into ten groups, and the CNN model is evaluated in ten steps.

At the second stage of evaluating the effectiveness of the proposed method, all unknown genes were ranked according to their probability of belonging to the class of driver genes. For the top ten genes in the ranked list, literature sources were searched to test whether our predictions were consistent with existing studies. In the same way, unknown genes were ranked using the SVM and 20/20+ methods and the results were compared with those obtained by the driveNet method.

The method was implemented using the open Keras library [15], which provides interaction with artificial neural networks. Keras is aimed at quickly working with deep learning networks and supports various neural network libraries, including TensorFlow. In the present study, the CNN architecture is defined using the following hyperparameters:

1. Number of CONV layers (ncl);
2. Number of FC layers (nfl);
3. Number of CONV layer nodes (ncn);
4. Number of FC layer nodes (nfn).

These hyperparameters were determined using a grid search, where the parameter ncl took values from the set {1,2,3,4}, nfl – from the set {1,2,3}, ncn – from the set {12,24,48} and nfn – from the set {24,48,96}. The optimal values of the parameters ncl, nfl, ncn, and nfn were set to be 2, 1, 24, and 48, respectively. The number of neighbours used in the *k*-nearest neighbours algorithm was also determined using a grid search, where *k* was selected from the set {3,5,7,9,11,13,15}. As a result, $k = 9$ and $k = 7$ were respectively selected for the BRCA and COAD datasets.

According to [14], a forest of trees consisting of 200 elements was used to implement the 20/20+ method. The SVM-based

method used linear and RBF kernel functions. A search was made for the optimal value of the penalty parameter C from the set {0.1,0.01,0.001,1,10,100,1000} and the parameter γ from the set {1⁄12,0.001,0.0001,0.00001}. As a result, for the BRCA and COAD datasets the SVM model showed the best results using the RBF kernel at $C = 1$, $\gamma = 0.0001$.

Figures 4 and 5 present the results of constructing ROC curves and the corresponding AUC values obtained by the proposed driveNet method and the 20/20+ and SVM methods for the BRCA and COAD data sets, respectively.

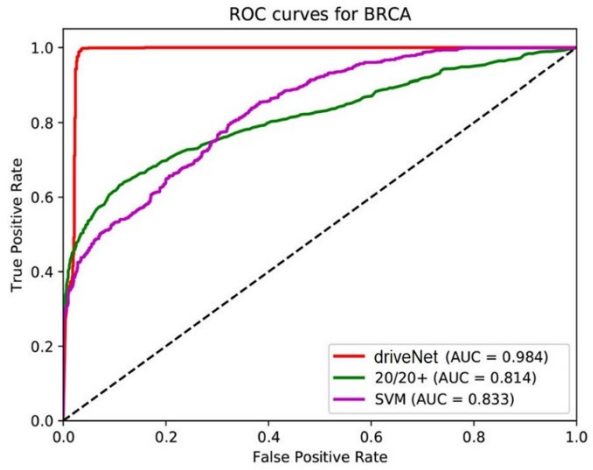

Fig. 4. ROC curves for the three methods obtained for the BRCA data set.

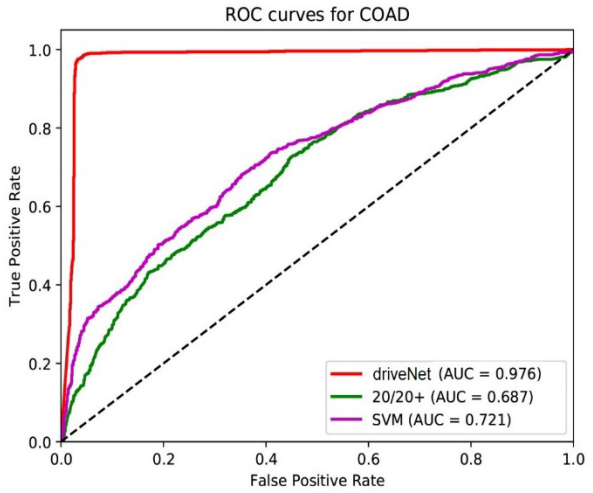

Fig. 5. ROC curves for the three methods obtained for the COAD data set.

In Figs. 4 and 5 the red, green and purple lines correspond to the ROC curves for the driveNet, 20/20+ and SVM methods, respectively. The AUC value for the driveNet is 0.984 for BRCA, which is 15.1 % higher than the corresponding values for the other two methods. The AUC value for the driveNet is 0.976 for COAD, which is 25.5 % higher than the corresponding values for the other two methods.

The proposed method was compared with the 20/20+ and SVM and the proportion of driver genes predicted as such for the second set of cancer driver genes from [14] was assessed. Figure 6 shows the average values and standard deviations of the proportion of driver genes from [14] predicted for two data sets. According to Fig. 6, the proposed method is the most accurate and 59 % and 61 % of driver genes are predicted for the BRCA and COAD datasets, respectively.

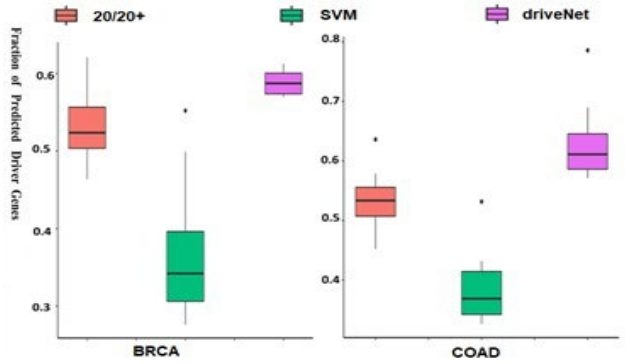

Fig. 6. Fraction of predicted driver genes from those known according to [14].

In order to evaluate the biological effectiveness of the proposed driveNet method, unknown genes were ranked according to the probability of belonging to the class of driver genes. As in the cross-validation procedure described previously, five datasets were used to train the model and unknown genes were ranked according to the average probabilities. The same ranking was performed using two alternative methods. The results were assessed by the number of genes that were investigated as drivers in the existing literature. The analysis used the CancerMine database [16], which is a specialised database for text mining in literature sources, and through which a number of significant genes were studied based on current literature reviews.

TABLE II
DESCRIPTION OF THE TOP TEN PREDICTED DRIVER GENES

BRCA		COAD	
GENE	MENTIONED IN LITERATURE	GENE	MENTIONED IN LITERATURE
PTEN	+	AMER1	
HCFC1	+	SOX9	+
UTRN	+	NRAS	+
ZNF517		MTOR	+
STAG2	+	ATM	+
ZFP36L1	+	ADAMTSL3	
ZNF91		ELMO1	+
VPS13C		TG	
DST		LAMA3	
FBXW7	+	KMT2A	

Table II shows the ten driver genes for BRCA and COAD predicted using the driveNet method. Six of the ten genes in the first column have been reported in the literature as potential driver genes for BRCA. It has been discovered that the DST gene has the potential to transform a tumour in the initial stages of development into a malignant breast tumour. Five of the ten

genes in the second column of Table II have been reported as drivers of COAD in the existing literature. Among the remaining five genes presented in Table II, AMER1 and ADAMTSL3 are frequently mutated genes in COAD. The LAMA3 gene has been predicted as a biomarker for the diagnosis of COAD in early developmental stages. KMT2A refers to the KMT2 family, which is related to COAD.

According to the prediction results for two compared methods, the proposed method is the most effective in predicting new cancer-related genes.

III. Discussion

The use of machine learning algorithms to predict cancer driver genes has enabled important scientific advances and has an excellent potential to go further. Nonetheless, to accelerate the computational discovery of new cancer drivers, several challenges remain to be explored. The prediction of driver genes and driver mutations is an inherent class imbalance problem due to the low number of known drivers compared to the passenger ones. Most supervised algorithms work more effectively for balanced train sets. It would be an advantage to apply a semi-supervised learning strategies to take into account unknown mutations. There is also a large body of works using network analysis without interface with ML. Thus, a natural direction in this domain is the use of graph-based ML algorithms, which can directly process graph-structured data in an end-to-end manner without the need for complex feature engineering. Graph Neural Networks (GNNs) can better explore network topological information for node-level or edge-level prediction tasks in contrast to traditional network analysis. Integrating multi-omics data and PPI networks into a learning framework based on Graph Convolutional Networks (GCNs) is promising in obtaining more accurate models than tools exploring network-based analysis, ML-based classification of omics data, or a combination of both. Further exploration of GNNs with multimodal data, proposing strategies to improve robustness to structural noise and class imbalance, may enable precise prediction of cancer driver genes.

IV. Conclusion

This article describes a method for predicting cancer-related genes based on deep convolutional network. The method allows combining convolutional networks as a classification model with gene similarity networks in a way that simultaneously takes into account the functional impact of mutations and the similarity of gene expression profiles. Integrating two data sources improves the accuracy of disease driver gene prediction. Distinctive features of the method are the use of gene expression data to determine the topological structure of the network, the efficiency of predicting driver genes with limited information about genes associated with the disease, the possibility of jointly including information on mutations and gene expression to improve the efficiency of classification model. Experiments performed on cancer data showed the advantages of the proposed method compared to analogues in terms of prediction efficiency using both cross-validation and de novo predictions (for unknown genes).

References

[1] 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, Jun. 2013. https://doi.org/10.1038/nature12213

[2] 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. https://doi.org/10.1093/bioinformatics/btt395

[3] L. Cowen, T. Ideker, B. J. Raphael, and R. Sharan, "Network propagation: a universal amplifier of genetic associations," *Nature Reviews Genetics*, vol. 18, no. 9, pp. 551–562, Jun. 2017. https://doi.org/10.1038/nrg.2017.38

[4] J. P. Hou and J. Ma, "DawnRank: discovering personalized driver genes in cancer," *Genome Medicine*, vol. 6, no. 7, Jul. 2014, Art. no. 56. https://doi.org/10.1186/s13073-014-0056-8

[5] W. F. Guo *et al*., "Discovering personalized driver mutation profiles of single samples in cancer by network control strategy," *Bioinformatics*, vol. 34, no. 11, pp. 1893–1903, Jun. 2018. https://doi.org/10.1093/bioinformatics/bty006

[6] F. Li *et al*., "Effects of multi-omics characteristics on identification of driver genes using machine learning algorithms," *Genes*, vol. 13, no. 5, Apr. 2022, Art. no. 716. https://doi.org/10.3390/genes13050716

[7] C. J. Tokheim, N. Papadopoulos, K. W. Kinzler, B. Vogelstein, and R. Karchin, "Evaluating the evaluation of cancer driver genes," in *Proceedings of the National Academy of Sciences*, vol. 113, no. 50, Nov. 2016, pp. 14330–14335. https://doi.org/10.1073/pnas.1616440113

[8] J. F. Cutigi *et al*., "Combining mutation and gene network data in a machine learning approach for false-positive cancer driver gene discovery," in *Advances in Bioinformatics and Computational Biology: 13th Brazilian Symposium on Bioinformatics, BSB 2020*, São Paulo, Brazil, Nov. 23–27, 2020, pp. 81–92. https://doi.org/10.1007/978-3-030-65775-8_8

[9] S. Zhang, H. Tong, J. Xu, and R. Maciejewski, "Graph convolutional networks: a comprehensive review," *Computational Social Networks*, vol. 6, no. 1, Nov. 2019, Art. no. 11. https://doi.org/10.1186/s40649-019-0069-y

[10] T. Cover and P. Hart, "Nearest neighbor pattern classification", *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967. https://doi.org/10.1109/TIT.1967.1053964

[11] P. C. Ng and S. Henikoff, "SIFT: Predicting amino acid changes that affect protein function," *Nucleic Acids Research*, vol. 31, no 13, pp. 3812–3814, Jul. 2003. https://doi.org/10.1093/nar/gkg509

[12] National Cancer Institute, "The cancer genome atlas (TCGA)". [Online]. Available: http://cancergenome.nih.gov/abouttcga. Accessed on: Sep. 6, 2023.

[13] Z. Sondka *et al*., "The COSMIC cancer gene census: describing genetic dysfunction across all human cancers," *Nature Reviews Cancer*, vol. 18, no. 11, pp. 696–705, Oct. 2018. https://doi.org/10.1038/s41568-018-0060-1

[14] M. H. Bailey *et al*., "Comprehensive characterization of cancer driver genes and mutations," *Cell*, vol. 173, no. 2, pp. 371–385, Apr. 2018. https://doi.org/10.1016/j.cell.2018.02.060

[15] A. Gulli and S. Pal, "Deep learning with Keras," Packt Publishing Ltd, 2017, 318 p. https://www.packtpub.com/product/deep-learning-with-keras/9781787128422.

[16] J. Lever, E. Y. Zhao, J. Grewal, M. R. Jones, and S. J. M. Jones, "CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer," *Nature Methods*, vol. 16, no 6, pp. 505–507, May 2019. https://doi.org/10.1038/s41592-019-0422-y

Natalia Novoselova received the PhD degree in Computer Sciences from the United Institute of Informatics Problems (UIIP), National Academy of Sciences of Belarus (NASB) in 2008.
Since 2000, she has been a Senior Scientific Researcher at the Laboratory of Bioinformatics, United Institute of Informatics Problems NASB in Minsk, Belarus. Her research interests include data mining methods, notably the neural network, genetic algorithms and fuzzy systems and their application to analysis of medical and biological data. She is an author of more than 30 research publications in these areas.
Contact information: Department of Bioinformatics, United Institute of Informatics Problems, Surganova Str. 6, Minsk, 220012, Belarus. Phone: +375-17- 3485092.
E-mail: novos65@gmail.com

Igor Tom received the PhD degree in Computer Science from the Institute of Engineering Cybernetics, National Academy of Sciences of Belarus (NASB) in 1986.
Since 1999, he has been the Head of the Bioinformatics Department at the United Institute of Informatics Problems of NASB. His current research interests are in the fields of the development of intelligent methods of data analysis and information technologies for medical and industrial applications. He is the author and co-author of more than 170 scientific publications, including 50 full papers in journal.
E-mail: ietom143@gmail.com

.