frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 04 August 2020
doi: 10.3389/fbioe.2020.00766

Pan-Cancer Classification Based on Self-Normalizing Neural Networks and Feature Selection

Junyi Li[1†], Qingzhe Xu[1†], Mingxiao Wu[1], Tao Huang[2] and Yadong Wang[1*]*

[1] Department of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China, [2] Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China

Cancer is a one of the severest diseases and cancer classification plays an important role in cancer diagnosis and treatment. Some different cancers even have similar molecular features such as DNA copy number variant. Pan-cancer classification is still non-trivial at molecular level. Herein, we propose a computational method to classify cancer types by using the self-normalizing neural network (SNN) for analyzing pan-cancer copy number variation data. Since the dimension of the copy number variation features is high, the Monte Carlo feature selection method was used to rank these features. Then a classifier was built by SNN and feature selection method to select features. Three thousand six hundred ninety-four features were chosen for the prediction model, which yields the accuracy value is 0.798 and macro F1 is 0.789. We compared our model to random forest method. Results show the accuracy and macro F1 obtained by our classifier are higher than those obtained by random forest classifier, indicating the good predictive power of our method in distinguishing four different cancer types. This method is also extendable to pan-cancer classification for other molecular features.

Keywords: cancer classification, pan-cancer, self-normalizing neural network, copy number variation, feature selection

OPEN ACCESS

Edited by:
Min Tang,
Jiangsu University, China

Reviewed by:
Guimin Qin,
Xidian University, China
Jie Yan,
Yale University, United States
Mao Peng,
Westerdijk Fungal Biodiversity Institute, Netherlands

Correspondence:
Tao Huang
tohuangtao@126.com
Yadong Wang
ydwang@hit.edu.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *05 May 2020*
Accepted: *17 June 2020*
Published: *04 August 2020*

Citation:
Li J, Xu Q, Wu M, Huang T and Wang Y (2020) Pan-Cancer Classification Based on Self-Normalizing Neural Networks and Feature Selection. Front. Bioeng. Biotechnol. 8:766. doi: 10.3389/fbioe.2020.00766

BACKGROUND

Cancer is a one of the severest diseases which cause abnormal cell growths or tumors that metastasize to other parts of human body (Mayer et al., 2017). There are around 8 million human deaths related to cancer each year (Wild et al., 2014). Cancer classification is important for cancer diagnosis and drug discovery and can help improving treatment of patients and their life quality (Lu and Han, 2003). To decrease the effect of cancer to human health, tremendous research has been done to the cancer diagnosis and treatment, among which molecular-feature-based cancer classification is an important perspective. Due to the drop in the cost of sequencing technology in recent years, the output of sequencing data has increased dramatically. This provides adequate data for cancer analysis. Copy number variance (CNV) has also been shown to be associated with different cancers (Greenman et al., 2007; Wang et al., 2013). Some different cancers even have similar CNV patterns and mechanisms (Hoadley et al., 2018). We focus on CNV data analysis in this study. We aim to find out an applicable computational method to classify different cancer types. At present, some machine learning models are widely used in data analysis. Some models have been used to analyze the CNV data for cancer analysis (Ostrovnaya et al., 2010; Ding et al., 2014). The utility of machine learning in revealing relationships between recurrent constitutional CNVs and cancers shows CNV data analysis is applicable to multi-type of cancers with a significant molecular component.

Deep learning has recently been widely used in computational scientific areas such as computer vision, natural language processing, computational biology (LeCun et al., 2015; Najafabadi et al., 2015; Angermueller et al., 2016; Sultana et al., 2020). The essence of deep learning algorithms is the domain independent idea of using hierarchical layers of learned abstraction to efficiently accomplish a complicated task. It uses many layers of convolutional or recurrent neural networks. The feed-forward neural network (FNN) is suitable for data without sequential features. However, there are some drawbacks of the FNNs. For instance, internal covariate shift (Ioffe and Szegedy, 2015) might causes the low training speed and poor generalization (Bengio et al., 1994; Pascanu et al., 2012, 2013). FNN might leads to invalid gradient too (Klambauer et al., 2017). Therefore, normalization is used and the self-normalizing neural network (SNN) (Klambauer et al., 2017) is proposed to overcome these short backs. SNNs make it possible for deep network applications on general data such as sequencing CNV data and SNNs have yielded the best results on some drug discovery and astronomy tasks.

In this study, we use a SNN-based prediction model to classify and analyze cancer patients with four cancers (LUAD, OV, LIHC, and BRCA). The data we used come from CNV data of The Cancer Genome Atlas (TCGA) (Grossman et al., 2016). We integrate a method which was used by Pan et al. (2018) to identify atrioventricular septal defect in Down syndrome patients to build our prediction model. Since the CNV data has a very high dimension, feature selection method is applied to identify important CNV features. Then a deep SNN model is trained based on these CNV features to perform pan-cancer classification. The normally used classification algorithm random forest (Cutler et al., 2012) is also used to compare with our model for its predictive ability in four different types of patient samples.

METHODS

Data Retrieval and Preprocessing

We download and collate the copy number variation data of 518 Lung adenocarcinoma (LUAD) patients, 597 Ovarian serous cystadenocarcinoma (OV) patients, 372 Hepatocellular carcinoma (LIHC) patients and 597 Breast cancer (BRCA) patients from TCGA database (Grossman et al., 2016), including the information of the copy number variation of probes. We use GISTIC2.0 (Mermel et al., 2011) to analyze the data. GISTIC2.0 can identify the key drivers of somatic copy number alterations (SCNAs) by the frequency and magnitude of mutation events. By using GISTIC2.0, we can select more important copy number variant genes, and then model the molecular information data of cancer patients more precisely. From the result generated form GISTIC2.0, we get a table which has 23,109 features. A series of discrete values is used to represent the specific type of copy number variation.

Approach for Cancer Classification

Feature Analysis

Since the dimensions of the CNV features are high, in order to avoid over-fitting, we need to select some features that can effectively classify patients. Therefore, we employed Monte Carlo Feature Selection (MCFS) (Draminski et al., 2008) and Incremental Feature Selection (IFS) methods as we used these two method before (Pan et al., 2018).

Monte Carlo Feature Selection method is proposed to improve a feature ranking obtained from an ensemble of decision trees. The general idea is to select s subset of the original d features, each with m features randomly selected. We repeat the selection process for s times, so that s feature subsets and a total of $t \times s$ tree classifier was obtained. Each feature f is assigned a score called relative importance (RI_f) which is assigned greater to feature f if it contributes more in the classification using the tree classifiers. RI of f is estimated by the Equation (1):

$$RI_f \;=\; \sum\nolimits_{\tau=1}^{s^*t} (wAcc)^u \sum\nolimits_{n_f(\tau)} IG(n_f\,(\tau)) \, \left(\frac{no.\ in\ n_f(\tau)}{no.\ in\ \tau}\right)^v \quad (1)$$

$\boldsymbol{wAcc}$ is the weighted accuracy and $\boldsymbol{IG}(\boldsymbol{n_f}\,(\boldsymbol{\tau}))$ is the information gain of node $\boldsymbol{n_f}(\boldsymbol{\tau})$. no.in $\boldsymbol{n_f}(\boldsymbol{\tau})$ is the number of patients in $\boldsymbol{n_f}(\boldsymbol{\tau})$ and no.in $\boldsymbol{\tau}$ is the number of patients in tree $\boldsymbol{\tau}$. u and v are a fixed real number.

The $\boldsymbol{wAcc}$ is defined by Draminski as Equation (2):

$$wAcc = \frac{1}{c} \sum\nolimits_{i\,=\,1}^{c} \frac{n_{ii}}{n_{i1}+n_{i2}+\ldots+n_{ic}} \quad (2)$$

In Equation (2), c is the number of classes and $\boldsymbol{n_{ij}}$ is the number of patients from class i that are classified as class j. The $\boldsymbol{IG}(\boldsymbol{n_f}\,(\boldsymbol{\tau}))$ is defined by Equation (3):

$$IG\left(n_f\,(\tau)\right) = Entropy\,(T)\,-Entropy\left(T,\,f\right) \quad (3)$$

In Equation (3), T is the class label of node $\boldsymbol{n_f}(\boldsymbol{\tau})$, Entropy(T) is the entropy of the frequency table of T and Entropy (T, f) is the entropy of the frequency table of the two variables T and f.

We used the MCFS method of Draminski and obtained a ranked feature list according to their RI values evaluate by the algorithm, which can be defined as Equation (4).

$$F = \left[f_1,\, f_2,\, \ldots,\, f_M\right] \quad (4)$$

And in Equation (4) M means the 23,109 CNV features.

Then we aimed to select a subgroup of CNV features to build a classification model. Therefore, in order to avoid training all CNV feature sets, we used Incremental Feature Selection method on previous obtained feature list. We first determine the approximate feature interval from which we can find optimal features. We defined CNV feature subsets as $S_1^1, S_2^1, \ldots, S_l^1$, where $S_i^1\; = f_1, f_2, \ldots, f_{i^*k}$, i.e., and the ith feature subset had the first i

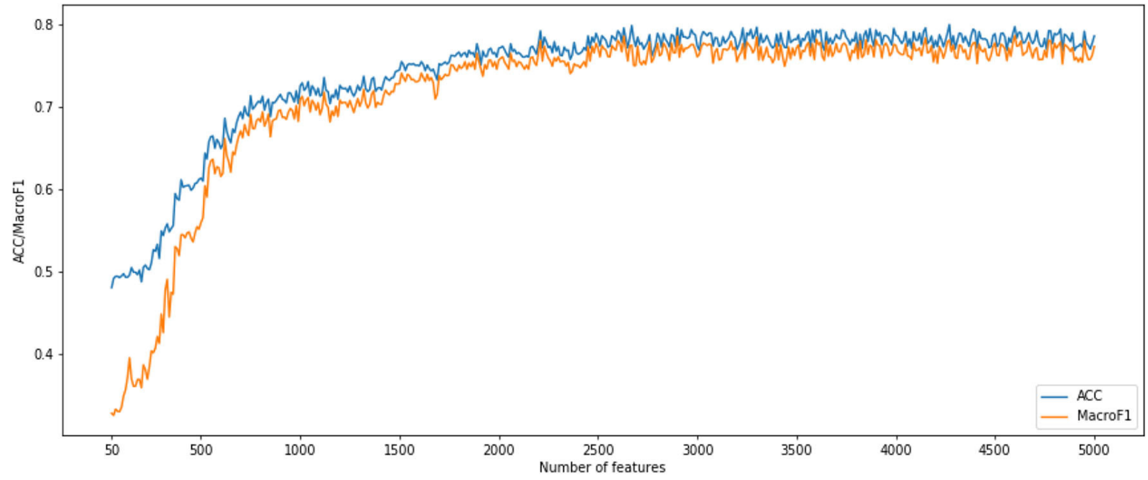

FIGURE 1 | Incremental feature selection (IFS) curves derived from the IFS method and SNN algorithm. IFS curve with *X*-values from 50 to 5,000.

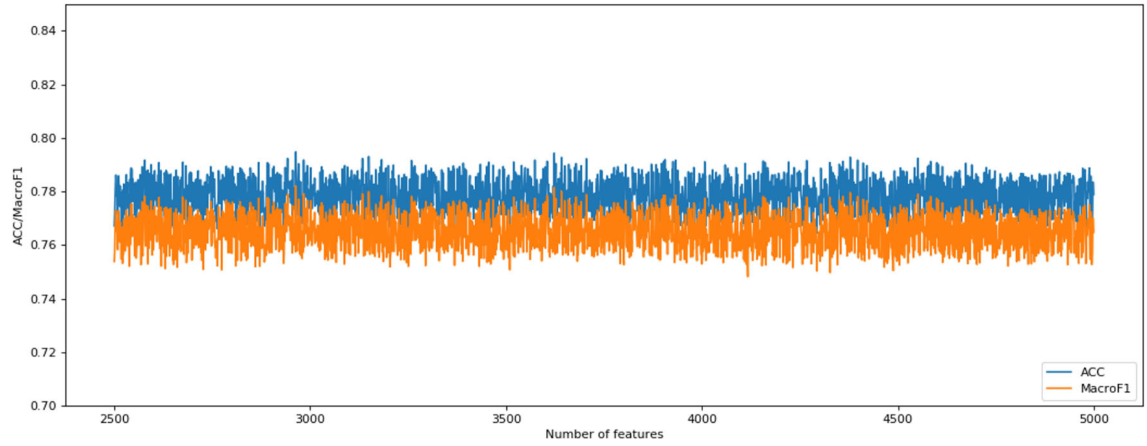

FIGURE 2 | Incremental feature selection (IFS) curves derived from the IFS method and SNN algorithm. IFS curve with X-values of 2501–4,999 for SNN algorithm.

times k features in the original M CNV feature list. Classification model was built by using features in each feature subset of corresponding patient samples in dataset. To estimate the CNV feature interval, we tested performances of different classification model based on different subsets. The feature subset was selected when it had the best performance.

Classification Methods

We need an algorithm to classify pan-cancer patients based on the selected subset of CNV features. Here, neural network SNN was used and RF method was applied for comparison.

(a) Self-Normalizing Neural Network Algorithm

SNN is proposed to enable high-level abstract representations through keeping neuron activations converge toward zero mean and unit variance (Klambauer et al., 2019). Klambauer et al. proposed a Scale ELU (SELU) function as activation function.

$$selu\,(x) = \lambda \begin{cases} x, & x > 0 \\ \alpha e^{x} - \alpha, & x \leq 0 \end{cases} \tag{5}$$

where scale $\lambda = 1.0507$ and $\alpha = 1.6733$ (see Klambauer et al., 2017 for details on the derivation of these two parameters).

By using the Banach fixed-point theorem, Klambauer et al. prove that activations close to zero mean and unit variance that are propagated through many network layers will converge toward zero mean and unit variance. A specific method to initialize SNNs and alpha dropout (Klambauer et al., 2017) are also proposed to make SNNs have a fixed point at zero mean and unit variance. In this study, the SNN classifiers those we constructed have three hidden layers with 200 hidden nodes of each layer.

(b) Random Forest Algorithm

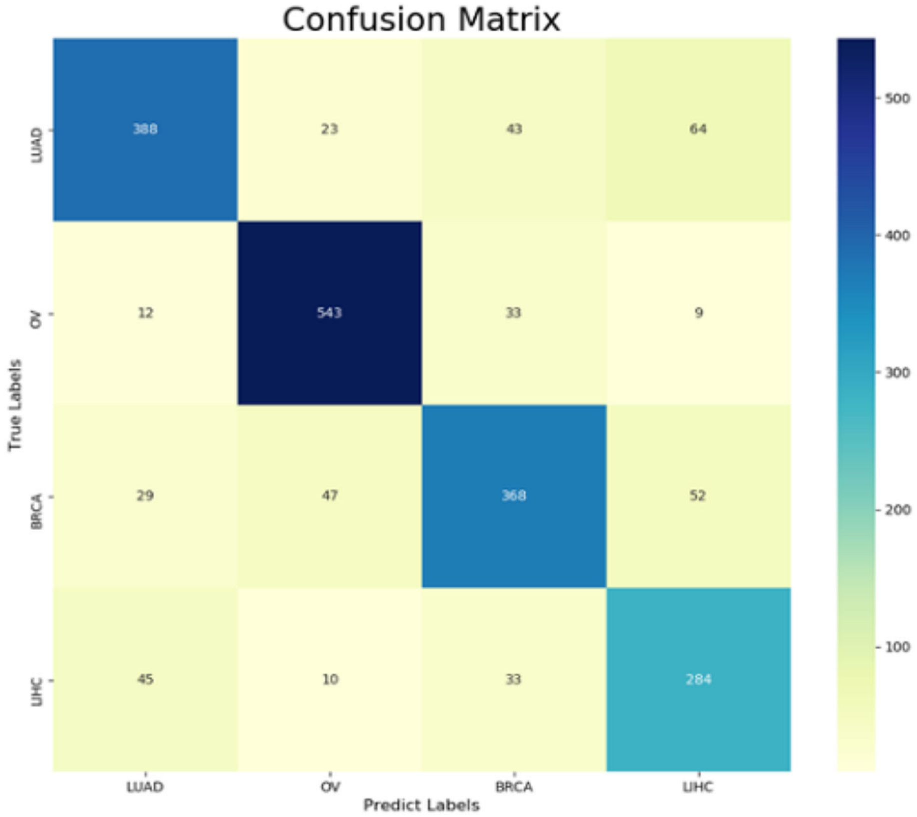

FIGURE 3 | Confusion matrix from pan-cancer classification by using SNN and features selection.

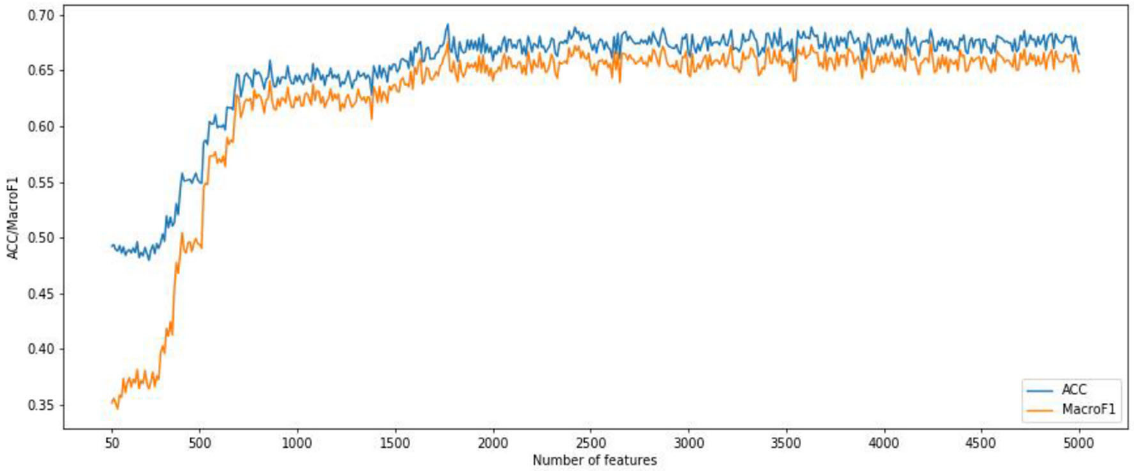

FIGURE 4 | Incremental feature selection (IFS) curves derived from the IFS method and RF algorithm. IFS curve with *X*-values from 50 to 5,000.

The random forest (RF) method is a supervised classification and regression algorithm (Cutler et al., 2012). The RF method builds multiple decision trees and merges them together to get a more accurate prediction. It adds additional randomness to the model when it growing the trees. Instead of searching for the most important feature when splitting a node, it searches for the best feature among a random subset of features. This generally results in a better model. The RF method has been widely used in machine learning area and is applied here to compare our model.

Performance Evaluation

Since pan-cancer classification is a multi-classification problem, we use accuracy (ACC) to measure the performance. There are also precision and recall to measure performance in a binary classification problem. One measurement closely related to these two values is F-score, which is a comprehensive indicator of precision and recall. That means, F-score is a parameter used to adjust the ratio of these two parts. When this parameter is 1, it degenerates into a harmonic average called F1-score. The multi-classification evaluation was split into multiple binary classification problems, and each F1-score was calculated. The average of the F1 scores was defined as Macro F1. To evaluate prediction of SNN classifier, we performed a 10-fold cross-validation (Kohavi, 1995; Chen et al., 2017, 2018).

RESULTS

To evaluate the best features for discriminating four types of cancer samples, a MCFS method was used to rank all features according to their RI values by using Monte Carlo method and decision trees. We selected the top 5,000 CNV features and applied IFS method.

After using MCFS for CNV feature sorting, we obtained two feature subset series. For the first CNV feature subsets, the parameter k is set to 10. That means, the i-th feature subset contains the first 10 times i features in the original CNV feature list. We constructed an SNN-based classification model on each feature subset, performed a 10-fold cross-validation and calculated its accuracy and macro F1 values. To show the changes of accuracy and macro F1 values, an IFS curve was generated as **Figure 1**. In **Figure 1**, the accuracy and macro f1 values are the Y axis and the number of features is the X axis. Both curves become stable after number of features >2,500 and them reached acceptable values. Therefore, we selected the number interval as [2,500, 4,999] for classifier to select the best number of features.

The following CNV feature subset is constructed by using the number of features in the number interval [2,500, 4,999]. By testing all of these subsets, we obtained the corresponding accuracy and macro F1 values. We also plotted the IFS curves to show these values in **Figure 2**. The best accuracy and macro

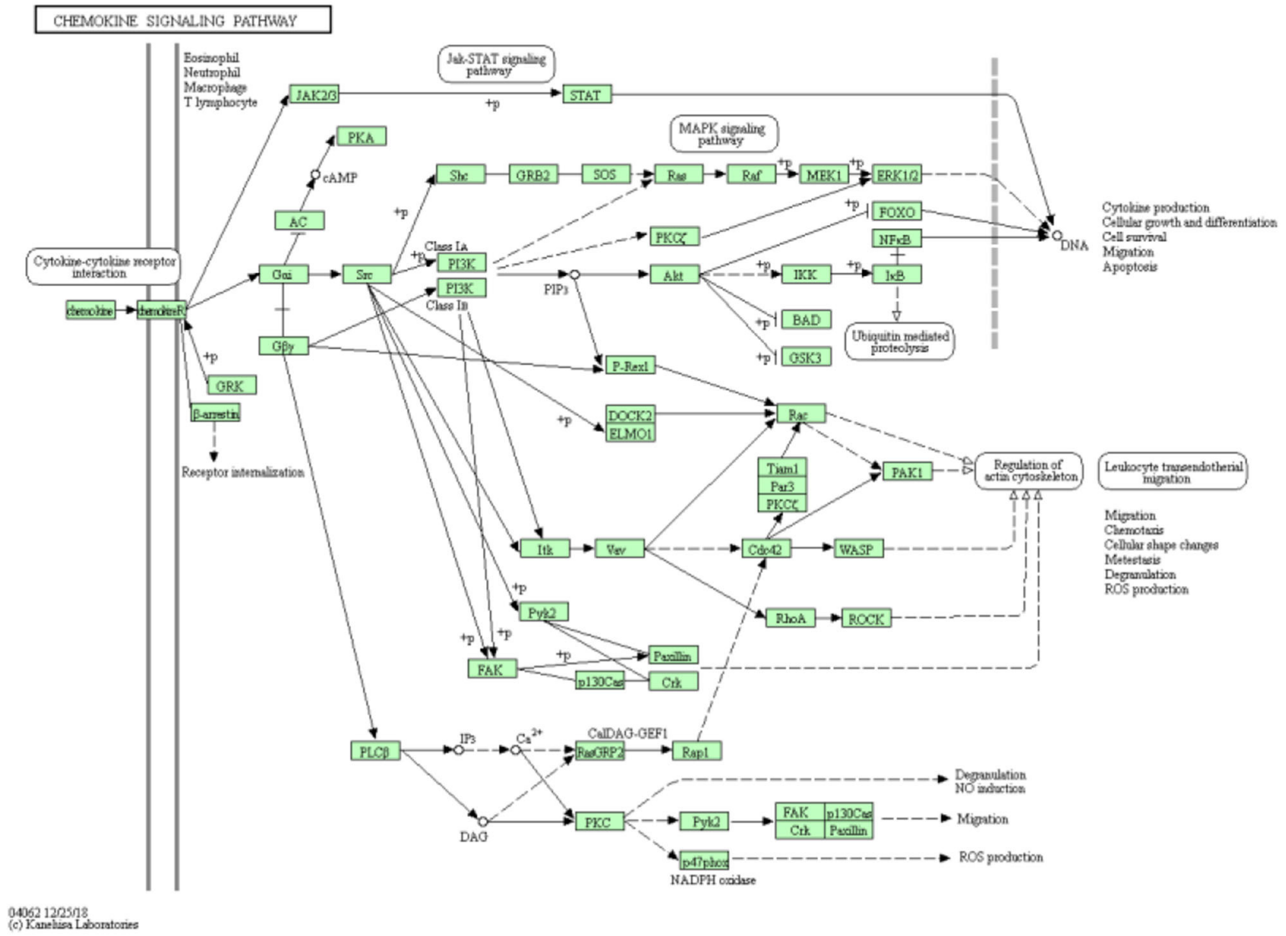

FIGURE 5 | Chemokine signaling pathway from KEGG has the highest counts for selected feature genes.

F1 values were generated when using the first 3,694 features to construct the SNN-based classification model. Thus, these first 3,694 genes were select for the final model. In the meantime, we used RF method as a comparison. The RF generated accuracy and macro F1 are much lower than the SNN one, which proves the efficiency of the deep SNN classifier. Therefore, we obtained the best feature subset and the optimal SNN-based model. Its ACC is 0.798 and the corresponding macro F1 is 0.789. **Figure 3** is confusion matrix and shows the good classification result from our model.

We also implemented the RF algorithm to construct a classifier on the CNV features subset obtained from the IFS method and evaluate each classifier through a 10-fold cross-validation test. Since the fast speed of RF method, which promised all CNV feature sets were tested. In order to compare the classification feature selection results, the IFS curves of accuracy and macro F1 were plotted in **Figure 4**. It can be seen that the optimal accuracy value is 0.689 and the macro F1 is 0.667 when using the first 1,693 features in the CNV feature list. Therefore, the first 1,693 features and RF algorithms can construct the best RF classification model. It can be seen that the accuracy and macro F1 obtained by the best RF classifier are much lower than those obtained by the best SNN-based classification model. That means our SNN-based model is effective in pan-cancer classification analysis.

DISCUSSION

DNA copy number variation is a straight-forward mechanism, which provides insight into genomic instability and structural dynamism in cancer researches. We applied Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis to the first 200 selected features and checked whether these were significant pathway information as shown as **Figure 5**. The highest counts are on the Chemokine signaling pathway, where chemoattractant proteins play an important role in controlling leukocyte migration during development, homeostasis, and inflammation. These processes are closely related to the occurrence and development of various cancers.

CONCLUSIONS

In this study, we use machine learning method for CNV-based pan-cancer classification. Considering the high dimension of data, MCFS and IFS are used to classify four different cancer patients effectively. And the feature subsets generated from IFS method are classified by integrating SNN method. Comparison experiments show that our SNN-based classification method has significant advantages over random forest in cancer classification. We demonstrate the advantages and potential of this method for copy number variant data. We suggest that this model can be extended and transferred to other pan-cancer classification fields. For future research, we will improve the models of other complex and large-scale data and expand our training data sets to further improve classification results.

DATA AVAILABILITY STATEMENT

The data and code are available at https://github.com/KohTseh/CancerClassification.

AUTHOR CONTRIBUTIONS

JL and QX leaded the method application, experiment conduction, the result analysis, and drafted the manuscript. QX and MW participated in the data extraction and preprocessing. TH and YW provided theoretical guidance and the revision of this paper. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the grants from the National 863 Key Basic Research Development Program (2014AA021505) and the startup grant of Harbin Institute of Technology (Shenzhen). National Natural Science Foundation of China (31701151), National Key R&D Program of China (2018YFC0910403), Shanghai Municipal Science and Technology Major Project (2017SHZDZX01), Shanghai Sailing Program (16YF1413800) and The Youth Innovation Promotion Association of Chinese Academy of Sciences (CAS) (2016245).

REFERENCES

Angermueller, C., Pärnamaa T, Parts, L., and Stegle, O. (2016). Deep learning for computational biology. *Mol. Syst. Biol.* 12:878. doi: 10.15252/msb.20156651

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural. Netw*. 5, 157–166. doi: 10.1109/72.279181

Chen, L., Li, J., Zhang, Y. H., Feng, K., Wang, S., Zhang, Y., et al. (2018). Identification of gene expression signatures across different types of neural stem cells with the Monte-Carlo feature selection method. *J. Cell Biochem.* 119, 3394–3403. doi: 10.1002/jcb.26507

Chen, L., Wang, S., Zhang, Y.-H., Li, J., Xing, Z.-H., Yang, J., Huang, T., and Cai, Y.-D. (2017). Identify key sequence features to improve CRISPR sgRNA efficacy. *Biochim. Biophys. Acta* 5, 26582–26590. doi: 10.1109/ACCESS.2017.2775703

Cutler, A., Cutler, D. R., and Stevens, J. R. (2012). "Random forests," in *Ensemble Machine Learning*, eds C. Zhang and Y. Ma (Golden Valley, MN: Springer), 157–175. doi: 10.1007/978-1-4419-9326-7_5

Ding, X., Tsang, S. Y., Ng, S. K., and Xue, H. (2014). Application of machine learning to development of copy number variation-based prediction of cancer risk. *Genomics Insights*. 7, 1–11. doi: 10.4137/GEI.S15002

Draminski, M., Rada-Iglesias, A., Enroth, S., Wadelius, C., Koronacki, J., and Komorowski, J. (2008). Monte Carlo feature selection for supervised classification. *Bioinformatics* 24, 110–117. doi: 10.1093/bioinformatics/btm486

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., et al. (2007). Patterns of somatic mutation in human cancer genomes. *Nature* 446, 153–158. doi: 10.1038/nature05610

Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A., et al. (2016). Toward a shared vision for cancer genomic data. *N. Engl. J. Med*. 375, 1109–1112. doi: 10.1056/NEJMp1607591

Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., Drill, E., et al. (2018). Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* 173, 291–304.e296. doi: 10.1016/j.cell.2018.03.022

Ioffe, S., and Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. *arXiv*.

Klambauer, G., Hochreiter, S., and Rarey, M. (2019). Machine learning in drug discovery. *J. Chem. Inf. Model.* 59, 945–946. doi: 10.1021/acs.jcim.9b00136

Klambauer, G., Unterthiner, T., Mayr, A., and Hochreiter, S. (2017). "Self-normalizing neural networks" in *Advances in Neural Information Processing Systems* (Long Beach, CA), 972–81.

Kohavi, R. (1995). *A study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*. Montreal, CA: IJCAI.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Lu, Y., and Han, J. J. (2003). Cancer classification using gene expression data. *Inf. Syst*. 28, 243–268. doi: 10.1016/S0306-4379(02)00072-8

Mayer, D. K., Nasso, S. F., and Earp, J. A. (2017). Defining cancer survivors, their needs, and perspectives on survivorship health care in the USA. *Lancet Oncol*. 18, e11–e18. doi: 10.1016/S1470-2045(16)30573-3

Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., and Getz, G. (2011). GISTIC2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* 12:R41. doi: 10.1186/gb-2011-12-4-r41

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., and Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *J. Big Data* 2:1. doi: 10.1186/s40537-014-0007-7

Ostrovnaya, I., Nanjangud, G., and Olshen, A. B. (2010). A classification model for distinguishing copy number variants from cancer-related alterations. *BMC Bioinformatics* 11:297. doi: 10.1186/1471-2105-11-297

Pan, X., Hu, X., Zhang, Y. H., Feng, K., Wang, S. P., Chen, L., Huang, T., and Cai, Y. D. (2018). Identifying patients with atrioventricular septal defect in down syndrome populations by using self-normalizing neural networks and feature selection. *Genes* 9:208. doi: 10.3390/genes9040208

Pascanu, R., Mikolov, T., and Bengio, Y. J. C. (2012). Understanding the exploding gradient problem. *Arvix.* 2:417.

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). "On the difficulty of training recurrent neural networks," in *International Conference on Machine Learning*, (Atlanta, GA), 1310–1318.

Sultana, J., Rani, M. U., and Farquad, M. (2020). "An extensive survey on some deep-learning applications," in *Emerging Research in Data Engineering Systems and Computer Communications*, eds P. V. Krishna and M. S. Obaidat (Springer), 511–519. doi: 10.1007/978-981-15-0135-7_47

Wang, Q., Jia, P., Li, F., Chen, H., Ji, H., Hucks, D., et al. (2013). Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. *Genome Med.* 5: 91. doi: 10.1186/gm495

Wild, C. P., Stewart, B. W., and Wild, C. (2014). *World Cancer Report 2014*. Geneva: World Health Organization.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.