A Convolutional Neural Network Based Ensemble Method for Cancer Prediction Using DNA Methylation Data

Chao Xia
School of Biomedical Engineering, Shanghai Jiao Tong University,
Shanghai 200240, China.
xiabc612@gmail.com

Yawen Xiao
Department of Automation, Shanghai Jiao Tong University,
Shanghai 200240, China.
foreverxyw@sjtu.edu.cn

Jun Wu
School of Life Sciences, East China Normal University,
Shanghai, China.
jwu@bio.ecnu.edu.cn

Xiaodong Zhao
Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University,
Shanghai, 200240, China.
xiaodongzhao@sjtu.edu.cn

Hua Li
School of Biomedical Engineering, Shanghai Jiao Tong University,
Shanghai 200240, China.
kaikaixinxin@sjtu.edu.cn

ABSTRACT

Cancer is a deadly disease all over the world and its morbidity is increasing at an alarming rate in recent years. With the rapid development of computer science and machine learning technologies, computer-aid cancer prediction has achieved increasingly progress. DNA methylation, as an important epigenetic modification, plays a vital role in the formation and progression of cancer, and therefore can be used as a feature for cancer identification. In this study, we introduce a convolutional neural network based multi-model ensemble method for cancer prediction using DNA methylation data. We first choose five basic machine learning methods as the first stage classifiers and conduct prediction individually. Then, a convolutional neural network is used to find the high-level features among the classifiers and gives a credible prediction result. Experimental results on three DNA methylation datasets of Lung Adenocarcinoma, Liver Hepatocellular Carcinoma and Kidney Clear Cell Carcinoma show the proposed ensemble method can uncover the intricate relationship among the classifiers automatically and achieve better performances.

CCS Concepts

•Applied computing→Computational genomics

Keywords

DNA methylation; convolutional neural network; cancer prediction; machine learning.

1. INTRODUCTION

In recent years, cancer incidence is growing at an alarming rate all over the world. Almost every tissue in the body can give rise to malignancy, but the commonality of all cancers is multiple genetic and epigenetic changes leading to the unlimited cell proliferation [1][2]. Previous researches have shown that some abnormal epigenetic markers can activate or inhibit gene expression, which may increase the risk of cancer [3]. As an important epigenetic modification, DNA methylation plays a vital role in the formation and progression of cancer [4]. Besides, it can be easily translated from laboratory research to clinical diagnosis [5]. Therefore, exploring the relationship between DNA methylation and cancer can help healthcare detect cancer in an early stage and assess the therapeutic effect.

With the rapid development of computer science and techniques, machine learning has a broad application in the fields of cancer diagnosis, and several prediction algorithms have been introduced by scholars. Y.-C. Chen et al. [6] used gene expression data and artificial neural network to assess the survival of non-small cell lung cancer (NSCLC) patients. The overall accuracy of survival prediction was more than 80%. Cho et al. [7] used a major voting method to integrate four classifiers for cancer prediction. Although the major voting method can absorb the advantages of different classifiers, the complex relationship and high-level features between classifiers were hard to discover. Besides gene expression data. Hao et al. [8] built a machine learning model to evaluate cancer diagnosis with DNA methylation data and got a prediction accuracy comparable to typical diagnostic methods.

In recent years, convolutional neural network (CNN) [9][10] has made tremendous progress in various areas and got state-of-the-art performance, which owing to its ability to learn the hierarchical representation of the input data. However, this method is a data-driven model that requires a large number of training samples. Limited by the resources like patient samples and experiment cost, the datasets in the biomedical field are much smaller. To solve this problem, several methods have been introduced to bridge the gap between the large-scale network and small dataset. An effective method is transfer learning, which first training the selected network on public large-scale dataset and fine-tuning the pertained models on medical dataset. Besides, ensemble learning and model fusion can also alleviate this problem. For example, Shi et al. [11] used a stacked method to improve the performance of deep polynomial network and obtained superior classification results on small ultrasound dataset.

ICMLC '19, February 22–24, 2019, Zhuhai, China
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6600-7/19/02…$15.00
DOI: https://doi.org/10.1145/3318299.3318372

Motivated by the aforementioned methods, in this paper, we introduce a convolutional neural network based ensemble method to explore internal relationships between DNA methylation and cancer. We first choose five basic classification methods, which are Naive Bayesian Classifier (NBC), k-Nearest Neighbor (KNN), Decision Tree (DT), Random Forest (RF) and Gradient Boosting Decision Tree (GBDT) as the first stage classifiers. Then, a convolutional neural network is used to integrate the classification results of the first stage classifiers. Due to the limitation of the dataset scale, we choose the most informative feature from the whole methylation site to avoid over-fitting. Experiment results on three public Methylation450K datasets show the effectiveness of the proposed methods.

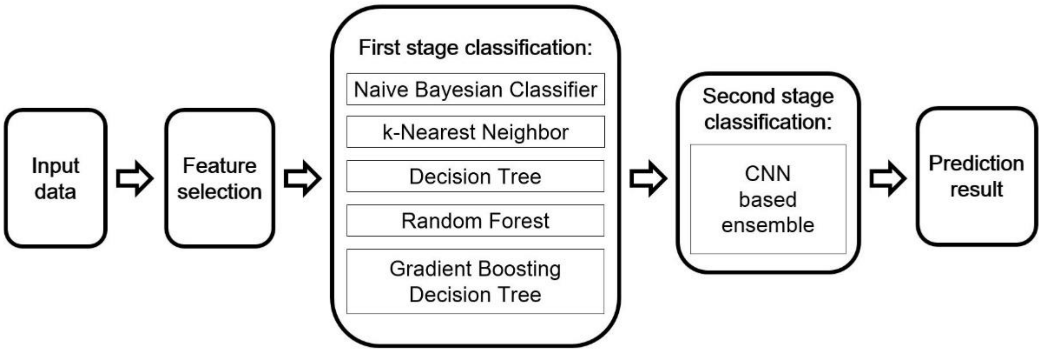

Input data
Feature selection
First stage classification:
Naive Bayesian Classifier
k-Nearest Neighbor
Decision Tree
Random Forest
Gradient Boosting Decision Tree
Second stage classification:
CNN based ensemble
Prediction result

Figure 1. Flowchart of the proposed convolutional neural network based ensemble method.

There are two main contributions in this paper. Firstly, instead of using the frequently used gene expression data, we attempt to use a machine learning method to predict cancer from DNA methylation status data. This epigenetic modification plays important roles in the development of cancers and is viewed as an early signal of cancer detection [4][12]. Then, a convolutional neural network based ensemble method is introduced to relieve the drawback of training large-scale deep neural network on small-scale datasets.

2. METHOD

In this paper, we introduce a convolutional neural network based ensemble method for cancer prediction using DNA methylation data. We first conduct t-test to choose a set of significantly differential methylation points. Then, the selected feature was feed into Naive Bayesian Classifier, k-Nearest Neighbor, Decision Tree, Random Forest and Gradient Boosting Decision Tree five basic classifiers for the first stage classification. Here we use S-fold cross validation method by dividing the whole datasets into S groups and choose S-1 groups as training sets, the left one as test sets at each time. Finally, a convolutional neural network is used to ensemble the predictions of the first stage classifiers and extract the internal relationship among different classifiers to predict a more reliable result. The flowchart of the proposed ensemble method is shown in Fig. 1.

2.1 Feature Selection

Feature selection is an important technique in machine learning, which aims to find a most significant subset of a special problem. Feature selection can not only reduce the model complexity, computational consumption but also improve classification accuracy. This technique is especially useful in the area with too many attributes and small samples, where the clinical DNA methylation dataset is a representation.

Compared with gene expression array data, 450K methylation array data contains more features. By contrast, the sample number is tiny, which makes it more challenging to develop a robust classification model. Here we use t-test to select significantly differential methylation point between normal and cancer samples. By setting the thresholds of p-value, the significant differences methylation points are selected.

2.2 First Stage Classification Methods

Over the past decade, machine learning technology has been widely used in high-throughput sequencing data analysis [13][14]. Here we use machine learning methods to establish a classifier for DNA Methylation450K datasets to distinguish cancer from normal samples. Specially, we use the selected significant differences methylation point and Naive Bayesian Classifier, k-Nearest Neighbor, Decision Tree, Random Forest and Gradient Boosting Decision Tree five classical classifiers to build the first-stage classification models. These classification methods have been proven to be highly accurate in different cancer prediction [15][16][17].

NBC is a statistical classifier, which based on the Bayes theorem. NBC predict class membership probabilities and is robust to the noise of the input data. In practice, NBC is a popular method for text categorization, besides, for medical diagnosis, it also provides performances equivalent to other machine learning techniques with low computational effort and high speed [18].

kNN is a non-parametric method that classifies the samples according to the distance of the input feature. For a test sample, kNN compares it with training samples that are similar to it and assign a most common class among its K-nearest neighbors. However, kNN is sensitive to the noises in the feature which may requires effective feature selection before classification. This method is also called as the lazy learner method.

DT is one of the earliest and widely used machine learning method, which uses a tree-like structure for prediction. DT consists of nodes and directed edges, where nodes include internal nodes and leaf nodes. Each internal node represents a decision condition while each leaf node represents a decision outcome. The paths from root to leaf represent classification rules. DT is a widely used classification method in many fields, however, it is easily overfitting in distinguish cancer and normal samples on DNA Methylation450K datasets.

RF and GBDT are two ensemble machine learning methods, which consist of several decision tree predictors. The predicted label of random forest is determined by the majority voting of the individual decision tree output. Gradient Boosting is a kind of Boosting method, which builds the model on the gradient descent direction of the loss function. These two ensemble model can relief the problem of overfitting caused by a single decision tree. However, they may produce a result tending to the category with more samples.

Above all, several machine learning methods are available for cancer prediction, but each classifier has its advantages and drawbacks, no one can outperform others in all the aspects. A classifier may get a good performance on several datasets but fail in others. A good way to solve this problem is to use an ensemble strategy to take the advantages of the multiple methods and avoid their shortcomings.

2.3 Convolutional Neural Network Based Ensemble Model

Motivated by the aforementioned problems of basic machine learning methods, we introduce a convolutional neural network based ensemble method for cancer prediction on DNA methylation450K data. Convolutional neural network is an extensively employed network structure in deep learning [19], which is inspired by the information processing mechanism in biological visual system. Due to the mechanism of "local receptive fields" and "shared weights" [9], convolutional neural network can reduce the number of neurons in the neural network and speed up the training process. In 1990, LeCun introduced a convolutional neural network model LeNet-5 for handwritten numeral recognition [20]. After decades of development, remarkable achievements have been made in the area of natural language processing and computer vision. Some improved neural networks were also proposed, such as AlexNet [10], VGGNet [21], GoogleNet [22] and ReNet [23]. However, these excellent performances are usually based on a large number of training samples, clinical cancer datasets derived from DNA methylation are relatively limited and difficult to meet this requirement.

To relief the aforementioned problems, we adopt a two-stage classification method and construct a convolutional neural network to stack the prediction results of multiple classifiers. Our CNN based ensemble model structure is summarized in Fig2, consisting of two convolution layers, a max-pooling layer and fully connection layer. Some frequently-used training strategies, such as dropout [24] and batch normalization [25] are employed to avoid over-fitting.

In convolutional neural network, convolution layer is used for feature extraction. Convolutional layer performs a sliding window operation across the whole data and conducts convolution operation at each position. For a convolution layer with N_{conv} filters, each of length k, the input data with the length of b can produce a feature map of $N_{conv} \times (b-k+1)$. A deep convolutional neural network with multiple convolutional layers can extract high-level features from low-level features iteratively. We define this step as f_{conv}.

Rectified Linear Units (ReLU) operation is used after every Convolution operation. ReLU is a non-linear operation, defined as eq.(1), which clamps all negative values to zero. The purpose of ReLU is to improve the nonlinearity of the network, which can improve the network performances. Compared with other nonlinear functions such as tanh or sigmoid, ReLU can improve training speed significantly and perform better in most situations. This step is defined as f_{ReLU}.

$$\boldsymbol{f_{ReLU}(z) = \mathbf{max}(0, z)} \tag{1}$$

Pooling operation in the neural network aims to reduce the dimensionality of feature map while keeping the most important information. The most commonly used Pooling method is Max Pooling, which takes the largest element from the selected window. Pooling operation can reduce the parameters in the network and thus control over-fitting. This step is defined as $f_{maxpool}$. Another approach to reduce over-fitting is Dropout [24].

After convolution and pooling operation, Fully Connected layer is added to map the high-level features into the output class label with a sigmoid activation function. We define this step as f_{full}. For the input data with n samples $X^{(n)}$, the whole network output form can be written as follows:

$$\boldsymbol{f(X^{(n)}) = f_{full}(f_{maxpool}(f_{ReLU}(f_{conv}(X^{(n)}))))} \tag{2}$$

For the input data $X^{(n)}$, we define a loss function to describe the difference between the prediction output $f(X^{(n)})$ and true label $y^{(n)}$. The loss function L for the whole training datasets with N_{samp} samples can be defined as:

$$\mathbf{L} = \boldsymbol{\sum_{n=1}^{N_{samp}} loss(\mathbf{f}(X^{(n)}), y^{(n)})} \tag{3}$$

We use Θ to describe all the parameters in the network. The stochastic gradient descent (SGD) algorithm [26] is used in the training process to minimize the loss function by update the network parameters Θ as follows:

$$\Theta = \Theta - \eta \ \boldsymbol{\frac{\partial L}{\partial \Theta}} \tag{4}$$

where η is the learning rate, setting to 0.001.

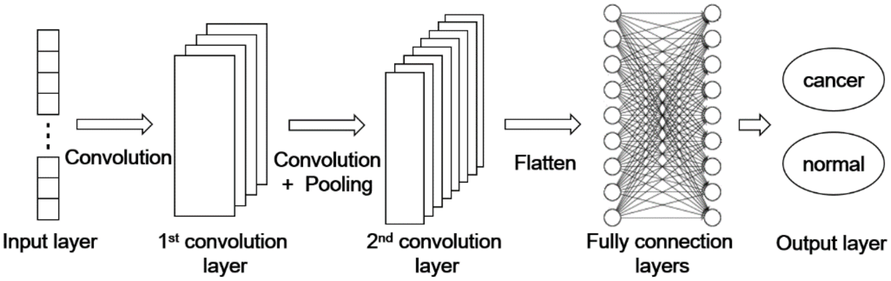


Figure 2. The architecture of the proposed convolutional neural network.

3. EXPERIMENT RESULTS

3.1 Dataset

In this section, we evaluated the proposed method on DNA Methylation450K datasets, including Lung Adenocarcinoma (LUAD), liver hepatocellular carcinoma (LIHC) and Kidney Clear Cell Carcinoma (KIRC) three types of common cancer data. These datasets, including all cancer stages, were collected from the patient with different ages, genders and races. After the data clearing, there are approximate 400 thousands of features in each sample. The DNA methylation data were downloaded from the TCGA-XENA project web page [27]. In our method, due to the tiny datasets scale and high-dimensional samples, we first use t-test for feature selection and dimensionality reduction. In this way, we select less than 100000 features with significant differences from the whole methylation450K data. The specific information of the datasets is shown in Table 1.

Table 1. Detail information of datasets.

Datasets	Features	Tumor samples	Normal samples	Total
LUAD	395995	460	32	492
LIHC	395911	379	50	429
KIRC	395808	320	160	480

3.2 Prediction Results of the Proposed Method

We first use five classical classification methods individually for the first stage classification, which are Naive Bayesian, k-nearest neighbor, decision tree, random forest and gradient boosting decision tree. To prevent over-fitting, we choose 5-fold cross validation technique. Then, an ensemble method based on the convolutional neural network was used to integrate the first stage prediction.

Since the problem we devoted to aiming to classify the input DNA methylation data into normal class and cancer class, which belonging to binary classification task. Thus, we choose a 'sigmoid' function rather than a 'softmax' function for the output layer. Other hyperparameters, such as a stochastic gradient descent optimizer, a learning rate of 0.001, a batch size of 32 and a momentum rate of 0.9, are chosen according to the preliminary experiments.

Table 2. Prediction result of different methods on three datasets.

Methods	LUAD (%)	LIHC (%)	KIRC (%)
NB	98.98	98.37	98.54
KNN	98.37	92.77	87.92
DT	98.17	96.74	98.96
RF	97.36	92.31	95.83
GB	97.15	95.34	99.38
Majority voting	98.37	97.44	99.17
CNN based ensemble	**99.39**	**98.83**	**99.58**

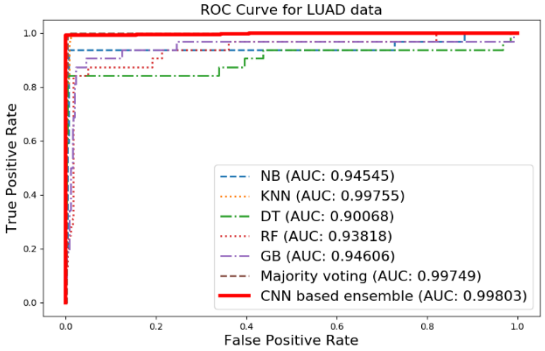
To evaluate each prediction method, we consider the data derived from cancer cells as positive samples, data from normal cells as negative samples. We compare the prediction results of each individual machine learning method, majority voting and the proposed ensemble method on LUAD, LIHC, KIRC three datasets. The prediction results are shown in Table 2. From this table, we can see that our CNN based ensemble method significantly outperforms the individual classification method in all data sets. Some single classifier may get a relatively good performance on one dataset but performs poor in another. In comparison, our CNN based ensemble method can get a more stable performance on different datasets.

The receiver operating characteristic (ROC) curves of the three cancer datasets are shown in Fig. 3. In statistic, the ROC curve use a dynamic threshold to illustrate the performance of different predictors, which takes true positive rate (TPR) as the vertical axis and false positive rate (FPR) as the horizontal axis. TPR and FPR are calculated as (5) and (6), where TP, FP, FN, TN means true positive, false positive, false negative, true negative respectively. The TPR is also known as sensitivity or recall. The area under the curve is an important measurement of classifier performance. Due to the imbalance distribution of different classes in the datasets, we further use precision-recall (PR) curve to evaluate. The PR curve uses precision as the vertical axis and recall as the horizontal axis. Precision (P) and recall (R) are calculated as (7) and (5). Area under the PR curve are also calculated.

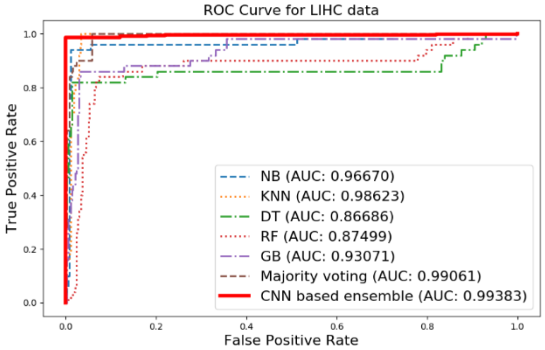
$$\text{TPR, recall}=\frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \tag{5}$$

$$\text{FPR}=\frac{\boldsymbol{FP}}{\boldsymbol{TN+FP}} \tag{6}$$

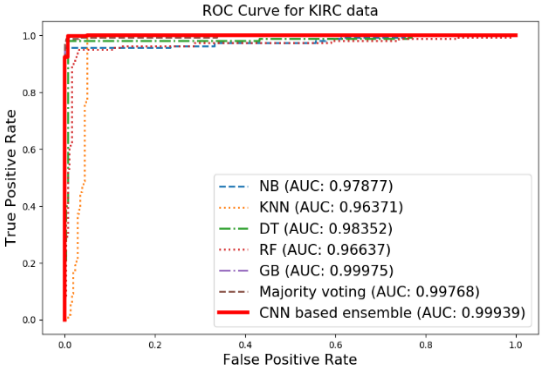
$$\text{precision}=\frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \tag{7}$$



(a). The ROC curves of different methods on LUAD datasets.



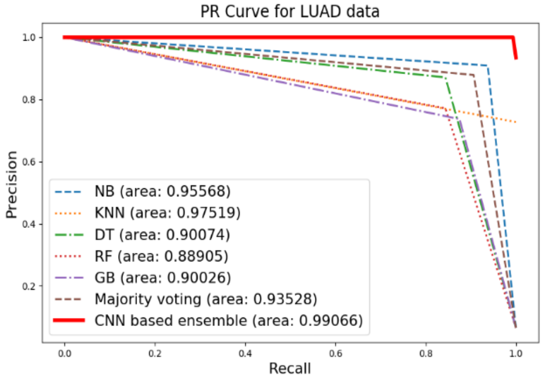
(b). The ROC curves of different methods on LIHC datasets.



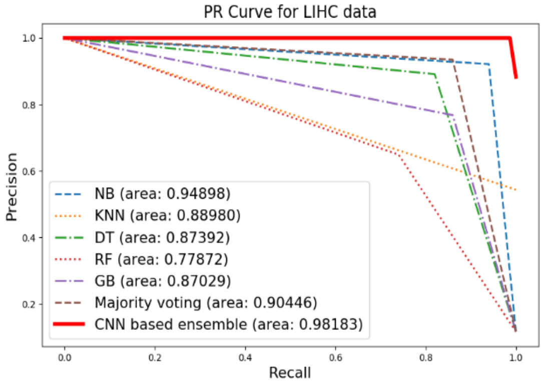
(c). The ROC curves of different methods on KIRC datasets.

Figure 3. The ROC curves of different methods on three datasets: (a).LUAD; (b).LIHC; (c).KIRC.

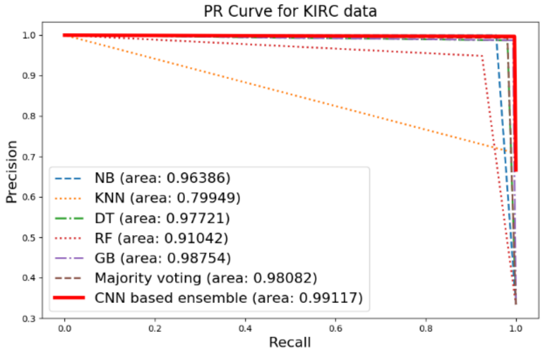
As shown in Fig. 3, we can know the performances of five kinds of single classifier, majority voting method and our CNN based ensemble method on three cancer datasets. The CNN based ensemble method can get higher AUC scores than single classifier operating alone. Besides, CNN based ensemble method performs better than majority voting method, which may owing to its ability to find the intrinsic feature of input data.



(a). The PR curves of different methods on LUAD datasets.



(b). The PR curves of different methods on LIHC datasets.



(c). The PR curves of different methods on KIRC datasets.

Figure 4. The PR curves of different methods on three datasets: (a).LUAD; (b).LIHC; (c).KIRC.

The PR curves of the three cancer datasets are shown in Fig. 4 respectively. For imbalance datasets, PR curves can provide a more reliable evaluation. In all three cancer datasets, our convolutional neural network based ensemble method obtains an area that is bigger than each single classifier as well as the majority voting method, which is inconsistent with the results of ROC curve.

4. CONCLUSIONS

In this paper, we explore the problem of cancer prediction by using DNA methylation data and introduce a convolutional neural network based two-stage classification method. Taking the limitation of small-scale clinical datasets into consideration, we first conduct t-test for feature selection and use Naive Bayesian Classifier, k-Nearest Neighbor, Decision Tree, Random Forest and Gradient Boosting Decision Tree five classical methods as the first stage classifier. Then a convolutional neural network based model was introduced to ensemble the prediction results of the first stage classifiers and predict a more reliable result. Experiment results on LUAD, LIHC, KIRC three DNA Methylation450K datasets show the proposed ensemble model outperforms each single classifier and the majority voting method in various evaluation metrics. These experiment results indicate that the convolutional neural network based multi-model ensemble method can learn the intricate relationship among the classifiers automatically and achieve better prediction results.

5. ACKNOWLEDGMENTS

This work was supported by Development Program for Basic Research of China (2014YQ09070904), National Natural Science Foundation of China (31671299), Shanghai Science and Technology Committee Program (17JC1400804), Medical engineering cross fund (YG2015MS33).

6. REFERENCES

[1] Jones, A., Teschendorff, A. E., Li, Q., Hayward, J. D., Kannan, A., Mould, T., ... & Lee, S. H. (2013). Role of DNA methylation and epigenetic silencing of HAND2 in endometrial cancer development. *PLoS medicine*, 10(11), e1001551.

[2] Moore, L. D., Le, T., & Fan, G. (2013). DNA methylation and its basic function. *Neuropsychopharmacology*, 38(1), 23.

[3] Terry, M. B., McDonald, J. A., Wu, H. C., Eng, S., & Santella, R. M. (2016). Epigenetic biomarkers of breast cancer risk: Across the breast cancer prevention continuum. In *Novel Biomarkers in the Continuum of Breast Cancer* (pp. 33-68). Springer, Cham.

[4] Lai, H. C., Wang, Y. C., Yu, M. H., Huang, R. L., Yuan, C. C., Chen, K. J., ... & Chao, T. K. (2014). DNA methylation as a biomarker for the detection of hidden carcinoma in endometrial atypical hyperplasia. *Gynecologic oncology*, 135(3), 552-559.

[5] Dehan, P., Kustermans, G., Guenin, S., Horion, J., Boniver, J., & Delvenne, P. (2009). DNA methylation and cancer diagnosis: new methods and applications. *Expert review of molecular diagnostics*, 9(7), 651-657.

[6] Chen, Y. C., Ke, W. C., & Chiu, H. W. (2014). Risk classification of cancer survival using ANN with gene expression data from multiple laboratories. *Computers in biology and medicine*, 48, 1-7.

[7] Cho, S. B., & Won, H. H. (2003, January). Machine learning in DNA microarray analysis for cancer classification. In *Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics 2003-Volume 19* (pp. 189-198). Australian Computer Society, Inc..

[8] Hao, X., Luo, H., Krawczyk, M., Wei, W., Wang, W., Wang, J., ... & Jafari, M. (2017). DNA methylation markers for diagnosis and prognosis of common cancers. *Proceedings of the National Academy of Sciences*, 114(28), 7414-7419.

[9] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324.

[10] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

[11] Shi, J., Zhou, S., Liu, X., Zhang, Q., Lu, M., & Wang, T. (2016). Stacked deep polynomial network based representation learning for tumor classification with small ultrasound image dataset. *Neurocomputing*, 194, 87-94.

[12] Liu, H., Li, Z., Ding, J., Liu, J., & Zhang, Y. (2013, July). Identification of the differential DNA methylation markers among cancers. In *Fuzzy Systems and Knowledge Discovery (FSKD), 2013 10th International Conference on* (pp. 730-734). IEEE.

[13] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3), 389-422.

[14] Wang, L., Zhu, J., & Zou, H. (2008). Hybrid huberized support vector machines for microarray classification and gene selection. *Bioinformatics*, 24(3), 412-419.

[15] Xiao, Y., Wu, J., Lin, Z., & Zhao, X. (2018). A deep learning-based multi-model ensemble method for cancer prediction. *Computer methods and programs in biomedicine*, 153, 1-9.

[16] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13, 8-17.

[17] Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial intelligence in medicine*, 34(2), 113-127.

[18] Dumitru, D. (2009). Prediction of recurrent events in breast cancer using the Naive Bayesian classification. *Annals of the University of Craiova-Mathematics and Computer Science Series*, 36(2), 92-96.

[19] Hinton, G. E., & Salakhutdinov, R. R. (2006). *Reducing the dimensionality of data with neural networks.* science, 313(5786), 504-507.

[20] LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., & Jackel, L. D. (1990). Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems* (pp. 396-404).

[21] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv*:1409.1556.

[22] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9).

[23] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[24] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1), 1929-1958.

[25] Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv*:1502.03167.

[26] Bottou, L. (2004). Stochastic learning. In *Advanced lectures on machine learning* (pp. 146-168). Springer, Berlin, Heidelberg.

[27] https://xenabrowser.net/datapages/.