 doi:10.20944/preprints202103.0324.v1

Title: Artificial Neural Network Analysis of Geo database in diagnosing Papillary Thyroid Carcinoma

Author list: Zhoujing Zhang1, Di Xu2, Ozioma Akakuru2, Wenjing Xu1, Yewei Zhang1

1: School of Medicine Southeast University

2: Ningbo Institute of Materials Technology&Engineering Chinese Academy of Sciences

Abstract: The diagnosis of papillary thyroid carcinoma has always been a concerned and challenging issue and it is very important and meaningful to have a definite diagnosis before the operation. In this study, we tried to use an artificial intelligence algorithm instead of medical statistics to analyze the genetic fingerprint from gene chip results to identify papillary thyroid carcinoma. We trained 20 artificial neural network models with differential genes and other important genes related to the cell metabolic cycle as the list of input features, and apply them to the diagnosis of papillary thyroid cancer in the independent validation data set. The results showed that when we used the DEGs and all genes lists as input features the models got the best diagnostic performance with AUC=98.97% and 99.37% and the accuracy were both 96%. This study revealed that the proposed artificial neural network models constructed with genetic fingerprints could achieve a prediction of papillary thyroid carcinoma. Such models can support clinicians to make more accurate clinical diagnoses. At the same time, it provides a novel idea for the application of artificial intelligence in clinical medicine.

Keywords： Papillary thyroid carcinoma; bioinformatics analysis; artificial neural network; diagnosis

Corresponding author contact details： Ph.D. Yewei Zhang

Department of General Surgery, Zhongda Hospital, School of Medicine, Southeast University, Nanjing, 210009, China

E-mail addresses: zhangyewei@seu.edu.cn

 doi:10.20944/preprints202103.0324.v1

Artificial Neural Network Analysis of Geo database in diagnosing Papillary Thyroid Carcinoma

Introduction

Thyroid carcinoma is the most common malignant tumor of the endocrine system. In recent decades, the incidence of thyroid cancer has been reported to be rising worldwide [1]. Among all the pathological types of thyroid cancer, differentiated thyroid cancer is the most common one, and among them, papillary thyroid cancer is more common [2]. The good prognosis of papillary thyroid carcinoma is directly related to the accurate diagnosis of papillary thyroid carcinoma in its early stage. Although the early diagnosis rate is extremely high, there are still numbers of papillary thyroid carcinomas are hidden. The image characteristics of the nodules in B-scan ultrasonography examination are not obvious, and the fine needle aspiration pathological examination also cause misjudgment [3].

In recent years, with the advent of gene chip technology and the big data, bioinformatic analysis have been widely used to search genetic alterations at the genome level. The huge amount of data provides a lot of resources that can be analyzed for the diagnosis and prognosis of various diseases [4]. The traditional method is to use medical statistics to analyze the data, and fit the data to the result through the logic of statistics, and the obtained fitting formula can be used to judge other data sets, so that a new clinical detection method or a starting point of pathway research were excavated [5].

Artificial Neural Network (ANN), referred to as Neural Network (NN), in the field of machine learning and cognitive science, is a kind of imitating biological neural network (animal's central nervous system, especially the mathematical model or calculation model of the structure and function of the brain), used to estimate or approximate the function [6]. Modern neural network is a non-linear statistical data modeling tool. Neural network is usually optimized through a learning method based on mathematical statistics, so it is also a practical application of mathematical statistics. It used a rough trial and error method to create a mathematical formula to fit every logic problem, including comparison of pure numbers, image identification or chess strategies [7]. Artificial neural networks have achieved breakthrough developments in various fields in recent decades, and have good applications in the medical

 doi:10.20944/preprints202103.0324.v1

field. Many researchers had used it successfully in the diagnosis of many diseases and the prediction of treatment effects [8-11].

In this research, we try to combine bioinformatics analysis and artificial intelligence methods to pre-screen the gene chip data from GEO database through statistical methods, and then train the ANN network through the above data, and then apply it to an independent data set for verification. The above strategies can save the time of ANN network learning, and at the same time can improve the accuracy of predicting disease events under the guidance of statistical methods.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2021 doi:10.20944/preprints202103.0324.v1

Materials and Methods

Ethics statement

All study procedures followed the guidelines outlined in the Declaration of Helsinki. The study protocol was approved by the Ethics Committee of Ningbo No2. Hospital（Huamei hospital, Chinese Academy of Sciences）and IEC for Clinical Research of Zhongda Hospital, Affiliated to Southeast University and all subjects provided written informed consent prior to participating in the study.

Data sets processing of Genes expression and Features

GEO (http://www.ncbi.nlm.nih.gov/geo) [12] is a public data repository of high throughout gene expression data, chips and microarrays. GEO2R online analysis tool (https://www.ncbi.nlm.nih.gov/geo/geo2r/) was used to detect the DEGs between thyroid carcinoma samples without and with lymph node metastasis. 10 gene expression datasets (GSE29265, GSE33630, GSE29315, GSE65074 defined as train set, GSE3467, GSE53157, GSE129562, GSE27155, GSE60542 defined as test set [13-20]) were selected and all data were freely available online, and the download data format is MINIML (more details of database see Table S1 in Supporting Information).

Statistical analysis

Batch effects between data sets were estimated and corrected using combat function of R package sva was used to remove batch effect [21]. Limma package of R software was used to study the differential expression of genes(DEGs), and adj p-value < 0.05 were considered statistically significant. The receiver operating characteristic (ROC) curves and area under the curve (AUC) were used to determine the diagnostic efficacy of different methods (logistic regression (LR) analysis, completely connected Neural Networks and convolutional neural network), the picture and the AUC were obtained with SPSS statistical package and an in-house Python script. The R version is 3.5.3, the SPSS version is 23 and Python version is 3.8.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2021 doi:10.20944/preprints202103.0324.v1

Covariant and Feature selection

Due to the inconsistency of clinical data among different data sets in geo database, we focused on genes with different expression as covariates or features in this study. Firstly, we take the whole genes (8647) list from the database and DEGs (198) (Fig. S1 C) in the training set as the covariates of the multivariate analysis of variance and the features of ANN training. Secondly, there are a lot of important metabolic or physiological processes in the tumorigenesis and development cycle which we also verified as references. Here we listed them as: DNA Damage Repair (DDR) Genes (232) [22], genes involved in anaerobic energy metabolism (590) [23], lipid metabolism related genes (751) [24], epithelial mesenchymal transition (EMT) - associated signature genes (1263) [25], Hypoxia-related genes (75) [26], immune signature (IS) genes (1959) [27], N6-methyladenosine(m6a) related genes (20) [28] and ferroptosis related genes (24) [29].

Data collation

A Python program is used to transfer and collect the data of the gene chip results which have been transformed the probe ID and removed the batch effects. The newly generated CSV file was the input file of ANN module for training and Predict program. The code of this program was detailed in the Supporting Information.

Automated diagnosis and ANN workflow

All programming related to the ANN was implemented in Python (Pytorch library; https://pytorch.org/). We train a model for each gene list with Deep Neural Networks (DNN). At the same time, because the genes in the list are related to some extent, in order to improve the efficiency of model fitting, we train a model for each gene list with Convolutional Neural Networks (CNN) too. The learning was done for appropriate epochs when the loss rate entered the platform period with a learning rate of 0.01, wherein 90% of the data set were used for training and 10% for self-validation to calculate the loss rate. After training, a test program run independently to test the data set used for verification, each individual was fed to the ANN fitting formula to generate a score between 0 and 1. (0: most likely normal, 1: most likely cancer). At the same time, ROC curve is automatically generated and AUC is automatically calculated by Python. All codes and operation parameters are shown in the supporting information.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2021 doi:10.20944/preprints202103.0324.v1

Results

Removing batch effects in analysis of genes expression from different data sets

The data sets from the geo database are often independent experimental results. The commercial chips used for different batches of testing are different, so there is a batch effect between different batches of experimental data. In order to expand the sample size and make the experimental data more reliable, we used the sav package in R to remove batch effects of gene chip results of all selected data sets before performing subsequent analysis. In order to simulate the actual clinical test situation, we divided the data set into two groups (train set and test set) and removed the batch effects respectively (Fig. 1 A, B and C). At the same time, we screening out DEGs in train set with Log2 (fold change) ≥1 or ≤-1 and adj p-value < 0.05(Fig. S1 A and Table S2). A heat map was also drawn based on the expression of genes in the mentioned gene set (Fig S1 A and B).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2021 doi:10.20944/preprints202103.0324.v1

ANN framework

To determine the diagnostic efficiency of artificial intelligence analysis of gene chip results in distinguishing papillary thyroid carcinoma from normal thyroid tissue, two classic architecture of ANN were built. Deep Neural Networks (Fig. 2A), it is the simplest neural network and uses the most network parameters and the most calculations to fit almost all the logic problems with function. The hidden layer is the most important intermediate in this network. For complex problems, especially there are too many inputs features the hidden layer often determines the final fitting function. Fewer hidden layers tend to have poor fit of the fitting function, and more hidden layers may lead to over-fitting, resulting in high diagnostic accuracy in the train set, but extremely poor diagnostic accuracy in the test set. In this study, 10 gene lists were selected as input, of which 9 we used 3 hidden layers. When using all chip genes (8647) as input feature values, we only used 1 hidden layer to avoid gradient explosion. The number of nodes in each hidden layer increases or decreases according to the number of input features from different gene list. The above parameters are listed in Table S3. Because there were connections between the expression levels of different genes, in order to improve computational efficiency and diagnostic accuracy, we also tried to build a convolutional neural network to validate our data set (Fig. 2B). Convolutional Neural Networks are often used to process data information in which input feature values such as images are related to neighboring or specific distances. The network structure is divided into three parts: convolution, activation, and pooling. The output result is the specific feature space of each data set. The most important work in the whole process is how to iteratively adjust the network weights through the training data, which is the backward propagation algorithm. In this study, the convolutional neural network we built was convolved twice, and the number of convolution kernels in each layer of convolution depended on the number of features in different gene list. All the parameters are listed in Table S4.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2021 doi:10.20944/preprints202103.0324.v1

Training of ANN models

In order to identify papillary thyroid cancer using artificial intelligence, we first trained 10 DNNs and 10 CNNs. As we previously described, the self-validation set in training step is 10% inside the training set. The loss rate is collected every 5 cycles. When the loss rate convergence curve reached the platform stage (Fig. 3A-J and Fig. 4A-J), we stop training. The training results were saved in every appropriate cycle, and this training result is validated against independent test sets later.

 doi:10.20944/preprints202103.0324.v1

Validation of ANN models

For validation of models, sensitivity, specificity, and area under the receiver operating characteristic curve of the predictive models were calculated to explore their predictive power. In order to compare with the method of medical statistics, Logistic regression was used to calculate the sensitivity and specificity, draw the ROC curve and calculate the AUC. However, because there are too many genes as covariates in the analysis, we used hierarchical calculation in the whole gene list model, DEGs list model, DDR gene list model, lipid metabolism related genes list model and EMT gene list model. These LR regression related calculation results have little reference value, but we still showed them as comparison (Fig. 5A-J) (Table 1). Almost in every genes list training the DNN models were doing better than CNN models. The best models were to use the whole gene and DEGs as the input features training model, their AUC of prediction using independent prediction set were 98.97% and 99.37% and the accuracy were both 96%.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2021 doi:10.20944/preprints202103.0324.v1

Discussion

The number of malignant tumors has been increasing both in the incidence rate and mortality rate in the past decade [30]. This phenomenon can be explained by many factors: unhealthy life style, environmental pollution, and radiation and immunity dysfunction [31-33]. It is crucial to be able to make accurate diagnosis and prognostic judgments of malignant tumors in the early stages of the occurrence and development of malignant tumors in all diagnosis and treatment links [34]. As the most important basis for formulating treatment plans, it is possible to decide whether or not to operate, and the postoperative of chemotherapy [35], some are even effective drugs that can accurately determine malignant tumors [36]. Among evidences, with the development of gene chip technology [37], the information provided by individual gene chip results has been paid more and more attention, and it can even be referred to together with the results of imaging or laboratory science in recent years [38].

In previous studies, many researchers also said that ANN model was used to analyze different clinical data, among which convolution neural network was used to analyze medical imaging and pathology [39-42]. However, there are few studies using ANN model to analyze bioinformatics data. In this study, we take papillary thyroid carcinoma as an example, which is the most common malignant tumor [43]. There is still a lot of room for the diagnosis of this solid tumor. We trained ANN model to analyze the expression difference between tumor tissue and normal thyroid tissue, and then applied it to other independent model data sets, and found that it can effectively detect papillary thyroid carcinoma from normal thyroid tissue: the AUCs of ANN models trained by DEGs and whole genes for independent verification set were 98.97% and 99.37%, the accuracy which is more commonly used in ANN network verification of prediction were both 96%.

Although the above results were surprising satisfactory, there are many shortcomings in the entire network training and data analysis process. Above the using of the ANN models, we found that when we used some special genes list as input features the loss rate of the training models were extremely difficult to converge. Sometimes no matter how we adjust the parameters, the model would not converge. In our research when we used m6a-related genes as input feature values, the loss rate of the DNN model will fluctuate if it was moderate (when reached 9000 epochs), and it was completely impossible to verify independently data sets to

 doi:10.20944/preprints202103.0324.v1

predict. When we used the trained ANN models to predict independent data, the rate of events(cancer) predicted by the model was always close to the situation of the train set we were training on, and this situation was inaccurate in clinical applications, because the entire population when used as a sample, the incidence rate will only be a particularly low rate, especially for malignant tumors prediction. These shortcomings may be solved by optimizing the codes and model parameters or expanding the sample size, but some of them may also be the dead defect of ANN model.

Although the application of artificial intelligence in medicine had many problems to be solved, we believe our study provided another influential method for the detection of gene-expression autographs, and the discovery of new feature lists that could subdivide the population according to their disease susceptibility risk.

Reference

1.Y. Mao, M. Xing Recent incidences and differential trends of thyroid cancer in the USA Endocr-Relat Cancer, 23 (4) (2016), pp. 313-322.

2. Lundgren CI, Hall P, Dickman PW, Zedenius J. Clinically significant prognostic factors for differentiated thyroid carcinoma: a population based, nested case-control study. Cancer. 2006;106(3):524–531.

3. Suen KC. Fine-needle aspiration biopsy of the thyroid. CMAJ. 2002 Sep 3;167(5):491-5. PMID: 12240817; PMCID: PMC121968.

4.Leipzig J. A review of bioinformatic pipeline frameworks. Brief Bioinform. 2017 May 1;18(3):530-536. doi: 10.1093/bib/bbw020. PMID: 27013646; PMCID: PMC5429012.

5.Eyileten C, Wicik Z, De Rosa S, Mirowska-Guzel D, Soplinska A, Indolfi C, Jastrzebska-Kurkowska I, Czlonkowska A, Postula M. MicroRNAs as Diagnostic and Prognostic Biomarkers in Ischemic Stroke-A Comprehensive Review and Bioinformatic Analysis. Cells. 2018 Dec 6;7(12):249. doi: 10.3390/cells7120249. PMID: 30563269; PMCID: PMC6316722.

6.Giddings MC, Shah AA, Freier S, Atkins JF, Gesteland RF, Matveeva OV. Artificial neural network prediction of antisense oligodeoxynucleotide activity. Nucleic Acids Res. 2002 Oct 1;30(19):4295-304. doi: 10.1093/nar/gkf557. PMID: 12364609; PMCID: PMC140555.

7.Zhang Z. When doctors meet with AlphaGo: potential application of machine learning to clinical medicine. Ann Transl Med. 2016 Mar;4(6):125. doi: 10.21037/atm.2016.03.25. PMID: 27127778; PMCID: PMC4828734.

8.Jujjavarapu SE, Deshmukh S. Artificial Neural Network as a Classifier for the Identification of Hepatocellular Carcinoma Through Prognosticgene Signatures. Curr Genomics. 2018 Sep;19(6):483-490. doi: 10.2174/1389202919666180215155234. PMID: 30258278; PMCID: PMC6128386.

9.Hou Q, Bing ZT, Hu C, Li MY, Yang KH, Mo Z, Xie XW, Liao JL, Lu Y, Horie S, Lou MW. RankProd Combined with Genetic Algorithm Optimized Artificial Neural Network Establishes a Diagnostic and Prognostic Prediction Model that Revealed C1QTNF3 as a Biomarker for Prostate Cancer. EBioMedicine. 2018 Jun;32:234-244. doi: 10.1016/j.ebiom.2018.05.010. Epub 2018 Jun 1. PMID: 29861410; PMCID: PMC6021271.

10.Yanagawa M, Niioka H, Hata A, Kikuchi N, Honda O, Kurakami H, Morii E, Noguchi M,

Watanabe Y, Miyake J, Tomiyama N. Application of deep learning (3-dimensional convolutional neural network) for the prediction of pathological invasiveness in lung adenocarcinoma: A preliminary study. Medicine (Baltimore). 2019 Jun;98(25):e16119. doi: 10.1097/MD.0000000000016119. PMID: 31232960; PMCID: PMC6636940.

11.Mehdy MM, Ng PY, Shair EF, Saleh NIM, Gomes C. Artificial Neural Networks in Image Processing for Early Detection of Breast Cancer. Comput Math Methods Med. 2017;2017:2610628. doi: 10.1155/2017/2610628. Epub 2017 Apr 3. PMID: 28473865; PMCID: PMC5394406.

12.Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002 Jan 1;30(1):207-10. doi: 10.1093/nar/30.1.207. PMID: 11752295; PMCID: PMC99122.

13.Tomás G, Tarabichi M, Gacquer D, Hébrant A, Dom G, Dumont JE, Keutgen X, Fahey TJ 3rd, Maenhaut C, Detours V. A general method to derive robust organ-specific gene expression-based differentiation indices: application to thyroid cancer diagnostic. Oncogene. 2012 Oct 11;31(41):4490-8. doi: 10.1038/onc.2011.626. Epub 2012 Jan 23. PMID: 22266856.

14.He H, Jazdzewski K, Li W, Liyanarachchi S, Nagy R, Volinia S, Calin GA, Liu CG, Franssila K, Suster S, Kloos RT, Croce CM, de la Chapelle A. The role of microRNA genes in papillary thyroid carcinoma. Proc Natl Acad Sci U S A. 2005 Dec 27;102(52):19075-80. doi: 10.1073/pnas.0509603102. Epub 2005 Dec 19. PMID: 16365291; PMCID: PMC1323209.

15.Pita JM, Banito A, Cavaco BM, Leite V. Gene expression profiling associated with the progression to poorly differentiated thyroid carcinomas. Br J Cancer. 2009 Nov 17;101(10):1782-91. doi: 10.1038/sj.bjc.6605340. Epub 2009 Oct 6. PMID: 19809427; PMCID: PMC2778548.

16.Lee S, Bae JS, Jung CK, Chung WY. Extensive lymphatic spread of papillary thyroid microcarcinoma is associated with an increase in expression of genes involved in epithelial-mesenchymal transition and cancer stem cell-like properties. Cancer Med. 2019 Nov;8(15):6528-6537. doi: 10.1002/cam4.2544. Epub 2019 Sep 9. PMID: 31498560; PMCID: PMC6825983.

17.Giordano TJ, Kuick R, Thomas DG, Misek DE, Vinco M, Sanders D, Zhu Z, Ciampi R, Roh M, Shedden K, Gauger P, Doherty G, Thompson NW, Hanash S, Koenig RJ, Nikiforov YE.

Molecular classification of papillary thyroid carcinoma: distinct BRAF, RAS, and RET/PTC mutation-specific gene expression profiles discovered by DNA microarray analysis. Oncogene. 2005 Oct 6;24(44):6646-56. doi: 10.1038/sj.onc.1208822. PMID: 16007166.

18.Tarabichi M, Saiselet M, Trésallet C, Hoang C, Larsimont D, Andry G, Maenhaut C, Detours V. Revisiting the transcriptional analysis of primary tumours and associated nodal metastases with enhanced biological and statistical controls: application to thyroid cancer. Br J Cancer. 2015 May 12;112(10):1665-74. doi: 10.1038/bjc.2014.665. PMID: 25965298; PMCID: PMC4430711.

19.Delys L, Detours V, Franc B, Thomas G, Bogdanova T, Tronko M, Libert F, Dumont JE, Maenhaut C. Gene expression and the biological phenotype of papillary thyroid carcinomas. Oncogene. 2007 Dec 13;26(57):7894-903. doi: 10.1038/sj.onc.1210588. Epub 2007 Jul 9. PMID: 17621275.

20.Handkiewicz-Junak D, Swierniak M, Rusinek D, Oczko-Wojciechowska M, Dom G, Maenhaut C, Unger K, Detours V, Bogdanova T, Thomas G, Likhtarov I, Jaksik R, Kowalska M, Chmielik E, Jarzab M, Swierniak A, Jarzab B. Gene signature of the post-Chernobyl papillary thyroid cancer. Eur J Nucl Med Mol Imaging. 2016 Jul;43(7):1267-77. doi: 10.1007/s00259-015-3303-3. Epub 2016 Jan 26. PMID: 26810418; PMCID: PMC4869750.

21.Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012 Mar 15;28(6):882-3. doi: 10.1093/bioinformatics/bts034. Epub 2012 Jan 17. PMID: 22257669; PMCID: PMC3307112.

22.Cleary JM, Aguirre AJ, Shapiro GI, D'Andrea AD. Biomarker-Guided Development of DNA Repair Inhibitors. Mol Cell. 2020 Jun 18;78(6):1070-1085. doi: 10.1016/j.molcel.2020.04.035. Epub 2020 May 26. PMID: 32459988; PMCID: PMC7316088.

23.Yang J, Ueharu H, Mishina Y. Energy metabolism: A newly emerging target of BMP signaling in bone homeostasis. Bone. 2020 Sep;138:115467. doi: 10.1016/j.bone.2020.115467. Epub 2020 Jun 5. PMID: 32512164; PMCID: PMC7423769.

24.Lu Q, Guo P, Liu A, Ares I, Martínez-Larrañaga MR, Wang X, Anadón A, Martínez MA. The role of long noncoding RNA in lipid, cholesterol, and glucose metabolism and treatment of obesity syndrome. Med Res Rev. 2020 Dec 24. doi: 10.1002/med.21775. Epub ahead of print.

PMID: 33368430.

25.Mohammadinejad R, Biagioni A, Arunkumar G, Shapiro R, Chang KC, Sedeeq M, Taiyab A, Hashemabadi M, Pardakhty A, Mandegary A, Thiery JP, Aref AR, Azimi I. EMT signaling: potential contribution of CRISPR/Cas gene editing. Cell Mol Life Sci. 2020 Jul;77(14):2701-2722. doi: 10.1007/s00018-020-03449-3. Epub 2020 Feb 1. PMID: 32008085.

26.Namba Y, Sogawa C, Okusha Y, Kawai H, Itagaki M, Ono K, Murakami J, Aoyama E, Ohyama K, Asaumi JI, Takigawa M, Okamoto K, Calderwood SK, Kozaki KI, Eguchi T. Depletion of Lipid Efflux Pump ABCG1 Triggers the Intracellular Accumulation of Extracellular Vesicles and Reduces Aggregation and Tumorigenesis of Metastatic Cancer Cells. Front Oncol. 2018 Oct 10;8:376. doi: 10.3389/fonc.2018.00376. PMID: 30364132; PMCID: PMC6191470.

27.Xie X, Stubbington MJ, Nissen JK, Andersen KG, Hebenstreit D, Teichmann SA, Betz AG. The Regulatory T Cell Lineage Factor Foxp3 Regulates Gene Expression through Several Distinct Mechanisms Mostly Independent of Direct DNA Binding. PLoS Genet. 2015 Jun 24;11(6):e1005251. doi: 10.1371/journal.pgen.1005251. PMID: 26107960; PMCID: PMC4480970.

28.Wang T, Kong S, Tao M, Ju S. The potential role of RNA N6-methyladenosine in Cancer progression. Mol Cancer. 2020 May 12;19(1):88. doi: 10.1186/s12943-020-01204-7. PMID: 32398132; PMCID: PMC7216508.

29.Lin X, Ping J, Wen Y, Wu Y. The Mechanism of Ferroptosis and Applications in Tumor Treatment. Front Pharmacol. 2020 Jul 22;11:1061. doi: 10.3389/fphar.2020.01061. PMID: 32774303; PMCID: PMC7388725.

30.Yi D, Song P, Huang T, Tang X, Sang J. A meta-analysis on the effect of operation modes on the recurrence of papillary thyroid microcarcinoma. Oncotarget. 2017 Jan 24;8(4):7148-7156. doi: 10.18632/oncotarget.12698. PMID: 27756889; PMCID: PMC5351696.

31.Kitahara CM, Sosa JA. The changing incidence of thyroid cancer. Nat Rev Endocrinol. 2016 Nov;12(11):646-653. doi: 10.1038/nrendo.2016.110. Epub 2016 Jul 15. PMID: 27418023.

32.Liu FC, Lin HT, Lin SF, Kuo CF, Chung TT, Yu HP. Nationwide cohort study on the epidemiology and survival outcomes of thyroid cancer. Oncotarget. 2017 Jul 22;8(45):78429-78451. doi: 10.18632/oncotarget.19488. PMID: 29108240; PMCID: PMC5667973.

33.Li C, Zhou L, Dionigi G, Li F, Zhao Y, Sun H. THE ASSOCIATION BETWEEN TUMOR TISSUE CALCIFICATION, OBESITY, AND THYROID CANCER INVASIVENESS IN A COHORT STUDY. Endocr Pract. 2020 Jun 23. doi: 10.4158/EP-2020-0057. Epub ahead of print. PMID: 32576045.

34.Huang X, Lin X, Zeng J, Wang L, Yin P, Zhou L, Hu C, Yao W. A Computational Method of Defining Potential Biomarkers based on Differential Sub-Networks. Sci Rep. 2017 Oct 30;7(1):14339. doi: 10.1038/s41598-017-14682-5. PMID: 29085035; PMCID: PMC5662748.

35.Meng F, Sun Y, Lee RJ, Wang G, Zheng X, Zhang H, Fu Y, Yan G, Wang Y, Deng W, Parks E, Kim BYS, Yang Z, Jiang W, Teng L. Folate Receptor-Targeted Albumin Nanoparticles Based on Microfluidic Technology to Deliver Cabazitaxel. Cancers (Basel). 2019 Oct 16;11(10):1571. doi: 10.3390/cancers11101571. PMID: 31623082; PMCID: PMC6827099.

36.Decker B, Karyadi DM, Davis BW, Karlins E, Tillmans LS, Stanford JL, Thibodeau SN, Ostrander EA. Biallelic BRCA2 Mutations Shape the Somatic Mutational Landscape of Aggressive Prostate Tumors. Am J Hum Genet. 2016 May 5;98(5):818-829. doi: 10.1016/j.ajhg.2016.03.003. Epub 2016 Apr 14. PMID: 27087322; PMCID: PMC4863563.

37.Li H, Liu S, Yang X, Du Y, Luo J, Tan J, Sun Y. Cellular Processes Involved in Jurkat Cells Exposed to Nanosecond Pulsed Electric Field. Int J Mol Sci. 2019 Nov 21;20(23):5847. doi: 10.3390/ijms20235847. PMID: 31766457; PMCID: PMC6929111.

38.Azizipour N, Avazpour R, Rosenzweig DH, Sawan M, Ajji A. Evolution of Biochip Technology: A Review from Lab-on-a-Chip to Organ-on-a-Chip. Micromachines (Basel). 2020 Jun 18;11(6):599. doi: 10.3390/mi11060599. PMID: 32570945; PMCID: PMC7345732.

39.Acs B, Rantalainen M, Hartman J. Artificial intelligence as the next step towards precision pathology. J Intern Med. 2020 Jul;288(1):62-81. doi: 10.1111/joim.13030. Epub 2020 Mar 3. PMID: 32128929.

40.Wong DJ, Gandomkar Z, Wu WJ, Zhang G, Gao W, He X, Wang Y, Reed W. Artificial intelligence and convolution neural networks assessing mammographic images: a narrative literature review. J Med Radiat Sci. 2020 Jun;67(2):134-142. doi: 10.1002/jmrs.385. Epub 2020 Mar 5. PMID: 32134206; PMCID: PMC7276180.

41.Joy Mathew C, David AM, Joy Mathew CM. Artificial Intelligence and its future potential in lung cancer screening. EXCLI J. 2020 Dec 11;19:1552-1562. doi: 10.17179/excli2020-3095.

PMID: 33408594; PMCID: PMC7783473.

42.Avanzo M, Stancanello J, Pirrone G, Sartor G. Radiomics and deep learning in lung cancer. Strahlenther Onkol. 2020 Oct;196(10):879-887. doi: 10.1007/s00066-020-01625-9. Epub 2020 May 4. PMID: 32367456.

43.Limaiem F, Rehman A, Mazzoni T. Papillary Thyroid Carcinoma. 2021 Jan 6. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2021 Jan–. PMID: 30725628.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2021 doi:10.20944/preprints202103.0324.v1

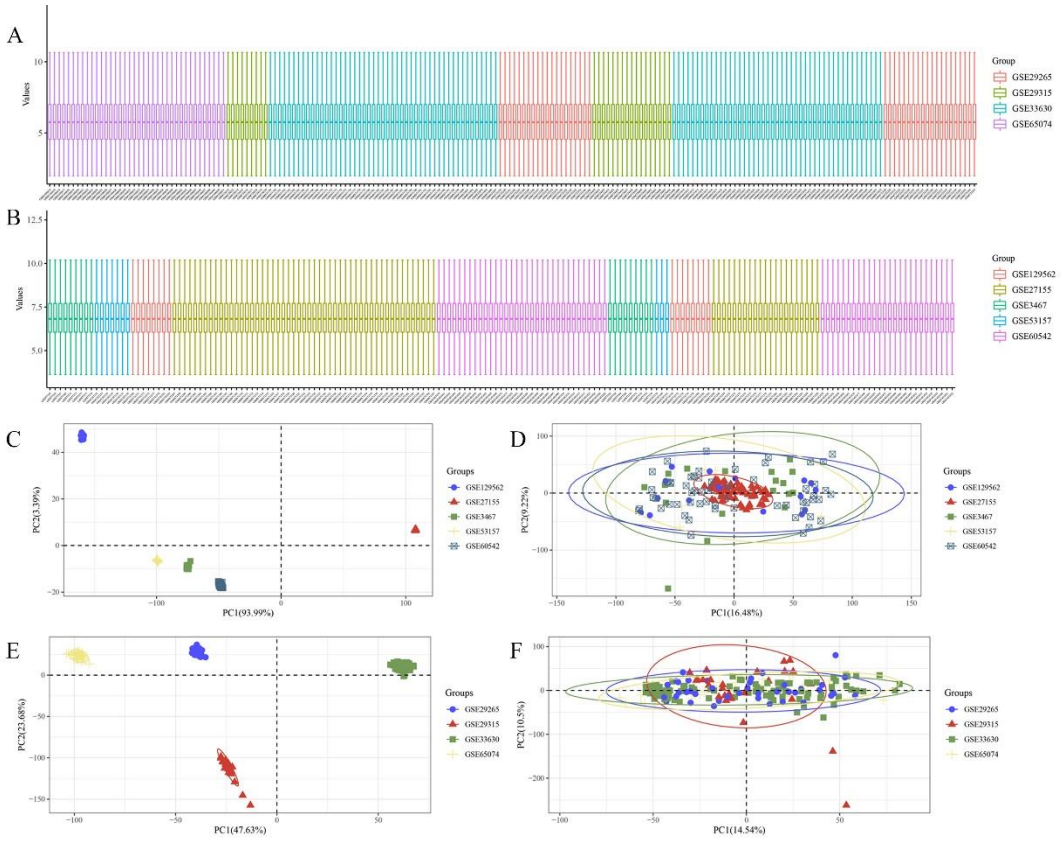

Figure 1 A, B: Box plot after data standardization, different colors represent different data sets. C, E: PCA results before batch removal for multiple data sets. Different colors represent different data sets. As shown in the schematic diagram, the three data sets are separated without any intersection; D, F: PCA results after batch removal, as shown in the schematic diagram Shows the intersection of three data sets, which can be used as a batch of data for subsequent analysis.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2021 doi:10.20944/preprints202103.0324.v1

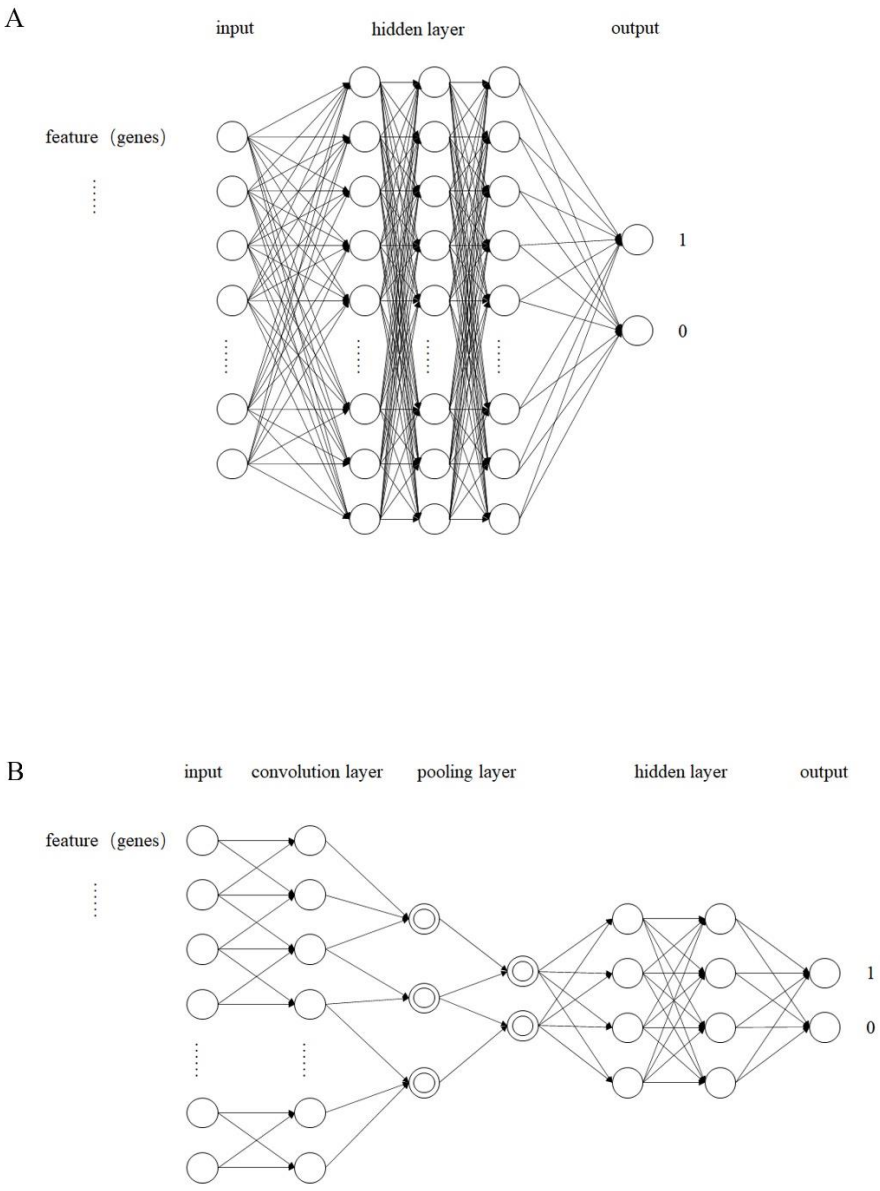

Figure 2 Sketch map of artificial neural network models. A: Deep Neural Networks. B: Convolutional Neural Networks.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2021 doi:10.20944/preprints202103.0324.v1

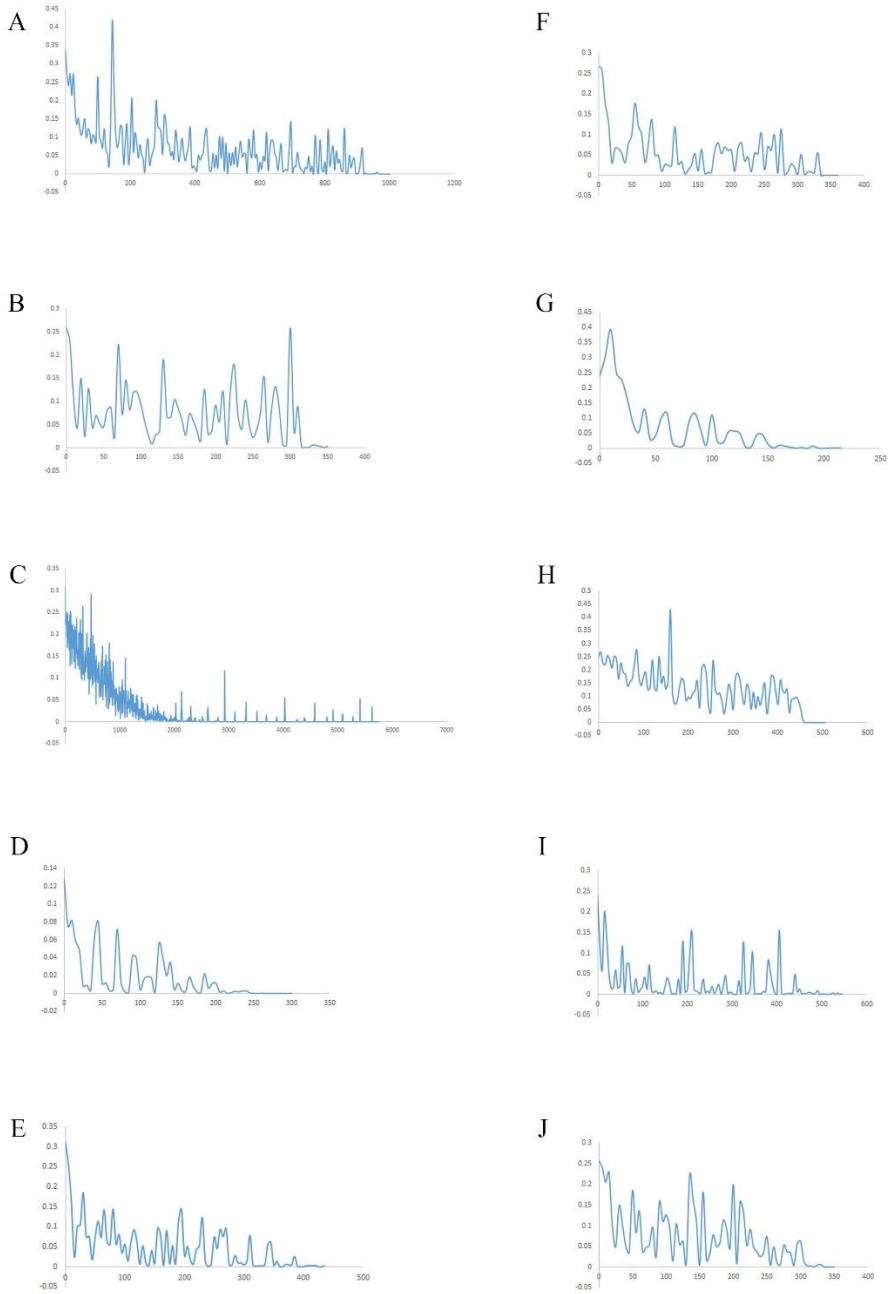

Figure 3 The loss rate in the training model cycles using the DNN models: From A-J, DNA Damage Repair (DDR) Genes (232) list, epithelial mesenchymal transition (EMT) -associated signature genes (1263) list, N6-methyladenosine(m6a) related genes (20) list, DEGs (198) list, immune signature (IS) genes (1959) list, genes involved in anaerobic energy metabolism (590) list, all genes (8647) list, Hypoxia-related genes (75) list, ferroptosis related genes (24) list and lipid metabolism related genes (751).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2021 doi:10.20944/preprints202103.0324.v1

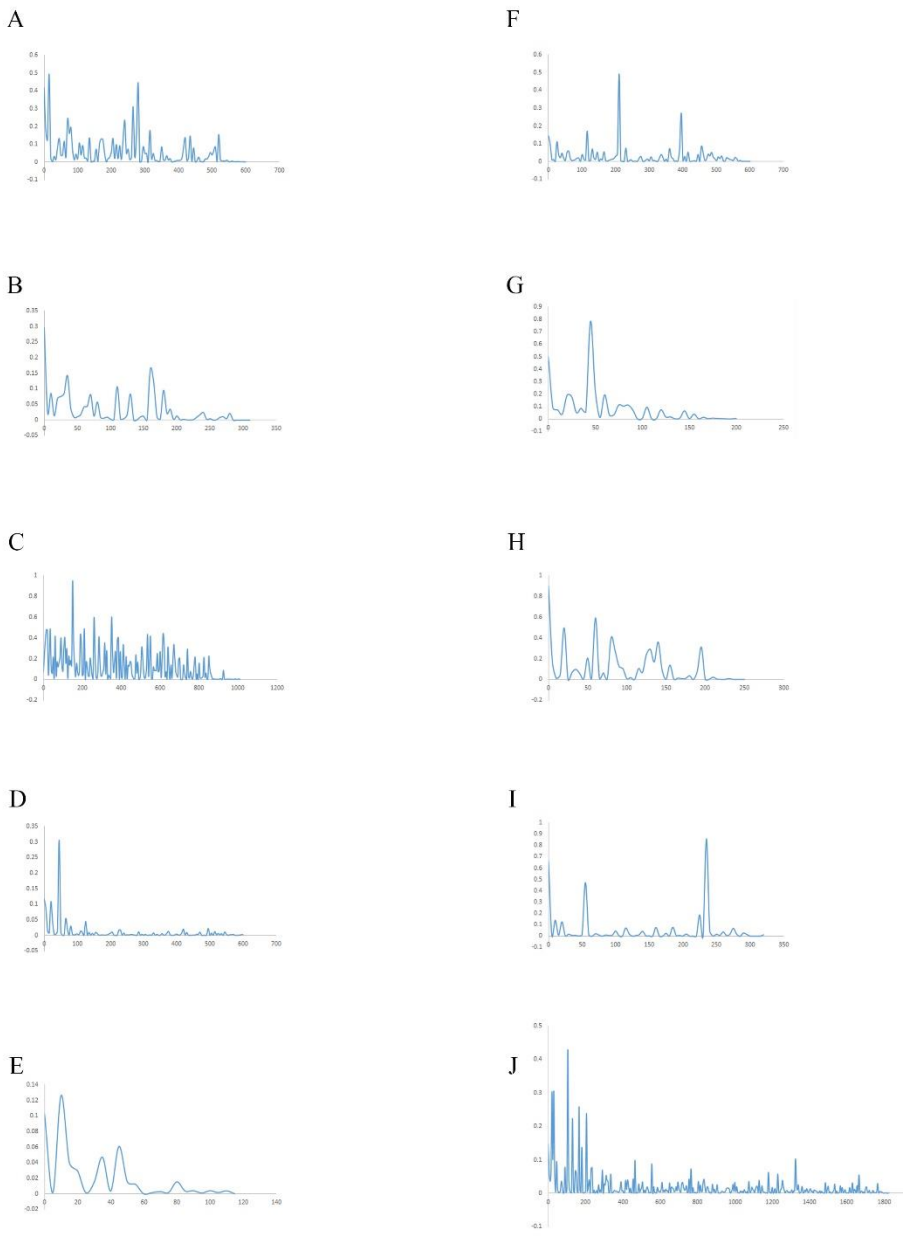

Figure 4 The loss rate in the training model cycles using the CNN models: From A-J, DNA Damage Repair (DDR) Genes (232) list, epithelial mesenchymal transition (EMT) -associated signature genes (1263) list, N6-methyladenosine(m6a) related genes (20) list, DEGs (198) list, immune signature (IS) genes (1959) list, genes involved in anaerobic energy metabolism (590) list, all genes (8647) list, Hypoxia-related genes (75) list, ferroptosis related genes (24) list and lipid metabolism related genes (751).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2021 doi:10.20944/preprints202103.0324.v1

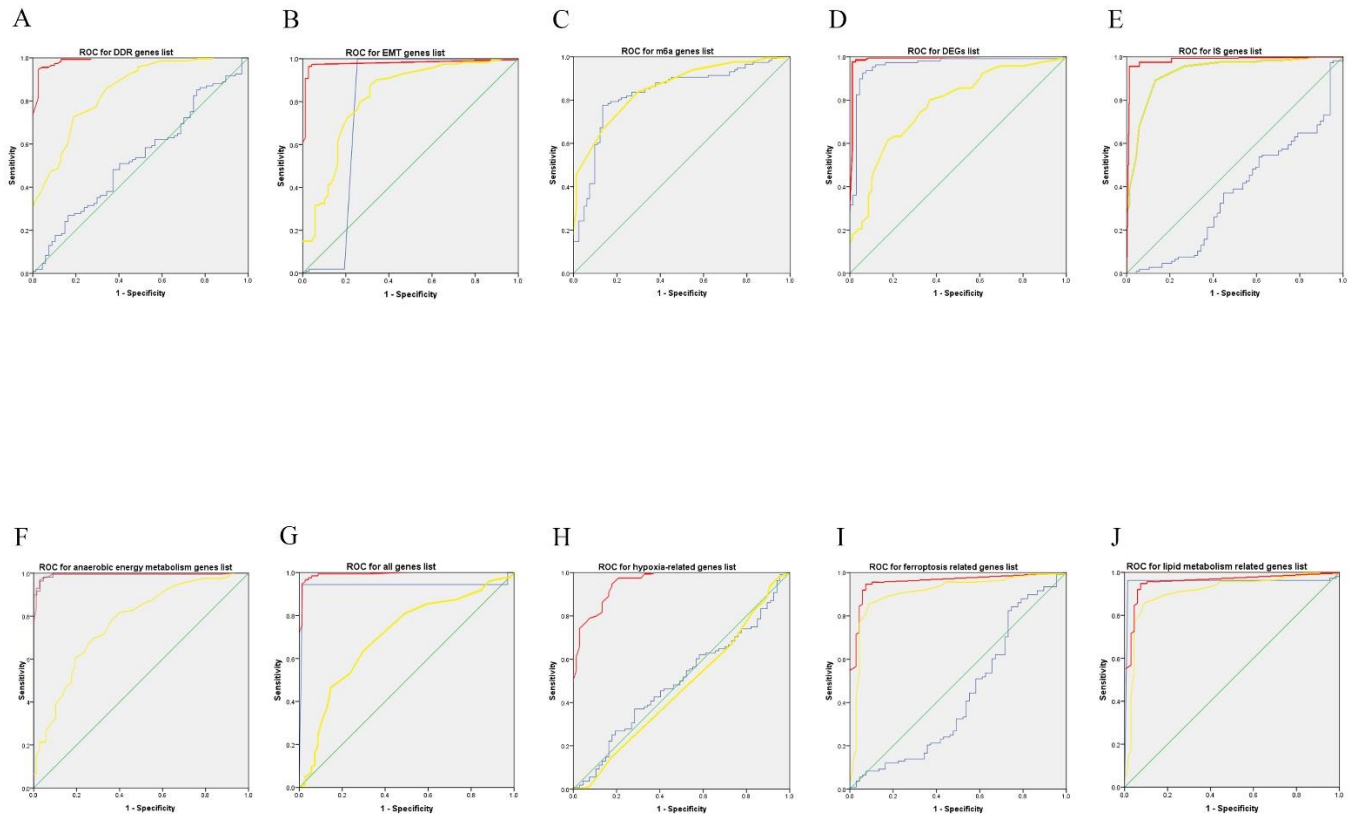

Figure 5 Comparison of ROC for DNN predict models (red curves), CNN predict models (yellow curves) and logistic regression (blue curves). No matter how to adjust the parameters DNN predict models for m6a genes list could not predict, so the red curve in C was missing. We put one of the confusion matrices into the supporting file.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 March 2021 doi:10.20944/preprints202103.0324.v1

Table 1 Performance comparison of ANN models and logistic regression in predicting independent test data set validation.

	Sensitivity (%)	Specificity (%)	AUC (%)	Accuracy (%)
DNN model for DDR genes	98.51	93.75	98.88	94.29
DNN model for EMT genes	88.89	98.51	97.96	92.57
DNN model for m6a related genes	-	-	-	-
DNN model for DEGs	99.07	91.04	98.97	96.00
DNN model for IS genes	95.37	94.03	98.46	94.86
DNN model for anaerobic energy metabolism genes	99.07	44.78	99.26	82.29
DNN model for all genes	94.44	98.51	99.37	96.00
DNN model for Hypoxia-related genes	97.22	68.66	95.62	86.29
DNN model for ferroptosis related genes	87.96	94.03	95.79	90.29
DNN model for lipid metabolism related genes	86.11	98.51	99.61	90.85
CNN model for DDR genes	67.30	93.75	84.33	69.71
CNN model for EMT genes	30.56	94.03	81.83	54.86
CNN model for m6a related genes	97.22	25.37	85.49	69.71
CNN model for DEGs	20.37	95.52	77.08	49.14
CNN model for IS genes	98.15	35.82	92.78	74.29
CNN model for anaerobic energy metabolism genes	12.96	98.51	76.65	45.71
CNN model for all genes	100*	0*	69.82*	61.71*
CNN model for Hypoxia-related genes	0.93	92.54	46.65	36.00
CNN model for ferroptosis related genes	99.07	5.97	90.88	63.42
CNN model for lipid metabolism related genes	25.93	95.52	86.20	52.57
LR for DDR genes	-	-	-	53.40
LR for EMT genes	-	-	-	78.00
LR for m6a related genes	-	-	-	83.10
LR for DEGs	-	-	-	91.60
LR for IS genes	-	-	-	36.70
LR for anaerobic energy metabolism genes	-	-	-	52.70
LR for all genes	-	-	-	93.90
LR for Hypoxia-related genes	-	-	-	49.80
LR for ferroptosis related genes	-	-	-	99.30
LR for lipid metabolism related genes	-	-	-	95.60

* when we used CNN model for all genes to predict the test data set all individuals were recognized as 1 no matter how to adjust the parameters