Send Orders for Reprints to reprints@benthamscience.net

Current Genomics, 2022, *23*, 94-108

RESEARCH ARTICLE

Comparisons of Forecasting for Survival Outcome for Head and Neck Squamous Cell Carcinoma by using Machine Learning Models based on Multi-omics

Liying Mo[1,#], Yuangang Su[1,2,#], Jianhui Yuan[1,3], Zhiwei Xiao[4], Ziyan Zhang[5], Xiuwan Lan[1,#] and Daizheng Huang[1,3,*]

[1]*School of Basic Medical Sciences, Guangxi Medical University, Nanning, Guangxi, China;* [2]*Research Centre for Regenerative Medicine, Guangxi Key Laboratory of Regenerative Medicine, Guangxi Medical University, Nanning, Guangxi, China;* [3]*The Laboratory of Biomedical Photonics and Engineering, Guangxi Medical University, Nanning, China;* [4]*School of Information and Management, Guangxi Medical University, Nanning, Guangxi, China;* [5]*Life Sciences Institute, Guangxi Medical University, Nanning, Guangxi, China*

ARTICLE HISTORY

Received: December 03, 2021
Revised: January 13, 2022
Accepted: January 19, 2022

DOI:
10.2174/1389202923666220204153744

Abstract: ***Background*:** Machine learning methods showed excellent predictive ability in a wide range of fields. For the survival of head and neck squamous cell carcinoma (HNSC), its multi-omics influence is crucial. This study attempts to establish a variety of machine learning multi-omics models to predict the survival of HNSC and find the most suitable machine learning prediction method.

***Methods*:** The HNSC clinical data and multi-omics data were downloaded from the TCGA database. The important variables were screened by the LASSO algorithm. We used a total of 12 supervised machine learning models to predict the outcome of HNSC survival and compared the results. *In vitro* qPCR was performed to verify core genes predicted by the random forest algorithm.

***Results*:** For omics of HNSC, the results of the twelve models showed that the performance of multi-omics was better than each single-omic alone. Results were presented, which showed that the Bayesian network(BN) model (area under the curve [AUC] 0.8250, F1 score=0.7917) and random forest(RF) model (area under the curve [AUC] 0.8002,F1 score=0.7839) played good prediction performance in HNSC multi-omics data. The results of *in vitro* qPCR were consistent with the RF algorithm.

***Conclusion*:** Machine learning methods could better forecast the survival outcome of HNSC. Meanwhile, this study found that the BN model and the RF model were the most superior. Moreover, the forecast result of multi-omics was better than single-omic alone in HNSC.

Keywords: Machine learning models, multi-omics integration, head and neck squamous cell carcinoma, survival prediction, bayesian network, random forest.

1. INTRODUCTION

Cancer is a major public health problem and causes 1 in 6 deaths around the world [1]. Head and neck squamous cell carcinoma (HNSC), which arises from multiple anatomic subsites in the head and neck region, is the seventh most common cancer worldwide. There is marked heterogeneity of tumors arising from the mucosal epithelium of the upper aerodigestive tract [2, 3]. The risk factors for the development of cancers of the oral cavity, oropharynx, hypopharynx, and larynx include tobacco exposure and alcohol dependence, and infection with oncogenic viruses is associated with cancers developing in the nasopharynx, palatine, and lingual tonsils of the oropharynx [4]. The high level of heterogeneity in HNSC, along with the complex etiological factors, makes the prognosis prediction deeply challenging. In the treatment of HNSC, a multispecialty team to evaluate the treatment choice is very important since the head and neck cancers differ from the patients' statement, molecular change, and other environmental factors, such as alcohol and smoking. Tobacco smoking and alcohol drinking are used as all-cause mortality to diagnose HNSC [3, 5]. Surgery, radiation, and chemotherapy in various combinations are utilized for the treatment of HNSC [6]. But all of these treatments are associated with toxicity which can lead to different degrees of late organ dysfunction or other serious adverse reactions [7]. The main method of evaluating cancer development and providing survival estimation prognosis prediction is the main method of evaluating cancer development and survival estimation, mainly based on patients'

*Address correspondence to this author at the School of Basic Medical Sciences, Guangxi Medical University, Nanning, Guangxi, China; The Laboratory of Biomedical Photonics and Engineering, Guangxi Medical University, Nanning, China; Tel: +867715358270;
E-mail: huangdaizheng@gxmu.edu.cn

# *These authors contributed equally to this work.*

1875-5488/22 $65.00+.00

clinical features and molecular profile [8]. Some studies [9, 10] applied public databases such as TCGA and GEO datasets to identify biomarkers associated with the prognosis of cancer patients and predict the clinical outcome. Ghafouri-Fard S*et al.* [11] found the role of miRNA as prognostic biomarkers in HNSC. Although the single genomic analysis approaches have contributed towards the identification of cancer-specific mutations and molecular subtyping of tumors [12], single-omic only considers the role of single molecular biological information. For HNSC, there was little study of multi-omics data to its survival. Therefore, we attempted to apply multi-omics molecular biology information to predict the survival of HNSC, at the same time, apply machine learning methods as a predictive tool.

Currently, many studies analyzed diseases not only from the level of gene expression alone but also from the multi-omics level. Multi-omics (mRNA, miRNA, DNA methylation, and copy number variation) is part of the prognostic effect. Multi-omics integration analysis and deep learning are used to predict high-grade patient survival and prognosis risk biomarkers [8, 13, 14]. HNSC is involved in a variety of complex mechanisms at the molecular level *in vivo*, and it is difficult to understand the development of cancer from gene-level alone, thereby making assessments to the patients. Integrative analyses that use information across the multi-omics profiling modalities promise to deliver more comprehensive insights into the survival prediction of cancer [15]. In the field of precision oncology, genomics approaches analyses have helped reveal several key mechanisms in cancer development, and several findings have been implemented in clinical oncology to help guide treatment decisions [16]. Moreover, multi-omics approaches can dissect the cellular response to chemo-/ immunotherapy as well as discover molecular candidates with diagnostic/prognostic value [17]. In summary, multi-omics integration models driven by multi-omics data may help overcome the chemo-/immunotherapy resistance phenotype of cancer cells, rendering them vulnerable to targeted therapies and ultimately improving the quality of life of patients [17].

With the development of artificial intelligence, machine learning (ML) methods have already achieved some success in the field of medical-related research. In particular, some prior approaches [18, 19] have shown that various ML techniques could be used in an automated way to predict trends in various chemical and material systems. The classification of genomics data can be performed through machine learning algorithms to find significant features related to survival as well. Combined ML, which includes random forest (RF), K-nearest neighbor (KNN), and artificial neural networks (ANN), is used to identify prognostic biomarkers in colorectal cancer; the performance of using RF is better [20]. Kaplan-Meier (KM), LASSO, and COX regression are performed to analyze the effect of CA9 on the survival of tongue squamous cell carcinoma (TSCC) [21]. Multi-omics data integration through machine learning (autoencoder and XGboost model) to construct an accurate and robust cancer prognosis prediction could cause abnormal C-index values fluctuations due to the neglect of tumor purity and known clinical data that affect the occurrence and development of tumors [22]. Therefore, the survival prediction that integrates multiple omics data and clinical data may acquire a robust and reliable prognostic prediction result. Fujino *et al*. [23] applied LASSO regression to predict the future visual field progression in glaucoma patients. A new algorithm based on LASSO called TG-LASSO was developed, which could predict clinical drug response of cancer patients and identify genes related to drug response, including known targets genes and pathways related to the drug action mechanisms [24]. Likewise, BN modeling has been used to develop decision-support tools in various oncologic diagnoses [25]. Burghardt *et al.* [26] indicated that BN could explain the data set by defining the phenotype and pathogenicity of the given mutation position and the conditional probability of the residual substitution. Moreover, myypbc3 mutant disease was predicted *via* the neural/Bayesian network. Bellot *et al.* [27] used deep learning to do the performance about the genomic prediction of complex human traits.

In summary, a growing body of research has applied machine learning, including LASSO algorithm, cox regression, BN, and neural network to analyze tumor data of integrated multi-omics. This study built up a combination of DNA methylation, a gene expression data analysis, a copy number variation, and a miRNA data analysis as a multi-omics integration. The aim of this study was to use machine learning models to forecast the survival outcome of HNSC and compare the prediction performance in each omics. This comparison may be helpful in describing the hierarchal relationships between prognostic and outcome variables. Meanwhile, we aim to determine which machine learning model would be suitable for clinical use with decision curve analysis. Multiple machine learning methods including LASSO and BN combined model, naive bayesian (NB), logistic regression (LR), generalized linear model (GLM), K-nearest neighbor (KNN), decision tree (DT), RF, bootstrap aggregating (bagging), Adaboost, gradient boosting decision tree (GBDT), neural networks (NN) and support vector machine (SVM), were used for prediction and their performances were compared.

2. MATERIALS AND METHODS

2.1. The Datasets Source and Data Pre-processing

RNA-sequencing data (IlluminaHiSeq_RNASeqV2; Level 3), miRNA-seq data (IlluminaHiSeq_miRNASeq; Level 3), DNA methylation data (HumanMethylation450; Level 3), copy number variation(CNV) data (Affymetrix SNP 6.0 array; Level 3), and corresponding clinical information from HNSC were obtained from The Cancer Genome Atlas (TCGA) database (https://portal.gdc.cancer.gov/), in which the method of acquisition and application complied with the guidelines and policies. The clinical information of HNSC samples was downloaded as well. Meanwhile, the tumor samples of the multi-omics were selected by filtering out the samples according to the nomenclature of TCGA sample IDs.

For the downloaded dataset, data pre-processing and dimensionality reduction were required. The gene expression data and miRNA expression data were identified by comparison with tumor tissues and normal tissues expression and using the edgeR, and DeSeq2 packages in R. Meanwhile, the chi-square test, and Kruskal-Wallis test was used to reduce the number of genes in order to obtain DEGs

in CNV data of HNSC. And for methylation data of HNSC, the limma package in R was used to filter DEGs.

Unless otherwise specified in the analysis of this paper, the programming language used is R (version 4.0.1).

2.2. Machine Learning

2.2.1. Least Absolute Shrinkage and Selection Operator

The least absolute shrinkage and selection operator (LASSO) is a regression-based algorithm that permits a large number of covariates in the model and penalizes the absolute value of the regression coefficient [28]. It is a linear regression method that uses L1 regularization, which can achieve the purpose of sparseness and feature selection. The LASSO regression is applied to data dimensionality reduction and feature selection due to its outstanding feature extraction and robust cancer prognosis [29]. Formula 1 described the representation method of the minimum residual sum of squares of the LASSO algorithm.

$$\arg\min_{\beta}\left\{\sum_{i=1}^{q}\left(y_i-\sum_{j=1}^{m}x_{ij}\beta_j\right)^2\right\}\text{subject}\sum_{j=1}^{m}|\beta_j|=\gamma \quad (1)$$

2.2.2. Bayesian Network

Bayesian network (BN) is a multi-layered network of connections between clinical factors in a multi-omics data set that provides a multivariate mapping of complex data [30]. BN is a directed acyclic graph. Its nodes represent some random variables (Fig. **1**). Some of these random variables are observable, and some are unobservable. Meanwhile, BN is a probabilistic graph model with a clear and transparent representation of the causal relationship between variables. Importantly, because the BN uses the posterior information of the data sets itself, it protects against overinterpretation of the data. Survival predictions based on BN models have been developed for a number of tumors to improve prognostic estimates and guide clinical decision-making for appropriate treatment [31, 32]. The BNs is one of the deep learning model methods, which also has the deep learning model's advantages. BNs with proper external validation could be useful as clinical decision support tools and provide clinicians and patients with information germane to the treatment of HNSC.

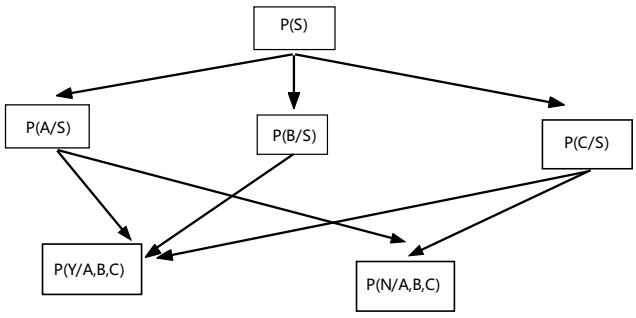

Fig. (1). A simple BN network diagram.

2.2.3. Naive Bayesian

Naive bayesian (NB) is one of the most widely used classification algorithms. NB is a method based on Bayes' theorem and assumes that the feature conditions are independent of each other. First, through the given training set, we need to take the independence between feature words as the premise and learn the joint probability distribution from input to output. Then based on the learned model, we would input X to find the output Y that maximizes the posterior probability.

2.2.4. Logistic Regression

Logistic regression(LR) is a probability-based pattern recognition algorithm. LR is based on linear regression theory, but it introduces nonlinear factors through the sigmoid function, which could easily handle 0/1 classification problems. In practical applications, LR could be said to be one of the most widely used machine learning algorithms.

2.2.5. Generalized Linear Model

The generalized linear model (GLM) is based on the exponential distribution family, and the prototype of the exponential distribution family is Formula 2.

$$P(y;\eta)=b(y)\cdot\exp\left(\eta^T T(y)-a(\eta)\right) \quad (2)$$

Where η is a natural parameter, it may be a vector. $T(y)$ is called a sufficient statistic.

2.2.6. K-Nearest Neighbor

The K-nearest neighbor(KNN) classification algorithm is one of the simplest data mining technology methods. The core idea of the KNN algorithm is that if most of the K nearest neighbors of a sample in the feature space belong to a certain category, the sample also belongs to this category and has the characteristics of the samples in this category.

2.2.7. Decision Tree

A decision tree (DT) is a basic classification and regression method composed of nodes and directed edges. The DT reflected the mapping relationship between features and tags as well. DT learning is a process of recursively selecting the optimal feature and segmenting the training data according to the feature so that each sub-data set has the best classification process.

2.2.8. Random Forest

Random forest (RF) is an integrated learning method based on decision trees. At the same time, RF is also an improvement to the bagging algorithm. The process of RF is shown in Fig. (**2**).

2.2.9. Bootstrap Aggregating

Bootstrap aggregating (bagging) is an ensemble method that reduces generalization error by combining several models. The core idea is to train several different models separately and then let all models vote on the output of the test example. This is an example of a conventional strategy in machine learning known as model averaging.

2.2.10. Adaboost

Adaboost is an iterative algorithm whose core idea is to train different classifiers (weak classifiers) for the same training set and then combine these weak classifiers to form a stronger final classifier (strong classifier).

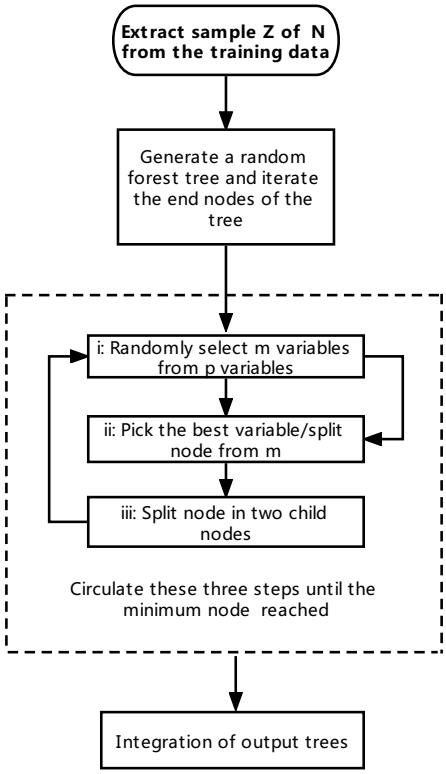

Fig. (2). The process of RF.

2.2.11. Gradient Boosting Decision Tree

Gradient boosting decision tree (GBDT) is also an ensemble method. The main idea of GBDT is that each time a model is established, the gradient descent direction of the model loss function is established before. The loss function is to evaluate the performance of the model (generally, the degree of fit and regularity); the smaller the loss function, the better the performance. So the best way to improve the performance of the model is to make the loss function descend along the gradient direction.

2.2.12. Neural Networks

Neural networks (NN) is a two-stage regression or classification model, which is a complex network system formed by a large number of simple processing units (called neurons) widely connected to each other. It is a highly complex nonlinear dynamic learning system.The network diagram is shown in Fig. (**3**).

2.2.13. Support Vector Machine

Support vector machine (SVM) is a generalized linear classifier that classifies data binary in a supervised learning manner, and its decision boundary is the maximum-margin hyperplane that solves the learning sample. SVM could perform nonlinear classification through the kernel method and is a classifier with sparsity and robustness.

2.2.14. Evaluation Index of Performance: AUC, F1 Score

Since the accuracy rate cannot fully evaluate the performance of the models, this study considered other evaluation indicators, namely the AUC value and F1score. AUC value and F1score are the performance indicator to measure the pros and cons of machine learning. The F1 score takes into account both the precision rate and the recall rate. AUC is the abbreviation of the area under the ROC curve. As the name implies, the value of AUC is the size of the area under the ROC curve. The definitions of AUC value and F1 value are given below:

AUC value: $sensitive = \frac{TP}{P}$; $secitificity = \frac{TN}{FP+TN}$;

The ROC curve is drawn by two variables. The abscissa is *1-specificity*, and the ordinate is sensitivity.

F1 score: $precision = \frac{TP}{TP+FP}$; $recall = \frac{TP}{TP+FN}$;

$$F1 = \frac{2}{1/\text{precision}+1/\text{recall}}$$

The meaning of these characters is shown here: *TP* represented the actual number of positive samples predicted as positive samples, *TN* represented the actual number of negative samples predicted as negative samples, *FP* represented actually negative samples were predicted to be the number of positive samples, and *FN* represented the actual positive samples were predicted to be the number of negative samples.

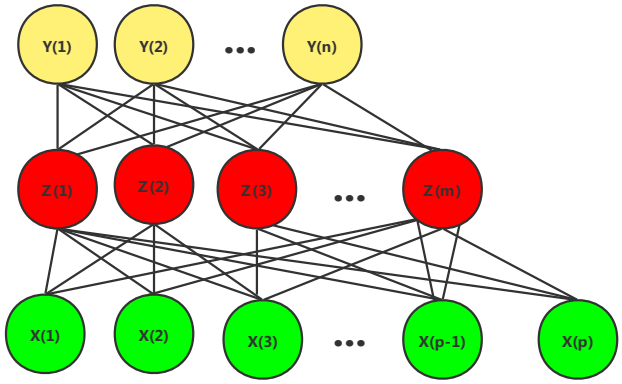

Fig. (3). The single hidden layer neural network diagram. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

2.3. Survival Prediction Process

By preprocessing the downloaded TCGA clinical data and omics data, the 490 HNSC samples shared by multi-omics were obtained. Likewise, DEGs were also obtained separately from each single-omic through preprocessing.

After the data preprocessing, the Lasso algorithm was used to select important variables for the survival outcome of HNSC from mRNA data, miRNA data, DNA methylation data, and CNV data. Random forest was used to calculate the ratio of each screened important variable. The random forest algorithm was used to calculate the ratio of each screened important variable. The four single omics were integrated, and machine learning models were used to predict HNSC survival outcomes. Likewise, using single-omic data as model input was also performed to predict survival outcomes. Among them, the 490 HNSC samples were randomly divided into 3 groups, of which 2/3 were used as the training set, and 1/3 were used as the test set. All mentioned models were operated 10 times.

The test results were measured and compared with performance indicators to find out which machine learning algorithms were effective and which omics were the most accurate for predicting HNSC survival. The flowchart for the main process of the study is presented in Fig. (**4**).

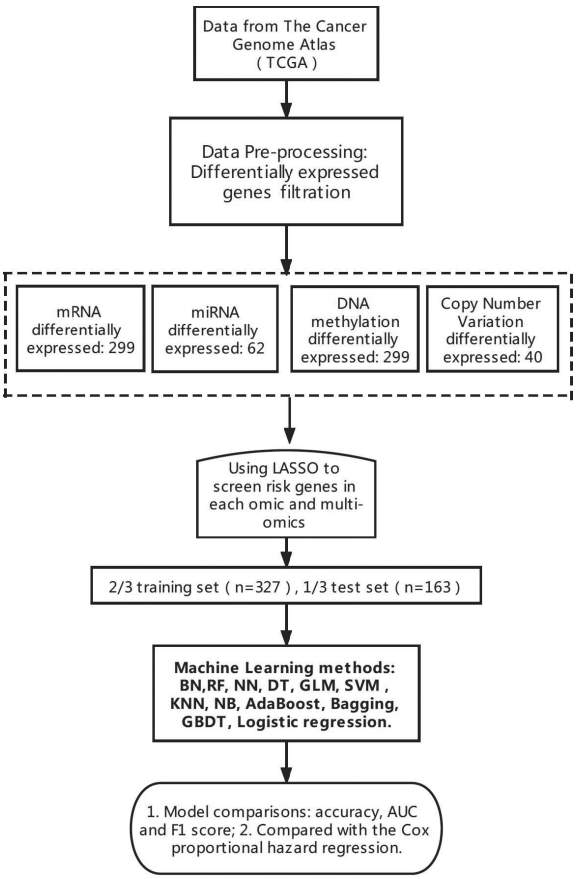

Fig. (4). The main process of the research.

2.4. *In vitro* Experimental

2.4.1. Cell Lines and Culture

A normal human immortalized keratinocytes (Hacat) cell line and three HNSC cell lines (Cal-27, SCC-9, and FaDu) were used in the present study. All cell lines were obtained from the Cell Bank of the Chinese Academy of Sciences. Hacat, Cal-27, and FaDu cell lines were cultured in Dulbecco's Modified Eagle Medium(DMEM), and SCC-9 was cultured in Dulbecco's Modified Eagle Medium/Nutrient Mixture F-12(DMEM/F12) in 5% CO_2 at 37°C. All media was supplemented with 10% fetal bovine serum (FBS) and 1% penicillin-streptomycin. All cell culture reagents were purchased from Gibco, Thermo Fisher Scientific company.

2.4.2. Quantitative Real-time PCR(qPCR) Assay

Cells were seeded at a density of 10^5 cells per well in a 6-well plate and cultured overnight. Total RNA was extracted from cultured cells using TRIzol reagent (Invitrogen). Complementary DNA (cDNA) was synthesized using Transcriptor First Strand cDNA Synthesis Kit (Roche), in accordance with the manufacturer's instructions. Quantitative reverse-transcription PCR was performed with Fast Start Essential DNA Green Master (Roche) and special primer sequences (Table **1**). Relative mRNA expression was quantified by the comparative Ct (ΔCt) method and normalized to the internal control gene, *ACTB*.

3. RESULTS

3.1. The Datasets Source and Data Pre-processing

The HNSC multi-omics data downloaded by TCGA included mRNA expression data, miRNA expression data, DNA methylation data, CNV data, and 528 clinical data containing clinical information. The multi-omics samples downloaded from TCGA were screened and compared, and 490 tumor samples that the multi-omics shared were obtained (Fig. **5**). We obtained 299 mRNA genes, 62 miRNA genes, 40 CNV genes, and 299 DNA methylation genes from data preprocessing. Fig. (**6**) shows the 40 top genes from the four single-omic data after DEGs.

3.2. Machine Learning Results

The parameter settings of the machine learning method used are shown in Table **2**.

Before utilizing the machine learning model to predict the survival outcome of HNSC, the LASSO algorithm was applied to select core genes in each single-omic. The LASSO algorithm took each single-omic data as the model input and took the event data of the clinical information as the model output. The results of each single-omics core gene obtained through the LASSO algorithm are shown in Table **3**. Core genes that LASSO selected were integrated.

Table 1. Primers sequences.

Gene	Primer Forward (5'→3')	Primer Reverse (5'→3')
AQP5	GCCACCTTGTCGGAATCTACT	CCTTTGATGATGGCCACACG
ACTN3	GCCCGATCGAGATGATGATGG	GGCAGTGAAGGTTTTCCGCT
TAC1	GGGACTGTCCGTCGCAAAAT	ACAGGGCCACTTGTTTTTCA
ZFR2	ATGGCTACCTACCAGGACAGT	GTATCCCGAGGACAAGGTGC
MMP11	GATCGACTTCGCCAGGTACT	CAGTGGGTAGCGAAAGGTGT
ACTB	TCACCATGGATGATGATATCGC	ATAGGAATCCTTCTGACCCATGC

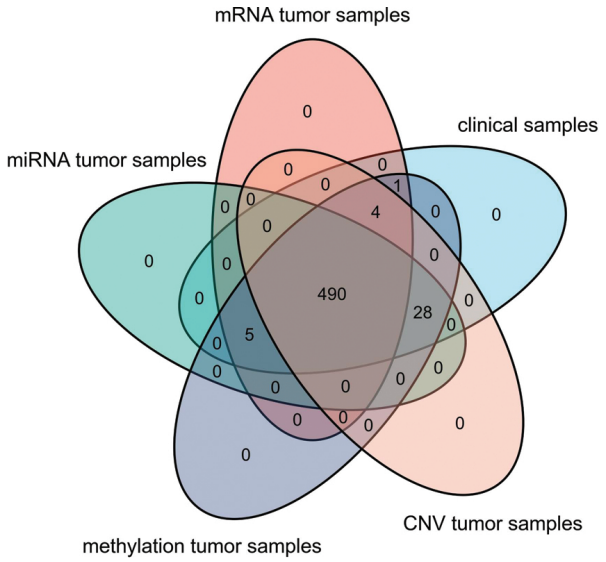

Fig. (5). The screened out of the shared tumor samples from TCGA. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

The data of each single-omic core gene selected from the LASSO algorithm were performed by the six machine learning models. The machine learning models took each single-omic core gene obtained from the LASSO algorithm as the model input and took the event data of the clinical information as the model output. Fig. (**7**) shows the results. A comparison of Fig. (**7**) reveals that among the HNSC omics model predictions, the prediction effect of mRNA was the best. The BN model (area under the curve [AUC] 0.7687, F1 score=0.7290) and RF model (area under the curve [AUC] 0.7307, F1 score=0.7327) had high predictive capacity, which was superior to that of other machine learning models. Besides, these 12 machine learning models showed the worst performance in survival outcome prediction in CNV data. In addition, it was worth noting that we found that GBDT made a similar predictive performance in every single omics data. In recent years, studies have shown that the GBDT model was rarely used for the prediction of genomic data. This could be most likely the reason for the low sensitivity of GBDT to genome-type data. Overall, the prediction performance of BN([AUC] 0.7687&F1 score=0. 7290 in mRNA, [AUC] 0.6341 & F1 score=0.6634 in miRNA, [AUC] 0.6325 & F1 score=0.6602 in methylation, [AUC] 0.5980 & F1 score=0.6411 in CNV) and RF([AUC] 0.7307 & F1 score=0.7327 in mRNA, [AUC] 0.6329 & F1 score=0.6298 in miRNA, [AUC] 0.6263 & F1 score=0.6701 in methylation, [AUC] 0.5253 & F1 score=0.6345 in CNV) in the four omics showed the best. Meanwhile, the predictive performance of miRNA data and methylation data were average, with AUC and F1 score less than 0.7. For the prediction performance results of CNV, it may have little effect on the survival outcome of HNSC.

The integrated core genes were selected again by the LASSO algorithm, and then the 36 multi-omics genes that affect the occurrence of HNSC could be obtained (Fig. **8A**). The 36 core genes that secondary selected were calculated to calculate the contribution of each gene (Fig. **8B**).

The 36 genes that were secondarily screened by LASSO were integrated. Meanwhile, the integrated multi-omics was used as the input of machine learning, and the event data of the clinical information were used as the output to predict the survival outcome of HNSC. Fig. (**9**) presents the results of machine learning methods performed in multi-omics. In Fig. (**7**), the multi-omics performed the best when compared to each single-omic. The F1 score of the 12 machine learning models in multi-omics was 0.7917 on BN, 0.7839 on RF, 0.6989 on NN, 0.6502on DT, 0.6061 on GLM, 0.7647 on SVM, 0.7653 on LR, 0.7405 on Bagging, 0.4394 on NB, 0.7196 on KNN, 0.7729 on Adaboost and 0.6797 on GBDT. The AUC of the 12 machine learning models in multi-omics were 0.8250 on BN, 0.8002on RF, 0.7207on NN, 0.5588 on DT, 0.6675 on GLM , 0.6826 on SVM, 0.7041 on LR, 0.6200 on Bagging, 0.5371 on NB, 0.6909 on KNN, 0.6910 on Adaboost and 0.7342 on GBDT. Except for the machine learning performance in DT and NB, the multi-omics data had the best forecast of HNSC survival outcome. Furthermore, BN and RF played the best predictive effect, whether it was the result of single-omic or multi-omics. Overall, these results suggested that the application of multi-omics data to predict the survival outcome of HNSC was better than the application of single-omic data alone. Likewise, the prediction performance of the BN model was better than other machine learning models as well. Together these results provide important insights that applying the LASSO algorithm to select the contributing variables and BN model to multi-omics data to predict the survival outcome may improve performance.

3.3. qPCR Results

Through the LASSO algorithm, we selected 5 genes to verify *in vitro*; we detected their mRNA expression levels in a normal Hacat cell line and three HNSC cell lines. Cells were cultured and detected *via* RT-qPCR. The mRNA expression levels of *AQP5*, *ACTN3*, *TAC1*, *ZFR2*, and *MMP11* were evaluated (Fig. **10**). We found that the mRNA expression of *MMP11* and *ZFR2* was significantly increased; meanwhile, the mRNA expression of the three genes *AQP5*, *ACTN3*, and *TAC1* was significantly decreased in the HNSC cell line. These findings were consistent with our model prediction of gene expression levels in HNSC.

4. DISCUSSION

Although HNSC's recent advances have brought substantial outcomes, it is still cancer with poor long-term survival due to the lack of specific therapeutic targets to predict its survival outcome [2]. Therefore, it is crucial to identify a robust method to predict the survival outcome of HNSC to evaluate its development and provide the survival estimation.

This study set out to assess the application of 12 supervised machine learning models in multi-omics integration data to predict the survival outcome in HNSC. The obvious finding to emerge from the analysis was that the multi-omics prediction shows good performance compared with each single-omic data prediction effect. The prediction accuracy of multi-omics integrated data was better than that of single-omics prediction alone in general. Furthermore, the

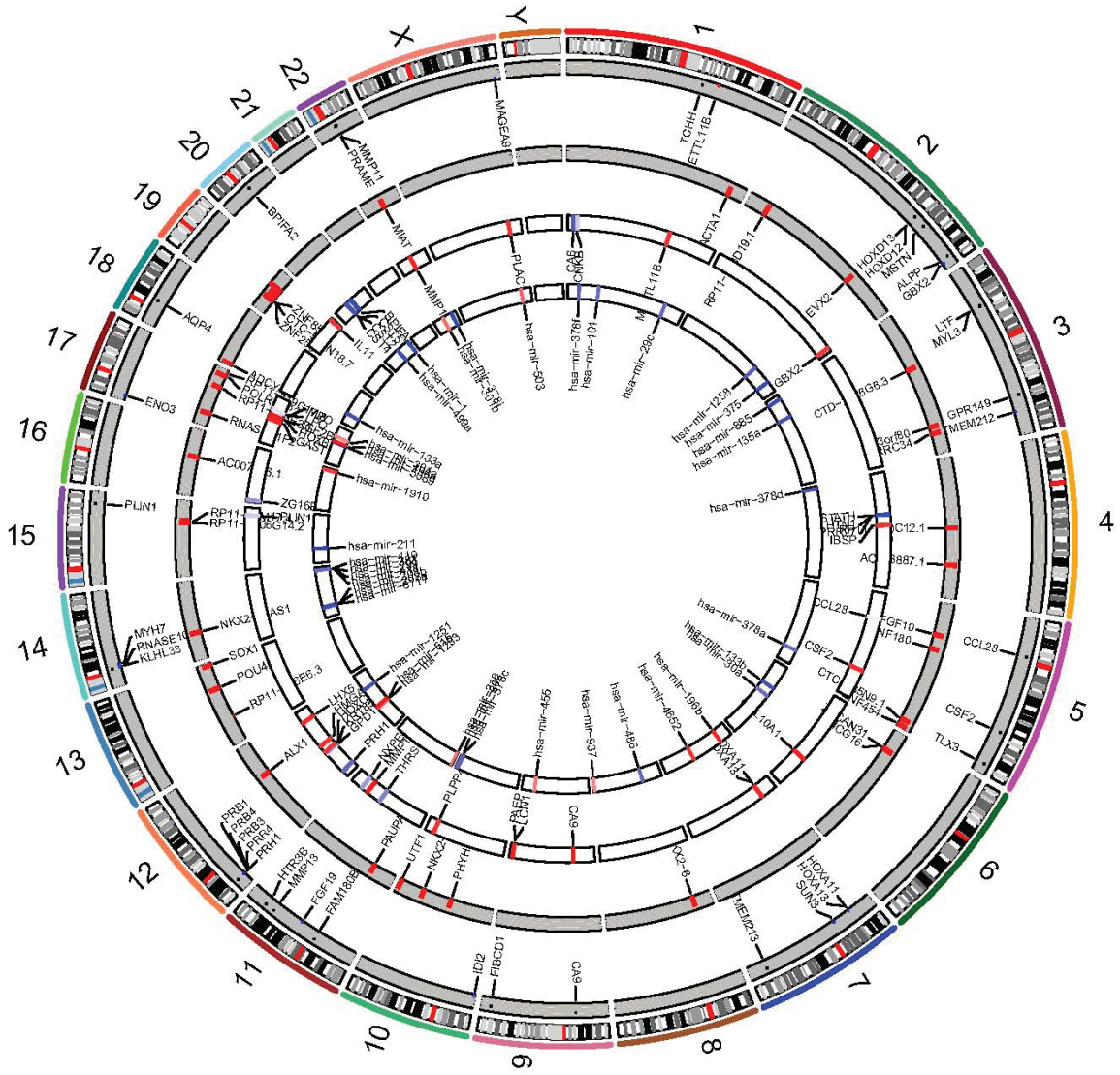

Fig. (6). The results of DEGs in TCGA. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

Table 2. Parameter in machine learning models.

Method	Parameter
Bayesian Networks, BN	Hill Climbing, maximum likelihood estimation method
Random Forest, RF	Select the number of trees corresponding to the smallest OOB error, mtry=sqrt(M), where M is the total number of features
Adaboost	mfinal = 100, control
Bootstrap aggregating, bagging	i in 1:20, mfinal = i
Naive Bayesian, NB	method="class", minsplit=1, cpfor pruning
K-Nearest Neighbor, KNN	distance = 1, kernel = "triangular"
Logistic Regression, LR	family = “binomia”, type = "response"
Gradient Boosting Decision Tree, GBDT	distribution = "bernoulli", n.trees = 1000,interaction.depth = 7,shrinkage = 0.01,cv.folds = 10
Neural Networks, NN	linout=F, size=10, decay=0.001, iteration ordinal number=1000,The hidden node n2 and the input node n1 in the three-layer NNET were related by n2=2n1+1
Generalized Linear Model, GLM	family = "binomial", eliminate variables with p<0.05
Decision Tree, DT	method=class, parms=default
Support Vector Machine, SVM	method="C-classification", kernel="radial", cost=10, gamma=0.1

Table 3. Single-omiccore genes selected by LASSO algorithm.

Data Type	Number of Genes	Gene Name
mRNA	21	*CST4,MSTN,ADH4,ACTN3,SLC13A2,TMEM210,MUC19,TAC1,METTL21C,AQP5,ADIPOQ,KCNJ16,CKM,DYNAP,GPRC6A,C14orf180,HOXB9,CGB5,ZFR2,PTX4,COX7B2*
miRNA	7	hsa-mir-378c,hsa-mir-411,hsa-mir-375,hsa-mir-499a,hsa-mir-503,hsa-mir-301a,hsa-mir-4776
Methylation	11	*RP11-266E6.3,AC007906.1,RP11-24M17.4,LRRC34,AC133644.2 RNU2-37P,TCF24,CTD-2540M10.1,SLITRK1,AC093787.1,HIST1H4A*
Copy number variation, CNV	4	*FGF19,MMP11,PLIN1,HOXD13*

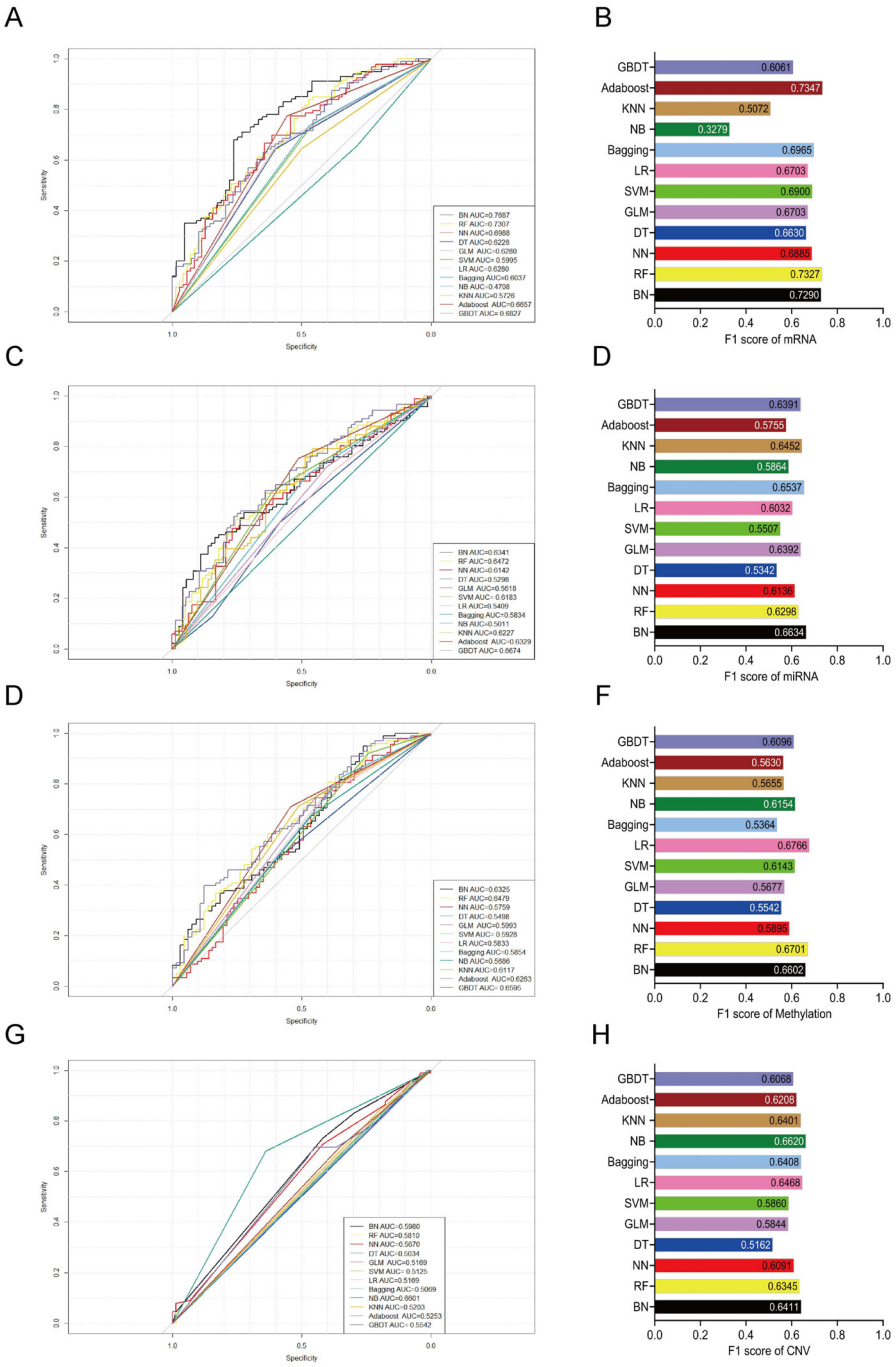

Fig. (7). Comparison of single-omic prediction results applied by the machine learning models. (A, C, E, F) showed the ROC curve. (B,D,F,H) showed the F1 score. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

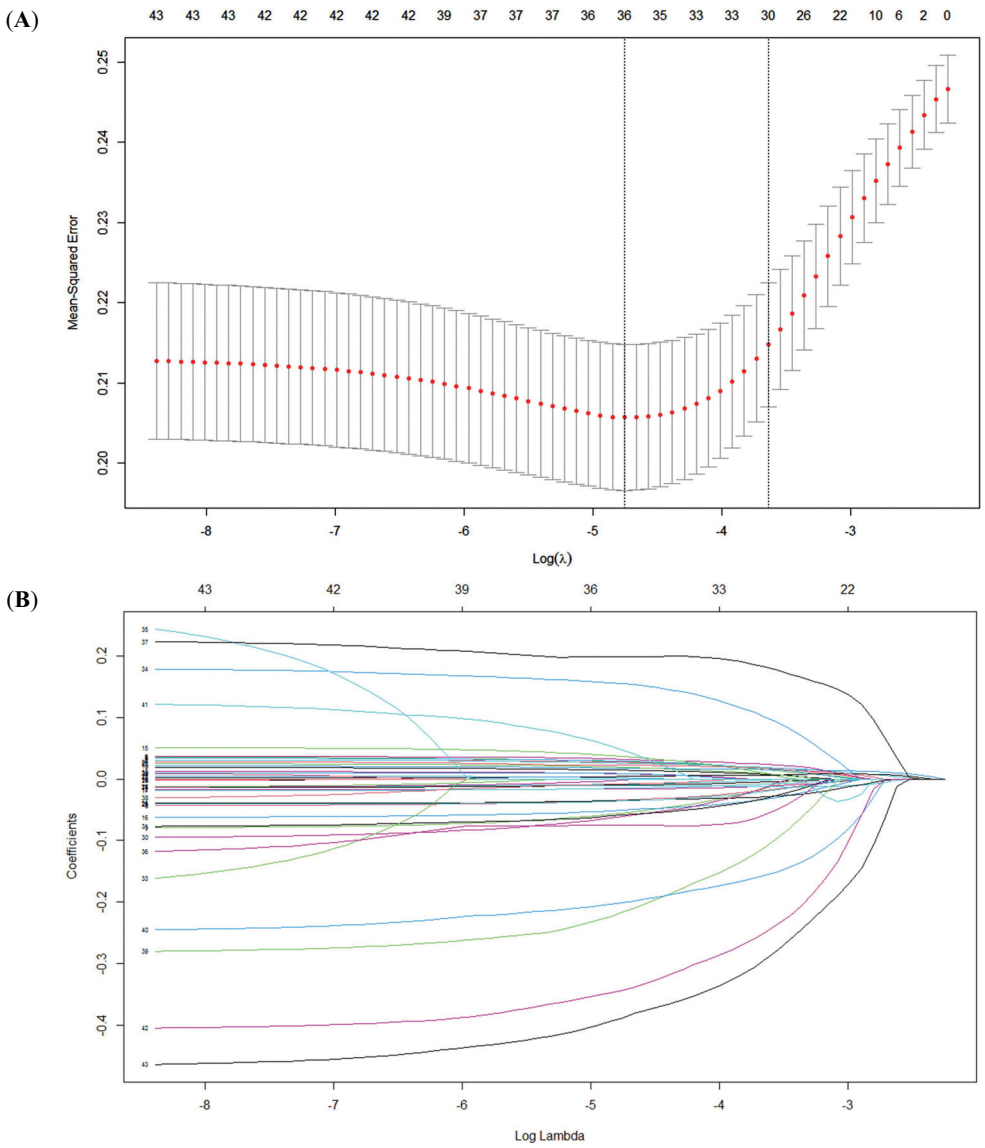

Fig. (8). The secondary selected results through LASSO. (**A**). To determine the penalty value at the lowest point of between lines determines. (**B**) Showed the contribution of variables. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

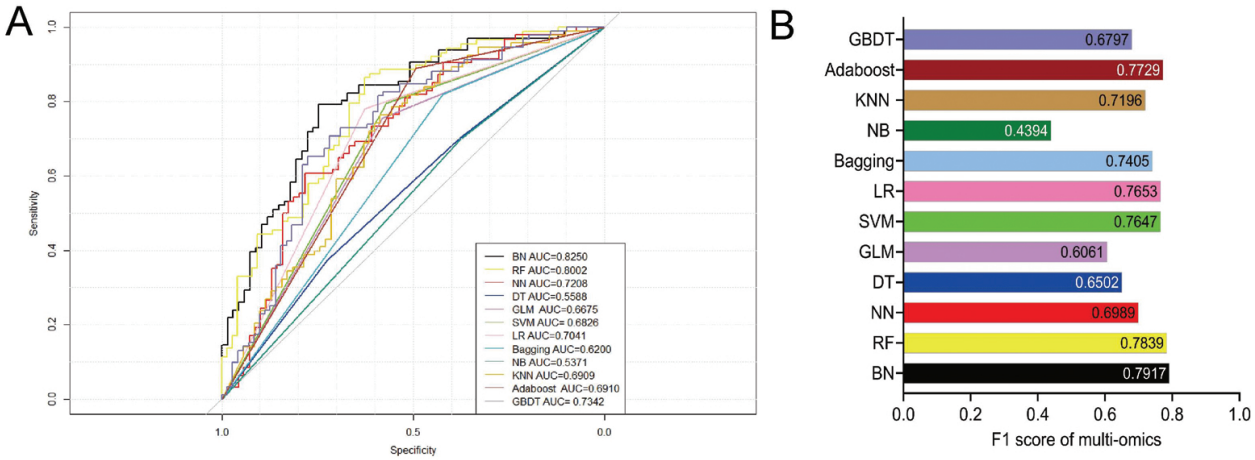

Fig. (9). The results of the machine learning models prediction in multi-omics. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

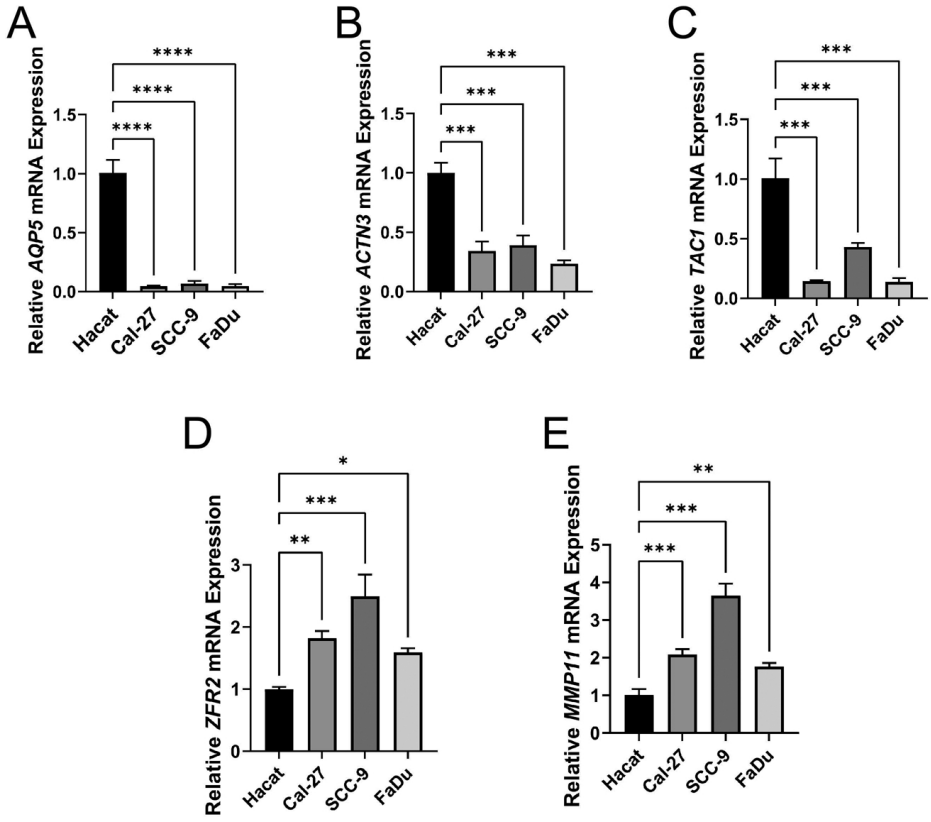

Fig. (10). The mRNA expression of 5genes. (A)*AQP5* expression;(B)*ACTN3*expression; (C)*TAC1* expression; (D)*ZFR2* expression; (E)*MMP11* expression. The results were presented as mean ± SEM. **p*< 0.05, ***p*<0.01,and ****p*< 0.001 as compared with the Hacat cell line group.

prediction effect of mRNA in six machine learning models was better than other single-omic, namely miRNA, methylation, and CNV. The result may partly be explained by the mRNAs playing a key role in the development of HNSC related pathways and protein expression. And the reason why the multi-omics integration data produced the best prediction performance is probably due to the multi-omics data integrating the molecular level information from each single-omic data in HNSC. The findings suggested that multi-omics data could more accurately reflect the relationship between molecular level and HNSC survival outcome than single-omics data. We can infer that integrating more data related to HNSC survival outcomes can get better prediction performance.

In addition, by comparing the performance metrics with other machine learning models, as shown in Fig. (**7**) and Fig. (**9**), the results found that the prediction efficiency of the BN model and the RF model was better than that of other machine learning models. Moreover, in all types of data, the BN model and the RF model played the best predictive effect among the 12 machine learning models. In the integrated multi-omics data, the AUC and the F1 score of the BN model were 0.8250 and 0.7917. Meanwhile, AUC and the F1 score of the RF model were 0.8002 and 0.7839. These results suggested that the BN model and the RF model could be suitable models for survival outcomes in HNSC. In terms of overall performance, the BN model and the RF model performed the best in predicting HNSC survival outcomes. The results indicated that the BN model and the RF model might be the most robust models to use in predicting overall survival from omics data. In summary, the BN model and the RF model were more suitable for HNSC survival prediction, especially in HNSC data that combined with multiple omics.

The current study found that the Cox proportional hazards regression was widely used to predict survival, and the prediction results were reliable [33, 34]. In general, therefore, it seems that the Cox proportional hazards regression could be used to compare the performance with the previously mentioned models to further confirm the predictive performance of BN in HNSC (Fig. **11**). Surprisingly, the 0.82 C-index value in the Cox proportional hazard regression indicated that the performance of the Cox proportional hazard regression was consistent with BN. The weight values that represented the influencing factors of multi-omics genes on the overall HNSC event in the last column of Fig. (**11**) were different from the results predicted by the LASSO algorithm in Fig. (**8**). For the inconsistent result, the importance of core genes was calculated again by RF (Fig. **12**). This finding was unexpected and suggested that the cox model and the BN model may have the same prediction of survival results, but the cox model and the LASSO regression model may have inconsistent results in the variable screening. This disappointing result might be explained by the fact that different model parameters cannot be ruled out. To further verify the results we found, cross-validation with other data sources and further experiments *in vitro* were performed.

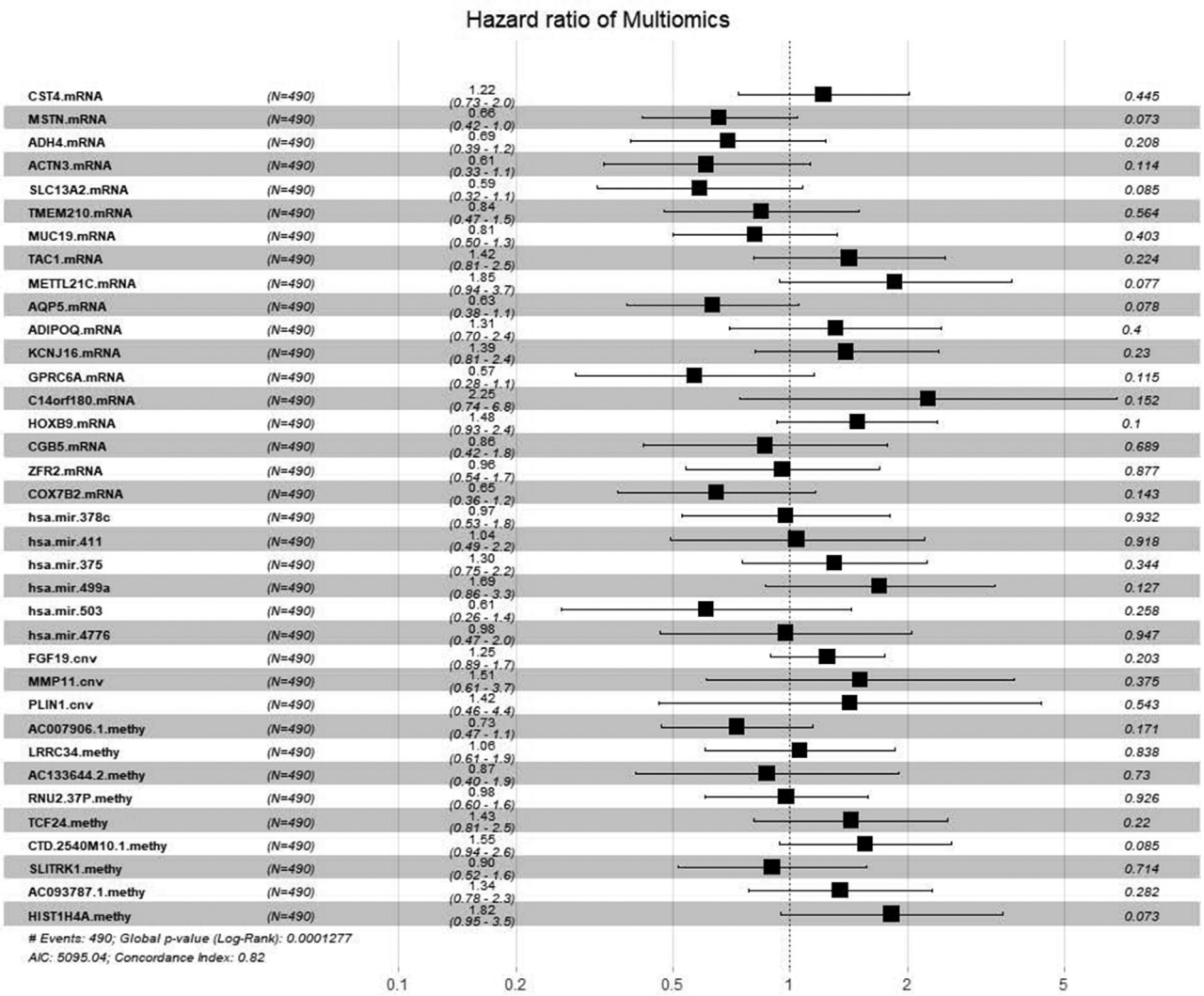

Fig. (11). The prediction results inCox proportional hazard regression.

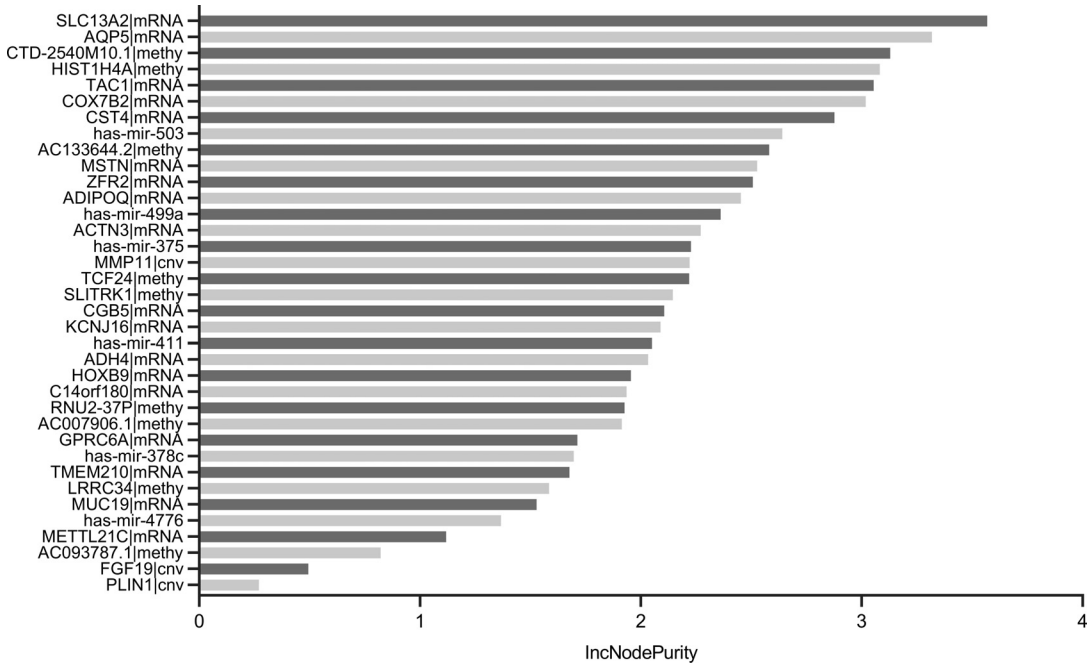

Fig. (12). The feature of each core gene was obtained through a Random forest.

Gene Ontology (GO) is a tool widely used to annotate the functional relationship between genes and gene products [35], which includes molecular functions (MF), biological pathways (BP), and cellular components (CC). To cross-validate the feature importance with the RF algorithm results, the GO functional enrichment was performed (Fig. **13**). The functional enrichment pointed that HNSC mainly regulated the cysteine-type endopeptidase inhibitor activity, cornification, negative regulation of proteolysis, antimicrobial humoral response, myosin complex, RNAi effector complex, respiratory chain complex, and mRNA binding in GO. Compared with the RF algorithm results, the selected genes from mRNA expression showed an important role in those pathways and revealed that the RF algorithm results made a good performance. In summary, the result of enrichment analysis was roughly similar to mentioned in the literature [36, 37], which demonstrated a partial overlap with the feature importance results obtained by the previous RF.

The results of *in vitro* verification of the core genes selected by the random forest suggested that these genes could serve as therapeutic targets and poor prognostic factors for HNSC. At the same time, in the Cox proportional hazard regression and BN models, these genes indicated that they have a greater impact on the survival outcome of HNSC. Importantly, prior studies have noted *TAC1*wasa powerful epigenetic biomarker in HNSC [38]. Meanwhile, *ZFR2*, *AQP5*, and *ACTN3* were found on the association between tumors such as cervical cancer, prostate cancer, acute myeloid leukemia, colorectal cancer, and breast cancer [39-43]. However, the *MMP11* gene has not been studied in HNSC. The result of this study may suggest that *MMP11*could serve as a novel biomarker for the diagnosis in HNSC, while the matrix metalloproteases (MMP) family were related to pan-cancer, especially with HNSC [44].

These findings may be somewhat limited firstly by the lack of predicting the combination of other personal factors such as smoking condition and alcohol condition with multi-omics data, and the lack of combined predictions of multi-omics data and data that fully describe the prognosis of cancer, such as TMN stage, radiotherapy, and chemotherapy. Moreover, at the molecular level, we neglected to combine proteomics data with multi-omics to predict the survival outcome of HNSC. This may be one of the reasons that the AUC that describes the prediction performance was lower than 0.8. Secondly, this research verified only part of the screened core genes by *in vitro* experiments. Despite these flaws, these results further support the idea of applying LASSO and BN combined models to multi-omics integration data to predict the survival outcome. Furthermore, these results show that the use of machine learning methods, especially BN and RF methods, is robust and accurate in predicting the survival outcome of HNSC.

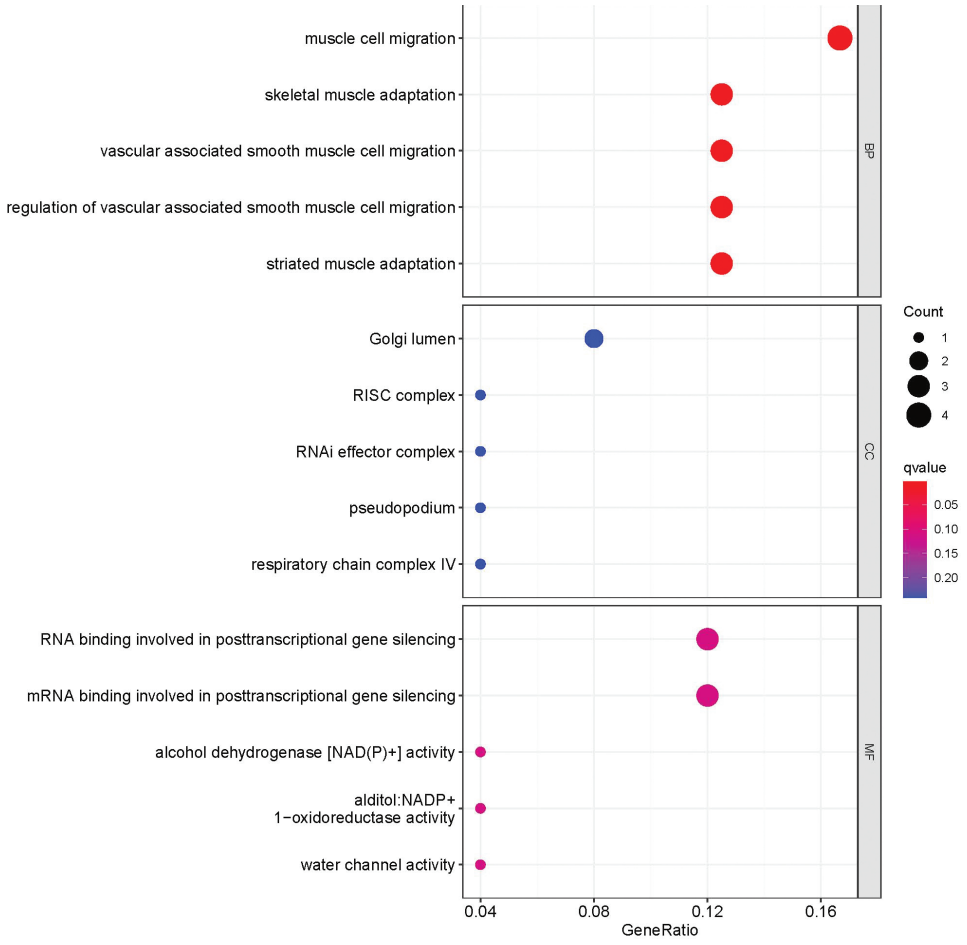

Fig. (13). The result of gene ontology functional enrichment. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

CONCLUSION

Applying multi-omics integration data to machine learning has important implications for predicting survival outcomes. One of the strengths of this study is that it is the multi-omics integration data machine learning analysis to date. It applied the LASSO and 12 machine learning models across four HNSC single-omics types, including mRNA, miRNA, methylation, and CNV, to predict the affecting HNSC variables and survival outcome. This study set out to explore whether the LASSO and 12 supervised machine learning models based on multi-omics integrated data could be robust in predicting the survival outcome of HNSC. Despite some limitations, the findings of this study are still valuable. The machine learning models, especially the BN model and the RF model, are expected to become a practical prediction model for tumor survival and prognosis. The multi-omics integration data could bring more information about the molecular level to better predict survival outcomes. Furthermore, better clinical services may bring new ideas to the precise prognosis and treatment of tumors.

AUTHORS' CONTRIBUTIONS

Daizheng Huang contributed to the study's conception and design. Data collection and analysis were performed by Liying Mo. Cell culture and qPCR experiments were performed by YuangangSu. The first draft of the manuscript was written by Liying Mo, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

No animals/humans were used for studies that are the basis of this research.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

All data generated or analyzed during this study are included in this article.

FUNDING

The authors would like to thank the BaGui scholar program of Guangxi Province, Guangxi Natural Science Foundation Innovation Research Team under Grant 2019 GXNSFGA245002, Guangxi Natural Science Foundation under Grant 2018GXNSFAA281133, and the National Natural Science Foundation of China under Grant 81860604.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer statistics, 2020. *CA Cancer J. Clin.,* **2020**, *70*(1), 7-30. http://dx.doi.org/10.3322/caac.21590 PMID: 31912902

[2] Chow, L.Q.M. Head and neck cancer. *N. Engl. J. Med.,* **2020**, *382*(1), 60-72. http://dx.doi.org/10.1056/NEJMra1715715 PMID: 31893516

[3] Beynon, R.A.; Lang, S.; Schimansky, S.; Penfold, C.M.; Waylen, A.; Thomas, S.J.; Pawlita, M.; Waterboer, T.; Martin, R.M.; May, M.; Ness, A.R. Tobacco smoking and alcohol drinking at diagnosis of head and neck cancer and all-cause mortality: Results from head and neck 5000, a prospective observational cohort of people with head and neck cancer. *Int. J. Cancer,* **2018**, *143*(5), 1114-1127. http://dx.doi.org/10.1002/ijc.31416 PMID: 29607493

[4] Torre, L.A.; Bray, F.; Siegel, R.L.; Ferlay, J.; Lortet-Tieulent, J.; Jemal, A. Global cancer statistics, 2012. *CA Cancer J. Clin.,* **2015**, *65*(2), 87-108. http://dx.doi.org/10.3322/caac.21262 PMID: 25651787

[5] Mourad, M.; Jetmore, T.; Jategaonkar, A.A.; Moubayed, S.; Moshier, E.; Urken, M.L. Epidemiological trends of head and neck cancer in the united states: A seer population study. *J. Oral Maxillofac. Surg.,* **2017**, *75*(12), 2562-2572. http://dx.doi.org/10.1016/j.joms.2017.05.008 PMID: 28618252

[6] Colevas, A.D.; Yom, S.S.; Pfister, D.G.; Spencer, S.; Adelstein, D.; Adkins, D.; Brizel, D.M.; Burtness, B.; Busse, P.M.; Caudell, J.J.; Cmelak, A.J.; Eisele, D.W.; Fenton, M.; Foote, R.L.; Gilbert, J.; Gillison, M.L.; Haddad, R.I.; Hicks, W.L., Jr; Hitchcock, Y.J.; Jimeno, A.; Leizman, D.; Maghami, E.; Mell, L.K.; Mittal, B.B.; Pinto, H.A.; Ridge, J.A.; Rocco, J.; Rodriguez, C.P.; Shah, J.P.; Weber, R.S.; Witek, M.; Worden, F.; Zhen, W.; Burns, J.L.; Darlow, S.D. NCCN guidelines insights: Head and neck cancers, version 1.2018. *J. Natl. Compr. Canc. Netw.,* **2018**, *16*(5), 479-490. http://dx.doi.org/10.6004/jnccn.2018.0026 PMID: 29752322

[7] Marur, S.; Forastiere, A.A. Head and neck squamous cell carcinoma: Update on epidemiology, diagnosis, and treatment. *Mayo Clin. Proc.,* **2016**, *91*(3), 386-396. http://dx.doi.org/10.1016/j.mayocp.2015.12.017 PMID: 26944243

[8] Chaudhary, K.; Poirion, O.B.; Lu, L.; Garmire, L.X. Deep learning-based multi-omics integration robustly predicts survival in liver cancer. *Clin. Cancer Res.,* **2018**, *24*(6), 1248-1259. http://dx.doi.org/10.1158/1078-0432.CCR-17-0853 PMID: 28982688

[9] Chen, L.; Lu, D.; Sun, K.; Xu, Y.; Hu, P.; Li, X.; Xu, F. Identification of biomarkers associated with diagnosis and prognosis of colorectal cancer patients based on integrated bioinformatics analysis. *Gene,* **2019**, *692*, 119-125. http://dx.doi.org/10.1016/j.gene.2019.01.001 PMID: 30654001

[10] Ren, N.; Liang, B.; Li, Y. Identification of prognosis-related genes in the tumor microenvironment of stomach adenocarcinoma by TCGA and GEO datasets. *Biosci. Rep.,* **2020**, *40*(10), BSR20200980. http://dx.doi.org/10.1042/BSR20200980 PMID: 33015704

[11] Ghafouri-Fard, S.; Gholipour, M.; Taheri, M.; Shirvani Farsani, Z. MicroRNA profile in the squamous cell carcinoma: Prognostic and diagnostic roles. *Heliyon,* **2020**, *6*(11), e05436. http://dx.doi.org/10.1016/j.heliyon.2020.e05436 PMID: 33204886

[12] Hu, F.; Zeng, W.; Liu, X. A gene signature of survival prediction for kidney renal cell carcinoma by multi-omic data analysis. *Int. J. Mol. Sci.,* **2019**, *20*(22), E5720. http://dx.doi.org/10.3390/ijms20225720 PMID: 31739630

[13] Vantaku, V.; Dong, J.; Ambati, C.R.; Perera, D.; Donepudi, S.R.; Amara, C.S.; Putluri, V.; Ravi, S.S.; Robertson, M.J.; Piyarathna, D.W.B.; Villanueva, M.; von Rundstedt, F.C.; Karanam, B.; Ballester, L.Y.; Terris, M.K.; Bollag, R.J.; Lerner, S.P.; Apolo, A.B.; Villanueva, H.; Lee, M.; Sikora, A.G.; Lotan, Y.; Sreekumar, A.; Coarfa, C.; Putluri, N. Multi-omics integration analysis robustly

predicts high-grade patient survival and identifies cpt1b effect on fatty acid metabolism in bladder cancer. *Clin. Cancer Res.,* **2019**, *25*(12), 3689-3701. http://dx.doi.org/10.1158/1078-0432.CCR-18-1515 PMID: 30846479

[14] Yin, Z.; Yan, X.; Wang, Q.; Deng, Z.; Tang, K.; Cao, Z.; Qiu, T. Detecting prognosis risk biomarkers for colon cancer through multi-omics-based prognostic analysis and target regulation simulation modeling. *Front. Genet.,* **2020**, *11*, 524. http://dx.doi.org/10.3389/fgene.2020.00524 PMID: 32528533

[15] Argelaguet, R.; Velten, B.; Arnol, D.; Dietrich, S.; Zenz, T.; Marioni, J.C.; Buettner, F.; Huber, W.; Stegle, O. Multi-omics factor analysis-a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.,* **2018**, *14*(6), e8124. http://dx.doi.org/10.15252/msb.20178124 PMID: 29925568

[16] Olivier, M.; Asmis, R.; Hawkins, G.A.; Howard, T.D.; Cox, L.A. The need for multi-omics biomarker signatures in precision medicine. *Int. J. Mol. Sci.,* **2019**, *20*(19), E4781. http://dx.doi.org/10.3390/ijms20194781 PMID: 31561483

[17] Chakraborty, S.; Hosen, M.I.; Ahmed, M.; Shekhar, H.U. Onco-Multi-OMICS approach: A new frontier in cancer research. *BioMed Res. Int.,* **2018**, *2018*, 9836256. http://dx.doi.org/10.1155/2018/9836256 PMID: 30402498

[18] Keyvanpour, M.R.; Shirzad, M.B. An analysis of QSAR research based on machine learning concepts. *Curr. Drug Discov. Technol.,* **2021**, *18*(1), 17-30. http://dx.doi.org/10.2174/1570163817666200316104404 PMID: 32178612

[19] Raza, A.; Bardhan, S.; Xu, L.; Yamijala, S.S.R.K.C.; Lian, C.; Kwon, H.; Wong, B.M. A machine learning approach for predicting defluorination of Per- and Polyfluoroalkyl Substances (PFAS) for their efficient treatment and removal. *Environ. Sci. Technol. Lett.,* **2019**, *6*(10), 624-629. http://dx.doi.org/10.1021/acs.estlett.9b00476

[20] Maurya, N.S.; Kushwaha, S.; Chawade, A.; Mani, A. Transcriptome profiling by combined machine learning and statistical R analysis identifies TMEM236 as a potential novel diagnostic biomarker for colorectal cancer. *Sci. Rep.,* **2021**, *11*(1), 14304. http://dx.doi.org/10.1038/s41598-021-92692-0 PMID: 34253750

[21] Guan, C.; Ouyang, D.; Qiao, Y.; Li, K.; Zheng, G.; Lao, X.; Zhang, S.; Liao, G.; Liang, Y. CA9 transcriptional expression determines prognosis and tumour grade in tongue squamous cell carcinoma patients. *J. Cell. Mol. Med.,* **2020**, *24*(10), 5832-5841. http://dx.doi.org/10.1111/jcmm.15252 PMID: 32299152

[22] Chai, H.; Zhou, X.; Zhang, Z.; Rao, J.; Zhao, H.; Yang, Y. Integrating multi-omics data through deep learning for accurate cancer prognosis prediction. *Comput. Biol. Med.,* **2021**, *134*, 104481. http://dx.doi.org/10.1016/j.compbiomed.2021.104481 PMID: 33989895

[23] Fujino, Y.; Murata, H.; Mayama, C.; Asaoka, R. Applying "Lasso" regression to predict future visual field progression in glaucoma patients. *Invest. Ophthalmol. Vis. Sci.,* **2015**, *56*(4), 2334-2339. http://dx.doi.org/10.1167/iovs.15-16445 PMID: 25698708

[24] Huang, E.W.; Bhope, A.; Lim, J.; Sinha, S.; Emad, A. Tissue-guided LASSO for prediction of clinical drug response using preclinical samples. *PLOS Comput. Biol.,* **2020**, *16*(1), e1007607. http://dx.doi.org/10.1371/journal.pcbi.1007607 PMID: 31967990

[25] Nandra, R.; Parry, M.; Forsberg, J.; Grimer, R. Can a bayesian belief network be used to estimate 1-year survival in patients with bone sarcomas? *Clin. Orthop. Relat. Res.,* **2017**, *475*(6), 1681-1689. http://dx.doi.org/10.1007/s11999-017-5346-1 PMID: 28397168

[26] Burghardt, T.P.; Ajtai, K. Neural/Bayes network predictor for inheritable cardiac disease pathogenicity and phenotype. *J. Mol. Cell. Cardiol.,* **2018**, *119*, 19-27. http://dx.doi.org/10.1016/j.yjmcc.2018.04.006 PMID: 29654880

[27] Bellot, P.; de Los Campos, G.; Pérez-Enciso, M. Can deep learning improve genomic prediction of complex human traits? *Genetics,* **2018**, *210*(3), 809-819. http://dx.doi.org/10.1534/genetics.118.301298 PMID: 30171033

[28] McEligot, A.J.; Poynor, V.; Sharma, R.; Panangadan, A. Logistic LASSO regression for dietary intakes and breast cancer. *Nutrients,* **2020**, *12*(9), E2652. http://dx.doi.org/10.3390/nu12092652 PMID: 32878103

[29] Ying, Q.; Chen, C.; Xiaoyi, L. Multi-feature fusion combined with machine learning algorithms to quickly screen uveitis. *J. Xinjiang Univ.,* **2021**, *38*(4), 439-449. http://dx.doi.org/10.13568/j.cnki.651094.651316

[30] Stojadinovic, A.; Bilchik, A.; Smith, D.; Eberhardt, J.S.; Ward, E.B.; Nissan, A.; Johnson, E.K.; Protic, M.; Peoples, G.E.; Avital, I.; Steele, S.R. Clinical decision support and individualized prediction of survival in colon cancer: Bayesian belief network model. *Ann. Surg. Oncol.,* **2013**, *20*(1), 161-174. http://dx.doi.org/10.1245/s10434-012-2555-4 PMID: 22899001

[31] Stojadinovic, A.; Eberhardt, C.; Henry, L.; Eberhardt, J.; Elster, E.A.; Peoples, G.E.; Nissan, A.; Shriver, C.D. Development of a Bayesian classifier for breast cancer risk stratification: A feasibility study. *Eplasty,* **2010**, *10*, e25.

[32] Wang, Q.; Chen, R.; Cheng, F.; Wei, Q.; Ji, Y.; Yang, H.; Zhong, X.; Tao, R.; Wen, Z.; Sutcliffe, J.S.; Liu, C.; Cook, E.H.; Cox, N.J.; Li, B. A Bayesian framework that integrates multi-omics data and gene networks predicts risk genes from schizophrenia GWAS data. *Nat. Neurosci.,* **2019**, *22*(5), 691-699. http://dx.doi.org/10.1038/s41593-019-0382-7 PMID: 30988527

[33] Matsuo, K. Survival outcome prediction in cervical cancer: Cox models *vs.* deep-learning model. *Am. J. Obstet. Gynecol.,* **2019**, *220*(4), 381. http://dx.doi.org/10.1016/j.ajog.2018.12.030

[34] Shen, Y.; Peng, X.; Shen, C. Identification and validation of immune-related lncRNA prognostic signature for breast cancer. *Genomics,* **2020**, *112*(3), 2640-2646. http://dx.doi.org/10.1016/j.ygeno.2020.02.015 PMID: 32087243

[35] The Gene Ontology Consortium. Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Res.,* **2017**, *45*(D1), D331-D338. http://dx.doi.org/10.1093/nar/gkw1108 PMID: 27899567

[36] Jin, Y.; Yang, Y. Identification and analysis of genes associated with head and neck squamous cell carcinoma by integrated bioinformatics methods. *Mol. Genet. Genomic Med.,* **2019**, *7*(8), e857. http://dx.doi.org/10.1002/mgg3.857 PMID: 31304688

[37] Wang, M.; Zhong, B.; Li, M.; Wang, Y.; Yang, H.; Du, K. Identification of potential core genes and pathways predicting pathogenesis in head and neck squamous cell carcinoma. *Biosci. Rep.,* **2021**, *41*(5), BSR20204148. http://dx.doi.org/10.1042/BSR20204148 PMID: 33982750

[38] Misawa, K.; Mima, M.; Imai, A.; Mochizuki, D.; Misawa, Y.; Endo, S.; Ishikawa, R.; Kanazawa, T.; Mineta, H. The neuropeptide genes *SST*, *TAC1*, *HCRT*, *NPY*, and *GAL* are powerful epigenetic biomarkers in head and neck cancer: A site-specific analysis. *Clin. Epigenetics,* **2018**, *10*(1), 52. http://dx.doi.org/10.1186/s13148-018-0485-0 PMID: 29682090

[39] Direito, I.; Madeira, A.; Brito, M.A.; Soveral, G. Aquaporin-5: From structure to function and dysfunction in cancer. *Cell. Mol. Life Sci.,* **2016**, *73*(8), 1623-1640. http://dx.doi.org/10.1007/s00018-016-2142-0 PMID: 26837927

[40] Rubicz, R.; Zhao, S.; Geybels, M.; Wright, J.L.; Kolb, S.; Klotzle, B.; Bibikova, M.; Troyer, D.; Lance, R.; Ostrander, E.A.; Feng, Z.; Fan, J.B.; Stanford, J.L. DNA methylation profiles in African American prostate cancer patients in relation to disease progression. *Genomics,* **2019**, *111*(1), 10-16. http://dx.doi.org/10.1016/j.ygeno.2016.02.004 PMID: 26902887

[41] Yang, X.; Pang, Y.; Zhang, J.; Shi, J.; Zhang, X.; Zhang, G.; Yang, S.; Wang, J.; Hu, K.; Wang, J.; Jing, H.; Ke, X.; Fu, L. High expression levels of *ACTN1* and *ACTN3* indicate unfavorable prognosis in acute myeloid leukemia. *J. Cancer,* **2019**, *10*(18), 4286-4292. http://dx.doi.org/10.7150/jca.31766 PMID: 31413748

[42] Zhang, L.; Jiang, Y.; Lu, X.; Zhao, H.; Chen, C.; Wang, Y.; Hu, W.; Zhu, Y.; Yan, H.; Yan, F. Genomic characterization of cervical cancer based on human papillomavirus status. *Gynecol. Oncol.,* **2019**, *152*(3), 629-637. http://dx.doi.org/10.1016/j.ygyno.2018.12.017 PMID: 30581036

[43] Zhu, Z.; Jiao, L.; Li, T.; Wang, H.; Wei, W.; Qian, H. Expression of AQP3 and AQP5 as a prognostic marker in triple-negative breast cancer. *Oncol. Lett.,* **2018**, *16*(2), 2661-2667. http://dx.doi.org/10.3892/ol.2018.8955 PMID: 30013662

[44] Gobin, E.; Bagwell, K.; Wagner, J.; Mysona, D.; Sandirasegarane, S.; Smith, N.; Bai, S.; Sharma, A.; Schleifer, R.; She, J.X. A pan-cancer perspective of Matrix Metalloproteases (MMP) gene expression profile and their diagnostic/prognostic potential. *BMC Cancer,* **2019**, *19*(1), 581. http://dx.doi.org/10.1186/s12885-019-5768-0 PMID: 31200666