Hindawi
Scientific Programming
Volume 2021, Article ID 4907687, 12 pages
https://doi.org/10.1155/2021/4907687

Research Article

Application of Data Mining in WITMED: Identification of Prognostic Genes in Oral Cancer

Gang Liu,[1] **Wei Wang**,[1] **He Qin**,[1] **Qingguo Zhou**,[1] **Jianbing Ma**,[2] **Xiaokang Zhou**,[3,4] **and Haiyan Zhen**[5]

[1]*School of Information Science & Engineering, LanZhou University, LanZhou, China*
[2]*School of Computer Science, ChengDu University of Information Technology, ChengDu, China*
[3]*The Faculty of Data Science, Shiga University, Tokyo, Japan*
[4]*RIKEN Center for Advanced Intelligence Project, Tokyo, Japan*
[5]*The First Hospital of Lanzhou University, LanZhou, China*

Correspondence should be addressed to Gang Liu; andyliu@lzu.edu.cn and Haiyan Zhen; 1041101942@qq.com

Received 26 August 2021; Accepted 23 November 2021; Published 18 December 2021

Academic Editor: Zhu Xiao

In recent years, the booming development of big data, cloud computing, Internet of Things, and other technologies provides conditions for the popularization and application of smart city. The combination of big data and medical information produces the emerging field of WITMED (Wise Information Technology of Med). WITMED is essential for the prospering growth of smart cities, which assumed a high quality of medical service is the most challenging goal for the city government. In this paper, the main attention is paid to the method of targeted gene therapy, which provides a new method for the treatment of oral cancer in inhibiting the growth, differentiation, invasion, and metastasis of oral cancer cells; therefore, the physical and psychological adverse effects of surgery and chemotherapy on patients are reduced and the survival and prognosis of patients are improved. Targeted gene therapy methods need to select the appropriate gene; that is, data mining methods are used to analyze a large number of complex genetic data from smart cities to obtain appropriate genetic markers, which makes the effect of targeted gene therapy better, and also provide some reference for the research of oral cancer gene direction and provide some basis for clinical treatment.

1. Introduction

Smart city is the use of advanced information technology to realize the intelligent management and operation of the city and then create a better life for the people in the city and promote the harmonious and sustainable growth of the city. WITMED is an important part of smart city. Through information technology, medical infrastructure is integrated with IT infrastructure. WITMED takes medical cloud data center as the core, crosses the spatial and temporal limitations of the original medical systems, and makes intelligent decisions on this basis to realize the medical system with optimized medical services. For example , through machine learning and other technologies, precise treatment is realized to help improve the efficiency of diagnosis and treatment of doctors and improve the quality of medical services [1].

Oral cancer is one of the most common cancers in the world, and a major health problem all over the world, with higher morbidity and mortality. Oral squamous cell carcinoma is a common malignant tumor in the head and neck, accounting for more than 90% of oral cancer cases. There are more than 300000 new cases in the world every year. In recent years, statistics of oral squamous cell carcinoma show that the incidence of oral squamous cell carcinoma is increasing [2].

In 2012, there were more than 440,000 new cases of oral and oropharyngeal cancer in the world, and more than 240,000 oral cancer and oropharyngeal cancer deaths, respectively, accounting for 3.1% and 3.0% of the new and dead cancer cases around the world [3]. According to GLOBOCAN estimates, the incidence of oral cancer and oropharyngeal cancer is the highest in Melanesia, followed

by Central and South Asia and Western Europe. The incidence rate was more than 10/100,000 every year in these two regions. Annual incidence rates are the lowest in East Asia and West Africa, about 2/100,000. Asia is one of the most serious regions of oral cancer and oropharyngeal cancer. In South Asia, the incidence and mortality rates of oral cancer and oropharyngeal cancer is the highest in Bangladesh. It can be seen that nearly two-thirds of oral cancer and oropharyngeal cancer cases live in underdeveloped countries. According to statistics, from 2005 to 2013, there were more than 280,000 new cases and more than 130,000 deaths in China, which were related to oral cancer and oropharyngeal cancer [2]. In the next 20 years, the incidence rate of oral cancer will increase from 2.26/100,000 to 3.21/100,000 people in the world [4, 5].

With the development of digital technology and smart cities, a great amount of data are produced in our real world, including daily life data, academic data produced in schools, and scientific experiment data produced in experiments [6]. For so many data, how to pass data to find useful information in it, and benefit to construction of smart cities, is a hot topic in today's technology research; it also promotes the rapid development of machine learning. For traditional machine learning, data dimensionality reduction is mainly used to learn low-dimensional feature representations from high-dimensional data [7]. For deep learning, images are mainly applied, including target detection and anomaly detection [8, 9].

Big data storage and processing platform is used to extensively collect and deeply utilize data in WITMED, with patient data as the core, and the medical historical data is modeled and analyzed by using data mining, to achieve the purpose of detecting early diseases and predicting health risks, at the same time, for medical staff to provide reference for diagnosis and treatment.

Predictive analytics rely on historical data and utilize advanced statistical or machine learning techniques to simulate the behavior or pattern so that it is possible to predict the likelihood of possible future trends or patterns in data. To sum up, it predicts what will happen in the future by learning the relevance of historical patterns and available data. Predictive analytics have been widely used for different applications including predictive maintenance, prediction of price, supply-demand trend, or prediction of likelihood of any outcome. State-of-the-art predictive modeling techniques include model based on statistical regression, Decision Trees, and Neural Network or Deep Neural Network-based models [10].

With the development of molecular biology, people have a more in-depth understanding of genetic sequencing and genetic markers; if scientists can analyze and research oral cancer from the perspective of molecular biology through massive genetic data, the corresponding research results will be helpful for the treatment of early diagnosis and prognosis of oral cancer to facilitate patients' medical decision-making process [11–13]. Since human genetic data is high-dimensional, the first problem to be solved for data mining of genes is dimension reduction. In this paper, Cox univariate regression analysis and Least Absolute Shrinkage and Selection Operator (LASSO) regression analysis were used to conduct data mining analysis on the genetic expression of oral cancer patients, and risk genes that have an impact on the prognosis of oral cancer are screened out. Second, the obtained risk genes were used to build the prognostic model and verify whether the risk genes screened by the two methods have reference value for the prognosis of oral cancer through survival analysis and ROC curve (AUC value). Third, the 23 genes closely related to the prognosis of oral cancer were obtained by the LASSO method, and the validation set was used to verify the reference value of the obtained genes for the prognosis of oral cancer and combined with the independent external data set to do double-blind verification for the screened genes. Finally, STRING genetic function analysis and literature review were used to further verify that the genes screened by us are closely related to the prognosis of oral cancer, which can provide a certain reference and theoretical basis for the future clinical research, treatment, diagnosis, and prognosis of oral cancer based on molecular biology.

The rest of this paper is organized as follows. In Section 2, we describe the system model and system architecture of prognostic analysis of oral cancer based on LASSO algorithm. In Section 3, we evaluate our model through validation set. We perform a functional analysis of the genes selected by the model in Section 4 and conclude in Section 5.

2. Data Analysis Process

In this section, we describe the establishment and data analysis process of prognostic model, including data collection and processing, differential expression of genes, Cox regression analysis, and LASSO regression analysis. The framework of the model is shown in Figure 1.

2.1. Data Collection and Processing. In this paper, the samples of head and neck cancer were downloaded respectively in three open source websites of the Xena Functional Genomics Explorer (xenabrowser, xenabrowser.net/datapages/), cBioportal (http://www.cbioportal.org/), and National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov/), a total of 566 standardized TCGA head and neck cancer samples are gotten from xenabrowser, and 485 oral cancer samples were sorted according to requirements. 514 head and neck cancer samples were downloaded from cBioportal and 385 oral cancer samples were sorted. 103 oral cancer samples were downloaded from NCBI and 74 samples containing tumor tissues were sorted out. The data obtained on xenabrowser is used for the establishment of data mining model and the data obtained on cBioportal and NCBI are used as independent data sets for result verification.

The raw data obtained from the open-source databases on the xenabrowser, cBioportal, and NCBI websites are mainly structured data.

The original data were processed as follows. (1) Data cleaning: in order to preserve the authenticity of the data, we chose to delete the data without many features and fill other

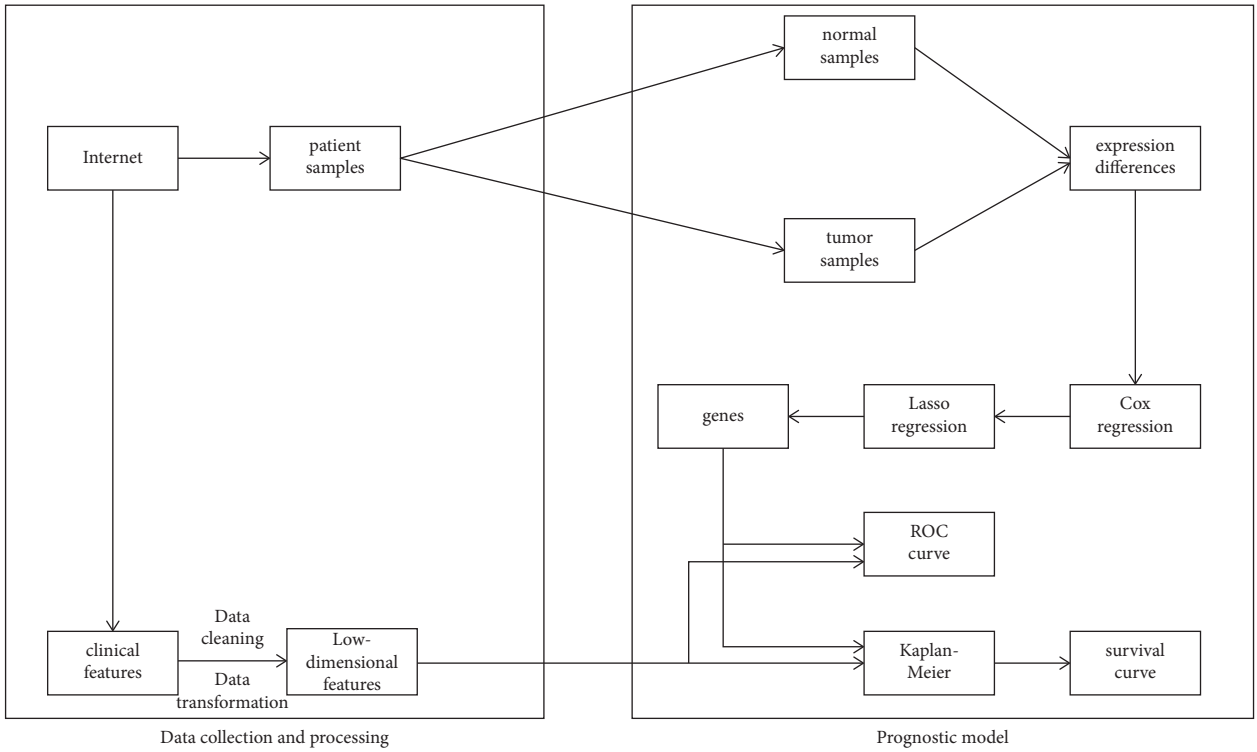

Figure 1: The mode framework for identification of prognostic genes in oral cancer based on data mining.

data with few missing features by the revised average method. Through data cleaning, 141-dimensional clinical data were organized into 80-dimensional data. (2) Data transformation: compared with the original 141-dimensional clinical feature data, 32-dimensional clinical features obtained by data transformation and discretization had improved the degree of analysis.

2.2. Model Building. First, we need to transform the clinical information data and gene data of these samples into data frames for data mining analysis. Second, data preprocessing was performed; that is, the patient samples with missing survival attribute values were removed and 453 patient samples were retained, and then the tumor-containing tissue and the normal solid tissue were separated; meanwhile, 412 patient samples with the tumor tissue and 41 normal solid tissue samples were obtained. Finally, the 412 patient samples with tumor tissues were randomly divided into training samples and validation samples at a ratio of 1 : 1. After data preprocessing, the data was divided into 3 groups: normal samples, training samples, and validation samples.

2.3. Gene Expression Differences. Data analysis of gene differential expression is to screen genes with differences in genetic expression, and those with "significant differences in expression" are screened out. In order to analyze the differences in genetic expression between the 200 training samples and 41 normal samples, the "Limma" package version 3.42.2 (39) in *R* was used. According to the adjusted *P* value (adj. *P*. val is less than 0.001), 2146 genes were considered to be differentially expressed genes. Some of the selected genes are shown in Table 1.

2.4. Cox Regression Analysis. After the differential expression analysis of genes, 2146 differentially expressed genes were obtained. Then, these 2146 genes were combined with survival data from the clinical characteristics of 200 training samples, and Cox regression was used to analyze the degree of correlation between each differentially expressed gene and the survival of oral cancer patients [14]. The Wald-Test *P* value and hazard ratio (HR) were calculated by each gene to filter genes which are significantly associated with the survival of oral cancer patients. According to statistical principles, threshold is set to $P < 0.05$ and 314 genes highly related to the survival of oral cancer patients were filtered. Some of the selected genes and related parameters are shown in Table 2.

2.5. LASSO Regression Algorithm. In order to achieve more accurate genetic screening, LASSO regression method was used to reduce dimension and regression analysis for gene data of training samples [15].

Genes screening based on LASSO regression: LASSO regression is a linear model with penalty term where L1 norm is the absolute value, and K-fold cross-validation was used to select the penalty parameter λ, and the value of K was 10. Through 10-fold cross-validation, the appropriate penalty parameter λ was determined [14, 16]. The process of parameter λ selection is shown in Figure 2. After the 10-fold

Table 1: Some significant differentially expressed genes screened out by gene differential expression analysis.

Symbol	logFC	AveExpr	t	P. value	adj. P. val	B
CAB39L	2.268	7.155	18.160	$2.400E-52$	$2.500E-48$	108.300
GLT25D1	−1.428	11.830	−17.970	$1.400E-51$	$9.500E-48$	106.500
GPRIN1	−2.329	8.533	−16.110	$4.700E-44$	$8.800E-41$	89.380
COL4A1	−2.922	12.660	−15.880	$4.100E-43$	$6.500E-40$	87.240
MYBL2	−2.335	10.560	−15.780	$1.000E-42$	$1.500E-39$	86.350
CDCA5	−2.090	9.803	−15.740	$1.500E-42$	$2.000E-39$	85.990
HOXA10	−3.953	6.401	−15.620	$4.600E-42$	$5.800E-39$	84.860
NETO2	−2.080	9.347	−15.470	$1.800E-41$	$1.900E-38$	83.510
GPD1L	2.498	8.613	15.450	$2.100E-41$	$2.100E-38$	83.330
FOXM1	−2.243	10.710	−15.290	$9.200E-41$	$8.200E-38$	81.890
AGFG2	1.969	8.630	15.190	$2.300E-40$	$2.000E-37$	80.990
CENPA	−2.186	7.679	−15.070	$6.800E-40$	$5.400E-37$	79.910
KIF2C	−2.087	9.666	−15.030	$9.700E-40$	$7.400E-37$	79.570
EME1	−1.985	6.424	−14.870	$4.300E-39$	$3.100E-36$	78.100
LOXL2	−3.116	10.050	−14.860	$4.900E-39$	$3.500E-36$	77.960
COL4A2	−2.610	12.920	−14.820	$6.400E-39$	$4.400E-36$	77.700
IL11	−4.011	6.292	−14.820	$6.900E-39$	$4.500E-36$	77.630
TPX2	−2.096	11.010	−14.780	$1.000E-38$	$6.300E-36$	77.240
KIF14	−2.356	8.229	−14.640	$3.600E-38$	$1.900E-35$	76.000

Table 2: Some genes that are significantly related to the survival of oral cancer patients screened by Cox regression.

Symbol	P	hr	coef	Low	Up
STC2	0.003	1.239	0.215	1.076	1.428
CLEC3B	0.006	0.806	−0.216	0.690	0.941
COL7A1	0.038	0.825	−0.192	0.688	0.989
CGNL1	0.031	0.891	−0.116	0.802	0.990
HAPLN1	0.047	0.884	−0.123	0.783	0.998
TPBG	0.045	1.318	0.276	1.006	1.726
RORC	0.049	0.902	−0.103	0.814	1.000
HMGA2	0.014	1.111	0.105	1.021	1.208
SELENBP1	0.030	0.874	−0.134	0.774	0.988
CENPI	0.042	1.273	0.242	1.009	1.608

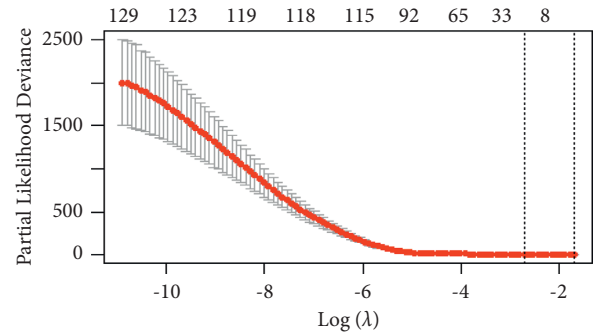

Figure 2: The selection process of 10-fold cross-validation penalty parameter λ.

cross-validation, λ is equal to 0.04098355; then λ was substituted into the LASSO regression equation, and 23 genes with the highest survival rate of oral cancer patients were finally obtained. The results are shown in Table 3.

Establishment of the prognostic index based on LASSO regression: as an important indicator of the integration of risk genes, a PI value can be determined for each patient with oral cancer. The PI was obtained by linearly fitting the product of the expression and the coefficient corrected by LASSO of each gene [14]. The formula of the prognosis index is shown as follows:

$$\mathrm{PI} = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n. \quad (1)$$

In the formula, Xi is the expression of the i-th gene, and βi is the regression coefficient of the i-th gene. Through linear fitting of the product of expression and regression coefficient of the 23 genes in each sample, the PI of each patient was calculated, and the patients were sorted from lowest to highest according to their PI value. Based on the median PI value, the patients were divided into high-risk and low-risk groups.

Based on LASSO regression analysis, 23 genes were selected finally. These 23 genes and the corresponding LASSO regression coefficients were used to construct a multivariate linear model. Because the expression of each gene in different samples was different, a prognostic index can be generated for each sample. The distribution of PI value is shown in Figure 3.

Next, the prognostic model constructed by the LASSO regression method was tested on the training samples to observe whether the samples of high-risk patients could be distinguished from the samples of low-risk patients. The Kaplan Meier method, which was combined with the division of high and low risk of the patient samples, the survival status, and survival time in the clinical characteristics of the samples, was used to draw the survival curve of the training samples [14]. ROC curve was used to further verify the scientific and feasibility of the prognosis model constructed by LASSO regression. The survival time of 4 years was selected for ROC curve analysis. If the AUC value is more than 0.5, it indicated that the prognostic model obtained under the LASSO regression method performs well for the mining and analysis of prognostic risk genes of oral

Table 3: 23 risk genes negatively related to oral cancer survival.

Symbol	Gene name	HR	*P*-value
ALG3	Alpha-1, 3-mannosyltransferase	1.085	$3.163E-03$
APOL1	Apolipoprotein L1	1.023	$2.225E-02$
B4GALNT1	Beta-1, 4-N-acetyl-galactosaminyltransferase1	1.016	$4.053E-03$
BID	BH3interacting domain death agonist	1.029	$2.416E-03$
C21ORF33	Chromosome 21 open reading frame 33	1.084	$4.658E-02$
C3ORF26	Chromosome 3 open reading frame 26	1.059	$2.108E-02$
CLYBL	Citrate lyase beta like	1.003	$6.712E-03$
UPRT	Uracil phosphoribosyltransferase homolog	1.119	$6.376E-03$
ELOVL6	ELOVL fatty acid elongase6	1.078	$7.762E-03$
FMNL3	Formin like 3	1.128	$5.750E-03$
TMEM144	Transmembrane protein 144	1.322	$7.570E-05$
SLC25A4	Solute carrier family 25 member 4	1.086	$8.077E-04$
LRG1	Leucine rich alpha-2-glycoprotein1	1.106	$6.804E-03$
P4HA1	Prolyl 4-hydroxylase subunit alpha1	1.055	$1.878E-03$
MIAT	Myocardial infarction associated transcript (non-protein-coding)	1.005	$1.982E-02$
NAA38	N (alpha)-acetyltransferase38, NatC auxiliary subunit	0.970	$4.258E-02$
HS3ST1	Heparan sulfate-glucosamine3-sulfotransferase1	0.922	$1.481E-03$
SLC5A1	Solute carrier family 5 member 1	0.971	$8.087E-03$
TNFRSF25	TNF receptor superfamily member 25	0.920	$1.200E-03$
GRAP	GRB2-related adaptor protein	0.930	$1.151E-02$
OSR2	Odd-skipped related transcription factor 2	0.986	$5.556E-3$
PDHB	Pyruvate dehydrogenase (lipoamide) beta	0.920	$1.904E-02$
CELSR3	Cadherin EGF LAG seven-pass G-type receptor 3	0.971	$1.152E-02$

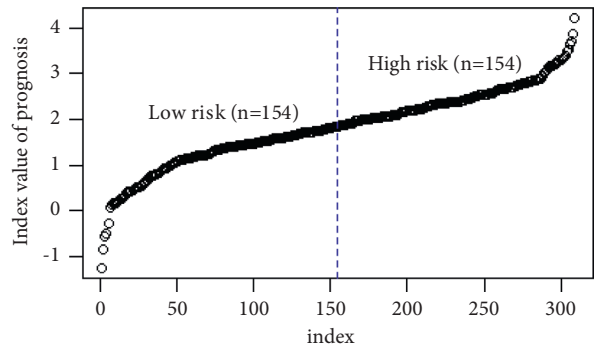

Figure 3: Distribution of PI values under LASSO regression analysis.

cancer, and the analysis results are shown in Figure 4. It can be seen that the high-risk group was clearly distinguished from the low-risk group, log rank *P* is less than 0.001, and the AUC value of LASSO regression is 0.963, which indicates that the model constructed by the LASSO method performs well.

The prognostic values of all samples were sorted from the lowest to the highest, and the median of the prognostic value was taken as a reference. The samples of patients larger than the selected median were considered high-risk patients, and those less than the median were considered low-risk patients [14]. The genetic expression profile of patient samples is shown in Figure 5.

3. Result Verification

3.1. Verification on Validation Set. Kaplan Meier method was used to verify whether the 23 genes screened by the LASSO regression model could distinguish high-risk patients from low-risk patients in the 212 validation samples. It was also necessary to use the ROC curve to further verify the scientific and feasibility of the LASSO model. The analysis results are shown in Figure 6. It shows that these genetic biomarkers could still classify patients with oral cancer in the validation samples into high-risk and low-risk categories.

3.2. Validation Based on Clinical Information. Through the screening and analysis of clinical data, the patient's drinking history, gender, tumor status, age, smoking history, and cancer status were closely associated with this research [17]. Then, the above 6 clinical factors in the clinical information of 485 samples were taken as univariate, and Cox univariate regression analysis was sequentially used on the selected 6 clinical features. The Log-rank *P* value and HR value of each clinical feature were calculated sequentially and the final results are shown in Table 4.

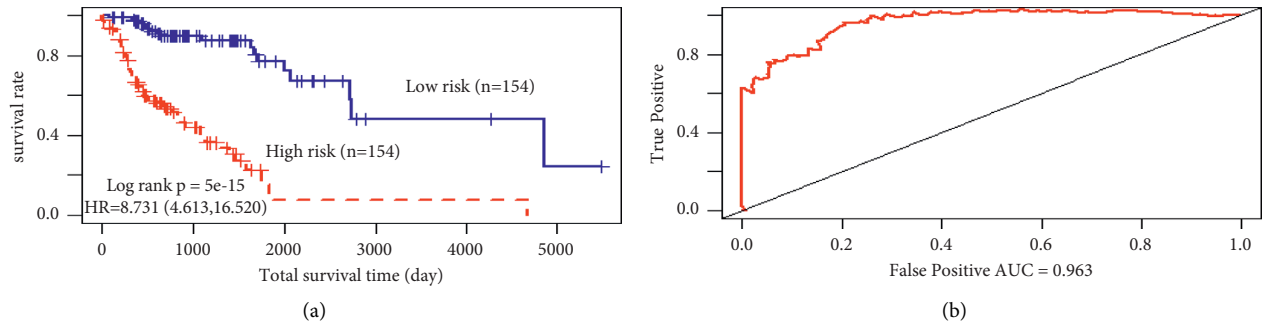

Figure 4: Integrative model for predicting outcome. (a) Survival time curve of training samples under LASSO regression model. (b) ROC curve of the training samples under LASSO regression model.

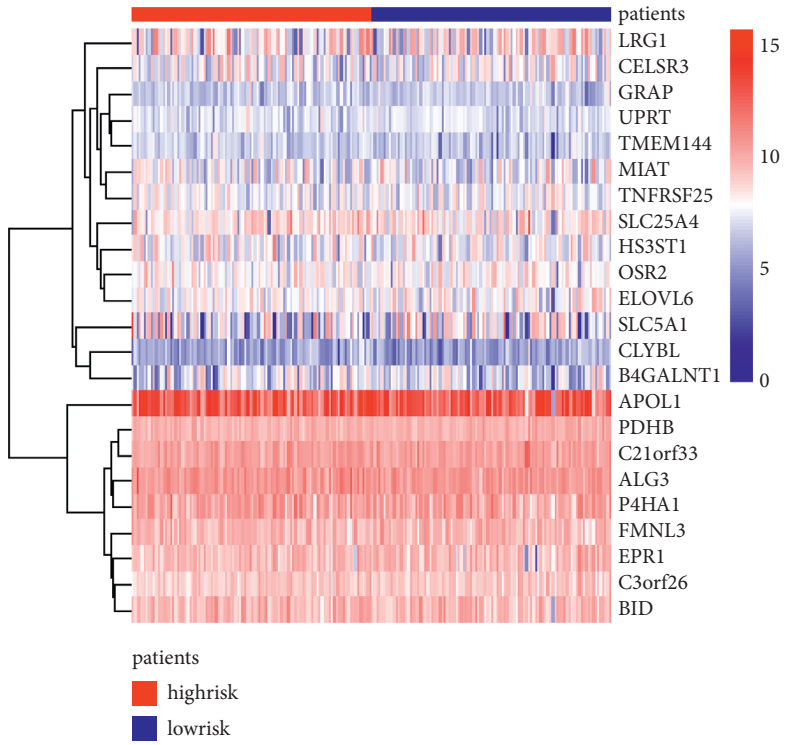

Figure 5: Genetic expression profile of patients under LASSO regression.

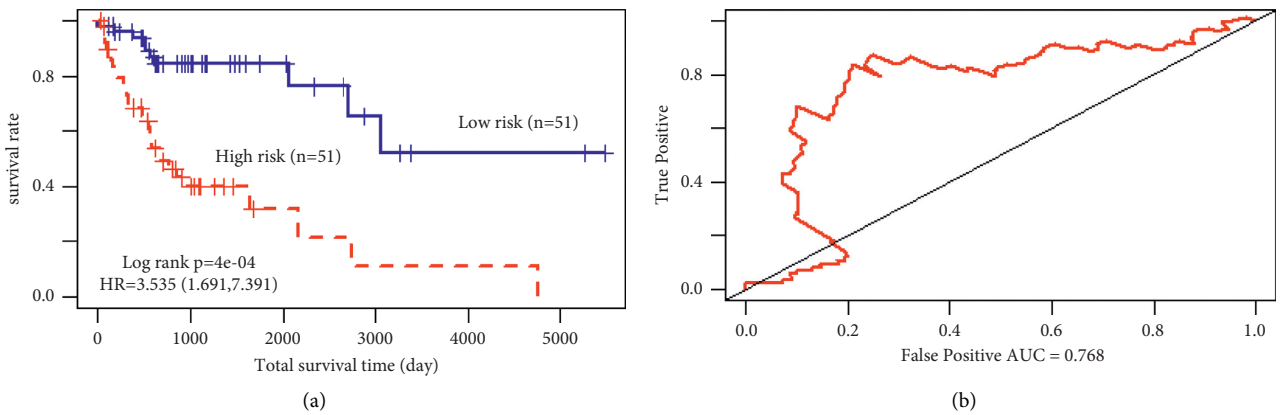

Figure 6: Integrative model for validating the results. (a) Survival analysis curve of validation samples. (b) ROC survival curve of the validation samples.

Table 4: Clinical information statistics table.

Factors	Patients (death)	Log-rank test P value	HR (95% CI)
Alcohol history documented		$7.000E-16$	6.922 (4.058, 11.81)
Yes	111/309		
No	81/166		
Gender		$8.000E-07$	4.384 (2.321, 8.279)
Male	128/346		
Female	68/139		
Age		$5.000E-13$	5.704 (3.381, 9.625)
>61	109/229		
≤61	87/255		
Smoking history		$3.000E-10$	5.304 (2.979, 9.443)
Yes	110/237		
No	86/248		
Pathologic T		$1.000E-10$	13.840 (4.919, 38.93)
T1–T3	109/288		
T4	84/182		
Cancer status		0.005	1.926 (1.209, 3.069)
With tumor	102/121		
Tumor free	66/334		

The results of Cox univariate regression analysis showed that the Log-rank test P of six clinical factors, such as drinking history, sex, tumor status, age, smoking history, and cancer stage, was less than 0.05. Therefore, we can see that these six clinical factors are significantly related to the survival of oral cancer patients. The 6 clinical factors of drinking history, gender, tumor status, and age were used as variables, and the patient samples were divided into two groups; then, the 23 genetic markers screened by the LASSO method in the training samples were used to analyze the survival curve of each clinical factor based on the Kaplan Meier method. The analysis results are shown in Figure 7. It can be seen from the above figures that the six clinical information features we selected are significantly associated with the survival of the oral cancer patients of the research in this paper.

3.3. Comparative Verification Based on Other Data Sets. Using single data set to analyze the test results was often not convincing enough, so other data sets needed to be used to verify the results. The first validation set was from cBioportal, and the data of 385 oral cancer samples were sorted. The survival curve, ROC curve, and AUC value were used to verify the results of the LASSO regression algorithm, and the results are shown in Figure 8. The second validation set was from NCBI, and the data of 74 oral cancer samples were sorted. The survival curve, ROC curve, and AUC value were also used to verify the results of the LASSO regression algorithm, and the results are shown in Figure 9. Therefore, it could be seen that these genes screened by LASSO regression analysis still have good results on other independent data sets and can also better distinguish high-risk and low-risk patients with oral cancer.

4. Genetic Function Analysis

4.1. Genetic Function Analysis Based on String. In order to further analyze and research the relationship between the 23 genes obtained by LASSO regression and oral cancer, we explore the biological activity relationship between these genes and how they affect the survival prognosis of oral cancer patients. We used STRING to analyze genetic function and obtained genetic function network pathway diagram which is shown in Figure 10. It can be seen from the above genetic function network pathway diagram that most of the 23 genes are involved in cell metabolism, the synthesis of biological enzymes, and some life activities associated with cell apoptosis. The life activities of these cells are closely associated with the generation, proliferation, and metastasis of cancer cells. Some of the results are shown in Table 5.

Further analysis results of cell components showed that some genes are involved in cell metabolism, cell apoptosis, and other processes, some genes are involved in the synthesis, metabolism of biological enzymes, and the synthesis and metabolism of nucleotide which affect some life activities of cells, and another part of genes are involved in some activities of mitochondria, and those mitochondrial activities are the energy source of cell life activities. Details are shown in Table 6.

According to some data obtained by STRING genetic function analysis, we can see that there are two pathways in these genes from the genetic function network pathway diagram. The first pathway is composed of 7 genes, namely, PDHA2, DLAT, PDHB, HS3ST1, PDHA1, PDHX, and PDHAX. The second pathway is composed of 3 genes, namely, TNFRSF25, CASP8, and BID. PDHA1 in the first pathway has an extremely important effect on the proliferation of tumor cells. In addition, HS3ST1 is related to the onset of inflammation. The BID in the second pathway is related to cell apoptosis and DNA damage response.

These genes play a vital role in cell mutation, proliferation, and DNA response. Therefore, it is very likely that they have an important impact on the formation and metastasis of cancer cells. In particular, PDHA1 directly and independently affects the prognosis and survival of oral cancer patients.

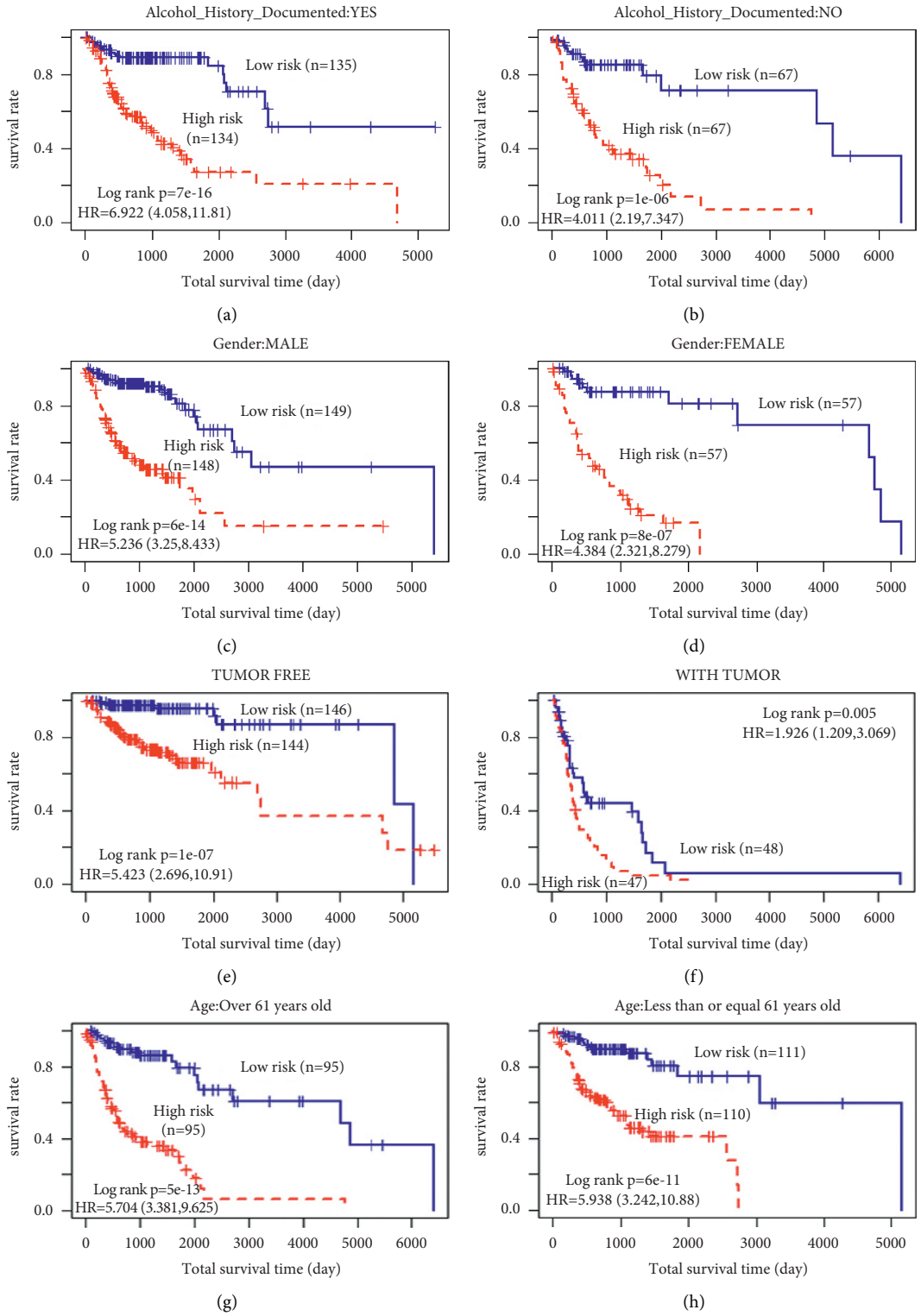

Figure 7: Survival curves of the patients. Survival curves of the patients based on (a, b) alcohol history documented, (c, d) gender, (e, f) cancer status, or (g, h) age, stratified by risk. HR: hazard ratio.

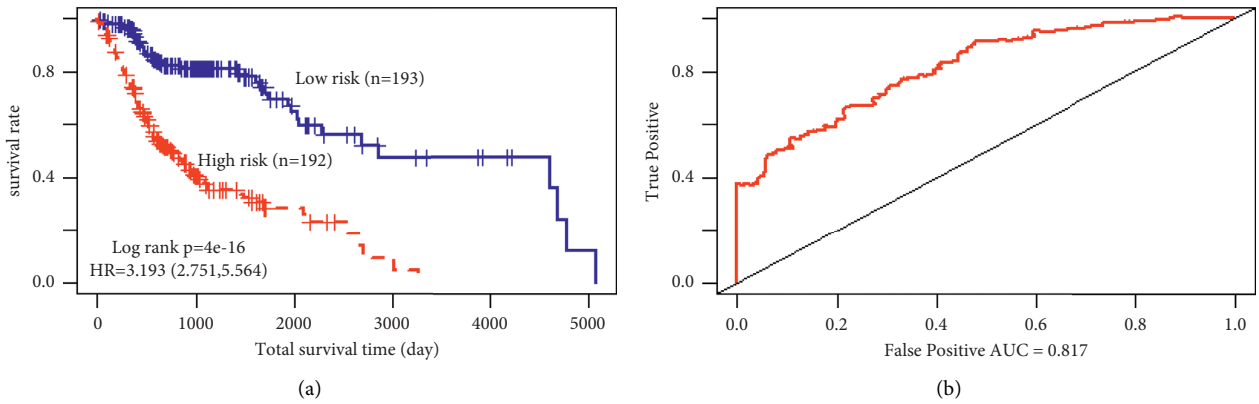

Figure 8: Integrative model for the first comparative verification data set. (a) Survival curve on the first data set. (b) ROC curve on the first data set.

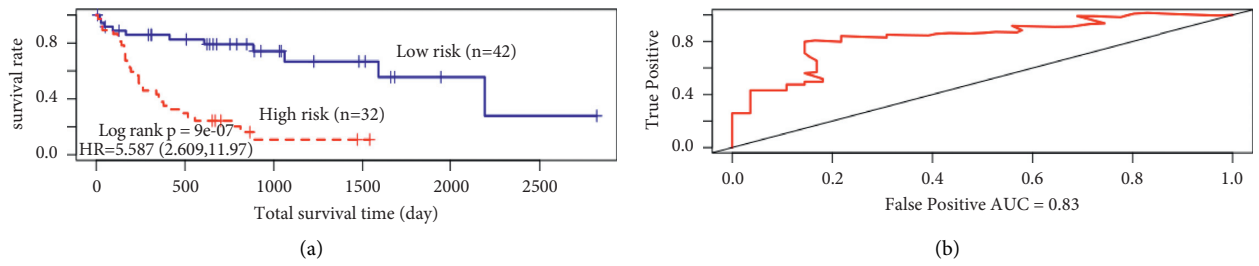

Figure 9: Integrative model for the second comparative verification data set. (a) Survival curve on the second data set. (b) ROC curve on the second data set.

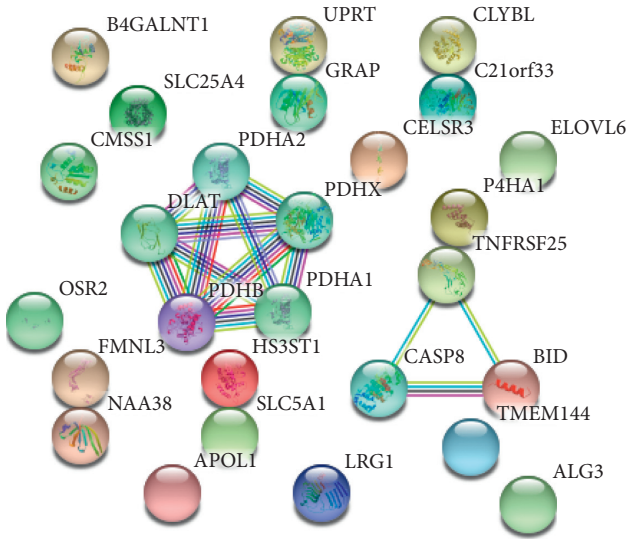

Figure 10: Genetic function network pathway diagram.

4.2. Analysis of Gene Function Based on the Literature

4.2.1. Risk Genes Associated with Oral Cancer. Among the available literature, we found that the prolyl 4-hydroxylase subunit alpha1 (P4HA1) in these 23 genes has a great correlation with the poor prognosis of oral cancer, and P4HA1 is a protein encoded by genes, which is involved in the hydroxylation of proline residues in posttranslational collagen synthesis and takes some responsibility in the prognostic information of polygenic hypoxia signals. The high level of P4HA1-mRNA, as a single gene substitute index of hypoxia, is an independent prognostic indicator of overall survival and local recurrence in oral cancer patients [18].

LRG1 (leucine rich alpha-2-glycoprotein1) is a pleiotropic protein that plays a pathogenic role in a variety of human diseases. The results showed that TGF-β is expressed in oral squamous cell carcinoma. It is of great significance that Lrg1 can control TGF-β pathway in oral squamous cell carcinoma [19].

4.2.2. Risk Genes Related to Other Cancers. FMNL3 belongs to the vertebrate-specific actin polymerization factor superfamily and has a wide range of biological functions in cell and tissue development. In the research of human cancer, FMNL3 is identified as overexpressed in lymphoid malignancies and melanoma and is associated with oncogenic signaling pathways that regulate cancer cell invasion and migration. There are literatures suggesting that increased expression of FMNL3 is associated with the development, metastasis, and poor prognosis of colorectal cancer (CRC) patients [20].

Table 5: Temperature and wildlife count in the three areas covered by the study.

Pathway ID	Description	Observed gene	False discovery rate	Matching proteins in your network (labels)
GO: 1901566	Organic nitrogen compound biosynthesis process	10	$1.700E-04$	ALG3, B4GALNT1, DLAT, ELOVL6, HS3ST1, PDHA1, PDHA2, PDHB, PDHX, UPRT
GO: 0006091	Production of precursor metabolites and energy	6	$2.600E-04$	SLC25A4, DLAT, PDHA1, PDHA2, PDHB, BID
GO: 0006006	Glucose metabolism process	4	$3.600E-04$	DLAT, PDHA1, PDHA2, PDHB
GO: 0016999	Metabolic process of antibiotics	4	$5.100E-04$	DLAT, PDHA1, PDHA2, PDHB
GO: 0032787	Metabolic process of monocarboxylic acid	6	$7.400E-04$	DLAT, ELOVL6, PDHA1, PDHA2,PDHB, PDHX
GO: 0008637	Changes of apoptotic mitochondria	2	$3.570E-02$	BID, SLC25A4
GO: 0044281	Metabolic process of small molecules	10	$1.300E-03$	APOL1, BID, DLAT, UPRT, P4HA1, PDHA1, PDHA2, PDHB, PDHX, ELOVL6
GO: 0060544	Regulation of necrosis process	2	$3.300E-03$	CASP8, SLC25A4
GO: 0008637	Changes of apoptotic mitochondria	2	$3.570E-02$	BID, SLC25A4

Table 6: Partial display table of cellular component (GO).

Pathway ID	Description	Observed gene count	False discovery rate	Matching proteins in your network (labels)
GO: 0045254	Pyruvate dehydrogenase complex	5	$5.570E-10$	DLAT, PDHA1, PDHA2, PDHB, PDHX
GO: 1990204	Oxidoreductase complex	6	$4.510E-07$	DLAT, P4HA1, PDHA1, PDHA2, PDHB, PDHX
GO: 0005739	Mitochondrion	10	$9.300E-04$	BID, CASP8, CLYBL, DLAT, P4HA1, PDHA1, PDHA2, PDHB, PDHX, SLC25A4
GO: 0044429	Mitochondrial part	8	$1.600E-03$	SLC25A4, CASP8, DLAT, PDHA1, PDHA2, PDHB, PDHX, BID
GO: 1902494	Catalytic complex	8	$6.900E-03$	CASP8, DLAT, NAA38, P4HA1,PDHA1,PDHA2, PDHB, PDHX
GO: 0005759	Mitochondrial matrix	5	$8.600E-03$	DLAT, PDHA1, PDHA2, PDHB, PDHX

B_1, 4-N-acetyl-galactosaminyltransferase1 (B4GALNT1) is one of the family members of glycosyltransferase, which is a key enzyme in the synthesis of ganglioside GM2, GD2, and glycolipid GA2. The researches have shown that B4GALNT1 is a key gene of clear cell renal cell carcinoma (ccRCC) metastasis and may become a new diagnostic marker and therapeutic target for ccRCC [21].

LRG1 is activated by HIF-1α to regulate angiogenesis and epithelial-mesenchymal transition (EMT) in colon cancer. According to reports, LRG1 is a potential noninvasive diagnostic and prognostic biomarker in colon cancer [22].

Some scholars wrote in the literatures that MIAT is partly involved in the development of AML (acute myeloid leukemia) through the negative regulation of miR-495; thus, it provided a promising target for the treatment of AML [23].

4.2.3. Risk Genes Associated with Other Diseases. SLC25A4 (A1, member of solute carrier family 4) is an important type of transmembrane glycoprotein, which plays an important role in maintaining the stability of Erythrocyte membrane structure and regulating energy metabolism [24].

GRAP is a low-abundance signaling protein that is enhanced in the samples of diabetic renal tubules and is predicted to be a new component of the TGF-β signaling pathway from biological information analysis [25].

According to the literature, the HS3ST1 gene regulates the inflammation of antithrombin and is related to atherosclerosis. HS3ST1 is a heparan sulfate with a specific Penta saccharide motif and can bind to the anticoagulant protein antithrombin (AT) [26].

TNFRSF25 (tumor necrosis factor receptor superfamily member 25) is the receptor of TNFSF12, APO3L, and TWEAK. It is pointed out in the literature that the methylation level of the TNFRSF25 promoter can be used as an epigenetic biomarker for patients with rheumatoid arthritis (RA) [27].

CELSR3 is an atypical receptor of 7-pass Cadherin and also is an epithelial marker that is downregulated in non-cystic fibrosis primary human bronchial epithelial cells. The

results have shown that the loss of celsr3 function may lead to fibrosis phenotype of noncystic fibrosis primary human bronchial epithelial cells [28–30].

NAA38 is a component of NatCN terminal acetylation complex. It is reported in the literature that the destruction of NAA38 will affect the stability of NRF2 and the expression of glutathione biosynthesis genes, thereby changing the sensitivity of hypertrophy [31].

Autosomal dominant mutation of ANT1 gene (SLC25A4) leads to autosomal dominant inheritance progressive external ophthalmoplegia. It is viewed in the literature that this recessive mutation was described in patients with rare hypertrophic cardiomyopathy, lactic acidosis, and exercise intolerance [24].

Apolipoprotein L1 (APOL1) is a substructure of the cell. It is pointed out in the literature that APOL1 may have an impact on human kidney diseases by participating in the fusion or fission of mitochondrial. The literature also pointed out that the fusion/fission pathway of mitochondria may be a therapeutic target for APOL1-nephropathy [32–34].

4.2.4. Risk Genes Related to Other Life Activities. OSR2 controls the production of tooth organs through the antagonism of secreted Wnt antagonists. The absence of Osr2 can prevent the growth and development of molar organs, including normal continuous bud-shaped to cap-shaped and then to bell-shaped teeth [35].

It can be seen that the genes obtained by mining analysis are basically involved in the synthesis of proteins and enzymes related to cell metabolism and some genes are related to mitochondrial activity. Among them, P4HA1 has been determined to be related to the poor prognosis of oral squamous cell carcinoma. While OSR2 is related to the synthesis of dental organs, and MIAT is overexpressed in the cell lines of patients with acute myeloid leukemia. The other genes mentioned in the literature have a certain connection with the poor prognosis of other cancers, and the remaining part of the genes may be related to human life activities or other diseases.

5. Conclusions

In the whole process of data mining, data analysis is the key. Through the means and methods of data analysis, the effectiveness of data mining can be improved and the accuracy of conclusions can also be ensured. WITMED based on big data has played a huge advantage in medical testing, medical image analysis, clinical diagnosis, and other fields. Big data analysis provides new methods for medical testing and clinical diagnosis and promotes the development of the medical industry. In the future, data analysis will continue to be integrated into the overall context of the smart cities, and the development of various technologies will be improved through various means.

In a word, the 23 genes obtained by data mining in this paper are closely associated with the prognosis of oral cancer, which can provide certain reference and theoretical basis for clinical research, treatment, diagnosis, and prognosis of oral cancer based on molecular biology.

Data Availability

In this paper, the samples of head and neck cancer were downloaded, respectively, in three open source websites of the Xena Functional Genomics Explorer (xenabrowser, xenabrowser.net/datapages/), cBioportal (http://www.cbioportal.org/), and National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov/).

Conflicts of Interest

The authors declare that they have no known conflicts of interest or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partially supported by National Key R&D Program of China under Grant no. 2020YFC0832500, National Natural Science Foundation of China under Grant no. 61402210, Program for New Century Excellent Talents in University under Grant no. NCET-12-0250, and the Key Project of China Ministry of Education for Philosophy and Social Science: Big Data Driven Risk Research on City's Public Safety, under Grant no. 16JZD023. The authors also gratefully acknowledge the support of Strategic Priority Research Program of the Chinese Academy of Sciences with Grant no. XDA03030100 used for this research.

References

[1] W. U. Xibo and Z. Yang, "The Concept of Smart City and Future City Development," *Urban Studies*, vol. 17, pp. 50–56, 2010.

[2] R. Zhou, "Research progress in the diagnosis of oral cancer," *Medicine and Pharmacy of Yunnan*, vol. 34, no. 3, pp. 262–265, 2013.

[3] S. N. Rogers, J. S. Brown, J. A. Woolgar, and D. Lowe, P. Magennis, R. J. Shaw, D. Sutton, D. Errington, and D. Vaughan, Survival following primary surgery for oral cancer," *Oral Oncology*, vol. 45, no. 3, pp. 201–211, 2009.

[4] F. Bray, J. Ferlay, I. Soerjomataram, and R. L. Siegel, "Global cancer statistics 2018: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries: global cancer statistics 2018," *CA: A Cancer Journal for Clinicians*, vol. 68, pp. 394–424, 2018.

[5] S. Arya, D. Chaukar, and P. Pai, "Imaging in oral cancers," *Indian Journal of Radiology and Imaging*, vol. 22, pp. 195–208, 2012.

[6] X. Zhou, W. Liang, R. Huang, K. I. K. Wang, and Q. Jin, "Academic influence aware and multidimensional network analysis for research collaboration navigation based on scholarly big data," *IEEE Transactions on Emerging Topics in Computing*, vol. 9, 2018.

[7] X. Zhou, Y. Hu, W. Liang, Q. Jin, and J. Ma, "Variational LSTM enhanced anomaly detection for industrial big data," *IEEE Transactions on Industrial Informatics*, vol. 17, 2020.

[8] X. Zhou, W. Liang, S. Shimizu, J. Ma, and Q. Jin, "Siamese neural network based few-shot learning for anomaly detection

in industrial cyber-physical systems," *IEEE Transactions on Industrial Informatics*, vol. 17, 2020.

[9] X. Zhou, X. Xu, W. Liang et al., "Intelligent small object detection based on digital twinning for smart manufacturing in industrial CPS," *IEEE Transactions on Industrial Informatics*, vol. 18, 2021.

[10] E. Adi, A. Anwar, Z. Baig, and S. Zeadally, "Machine learning and data analytics for the iot," *Neural Computing & Applications*, vol. 32, 2020.

[11] J. C. Sowder, R. B. Cannon, L. O. Buchmann et al., "Treatment-related determinants of survival in early-stage (T1-2N0M0) oral cavity cancer: A population-based study," *Head & Neck*, vol. 39, p. 876, 2017.

[12] A. Chatterjee, S. Ghosh Laskar, and D. Chaukar, "Management of early oral cavity squamous cancers," *Oral Oncology*, vol. 104, Article ID 104627, 2020.

[13] X. Zhou, L. Yue, and W. Liang, "CNN-RNN based intelligent recommendation for online medical pre-diagnosis support," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, 2020.

[14] C. C. Yu, "Let-7d functions as novel regulator of epithelial-mesenchymal transition and chemoresistant property in oral cancer," *Oncology Reports*, vol. 26, pp. 1003–1010, 2011.

[15] Z. Y. Algamal and M. H. Lee, "Penalized logistic regression with the adaptive lasso for gene selection in high-dimensional cancer classification," *Expert Systems with Applications*, vol. 42, pp. 9326–9332, 2015.

[16] A. H. Zadeh, Q. Alsabi, J. E. Ramirez-Vick, and N. Nosoudi, "Characterizing basal-like triple negative breast cancer using gene expression analysis: a data mining approach," *Expert Systems with Applications*, vol. 148, Article ID 113253, 2020.

[17] A. Fujimoto, M. Furuta, Y. Shiraishi et al., "Whole-genome mutational landscape of liver cancers displaying biliary phenotype reveals hepatitis impact and molecular diversity," *Nature Communications*, vol. 6, p. 6120, 2015.

[18] M. Kappler, J. Kotrba, T. Kaune et al., "P4HA1: a single-gene surrogate of hypoxia signatures in oral squamous cell carcinoma patients," *Clinical and Translational Radiation Oncology*, vol. 5, pp. 6–11, 2017.

[19] Deleheibateer, Y. Peng, and X. Zhou, "Correlation analysis and clinical significane of TGF-β, smads expression and infiltration pattern (Y-K classification) in oral squamous cell carcinoma," *Journal of Inner Mongolia Medical University*, vol. 41, no. 5, p. 10, 2019.

[20] X. Gao, X. Wang, K. Cai et al., "MicroRNA-127 is a tumor suppressor in human esophageal squamous cell carcinoma through the regulation of oncogene FMNL3l," *European Journal of Pharmacology*, vol. 791, pp. 603–610, 2016.

[21] M. Aristidis and H. Carl-Henrik, "Mechanisms of TGFβ-induced epithelial–mesenchymal transition," *Journal of Clinical Medicine*, vol. 5, p. 63, 2016.

[22] C. Chen, M. Zimmermann, I. Tinhofer, A. M. Kaufmanne, and A. Albersb, "Epithelial-to-mesenchymal transition and cancer stem(-like) cells in head and neck squamous cell carcinoma," *Cancer Letters*, vol. 338, pp. 47–56, 2013.

[23] G. Wang, X. Li, L. Song, H. Pan, and L. Sun, "Long noncoding rna miat promotes the progression of acute myeloid leukemia by negatively regulating miR-495," *Leukemia Research*, vol. 87, Article ID 106265, 2019.

[24] A. Tosserams, C. Papadopoulos, C. Jardel, and P. de Lonlay, "Two new cases of mitochondrial myopathy with exercise intolerance, hyperlactatemia and cardiomyopathy, caused by recessive SLC25A4 mutations," *Neuromuscular Disorders*, vol. 26, p. S176, 2016.

[25] T. D. Cummins, M. T. Barati, S. C. Coventry, S. A. Salyerb, J. B. Kleinab, and D. W. Powell, "Quantitative mass spectrometry of diabetic kidney tubules identifies grap as a novel regulator of TGF-β signaling," *Biochimica et Biophysica Acta (BBA) - Proteins & Proteomics*, vol. 1804, pp. 653–661, 2010.

[26] N. C. Smits, T. Kobayashi, P. K. Srivastava et al., "HS3ST1 genotype regulates antithrombin's inflammo modulatory tone and associates with atherosclerosis," *Matrix Biology: Journal of the International Society for Matrix Biology*, vol. 63, no. 69, 2017.

[27] B. Golbargi, N. Ali, S. Mastury, and F. Golbargi, "Analysis of the epigenetic regulation of TNF receptor superfamily 25 (TNFRSF25) in rheumatoid arthritis," *Gene Reports*, vol. 16, Article ID 100424, 2019.

[28] A. A. Lewis, J. T. Mahoney, N. Wilson, and S. E. Brockerhoff, "Identification of amacrine subtypes that express the atypical cadherin celsr3," *Experimental Eye Research*, vol. 30, pp. 51–57, 2015.

[29] G. Chai, L. Zhou, and M. Manto, "Celsr3 is required in motor neurons to steer their axons in the hindlimb," *Nature Neuroscience*, vol. 17, pp. 1171–1179, 2014.

[30] A. Lewis, N. Wilson, G. Stearns, N. Johnson, R. Nelson, and S. E. Brockerhoff, "Celsr3 is required for normal development of gaba circuits in the inner retina," *PLoS Genetics*, vol. 7, Article ID 1002239, 2011.

[31] J. Y. Cao, A. Poddar, L. Magtanong et al., "A genome-wide haploid genetic screen identifies regulators of glutathione abundance and ferroptosis sensitivity," *Cell Reports*, vol. 26, pp. 1544–1556, 2019.

[32] L. Ma, H. C. Ainsworth, H. C. Ainsworth et al., "APOL1 kidney-risk variants induce mitochondrial fission," *Kidney International Reports*, vol. 5, pp. 4–34, 2020.

[33] S. Tzur, S. Rosset, R. Shemer et al., "Missense mutations in the APOL1 gene are highly associated with end stage kidney disease risk previously attributed to the MYH9 gene," *Human Genetics*, vol. 128, pp. 345–350, 2010.

[34] O. A. Olabisi, J.-Y. Zhang, L. VerPlank et al., D. J. Friedman and M. R. Pollak, APOL1 kidney disease risk variants cause cytotoxicity by depleting cellular potassium and inducing stress-activated protein kinases," *Proceedings of the National Academy of Sciences*, vol. 113, pp. 830–837, 2016.

[35] S. Jia, H. J. E. Kwon, Y. Lan, J. Zhou, H. Liu, and R. Jiang, "Bmp4-Msx1 signaling and Osr2 control tooth organogenesis through antagonistic regulation of secreted Wnt antagonists," *Developmental Biology*, vol. 420, pp. 110–119, 2016.