scientific reports

Check for updates

OPEN

Construction and analysis of a conjunctive diagnostic model of HNSCC with random forest and artificial neural network

Yao Luo[1,2], Liu-qing Zhou[1,2], Fan Yang[1,2], Jing-cai Chen[1], Jian-jun Chen[1]✉ & Yan-jun Wang[1]✉

Head and neck squamous cell carcinoma (HNSCC) is a heterogeneous tumor that is highly aggressive and ranks fifth among the most common cancers worldwide. Although, the researches that attempted to construct a diagnostic model were deficient in HNSCC. Currently, the gold standard for diagnosing head and neck tumors is pathology, but this requires a traumatic biopsy. There is still a lack of a noninvasive test for such a high—incidence tumor. In order to screen genetic markers and construct diagnostic model, the methods of random forest (RF) and artificial neural network (ANN) were utilized. The data of HNSCC gene expression was accessed from Gene Expression Omnibus (GEO) database; we selected three datasets totally, and we combined 2 datasets (GSE6631 and GSE55547) for screening differentially expressed genes (DEGs) and chose another dataset (GSE13399) for validation. Firstly, the 6 DEGs (CRISP3, SPINK5, KRT4, MMP1, MAL, SPP1) were screened by RF. Subsequently, ANN was applied to calculate the weights of 6 genes. Besides, we created a diagnostic model and nominated it as neuralHNSCC, and the performance of neuralHNSCC by area under curve (AUC) was verified using another dataset. Our model achieved an AUC of 0.998 in the training cohort, and 0.734 in the validation cohort. Furthermore, we used the Cell-type Identification using Estimating Relative Subsets of RNA Transcripts (CIBERSORT) algorithm to investigate the difference in immune cell infiltration between HNSCC and normal tissues initially. The selected 6 DEGs and the constructed novel diagnostic model of HNSCC would make contributions to the diagnosis.

Head and neck squamous cell carcinomas (HNSCC) mostly derive from the mucosal epithelium in the oral cavity, pharynx and larynx and rank fifth in the world most common tumors. Every year, there has approximately 540,000 new cases and estimated 108,500 deaths from HNSCC in the United States[1]. There have been great advances in surgical techniques and development of adjuvant therapy for HNSCC. Nonetheless, under these current treatment strategies, the 5-survival rate of HNSCC patients is still only 40–50% and remains dissatisfactory[2]. This poor prognosis is connected with metastasis and recurrence, additionally, the high rate of resistance of chemotherapy and locoregional recurrence existing in HNSCC patients[3,4]. The clinical stage (TNM stage) is considered as the most important prognostic factor for patients with HNSCC; however, the survival rate of patients is variable with the same stage[4,5]. Thus, there is an urgent request to hunt for new prognostic biomarkers to modify this situation[6,7]. Recently, the widespread use of microarray technology has made the study of disease mechanisms more convenient. However, the main puzzle of constructing a classification model by utilizing the gene expression data was how to search out the classification index or significant feature. Thus, certain machine learning methods such as random forest (RF)[8,9], artificial neural network (ANN)[10] and CICERSORT software were applied to handle this problem[11]. In this study, the diagnostic model of HNSCC was established by combining the methods above with microarray in Gene Expression Omnibus (GEO) database (The analysis process was shown in Fig. 1).

[1]Department of Otorhinolaryngology, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430022, China. [2]These authors contributed equally: Yao Luo, Liu-qing Zhou and Fan Yang ✉email: ylly80331@163.com; yjwang@hust.edu.cn

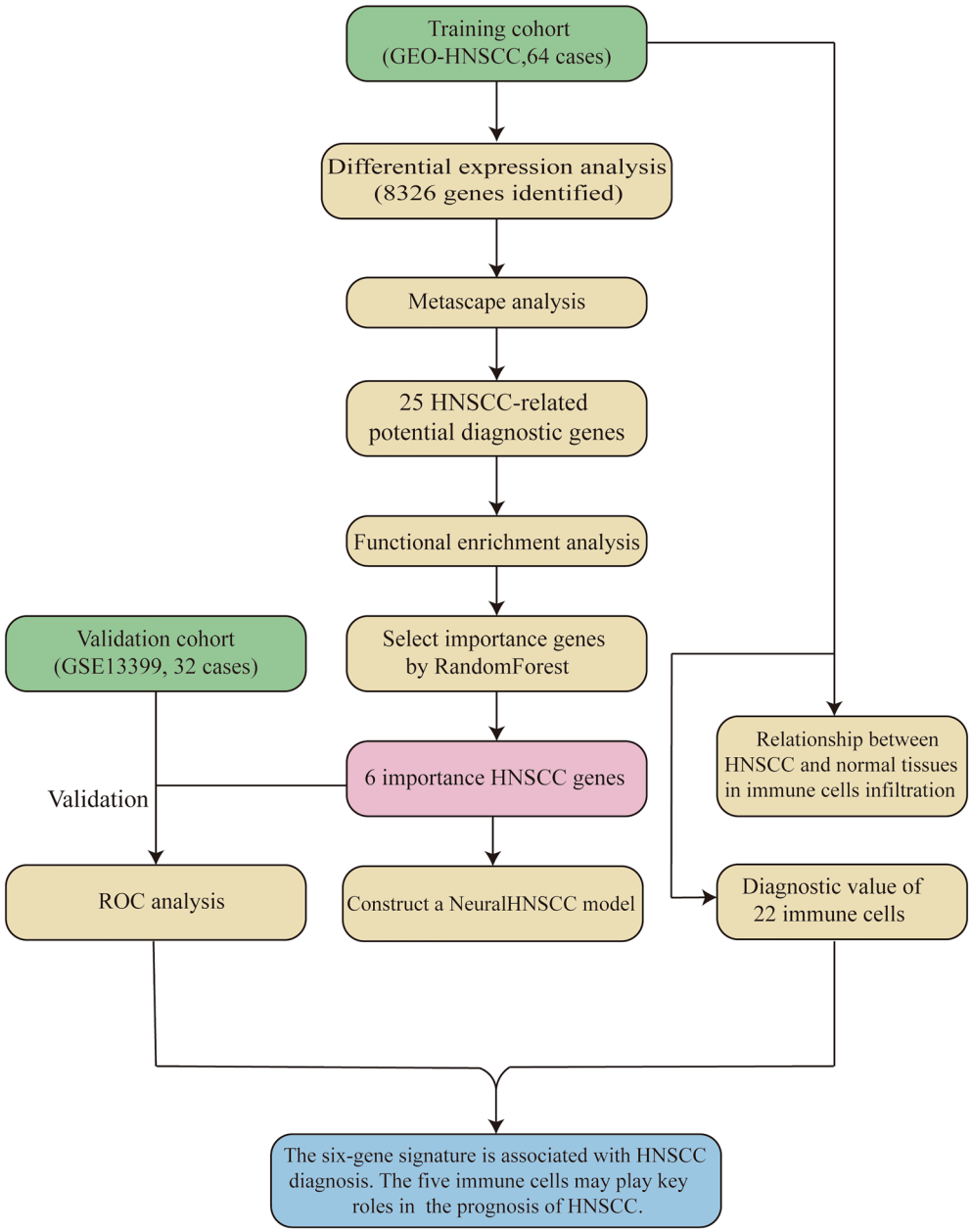

Figure 1. Flow chart of this study.

Materials and methods

Data download. In this study, an extensive search with the key words "HNSCC, human" were conducted through the NCBIGEO platform (https://www.ncbi.nlm.nih.gov/geo/). Then, a total of three datasets were screened, which were demonstrated in Table 1. And we combined two datasets (GSE6631 and GSE55547) as a training dataset. Meanwhile, GSE13399 was used as a validation dataset. The goal of training cohort was to identify the weights of candidate differentially expressed genes and establish the diagnosis model of HNSCC. The effectiveness verification of the classification score model was tested on the validation dataset.

Differentially expressed genes (DEGs) screening and enrichment analysis. The differential analysis was conducted on 38 HNSCC and 26 normal samples of microarray datasets GSE6631 and GSE55547 with a cutoff value of adjust.P.Value (adj.P.Val) < 0.05 and logFoldChang (logFC) > 2.0 by using the Limma package[12]. The heatmap and volcano of DEGs was visualized using the pheatmap software package and ggplot2 software package separately. In order to uncover the Functional or Pathway enrichment analysis of the DEGs, login in http://metascape.org/gp/index.html website to use the Metascape dataset. The specific method is to enter all differential gene names within the gene list, and the prerequisite of the term is that the p-value of term should below 0.01 and the term should contain the count of DEGs is 3 at least.

Data	Sample size	Organization type	Data type
GSE6631	44(Normal: 22; Disease: 22)	Normal tissue: 22 Head and neck squamous cell carcinoma: 22	Microarray
GSE55547	20(Normal: 4; Disease: 16)	Normal benign uvula/tonsil tissue: 4 HPV-negative oropharyngeal squamous cell carcinoma: 16	Microarray
GSE13399	32(Normal: 16; Disease: 16)	Normal tonsil tissue: 16 Head and neck squamous cell carcinoma: 16	Microarray

Table 1. Data download.

Furthermore, R package clusterProfiler was applied for Gene Ontology (GO) function enrichment analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis of DEGs[13,14] and identified three types of significantly enriched GO terms [include biological process (BP), cellular component (CC), and molecular function (MF)] and significantly enriched pathways (both q-value < 0.05). Likewise, the functional analysis of DEGs in GO terms and KEGG pathways were conducted by the Cluster Profiler R package[13]. The R of the KEGG was in the Supplementary File S1. The threshold of p-value was set less than 0.05 and was corrected using false discovery rate (FDR). The results were visualized using the R packages, enrichplot ggplot2 GOplot package[15].

Protein–protein interaction (PPI) analysis. Through Search Tool for the Retrieval of Interacting Genes (STRING) software, Protein–protein interaction (PPI) of DEGs was analyzed. In the PPI network, two or more protein molecules form complexes through non-covalent bonds. STRING is available at https://cn.string-db.org/.

Random forest (RF) classification. Random forest model of DEGs was constructed using the randomForest package[16]. The DEGs were put into the random forest classifier. The effective estimation of our RF prediction error based on the out-of-bag (OOB) error was established, therefore, OOB error estimation was used to optimize the parameters in this research. The optimal parameters of mtry (number of optimal variables in binary tree in node) and ntree were considered in the construction of RF model, and the best variable number for mtry and ntree was set as 5 and 500 respectively. These genes with importance values greater than 2 and ranked in the top 6 were selected for the following analysis, known as disease-specific genes. The unsupervised hierarchical clustering of six significant genes from the combined (GSE6631 and GSE55547) datasets was reclassified using the software package pheatmap to draw a heatmap.

Neural network to build disease classification model. The combined datasets (GSE6631 and GSE55547) were selected to proceed with the next neural network model training. The effectiveness verification of the classification score model was tested on the another independent GSE13399 dataset. The R software package neuralnet and NeuarlNetTools (version R-4.1.1) were used to construct the artificial neural network model[17]. The constitution of a classification model of HNSCC depend on the obtained gene weight information through referring the five hidden layers as model parameters. The model results of five-fold cross-validation were calculated using the confusion matrix function, and the validation results of AUC classification performance were calculated using the pROC[18] software package. The R of the ROC was in the Supplementary File S2. The method of the classification score of established disease neural network model was as follows: neuralHNSCC = $\sum$ (*Gene Expression* × *Neural Network Weight*).

Evaluation of tumor infiltrating immune cells. Gene expression matrix data was uploaded to CIBERSORT (https://cibersort.stanford.edu/) for assessing the abundance of immune infiltrates[11]. A correlation heat map was constructed to visualize the correlation of 22 types of infiltrating immune cells by utilizing "Corrplot" package and "barplot" package was explored by visualized analysis for the proportion of 22 immune cells between the normal group and experimental group. Furthermore, the visualization of the differential expression of immune infiltrating cells in normal and experimental group was used by "vioplot" package.

Results

Verification of DEGs in HNSCC. The expression microarray data of datasets GSE6631 and GSE55547 was downloaded from GEO. Overall, 8326 gene symbols were annotated, in which the distribution of DEGs (adj.P.value < 0.05, |logFC| > 2) were expressed as a volcano plot, concluding 11 upregulated genes and 14 downregulated genes (Fig. 2A). In addition, the heat map of the differential genes was shown in Fig. 2B.

Functional enrichment analysis for DEGs. Metascape was performed for the functional enrichment analysis of these upregulated genes. As indicated in (Fig. 3A), the most commonly enriched terms were external encapsulating structure organization and the biomineral tissue development. Enriched terms have been selected and rendered as a network plot, where terms with a similarity > 0.3 are connected by edges. Nodes are colored to represent their cluster memberships (Fig. 3B).

Subsequently, to further analyze the function of the identified DEGs, GO analysis of DEGs was performed using the online software R. The GO terms comprised three parts: biological process (BP), cellular component (CC) and molecular function (MF).The HNSCC results from the GO analysis indicated that the related BP

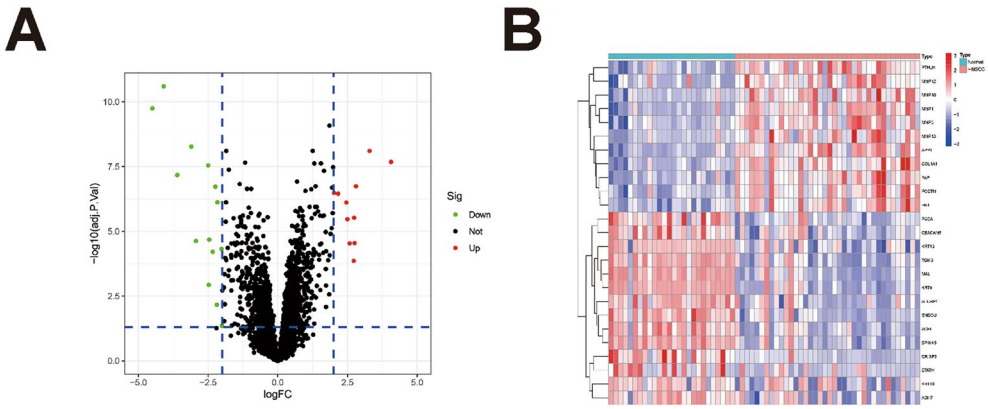

Figure 2. (**A**)Volcano plot of DEGs. Green dots indicate down-regulated genes and red dots indicate up-regulated genes. (**B**) The expression heatmap of DEGs.

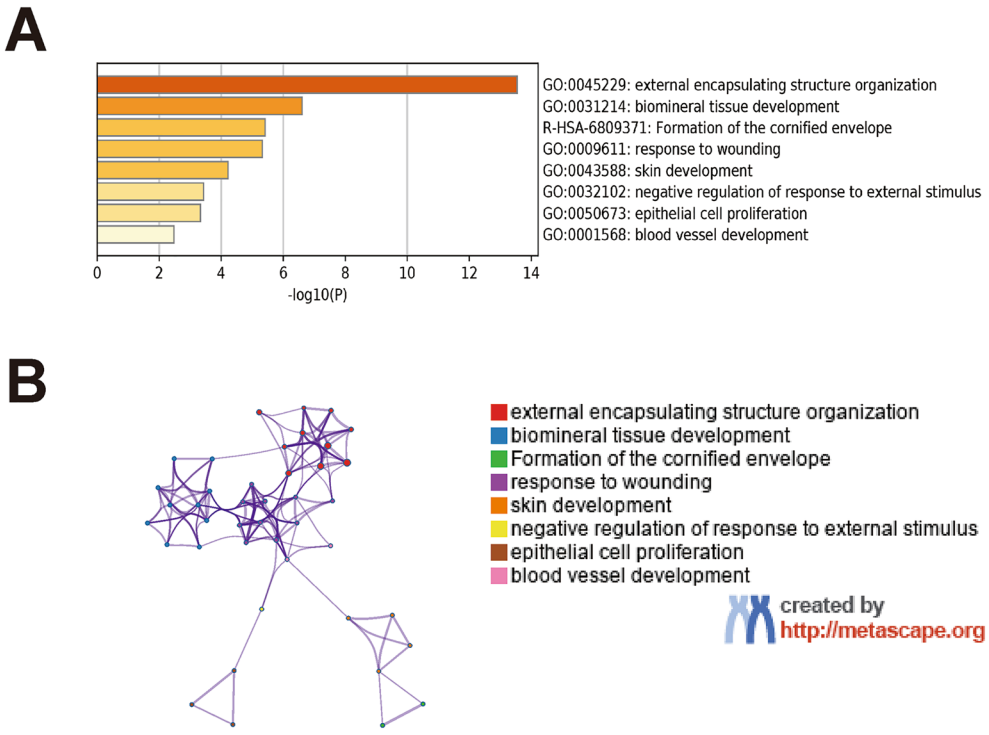

Figure 3. DEGs from the GSE6631 and GSE55547 datasets. (**A**) Metascape enrichment analysis of the DEGs. (**B**) Enrichment networks of DEGs, expressed as cluster memberships, were acquired using the Metascape dataset.

involved in HNSCC included extracellular matrix organization, extracellular structure organization. The CC involved included apical part of cell, and intermediate filament. The MF included metalloendopeptidase activity, metallopeptidase activity, and endopeptidase activity (Fig. 4A,B). The KEGG pathway analysis discovered that the all DEGs were mainly enriched in ECM−receptor interaction, IL−17 signaling pathway, and Relaxin signaling pathway (Fig. 4C,D). Figure 4E shows the GO enriched terms and the significant DEGs involved. Besides, KEGG pathway enrichment analysis of DEGs was also performed (Fig. 4F), showing the results of the significantly enriched biological pathways involved and the corresponding DEGs. In addition, the results of the clustering functional analysis of enriched pathways in GO terms and KEGG pathways were showed in Fig. 4G,H.

In the PPI network, 22 excellent proteins were identified to construct the network using STRING software, which consisted of 55 edges and 25 nodes. The PPI network exhibited that MMP1, MMP3, MMP10, MMP12, MMP13, COL1A1, POSTN, FN1 and SPP1 proteins had nodes with the high connectivity and were relatively more critical (Fig. 5).

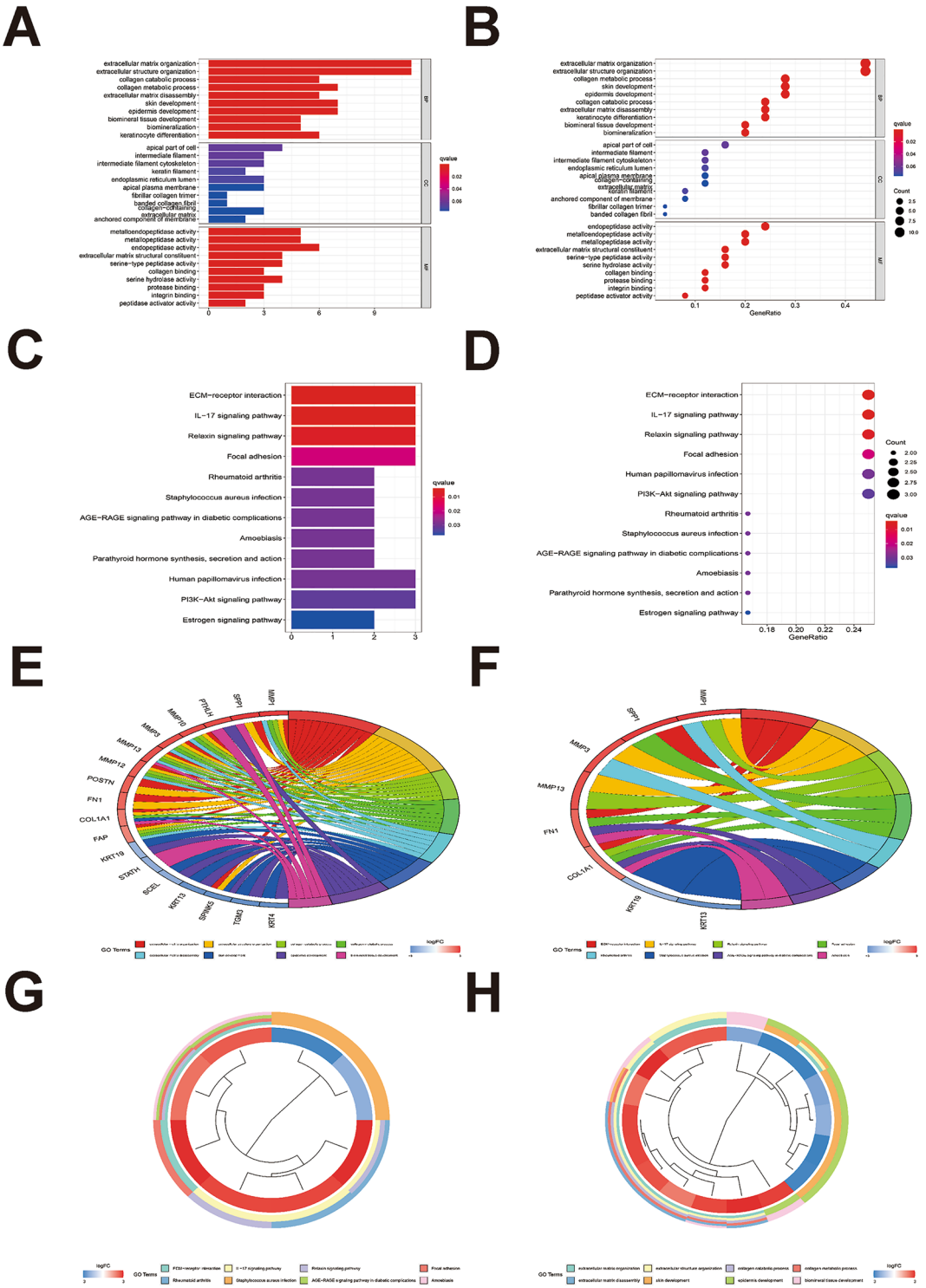

Figure 4. (**A**, **B**) The bar chart and bubble chart of GO enrichment analysis. (**C**, **D**) The 12 most significant KEGG pathways were shown in the bar chart and bubble chart. (**E, F**) Ring plot of GO and KEGG: a plot reveals the relationship between DEGs and their associated pathways. On the left side are DEGs, the color represents upregulation (red) or downregulation (blue). Different colored bands on the right represent different paths. The connecting line shows that this gene is involved in this pathway. (**G, H**) Cluster diagram of GO and KEGG: the top eight significant GO terms and KEGG pathways enriched by DEGs. The color of the inner ring indicates upregulation (red) or downregulation (blue).

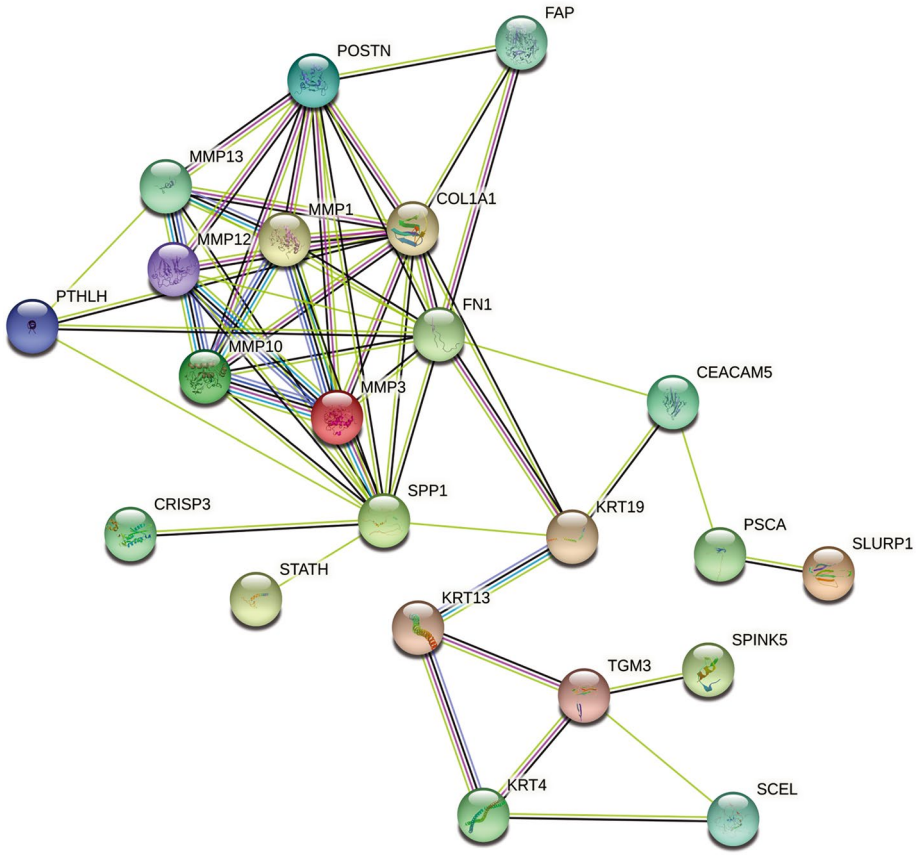

Figure 5. Protein–protein interaction network. Nodes represent genes, and the information inside the node indicates protein structure: empty nodes, protein of unknown 3D structure; filled nodes, some 3D structure is known or predicted; and the lines represent interactions between genetically encoded proteins and the different colors of the lines represent various evidence of interactions between proteins.

Random forest screening for DEGs. Next, DEGs were input into the random forest classifier. In order to find the optimal parameter ntree, the point with the lowest OOB error rate on RF was found, and the number of trees it corresponded to was 83. As the Fig. 6A illustrated, the error of the model was stable. To rank the variables which input in the random forest model, based on the MeanDecreaseGini, the 25 genes were showed from the most significant ones to the least ones (Fig. 6B). The number of variables corresponding to the point with the lowest out-of-bag error rate in the graph was six (Fig. S1), so the genes were selected using the criterion of variable importance greater than 2 and ranking above the 6th. Finally, the top 6 genes that the importance values greater than 2 were selected for subsequent model construction. As Fig. 6B showed that CRISP3 was foremost ranked, and SPINK5, KRT4, MMP1, SPP1, and MAL were followed successively. In addition, the heat map of the top 6 DEGs were constructed (Fig. 6C).

Construction of the artificial neural network model. An artificial neural network model was developed by the R software package neuralnet and NeuarlNetTools on the basis of the combined datasets (GSE6631 and GSE55547). The area under the ROC curve (AUC) of the training cohort was 0.998 showing the excellence classification performance of model (Fig. 7A). The GSE13399 dataset was utilized to test the constructed classification score model for effectiveness verification. The AUC verification result of neuralHNSCC was 0.734 (Fig. 7B). The neural network of the 6 DEGs was displayed in Fig. 7C.

Evaluation of infiltrating immune cells associated with HNSCC. Using the CIBERSORT algorithm, the difference in immune infiltration between HNSCC and normal tissues in 22 immune cells were investigated. The results contained 38 HNSCC patients and 26 normal tissues were visualized in Fig. 8A. As shown in Fig. 8B, the proportion of different immune cells in tumor tissue was weakly to moderately correlated in the GEO cohort. The heatmap demonstrated that the highest positive correlation was in activated memory $CD4^+$ T cells and delta gamma T cells. The violin plot indicated that, in the HNSCC samples, naïve B cells ($p = 0.045$), M0 macrophages ($p < 0.001$), activated dendritic cells ($p = 0.028$) and activated mast cells ($p = 0.002$) in the combined datasets (GSE6631 and GSE55547) infiltrated more, while resting mast cells ($p < 0.001$) infiltrated less (Fig. 8C).

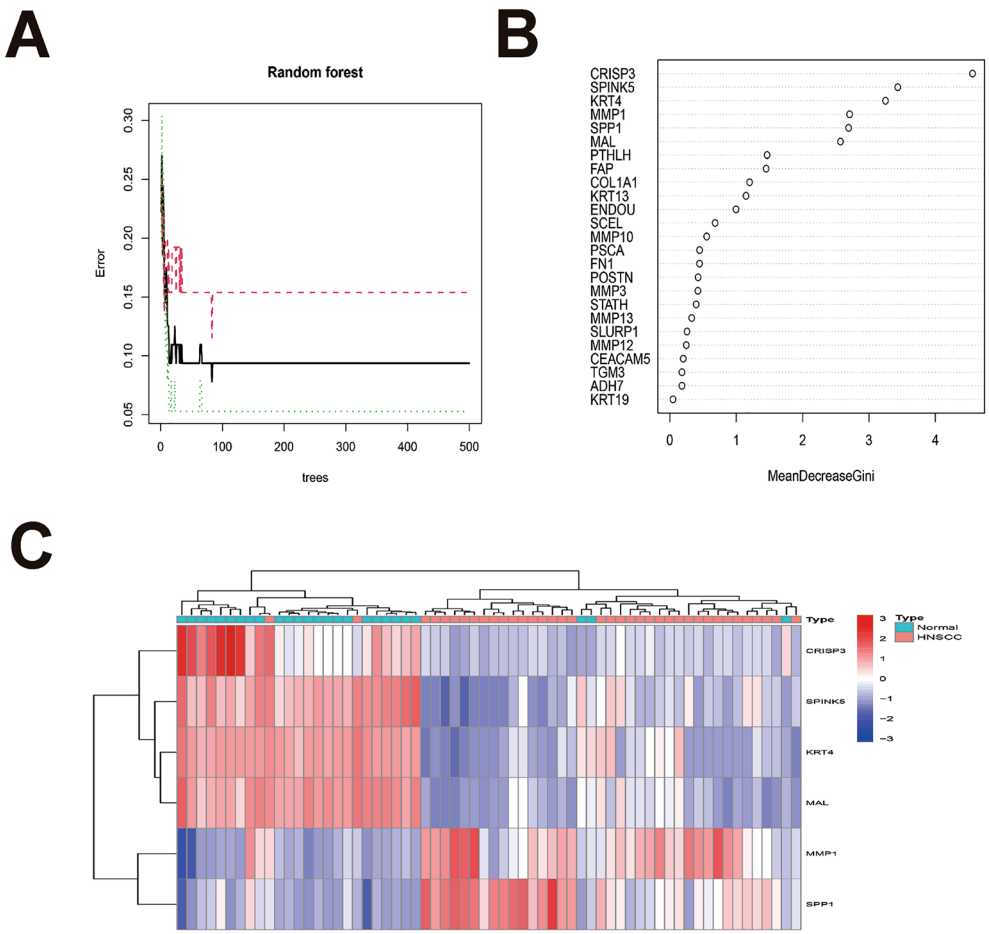

Figure 6. (**A**) The effect of the number of decision trees on the error rate. The x-axis indicates the number of decision trees and the y-axis indicates the error rate. When the number of decision trees is above 100, the error rate is relatively stable. Red color indicates error rates in HNSCC samples, green color indicates error rates in normal samples, and the black color indicates error rates in all samples. (**B**) Results of application of Gini coefficient method in random forest classifier. The x-axis represents the genetic variable and the y-axis represents the importance index. (**C**) Heatmap shows the hierarchical clustering results generated for six significant genes generated by random forest in the combined datasets (GSE6631 and GSE55547).

Discussion

In the present study, we calculated HNSCC-related DEGs and applied random forest classifier to acquire six DEGs in HNSCC. In addition, the neural network models were utilized to measure the prediction weights of related genes in HNSCC. Classification model score neural network related to HNSCC was constituted, and the classification efficiency of model scores on pooled sample datasets was assessed. Moreover, The AUC of the training cohort indicated a high accuracy of the classification model. The GSE13399 dataset showed excellent performance in the verification results. Furthermore, we probed the differential expression of the immune cells between normal and HNSCC group, and the latent roles of the infiltrated immune cells in the pathogenesis of HNSCC.

For the selected 6 DEGs, cysteine-rich secretory protein 3 (CRISP3) was foremost ranked. CRISP3 is a glycoprotein that belongs to a family of cysteine-rich secretory proteins (CRISPs), which was identified in human neutrophilic granulocytes initially[19]. It has been reported that CRISP3 is found to be one of the highly up-regulated proteins during the transition of prostate epithelial cells to prostate cancer in healthy individuals. In addition, it shows that CRISP3 is able to improve cell motility and invasiveness in both human and mouse prostate cancer cell lines[20]. Additionally, Wang et al. reveal that lower expression of CRISP3 was associated with a significantly improved DFS (disease-free survival) and OS (overall survival) in mammary carcinoma, and may provide an unprecedented approach for the treatment[21]. Furthermore, it was reported that in oral squamous cell carcinoma (OSCC), CRISP3 was down-regulated in tumor tissues, and its DNA copy number loss was found in T1/T2 classification and down-regulated CRISP3 may be a protective biomarker in OSCC[22]. However, the relevant research to further investigate its role in the development and progression of OSCC is needed.

SPINK5, serine peptidase inhibitor Kazal type 5, authorized as a lymphoepithelial Kazal type related inhibitor, is a member of the Kazal type family of serine protease inhibitors[23]. Previous studies have illustrated that SPINK5 is in connection with Netherton syndrome (NS) and may play an active biological role in the coagulation

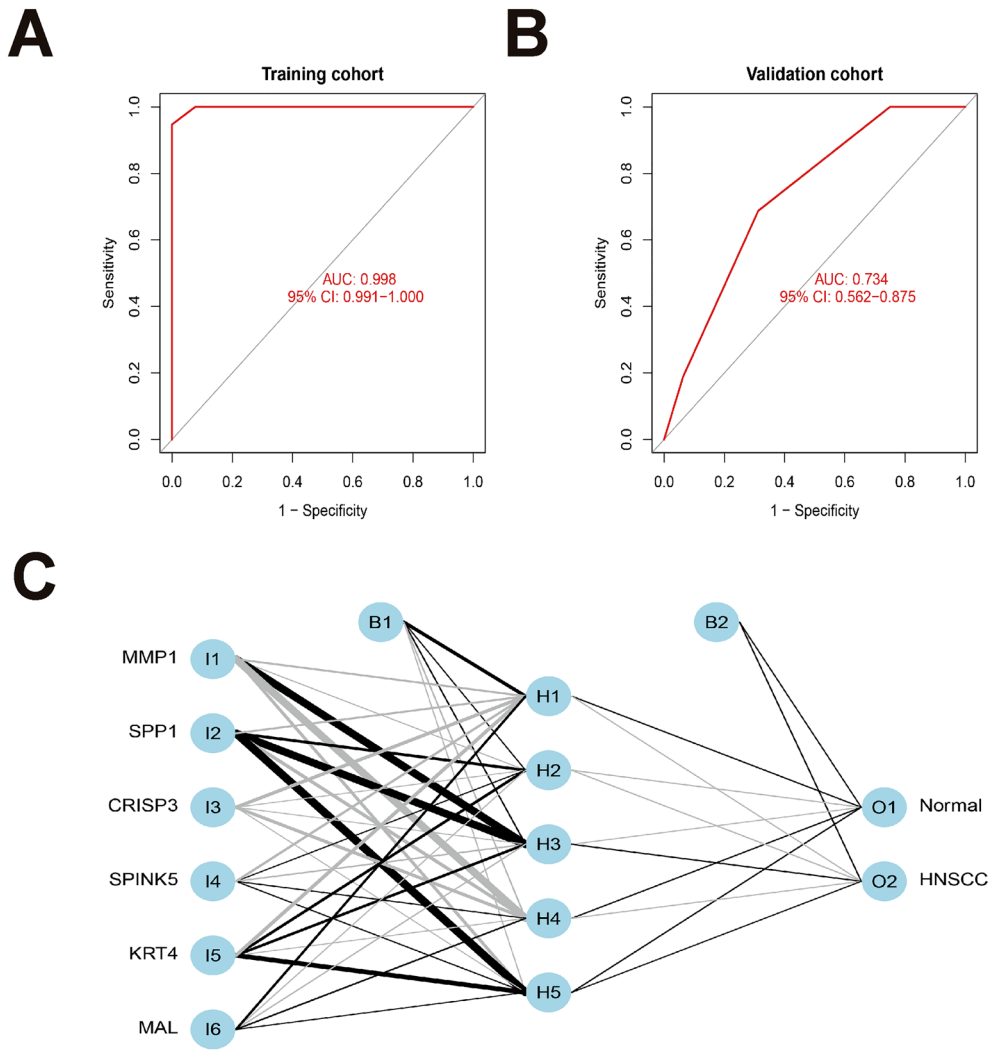

Figure 7. (**A**) AUC of the training cohort shows the classification performance of model. (**B**) Graph of AUC verification result. AUC verification results in the GSE13399 dataset. The AUC value is the area under the ROC curve. (**C**) Visualization results of neural network.

process[24,25]. Later studies revealed SPINK5 may be associated with tumor biological behavior. There also had research reported that SPINK5 was down-regulated in HNSCC tissues compared with adjacent normal tissues[26], subsequently the team showed that down-regulated SPINK5 promoted the proliferation, cluster formation and invasion of HNSCC cells. Liu et al.[27], discovered SPINK5 expression was decreased in Laryngeal Squamous Cell Carcinoma (LSCC) tissues and uncovered the role of increased SPINK5 in LSCC was associated with better survival time and prognosis, which is subtype carcinoma of HNSCC. Besides, SPINK5 was downregulated in esophageal squamous cell carcinoma (ESCC) and non-small cell lung carcinoma (NSCLC) compared with in normal squamous epithelium, and SPINK5 may be a protective gene in ESCC and NSCLC[28,29]. In our research, SPINK5 was down-regulated in HNSCC and significantly associated with HNSCC, and these findings demonstrated that SPINK5 was an antitumor gene and deserved more research in HNSCC and its subtypes in the future.

KRT4, keratin 4, one of the Keratin gene family members, is the major protein found in the epidermis and hair follicles. As intermediate filament proteins, it plays several important roles within the cell. KRT4 encodes a type II cytokeratin, cytokeratin 4 (CK4), found specifically in the esophageal epithelial differentiation layer. The previous research has found that KRT4 was down-regulated in the epithelium of HNSCC and the decreased expression may be relevant to local recurrence in HNSCC[30,31]. Furthermore, another research found KRT4 participated in the local recurrence of HNSCC[32]. CK4 is a predictive biomarker for chemoradiotherapy and surgery in esophageal cancer[33]. In our study, KRT4 was down-regulated in HNSCC, however, the mechanism of KRT4 and CK4 encoded by it are involved in the pathogenesis of HNSCC is still unknown, more studies should be need.

MMP1, matrix metallopeptidase 1, is a member of matrix metalloproteinases (MMPs) comprise a family of endopeptidases which include more than 28 human matrix metalloproteinases, regulate the tumor microenvironment by degrading extracellular matrix (ECM) components[34]. In the previous research, MMP1 was verified to be increased in uveal melanoma and cervical squamous cell carcinoma[35]. It was also reported that MMP1

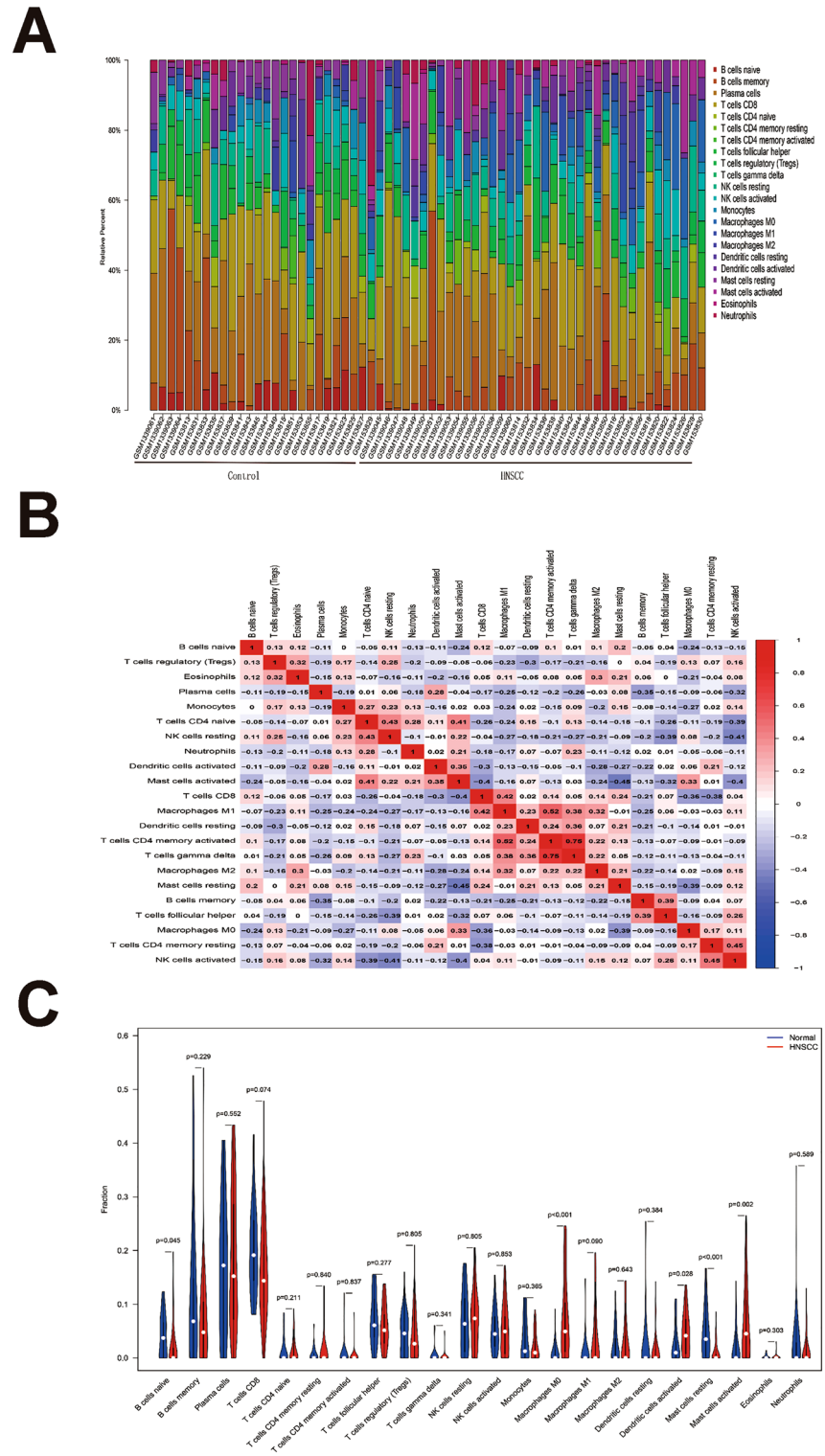

Figure 8. Composition of infiltrating immune cells between paired tumor and adjacent normal tissue in GEO cohort, $p < 0.05$ for CIBERSORT for all eligible samples. Twenty-five immune cells from the GEO group were screened for analysis. (**A**) Fractions of immune cells from 38 tumor samples and 26 normal samples in GEO. (**B**) Correlation heat map between tumor-infiltrating immune cells. Red: positive correlation; blue: negative correlation. The deeper the color, the stronger the association among them. (**C**) Violin plot showing the differentially infiltrating immune cells.

was up-regulated in OSCC and accelerated the growth of the tumor and the motility of the cell in OSCC[36]. And other studies also suggested that MMP1 might be involved in the invasion, metastasis, and poor prognosis of OSCC[37–40]. In our study, MMP1 was discovered to be increased in HNSCC and was supposed to be associated with the poor prognosis in HNSCC, which was in accordance with the results of the previous research.

SPP1, secreted phosphoprotein 1, is a secreted phosphoglycoprotein involved in many biological functions, including cell adhesion, migration and invasion[40–42]. And in the meanwhile, overexpressed SPP1 participates in downregulating the epithelial biomarkers and upregulating the mesenchymal biomarkers. These results indicate that SPP1 promotes the epithelial-to-mesenchymal transition (EMT), which is consistent with the tumor promoting characteristics of SPP1. EMT is related to acquire the invasiveness of cancer cells, thus leading to the metastasis of cancer and resistance of chemotherapy[43]. High expression of SPP1 was primarily owing to its decreased methylation in lung cancer, besides, SPP1 can affect the metastasis and chemoresistance of lung cancer cells and thus was associated with poor prognostic and survival in patient with lung cancer[44]. Similarly, it has been reported to upregulate in many cancers, such as osteosarcoma, gastric cancer, oral squamous cell carcinoma, lung cancer, and overexpressed SPP1 was associated with poor prognosis and survival in cancers[45–48]. In this study, SPP1 was found to up-regulate in HNSCC tissues and was supposed to be a promoting-cancer biomarker.

MAL, Myelin and lymphocyte protein, encodes a membrane protein that is regarded as a central component of the complete protein machinery for apical transport[49]. Furthermore, MAL protein may be involved in the cell polarity. The earlier studies have suggested that methylation of the MAL promoter participated in the inactivation mechanism in HNSCC. And Beder et al. also indicated that MAL expressed selectively decreased or lost in HNSCC metastatic tumor cells in comparison to primary tumor cells, showing that MAL gene may be relevant to suppress metastasis in HNSCC[50,51]. In conclusion, our results were in accordance with the earlier research, the MAL expression was decreased in HNSCC patients, indicating that the MAL gene may be a new candidate tumor-suppressor gene for HNSCC.

Furthermore, immune cells are an important ingredient in the tumor microenvironment (TME) and their infiltration is deemed to play an important role in the biological behavior of a variety of cancers[52,53]. The immune cells in the tumor microenvironment may have functions for promoting or suppressing tumor[54]. The previous studies have demonstrated that HNSCC cells had immunosuppressive properties and had capacity to evade the recognition of immune system[55,56]. In this research, the proportions of naïve B cells, M0 macrophages, activated dendritic cells and activated mast cells significantly increased in HNSCC than in normal tissues, while resting mast cells significantly declined in HNSCC tissues than in normal tissues. The function of naïve B cells was debatable. The previous studies on the role of naïve B cells in HNSCC have indicated that the function of naïve B cells associated with a better survival of HNSCC[57–60], however, another study have confirmed that naïve B cells were associated with tumorigenesis and progression of HNSCC[57]. Therefore, the function of naïve B cells in the pathogenesis of HNSCC needed more research to define. Besides, Ge et al. found that M0 macrophages were indicated to be connected with lymph node stage in colorectal cancer (CRC). M0 macrophages had the highest fraction in N1 stage of CRC, while N2 stage tumors showed the lowest fraction ($p < 0.05$). This result indicated that the tumor-infiltrating immune cells changed in different tumor stages and displayed complicated functions in tumor progression[61]. In this study, the proportion of M0 macrophages was higher in HNSCC group, and Liang et al. supposed higher proportion was associated with the poor survival in HNSCC[62]. Additionally, the functions of dendritic cells (DCs) in HNSCC could vary in the different stages of tumor development and the activation state or the polarization of TME, which could stimulate or suppress immune response[63,64]. And Jin et al. reported that resting mast cells infiltrated less in patients in HNSCC of advanced T stage and they speculated that resting mast cells may have the prospective functions of inhibiting HNSCC malignant progression[65]. In addition, our results showed that activated dendritic cells displayed a higher infiltration in HNSCC, so we supposed that the activated dendritic cells may be a tumor promotor in HNSCC. According to the previous research, accumulation of mast cells is associated with an increase in neovascularization, expression of mast cell VEGF, tumor aggressiveness and poor prognosis[66]. Moreover, Jin et al. also reported that resting mast cells may inhibit the progression of HNSCC[65]. In summary, the function of infiltrated immune cells of TME was in dispute and more research was required to probe the exact function of immune cells in HNSCC.

Nevertheless, we have to acknowledge some limitations existed. Firstly, our total sample size was not large enough, the sample size of each dataset was small, in order to obtain a larger sample size in the training dataset, the two datasets were combined. It was not the most appropriate dataset even if the batch effect was eliminated by the Combat[67]. The datasets we applied were all microarray datasets but no RNA-seq dataset, and it would be more comprehensive and persuasive that a diagnostic model containing RNA-seq data should be established in the future. Besides, the clinical information on patients was incomplete, it remained to be examined whether the model we obtained was fully applicable to patients with HNSCC in clinical practice. And another deficiency was that the tissue source of gene expression profiling for one of our training datasets was oropharyngeal carcinoma, a category of HNSCC, thus, the DEGs selected by our model may be more strongly associated with oral squamous cell carcinoma. Moreover, all the data from GEO were collected from Western countries, which may lead to a biased analysis result. However, our model could serve as a supplement to the existing clinical diagnosis and treatment methods.

In this study, we used microarray datasets to construct a novel diagnostic model for HNSCC based on machine learning algorithms (ML), and the microarray data from GEO showed an excellent diagnostic performance. The novelty of our scoring model was reflected in its comprehensive consideration of two aspects of the genes and their weights that are essential for classification. However, further studies will be required to validate this model, and further analysis with more comprehensive clinical data and from other countries not contained in the GEO may obtain more accurate diagnosis. In recent years, the application of ML techniques in the diagnosis and prediction of diseases including HNSCC has increased significantly[68,69]. With the application of the ML technology in genomic data analysis, a series of diagnosis or prognostic prediction models have been

generated. ML has great potentialities to make contributions to true precision medicine. The proportions of 22 immune cells in tumor microenvironment of HNSCC was disclosed and the clinical function of immune cells was emphasized. Through the combination of rigorous algorithm and genomic data, the results of the present study revealed that the 6 key genes and 5 infiltrated immune cells of TME may involve in the pathogenesis of HNSCC. These findings have contributions to the strategies in clinical immunotherapy and individual-based treatment for HNSCC patients. Whereas, further researches are demanded to confirm these findings.

Data availability

The datasets used and/or analyzed during the current study available from the corresponding author on reasonable request.

Received: 25 August 2022; Accepted: 30 March 2023
Published online: 25 April 2023

References

1. Siegel, R. L., Miller, K. D., Fuchs, H. E. & Jemal, A. Cancer statistics. *CA Cancer J. Clin.* **71**(1), 7–33 (2021).
2. Leemans, C. R., Braakhuis, B. J. & Brakenhoff, R. H. The molecular biology of head and neck cancer. *Nat. Rev. Cancer* **11**(1), 9–22 (2011).
3. Marur, S. & Forastiere, A. A. Head and neck squamous cell carcinoma: Update on epidemiology, diagnosis, and treatment. *Mayo Clin. Proc.* **91**(3), 386–396 (2016).
4. Lydiatt, W. M. *et al.* Head and Neck cancers-major changes in the American Joint Committee on cancer eighth edition cancer staging manual. *CA Cancer J. Clin.* **67**(2), 122–137 (2017).
5. Takes, R. P. *et al.* Distant metastases from head and neck squamous cell carcinoma. Part I. Basic aspects. *Oral Oncol.* **48**(9), 775–779 (2012).
6. Conley, B. A. Treatment of advanced head and neck cancer: What lessons have we learned?. *J. Clin. Oncol.* **24**(7), 1023–1025 (2006).
7. Gavrielatou, N., Doumas, S., Economopoulou, P., Foukas, P. G. & Psyrri, A. Biomarkers for immunotherapy response in head and neck cancer. *Cancer Treat. Rev.* **84**, 101977 (2020).
8. Kursa, M. B. Robustness of Random Forest-based gene selection methods. *BMC Bioinform.* **13**(15), 8 (2014).
9. Cai, Z. *et al.* Classification of lung cancer using ensemble-based feature selection and machine learning methods. *Mol. Biosyst.* **11**(3), 791–800 (2015).
10. Chen, Y. C., Ke, W. C. & Chiu, H. W. Risk classification of cancer survival using ANN with gene expression data from multiple laboratories. *Comput. Biol. Med.* **48**, 1–7 (2014).
11. Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat Genet.* **48**(10), 1193–1203 (2016).
12. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**(7), e47 (2015).
13. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
14. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**(1), 27–30 (2000).
15. Ginestet, C. ggplot2: Elegant graphics for data analysis. *J. R. Stat. Soc. A Stat.* **174**(1), 245–246 (2011).
16. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002; 23.
17. Vallejo, G., Mestre-Citrinovitz, A. C., Winterhager, E. & Saragüeta, P. E. CSDC2, a cold shock domain RNA-binding protein in decidualization. *J. Cell Physiol.* **234**(1), 740–748 (2018).
18. Robin, X. *et al.* pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform.* **17**(12), 77 (2011).
19. Udby, L. *et al.* An ELISA for SGP28/CRISP-3, a cysteine-rich secretory protein in human neutrophils, plasma, and exocrine secretions. *J. Immunol. Methods* **263**(1–2), 43–55 (2002).
20. Volpert, M. *et al.* CRISP3 expression drives prostate cancer invasion and progression. *Endocr. Relat. Cancer.* **27**(7), 415–430 (2020).
21. Wang, Y. *et al.* Low expression of CRISP3 predicts a favorable prognosis in patients with mammary carcinoma. *J. Cell Physiol.* **234**(8), 13629–13638 (2019).
22. Ko, W. C. *et al.* Copy number changes of CRISP3 in oral squamous cell carcinoma. *Oncol. Lett.* **3**(1), 75–81 (2012).
23. Wapenaar, M. C. *et al.* The SPINK gene family and celiac disease susceptibility. *Immunogenetics* **59**(5), 349–357 (2007).
24. Ramesh, K., Matta, S. A., Chew, F. T. & Mok, Y. K. Exonic mutations associated with atopic dermatitis disrupt lympho-epithelial Kazal-type related inhibitor action and enhance its degradation. *Allergy* **75**(2), 403–411 (2020).
25. Ramesh, K. *et al.* Homologous Lympho-epithelial Kazal-type inhibitor domains delay blood coagulation by inhibiting factor X and XI with differential specificity. *Structure* **26**(9), 1178-1186.e3 (2018).
26. Liu, J. *et al.* SPINK5 is a prognostic biomarker associated with the progression and prognosis of laryngeal squamous cell carcinoma identified by weighted gene co-expression network analysis. *Evol. Bioinform. Online* **4**(18), 11769343221077118 (2022).
27. Wang, Q. *et al.* A novel tumor suppressor SPINK5 targets Wnt/β-catenin signaling pathway in esophageal cancer. *Cancer Med.* **8**(5), 2360–2371 (2019).
28. Zhang, J. *et al.* SPINK5 is a tumor-suppressor gene involved in the progression of nonsmall cell lung carcinoma through negatively regulating PSIP1. *J. Healthc. Eng.* **25**(2022), 2209979 (2022).
29. Sakamoto, K. *et al.* Down-regulation of keratin 4 and keratin 13 expression in oral squamous cell carcinoma and epithelial dysplasia: A clue for histopathogenesis. *Histopathology* **58**(4), 531–542 (2011).
30. Chung, J. Y. *et al.* A multiplex tissue immunoblotting assay for proteomic profiling: A pilot study of the normal to tumor transition of esophageal squamous cell carcinoma. *Cancer Epidemiol. Biomark. Prev.* **15**(7), 1403–1408 (2006).
31. Schaaij-Visser, T. B. *et al.* Differential proteomics identifies protein biomarkers that predict local relapse of head and neck squamous cell carcinomas. *Clin. Cancer Res.* **15**(24), 7666–7675 (2009).
32. Takashima, K. *et al.* CD24 and CK4 are upregulated by SIM2, and are predictive biomarkers for chemoradiotherapy and surgery in esophageal cancer. *Int. J. Oncol.* **56**(3), 835–847 (2020).
33. Visse, R. & Nagase, H. Matrix metalloproteinases and tissue inhibitors of metalloproteinases: Structure, function, and biochemistry. *Circ. Res.* **92**(8), 827–839 (2003).
34. Egeblad, M. & Werb, Z. New functions for the matrix metalloproteinases in cancer progression. *Nat. Rev. Cancer.* **2**(3), 161–174 (2002).
35. Wang, C., Mao, C., Lai, Y., Cai, Z. & Chen, W. MMP1 3'UTR facilitates the proliferation and migration of human oral squamous cell carcinoma by sponging miR-188-5p to up-regulate SOX4 and CDK4. *Mol. Cell Biochem.* **476**(2), 785–796 (2021).
36. Wang, T., Zhang, Y., Bai, J., Xue, Y. & Peng, Q. MMP1 and MMP9 are potential prognostic biomarkers and targets for uveal melanoma. *BMC Cancer* **21**(1), 1068 (2021).

37. Yen, C. Y. *et al.* Matrix metalloproteinases (MMP) 1 and MMP10 but not MMP12 are potential oral cancer markers. *Biomarkers* **14**(4), 244–249 (2009).
38. Reis, P. P. *et al.* A gene signature in histologically normal surgical margins is predictive of oral carcinoma recurrence. *BMC Cancer* **11**(11), 437 (2011).
39. Zhao, C., Zou, H., Zhang, J., Wang, J. & Liu, H. An integrated methylation and gene expression microarray analysis reveals significant prognostic biomarkers in oral squamous cell carcinoma. *Oncol. Rep.* **40**(5), 2637–2647 (2018).
40. Elias, E. G., Hasskamp, J. H. & Sharma, B. K. Cytokines and growth factors expressed by human cutaneous melanoma. *Cancers (Basel).* **2**(2), 794–808 (2010).
41. Packer, L. *et al.* Osteopontin is a downstream effector of the PI3-kinase pathway in melanomas that is inversely correlated with functional PTEN. *Carcinogenesis* **27**(9), 1778–1786 (2006).
42. Zhou, Y. *et al.* Osteopontin expression correlates with melanoma invasion. *J. Invest. Dermatol.* **124**(5), 1044–1052 (2005).
43. Shih, J. Y. & Yang, P. C. The EMT regulator slug and lung carcinogenesis. *Carcinogenesis* **32**(9), 1299–1304 (2011).
44. Qin, X. *et al.* Cisplatin-resistant lung cancer cell-derived exosomes increase cisplatin resistance of recipient cells in exosomal miR-100-5p-dependent manner. *Int. J. Nanomed.* **15**(12), 3721–3733 (2017).
45. Dalla-Torre, C. A. *et al.* Effects of THBS3, SPARC and SPP1 expression on biological behavior and survival in patients with osteosarcoma. *BMC Cancer* **5**(6), 237 (2006).
46. Junnila, S. *et al.* Gene expression analysis identifies over-expression of CXCL1, SPARC, SPP1, and SULF1 in gastric cancer. *Genes Chromosomes Cancer.* **49**(1), 28–39 (2010).
47. Huang, C. F., Yu, G. T., Wang, W. M., Liu, B. & Sun, Z. J. Prognostic and predictive values of SPP1, PAI and caveolin-1 in patients with oral squamous cell carcinoma. *Int. J. Clin. Exp. Pathol.* **7**(9), 6032–6039 (2014).
48. Tang, H., Chen, J., Han, X., Feng, Y. & Wang, F. Upregulation of SPP1 is a marker for poor lung cancer prognosis and contributes to cancer progression and cisplatin resistance. *Front. Cell Dev. Biol.* **29**(9), 646390 (2021).
49. Marazuela, M., Acevedo, A., Adrados, M., García-López, M. A. & Alonso, M. A. Expression of MAL, an integral protein component of the machinery for raft-mediated pical transport, in human epithelia. *J. Histochem. Cytochem.* **51**(5), 665–674 (2003).
50. Lind, G. E. *et al.* Hypermethylated MAL gene—A silent marker of early colon tumorigenesis. *J. Transl. Med.* **17**(6), 13 (2008).
51. Beder, L. B. *et al.* T-lymphocyte maturation-associated protein gene as a candidate metastasis suppressor for head and neck squamous cell carcinomas. *Cancer Sci.* **100**(5), 873–880 (2009).
52. Binnewies, M. *et al.* Understanding the tumor immune microenvironment (TIME) for effective therapy. *Nat. Med.* **24**(5), 541–550 (2018).
53. Roma-Rodrigues, C., Mendes, R., Baptista, P. V. & Fernandes, A. R. Targeting tumor microenvironment for cancer therapy. *Int. J. Mol. Sci.* **20**(4), 840 (2019).
54. Xue, Y. *et al.* Tumor-infiltrating M2 macrophages driven by specific genomic alterations are associated with prognosis in bladder cancer. *Oncol. Rep.* **42**(2), 581–594 (2019).
55. Johnson, D. E. *et al.* Head and neck squamous cell carcinoma. *Nat. Rev. Dis. Primers.* **6**(1), 92 (2020).
56. Ferris, R. L. Immunology and immunotherapy of head and neck cancer. *J. Clin. Oncol.* **33**(29), 3293–3304 (2015).
57. Song, J. *et al.* Patterns of immune infiltration in HNC and their clinical implications: A gene expression-based study. *Front. Oncol.* **4**(9), 1285 (2019).
58. Zhou, H., He, Y., Li, L., Wu, C. & Hu, G. Identification novel prognostic signatures for Head and Neck Squamous Cell Carcinoma based on ceRNA network construction and immune infiltration analysis. *Int. J. Med. Sci.* **18**(5), 1297–1311 (2021).
59. Zhang, J. *et al.* Comprehensive characterization of the tumor microenvironment for assessing immunotherapy outcome in patients with head and neck squamous cell carcinoma. *Aging (Albany NY).* **12**(22), 22509–22526 (2020).
60. Guo, Y. *et al.* Identification of novel biomarkers for predicting prognosis and immunotherapy response in head and neck squamous cell carcinoma based on ceRNA network and immune infiltration analysis. *Biomed. Res. Int.* **6**(2021), 4532438 (2021).
61. Ge, P. *et al.* Profiles of immune cell infiltration and immune-related genes in the tumor microenvironment of colorectal cancer. *Biomed. Pharmacother.* **118**, 109228 (2019).
62. Liang, B., Tao, Y. & Wang, T. Profiles of immune cell infiltration in head and neck squamous carcinoma. *Biosci. Rep.* **40**(2), BSR20192724 (2020).
63. Engelhardt, J. J. *et al.* Marginating dendritic cells of the tumor microenvironment cross-present tumor antigens and stably engage tumor-specific T cells. *Cancer Cell* **21**(3), 402–417 (2012).
64. Tran Janco, J. M., Lamichhane, P., Karyampudi, L. & Knutson, K. L. Tumor-infiltrating dendritic cells in cancer pathogenesis. *J. Immunol.* **194**(7), 2985–2991 (2015).
65. Jin, Y. & Qin, X. Profiles of immune cell infiltration and their clinical significance in head and neck squamous cell carcinoma. *Int. Immunopharmacol.* **4**(82), 106364 (2020).
66. Cimpean, A. M. *et al.* Mast cells in breast cancer angiogenesis. *Crit. Rev. Oncol. Hematol.* **115**, 23–26 (2017).
67. Fortin, J. P. *et al.* Harmonization of multi-site diffusion tensor imaging data. *Neuroimage* **161**, 149–170 (2017).
68. Howard, F. M., Kochanny, S., Koshy, M., Spiotto, M. & Pearson, A. T. Machine learning-guided adjuvant treatment of head and neck cancer. *JAMA Netw. Open.* **3**(11), e2025881 (2020).
69. Leitheiser, M. *et al.* Machine learning models predict the primary sites of head and neck squamous cell carcinoma metastases based on DNA methylation. *J. Pathol.* **256**(4), 378–387 (2022).

Author contributions

Y.W. and J.C. designed the research; the researchers (Y.L. and L.Z.) independently searched the NCBIGEO platform extensively. Data were carefully checked by J.C., analyzed by Y.L. and F.Y.. Y.L. wrote the paper that was revised by Y.W. and L.Z.. All authors read and approved the final manuscript.

Funding

This work is supported by National Natural Science Foundation of China (81873693), Bethune Charitable Foundation ((Grant Nos. BQE-TY-SSPC (8)-E-01) and Health Commission of Hubei Province Scientific research project (Grant nos. WJ2021M250).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-32620-6.

Correspondence and requests for materials should be addressed to J.C. or Y.W.