scientific reports

Check for updates

OPEN

Early detection of oral squamous cell carcinoma using five tumor-associated autoantibodies and a Naive Bayes-based machine learning model

Lijuan Xu[1,6], Weihong Xie[2,6], Yuanlin Zou[3,4], Qian Yang[5], Zhong Zheng[3,4], Xiaoyue Zhang[3,4], Yihe Hou[3,4], Yuqi Liu[3,4], Meng Li[3,4], Hua Ye[3,4] & Peng Wang[3,4]✉

This study aimed to identify tumor-associated autoantibodies (TAAbs) with diagnostic potential for the early detection of oral squamous cell carcinoma (OSCC). Bioinformatics analyses were used to screen candidate genes. The candidate tumor-associated antigens (TAAs) were selected from the proteins encoded by the candidate genes. Serum levels of corresponding TAAbs were measured by enzyme-linked immunosorbent assay (ELISA) in 496 participants. Eight machine learning algorithms were employed to develop diagnostic models, and Shapley Additive exPlanations (SHAP) were applied to interpret the optimal model. Twelve candidate genes were identified, among which eight encoded proteins were confirmed to be overexpressed in OSCC. Based on mRNA expression evidence, all 12 encoded proteins were included as candidate TAAs. Of the corresponding autoantibodies, five TAAbs (anti-BLM, anti-BUB1, anti-KIF18A, anti-KIF2C, and anti-TPX2) demonstrated potential diagnostic performance in both the training and validation sets. Among the eight models constructed, the Naive Bayes (NB) model performed best, achieving an area under the receiver operating characteristic curve (AUC) of 0.75 (95% CI 0.70–0.80) in the training set and 0.66 (95% CI 0.57–0.75) in the validation set. SHAP analysis indicated anti-KIF2C contributed most to predictive performance. Five TAAbs were identified with diagnostic potential for OSCC. The NB model constructed based on these TAAbs demonstrated potential diagnostic performance.

Keywords OSCC, Bioinformatics, Tumor-associated antigen autoantibodies, Naive Bayes model

Oral cancer is one of the most common malignant tumors of the head and neck[1]. Approximately 90% of oral cancers are squamous cell carcinomas originating from the mucosal epithelium, referred to as oral squamous cell carcinoma (OSCC), which accounts for 38% of all head and neck cancers globally[2,3]. In 2022, there were 389,000 oral cancer new cases and 188,000 deaths worldwide[4]. And the incidence and mortality rates of oral cancer had shown an upward trend from 1990 to 2021[5]. Meanwhile, oral cancer can lead to disfigurement and loss of functions such as swallowing, speech, and taste, causing significant physiological and psychological burden on patients[6].

Despite substantial progress in the treatment of OSCC[7,8], patients with advanced OSCC still face poor responses to drug therapy, functional impairments from surgery, and significantly lower survival rates compared to those with early-stage disease[9]. Now, the main detection methods for OSCC are clinical examination and histopathological examination, but early lesions and precancerous changes are difficult to detect[10].

[1]Department of Clinical Laboratory, The First Affiliated Hospital of Zhengzhou University, Zhengzhou 450052, Henan Province, China. [2]Department of Oral and Maxillofacial Surgery, The First Affiliated Hospital of Zhengzhou University, Zhengzhou 450052, Henan Province, China. [3]College of Public Health, Zhengzhou University, Zhengzhou 450001, Henan Province, China. [4]Henan Key Laboratory of Tumor Epidemiology, Henan International Joint Laboratory of Tumor Markers and Molecular Imaging, Zhengzhou University, Zhengzhou 450052, Henan Province, China. [5]Prenatal Diagnosis Center, The Third Affiliated Hospital of Zhengzhou University/Maternal and Child Health Hospital of Henan Province, Zhengzhou 450052, Henan Province, China. [6]These authors contributed equally: Lijuan Xu and Weihong Xie. ✉email: wangpeng1658@hotmail.com

Histopathological biopsy can most accurately diagnose the disease at the tissue cell level and is considered the gold standard for oral cancer diagnosis. However, it is an invasive procedure, and has a long detection cycle, making it unsuitable for frequent use, early diagnosis, and long-term monitoring. Additionally, factors such as tissue size and depth variations, the sensitivity of cryotechnology, and observer subjectivity can affect the diagnostic results[11]. Therefore, there is a need to explore biomarkers for the diagnosis of OSCC. The human immune system can detect tumor-associated antigens (TAAs) that emerge in the early stages of tumor development, and immune responses generate tumor-associated antigen autoantibodies (TAAbs)[12]. For early cancer diagnosis, TAAbs are present in the serum before clinical symptoms appear[12]. They also have advantages such as higher titers than the corresponding antigens, good stability, and ease of detection, making TAAbs one of the most promising diagnostic biomarkers[13]. Notably, many studies have confirmed the feasibility and rationale of TAAbs for early tumor detection[13,14], and their presence in OSCC serum has also been verified[15,16]. In oral cancer specifically, growing evidence supports their diagnostic potential. For instance, one study employed an immunome protein array to identify 53 differentially expressed autoantibodies in OSCC, with further immunohistochemical validation indicating positive NUBP2 staining in 72% of cases[17]. Another study identified 10 key genes with high expression in oral cancer by integrating transcriptomic data from public databases and protein-protein interaction network analysis, and subsequent serological validation further confirmed a diagnostic biomarker panel composed of three autoantibodies (anti-AURKA, anti-CXCL10, and anti-FOXM1) that exhibited favorable diagnostic performance after logistic regression modeling[16]. Despite these advances, current approaches often rely on single-omic platforms, such as protein arrays or transcriptomic data alone, or employ conventional statistical methods that may not fully capture complex biomarker interactions. Moreover, given that single autoantibodies typically exhibit limited sensitivity, there is a recognized need for multi-target panels[18,19]. These aspects underscore the importance of developing more integrated and analytically robust strategies.

To address this, our study established a systematic technical route with distinct innovations: integrating transcriptomic and proteomic data, we used Weighted Gene Co-expression Network Analysis (WGCNA) and differentially expressed genes (DEGs) to identify biologically meaningful TAAs. We then validated corresponding TAAbs in a large cohort of 496 subjects via enzyme-linked immunosorbent assay (ELISA) and constructed eight machine learning models for diagnosis. Compared with previous studies, our multi-omics integration, WGCNA-aided screening, 496-sample validation and multi-model optimization significantly enhance biomarker reliability and diagnostic performance, facilitating OSCC early detection.

Results

Demographic characteristics of the study population

A total of 248 patients with OSCC and 248 normal controls (NC) were enrolled. Table 1 described the characteristics of all OSCC patients and NC. No statistically significant differences in age or gender distributions were observed between the OSCC and NC groups in either the training or validation sets (all $P > 0.05$). To ensure comparability, the demographic and clinical features of OSCC patients in the training and validation sets were assessed. No significant differences were observed between the two groups (all $P > 0.05$).

Identification of candidate genes across bioinformatics

Based on the WGCNA results, the top 100 genes from the brown module, which was positively correlated with OSCC (correlation coefficient $r = 0.41$, $P < 0.001$), were extracted as hub genes (Fig. 1A–D, Table S1). Considering the generation mechanism of TAAbs, this study focused solely on highly expressed DEGs. A total of 972 upregulated DEGs ($\log_2$FC > 1, adjusted $P < 0.05$) were identified from the GSE30784 dataset, and 523 from the GSE25099 dataset. By intersecting the hub genes with the upregulated DEGs from both datasets, 12 commonly upregulated candidate genes were obtained (Fig. 1E). The mRNA expression levels of these 12 candidate genes were further validated using the Clinical Proteomic Tumor Analysis Consortium (CPTAC) database, where all genes showed significantly higher expression in head and neck cancer tissues compared to adjacent normal tissues ($P < 0.05$; Fig. S1).

Functional enrichment analysis of the candidate genes

GO terms enrichment analysis was performed and categorized into biological process (BP), cellular component (CC), and molecular function (MF). The CC terms showed that these candidate genes were mainly enriched in the spindle, mitotic spindle, condensed chromosomes, and chromosomal regions (Fig. S2A). MF terms revealed predominant involvement in microtubule binding, tubulin binding, microtubule motor activity, cytoskeletal motor activity, and ATPase activity (Fig. S2B). Regarding BP, the genes were largely associated with sister chromatid segregation, mitotic nuclear division, chromosome segregation, nuclear division, and organelle fission (Fig. S2C). KEGG pathway enrichment analysis further indicated that these genes were significantly enriched in pathways related to the cell cycle, motor proteins, and oocyte meiosis (Fig. S2D).

Selection of candidate TAA

Analysis at the protein level showed that proteins encoded by 8 candidate genes exhibited significantly higher abundance in head and neck cancer tissues compared to normal tissues ($P < 0.05$), whereas protein expression data for the remaining four candidate genes—FOXM1, BUB1, CDC6, and KIF18A—were unavailable (Fig. S3). Given the previously observed high mRNA expression levels of all 12 genes, their encoded proteins were all included as candidate TAAs in this study.

The diagnostic value of individual TAAbs for OSCC detected by ELISA

Twelve candidate TAAs were used to assess the serum levels of TAAbs in training set. ELISA results revealed that the expression levels of six TAAbs (anti-FOXM1, anti-BLM, anti-BUB1, anti-KIF18A, anti-KIF2C, and

Variables	Training set		P	Validation set		P
	OSCC (n = 174)	NC (n = 174)		OSCC (n = 74)	NC (n = 74)	
Age						
Median (P25, P75)	59.0 (50.0, 68.0)	59.0 (51.0, 66.3)	0.758	59.0 (53.0, 66.0)	58.0 (50.0, 69.0)	0.765
Gender, n(%)						
Male	126(72.4)	130(74.7)	0.627	57(77.0)	53(71.6)	0.452
Female	48(27.6)	44(25.3)		17(23.0)	21(28.4)	
Smoke, n(%)						
Yes	69(39.7)			32(43.2)		
No	87(50.0)			34(45.9)		
Unknow	18(10.3)			8(10.8)		
Alcohol, n(%)						
Yes	49(28.2)			26(35.1)		
No	95(54.6)			34(45.9)		
Unknow	30(17.2)			14(18.9)		
Family tumor history, n(%)						
Yes	10(5.7)			5(6.8)		
No	133(76.4)			55(74.3)		
Unknow	31(17.8)			14(18.9)		
Tumor location, n(%)						
Tongue	57(32.8)			31(41.9)		
Buccal mucosa	23(13.2)			11(14.9)		
Others	79(45.4)			26(35.1)		
Unknow	15(8.6)			6(8.1)		
Differentiation, n(%)						
Low	14(8.0)			9(12.2)		
Moderate and high	124(71.3)			50(67.6)		
Unknow	36(20.7)			15(20.3)		
TNM stage, n(%)						
Ⅰ - Ⅱ	84(48.3)			30(40.5)		
Ⅲ-Ⅳ	89(51.1)			43(58.1)		
Unknow	1(0.6)			1(1.4)		
Depth of tumor invasion, n(%)						
T0-T1	44(25.3)			16(21.6)		
T2	62(35.6)			21(28.4)		
T3	43(24.7)			20(27.0)		
T4	24(13.8)			16(21.6)		
Unknow	1(0.6)			1(1.4)		
Lymphatic metastatic, n(%)						
Positive	48(27.6)			24(32.4)		
Negative	125(71.8)			49(66.2)		
Unknow	1(0.6)			1(1.4)		

Table 1. Characteristics of study participants. OSCC, oral squamous cell carcinoma; NC, normal control; P25, 25th Percentile; P75, 75th Percentile; TNM, tumor-node-metastasis. P values in the table compare OSCC and NC within the training and validation sets. Additionally, no significant differences were observed between OSCC patients in the training and validation sets for all listed variables (all P > 0.05).

anti-TPX2) were significantly higher in the OSCC group compared to the NC group (Table S2; Fig. 2). The AUCs of these six TAAbs ranged from 0.58 to 0.69 for OSCC detection (Table S3; Fig. 3). To further validate the diagnostic performance, these six autoantibodies were assessed in a validation set. The results showed that, except for anti-FOXM1 ($P = 0.224$), the remaining five TAAbs (anti-BLM, anti-BUB1, anti-KIF18A, anti-KIF2C, and anti-TPX2) remained significantly elevated in the OSCC group and retained diagnostic value (Table S2, 3; Fig. S4). In the validation set, these TAAbs exhibited AUCs of 0.62–0.65, with sensitivity ranging from 27.0% to 32.4%, specificity from 90.5% to 94.6%, accuracy from 58.8% to 62.2%, and Youden index from 0.176 to 0.243 (Table S3; Fig. S5). Among the significant TAAbs, anti-KIF2C demonstrated the highest discriminative performance, with AUCs of 0.69 (95% CI: 0.63–0.75) in the training set and 0.65 (95% CI: 0.56–0.74) in the validation set.

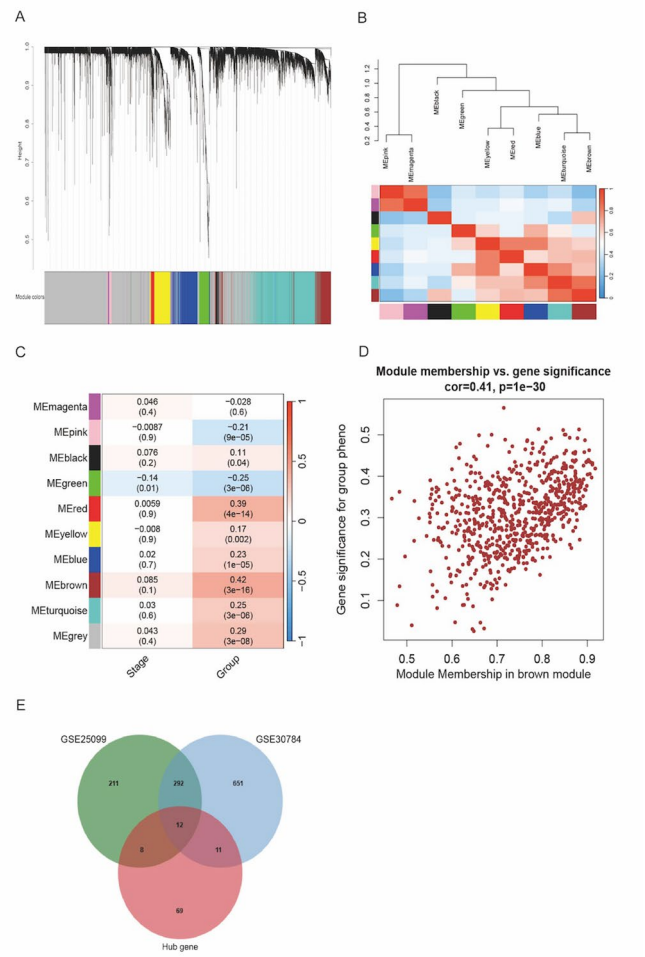

Fig. 1. Weighted gene co-expression network analysis and candidate genes screening. (**A**) Hierarchical clustering dendrogram of samples; (**B**) Heatmap and dendrogram of module-trait relationships; (**C**) Module eigengene adjacency network; (**D**) Module-phenotype correlation analysis; (**E**) Venn diagram of hub genes overlapping with upregulated genes from GSE30784 and GSE25099.

Development, evaluation, and interpretation of machine learning-based diagnostic models

Eight machine learning models were constructed using five TAAbs in the training set to evaluate their diagnostic performance. These models were further tested in the validation set. Based on a comprehensive assessment of diagnostic performance and model stability between training and validation set, the Naive Bayes (NB) model was identified as the optimal diagnostic model (Table 2), owing to its simplicity, robustness to limited sample sizes, and interpretability. In the training set, the NB model achieved an AUC of 0.75 (95% CI: 0.70–0.80), with a sensitivity of 55.7%, specificity of 81.0%, and accuracy of 68.4%. In the validation set, its AUC was 0.66 (95% CI: 0.57–0.75), with a sensitivity of 54.1%, specificity of 78.4%, and accuracy of 66.2% (Fig. 4A). The difference in AUC between the training and validation sets was not statistically significant according to the DeLong test ($P = 0.095$), indicating acceptable model stability. The calibration curve indicated acceptable agreement between predicted and observed probabilities, with a mean absolute error (MAE) of 0.015 and a mean squared error (MSE) of 0.0003 in the training set, and an MAE of 0.062 and an MSE of 0.00576 in the validation set (Fig. S6). The learning curve showed no evidence of substantial overfitting (Fig. S7). Subgroup analysis revealed that the NB diagnostic model demonstrated consistent diagnostic performance for OSCC across different tumor differentiation grades, tumor-node-metastasis (TNM) stages, and lymph node metastasis statuses in both the training and validation sets. No significant differences in diagnostic ability were observed among the subgroups (DeLong test, $P > 0.05$; Table 3).

To interpret the model's predictions, SHAP analysis was performed on the NB model. Among the five TAAbs, anti-KIF2C showed the highest SHAP value in training and validation sets, indicating it was the most influential predictor contributing to model output (Fig. 4B,C).

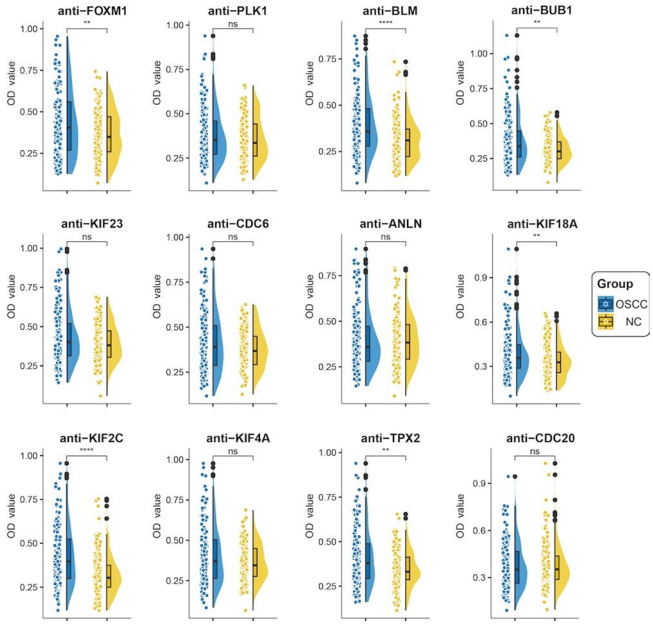

Fig. 2. Comparison of serum levels of 12 autoantibodies between OSCC and NC in the training set. OD, optical density; OSCC, oral squamous cell carcinoma; NC, normal control; **: $P < 0.01$; ****: $P < 0.0001$; ns: $P > 0.05$.

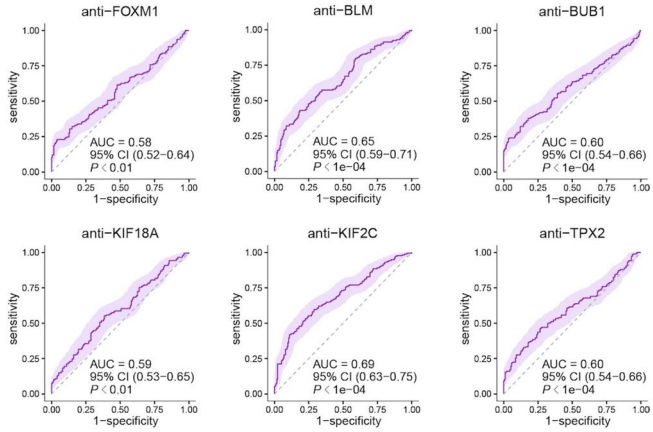

Fig. 3. ROC curve analysis of six autoantibodies (anti-FOXM1, anti-BLM, anti-BUB1, anti-KIF18A, anti-KIF2C, and anti-TPX2) for discriminating OSCC patients from NC in the training set. AUC, area under the ROC curve; ROC, receiver operating characteristic; CI, confidence interval.

Discussion

Early diagnosis of OSCC remains challenging due to the lack of reliable non-invasive biomarkers. While TAAbs have emerged as promising diagnostic tools across multiple cancer types, their study in OSCC remains limited. Therefore, the identification and validation of TAAbs with diagnostic potential is of critical importance for improving the early detection of OSCC. In this study, twelve candidate genes (FOXM1, PLK1, BLM, BUB1, KIF23, CDC6, ANLN, KIF18A, KIF2C, KIF4A, TPX2, and CDC20) were identified through comprehensive bioinformatics analyses, and their encoded proteins were selected as candidate TAAs. Subsequent experimental validation by ELISA demonstrated that five autoantibodies—anti-BLM, anti-BUB1, anti-KIF18A, anti-KIF2C, and anti-TPX2—exhibited significant diagnostic potential for OSCC in both training and validation cohorts. Among various machine learning algorithms tested, the NB model demonstrated robust diagnostic performance, with AUCs of 0.75 and 0.66 in training and validation sets, respectively. These findings not only identify novel diagnostic biomarkers for OSCC but also provide a foundation for developing non-invasive screening approaches, potentially facilitating earlier detection and improved clinical outcomes.

Firstly, this study combined multiple bioinformatics methods, including WGCNA, differential expression analysis and enrichment analysis. Previous studies that used bioinformatics to screen for TAAbs often relied

Models	AUC (95%CI)	Sensitivity (%)	Specificity (%)	Accuracy (%)	Youden's index	P
Training set						
LR	0.71(0.65–0.76)	52.9(45.2–60.5)	80.5(73.8–86.1)	66.7(61.4–71.6)	0.333(0.241–0.425)	0.393
SVM	0.83(0.79–0.87)	70.7(63.3–77.3)	83.9(77.6–89.0)	77.3(72.5–81.6)	0.546(0.465–0.632)	0.012
RF	0.72(0.67–0.78)	54.0(46.3–61.6)	81.0(74.4–86.6)	67.5(62.3–72.4)	0.351(0.259–0.448)	0.326
NB	0.75(0.70–0.80)	55.7(48.0-63.3)	81.0(74.4–86.6)	68.4(63.2–73.2)	0.368(0.276–0.460)	0.095
NN	0.82(0.78–0.86)	70.7(63.3–77.3)	80.5(73.8–86.1)	75.6(70.7–80.0)	0.511(0.420–0.603)	0.040
LDA	0.71(0.65–0.76)	54.0(46.3–61.6)	80.5(73.8–86.1)	67.2(62.0-72.2)	0.345(0.253–0.443)	0.408
FDA	0.69(0.64–0.75)	42.0(34.5–49.7)	89.1(83.5–93.3)	65.5(60.3–70.5)	0.310(0.224–0.397)	0.506
MDA	0.73(0.68–0.79)	58.6(50.9–66.0)	86.8(80.8–91.4)	72.7(67.7–77.3)	0.454(0.362–0.546)	0.042
Validation set						
LR	0.66(0.57–0.75)	48.6(36.9–60.6)	86.5(76.5–93.3)	67.6(59.4–75.0)	0.351(0.216–0.473)	
SVM	0.71(0.63–0.79)	50.0(38.1–61.9)	82.4(71.8–90.3)	66.2(58.0-73.8)	0.324(0.189–0.446)	
RF	0.67(0.58–0.76)	51.4(39.4–63.1)	83.8(73.4–91.3)	67.6(59.4–75.0)	0.351(0.216–0.486)	
NB	0.66(0.57–0.75)	54.1(42.1–65.7)	78.4(67.3–87.1)	66.2(58.0-73.8)	0.324(0.176–0.460)	
NN	0.72(0.64–0.80)	55.4(43.4–67.0)	82.4(71.8–90.3)	68.9(60.8–76.3)	0.378(0.243–0.514)	
LDA	0.66(0.57–0.75)	48.6(36.9–60.6)	86.5(76.5–93.3)	67.6(59.4–75.0)	0.351(0.216–0.473)	
FDA	0.65(0.57–0.74)	39.2(28.0-51.2)	87.8(78.2–94.3)	63.5(55.2–71.3)	0.270(0.135–0.405)	
MDA	0.62(0.53–0.72)	48.6(36.9–60.6)	75.7(64.3–84.9)	62.2(53.8–70.0)	0.243(0.095–0.378)	

Table 2. Diagnostic performance of the 8 machine learning algorithms in the training and validation sets. AUC, area under the receiver operating characteristic curve; LR, logistic regression; SVM, support vector machine; RF, random forest; NB, Naive Bayes; NN, neural networks; LDA, linear discriminant analysis; FDA, flexible discriminant analysis; MDA, Mixture discriminant analysis.

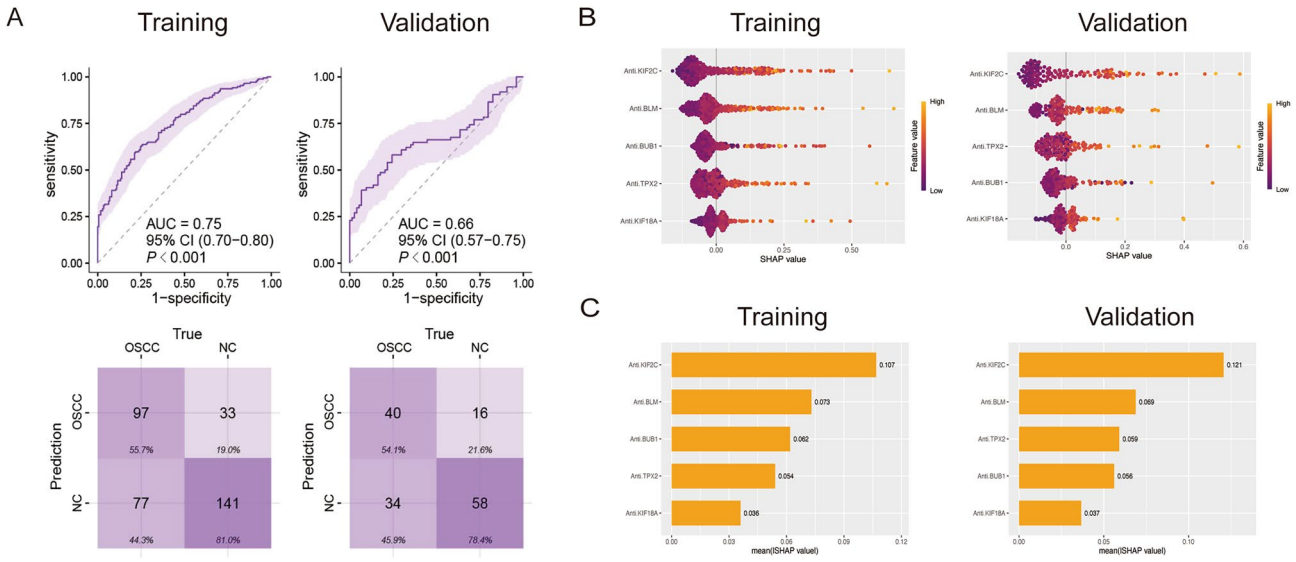

Fig. 4. Diagnostic performance and SHAP analysis of the Naive Bayes model. (**A**) ROC curves and confusion matrices in training and validation sets; (**B**) SHAP summary plot of each TAAb in training and validation sets; (**C**) SHAP bar plot of mean SHAP values for each TAAb in training and validation sets. SHAP: Shapley Additive exPlanations; AUC, area under the ROC curve; CI, confidence interval; ROC: Receiver operating characteristic; TAAb: tumor-associated autoantibody; OSCC, oral squamous cell carcinoma; NC, normal control.

on differential expression analysis to identify target genes[16,20]. This study utilized WGCNA, a bioinformatics method commonly used for identifying disease biomarkers, to analyze data from the TCGA database. It clusters genes based on their expression patterns from a global perspective and constructs co-expression networks to reveal functional modules associated with phenotypes[21,22]. WGCNA has been widely applied in cancer biomarker screening, such as pancreatic cancer[23], lung cancer[24], and gastric cancer[25]. Compared to differential expression analysis, WGCNA integrates gene expression with clinical features to more accurately identify key biomarkers related to diseases and focuses on gene interaction networks[26], which might be closely related to

Subgroups	*N*	AUC (95%CI)	Sensitivity (%)	Specificity (%)	Accuracy (%)	Youden's index	*P*
Training set							
Differentiation							0.766
Low	14	0.72(0.55–0.89)	50.0	86.2	83.5	0.362	
Moderate and High	124	0.74(0.69–0.80)	53.2	81.0	69.5	0.343	
TNM stage							0.962
Ⅰ-Ⅱ	84	0.75(0.68–0.82)	54.8	84.5	74.8	0.392	
Ⅲ-Ⅳ	89	0.75(0.69–0.82)	55.1	81.0	72.2	0.361	
Lymphatic metastatic							0.675
Yes	48	0.77(0.68–0.85)	58.3	81.0	76.1	0.394	
No	125	0.75(0.69–0.80)	55.2	81.0	70.2	0.362	
Validation set							
Differentiation							0.245
Low	9	0.80(0.55-1.00)	66.7	100.0	96.4	0.667	
Moderate and High	50	0.64(0.53–0.74)	34.0	93.2	69.4	0.272	
TNM stage							0.904
Ⅰ-Ⅱ	30	0.65(0.52–0.78)	33.3	97.3	78.8	0.306	
Ⅲ-Ⅳ	43	0.66(0.55–0.77)	53.5	82.4	71.8	0.359	
Lymphatic metastatic							0.198
Yes	24	0.73(0.60–0.87)	54.2	93.2	83.7	0.474	
No	49	0.62(0.51–0.73)	30.6	95.9	69.9	0.266	

Table 3. Diagnostic performance of the Naive Bayes model across clinical subgroups. ROC, receiver operating characteristic; AUC, area under the ROC curve; CI, Confidence interval; TNM, tumor-node-metastasis.

disease development. Furthermore, WGCNA employs topological algorithms to identify network topological structures, which is valuable for uncovering disease regulatory mechanisms[27]. Subsequently, to ensure the reliability of the included genes, this study performed differential expression analysis on the OSCC dataset selected from GEO database. At the same time, previous similar studies primarily focused on mRNA levels[16,20], whereas this study further validated candidate genes by analyzing their mRNA and corresponding protein expression levels using the CPTAC database, ensuring transcriptional-translational consistency to improve the reliability of downstream assays.

Autoantibodies were regarded as immunological biomarkers of aberrant cellular mechanisms existing during carcinogenesis. However, the specific mechanisms underlying the autoantibody response remain to be elucidated. In our study, the panel involved autoantibodies against BLM, BUB1, KIF18A, KIF2C, and TPX2, with significant differences observed in the levels of these autoantibodies. KIF2C had the highest contribution in the NB model. KIF2C/MCAK was involved in the regulation of microtubule dynamics and chromosome movement during mitosis by generating mechanical forces through microtubule-dependent ATP hydrolysis[28]. Overexpression of KIF2C leaded to excessive premature separation of spindle and sister chromatid[28], suggesting that cell cycle disorders might contribute to the development of cancer. Studies had shown that KIF2C was also highly expressed in all subtypes of breast cancer, suggesting that KIF2C might be a potential biomarker for breast cancer prognosis and immunotherapy[29]. The protein encoded by BLM is a DNA helicase of the RecQ family, and mutations in this gene are linked to Bloom syndrome, a condition that causes genetic instability and elevates the risk of cancer in Bloom syndrome patients[30]. Overexpression of BLM was linked to breast and ovarian cancers[31]. BUB1, a critical serine/threonine kinase in mitosis, was found to be abnormally expressed in several malignancies, including pancreatic cancer, gastric cancer and so on, suggesting its involvement in tumorigenesis[32,33]. A kinesin-driving protein called KIF18A was overexpressed in a number of cancers, but it was especially connected with tumor grade, migration, and prognosis in pancreatic, colorectal, and head and neck squamous cell carcinomas[34–36]. This suggested that KIF18A could be used as a diagnostic marker. TPX2 was highly expressed in prostate, colorectal, and pancreatic cancers, indicating its role in tumor development and poor prognosis, and its inhibition in oral squamous cell carcinoma presented a potential therapeutic strategy by regulating the Hh signaling pathway[37,38]. In summary, BLM, BUB1, KIF18A, KIF2C and TPX2 were all closely associated with cancer development, tended to be highly expressed in cancer, and were mostly associated with poor prognosis.

Considering the limitations of single TAAbs in terms of low diagnostic sensitivity, the combined detection of a set of TAAbs could effectively improve sensitivity while ensuring diagnostic specificity[19]. A previous study on oral cancer used logistic regression to construct an TAAbs diagnostic model but did not compare it with other machine learning methods[16]. Therefore, this study employed multiple machine learning methods for modeling and, after considering both the diagnostic value and stability of the models, the NB model was selected as the optimal model for diagnosing OSCC. In addition to its performance, the NB model offers advantages in interpretability and robustness, which enhance its feasibility for clinical application, particularly under limited sample sizes. Although many machine learning models can provide high accuracy in predictions, they are often viewed as "black boxes", making it difficult to understand how the model makes predictions based on

input features[39]. SHAP, by introducing Shapley value-based interpretation methods, provides transparency for complex models, which quantifies the contribution of each feature to the model's prediction[39,40]. By calculating the SHAP values for each feature, it becomes clear which TAAb expression level has the greatest impact on the diagnosis of oral cancer, making the model interpretable. This not only helps identify key biomarkers but also provides directions for subsequent biological research.

Although the sensitivity of the NB model in our study was approximately 55% with specificity around 80%, it is useful to contextualize these values by reported sensitivities and specificities of conventional serum tumor markers, which generally exhibit moderate sensitivity and relatively high specificity. For example, alpha-fetoprotein (AFP) for hepatocellular carcinoma at a cutoff of 20 μg/L shows a sensitivity of 41%–65% with specificity ranging from 80% to 94%[41]. Similarly, ovarian tumor markers such as HE4, CA125, CA19-9, and CEA exhibit sensitivities of 36%–64% with specificities between 71% and 97%[42]. Moreover, at commonly used cutoffs, SCCAg, ferritin, and CEA showed sensitivities of approximately 73%, 82%, and 66%, with corresponding specificities of 68%, 41%, and 61%, respectively[43]. These observations indicate that, in early cancer detection, serum biomarkers rarely achieve both high sensitivity and specificity simultaneously. Instead, a sensitivity of approximately 50%–70% combined with a specificity of ≥ 80% is generally considered clinically meaningful, particularly for non-invasive screening or adjunctive diagnostic applications. Prioritizing specificity while maintaining moderate sensitivity represents a practical compromise, balancing the risk of false positives with clinical utility. Overall, these results indicate that the current NB model has moderate diagnostic performance and may be more suitable for high-risk populations rather than for population screening. Further optimization is needed to improve its clinical utility.

There are a few limitations in our study. First, during the validation of the CPTAC database, samples of head and neck cancer tissues were used to validate the mRNA level expression and protein abundance of the screened OSCC candidate genes due to the limitations of the cancer types in the database itself. The expanded scope of validation may have led to the inclusion of imprecise candidate TAAs. Second, the specificity of the five identified autoantibodies as early non-invasive diagnostic markers for OSCC remains uncertain, as their expression in other malignancies was not assessed. In addition, this single-center, hospital-based design may introduce selection bias and limit external generalizability, highlighting the need for future multicenter studies with independent validation cohorts. In conclusion, five TAAbs were identified as promising diagnostic biomarkers for OSCC, and the NB model based on these TAAbs exhibited reliable diagnostic accuracy, with anti-KIF2C being the most influential marker.

Method

Serum samples and research design

In this study, 248 serum samples of OSCC patients were obtained from June 2021 to August 2023 from the Affiliated Hospital of Zhengzhou University. Inclusion criteria were: (1) newly diagnosed OSCC; (2) no prior surgical, radiotherapy, or chemotherapy treatment before blood collection; and (3) histopathologically confirmed squamous cell carcinoma. 248 serum samples from healthy controls were obtained from healthy people who had a physical examination at this hospital. The 248 controls were free from any cancer, gastrointestinal-related diseases and autoimmune diseases. To reduce potential confounding, cases and controls were frequency-matched by age (± 2 years) and sex.

Fasting blood samples were collected from all participants. After collection, samples were centrifuged at 3,000 rpm for 10 min at room temperature (approximately 26 °C) to separate the serum. The supernatant was aliquoted into 1.5-mL enzyme-free tubes (each 60 μL) and stored at − 80 °C. Each aliquot was thawed only once for analysis, avoiding repeated freeze–thaw cycles. All participants provided written informed consent, and the study was approved by the Medical Ethics Committee of Zhengzhou University. All experiments were performed in accordance with relevant named guidelines and regulations. A total of 496 participants (248 cases and 248 controls) were included in this study. The cases and controls were randomly divided into training and validation sets at a ratio of 7:3.

Screening of candidate TAAs

Candidate TAAs were screened through integrative bioinformatics analyses. Weighted Gene Co-expression Network Analysis (WGCNA) was performed on gene expression and clinical data of head and neck squamous cell carcinoma retrieved from the TCGA database via the UCSC Xena platform. After screening for OSCC-specific sites, 328 OSCC tissue samples and 32 adjacent non-tumor tissue samples were included. Co-expression modules significantly associated with OSCC were identified, and hub genes were selected based on network topology using the CytoHubba plugin in Cytoscape.

In parallel, differential expression analysis was conducted on OSCC-related datasets (GSE30784 and GSE25099) from the GEO database using the limma package in R, and *P* values were adjusted for multiple comparisons using the Benjamini–Hochberg false discovery rate method. Up-regulated DEGs were defined as those with $\log_2$fold change ($\log_2$FC) > 1 and adjusted $P < 0.05$. Candidate genes were identified by intersecting hub genes from WGCNA with up-regulated DEGs from GEO datasets. Finally, Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses were performed to explore the biological functions and pathways associated with the candidate genes.

Additionally, the mRNA expression levels and corresponding protein expression levels of these candidate genes were further examined using the Cancer Proteogenomic Data Analysis Site (cProSite). The mRNA expression analysis was conducted based on 110 head and neck cancer samples and adjacent normal tissues. Proteins encoded by these validated candidate genes were used as candidate TAAs.

Enzyme-linked immunosorbent assay

Serum levels of 12 candidate TAAbs were quantified using an indirect ELISA protocol, as previously described[44,45]. Recombinant proteins (CDC6, FOXM1, PLK1, BLM, CDC20, KIF4A, TPX2, KIF23, KIF18A, BUB1, ANLN, and KIF2C) were purchased from CUSABIO, PROTEINTECH, and CLOUD-CLONE (Wuhan, China). Protein concentration, purity, and molecular weight were confirmed by SDS-PAGE, confirming expected molecular weight and >85% purity. Microplates were coated with each antigen in carbonate buffer at optimized concentrations (0.125–0.25 μg/mL) and incubated overnight at 4 °C. After washing with Phosphate-Buffered Saline with Tween (PBST), plates were blocked with 2% Bovine Serum Albumin (BSA) in PBST for 2 h at 37 °C. Serum samples, diluted 1:100 in 1% BSA-PBST, were added (50 μL/well) and incubated for 1 h at 37 °C, followed by horseradish peroxidase (HRP)-conjugated secondary antibody (1:5000) for 1 h at 37 °C. Color development was achieved using TMB substrate solution, and the reaction was stopped with 10% H_2SO_4. Absorbance was measured at 450 nm with 620 nm as reference, and net optical density (OD) values were calculated (OD_{450} – OD_{620}).

The ELISA detection of autoantibodies and subsequent data processing were performed by two researchers (Lijuan Xu and Weihong Xie) who were blinded to sample status. To minimize plate-to-plate variation, serum samples from patients with OSCC and healthy controls were simultaneously detected on each ELISA plate. Each plate included 2 blank wells and 6 quality control (QC) wells using the same serum. Plates with blank OD > 0.1 were repeated[46]. The CV was calculated based on six QC replicates on each plate. When the CV exceeded 10%, the most extreme QC replicate was identified according to its deviation from the mean and removed prior to recalculation. The process was repeated until the CV fell below 10%. No more than one QC replicate was removed per plate. To control plate-to-plate variability, OD values were normalized by multiplying each plate's data by the ratio between the overall QC mean and the plate-specific QC mean. Background-corrected OD values were used for final analysis.

Statistical analysis

Statistical analyses and data visualization were performed using IBM SPSS Statistics version 26.0 and R version 4.3.1. Sample size calculation was conducted using PASS version 15.0. The Mann-Whitney U test was used to compare differences between groups. The diagnostic performance of autoantibodies was evaluated using receiver operating characteristic (ROC) curve analysis. The optimal cutoff value was determined in the training set by maximizing Youden's index under the constraint of a specificity ≥ 80%. This cutoff was then applied unchanged to the validation set to evaluate model performance. The evaluation indicators included area under the ROC curve (AUC), sensitivity, specificity, YI, and accuracy. Eight diagnostic models, including logistic regression (LR), support vector machine (SVM), random forest (RF), neural network (NN), Naive Bayes (NB), linear discriminant analysis (LDA), mixture discriminant analysis (MDA), and flexible discriminant analysis (FDA) were constructed using the training set and validated in the validation set. Each model's performance was assessed through 100 iterations of 10-fold cross-validation to ensure reliable results. The Delong test was used to compare AUCs. Additionally, Shapley Additive exPlanations (SHAP) values were calculated for the most effective prediction model using the fastshap R package, with 100 Monte Carlo repeats. The fastshap package is a model-agnostic tool that can efficiently compute approximate SHAP values for any supervised learning model[47]. For the final selected model, learning curves and calibration curves were generated in the training and validation sets to assess model stability and calibration. A P value < 0.05 was considered statistically significant.

Data availability

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

Received: 28 October 2025; Accepted: 9 April 2026
Published online: 15 April 2026

References

1. Mody, M. D., Rocco, J. W., Yom, S. S., Haddad, R. I. & Saba, N. F. Head and neck cancer. *Lancet* **398**, 2289–2299. https://doi.org/10.1016/S0140-6736(21)01550-6 (2021).
2. Lee, C. C. et al. Suppression of oral cancer by induction of cell cycle arrest and apoptosis using Juniperus communis extract. *Biosci. Rep.* **40**, BSR20202083. https://doi.org/10.1042/BSR20202083 (2020).
3. Dhanuthai, K. et al. Oral cancer: A multicenter study. *Med. Oral Patol. Oral Cir. Bucal.* **23**, e23–e29. https://doi.org/10.4317/medoral.21999 (2018).
4. Bray, F. et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **74**, 229–263. https://doi.org/10.3322/caac.21834 (2024).
5. Gbd Lip, O. et al. The global, regional, and national burden of adult lip, oral, and pharyngeal cancer in 204 countries and territories: a systematic analysis for the gobal burden of disease study 2019. *JAMA Oncol.* **9**, 1401–1416. https://doi.org/10.1001/jamaoncol.2023.2960 (2023).
6. Matsuda, Y., Jayasinghe, R. D., Zhong, H., Arakawa, S. & Kanno, T. Oral health management and rehabilitation for patients with oral cancer: a narrative review. *Healthc. (Basel).* **10**, 960. https://doi.org/10.3390/healthcare10050960 (2022).
7. Burtness, B. et al. Pembrolizumab alone or with chemotherapy versus cetuximab with chemotherapy for recurrent or metastatic squamous cell carcinoma of the head and neck (KEYNOTE-048): a randomised, open-label, phase 3 study. *Lancet* **394**, 1915–1928. https://doi.org/10.1016/S0140-6736(19)32591-7 (2019).
8. Chu, P. Y., Hsu, N. C., Lin, S. H., Hou, M. F. & Yeh, K. T. High nuclear protein kinase CbetaII expression is a marker of disease recurrence in oral squamous cell carcinoma. *Anticancer Res.* **32**, 3987–3991 (2012).
9. Hashim, D., Genden, E., Posner, M., Hashibe, M. & Boffetta, P. Head and neck cancer prevention: from primary prevention to impact of clinicians on reducing burden. *Ann. Oncol.* **30**, 744–756. https://doi.org/10.1093/annonc/mdz084 (2019).
10. Sannam Khan, R., Khurshid, Z., Akhbar, S. & Faraz Moin, S. Advances of salivary proteomics in oral squamous cell carcinoma (OSCC) detection: an update. *Proteomes* **4**, 41. https://doi.org/10.3390/proteomes4040041 (2016).

11. Mercadante, V., Paderni, C. & Campisi, G. Novel non-invasive adjunctive techniques for early oral cancer diagnosis and oral lesions examination. *Curr. Pharm. Des.* **18**, 5442–5451. https://doi.org/10.2174/138161212803307626 (2012).
12. Macdonald, I. K., Parsy-Kowalska, C. B. & Chapman, C. J. Autoantibodies: opportunities for early cancer detection. *Trends Cancer.* **3**, 198–213. https://doi.org/10.1016/j.trecan.2017.02.003 (2017).
13. Tan, E. M. & Zhang, J. Autoantibodies to tumor-associated antigens: reporters from the immune system. *Immunol. Rev.* **222**, 328–340. https://doi.org/10.1111/j.1600-065X.2008.00611.x (2008).
14. Liu, W. et al. Autoantibodies to tumor-associated antigens as biomarkers in cancer immunodiagnosis. *Autoimmun. Rev.* **10**, 331–335. https://doi.org/10.1016/j.autrev.2010.12.002 (2011).
15. Soussi, T. p53 Antibodies in the sera of patients with various types of cancer: a review. *Cancer Res.* **60**, 1777–1788 (2000).
16. Xie, W. et al. Identification of novel tumor-associated antigens and evaluation of a panel of autoantibodies in detecting oral cancer. *BMC Cancer.* **23**, 802. https://doi.org/10.1186/s12885-023-11247-w (2023).
17. Abdulla, R., Devasia Puthenpurackal, J., Pinto, S. M., Rekha, P. D. & Subbannayya, Y. Serum autoantibody profiling of oral squamous cell carcinoma patients reveals NUBP2 as a potential diagnostic marker. *Front. Oncol.* **13**, 1167691. https://doi.org/10.3389/fonc.2023.1167691 (2023).
18. Khurshid, Z. et al. Role of salivary biomarkers in oral cancer detection. *Adv. Clin. Chem.* **86**, 23–70. https://doi.org/10.1016/bs.acc.2018.05.002 (2018).
19. Zhang, J. Y. et al. Enhancement of antibody detection in cancer using panel of recombinant tumor-associated antigens. *Cancer Epidemiol. Biomarkers Prev.* **12**, 136–143 (2003).
20. Li, Z. et al. Discovery and validation of novel biomarkers for detection of cervical cancer. *Cancer Med.* **10**, 2063–2074. https://doi.org/10.1002/cam4.3799 (2021).
21. Comunale, M. A. et al. Linkage specific fucosylation of alpha-1-antitrypsin in liver cirrhosis and cancer patients: implications for a biomarker of hepatocellular carcinoma. *PLoS One.* **5**, e12419. https://doi.org/10.1371/journal.pone.0012419 (2010).
22. Zhu, J. et al. Analysis of serum haptoglobin fucosylation in hepatocellular carcinoma and liver cirrhosis of different etiologies. *J. Proteome Res.* **13**, 2986–2997. https://doi.org/10.1021/pr500128t (2014).
23. Ye, H. et al. TSPAN1, TMPRSS4, SDR16C5, and CTSE as novel panel for pancreatic cancer: a bioinformatics analysis and experiments validation. *Front. Immunol.* **12**, 649551. https://doi.org/10.3389/fimmu.2021.649551 (2021).
24. Ding, M., Li, F., Wang, B., Chi, G. & Liu, H. A comprehensive analysis of WGCNA and serum metabolomics manifests the lung cancer-associated disordered glucose metabolism. *J. Cell. Biochem.* **120**, 10855–10863. https://doi.org/10.1002/jcb.28377 (2019).
25. Wei, G. et al. Identification of hub genes and construction of an mRNA-miRNA-lncRNA network of gastric carcinoma using integrated bioinformatics analysis. *PLoS One.* **16**, e0261728. https://doi.org/10.1371/journal.pone.0261728 (2021).
26. Chen, R. et al. Elevated level of anterior gradient-2 in pancreatic juice from patients with pre-malignant pancreatic neoplasia. *Mol. Cancer.* **9**, 149. https://doi.org/10.1186/1476-4598-9-149 (2010).
27. Jiao, Y. et al. Whole-exome sequencing of pancreatic neoplasms with acinar differentiation. *J. Pathol.* **232**, 428–435. https://doi.org/10.1002/path.4310 (2014).
28. Sanhaji, M., Friel, C. T., Wordeman, L., Louwen, F. & Yuan, J. Mitotic centromere-associated kinesin (MCAK): a potential cancer drug target. *Oncotarget* **2**, 935–947. https://doi.org/10.18632/oncotarget.416 (2011).
29. Liu, S. et al. KIF2C is a prognostic biomarker associated with immune cell infiltration in breast cancer. *BMC Cancer.* **23**, 307. https://doi.org/10.1186/s12885-023-10788-4 (2023).
30. Lee, J. et al. Bloom syndrome patients and mice display accelerated epigenetic aging. *Aging Cell.* **22**, e13964. https://doi.org/10.1111/acel.13964 (2023).
31. Foretova, L. et al. Recommendations for preventive care for women with rare genetic cause of breast and ovarian cancer. *Klin. Onkol.* **32**, 6–13. https://doi.org/10.14735/amko2019S6 (2019).
32. Li, X., He, J., Yu, M., Zhang, W. & Sun, J. [BUB1 gene is highly expressed in gastric cancer:analysis based on Oncomine database and bioinformatics]. *Nan Fang Yi Ke Da Xue Xue Bao.* **40**, 683–692. https://doi.org/10.12122/j.issn.1673-4254.2020.05.11 (2020).
33. Eyuboglu, S., Alpsoy, S., Uversky, V. N. & Coskuner-Weber, O. Key genes and pathways in the molecular landscape of pancreatic ductal adenocarcinoma: A bioinformatics and machine learning study. *Comput. Biol. Chem.* **113**, 108268. https://doi.org/10.1016/j.compbiolchem.2024.108268 (2024).
34. Zhang, C. et al. Kif18A is involved in human breast carcinogenesis. *Carcinogenesis* **31**, 1676–1684. https://doi.org/10.1093/carcin/bgq134 (2010).
35. Feng, G., Xue, F., He, Y., Wang, T. & Yuan, H. The identification of stemness-related genes in the risk of head and neck squamous cell carcinoma. *Front. Oncol.* **11**, 688545. https://doi.org/10.3389/fonc.2021.688545 (2021).
36. Nan, K. et al. Integrated profiling delineated KIF18A as a significant biomarker associated with both prognostic outcomes and immune response in pancreatic cancer. *Immunotargets Ther.* **14**, 123–138. https://doi.org/10.2147/ITT.S497284 (2025).
37. Ma, S. et al. CEACAM1 as a molecular target in oral cancer. *Aging (Albany NY).* **15**, 8137–8154. https://doi.org/10.18632/aging.204960 (2023).
38. Shao, T., Jiang, X., Bao, G., Li, C. & Guo, C. Comprehensive analysis of the oncogenic role of targeting protein for Xklp2 (TPX2) in human malignancies. *Dis. Markers* **2022** (7571066). https://doi.org/10.1155/2022/7571066 (2022).
39. Wang, K. et al. Interpretable prediction of 3-year all-cause mortality in patients with heart failure caused by coronary heart disease based on machine learning and SHAP. *Comput. Biol. Med.* **137**, 104813. https://doi.org/10.1016/j.compbiomed.2021.104813 (2021).
40. Lundberg, S. M. et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.* **2**, 749–760. https://doi.org/10.1038/s41551-018-0304-0 (2018).
41. Gupta, S., Bent, S. & Kohlwes, J. Test characteristics of alpha-fetoprotein for detecting hepatocellular carcinoma in patients with hepatitis C. A systematic review and critical analysis. *Ann. Intern. Med.* **139**, 46–50. https://doi.org/10.7326/0003-4819-139-1-200307010-00012 (2003).
42. Chen, F., Shen, J., Wang, J., Cai, P. & Huang, Y. Clinical analysis of four serum tumor markers in 458 patients with ovarian tumors: diagnostic value of the combined use of HE4, CA125, CA19-9, and CEA in ovarian tumors. *Cancer Manag Res.* **10**, 1313–1318. https://doi.org/10.2147/CMAR.S155693 (2018).
43. Yuan, C., Yang, K., Tang, H. & Chen, D. Diagnostic values of serum tumor markers Cyfra21-1, SCCAg, ferritin, CEA, CA19-9, and AFP in oral/oropharyngeal squamous cell carcinoma. *Onco Targets Ther.* **9**, 3381–3386. https://doi.org/10.2147/OTT.S105672 (2016).
44. Ma, Y. et al. Using protein microarray to identify and evaluate autoantibodies to tumor-associated antigens in ovarian cancer. *Cancer Sci.* **112**, 537–549. https://doi.org/10.1111/cas.14732 (2021).
45. Xu, Y. W. et al. Autoantibodies as potential biomarkers for the early detection of esophageal squamous cell carcinoma. *Am. J. Gastroenterol.* **109**, 36–45. https://doi.org/10.1038/ajg.2013.384 (2014).
46. Porchas-Quijada, M. et al. IgG Anti-ghrelin Immune Complexes Are Increased in Rheumatoid Arthritis Patients Under Biologic Therapy and Are Related to Clinical and Metabolic Markers. *Front. Endocrinol. (Lausanne).* **10**, 252. https://doi.org/10.3389/fendo.2019.00252 (2019).
47. Wang, E., Shuryak, I. & Brenner, D. J. A competing risks machine learning study of neutron dose, fractionation, age, and sex effects on mortality in 21,000 mice. *Sci. Rep.* **14**, 17974. https://doi.org/10.1038/s41598-024-68717-9 (2024).

Author contributions

L.X. and W.X. contributed to the original draft writing, methodology, data curation, and formal analysis. Y.Z., Z.Z., X.Z., Q.Y., Y.H., Y.L., and M.L. contributed to data curation and writing—review and editing. H.Y. contributed to writing—review and editing. P.W. contributed to conceptualization, supervision, and writing—review and editing. All authors read and approved the final manuscript.

Funding

This work was supported by the Project of Tackling Key Problems in Science and Technology of Henan Province (grant number 252102311002), and Zhengzhou Major Project for Collaborative Innovation (grant number 18XTZX12007).

Declarations

Competing interests

The authors declare no competing interests.

Ethics approval and consent to participants

The study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of Zhengzhou University. Written informed consent was obtained from all participants prior to inclusion in the study.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-48821-8.

Correspondence and requests for materials should be addressed to P.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.