BMC Cancer

RESEARCH

Open Access

# Development and validation of a novel artificial intelligence algorithm for precise prediction the postoperative prognosis of esophageal squamous cell carcinoma

Zichen Wang[1†], Zhihan Xiao[2†], Tongyu Zhang[1], Meiyou Lu[1], Hai Li[3], Jing Cao[1], Jianan Zheng[1], Yichan Zhou[4], Juncheng Dai[5], Cheng Wang[5], Liang Chen[1] and Jing Xu[1*]

## Abstract

**Background** Esophageal squamous cell carcinoma (ESCC) is a highly aggressive malignancy, and current postoperative prognostic assessment methods remain unsatisfactory, underlining the urgent to develop a reliable approach for precision medicine. Given the similarities with gametogenesis, cancer/testis genes (CTGs) are acknowledged for regulation unrestrained multiplication and immune microenvironment during oncogenic processes. These processes are associated with advanced disease and poorer prognosis, indicating that CTGs could serve as ideal prognostic biomarkers in ESCC. The purpose of this study is to develop a novel clinically prognostic prediction system to facilitate the individualized postoperative care.

**Methods** We conducted LASSO regression analysis of protein-coding CTGs and clinical characteristics from 119 pathologically confirmed ESCC patients to recognize powerful predictive variables. We employed nine supervised machine learning classifiers and integrated best predictive machine learning classifiers by weighted voting method to construct an ensemble model called PPMESCC. Additionally, functional assay was conducted to examine the potential effect of top-ranking CTG HENMT1 in ESCC.

**Results** LASSO regression identified five CTGs and TNM stage as optimized prognostic features. Six machine learning classifiers were integrated to construct an ensemble model, PPMESCC, which exhibited outstanding performance in ESCC prediction. The AUC for PPMESCC was 0.9828 (95% confidence interval: 0.9608 to 0.9926), with an accuracy of 98.32% (95% CI: 96.64–99.16%) in the discovery cohort and 0.9057 (95% CI: 0.8897 to 0.9583) of AUC with an accuracy of 90% (95% CI: 89.08–93.28%) in validation cohort. In addition, the top-ranking CTG *HENMT1* encodes 2′-O-methyltransferase of piRNAs that was confirmed positively correlated with the proliferation capacity of ESCC cells. Then we systematically screen piRNAs associated with esophageal carcinoma based on GWAS, eQTL-piRNA, and i2OM databases, and successfully discovered 8 piRNAs potentially regulated by HENMT1.

†Zichen Wang and Zhihan Xiao contributed equally to this work.

*Correspondence:
Jing Xu
jingxu@njmu.edu.cn

Full list of author information is available at the end of the article

**Conclusion** The study highlights the clinical utility of PPMESCC algorithm in prognostic prediction that may facilitate to establish the personalized screening and management strategies for postoperative ESCC patients.

**Keywords** Esophageal squamous cell carcinoma, Artificial intelligence, Machine learning, Postoperative prognosis, Cancer/testis gene

## Background

Esophageal cancer (EC) is one of the most common cancers in the world, with greater than 600,000 new cases, and 540,000 deaths occurred annually worldwide [1]. In China, esophageal squamous cell carcinoma (ESCC) accounts for approximately 95% of EC patients and bears over half of the global burden [2]. Despite recent advances in surgical techniques and multimodal adjuvant therapies, the aggressive and heterogeneous nature of ESCC continues to seriously threaten patients' health [3], with the 5-year survival rate ranged from 15 to 25% [4]. Therefore, the establishment of an accurate individualized prediction system for ESCC prognosis will enable clinicians to make well-informed decisions regarding patient counseling, personalized surveillance, and selecting suitable postoperative adjuvant therapy.

Conventionally, the American Joint Commission on Cancer (AJCC) TNM staging classification is considered the primary determinant for prognostic prediction and treatment decision-making in ESCC [5]. Nevertheless, it has been observed that patients with comparable AJCC TNM staging may have divergent outcomes [6], and the variations are largely attributable to biological heterogeneity. Hence, constructing an integrated prediction model that incorporates molecular factors and clinical features is of paramount importance and received considerable attention. Recently, Tan et al. introduced a FENSAM-staging system that combines nine biomarkers and thirteen clinical characteristics [7]. Unfortunately, despite offering a relatively simplified framework for clinical application, this innovative approach merely exhibited similar predictive efficacy to the traditional AJCC TNM staging system. Future endeavors should focus on investigating the appropriate prognostic biomarkers that contribute to develop a model with enhanced predictive capability.

Cancer/testis genes (CTGs) are a group of genes that are typically expressed in the testes but can also aberrantly expressed in various types of cancers. With the functional commonalities in gametogenesis, it is believed that the abnormal expression of CTGs in cancer cells contribute to tumorigenesis by promoting cell proliferation, inhibiting apoptosis, and enhancing tumor cell survival [8]. Studies have shown that the upregulation of CTGs are correlated with adverse clinical outcomes of ESCC. For example, Chen et al. analyzed the expression of CTGs in ESCC patients and found that high expression of MAGE-A1, GAGE1, and SP17 were significantly associated with shorter survival and higher risk of recurrence. The presence of CTGs expression in ESCC tumors may indicate a more aggressive phenotype and resistance to the treatment [9]. In our prior investigation, we conducted a comprehensive evaluation of CTGs in ESCC and successfully identified ESCC specific protein-coding CTGs using transcriptomics data from multiple independent databases [10], which will serve as valuable biomarkers available for predicting survival of postoperative ESCC patients.

In the past, prognostic prediction models were predominantly constructed using univariate or multivariate statistical methodologies, such as Cox regression or logistic regression. The underlying linearity assumptions of these methodologies pose a considerable challenge in exploring nonlinear relationships between variables in the real world, resulting in hindering identification the optimal performing model [11, 12]. Machine learning (ML), as a subset of artificial intelligence (AI), enables computers to learn from data, identify patterns, and make predictions or take actions based on acquired knowledge. ML techniques manifest a potential solution to address the limitations of current analytical methods, which can effectively handle multidimensional variables, identify non-linear relationships between features and outcomes, and develop prediction models with improved accuracy and efficiency [13, 14]. In 2021, Abuhelwa et al. devised five ML models for prognosticating the survival outcomes of urothelial cancer patients treated with atezolizumab, an immune checkpoint inhibitor. Their results highlighted that the Gradient-boosted model exhibited superior performance compared to the other evaluated models [15]. Until now, few studies have reported prognostic model of ESCC with superior predictive efficiency, particularly for postoperative patients. The obstacle motivates us to implement AI procedures for creation an innovative ESCC prognostic model.

In this study, we aimed to establish and verify a novel ESCC-specific postoperative prognostic prediction system by leveraging CTGs and clinical variables through ML algorithms, which may contribute to personalized therapy and ultimately prolonging the lifespan of ESCC patients.

## Methods

### Participants

The study flowchart was shown in Fig. 1. We included two independent cohorts of patients with ESCC from

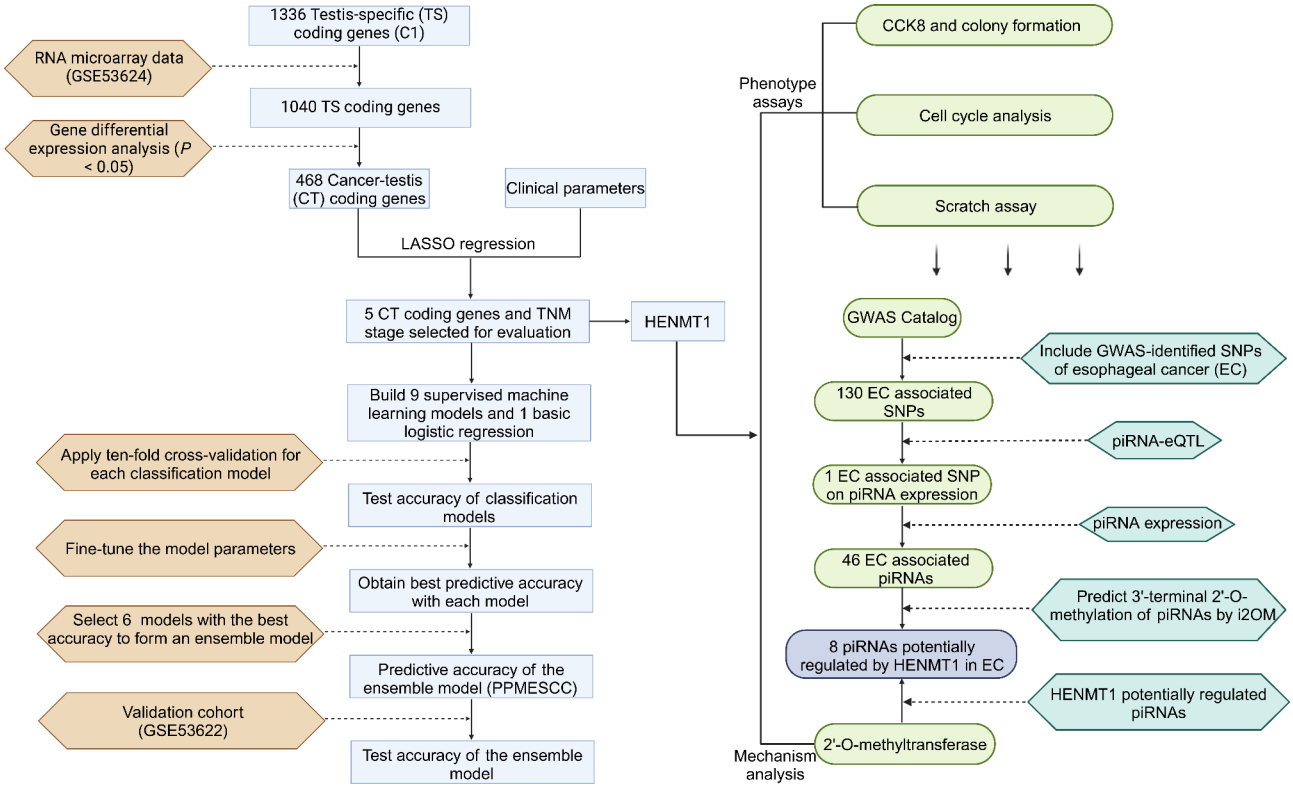


**Fig. 1** Flowchart of construction PPMESCC model and the functional exploration of HENMT1

the Gene Expression Omnibus database, including GSE53624 consisting of 119 patients and GSE53622 comprising 60 patients. 119 patients were assigned to the discovery cohort and 60 patients were assigned to the validation cohort. All patients had surgically confirmed primary ESCC and underwent esophagectomy (R0 resection), with follow-up data available. Individuals didn't receive chemotherapy or radiotherapy before tumor removal. Informed consent was obtained from the participants, and provide clinical and pathological information. In the current study, we employed 12 clinicopathological parameters including age at diagnosis, sex, tobacco use, alcohol use, tumor location, tumor grade, T stage, N stage, TNM stage, postoperative arrhythmia, pneumonia and adjuvant therapy. The study received approval from The First Affiliated Hospital of Nanjing Medical University (2022-SR-055).

### Feature selection

In the present study, we combined the RNA sequencing data from 119 ESCC samples with C1 genes discovered by our previous work [16] to systematically explore CT coding genes with ESCC tissue-specific expression. Ultimately, a total of 468 ESCC-specific CTGs and 12 clinical characteristics were eligible for variable selection. To eliminate redundant collinear features and diminish cost of clinical testing, we performed the most predictive features selection by LASSO regression [17]. LASSO added the L1 norm of the feature coefficients as a penalty term to the loss function, which forced the coefficients corresponding to those weak features to become zero. In this context, we considered features that had coefficients equal to zero as redundant and eliminated them, resulting in 6 features being selected for model development.

### Model development

In this study, nine types of supervised ML classifiers, including AdaBoost, Gradient Boosting Decision Trees (GBDT), Bagging, Decision Trees (DT), Extra Trees (ET), Gaussian Naive Bayes (GaussianNB), K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Random Forest (RF), were assessed. We also utilized Logistic Regression (LR) model as the baseline. Classifiers were trained using repeated 10-fold cross-validation of discovery dataset, and their predictive performances were also evaluated in the validation dataset.

In recent years, clinical evidence has indicated that the prediction results of individual basic classifiers for certain samples may be inaccurate. To address this issue, ensemble ML methods are recommended to integrate multiple individual ML classifiers. This approach often outperforms simple class label combination and has been widely applied to address complex scientific problems [18]. The weighted voting method is a powerful ensemble

ML strategy designed to enhance model performance by combining the predictions of multiple basic classifiers with assigned weights [19]. In the weighted voting method, the number of basic models included and the weight assigned to each model are critical factors that determine the overall performance of the ensemble. The optimal number of models and their corresponding weights are typically determined through a systematic evaluation process, which explores all possible combinations for each ML classifier to achieve the best predictions. The prediction probability of each sample in the weighted voting method is calculated according to the following form.

$$H(x) = c_{\arg\max_j} \sum_{i=1}^{T} w_i h_i^j(x)$$

$$h_i^j(x) \in [0,1], w_i \in [0.1, 0.8],\ T = 9$$

where $H$(x) is the final prediction probability, $T$ is the number of basic models, $w_i$ is the weight, $h_i^j(x)$ is class probability.

Then, we trained nine base classifier models and traverse all weight combinations. The weight of each basic classification model is changed from 0.1 to 0.8 and the sum of their weights is guaranteed to be 1. The weight changes by 0.1 each time. Finally, an ensemble model derived from six base models of best predictive performance (AdaBoost, GBDT, Bagging, DT, ET, and RF), named PPMESCC, was developed.

#### Model evaluation

The predictive performance of the models was evaluated by ROC curve, Kaplan-Meier curve and evaluation metrics including area under the ROC curve (AUC), accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1 score and Cohen's Kappa coefficient (Kappa). Herein, we framed the mortality prediction task as a binary classification problem. We selected the threshold of 46 months to designate the label of mortality risk by optimizing F1 score (0.9804) on the discovery cohort (Fig. 2I). Survival of less than 46 months was assigned to poor prognosis and high risk, otherwise it was assigned to good prognosis and low risk across all included ML methods. 51/68 and 27/33 patients above and below the 46-month cutoff were observed in the discovery and validation cohorts, respectively.

#### Cell culture and transfection

The human ESCC cell lines KYSE410 and KYSE150 were purchased from Genechem Co. Ltd. (Shanghai, China) and cultured in RPMI-1640 medium (Gibco, USA) with 10% fetal bovine serum (FBS, Gibco, USA) and 1% penicillin G sodium/streptomycin sulphate in a humidified atmosphere consisting of 95% air and 5% $CO_2$ at 37 °C. The full-length of HENMT1 (HENMT1 OE) was amplified and cloned into pcDNA3.1 (Thermo Fisher, USA). The primers (RiboBio, Guangzhou, China) for human HENMT1 were as follow: F: 5′-CCAGAATGGAGTT TCAGACC-3′ and R: 5′-GATTCTGTTGCCTTTCCT CC-3′. The transfection of the plasmid was performed using Lipofectamine 3000 reagent (Invitrogen, USA).

#### CCK8 and colony formation

The viability of treated cells was measured by Cell Counting Kit-8 assay (CCK8, Dojindo, Japan). Briefly, $5 \times 10^4$ transfected cells were seeded in 96-well plate and incubated with 10 ml of CCK8 solution at 37 °C for 2 h. The absorbance was measured at wavelengths of 450 nm. For colony formation assay, $1 \times 10^3$ ESCC cells were cultured in six-well plates and fixed with methanol for 20 min at room temperature after 2 weeks. Cell colonies were stained with 0.5% crystal violet for 30 min. Images of all wells were captured and counted by hand with the aid of imaging software.

#### Cell cycle and scratch assay

The cell cycle was conducted by flow cytometry. Transfected ESCC cells were suspended in 75% ethanol overnight and centrifuged at 1,000 rpm. Following washed in cold PBS 48 h, 50 mg/ml propidium iodide (PI) and 100 g/ml DNase-free RNase A was added for 30 min at 37 °C. Cell cycle distribution was further analyzed with Cell Quest software (Becton Dickinson, San Jose, CA) and Mod Fit LT (Verity Software House, Topsham, ME). In scratch assay, ESCC cells were inoculated and covered with a layer in six-well plates. After serum starvation for 24 h, a 200 L plastic pipette tip was used to scratch the monolayer. The distance that cells had migrated was photographed by a digital camera under Inverted microscope (Olympus) at the same position at 0 and 48 h for later calculation. Image Pro Plus 6.0 software (Media Cybernetics, Bethesda, MD, USA) was used to measure and calculate the distance that the cells had migrated.

#### Identification of HENMT1 potentially regulated piRNAs in ESCC

We obtained esophageal carcinoma-related single nucleotide polymorphisms (SNPs) from the GWAS catalog (https://www.ebi.ac.uk/gwas/). Next, the piRNA-eQTL database (http://njmu-edu.cn:3838/piRNA-eQTL/) was utilized to examine the association between esophageal carcinoma-related SNPs and piRNA expression. The piRNA base database (http://bigdata.ibp.ac.cn/piRBase/) and i2OM (i2om.lin-group.cn) database were applied to evaluate the sequences of piRNAs and 3'-end 2′-O-methylation (2OM) regulatory sites, respectively.

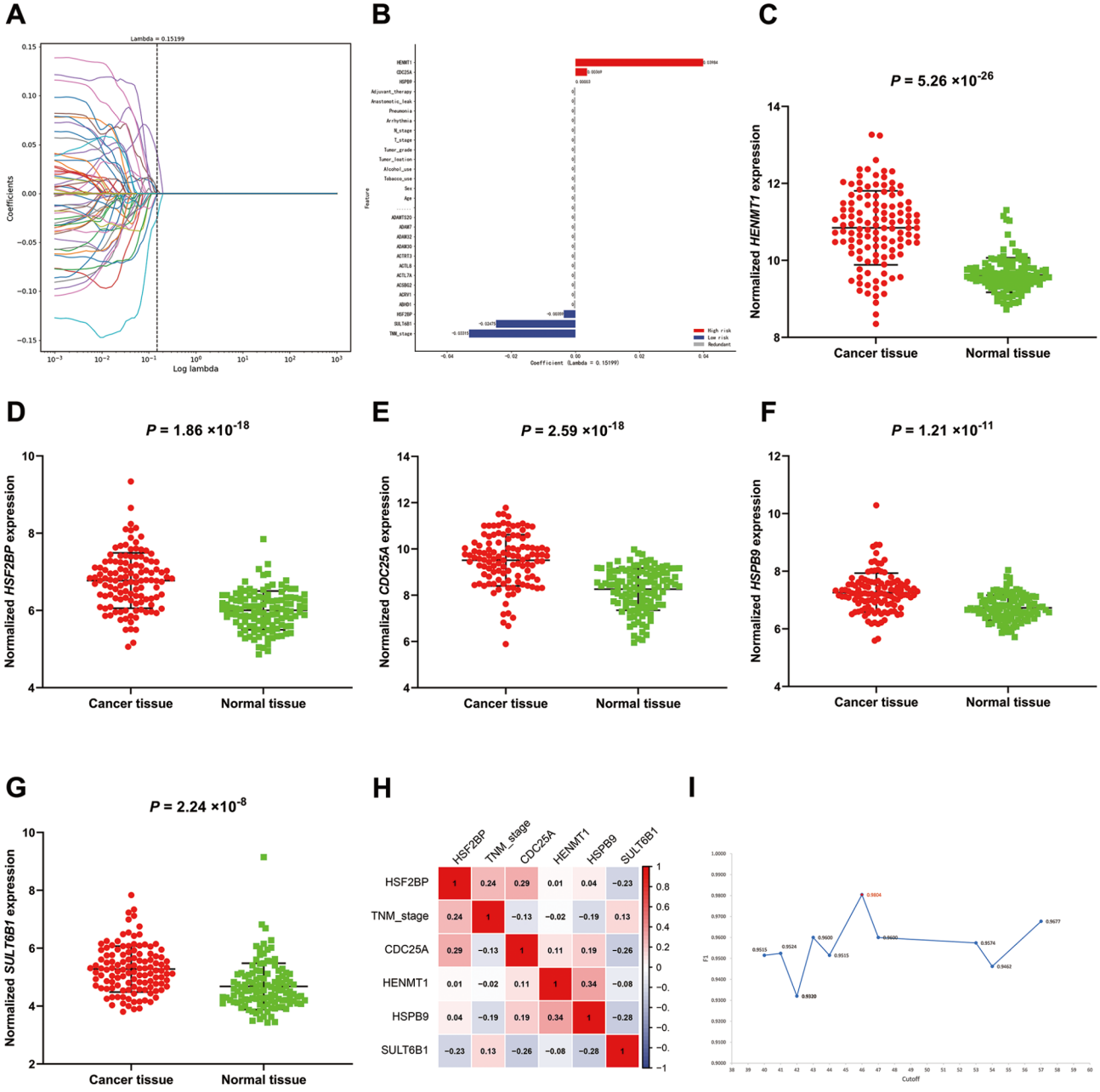


**Fig. 2** Feature selection and statistical analysis in the model. **A** LASSO variable trace profiles of the 468 CTGs and 12 clinical features. **B** Feature coefficient of LASSO with best lambda value 0.15199. Relatively high-risk and low-risk features are colored in red and blue, respectively. Gray features with coefficient 0 were considered redundant and removed. **C-G** Beeswarm plots representing the distribution of continuous features included in PPMESCC between ESCC tumor tissues and adjacent normal tissues. **H** Heatmap illustrating the correlation between continuous features included in PPMESCC using Spearman's correlation coefficient. **I** F1 score of PPMESCC on the discovery cohort in respect to different cutoff value. Each dot point indicates the corresponding F1 score of the OS. Dot colored with red indicates the highest F1 score (0.9804) with cutoff at 46 months

### Statistical analysis

Statistical analysis was performed in R (version 3.6.2). Comparison of continuous variables was achieved by the Mann–Whitney U test using R-package Table 1. Odds ratio and corresponding 95% CI from LR were calculated with R-package stats. Univariate and multivariate Cox regression was utilized to calculate the hazard ratio (HR) with R-package survival. The ROC curve and AUC analysis were conducted with R pROC package. Accuracy, sensitivity, specificity, PPV, NPV, Kappa, and F1 score were calculated with R caret and epiR packages. Survival curves were developed by Kaplan-Meier method with log-rank test, and plotted with R-package survival and survminer. The significance level was set at a two-sided *p* value below 0.05.

**Table 1** Clinical characteristics of ESCC patients from discovery and validation cohorts

| | Discovery cohort | | | Validation cohort | | |
|---|---|---|---|---|---|---|
| | 119 cases | | | 60 cases | | |
| | Overall (perc.) | HR (95% CI) | *P* value | Overall (perc.) | HR (95% CI) | *P* value |
| Age at diagnosis | | | | | | |
| ≤ 60y | 69(58.0) | 1(ref.) | | 30(50.0) | 1(ref.) | |
| > 60y | 50(42.0) | 5.11(2.01–12.98) | 0.001 | 30(50.0) | 3.85(1.36–10.89) | 0.011 |
| Sex | | | | | | |
| Male | 98(82.4) | 1(ref.) | | 48(80.0) | 1(ref.) | |
| Female | 21(17.6) | 8.80(1.52–50.86) | 0.015 | 12(20.0) | 1.23(0.40–3.73) | 0.719 |
| T stage | | | | | | |
| T1-2 | 28(23.5) | 1(ref.) | | 11(18.4) | 1(ref.) | |
| T3-4 | 91(76.5) | 0.97(0.28–3.38) | 0.956 | 49(81.7) | 2.99(0.55–16.39) | 0.206 |
| N stage | | | | | | |
| N0 | 54(45.4) | 1(ref.) | | 29(48.3) | 1(ref.) | |
| N1-3 | 65(54.6) | 0.79(0.21–2.99) | 0.732 | 31(51.7) | 4.84(0.569–41.13) | 0.149 |
| TNM stage | | | | | | |
| I-II | 53(44.5) | 1(ref.) | | 34(56.7) | 1(ref.) | |
| III-IV | 66(55.5) | 1.06(0.30–3.78) | 0.933 | 26(43.3) | 0.32(0.03–3.25) | 0.336 |
| Tumor grade | | | | | | |
| Well | 32(26.9) | 1(ref.) | | 17(28.3) | 1(ref.) | |
| Moderately | 64(53.8) | 2.33(0.79–6.90) | 0.127 | 34(56.7) | 0.81(0.20–3.34) | 0.776 |
| Poorly | 23(19.3) | 1.54(0.43–5.43) | 0.506 | 9(15.0) | 1.13(0.30–4.71) | 0.872 |
| Tumor location | | | | | | |
| Upper | 14(11.8) | 1(ref.) | | 6(10.0) | 1(ref.) | |
| Middle | 69(58.0) | 0.50(0.05–5.24) | 0.559 | 28(46.7) | 0.37(0.09–1.49) | 0.161 |
| Lower | 36(30.3) | 1.61(0.16–16.12) | 0.686 | 26(43.3) | 0.11(0.02–0.48) | 0.004 |
| Arrhythmia | | | | | | |
| Yes | 27(22.7) | 1(ref.) | | 16(26.7) | 1(ref.) | |
| No | 92(77.3) | 1.05(0.32–3.48) | 0.933 | 44(73.3) | 1.43(0.52–3.98) | 0.492 |
| Pneumonia | | | | | | |
| Yes | 12(10.1) | 1(ref.) | | 3(5.0) | 1(ref.) | |
| No | 107(89.9) | 0.31(0.07–1.36) | 0.121 | 57(95.0) | 1.08(0.12–9.69) | 0.947 |
| Adjuvant therapy | | | | | | |
| Yes | 69(58.0) | 1(ref.) | | 35(58.3) | 1(ref.) | |
| No | 24(20.2) | 0.23(0.05–1.15) | 0.074 | 21(35.0) | 0.32(0.09–1.13) | 0.076 |
| Unknown | 26(21.8) | 0.83(0.32–2.15) | 0.703 | 4(6.7) | 1.63(0.43–6.22) | 0.473 |
| Alcohol use | | | | | | |
| Yes | 74(62.2) | 1(ref.) | | 32(53.3) | 1(ref.) | |
| No | 45(37.8) | 1.84(0.65–5.22) | 0.251 | 28(46.7) | 0.69(0.25–1.88) | 0.469 |
| Tobacco use | | | | | | |
| Yes | 80(67.2) | 1(ref.) | | 34(56.7) | 1(ref.) | |
| No | 39(32.8) | 0.62(0.18–2.18) | 0.454 | 26(43.3) | 1.14(0.34–3.78) | 0.837 |

## Results

### Patient characteristics

The characteristics of patients in the discovery and independent validation cohorts were listed in Table 1. The median duration of follow-up was 32.2 (IQR 13.2–61.5) months in the discovery cohort and 39.3 (IQR 12.7–53.2) months in the independent validation cohort, respectively. Furthermore, the 1- and 3-year overall survival (OS) were 78.2% and 46.2% in the discovery cohort, whereas the validation cohort exhibited rates of 75% and 55%. In particularly, age was a critical variable significantly associated with the prognosis of ESCC. Patients with higher age at diagnosis (> 60 years old) had a lower OS rate in both discovery and validation cohorts. Besides, tumor grade, T stage, N stage, TNM stage, tobacco use, alcohol use, arrhythmia, pneumonia, and adjuvant therapy showed no remarkable association with ESCC prognosis, while sex and tumor location were observed to be statistically significantly in the discovery or validation cohorts, respectively.

### Feature selected by LASSO regression

According to our previous study, 1,336 coding genes were defined as testis-specific genes (C1 class). We performed a systematic analysis based on 1,336 testis-specific coding genes and RNA microarray data from 119 paired ESCC samples to identify CTGs of ESCC [10, 16]. Eventually, 468 ESCC specific CTGs (Additional file 1: Table S1) combined with 12 clinical characteristics consisted of 480 original features. We utilized LASSO regression to select the optimal features, removing less important parameters and reducing the correlation between variables (Fig. 2A-B). Ultimately, 6 variables, including HEN Methyltransferase 1 (*HENMT1*), Cell Division Cycle 25 A (*CDC25A*), Heat Shock Protein Family B Member 9 (*HSPB9*), Heat Shock Transcription Factor 2 Binding Protein (*HSF2BP*), Sulfotransferase Family 6B Member 1 (*SULT6B1*) (Fig. 2C-F) and TNM stage were selected for predicting OS according to the weighted coefficients in the discovery cohort. To prevent overfitting or uncertainty in the model, we examined the correlation between continuous variables by spearman method. In fact, as shown in Fig. 2H, we observed slight correlation between *HENMT1* and *HSPB9* among six variables.

### Model performance

Overall, 9 supervised ML models, including AdaBoost, GBDT, Bagging, DT, ET, GaussianNB, KNN, SVM and RF, and basic LR all showed varying but promising performance in predicting mortality risk of ESCC in the discovery and validation cohorts. Table 2 shows that in ten classifiers, the highest predictive accuracy and the AUC were 88.24% and 0.8848 with DT, 87.39% and 0.8775 with GBDT, 87.39% and 0.8578 with ET, 83.19% and 0.826 with Adaboost, 83.19% and 0.8113 with RF, 81.51% and 0.8137 with Bagging in the discovery cohort, respectively. KNN, GaussianNB, SVM and LR models perform slightly worse, yielding predictive accuracy and the AUC of 71.43% and 0.6985, 65.55% and 0.6299, 60.50% and 0.5956, 67.23% and 0.6422, respectively. Similarly, in validation cohort, the predictive accuracy and the AUC were 83.33% and 0.835 with ET, 81.67% and 0.8165 with GBDT, 80% and 0.7946 with Bagging, 78.33% and 0.7761 with DT, 78.33% and 0.7626 with RF, 70% and 0.7003 with AdaBoost, 68.33% and 0.6717 with KNN, 60% and 0.5758 with GaussianNB, 58.33% and 0.5842 with SVM, 58.33% and 0.564 with LR, respectively.

Using weighted voting method to integrate multiple basic classifier models can augment model performance. Thus, in our study, we integrated the top six best predictive models (AdaBoost, GBDT, Bagging, DT, ET and RF) to create an ensemble model called PPMESCC. As expected, the final forecast result of PPMESCC outperforms the basic classifier models and other ensemble strategies. The AUC for PPMESCC was 0.9828 (95% confidence interval: 0.9608 to 0.9926), with an accuracy of 98.32% (95% CI: 96.64–99.16%) in the discovery cohort. In the validation cohort, the AUC for PPMESCC to predict ESCC prognosis was 0.9057 (95% CI: 0.8897 to 0.9583), with an accuracy of 90% (95% CI: 89.08–93.28%) (Fig. 3A-B; Table 2).

Moreover, utilizing the time from surgery to death or discharge as the designated endpoint, the Kaplan-Meier analysis further confirmed that PPMESCC exhibited an impressive ability to stratify OS of postoperative ESCC patients. Within both the discovery and validation cohorts, ESCC patients identified as having poor prognosis according to PPMESCC demonstrated significantly lower chances of survival compared to those with a favorable prognosis (Fig. 3C, D; $p < 0.001$), emphasizing the accurate prognostic predictive power of PPMESCC for ESCC.

### HENMT1 promoted the malignant phenotypes of ESCC cells in vitro

To elucidate the biological function of *HENMT1* in ESCC, an overexpression plasmid was synthesized to upregulate *HENMT1* (*HENMT1 OE*). CCK8 and clone formation assays indicated that elevated expression of *HENMT1* in KYSE150 and KYSE410 cells dramatically enhanced cell viability and proliferative capacity, respectively (Fig. 4A, B). Besides, we further investigated the effect of *HENMT1* on cell cycle and migration by flow cytometric analysis and scratch assay. Consistently, our results determined that upregulated *HENMT1* promoted the cell cycle and wound healing (Fig. 4C, D). These findings together illustrate that CTG *HENMT1* plays an oncogenic role in ESCC progression via stimulated tumor proliferation.

### HENMT1 potentially regulated piRNAs in ESCC

Considering that HENMT1 is responsible for regulating the 2'-O-methylation of the 3'-end of piRNAs, we explored piRNAs that may be regulated by HENMT1 in ESCC. A systematic strategy was devised to search esophageal carcinoma-related SNPs in the GWAS catalog database (https://www.ebi.ac.uk/gwas/), and discovered 130 SNPs significant associated with the development of esophageal carcinoma (Additional file 1: Table S2). Subsequently, piRNA-eQTL analysis was conducted to further examine the association between positive SNPs with the piRNA expression (http://njmu-edu.cn:3838/piRNA-eQTL/). Enthusiastically, 46 piRNAs were identified linked to rs7141987 that exhibited aberrant expression in esophageal carcinoma (Additional file 1: Table S3). To investigate the process of 3'-end 2OM of piRNAs, the piRNA base database (http://bigdata.ibp.ac.cn/piRBase/) and i2OM motifs prediction system (i2om.lin-group.cn) were employed. Consequently, 8 piRNAs

**Table 2** Performance for mortality risk prediction of models in discovery and validation cohorts

| | | AUC (95% CI) | Accuracy (95% CI) | Sensitivity (95% CI) | Specificity (95% CI) | PPV (95% CI) | NPV (95% CI) | F1 | Kappa |
|---|---|---|---|---|---|---|---|---|---|
| **Discovery cohort (GSE53624)** | | | | | | | | | |
| | Adaboost | 0.8260 (0.8039–0.8480) | 83.19% (81.51–85.71%) | 78.43% (68.63–84.31%) | 86.76% (83.82–88.24%) | 81.63% (74.19–90.70%) | 84.29% (82.19–86.67%) | 0.800 | 0.655 |
| | GaussianNB | 0.6299 (0.6176–0.6495) | 65.55% (64.71–67.23%) | 45.10% (35.29–50.98%) | 80.88% (77.94–82.35%) | 63.89% (61.54–65.71%) | 66.27% (64.52–67.95%) | 0.529 | 0.270 |
| | GBDT | 0.8775 (0.8554–0.9044) | 87.39% (85.71–89.08%) | 90.20% (80.39–98.04%) | 85.29% (83.82–88.24%) | 82.14% (72.92–90.74%) | 92.06% (91.30-94.37%) | 0.860 | 0.746 |
| | KNN | 0.6985 (0.6789–0.7181) | 71.43% (69.75–73.95%) | 58.82% (49.02–70.59%) | 80.88% (79.41–83.82%) | 69.77% (63.46–81.25%) | 72.37% (69.32–75.36%) | 0.638 | 0.405 |
| | LR | 0.6422 (0.6250–0.6569) | 67.23% (66.39–68.07%) | 43.14% (31.37–54.90%) | 85.29% (82.35–88.24%) | 68.75% (62.50-77.78%) | 66.67% (63.64–69.33%) | 0.530 | 0.298 |
| | SVM | 0.5956 (0.5784–0.6201) | 60.50% (58.82–62.18%) | 52.94% (41.18–54.90%) | 66.18% (64.71–69.12%) | 54.00% (53.06–56.76%) | 65.22% (62.92–68.06%) | 0.535 | 0.192 |
| | Bagging | 0.8137 (0.7892–0.8309) | 81.51% (79.83–83.19%) | 80.39% (70.59–82.35%) | 82.35% (79.41–85.29%) | 77.36% (69.81–88.10%) | 84.85% (81.82–87.67%) | 0.789 | 0.624 |
| | DT | 0.8848 (0.8627–0.9020) | 88.24% (87.39–89.08%) | 90.20% (80.39–98.04%) | 86.76% (83.82–88.24%) | 83.64% (74.51–92.16%) | 92.19% (87.50-94.12%) | 0.868 | 0.762 |
| | ET | 0.8578 (0.8284–0.8627) | 87.39% (85.71–89.92%) | 74.51% (64.71–86.27%) | 97.06% (94.12–98.53%) | 95.00% (84.21–97.96%) | 83.54% (81.43–85.33%) | 0.835 | 0.736 |
| | RF | 0.8113 (0.7966–0.8309) | 83.19% (81.51–85.71%) | 66.67% (52.94–76.47%) | 95.59% (92.65–97.06%) | 91.89% (82.35–98.04%) | 79.27% (77.78–82.61%) | 0.773 | 0.645 |
| | PPMESCC | 0.9828 (0.9608–0.9926) | 98.32% (96.64–99.16%) | 98.04% (88.24–99.99%) | 98.53% (95.59–99.99%) | 98.04% (89.29–98.08%) | 98.53% (96.97-99.00%) | 0.980 | 0.966 |
| **Validation cohort (GSE53622)** | | | | | | | | | |
| | Adaboost | 0.7003 (0.6765–0.7672) | 70.00% (68.07–74.79%) | 70.37% (64.71–72.55%) | 69.70% (67.65–76.47%) | 65.52% (59.15-74.00%) | 74.19% (72.73–79.71%) | 0.679 | 0.398 |
| | GaussianNB | 0.5758 (0.5606–0.6044) | 60.00% (58.33–61.67%) | 33.33% (29.63–48.15%) | 81.82% (78.79–84.85%) | 60.00% (55.00-64.29%) | 60.00% (59.09–63.16%) | 0.429 | 0.158 |
| | GBDT | 0.8165 (0.7819–0.8480) | 81.67% (79.83–84.87%) | 81.48% (78.43–84.31%) | 81.82% (75.00-86.76%) | 78.57% (71.67–87.50%) | 84.38% (81.43–86.76%) | 0.800 | 0.631 |
| | KNN | 0.6717 (0.6520–0.7157) | 68.33% (65.55–72.27%) | 55.56% (47.06–66.67%) | 78.79% (73.53–82.35%) | 68.18% (60.47–77.42%) | 68.42% (66.25–72.37%) | 0.612 | 0.349 |
| | LR | 0.5640 (0.5370–0.5875) | 58.33% (55.00-61.67%) | 37.04% (29.63–40.74%) | 75.76% (69.70-78.79%) | 55.56% (50.00-66.67%) | 59.52% (56.82–60.98%) | 0.444 | 0.132 |
| | SVM | 0.5842 (0.5556–0.6077) | 58.33% (55.00-61.67%) | 59.26% (48.15–66.67%) | 57.58% (54.55–60.61%) | 53.33% (50.00-56.52%) | 63.33% (60.47–66.67%) | 0.561 | 0.167 |
| | Bagging | 0.7946 (0.7647–0.8186) | 80.00% (78.99–83.19%) | 74.07% (66.67–82.35%) | 84.85% (80.88–89.71%) | 80.00% (71.43–90.91%) | 80.00% (77.63–84.38%) | 0.769 | 0.593 |
| | DT | 0.7761 (0.7525–0.8039) | 78.33% (76.47–80.67%) | 70.37% (60.78–78.43%) | 84.85% (79.41–89.71%) | 79.17% (68.42-86.00%) | 77.78% (75.64–80.65%) | 0.745 | 0.558 |
| | ET | 0.8350 (0.7745–0.8603) | 83.33% (78.99–86.55%) | 85.19% (72.55–94.12%) | 81.82% (77.94–88.24%) | 79.31% (69.39–88.46%) | 87.10% (86.57–88.89%) | 0.821 | 0.666 |
| | RF | 0.7626 (0.7475–0.7917) | 78.33% (76.47–80.67%) | 55.56% (50.98–70.59%) | 96.97% (91.18–98.53%) | 93.75% (80.95–97.22%) | 72.73% (71.28–73.63%) | 0.698 | 0.546 |
| | PPMESCC | 0.9057 (0.8897–0.9583) | 90.00% (89.08–93.28%) | 96.30% (88.24–98.04%) | 84.85% (83.82–88.24%) | 83.87% (80.36–93.88%) | 96.55% (94.29–97.06%) | 0.897 | 0.801 |

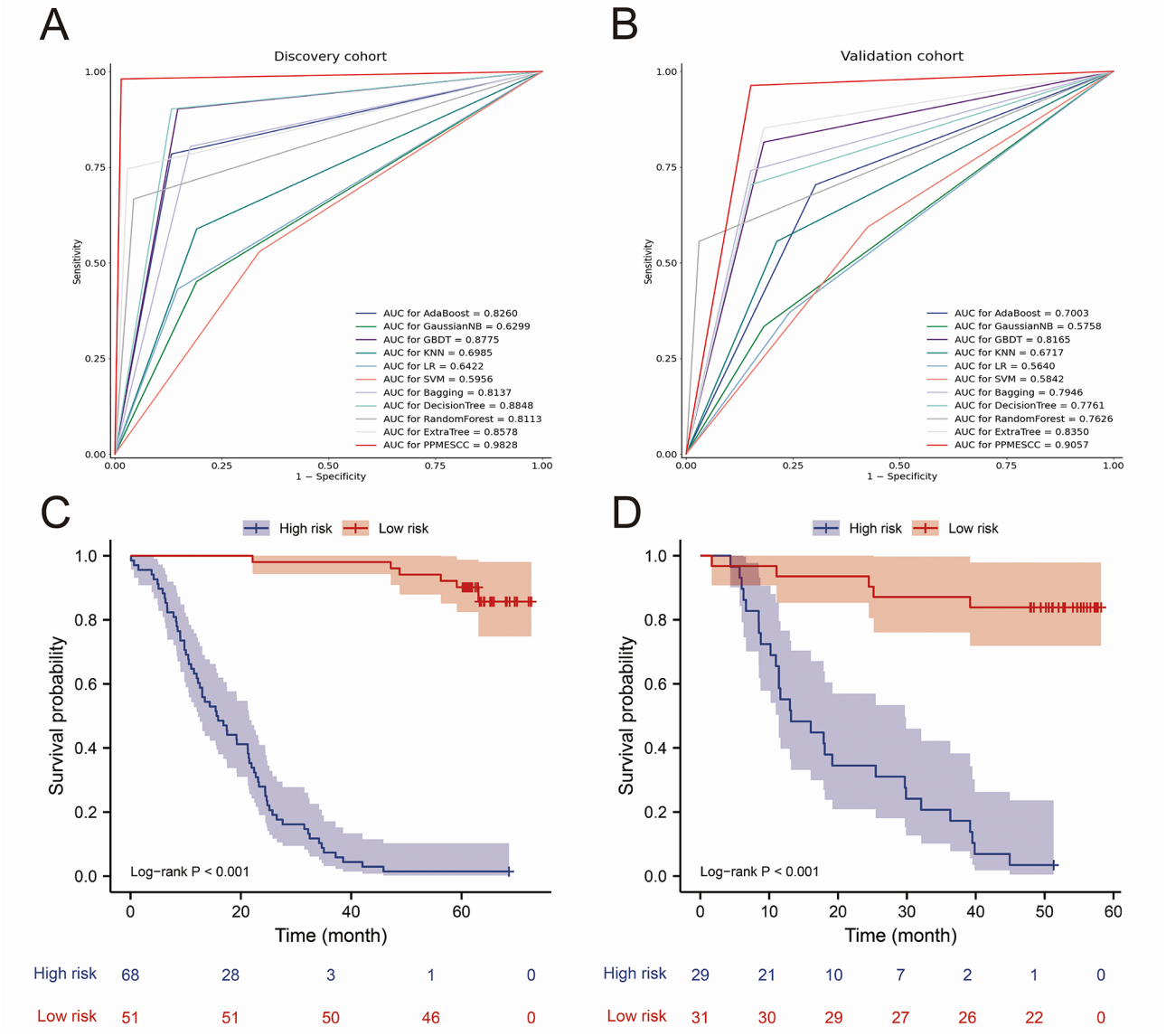


**Fig. 3** Predictive performances of models in the discovery and validation cohorts. **A-B** AUC exhibited the performance of prediction prognosis of models (AdaBoost, GBDT, Bagging, DT, ET, GaussianNB, KNN, SVM, RF, LR and PPMESCC). **C-D** Kaplan–Meier curves indicated OS of patients with high and low mortality risk. The red or blue numbers represent the probability of survival, and the red or blue areas indicate the 95% confidence interval for the survival probability

(piR-hsa-155556, piR-hsa-4380933, piR-hsa-4380226, piR-hsa-1873592, piR-hsa-4458529, piR-hsa-142679, piR-hsa-3661062 and piR-hsa-139945) were identified potentially regulated by HENMT1 in ESCC (Table 3).

## Discussion

By utilizing large datasets and advanced algorithms, ML has the tremendous potential to accurately predict cancer outcomes, thereby enabling clinicians to identify patients at greater risk of disease progression or recurrence, and formulate tailored treatment approaches. In this study, LASSO regression was employed to identify the key predictors for postoperative ESCC patients. Among 468 CTGs and 12 clinical parameters, the final model selected five CTGs (*HENMT1*, *CDC25A*, *HSPB9*, *HSF2BP* and *SULT6B1*) and TNM stage as the most influential variables. Initially, nine ML models and LR method displayed varying but promising performances to predict OS of ESCC. To build a model with robust predictive capacity, we integrated the top six best predictive models (Adaboost, GBDT, Bagging, DT, ET and RF) to create an ensemble model called PPMESCC. As expected, PPMESCC achieved an AUC of 0.9828 (95% confidence interval: 0.9608–0.9926) in identification of non-survivors with an accuracy of 98.32% (95% CI: 96.64–99.16%) in the discovery cohort. For validation cohort, PPMESCC

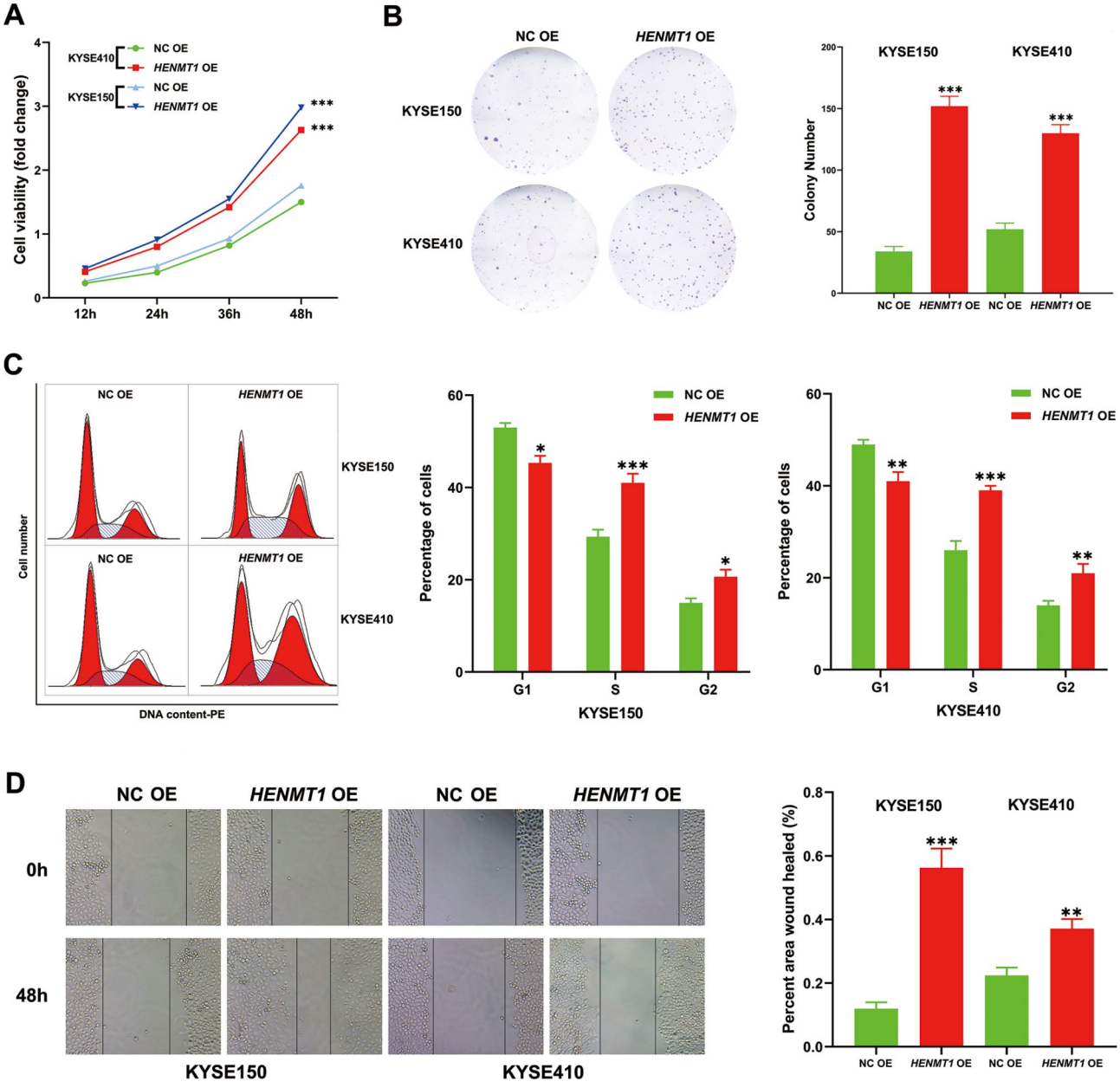


**Fig. 4** *HENMT1* promoted the proliferation and migration in vitro. **A** CCK8 assay was performed to determine the cell viability of ESCC cells. **B** Colony formations assay detected the cell proliferation capacity of ESCC cells. **C** Flow cytometry analysis of cell-cycle phase distribution. **D** Scratch assay was used to measure wound-healing capability of ESCC cells. *Indicates $p \leq 0.05$, **indicates $p \leq 0.01$, and ***indicates $p \leq 0.001$ by a two tailed Student's t test. Data show mean ± SE for all panels with error bars

revealed an AUC of 0.9057 (95% CI: 0.8897–0.9583) and an accuracy of 90.00% (95% CI: 89.08–93.28%) to predict prognosis of ESCC. Moreover, *HENMT1* showed the most significant expression among the valuable variables and was subsequently selected for functional investigation. In KYSE150 and KYSE410 cells, overexpressing *HENMT1* dramatically increased cell viability, proliferation, and migration capacities. Besides, HENMT1 functions as a piRNA 2'-O-methyltransferase, we conducted a thorough analysis of GWAS, eQTL-piRNA, and i2OM databases, eventually unearthing 8 piRNAs potentially regulated by HENMT1 in ESCC.

CTGs, highly expressed in cancer and testis tissues but scarcely or not expressed in other normal tissues, share the common traits of indefinite multiplication and diffusion, thus providing the inherent advantage in predicting tumor prognosis [20]. In 2019, we adopted a systematic screening strategy to screen CTGs in ESCC by integrating multiple public databases and RNA expression microarray data. The study found that CDCA5 could promote

**Table 3** The probability of 2′-O-methylation at the 3′ end of esophageal carcinoma-related piRNAs according to i2OMa

| pi-RNA | Length | Sequence | 2OM Position | Base | 3′ end 2OM | Probability |
|---|---|---|---|---|---|---|
| piR-hsa-155,556 | 25 | TCTCACACAGAAATCGCACCCGTT**A** | 25 | A | Yes | 0.739 |
| piR-hsa-4,380,933 | 25 | TCTCACACAGAAATCGCAACCGTC**A** | 25 | A | Yes | 0.616 |
| piR-hsa-1,873,592 | 27 | TCTCACACAGAAATCGCACCCGTCAC**A** | 27 | A | Yes | 0.551 |
| piR-hsa-142,679 | 25 | TCTCACACAGAAATCGCACCCGTG**A** | 25 | A | Yes | 0.729 |
| piR-hsa-4,458,529 | 25 | TCTCACACAGAAATCGCACCTGTC**A** | 25 | A | Yes | 0.737 |
| piR-hsa-3,661,062 | 26 | TCTCACACAGAAATCGCACCCGTCG**C** | 26 | C | Yes | 0.517 |
| piR-hsa-4,380,226 | 25 | TCTCACACAGAAATCGCACCCGCC**A** | 25 | A | Yes | 0.694 |
| piR-hsa-139,945 | 25 | TCTCACACAGAAATCGCACCCGTA**A** | 25 | A | Yes | 0.572 |

[a]i2OM is an online tool that available for prediction 2′-O-methylation in human RNA (i2om.lin-group.cn)

ESCC cells proliferation, invasion, migration, apoptosis resistance and reduce chemosensitivity to cisplatin [10]. In the present study, among 468 CTGs in ESCC, a total of 5 CTGs including *HENMT1*, *CDC25A*, *HSPB9*, *HSF2BP* and *SULT6B1* were applied to develop the postoperative predictive model. CDC25A is a member of the CDC25 family of phosphatases involved in the progression from G1 to the S phase of the cell cycle. Li et al. unveiled that CDC25A could be activated by FOXK1, which was positively correlated with TNM stage, invasion depth, and lymph node metastasis of ESCC [21]. Interestingly, Luo et al. revealed that downregulation of miR-339-5p in ESCC tissues stimulated the enhanced expression of CDC25A, which consequently led to radioresistance, local recurrence, and distant metastatic relapse [22]. Unfortunately, there is no literature regarding the functions of the other four CTGs in ESCC.

HENMT1, a small RNA 2′-O-methyltransferase, is responsible for the addition of a 2′-O-methyl group to the 3′-end of piRNAs, which shields them from uridylation and subsequent degradation. The lack of HENMT1 has been observed to cause piRNA instability, which in turn led to the derepression of retrotransposons and the precocious expression of the haploid germ cell programme in meiotic cells, resulting in malformed spermatids and male infertility [23]. Recently, Begik and colleagues performed a comprehensive analysis of human RNA modification-related protein expression patterns across 32 tissues, 10 species, and 13,358 paired tumor-normal human samples. The analysis demonstrates that HENMT1 was the top recurrently upregulated RNA modification-related protein in multiple types of cancer, particularly in stages III and IV patients [24]. In addition, HENMT1 was discovered as the key regulator of 3′-terminal 2′Ome of miR-21-5p in non-small cell lung cancer. Compared to non-methylated miR-21-5p, methylated miR-21-5p was more resistant to digestion by 3′→5′ exoribonuclease polyribonucleotide nucleotidyltransferase 1 and had higher affinity to Argonaute-2, which may contribute to its stability and inhibition programmed cell death protein 4 translation [25]. More importantly, in the present study, we found that HENMT1 showed the most aberrantly expression among five ESCC-specific CTG biomarkers. Functional assays further revealed that overexpression of HENMT1 significantly augmented the proliferation and migration capacities of ESCC cells, indicating that the anomalous expression of HENMT1 is closely implicated in the occurrence and progression of ESCC.

piRNAs, a novel group of noncoding RNAs spanning 24 to 30 nucleotides in length, are responsible for the regulation of numerous downstream genes through processes such as heterochromatin formation, DNA methylation, mRNA cleavage, and protein interactions [26]. Recent studies have shown that acting as tissue-specific molecules, piRNAs have both oncogenic and tumor suppressive functions in cancer progression, including regulating cancer cell proliferation, metastasis, chemoresistance, and stemness [27]. For instance, piR-651 has been found to upregulate Cyclin D1 and CDK4, two key regulators of G1-to-S phase transition, in both non-small cell lung carcinoma and breast cancer [28, 29]. Additionally, two upregulated (piR-34871 and piR-52200) and downregulated (piR-35127 and piR-46545) piRNAs were reported upon overexpression of RASSF1C in lung cancer, resulting in the downregulation of genes associated with cell proliferation through the AMPK pathway [30]. Nonetheless, the underlying functions of piRNAs in the development and progression of ESCC have yet to be fully investigated. Enthusiastically, in the past decade, genome-wide association studies (GWASs) have successfully identified multiple SNPs associated with human cancers. Expression quantitative trait locus (eQTL) analysis, a method for linking SNPs to gene expression, has been demonstrated to be a powerful approach in unraveling the underlying molecular mechanisms of cancers. In 2021, Xin et al. developed a user-friendly database, piRNA-eQTL (http://njmu-edu.cn:3838/piRNA-eQTL/), to provide an eQTL analysis between SNPs and piRNA expression using genotyping and piRNA expression data for 10,997 samples across 33 cancer types from The Cancer Genome Atlas (TCGA) [31]. In the current research, we initially searched the GWAS catalog database (https://www.ebi.ac.uk/gwas/) for esophageal carcinoma-related

SNPs and discovered 130 SNPs that exhibit a strong correlation with EC. Subsequently, the piRNA-eQTL database was utilized to examine the association between the 130 SNPs and piRNA expression. Notably, 46 piRNAs was identified linked to rs7141987 exhibited aberrant expression patterns, highlighting their potential functional role in EC.

2′-O-methylation is a ubiquitous post-transcriptional modification in RNAs that is catalyzed by 2'-O-methyltransferase, replacing the H on the 2'-hydroxyl group with a methyl group. It holds the ability to exert diverse effects on RNAs, including enhancing their hydrophobic properties, safeguarding them from nuclease cleavage, stabilizing helical conformations, and modulating RNA-protein/RNA interactions [32]. Recently, it has become evident that the roles of 2OM extend far beyond basic RNA stabilization, with these 2OM sites participating in the regulation of gene expression as well as various other cellular processes [33]. Since HENMT1 protein is primarily involved in regulating the 3'-end 2OM of piRNAs, the sequences of the 46 EC associated piRNAs were obtained from the piRNA base database (http://bigdata.ibp.ac.cn/piRBase/). We then used a powerful system called i2OM (i2om.lin-group.cn), developed by Yang et al. in 2023, that predicts the sequence motifs of 2OM in human RNA [34] to search 3'-end 2OM regulatory sites. Ultimately, 8 piRNAs were identified potentially regulated by HENMT1 in EC.

Harnessing ML methods to analyze extensive healthcare data yields considerable advantages in grasping and evaluating intricate information, and precisely forecasting the survival of cancer patients [35–37]. Prior research has endeavored to determine the practicability and efficacy of ML-driven approaches in prognosticating the survival outcomes of patients with ESCC. Zhang et al. devised a ML prediction model for ESCC patients' survival by 27 clinical features which effectively stratified ESCC patients into low-, intermediate-, and high-risk groups, with distinctly different 3-year OS probabilities of 80.8%, 58.2%, and 29.5%, respectively [38]. On the other hand, using 48 clinically proven molecules linked to ESCC progression, Li et al. developed a ML model for prognostic prediction in ESCC, resulting in 3-year survival rates of 42.4% and 63.1% for the high-risk and low-risk subgroups, respectively [39]. Nevertheless, these investigations relied solely on either biomarkers or clinical features to predict the prognosis of patients with ESCC, thereby limiting their predictive capability and applicability in current clinical practice. Interestingly, Zheng et al. constructed a six-lncRNA signature that, when combined with the TNM stage, demonstrated improved predictive ability for ESCC prognosis using the GSE53622 and GSE53624 datasets. This inspired us to develop a predictive model by integrating molecular biomarkers with clinical parameters [40]. In this study, we integrated 12 clinical features and 468 CT coding genes as research variables and employed nine ML approaches to develop prognosis prediction models including AdaBoost, GBDT, Bagging, DT, ET, GaussianNB, KNN, SVM, and RF. It is noteworthy that we innovatively created an ensemble model, PPMESCC, derived from six ML algorithms (AdaBoost, GBDT, Bagging, DT, ET, and RF) which exhibits exceptional predictive performance for ESCC, with an AUC of 0.983 in discovery cohort and 0.906 in validation cohort. Based on accurate predictions from PPMESCC, clinicians can devise personalized, evidence-based decisions, such as adjustments to adjuvant therapy regimens or the implementation of more frequent follow-up schedules for high-risk patients, which could ultimately optimize resource allocation and improve patient outcomes in ESCC.

Nonetheless, several limitations also need to be considered. First, the limited sample sizes may undermine the robustness and reliability of the findings. Future research should prioritize incorporating larger, more heterogeneous populations to confirm these findings and enhance their generalizability to broader clinical contexts. Second, although bioinformatics analyses and cell experiments have validated the results, further emphasis should be placed on in vivo animal studies and clinical trials to fully confirm the findings.

## Conclusion

In summary, our study has successfully developed and validated a ML prognostic model by CT coding genes and clinical feature, named PPMESCC. This model can serve as a reliable tool for accurate predicting the survival outcome of postoperative ESCC patients. The application of PPMESCC may potentially assist clinicians to promptly target the high-risk patients and make effective management strategy.

**Abbreviations**

| | |
|---|---|
| ESCC | Esophageal squamous cell carcinoma |
| LASSO | Least absolute shrinkage and selection operator |
| 2OM | 2′-O-methylation |
| CTGs | Cancer/testis genes |
| EC | Esophageal cancer |
| AJCC | American Joint Commission on Cancer |
| ML | Machine learning |
| AI | Artificial intelligence |
| OS | Overall survival |
| GBDT | Gradient Boosting Decision Trees |
| DT | Decision Trees |
| ET | Extra Trees |
| GaussianNB | Gaussian Naive Bayes |
| KNN | K-Nearest Neighbors |
| SVM | Support Vector Machines |
| RF | Random Forest |
| LR | Logistic Regression |
| ROC | Receiver operating characteristic curve |
| AUC | Area under the ROC curve |
| PPV | Positive predictive value |

| | |
|---|---|
| NPV | Negative predictive value |
| Kappa | Cohen's Kappa coefficient |
| CCK8 | Cell Counting Kit-8 assay |
| SNP | Single nucleotide polymorphism |
| GWAS | Genome-wide association study |
| eQTL | Expression quantitative trait locus |

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12885-025-13520-6.

**Supplementary Material 1**: **Table S1**. General description of 468 ESCC specific CTGs.

**Supplementary Material 2**: **Table S2**. SNPs associated with esophageal carcinoma according to GWAS catalog.

**Supplementary Material 3**: **Table S3**. 46 piRNAs associated with Esophageal carcinoma-related SNPs according to piRNA-eQTL.

**Supplementary Material 4**: **Table S4**. 4 The univariate and multivariate COX regression analysis of the five CTGs in the PPMESCC model.

**Supplementary Material 5**: **Figure S1**. Time-dependent ROC of the PPMESCC model.

**Supplementary Material 6**: **Figure S2**. Correlation between the expression of five CTGs and the prognosis of ESCC.

**Supplementary Material 7**: **Figure S3**. Stratified analysis of the expression of five CTGs according to TNM stages.

**Supplementary Material 8**: **Figure S4**. 4 ROC Curves of PPMESCC models with different numbers of predictors.

**Acknowledgements**
We are grateful to the contributors to the public databases used in this study and all the authors of the study.

**Author contributions**
J.X. conceived the study concept and design, and provided expert knowledge and critically revised the paper; Z.W. and Z.X. conducted data analysis and interpretation, and wrote the paper; T.Z. and M.L. collected the dataset; J.Z. was responsible for the data curation; H.L., C.J., Y.Z., J.D., C.W. and L.C. performed the acquisition of data. All the authors of the manuscript have read and agreed with the presented findings and gave their consent for submission and publication. All authors reviewed the manuscript.

**Funding**
This research was funded by National Natural Science Foundation of China (Grant No. 81972175), Hengrui Medical Clinical Research Foundation of Collaborative Innovation Center for Cancer Personalized Medicine (Grant No. 2024-02) and the Clinical Capacity Enhancement Project of the First Affiliated Hospital of Nanjing Medical University (Grant No. 2021-02).

**Data availability**
No datasets were generated or analysed during the current study.

## Declarations

**Ethics approval and consent to participate**
The study was approved by the Ethics Committee of The First Affiliated Hospital of Nanjing Medical University (Protocol code: 2022-SR-055. Approval date: 20 January 2022). All experiments on humans and/or the use of human tissue samples strictly adhered to the Declaration of Helsinki.

**Consent for publication**
Consent to publish has been obtained from all authors.

**Competing interests**
The authors declare no competing interests.

**Author details**
[1]Department of Thoracic Surgery, The First Affiliated Hospital of Nanjing Medical University, No. 300, Guangzhou Road, Nanjing 210029, Jiangsu, China
[2]Department of Thoracic Surgery, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, No. 1095, Jiefang Avenue, Wuhan 430030, Hubei, China
[3]Department of Pathology, The First Affiliated Hospital of Nanjing Medical University, No. 300, Guangzhou Road, Nanjing 210029, Jiangsu, China
[4]Department of Geriatrics, The First Affiliated Hospital of Nanjing Medical University, No. 300, Guangzhou Road, Nanjing 210029, Jiangsu, China
[5]Department of Epidemiology and Biostatistics, Jiangsu Key Lab of Cancer Biomarkers, Prevention and Treatment, Collaborative Innovation Center for Cancer Personalized Medicine, School of Public Health, Nanjing Medical University, No. 101, Longmian Avenue, Nanjing 211166, Jiangsu, China

Received: 28 April 2024 / Accepted: 14 January 2025
Published online: 23 January 2025

**References**

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer statistics 2020: GLOBOCAN estimates of incidence and Mortality Worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71:209–49.
2. Arnold M, Soerjomataram I, Ferlay J, Forman D. Global incidence of oesophageal cancer by histological subtype in 2012. Gut. 2015;64:381–87.
3. Rogers JE, Sewastjanow-Silva M, Waters RE, Ajani JA. Esophageal cancer: emerging therapeutics. Expert Opin Ther Targets. 2022;26:107–17.
4. Huang FL, Yu SJ. Esophageal cancer: risk factors, genetic association, and treatment. Asian J Surg. 2018;41:210–15.
5. Rice TW, Gress DM, Patil DT, Hofstetter WL, Kelsen DP, Blackstone EH. Cancer of the esophagus and esophagogastric junction-major changes in the American Joint Committee on Cancer eighth edition cancer staging manual. CA Cancer J Clin. 2017;67:304–17.
6. Stiekema J, Boot H, Aleman BM, Wessels LF, van Sandick JW. Prognostication and prediction using gene expression profiling in oesophageal cancer. Eur J Surg Oncol. 2013;39:17–23.
7. Tan H, Zhang H, Xie J, Chen B, Wen C, Guo X, et al. A novel staging model to classify oesophageal squamous cell carcinoma patients in China. Br J Cancer. 2014;110:2109–15.
8. Simpson AJ, Caballero OL, Jungbluth A, Chen YT, Old LJ. Cancer/testis antigens, gametogenesis and cancer. Nat Rev Cancer. 2005;5:615–25.
9. Forghanifard MM, Gholamin M, Farshchian M, Moaven O, Memar B, Forghani MN, et al. Cancer-testis gene expression profiling in esophageal squamous cell carcinoma: identification of specific tumor marker and potential targets for immunotherapy. Cancer Biol Ther. 2011;12:191–97.
10. Xu J, Zhu C, Yu Y, Wu W, Cao J, Li Z, et al. Systematic cancer-testis gene expression analysis identified CDCA5 as a potential therapeutic target in esophageal squamous cell carcinoma. EBioMedicine. 2019;46:54–65.
11. Wang Q, Cao B, Peng L, Dai W, Jiang Y, Xie T, et al. Development and validation of a practical prognostic Coagulation Index for patients with esophageal squamous cell Cancer. Ann Surg Oncol. 2021;28:8450–61.
12. Song Q, Wu JA-O, Wang SA-O, Chen WH. Elevated preoperative platelet distribution width predicts poor prognosis in esophageal squamous cell carcinoma. Sci Rep. 2019;9:15234.
13. D'Ascenzo F, De Filippo O, Gallone G, Mittone G, Deriu MA, Iannaccone M, et al. Machine learning-based prediction of adverse events following an acute coronary syndrome (PRAISE): a modelling study of pooled datasets. Lancet. 2021;397:199–207.
14. Schwalbe N, Wahl B. Artificial intelligence and the future of global health. Lancet. 2020;395:1579–86.
15. Abuhelwa AY, Kichenadasse G, McKinnon RA, Rowland A, Hopkins AM, Sorich MJ. Machine learning for prediction of Survival outcomes with Immune-Checkpoint inhibitors in Urothelial Cancer. Cancers (Basel). 2021;13:2001.
16. Wang C, Gu Y, Zhang K, Xie K, Zhu M, Dai N, et al. Systematic identification of genes with a cancer-testis expression pattern in 19 cancer types. Nat Commun. 2016;7:10499.
17. Fu H, Zhu Y, Wang Y, Liu Z, Zhang J, Xie H, et al. Identification and validation of Stromal Immunotype Predict Survival and Benefit from Adjuvant

Chemotherapy in patients with muscle-invasive bladder Cancer. Clin Cancer Res. 2018;24:3069–78.
18. Liang Y, Reza S. Advanced defensive distillation with ensemble voting and noisy logits. Appl Intell. 2022;53:3069–94.
19. Pan Z, Zhou S, Liu T, Liu C, Zang M, Wang Q. WVDL: Weighted Voting Deep Learning Model for Predicting RNA-Protein binding sites. IEEE/ACM Trans Comput Biol Bioinform. 2023;20:3322–28.
20. Sineath RC, Mehta A. Preservation of fertility in Testis Cancer Management. Urol Clin North Am. 2019;46:341–51.
21. Li X, Lu J, Liu L, Li F, Xu T, Chen L, et al. FOXK1 regulates malignant progression and radiosensitivity through direct transcriptional activation of CDC25A and CDK4 in esophageal squamous cell carcinoma. Sci Rep. 2023;13:7737.
22. Luo A, Zhou X, Shi X, Zhao Y, Men Y, Chang X, et al. Exosome-derived mir-339-5p mediates radiosensitivity by targeting Cdc25A in locally advanced esophageal squamous cell carcinoma. Oncogene. 2019;38:4990–5006.
23. Hempfling AL, Lim SL, Adelson DL, Evans J, O'Connor AE, Qu ZP, et al. Expression patterns of HENMT1 and PIWIL1 in human testis: implications for transposon expression. Reproduction. 2017;154:363–74.
24. Begik O, Lucas MC, Liu H, Ramirez JM, Mattick JS, Novoa EM. Integrative analyses of the RNA modification machinery reveal tissue- and cancer-specific signatures. Genome Biol. 2020;21:97.
25. Liang H, Jiao Z, Rong W, Qu S, Liao Z, Sun X, et al. 3'-Terminal 2'-O-methylation of lung cancer mir-21-5p enhances its stability and association with Argonaute 2. Nucleic Acids Res. 2020;48(13):7027–40.
26. Wu Z, Yu X, Zhang S, He Y, Guo W. Novel roles of PIWI proteins and PIWI-interacting RNAs in human health and diseases. Cell Commun Signal. 2023;21:343.
27. Deng X, Liao T, Xie J, Kang D, He Y, Sun Y, et al. The burgeoning importance of PIWI-interacting RNAs in cancer progression. Sci China Life Sci. 2023. https://doi.org/10.1007/s11427-023-2491-7.
28. Li D, Luo Y, Gao Y, Yang Y, Wang Y, Xu Y, et al. piR-651 promotes tumor formation in non-small cell lung carcinoma through the upregulation of cyclin D1 and CDK4. Int J Mol Med. 2016;38:927–36.
29. Liu T, Wang J, Sun L, Li M, He X, Jiang J, et al. Piwi-interacting RNA-651 promotes cell proliferation and migration and inhibits apoptosis in breast cancer by facilitating DNMT1-mediated PTEN promoter methylation. Cell Cycle. 2021;20:1603–16.
30. Reeves ME, Firek M, Jliedi A, Amaar YG. Identification and characterization of RASSF1C piRNA target genes in lung cancer cells. Oncotarget. 2017;8:34268–82.
31. Xin J, Du M, Jiang X, Wu Y, Ben S, Zheng R, et al. Systematic evaluation of the effects of genetic variants on PIWI-interacting RNA expression across 33 cancer types. Nucleic Acids Res. 2021;49:90–7.
32. Xiong Q, Zhang Y, Small. RNA modifications: regulatory molecules and potential applications. J Hematol Oncol. 2023;16:64.
33. Wu H, Qin W, Lu S, Wang X, Zhang J, Sun T, et al. Long noncoding RNA ZFAS1 promoting small nucleolar RNA-mediated 2'-O-methylation via NOP58 recruitment in colorectal cancer. Mol Cancer. 2020;19:95.
34. Yang YH, Ma CY, Gao D, Liu XW, Yuan SS, Ding H. i2OM: toward a better prediction of 2'-O-methylation in human RNA. Int J Biol Macromol. 2023;239:124247.
35. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. Lancet Oncol. 2019;20:e262–73.
36. Ding D, Lang T, Zou D, Tan J, Chen J, Zhou L, et al. Machine learning-based prediction of survival prognosis in cervical cancer. BMC Bioinformatics. 2021;22:331.
37. Howard FM, Kochanny S, Koshy M, Spiotto M, Pearson AT. Machine learning-guided adjuvant treatment of Head and Neck Cancer. JAMA Netw Open. 2020;3:e2025881.
38. Zhang K, Ye B, Wu L, Ni S, Li Y, Wang Q, et al. Machine learning–based prediction of survival prognosis in esophageal squamous cell carcinoma. Sci Rep. 2023;13:13532.
39. Li MX, Sun XM, Cheng WG, Ruan HJ, Liu K, Chen P, et al. Using a machine learning approach to identify key prognostic molecules for esophageal squamous cell carcinoma. BMC Cancer. 2021;21:906.
40. Zheng ZJ, Li YS, Zhu JD, et al. Construction of the Six-lncRNA prognosis signature as a novel biomarker in esophageal squamous cell carcinoma. Front Genet. 2022;13:839589.

## Publisher's note