Submitted: 04 April, 2023 Accepted: 24 May, 2023 Published: 15 March, 2025 DOI:10.22514/ejgo.2025.034

European Journal of Gynaecological Oncology

ORIGINAL RESEARCH

Development and validation of a novel gene signatures based on a random forest algorithm and artificial neural network for predictive diagnosis of cervical squamous cell carcinoma

Guiling Wang[1], Wenzheng Nong[1], Qingchun Lu[1], Ping Du[1], Jinghua Gan[1,]*

[1] Department of Gynecology, Affiliated Hospital for Nationalities of Guangxi Medical University, 530000 Nanning, Guangxi, China

***Correspondence**
ganjinghua@gxmzyy.wecom.work
(Jinghua Gan)

Abstract

Background: Effective treatment of cervical carcinoma can be challenging due to the lack of specific symptoms in the initial phase, as well as patients often only seeking medical attention in the middle and late stages of the disease when symptoms become more apparent. This study aims to address these limitations by developing and validating a gene signature for predicting cervical squamous cell carcinoma (CESC) using both the random forest algorithm and artificial neural network. **Methods**: Potential predictive genes for CESC were identified by analyzing three matrix datasets containing tissues from individuals with normal cervical epithelium and patients with CESC. Then, the random forest algorithm and artificial neural network were used to construct predictive models for CESC diagnosis, which were validated using both an independent validation dataset and *in vitro* experiments. To confirm the validity of the identified genes, protein and mRNA expression of eight disease signature genes were detected in the two groups using Western blotting and real-time quantitative polymerase chain reaction. Additionally, immunoinfiltration analysis was performed. **Results**: A total of 241 differentially expressed genes (DEGs) were identified, based on which eight genes with the highest predictive ability were selected and used to construct a molecular prognostic scoring system, which demonstrated exceptional predictive accuracy (Area Under Curve (AUC) = 0.995). Validation using an independent dataset confirmed the model's remarkable predictive ability (AUC = 1.000). *In vitro* experiments demonstrated significant differences in the expression of the eight disease signature genes between the two groups. Immunoinfiltration analysis also revealed significant differences in immune cell infiltration, with squamous cell carcinoma of the cervix showing a higher degree of macrophage infiltration than normal cervical epithelium. **Conclusions**: Random forest algorithm and artificial neural network were used to obtain new gene signatures, based on which a molecular prognostic scoring system was developed to predict CESC and aid clinical decision-making.

Keywords

Cervical squamous cell carcinoma; Diagnosis and treatment; Artificial neural network; Prediction model; Random forest algorithm; Immunity

1. Introduction

Cervical carcinoma (CC) is the second most commonly diagnosed malignancy in women, with a global incidence of 604,127 million cases and 341,831 million deaths in 2020 [1]. Approximately 75% of CC patients are histologically diagnosed with cervical squamous cell carcinoma (CESC), which accounts for most CC cases [2]. Unfortunately, the lack of specific symptoms in the initial stages of CC often results in patients presenting with prominent symptoms, such as contact bleeding, when the disease has already progressed to the middle or late stages, whereby treatment outcomes are often unsatisfactory, leading to a high recurrence rate and a low 5-year survival rate [3].

Radiotherapy based on chemotherapy is the current standard treatment approach for cervical carcinoma (CC), and neoadjuvant chemotherapy is often administered prior to surgery or radiotherapy. However, while this treatment approach may show positive outcomes in the short term, there is no substantial evidence from domestic and international guidelines and norms to demonstrate that it ultimately improves patient prognosis [4]. Thus, there is an urgent need to develop a reliable diagnostic model for predicting cervical squamous cell

This is an open access article under the CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/).

©2025 The Author(s). Published by MRE Press.

carcinoma (CESC) diagnosis, as this may help achieve early disease detection. Currently, different research teams in different regions are exploring diagnostic biomarkers for CESC. Zhang J *et al.* [5] found that C-X-C Motif Chemokine Ligand 10 (CXCL10) could be used as a potential serum biomarker in the diagnosis of cervical squamous cell carcinoma with squamous cell carcinoma antigen (SCC-Ag).

At present, due to the complexity of CESC pathogenesis, few effective tools are available for the early and accurate diagnosis of CC. However, advancements in bioinformatics have provided new methods for clinical prediction, and machine learning techniques such as random forest algorithms and artificial neural networks have proven effective in discovering biomarkers and researching various disease types [6, 7]. The development of machine learning techniques has enabled the selection of the most significant differentially expressed genes (DEGs) and their transformation into statistical models, which can assist clinicians in selecting rational and effective treatment options [8].

Herein, we designed this study to construct and validate a gene signature based on the random forest algorithm and artificial neural network for predicting the diagnosis of cervical squamous cell carcinoma (CESC).

2. Materials and methods

2.1 Study design and data sets

Four matrix datasets (GSE9750 [9], GSE63514 [10], GSE122697 [11] and GSE7803 [12]) containing data on the tissues from patients with normal cervical epithelium and CESC were selected for analysis. The GSE9750, GSE63514 and GSE122697 datasets were used as the training group, while the GSE7803 dataset was used as the validation group. The details of the training and validation groups are presented in Table 1. Fig. 1 illustrates the flowchart of the study design. The raw data for the above datasets were sourced from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/).

The top 30 differentially expressed genes (DEGs) significantly contributing to cervical squamous cell carcinoma (CESC) diagnosis prediction were identified using the random forest algorithm. Gene expression scores were then computed based on the expression data of these DEGs in all samples in the training group. A neural network model was created to obtain weight values of the genes with an importance score >2 and the greatest predictive contribution to CESC diagnosis. Based on the obtained gene expression scores and weight values, a molecular prognostic scoring system was constructed. The validity of the gene model was confirmed using the publicly available GSE7803 dataset, and ethical review was not required as the dataset is publicly available in the GEO database.

2.2 Identification of DEGs

Both the training and validation group data were processed using R software (version 4.1.1, Statistics Department of the University of Auckland, Auckland, New Zealand), which facilitated data standardization and normalization. The steps of merging training set data are as follows: 1. Read and merge the gene information of each training set; 2. Take log2 for the data with large values; 3. Merge the data; 4. Output the corrected data result. Duplicate gene probes were eliminated, and the consolidated data was used for subsequent analysis (**Supplementary Fig. 1**). The R software "Limma" [13] package was used to eliminate the batch effect in normal and tumor samples of the training group when identifying DEGs based on logFC >2 (log2 (Fold Change) >2) and adj. *p*.value < 0.05. A volcano plot was then created to visualize the identified DEGs. Finally, "pheatmap" and "ggplot2" packages of R software are used to draw heat map and volcano map of DEGs [14].

2.3 Enrichment analysis of DEGs

To obtain comprehensive information on the biological functions and signaling pathways associated with the significant differentially expressed genes (DEGs) in the training group dataset, pathway enrichment analysis and gene ontology annotation were performed using the Metascape online database (http://metascape.org) [15]. The results were then grouped into clusters based on the similarity of the significant terms, and the most prominent term was selected to present each cluster. Gene Ontology (GO) circles showing biological functions are drawn through the R software "clusterProfiler", "org.Hs.eg.db", "enrichplot" and "GOplot" package [16–19].

2.4 Identification of DEGs for predictive diagnosis of CESC based on random forest algorithm and artificial neural network

The R software "randomForest" package was used to identify the top 30 genes that significantly impacted the diagnosis of CESC [20], which were then used to generate a gene score table reflecting their expression levels [21]. The expression values of the DEGs were converted into binary values of 1 or 0 based on specific conversion criteria. If the expression value of an upregulated gene in a sample was higher than the median expression value of that gene across all samples, it was assigned a value of 1; otherwise, it was assigned a value of 0. Conversely, if the downregulated gene was higher, it was assigned a value of 0; otherwise, it was assigned a value of 1. The R software "neuralnet" and "NeuralNetTools" [22] packages were used to develop a predictive model featuring a single input layer, a single hidden layer, and a single output layer. The hidden layer was configured with five hidden nodes, and the output layer was designed with two nodes using a softmax activation function. The cross-entropy error function was set, and the optimization process involved selecting the maximum weight value of the DEGs within the hidden layer with values optimized accordingly [23].

2.5 Development and validation of new gene signatures

The molecular prognostic scoring system is a novel scoring system that has proven successful in predicting patients with breast cancer [23] and was implemented in this study to predict

TABLE 1. Information about the data sets used for training and validation groups.

Group	GEO number	Platform ID	Number of normal samples	Number of tumor samples
Training Group 1	GSE9750	GPL96	24	42
Training Group 2	GSE63514	GPL570	24	28
Training Group 3	GSE122697	GPL10558	5	11
Validation group	GSE7803	GPL96	10	21

GEO: Gene Expression Omnibus.

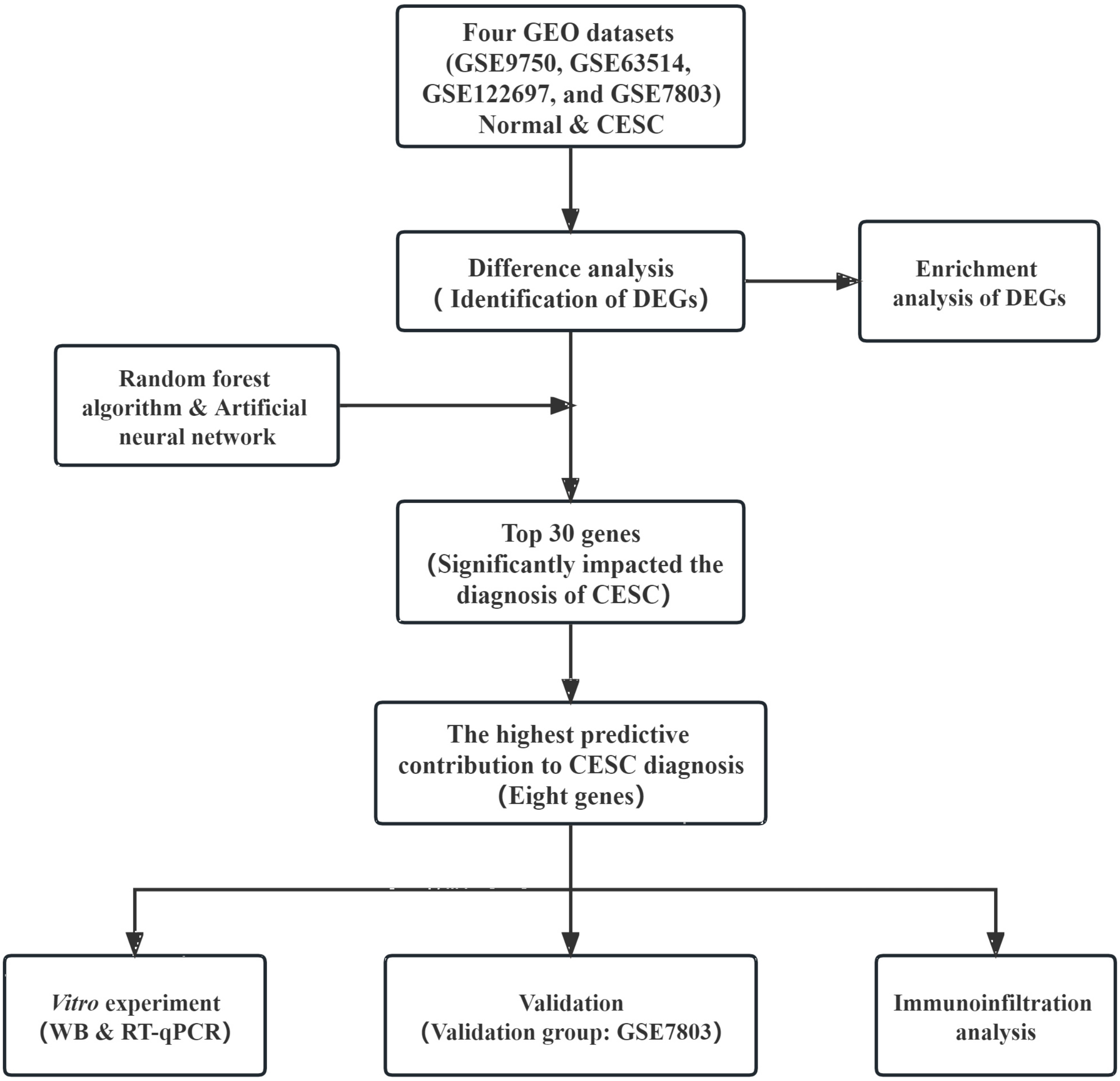

Four GEO datasets (GSE9750, GSE63514, GSE122697, and GSE7803) Normal & CESC
Difference analysis (Identification of DEGs)
Enrichment analysis of DEGs
Random forest algorithm & Artificial neural network
Top 30 genes (Significantly impacted the diagnosis of CESC)
The highest predictive contribution to CESC diagnosis (Eight genes)
Vitro experiment (WB & RT-qPCR)
Validation (Validation group: GSE7803)
Immunoinfiltration analysis

FIGURE 1. The flowchart of the study design. GEO: Gene Expression Omnibus; CESC: cervical squamous cell carcinoma; DEGs: differentially expressed genes; WB: Western blot; RT-qPCR: real-time quantitative polymerase chain reaction.

a new gene signature for the diagnosis of cervical squamous cell carcinoma (CESC). The scoring formula for evaluating each differentially expressed gene (DEG) was computed as follows: system score = (gene score × gene weight). The total system score of the eight DEGs with the highest predictive CESC diagnostic ability was then calculated. The publicly available GSE7803 dataset was utilized to validate the model, and the accuracy was assessed through the area under the curve (AUC) value of the receiver operating characteristic (ROC) curve generated using the "pROC" package in R. An AUC value greater than 0.8 indicated favorable accuracy, while an AUC value >0.9 indicated outstanding accuracy of the model [24].

2.6 Cell culture

The human cervical epithelial cell line HCerEpC (JZ-004978) and human cervical cancer cell line Hela (CL-0101) were cultured in MEM medium containing 10% fetal bovine serum and maintained in a 5% carbon dioxide (CO_2) incubator at 37 °C.

2.7 Western blot (WB) and real-time quantitative polymerase chain reaction (RT-qPCR)

Cell lysis buffer was used to lyse the collected HCerEpC and Hela cells, the proteins were extracted, and the protein concentration was determined. The protein levels of *MELK* (A10794, abclonal), *CRISP3* (14847-1-AP, proteintech), *GINS2* (A9172, abclonal), *DTL* (A12150, abclonal), *C1orf112* (PA5-55082, ThermoFisher), *KIF14* (A10275, abclonal), *SPINK5* (A20916, abclonal) and *CELSR3* (sc-293381) were detected by western blot (Primary Antibody Dilution Buffer for Western Blot & Secondary Antibody Dilution Buffer for Western Blot: Beyotime Biotechnology). Development was performed with the aid of a supersensitive luminescent substrate solution (Biosharp) using a protein gel imaging analyzer (Bio-Rad). Western blot was analyzed by Image J software (ImageJ 1.48v, National Institutes of Health, Bethesda, MD, USA). Total RNA was extracted from each cell and reverse-transcribed into cDNA using primer sequences (Sangon Biotech) listed in **Supplementary Table 1** and following the instructions of the detection kit (QIAGEN kit). The reaction procedures were as follows: PCR initial heat activation at 95 °C for 2 min; denaturation at 95 °C for 5 s, Combined annealing/extension at 60 °C for 30 s for a total of 40 cycles. β-Actin was used as an internal control. The mRNA expression of *MELK*, *CRISP3*, *GINS2*, *DTL*, *C1orf112*, *KIF14*, *SPINK5* and *CELSR3* was quantified using the $2^{-\Delta\Delta Ct}$ method [25].

2.8 Immuno infiltration analysis

The CIBERSORT R script v1.03 was used to calculate the infiltration score of each immune cell in the two groups of samples. The R software "corrplot" [26] and "vioplot" packages were used to analyze the correlation and difference of immune cells, following a correlation heat map and infiltration histogram were subsequently drawn.

2.9 Statistical analysis

All statistical analyses were performed using R language software (version 4.1.1). Statistical significance was set at $p < 0.05$.

3. Results

3.1 Identification of DEGs

A total of 241 DEGs were identified in the training dataset, and their expression profiles are displayed through heat maps and volcano plots (Fig. 2A,B). The heat map revealed that approximately half of the DEGs were highly expressed in CESC, while the rest exhibited low expression, thereby suggesting that both oncogenes and tumor suppressor genes were present among the DEGs. logFC of DEGs are listed in **Supplementary Table 2**.

3.2 Enrichment analysis of DEGs in the training dataset

Metascape (v3.5, The team of Yingyao Zhou, San Diego, CA, USA) was used to perform an enrichment analysis to better understand the functional and metabolic pathways associated with the DEGs. The results are depicted in a clustering network diagram and show the top 20 clusters in which the DEGs were significantly enriched (Fig. 3A).

The most significantly enriched biological processes were the "mitotic cell cycle process" and "keratinized envelope formation". Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis showed that the DEGs were primarily involved in the "vitamin D receptor pathway" and "Polo-like Kinase 1 (PID PLK1) pathway". Furthermore, GO circle diagrams were drawn to identify the regulatory role of the DEGs in the enrichment analysis. The majority of DEGs were upregulated in "mitotic sister chromatid separation, sister chromatid separation and organelle fission", while they were downregulated in "epidermal development and skin development" (Fig. 3B).

3.3 Random forest algorithm identification of disease signature genes

The expression data of 241 DEGs were integrated into the random forest algorithm classifier (Fig. 4A,B). A screening process was conducted to identify the eight disease signature genes with importance scores above 2, which were *SPINK5*, *CRISP3*, *MELK*, *CELSR3*, *GINS2*, *DTL*, *C1orf112* and *KIF14*. Except for *CRISP3* and *SPINK5*, the remaining six genes demonstrated high expression levels in CESC and low expression levels in the normal cervical epithelium (Fig. 4C). The neural network showed that *MELK*, *CRISP3*, *GINS2*, *DTL*, *C1orf112*, *KIF14*, *SPINK5* and *CELSR3* had good CESC diagnostic predicting (Fig. 4D).

3.4 Artificial neural network-based molecular prognostic scoring system

After converting the expression data of the eight disease-specific genes with importance scores greater than 2 into "gene scores", the weight values of each gene were optimized using an artificial neural network algorithm. The molecular prognostic score was calculated by summing the systematic scores ("gene score × gene weight") of these eight signature genes (Table 2). The molecular prognostic scores of 134 samples in the training group were then used as predictive values, with the presence or absence of CESC in patients as the true value. The AUC of the new gene model was 0.995, indicating the model's excellent predictive ability (Fig. 5).

TABLE 2. Gene weights of the eight disease signature genes in the training group.

Gene name	Gene weights	Gene name	Gene weights
MELK	15.559	*C1orf112*	15.404
CRISP3	4.960	*KIF14*	1.301
GINS2	2.765	*SPINK5*	22.798
DTL	23.042	*CELSR3*	15.393

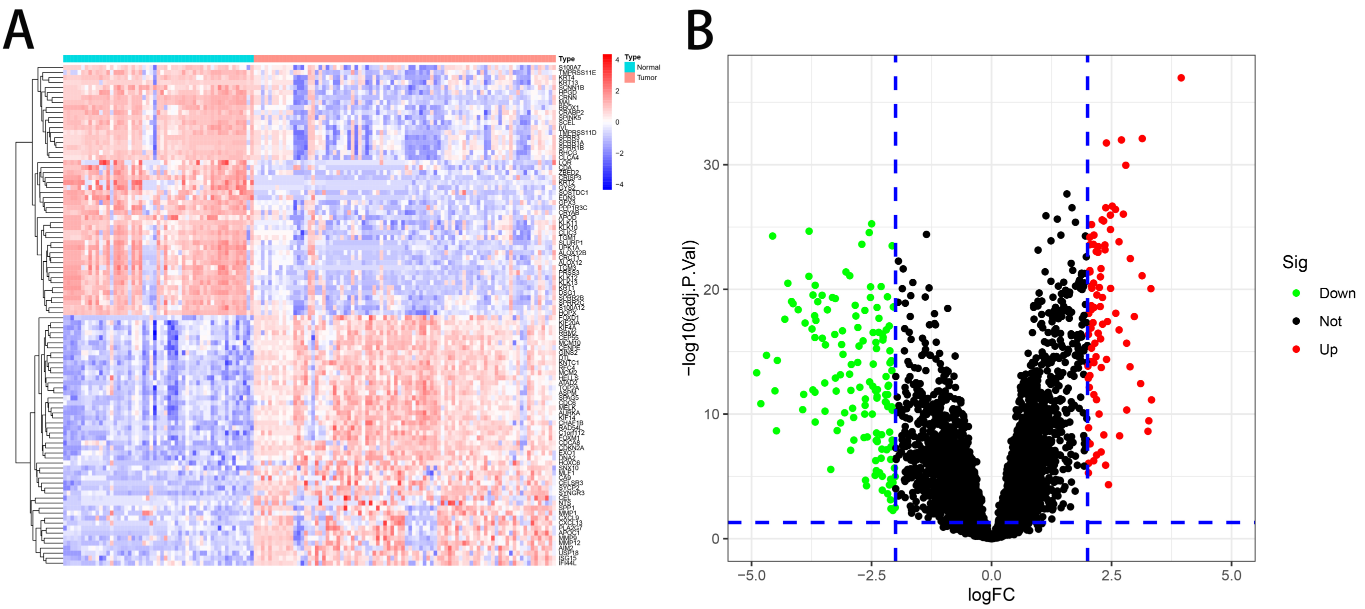


FIGURE 2. Expression of all differential genes in the training group. (A) Heat map (Red: High expression; Blue: Low expression). (B) Volcano map (Green: Down-regulated; Red: Up-regulated). logFC: log fold change; Sig: Significance Level.

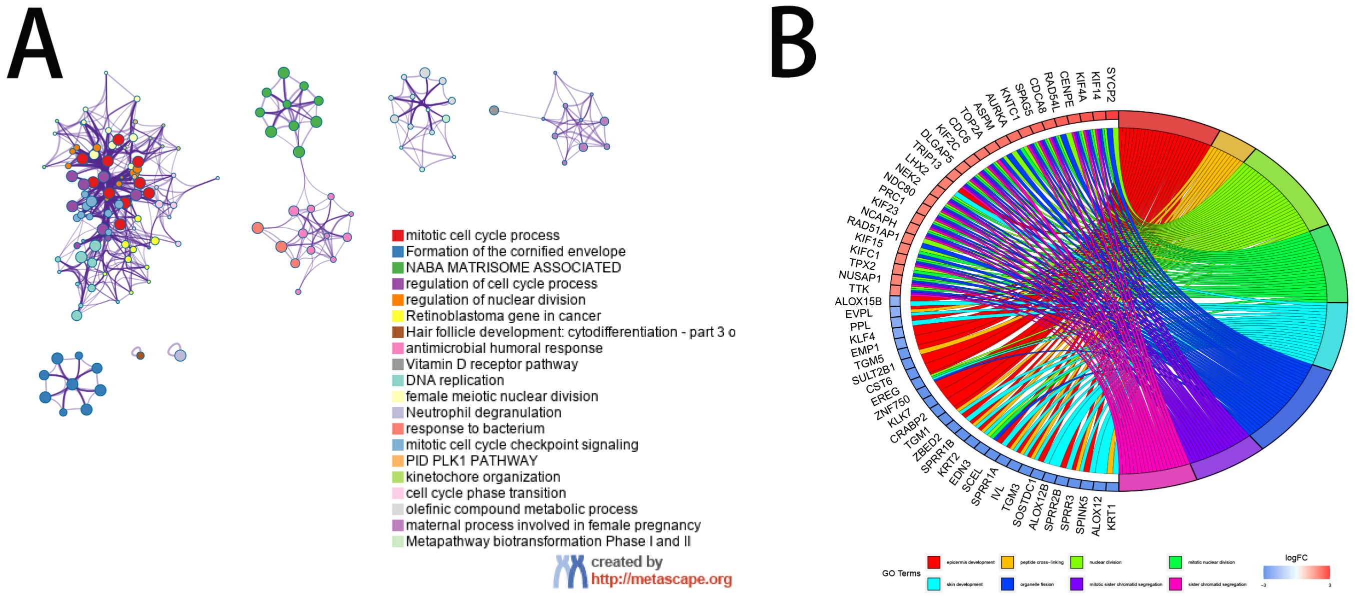

A
mitotic cell cycle process
Formation of the cornified envelope
NABA MATRISOME ASSOCIATED
regulation of cell cycle process
regulation of nuclear division
Retinoblastoma gene in cancer
Hair follicle development: cytodifferentiation - part 3 o
antimicrobial humoral response
Vitamin D receptor pathway
DNA replication
female meiotic nuclear division
Neutrophil degranulation
response to bacterium
mitotic cell cycle checkpoint signaling
PID PLK1 PATHWAY
kinetochore organization
cell cycle phase transition
olefinic compound metabolic process
maternal process involved in female pregnancy
Metapathway biotransformation Phase I and II
created by
http://metascape.org
B
SYCP2
KIF14
KIF4A
CENPE
RAD54L
CDCA8
SPAG5
KNTC1
AURKA
ASPM
TOP2A
CDC6
KIF2C
DLGAP5
TRIP13
LHX2
NEK2
NDC80
PRC1
KIF23
NCAPH
RAD51AP1
KIF15
KIFC1
TPX2
NUSAP1
TTK
ALOX15B
EVPL
PPL
KLF4
EMP1
TGM5
SULT2B1
CST6
EREG
ZNF750
KLK7
CRABP2
TGM1
ZBED2
SPRR1B
KRT2
EDN3
SCEL
SPRR1A
IVL
TGM3
SOSTDC1
ALOX12B
SPRR2B
SPRR3
SPINK5
ALOX12
KRT1
GO Terms
epidermis development
peptide cross−linking
nuclear division
mitotic nuclear division
skin development
organelle fission
mitotic sister chromatid segregation
sister chromatid segregation
logFC
−3
3

FIGURE 3. Enrichment analysis of all differential expressed genes in the training group. (A) Cluster network plot (Different colored squares represent different signaling pathways). (B) GO circle plot (Different colored squares represent different biological functions).

3.5 Validation of the prediction model

To validate the prediction model developed in the training set, an independent dataset (GSE7803) was used to evaluate its ability to diagnose CESC in other individuals. The random forest algorithm was utilized to select the top 30 DEGs and disease signature genes with importance scores greater than 2 from the validation set. The results showed that these genes were identical to those in the training set, demonstrating the scalability and stability of the random forest algorithm. Subsequently, the "gene score" and "molecular prognostic score" of GSE7803 were calculated using the same methodology in the testing set. The validation model generated a ROC curve with an AUC of 1.000, indicating that the prediction model is highly valid and stable (Fig. 6).

Next, we investigated the expression levels of the eight CESC-characteristic genes, namely *MELK*, *CRISP3*, *GINS2*, *DTL*, *C1orf112*, *KIF14*, *SPINK5* and *CELSR3*, in human cervical epithelial cells (HCerEpC) and human cervical carcinoma cells (Hela). The results revealed that the expression of the CRISP3 and SPINK5 protein and mRNA was high in HCerEpC and low in Hela cells, while the expression levels of MELK, GINS2, DTL, C1orf112, KIF14 and CELSR3 protein and mRNA were significantly lower in HCerEpC but significantly higher in Hela cells (Fig. 7A–I,8A–H, **Supplementary**

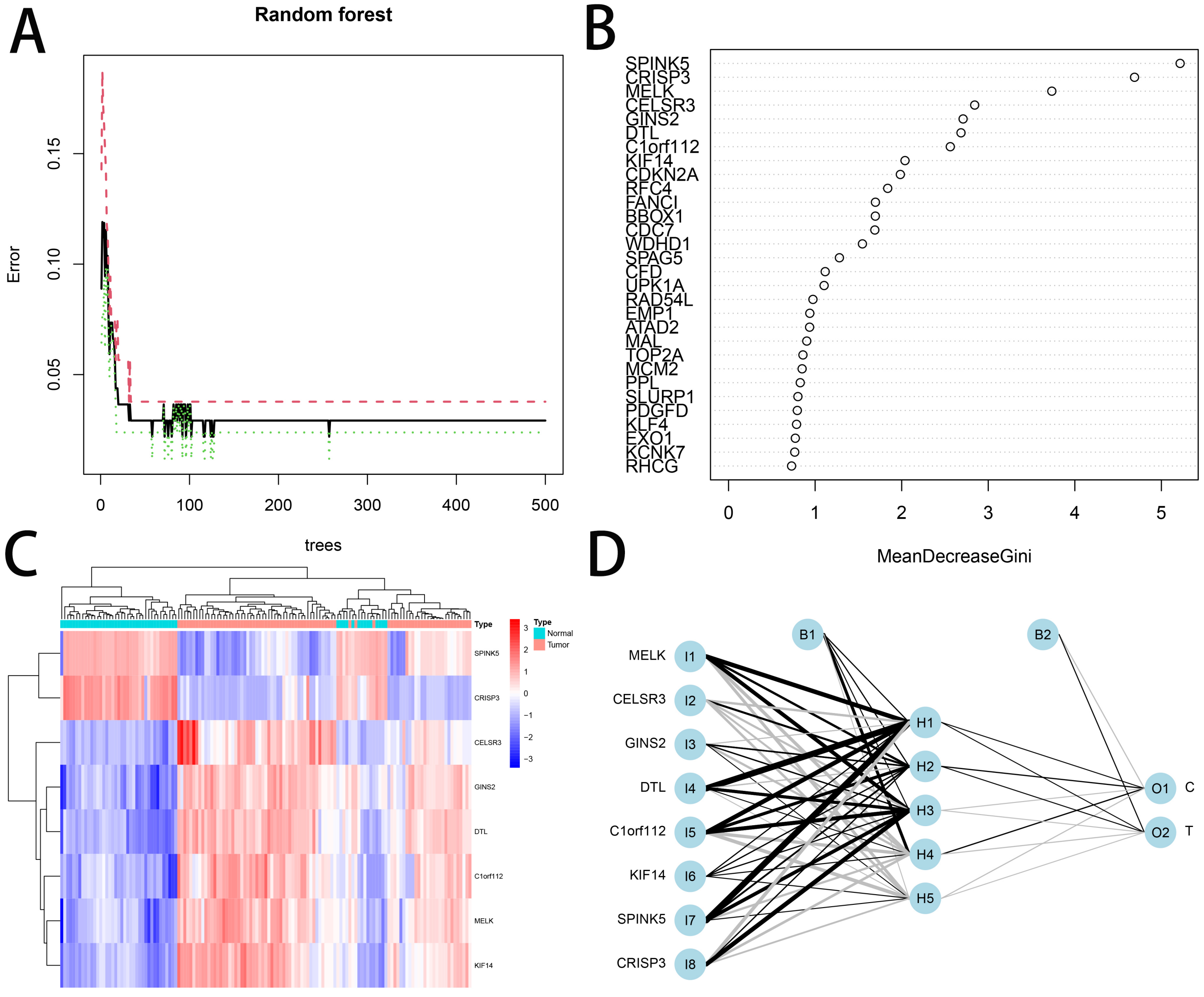


FIGURE 4. Random forest algorithm for identification of disease signature genes. (A) Random forest. (B) Random forest algorithm plot of the top 30 DEGs. (C) Heat map of DEGs with importance scores greater than 2. (D) Neural network model.

Table 3, **Supplementary Fig. 2**), indicating that these genes may have a role in the development and progression of CESC and could serve as potential targets for further research or therapeutic intervention.

3.6 Immunoinfiltration analysis

The analysis of immune infiltration in two groups based on the predicted model revealed a higher degree of macrophage infiltration in squamous cell carcinoma of the cervix than in the normal cervical epithelium (Fig. 9A). The quantitative comparison of immune infiltration in the two groups showed significant differences, based on the prediction model, in immune cells such as T cells CD8, T cells CD4 naive, T cells CD4 memory activated, Macrophages M0, Macrophages M1 and Dendritic cells resting ($p < 0.05$, Fig. 9B). Notably, Macrophages M0 were negatively correlated with Dendritic cells resting (−0.41), T cells CD8 (−0.33), and activated T cells CD4 memory (−0.14), while they were positively correlated with Macrophages M2 (0.37) (Fig. 9C). Collectively, these results suggest that the predicted model could be useful in analyzing immune infiltration in cervical cancer and identifying potential targets for immunotherapy.

4. Discussion

CESC is the most common pathological subtype of cervical cancer. Despite advancements in screening methods, over 50% of patients with CESC present at advanced stages, with limited treatment options and high recurrence and mortality rates [27]. Therefore, there is a need to establish a simple and efficient approach for the early diagnosis of CESC.

The random forest algorithm is a powerful tool for identifying disease-specific characteristic genes with high accuracy. The artificial neural network can further enhance the stability and dependability of the model due to its high tolerance for errors and scalability. Additionally, the molecular prognostic scoring system is a simple and effective tool for identifying heterogeneity and has been shown to be excellent in predicting disease prognosis [23]. This study established an innovative diagnostic model for CESC by integrating two machine learn-

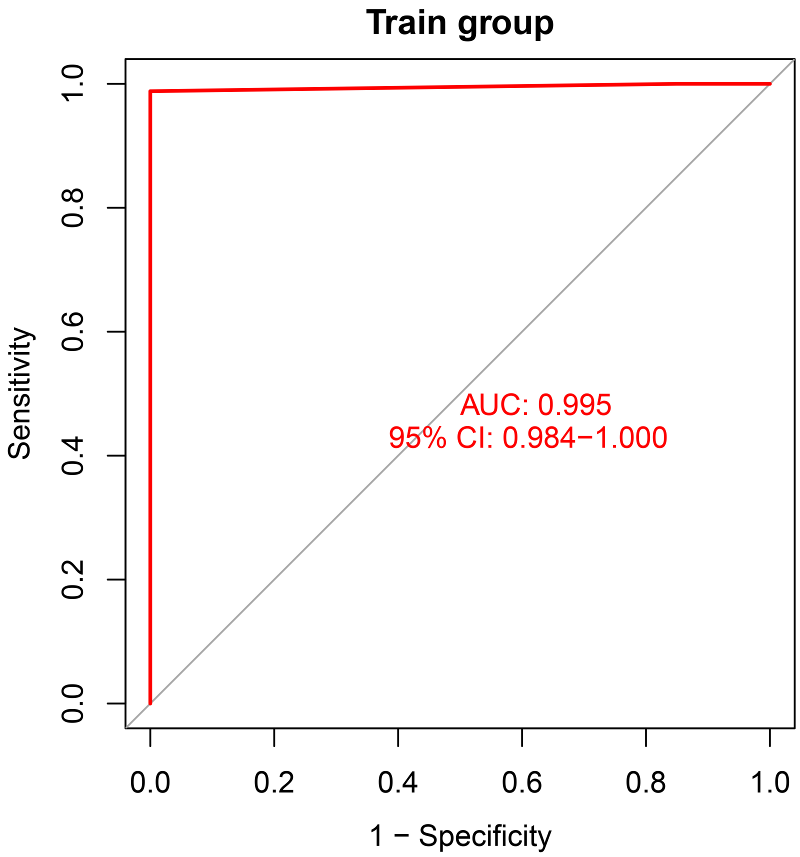


FIGURE 5. ROC plot of the training group evaluated of model accuracy. AUC: area under the curve; CI: confidence interval.

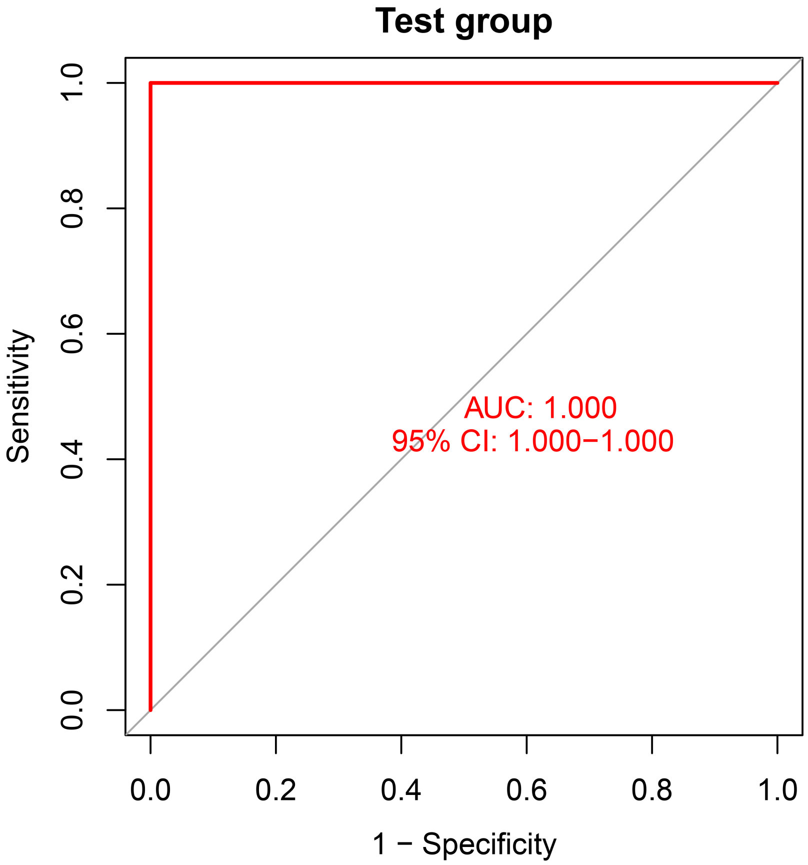


FIGURE 6. ROC plot of the testing group evaluated of model accuracy. AUC: area under the curve; CI: confidence interval.

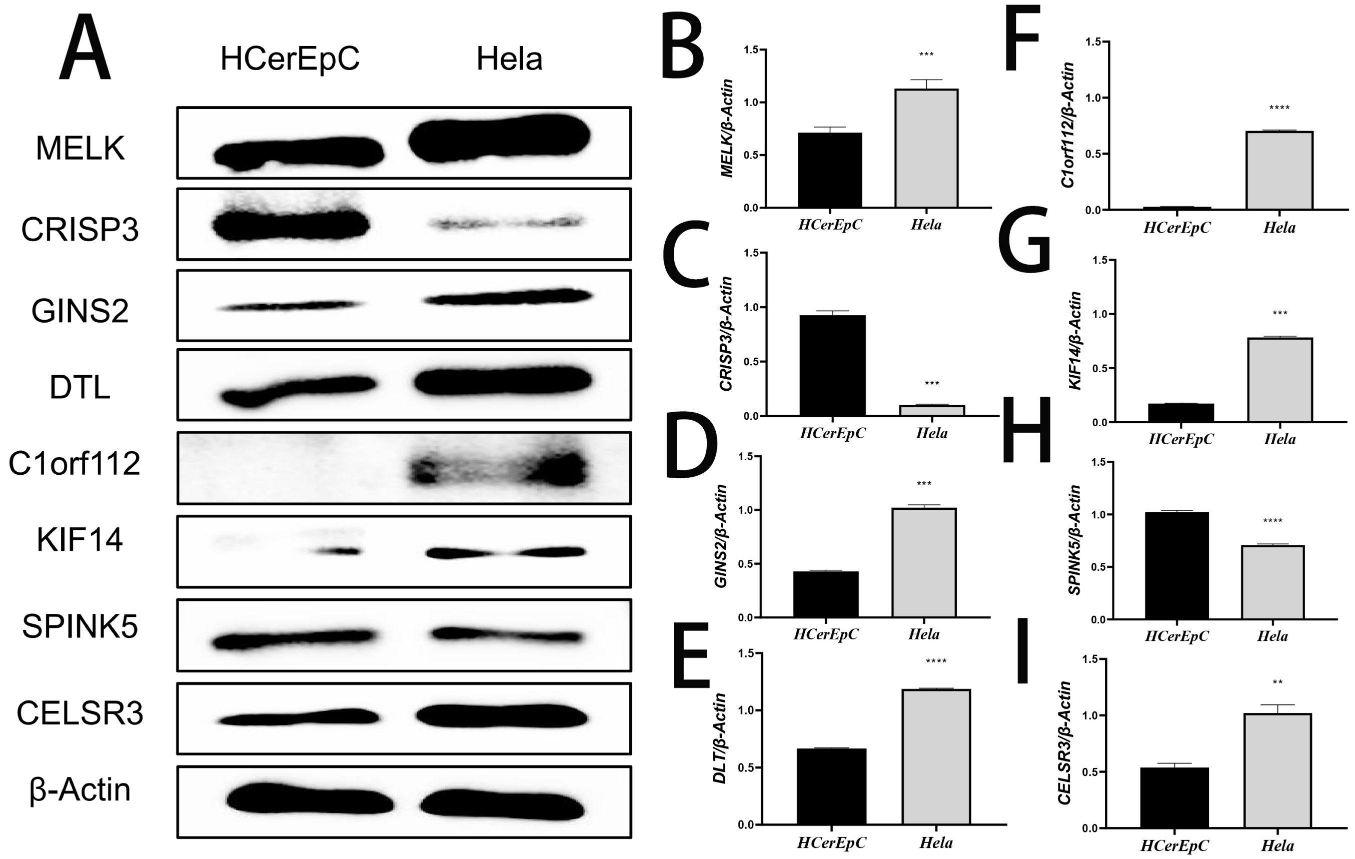


FIGURE 7. The expression of eight disease-specific proteins in human cervical epithelial cell line HCerEpC and human cervical cancer cell line Hela was detected by WB. (A) The total. (B) MELK. (C) CRISP3. (D) GINS2. (E) DTL. (F) C1orf112. (G) KIF14. (H) SPINK5. (I) CELSR3. **: $p < 0.01$; ***: $p < 0.001$; ****: $p < 0.0001$.

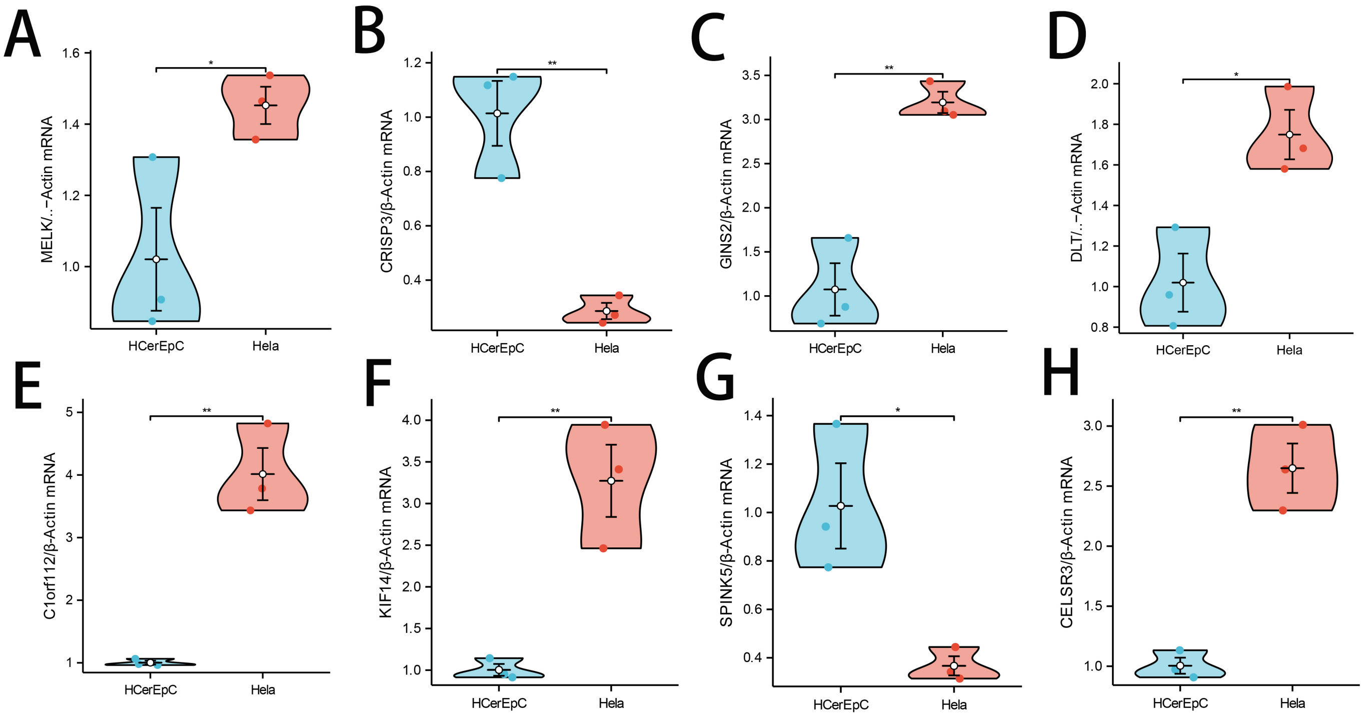
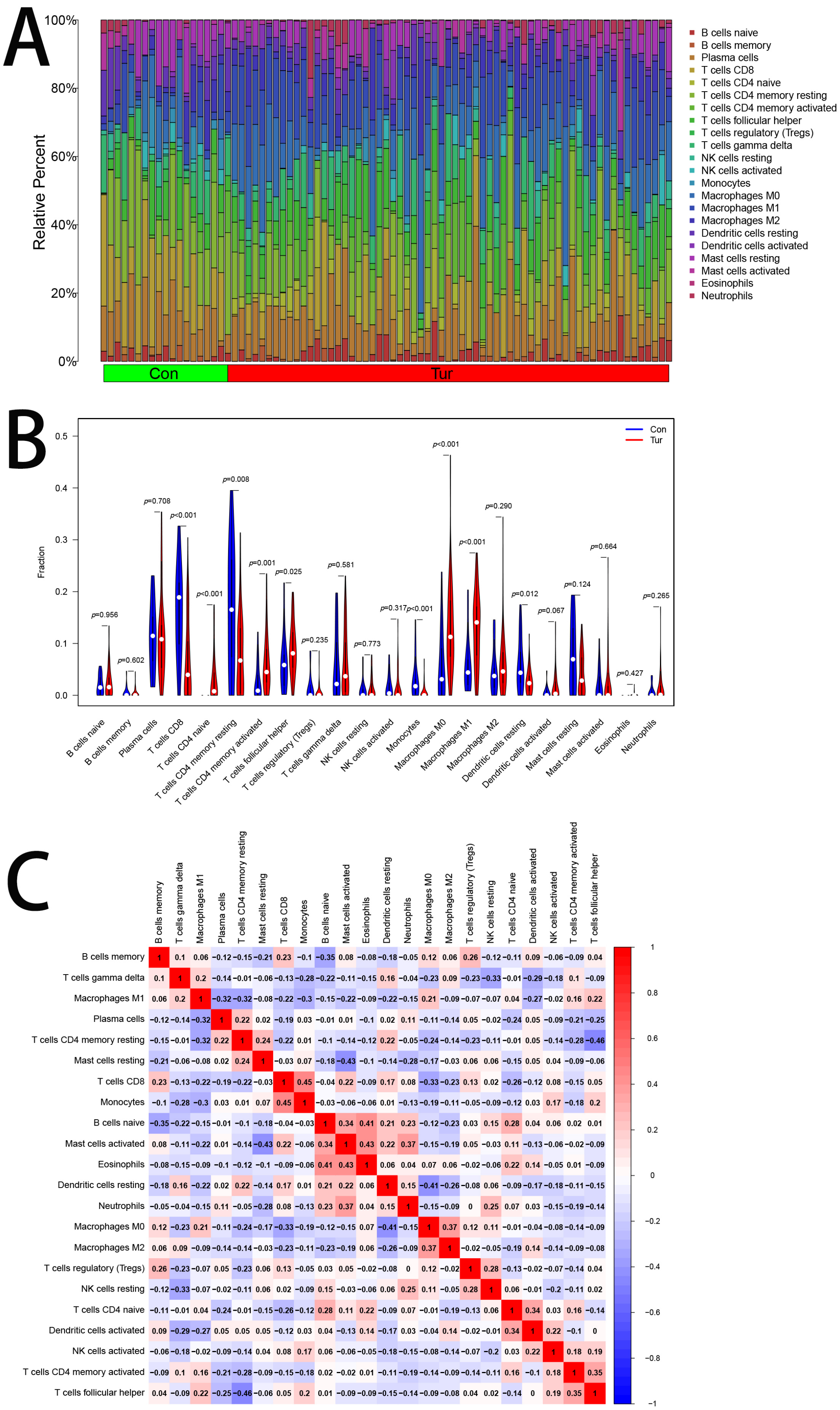


FIGURE 8. The expression of 8 disease characteristic genes in human cervical epithelial cells HCerEpC and human cervical cancer cells Hela was detected by RT-qPCR. (A) *MELK*. (B) *CRISP3*. (C) *GINS2*. (D) *DTL*. (E) *C1orf112*. (F) *KIF14*. (G) *SPINK5*. (H) *CELSR3*. *: $p < 0.05$; **: $p < 0.01$.



	B cells memory	T cells gamma delta	Macrophages M1	Plasma cells	T cells CD4 memory resting	Mast cells resting	T cells CD8	Monocytes	B cells naive	Mast cells activated	Eosinophils	Dendritic cells resting	Neutrophils	Macrophages M0	Macrophages M2	T cells regulatory (Tregs)	NK cells resting	T cells CD4 naive	Dendritic cells activated	NK cells activated	T cells CD4 memory activated	T cells follicular helper
B cells memory	1	0.1	0.06	−0.12	−0.15	−0.21	0.23	−0.1	−0.35	0.08	−0.08	−0.18	−0.05	0.12	0.06	0.26	−0.12	−0.11	0.09	−0.06	−0.09	0.04
T cells gamma delta	0.1	1	0.2	−0.14	−0.01	−0.06	−0.13	−0.28	−0.22	−0.11	−0.15	0.16	−0.04	−0.23	0.09	−0.23	−0.33	−0.01	−0.29	−0.18	0.1	−0.09
Macrophages M1	0.06	0.2	1	−0.32	−0.32	−0.08	−0.22	−0.3	−0.15	−0.22	−0.09	−0.22	−0.15	0.21	−0.09	−0.07	−0.07	0.04	−0.27	−0.02	0.16	0.22
Plasma cells	−0.12	−0.14	−0.32	1	0.22	0.02	−0.19	0.03	−0.01	0.01	−0.1	0.02	0.11	−0.11	−0.14	0.05	−0.02	−0.24	0.05	−0.09	−0.21	−0.25
T cells CD4 memory resting	−0.15	−0.01	−0.32	0.22	1	0.24	−0.22	0.01	−0.1	−0.14	−0.12	0.22	−0.05	−0.24	−0.14	−0.23	−0.11	−0.01	0.05	−0.14	−0.28	−0.46
Mast cells resting	−0.21	−0.06	−0.08	0.02	0.24	1	−0.03	0.07	−0.18	−0.43	−0.1	−0.14	−0.28	−0.17	−0.03	0.06	0.06	−0.15	0.05	0.04	−0.09	−0.06
T cells CD8	0.23	−0.13	−0.22	−0.19	−0.22	−0.03	1	0.45	−0.04	0.22	−0.09	0.17	0.08	−0.33	−0.23	0.13	0.02	−0.26	−0.12	0.08	−0.15	0.05
Monocytes	−0.1	−0.28	−0.3	0.03	0.01	0.07	0.45	1	−0.03	−0.06	−0.06	0.01	−0.13	−0.19	−0.11	−0.05	−0.09	−0.12	0.03	0.17	−0.18	0.2
B cells naive	−0.35	−0.22	−0.15	−0.01	−0.1	−0.18	−0.04	−0.03	1	0.34	0.41	0.21	0.23	−0.12	−0.23	0.03	0.15	0.28	0.04	0.06	0.02	0.01
Mast cells activated	0.08	−0.11	−0.22	0.01	−0.14	−0.43	0.22	−0.06	0.34	1	0.43	0.22	0.37	−0.15	−0.19	0.05	−0.03	0.11	−0.13	−0.06	−0.02	−0.09
Eosinophils	−0.08	−0.15	−0.09	−0.1	−0.12	−0.1	−0.09	−0.06	0.41	0.43	1	0.06	0.04	0.07	0.06	−0.02	−0.06	0.22	0.14	−0.05	0.01	−0.09
Dendritic cells resting	−0.18	0.16	−0.22	0.02	0.22	−0.14	0.17	0.01	0.21	0.22	0.06	1	0.15	−0.41	−0.26	−0.08	0.06	−0.09	−0.17	−0.18	−0.11	−0.15
Neutrophils	−0.05	−0.04	−0.15	0.11	−0.05	−0.28	0.08	−0.13	0.23	0.37	0.04	0.15	1	−0.15	−0.09	0	0.25	0.07	0.03	−0.15	−0.19	−0.14
Macrophages M0	0.12	−0.23	0.21	−0.11	−0.24	−0.17	−0.33	−0.19	−0.12	−0.15	0.07	−0.41	−0.15	1	0.37	0.12	0.11	−0.01	−0.04	−0.08	−0.14	−0.09
Macrophages M2	0.06	0.09	−0.09	−0.14	−0.14	−0.03	−0.23	−0.11	−0.23	−0.19	0.06	−0.26	−0.09	0.37	1	−0.02	−0.05	−0.19	0.14	−0.14	−0.09	−0.08
T cells regulatory (Tregs)	0.26	−0.23	−0.07	0.05	−0.23	0.06	0.13	−0.05	0.03	0.05	−0.02	−0.08	0	0.12	−0.02	1	0.28	−0.13	−0.02	−0.07	−0.14	0.04
NK cells resting	−0.12	−0.33	−0.07	−0.02	−0.11	0.06	0.02	−0.09	0.15	−0.03	−0.06	0.06	0.25	0.11	−0.05	0.28	1	0.06	−0.01	−0.2	−0.11	0.02
T cells CD4 naive	−0.11	−0.01	0.04	−0.24	−0.01	−0.15	−0.26	−0.12	0.28	0.11	0.22	−0.09	0.07	−0.01	−0.19	−0.13	0.06	1	0.34	0.03	0.16	−0.14
Dendritic cells activated	0.09	−0.29	−0.27	0.05	0.05	0.05	−0.12	0.03	0.04	−0.13	0.14	−0.17	0.03	−0.04	0.14	−0.02	−0.01	0.34	1	0.22	−0.1	0
NK cells activated	−0.06	−0.18	−0.02	−0.09	−0.14	0.04	0.08	0.17	0.06	−0.06	−0.05	−0.18	−0.15	−0.08	−0.14	−0.07	−0.2	0.03	0.22	1	0.18	0.19
T cells CD4 memory activated	−0.09	0.1	0.16	−0.21	−0.28	−0.09	−0.15	−0.18	0.02	−0.02	0.01	−0.11	−0.19	−0.14	−0.09	−0.14	−0.11	0.16	−0.1	0.18	1	0.35
T cells follicular helper	0.04	−0.09	0.22	−0.25	−0.46	−0.06	0.05	0.2	0.01	−0.09	−0.09	−0.15	−0.14	−0.09	−0.08	0.04	0.02	−0.14	0	0.19	0.35	1

FIGURE 9. Immunoinfiltration analysis. (A) Two groups of immune cell infiltration distribution. (B) Difference analysis of immune cells between two groups. (C) The correlation between immune cells.

ing techniques, the random forest algorithm and the artificial neural network, and applying the molecular prognostic scoring system. The combination of these techniques significantly improves the model's prediction accuracy, making it more applicable for clinical use and facilitating clinical decision-making and advancements. Further, the proposed model's validity was demonstrated through validation, and it was found to have outstanding predictive capabilities.

The present study yielded several significant findings. First, 241 differentially expressed genes (DEGs) were identified through consensus analysis of tissues from patients with normal cervical epithelium and cervical squamous cell carcinoma, and eight genes were identified as disease signature genes for cervical squamous cell carcinoma. Second, the prognostic model based on the molecular prognostic scoring system demonstrated excellent predictive ability. Third, the validation dataset confirmed the significant prediction ability of the model. Fourth, *in vitro* experiments revealed significant differences in the expression of the eight disease signature genes between the human cervical epithelial cell line HCerEpC and the human cervical cancer cell line Hela. Finally, immunoinfiltration analysis showed significant differences in multiple immune cell infiltrations between the two groups, with a higher degree of macrophage infiltration observed in squamous cell carcinoma of the cervix than in normal cervical epithelium.

Of the eight disease-characteristic genes obtained in this study, only one gene has been mentioned in cervical squamous cell carcinoma. Li X *et al.* [28] selected DTL as a biomarker of cervical squamous cell carcinoma through comprehensive multi-omics method, which is related to the development and diagnosis of CESC. The results are consistent with the present study.

Gene enrichment analysis revealed that many of the identified DEGs are involved in mitotic cell cycle processes and the formation of the keratinized envelope. Previous studies have reported that the human papillomavirus (HPV) type 16, a high-risk causative agent of CC, infects basal epithelial cells, interferes with their normal cell division, then colonizes viral DNA in the nucleus of the host cell, making them divide at a rate of 20–50 copies per cell [29, 30]. Studies have shown that altering the HPV viral genome can delay human foreskin keratinocyte (HFK) cell division and increase the number of viral genomes created by HFK, thereby disrupting the viral life cycle and leading to viral replication failure [30]. Targeting the HPV viral genome is a potential alternative therapy for CC. This present study identified several pathways, including the “vitamin D receptor pathway” and “PID PLK1 pathway”. Previous studies by Li X *et al.* [31] and Liu J *et al.* [32] showed that the main metabolites of vitamin D (25(OH)D, E2) are correlated with the clinical stage and differentiation degree of patients with CC and may contribute to the development and progression of CC to a certain extent. It has also been shown [33] that 1,25(OH)2D3 binding to the vitamin D receptor can inhibit the mitogenic effects of Insulin-like growth factor 1 (IGF-1) and IGF-2 and result in cell cycle arrest (G0/G1 phase). PLK1 plays multiple roles in the cell cycle, such as controlling mitotic entry and participating in cytoplasmic and meiotic divisions [34]. Previous studies [35, 36] have confirmed that PLK1 overexpression is associated with tumorigenic migration, and PLK1 inhibitors may be considered for the treatment of patients with cervical squamous cell carcinoma. Thus, proper interpretation of gene enrichment studies may improve our understanding of the molecular basis of CESC and serve as a theoretical foundation for developing novel substitute medications.

Macrophages are highly infiltrated in cervical squamous cell carcinoma, and the infiltration degree of M0 and M1 macrophages in cervical squamous cell carcinoma is significantly higher than that in normal cervical epithelium. Additionally, the infiltration degree of M2 macrophages in cervical squamous cell carcinoma is higher than that in normal cervical epithelium, although the difference between them was not significantly different. Macrophages [37] are immune effector cells with functional plasticity and can be classified into several types, including M1 macrophages, which promote inflammatory cytokines and chemokines and M2 macrophages, which secrete inhibitory cytokines [38]. Previous studies have confirmed that macrophages have dual effects on tumors, as they can both inhibit and promote tumor growth under specific environments [39]. In this study, we found that both M1-type pro-inflammatory macrophages and M2-type macrophages were highly infiltrated in cervical squamous cell carcinoma, and there was a significant difference in M1-type macrophage infiltration between cervical squamous cell carcinoma and normal cervical epithelial tissue. Previous studies have shown [40] that M1-type tumor suppressor macrophages promote tumor progression by transforming into M2-type macrophages. Given the significant role of macrophages in the tumor microenvironment and their immunomodulatory effects on tumors, targeting macrophages for immunotherapy may represent a novel therapeutic strategy for cervical squamous cell carcinoma.

Despite the promising results reported in this study, there are some limitations that must be considered. First, the sample size used to develop and validate the prediction model was relatively limited. Therefore, further studies with larger sample sizes are needed to validate the findings of this study. Second, although the model was validated using an independent dataset and *in vitro* experiments, further clinical validation is necessary to establish the clinical applicability of the model. Lastly, it is important to note that the developed model is intended only for predicting the diagnosis of CESC and cannot be used for other purposes.

5. Conclusions

In summary, this study developed a prediction model for CESC based on a molecular prognostic scoring system using the random forest algorithm and artificial neural network. The model's gene signatures demonstrated good predictive ability, which could aid clinical decision-making for the diagnosis of CESC. Additionally, the study revealed the potential of targeting macrophages for immunotherapy as a novel therapeutic strategy for cervical squamous cell carcinoma. Overall, this research provides valuable insights into the molecular mechanisms underlying CESC and suggests potential avenues for developing targeted therapies for this disease.

AVAILABILITY OF DATA AND MATERIALS

The datasets used and analyzed during the current study are available from Gene expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/).

AUTHOR CONTRIBUTIONS

GLW, JHG—designed the research study; wrote the manuscript. GLW, WZN—performed the research. QCL, PD—analyzed the data. All authors read and approved the final manuscript.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

ACKNOWLEDGMENT

Not applicable.

FUNDING

This research was funded by National Natural Science Foundation of China, grant number 82060469.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

SUPPLEMENTARY MATERIAL

Supplementary material associated with this article can be found, in the online version, at `https://oss.ejgo.net/files/article/1900457243270103040/attachment/Supplementary%20material.docx`.

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: A Cancer Journal for Clinicians. 2021; 71: 209–249.

[2] Zhen J, Liu YD, Song CH, Wu S, Bi XJ. The relationship between the expression of RNA-binding protein Musashi2 and cervical squamous-cell carcinoma. Tianjin Medical Journal. 2021; 49: 842–846.

[3] Hang S, Zhang H, Li W, Song YJ, Wang Q, Qi J. Expression of EMMPRIN in cervical cquamous cell carcinoma and its influence on the migration and invasion of cancer cells. Chinese Journal of Diagnostic Pathology. 2021; 28: 210–213. (In Chinese)

[4] Wei M, Chen S, Huang H. Research progress of comprehensive treatment mode for locally advanced cervical cancer. China Cancer. 2019; 28: 456–460. (In Chinese)

[5] Zhang J, Dong D, Wei Q, Ren L. CXCL10 serves as a potential serum biomarker complementing SCC-Ag for diagnosing cervical squamous cell carcinoma. BMC Cancer. 2022; 22: 1052.

[6] Worachartcheewan A, Shoombuatong W, Pidetcha P, Nopnithipat W, Prachayasittikul V, Nantasenamat C. Predicting metabolic syndrome using the random forest method. The Scientific World Journal. 2015; 2015: 581501.

[7] Zafeiris D, Rutella S, Ball GR. An artificial neural network integrated pipeline for biomarker discovery using Alzheimer's disease as a case study. Computational and Structural Biotechnology Journal. 2018; 16: 77–87.

[8] Bar-Yoseph H, Levhar N, Selinger L, Manor U, Yavzori M, Picard O, *et al*. Early drug and anti-infliximab antibody levels for prediction of primary nonresponse to infliximab therapy. Alimentary Pharmacology & Therapeutics. 2018; 47: 212–218.

[9] Scotto L, Narayan G, Nandula SV, Arias-Pulido H, Subramaniyam S, Schneider A, *et al*. Identification of copy number gain and overexpressed genes on chromosome arm 20q by an integrative genomic approach in cervical cancer: potential role in progression. Genes Chromosomes Cancer. 2008; 47: 755–765.

[10] den Boon JA, Pyeon D, Wang SS, Horswill M, Schiffman M, Sherman M, *et al*. Molecular transitions from papillomavirus infection to cervical precancer and cancer: role of stromal estrogen receptor signaling. Proceedings of the National Academy of Sciences. 2015; 112: E3255–E3264.

[11] Roychowdhury A, Samadder S, Das P, Mazumder DI, Chatterjee A, Addya S, *et al*. Deregulation of H19 is associated with cervical carcinoma. Genomics. 2020; 112: 961–970.

[12] Zhai Y, Kuick R, Nan B, Ota I, Weiss SJ, Trimble CL, *et al*. Gene expression analysis of preinvasive and invasive cervical squamous cell carcinomas identifies HOXC10 as a key mediator of invasion. Cancer Research. 2007; 67: 10163–10172.

[13] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, *et al*. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015; 43: e47.

[14] Zhang MY, Huo C, Liu JY, Shi ZE, Zhang WD, Qu JJ, *et al*. Identification of a five autophagy subtype-related gene expression pattern for improving the prognosis of lung adenocarcinoma. Frontiers in Cell and Developmental Biology. 2021; 9: 756911.

[15] Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, *et al*. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nature Communications. 2019; 10: 1523.

[16] Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, *et al*. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. The Innovation. 2021; 2: 100141.

[17] Ma X, Zhang X, Leng T, Ma J, Yuan Z, Gu Y, *et al*. Identification of oxidative stress-related biomarkers in diabetic kidney disease. Evidence-Based Complementary and Alternative Medicine. 2022; 2022: 1067504.

[18] Zhao Y, Huang T, Huang P. Integrated analysis of tumor mutation burden and immune infiltrates in hepatocellular carcinoma. Diagnostics. 2022; 12: 1918.

[19] Walter W, Sánchez-Cabo F, Ricote M. GOplot: an R package for visually combining expression data with functional analysis. Bioinformatics. 2015; 31: 2912–2914.

[20] Xin W, Zhang L, Zhang W, Gao J, Yi J, Zhen X, *et al*. An integrated analysis of the rice transcriptome and metabolome reveals root growth regulation mechanisms in response to nitrogen availability. International Journal of Molecular Sciences. 2019; 20: 5893.

[21] Li H, Lai L, Shen J. Development of a susceptibility gene based novel predictive model for the diagnosis of ulcerative colitis using random forest and artificial neural network. Aging. 2020; 12: 20471–20482.

[22] Beck MW. NeuralNetTools: visualization and analysis tools for neural networks. Journal of Statistical Software. 2018; 85: 1–20.

[23] Shimizu H, Nakayama KI. A 23 gene-based molecular prognostic score precisely predicts overall survival of breast cancer patients. EBioMedicine. 2019; 46: 150–159.

[24] Rice ME, Harris GT. Comparing effect sizes in follow-up studies: ROC area, Cohen's *d*, and *r*. Law and Human Behavior. 2005; 29: 615–620.

[25] Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(−Delta Delta C(T)) Method. Methods. 2001; 25: 402–408.

[26] Liu Z, Wang L, Xing Q, Liu X, Hu Y, Li W, *et al*. Identification of GLS as a cuproptosis-related diagnosis gene in acute myocardial infarction. Frontiers in Cardiovascular Medicine. 2022; 9: 1016081.

[27] Cohen PA, Jhingran A, Oaknin A, Denny L. Cervical cancer. The Lancet. 2019; 393: 169–182.

[28] Li X, Abdel-Maksoud MA, Iqbal I, Mubarak A, Farrag MA, Haris M, *et al*. Deciphering cervical cancer-associated biomarkers by integrated

multi-omics approach. American Journal of Translational Research. 2022; 14: 8843–8861.

[29] Campos SK. Subcellular trafficking of the papillomavirus genome during initial infection: the remarkable abilities of minor capsid protein L2. Viruses. 2017; 9: 370.

[30] Prabhakar AT, James CD, Das D, Otoa R, Day M, Burgner J, *et al*. CK2 phosphorylation of human papillomavirus 16 E2 on serine 23 promotes interaction with TopBP1 and is critical for E2 interaction with mitotic chromatin and the viral life cycle. mBio. 2021; 12: e0116321.

[31] Li X, Qin G, Tang J. Correlation analysis of serum 25-hydroxyvitamin D[25(OH)D] and estradiol levels with the incidence of cervical cancer. Jilin Medical Journal. 2021; 42: 1976–1977. (In Chinese)

[32] Liu J. Correlation between serum Hcy, Gal-9 and 1,25-(OH)2D; levels and grade of cervical lesions. China Medical Engineering. 2022; 30: 103–105. (In Chinese)

[33] Yin J, Ye Y. Research advances of vitamin D and its receptor in gynecologic cancer. Journal of Modern Oncology. 2022; 30: 1129–1133. (In Chinese)

[34] van de Weerdt BC, Medema RH. Polo-like kinases: a team in control of the division. Cell Cycle. 2006; 5: 853–864.

[35] Rizki A, Mott JD, Bissell MJ. Polo-like kinase 1 is involved in invasion through extracellular matrix. Cancer Research. 2007; 67: 11106–11110.

[36] Zhao C, Gong L, Li W, Chen L. Overexpression of Plk1 promotes malignant progress in human esophageal squamous cell carcinoma. Journal of Cancer Research and Clinical Oncology. 2010; 136: 9–16.

[37] Shapouri-Moghaddam A, Mohammadian S, Vazini H, Taghadosi M, Esmaeili S, Mardani F, *et al*. Macrophage plasticity, polarization, and function in health and disease. Journal of Cellular Physiology. 2018; 233: 6425–6440.

[38] Wu K, Lin K, Li X, Yuan X, Xu P, Ni P, *et al*. Redefining tumor-associated macrophage subpopulations and functions in the tumor microenvironment. Frontiers in Immunology. 2020; 11: 1731.

[39] Anderson NR, Minutolo NG, Gill S, Klichinsky M. Macrophage-based approaches for cancer immunotherapy. Cancer Research. 2021; 81: 1201–1208.

[40] Yunna C, Mengru H, Lei W, Weidong C. Macrophage M1/M2 polarization. European Journal of Pharmacology. 2020; 877: 173090.

How to cite this article: Guiling Wang, Wenzheng Nong, Qingchun Lu, Ping Du, Jinghua Gan. Development and validation of a novel gene signatures based on a random forest algorithm and artificial neural network for predictive diagnosis of cervical squamous cell carcinoma. European Journal of Gynaecological Oncology. 2025; 46(3): 34-45. doi: 10.22514/ejgo.2025.034.