WILEY | Hindawi

Research Article

Optimizing the Prognostic Model of Cervical Cancer Based on Artificial Intelligence Algorithm and Data Mining Technology

Yue Ma ⓘ,[1] Hongbo Zhu,[2] Zhuo Yang,[1] and Danbo Wang ⓘ[1]

[1]*Department of Gynecology, Cancer Hospital of China Medical University, Liaoning Cancer Hospital and Institute, Shenyang, Liaoning 110042, China*
[2]*The School of Information Science and Engineering, Shenyang Ligong University, Shenyang Liaoning 110159, China*

Correspondence should be addressed to Danbo Wang; wangdanbo@cancerhosp-ln-cmu.com

With the accumulation and development of medical multimodal data as well as the breakthrough in the theory and practice of artificial neural network and deep learning algorithm, the deep integration of multimodal data and artificial intelligence based on the Internet has become an important goal of the Internet of Medical Things. The deep application of the latest technologies in the medical field, such as artificial intelligence, machine learning, multimodal data, and advanced sensors, has a profound impact on the development of medical research. Artificial intelligence can achieve low-consumption and high-efficiency screening of specific markers due to its powerful data integration and processing capabilities, and its advantages are fully demonstrated in the construction of disease-related risk prediction models. In this study, multi-type cloud data were used as research objects to explore the potential of alternative CpG sites and establish a high-quality prognosis model of cervical cancer DNA methylation big data. 14,419 strict differentially methylated CpG sites (DMCs) were identified by ChAMP methylation analysis and presented these distributions based on different genomic regions and relation to island. Further, rbsurv and Cox regression analyses were performed to construct a prognostic model integrating these four methylated CpG sites that could adequately predict the survival of patients (AUC = 0.833, $P < 0.001$). The low- and high-risk patient groups, divided by risk score, showed significantly different overall survival (OS) in both the training ($P < 0.001$) and validation datasets ($P < 0.005$). Moreover, the model has an independent predictive value for FIGO stage and age and is more suitable for predicting survival time in patients with histological type (SCC) and histologic grade (G2/G3). Finally, the model exhibited much higher predictive accuracy compared to other known models and the corresponding expression of genes. The proposed model provides a novel signature to predict the prognosis, which can serve as a useful guide for increasing the accuracy of predicting overall survival of cervical cancer patients.

1. Introduction

The Internet of Things (IoT) is a perceptual network of everything connected based on computer network technology [1]. Its main supporting technologies are sensor network technology and Internet technology, which belongs to the lower and upper layers of network architecture, respectively. Artificial intelligence (AI) is a form of intelligence based on computer systems that simulate inanimate objects. AI is analogous to software and requires the IoT as the carrier, while the IoT is similar to hardware and requires artificial

intelligence to drive it. With the in-depth excavation of the IoT and AI technology, the IoT technology has been widely used in intelligent transportation, intelligent fire protection, intelligent agriculture, environmental protection, water monitoring, and medical and health care [2, 3]. Medicine as an important application field, the IoT technology is used to collect, transmit, and store diversified medical information, so as to realize the intelligence of resources, information sharing, and interconnection, also known as the Internet of Medical Things (IoMT). Simultaneously, AI is able to process, integrate, and analyze multisource and

multimode medical data to achieve large-scale multimodal disease data fusion and promote accurate analysis of disease data. Therefore, the effective integration of the two means is an important means to promote "4P" medicine, namely, preventive, predictive, personalized, and participatory medicine.

In the field of tumor research, with the rapid development of second-generation sequencing and computer analysis technology in recent years, the exponential growth of genome, transcriptome, epigenome, proteome, and other data has been recorded, and biomedical data has been boosted to the PB era [4]. The sharing of biomedical big data on a global scale has led to a fundamental revolution in medical research models [5]. Since IoMT and AI algorithm have spilled over into the field of medical research, medical researchers have begun to try to take advantage of AI technology, especially the classification algorithm of machine learning, to integrate, process, and analyze biomedical high-dimensional data of tumor so as to reveal the inherent mechanism of tumor development, thereby providing a theoretical basis for the realization of individualized precision diagnosis and treatment of tumor [6]. As is known, machine learning, an important component of AI algorithms, is aimed at generating an informed assessment by using numerical algorithms to detect relationships in information, which has the advantage of being able to computerize the hypothesis construction methods and, in some cases, optimize traditional statistical methods [7]. Machine learning possesses obvious advantages in mining and analyzing complex multimode big data, as well as potential clinical application value in constructing tumor-related risk models [8, 9]. Therefore, the application of relevant AI algorithms to analyze multilevel and multiform data is playing an increasingly important role in tumor diagnosis and prognosis evaluation. M. Gupta and B. Gupta fused gene sequencing and clinicopathological data and used machine learning to extract the genes which were the most significantly related to the occurrence of breast cancer and analyze the selected genes, which was conducive to reducing the cost and time of early diagnosis of breast cancer [10]. Zhang et al. applied machine learning algorithm to analyze the expression profile of lung cancer miRNA in The Cancer Genome Atlas (TCGA) database, thereby obtaining the characteristic miRNA and classification model related to lung cancer through difference analysis and model training of miRNA expression profile of lung cancer and healthy individuals, so as to achieve accurate diagnosis of lung cancer at the level of DNA computation [11]. At present, many studies have suggested that the construction of gene model is of great significance for the prediction of clinical outcome and treatment guidance of cervical cancer (CC) [12, 13]. Regrettably, there is no uniform and high-efficiency prediction model of gene that has been proved to be effective in practical application. DNA methylation is a stable target and allows for flexibility of assay development. Certain DNA methylation sites have been shown to affect the expression of genes and be associated with the prognosis of cervical cancer [14]. However, there is still a lack of thorough exploration of genome-wide DNA methylation data of cervical cancer based on AI algorithm. Therefore, it is an indispensable research direction

for accurate diagnosis and treatment of cervical cancer through analyzing and processing DNA methylation and clinical data by AI algorithm and discovering specific prognostic DNA methylation markers.

In the present study, we used AI algorithm to analyze genome-wide DNA methylation profile data of CC from TCGA database in order to obtain specific differentially methylated CpG sites (DMCs) and distribution features and construct a prediction model of DMCs to evaluate the prognosis of CC. Clinical information, single methylation site, and known valid prognostic biomarkers were compared with the proposed one to evaluate the performance of this prediction model. Further validation to identify this prediction model can accurately and effectively predict the prognosis of CC without relying on clinicopathological parameters.

2. Materials and Methods

2.1. Data Sources. The intensity data (IDAT) files of DNA methylation and clinical information were downloaded from TCGA database (http://cancergenome.nih.gov). The RSEM-normalized mRNA datasets and preprocessed mature miRNA-normalized expression profiles were accessed from the Firebrowse portal (http://firebrowse.org/). Both mRNA and miRNA data were transformed by the transformation log2 (Exp+1), where Exp was the original expression value. The inclusion criteria for samples were set as follows: (1) DNA methylation, gene expression, miRNA expression, and clinical information were available; (2) specimens were the primary tumor tissue; (3) complete prognostic follow-up data were available. Finally, a cohort of 299 patients with CC, which included 3 pairs of matched cancer and adjacent cervical tissues, was identified.

2.2. Identification of Differentially Methylated CpG Sites (DMCs). DMCs were identified using the Chip Analysis Methylation Pipeline (ChAMP) methylation analysis package. The algorithm of DMCs mainly applied the robust empirical Bayes machine learning [15]. In short, the procedure involved loading the data from the IDAT files, filtering it using predetermined settings, quality control, and normalization using the "Functional Normalization" method. A statistical cut-off of Benjamini-Hochberg (BH)/adjusted $P < 0.05$, and $|\Delta \beta| > 0.2$ was used to select associated DMCs. In addition, to identify the most significant and relevant cancer-specific DMCs, $|\Delta \beta| > 0.2$ and average tumor tissue methylation level > 0.6 or average normal tissue methylation level < 0.2 was used as a strict filtering criteria for selection.

2.3. Construction of the Prognostic Model for CC. All of the samples were randomly distributed to a training dataset, and the validation dataset ratio was set to 6 : 4. Cox proportional hazard regression analysis was used to develop the proposed hazard model. First, univariate Cox regression analysis was applied to screen the CpG sites for those significantly related to the prognosis of CC in the training dataset ($P < 0.05$). Following this, robust likelihood-based survival modeling (rbsurv in R) was used to identify the more
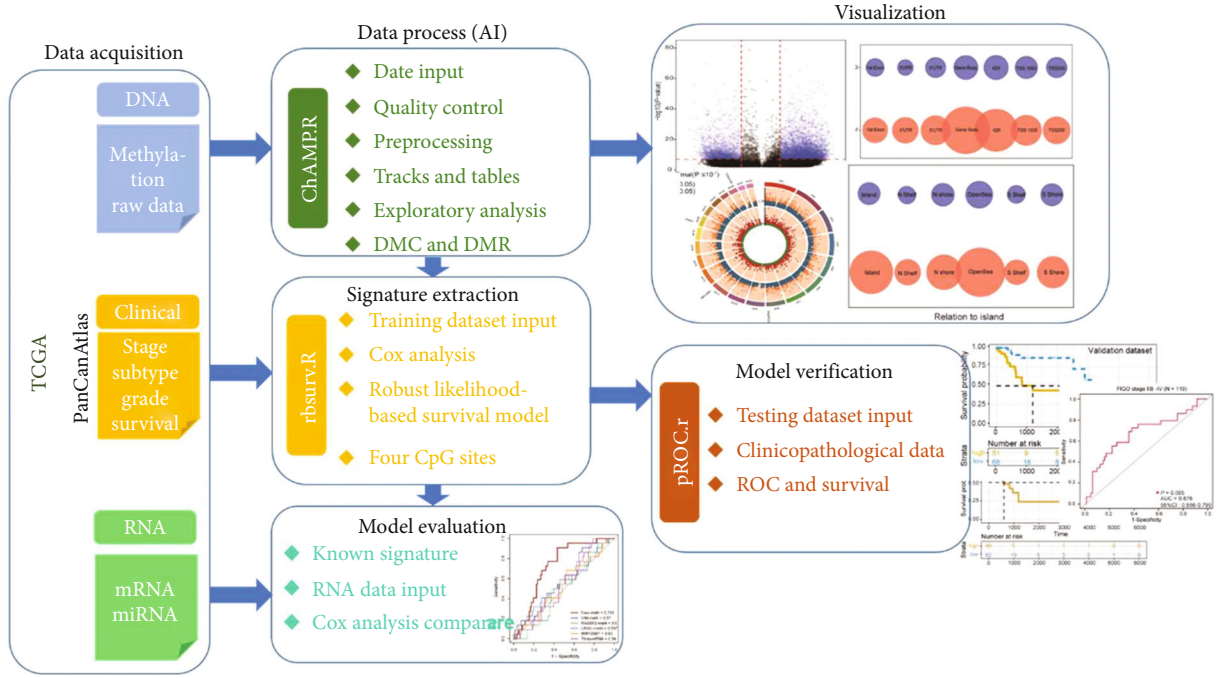
FIGURE 1: Flowchart of the study.

significant CpG sites from the results of the univariate analysis [16]. All of the alternative CpG sites were subjected to multivariable Cox regression analysis to further filter the markers associated with OS. According to $P < 0.05$, several markers were screened out as covariates to construct the model. The predicted risk score of the model was constructed to predict OS by using the regression coefficient (β) from the multivariate Cox regression model as follows:

$$\text{Risk score} = \sum_{i=1}^{n}(\text{Exp}(i) * \beta(i)). \quad (1)$$

Optimization risk score cut-off threshold values were selected based on the concordance index (C-index) by plotting cross-validated time-dependent receiver operating characteristic (ROC) curves.

2.4. Validation of the Prognostic Model for CC. Patients were classified into "high-risk" and "low-risk" groups based on their prognostic risk score cut-off value. The survival conditions of the two groups were compared with the log-rank tests conducted in the training dataset and validation dataset. The survival curves were then plotted using the "survival" R package. The effect of the model was further evaluated by differentiating subgroups according to different clinicopathological characteristics. The "pROC" package was used to perform ROC analysis. Moreover, by comparing with other known biomarkers and corresponding RNA, we establish the performance of the four-CpG site biomarker using the Z test.

3. Results

3.1. Characterization of the Study Population. The workflow was performed as indicated in Figure 1.

A total of 299 samples of patients diagnosed with CC were used in this study. The sample sizes of the training and validation datasets were 180 and 119, respectively. The median age of the patients at the point of initial diagnosis was 46.5 years (age range, 20–88 years). The median survival time was 2,888 days. The characteristics of the patients have been summarized in Table 1.

3.2. Identification of DMCs in CC. The methylation expression matrix was obtained by data washing as described, and 364,001 methylation sites were used for analysis. Focal analysis identified 34,389 DMCs in the CC samples using a fixed statistical cut-off (adjusted $P < 0.05$ and $|\Delta\beta| > 0.2$). Tumor tissue had a greater number of hypermethylated CpG sites as compared to hypomethylated sites (Figure 2(a)). Due to the strict filtering criteria, the number of DMCs dwindled to 14,419. All DMC distributions of the autosomal chromosomes were depicted using the $P < 1 \times 10^{-7}$ as a cut-off point in the circular plot (Figure 2(b)). The percentage of hypermethylated CpG sites located in the gene body region was far greater than that of other genomic regions. Conversely, in hypomethylated CpG sites, intergenic regions (IGRs) had similar proportions with gene body regions. We also found that the proportion of DMCs in opensea regions was the largest in each type of relation to island compared to that of other non-CpG island regions (Figure 2(c)).

3.3. Confirmation of the Four-CpG Biomarker Model Closely Related to the OS of Patients in the Training Dataset. A total of 1,707 CpG sites out of all the candidate sites were

TABLE 1: Clinicopathological characteristics of cervical cancer patients.

Characteristics	Groups	Total ($N = 299$)	Patients Training dataset ($N = 180$)	Validation dataset ($N = 119$)
Age at initial diagnosis	Median	46.5	47	46
	Range	20-88	24-81	20-88
	<50	177	102 (57.63)	75 (42.37)
	≥50	122	78 (63.93)	44 (36.07)
FIGO stage	I	158	96 (60.76)	62 (39.24)
	II	5	3 (60.00)	2 (40.00)
	IIA	21	13 (61.90)	8 (38.10)
	IIB	43	24 (55.81)	19 (44.19)
	III	44	23 (52.27)	21 (47.73)
	IV	21	15 (71.43)	6 (28.57)
	Unknown	7	6 (85.71)	1 (14.29)
Histological type	SCC	247	142 (57.49)	105 (42.51)
	Adenocarcinoma	46	33 (71.74)	13 (28.26)
	Adenosquamous	6	5 (83.33)	1 (16.67)
Histologic grade	G1	16	8 (50.00)	8 (50.00)
	G2	131	78 (59.54)	53 (40.46)
	G3	120	71 (59.17)	49 (40.83)
	Others	32	23 (71.88)	9 (28.12)
Neoplasm statue	Tumor free	186	112 (60.22)	74 (39.78)
	With tumor	71	45 (63.38)	26 (36.62)
	Unknown	42	23 (54.76)	19 (45.24)

G: grade; SCC: squamous cell carcinoma. Values are shown as n (%) unless otherwise specified.

identified in univariable Cox regression analysis ($P < 0.05$). After robust likelihood-based survival modeling analysis, the first 17 significant prognosis-related CpG sites were chosen ($P < 0.05$). Subsequently, in the multivariable Cox regression analysis, the four most significant CpG sites (cg06661994, cg07281370, cg07141215, and cg11256152) were screened out to construct a multivariate hazard ratio (HR) model.

The corresponding gene symbols of these four CpG sites were C20orf195, MIR125B1, TFAP4, and TRAPPC9, respectively. Only the methylation level of cg06661994 presented directly proportional to the risk of death, with an HR of 6.752, while the other three CpG sites were presented as protective factors. The site information of the four CpG sites and related their risk coefficients are shown in Table 2.

Hazard ratios (HRs) of the four-CpG biomarker by the Cox regression analysis were significantly correlated with the OS of all samples ($P < 0.001$, HR = 2.97, and 95% CI 1.97–4.47). The risk-score formula was as follows:

$$\text{Risk score} = 6.752 \times \beta \text{ value of } cg06661994 - 2.490 \\ \times \beta \text{ value of } cg07281370 - 3.328 \\ \times \beta \text{ value of } cg07141215 - 6.480 \\ \times \beta \text{ value of } cg11256152. \tag{2}$$

Furthermore, we evaluated the multivariable hazard model using the proportional hazard (PH) assumption.

The P value of the global model was 0.660, illustrating that the PH hypothesis was established (Table 3).

Accordingly, a Cox regression model was successfully established. The AUC was 0.833 ($P < 0.001$), demonstrating that a model composed of these four CpG sites presented high sensitivity and specificity in predicting the prognosis of patients with CC. Meanwhile, according to the C-index, 1.067 was selected as the optimal cut-off risk score for the model, which was more relevant in predicting survival (Figure 3(a)).

3.4. Validation of the Prognostic Value of the Four-CpG Biomarker Model. According to the optimal risk score cutoff value, the biomarker composed of four CpG sites was used to categorize the patients into either the high-risk ($n = 132$) or low-risk ($n = 167$) group. The individual methylation levels of these four CpG sites were assessed in both groups. The result illustrated that these four candidate CpG sites could differentiate between patients with high- and low-risk of CC (Figure 3(b)). Kaplan–Meier analysis was performed to verify the predictive value of the prognostic hazard model composed of the four-CpG biomarker model. The OS of the low-risk group was significantly better than that of the high-risk group in both the training and validation datasets (Figure 3(c)).

3.5. The Predictive Value of the Four-CpG Biomarker Model in Prognosis Based on Various Clinical Risk Factors. Several clinicopathological characteristics are associated with poor prognosis of CC including age, FIGO stage, tumor size,
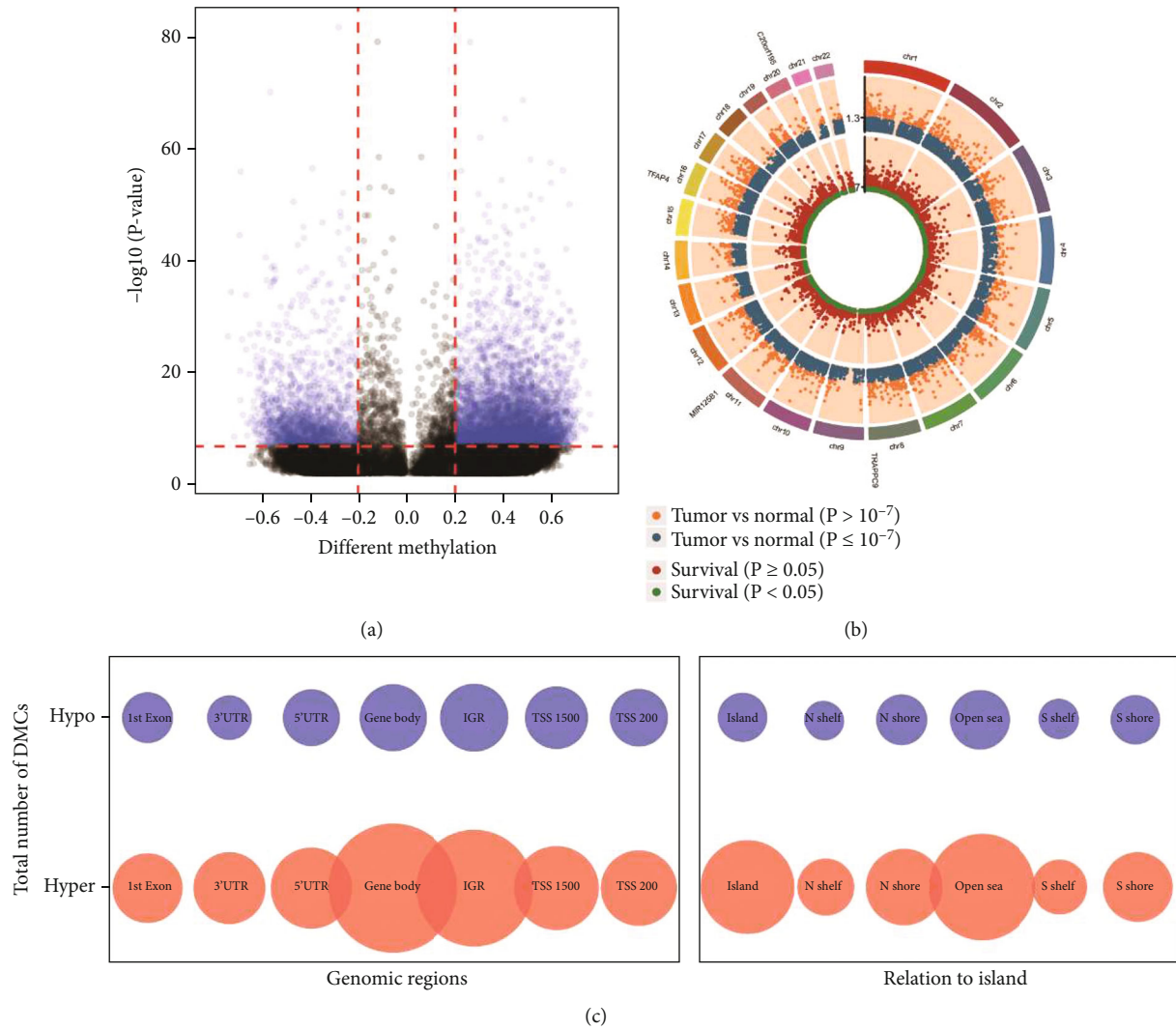
(a)

(b)

(c)

Figure 2: DMC analysis of cervical cancer and paracancerous samples. (a) Volcano plot of all DMCs ($|\Delta\beta| > 0.2$). (b) Circos plot of filtered DMCs. The outermost circle presented chromosome distribution, and the two innermost circles represent the DMCs bounded by $P < 1 \times 10^{-7}$ and univariable Cox regression analysis of the training dataset bounded by $P < 0.05$, respectively. (c) Bubble plot of filtered DMCs in different genomic regions and relation to island. Abbreviation: TSS: transcription start site; IGR: intergenic region; UTR: untranslated region.

lymph node metastasis, histological type, and histological grade. According to the different principles of therapy, 292 patients with known staging were divided into two sets: FIGO stages I–IIA2 ($n = 182$, 60.88%) and IIB–IV ($n = 110$, 36.79%). Five patients were diagnosed as FIGO stage II, according to TNM staging; three patients identified as T2a were assigned to the former set, and two patients diagnosed as T2b were assigned to the latter set. As determined by the Kaplan–Meier analysis, the low-risk patients had better survival ratio than the high-risk risk patients regardless of subgroup ($P < 0.001$). The AUC values of the four-CpG biomarker model in the two FIGO stage sets were 0.693 and 0.676, respectively (Figure 4).

Next, in the two subgroups of age, OS was also significantly increased in the low-risk group compared with that of the high-risk group ($P < 0.001$). Additionally, in squamous cell carcinoma, we found that the low-risk group had a longer OS ($P < 0.001$), while the AUC also revealed that

the model had a high diagnostic value. However, potentially due to the low number of adenocarcinomas, its diagnostic applicability could not be determined using the ROC curve even though Kaplan–Meier analysis showed differences between the two groups. With respect to the histological grade, due to the restrictions in the number of samples, G2 and G3 were selected to verify the predictive power. The patients in the high-risk group demonstrated shorter OS, and the AUCs were 0.627 and 0.762 for G2 and G3, respectively. A summary analysis has been presented in Table 4.

Thus, the four-CpG biomarker model revealed suitable applicability when patients were stratified with respect to different clinicopathological characteristics. This suggests that the biomarker has an independent predictive value of the OS for patients with CC.

3.6. Comparison of the Four-CpG Biomarker Model with Other Biomarkers. Several biomarkers and their prognostic

Table 2: The information of the four significantly survival-related CpG sites.

Probe ID	Gene symbol	Position	Genomic features	CGI	P value*	Coef#	P value#
cg06661994	C20orf195	Chr20: 62184363	TSS200	Shore	2.78E-03	6.752	3.79E-03
cg07281370	MIR125B1 (mir-125b-1)	Chr11: 121970768	TSS200	Opensea	6.82E-03	-2.499	2.78E-03
cg07141215	TFAP4	Chr16: 4308070	Body	Shelf	3.70E-05	-3.328	1.91E-02
cg11256152	TRAPPC9	Chr8: 141359786	Body	Shore	5.51E-04	-6.480	1.72E-04

CGI: CpG island; ∗ is in univariate Cox regression analysis; # is in multivariate Cox regression analysis.

Table 3: The proportional hazard (PH) assumption test of the four-CpG biomarker model.

Methylation site	Rho	Chisq	P
cg06661994	0.0374	0.0533	0.817*
cg07281370	0.1923	1.4901	0.222*
cg07141215	-0.0789	0.1533	0.695*
cg11256152	0.0438	0.0973	0.755
GLOBAL	NA	2.4169	0.660#

*All the P values of the five covariates were greater than 0.05, indicating that each variable met the PH assumption. #The P value of global (0.660) was also not statistically significant; the whole model met the PH assumption.

roles in CC have been illustrated in previous studies. The hypermethylation of VIM and RASSF2 has been reported as a favorable biomarker for the prognosis of cervical squamous cell carcinoma [17, 18]. Moreover, the methylation status of LRIG1, an important tumor suppressor, combined with gene loss is a poor prognostic factor in CC [19]. Therefore, to verify the reliability and stability of the four-CpG biomarker model in forecasting patient survival, ROC analyses were performed on these biomarkers, as well as the corresponding expression of the three-mRNAs (C20orf195, TFAP4, and TRAPPC9) and one microRNA (MIR125B1) in the validation dataset. The four-CpG biomarker model revealed optimal performance when compared with the other biomarkers ($P < 0.05$). The AUCs of all biomarkers have been shown in Figure 5. The results demonstrated that the four-CpG biomarker model played an important role in prognosis evaluation and presented better specificity and sensitivity in predicting the OS of patients with CC.

4. Discussion

The advantage of AI is that it can deal with complex and data-rich problems stably and flexibly, which makes its application in medical bioinformatics take a qualitative leap [20, 21]. In recent years, AI has shown higher accuracy in clinical prediction modeling of tumor genomics and is expected to become a promising tool for tumor diagnosis and prognosis evaluation [22, 23]. Numerous research results have suggested that using artificial intelligence algorithm to construct multigene models based on RNA or protein levels has the indispensable clinical value of earlier diagnosis and prognosis estimation in cervical cancer. An independent prognostic model composed by SPP1, EFNA1, MMP1, ITM2A, and DSG2 has shown high efficiency in dis-

tinguishing survival outcomes for cervical cancer patients [24]. And a combination of the four miRNAs (miR-502, miR-145, miR-142, and miR-33b) based on the public database was identified as an independent prognostic signature of cervical cancer [25]. However, due to the natural defects of tumors, such as the highly dynamic nature of RNA and proteins and the physicochemical fragility of biological specimens, problems such as unstable access to information and poor practicability sometimes occur. In this regard, DNA methylation information is relatively stable, especially the tumor-associated DNA methylation patterns are relatively conserved and DNA preservation is more stable than RNA and protein. And changes in DNA methylation of specific genes accumulate with disease progression, and detection of specific gene DNA methylation levels is helpful to determine disease progression [26, 27]. Therefore, it has great clinical significance to understand the status of genome-wide methylation profile characteristics of cervical cancer and construct an accurate and stable biomarker model for prognosis.

We analyzed genome-wide DNA methylation profile data of CC in TCGA database by using ChAMP analysis and obtained a set of DMCs. Interestingly, we found DMCs in gene body regions account for the highest proportion, which is related to regulatory mechanisms such as abnormal DNA methylation of gene bodies involved in regulating transcriptional activity and thereby increasing gene expression in tumors. The idea that gene body DNA methylation correlates with gene expression is widely accepted. Hypermethylation of DNA methylation sites cg13600622 and cg14204784 in gene body regions can be used as biomarkers to predict adverse outcomes in laryngeal squamous cell carcinoma [28]. In cervical cancer, abnormal DNA methylation in a large proportion of gene body regions also indicates that it also plays an important regulatory role in the development of CC. In addition, CpG islands, as classical methylation study regions, are expected to occupy an important proportion of DMCs in CC, because their abnormal hypermethylation is closely related to the inactivation of related genes. Furthermore, IGRs have been found to account for a large proportion of cervical cancer DMCs, and this phenomenon has also been observed in colorectal laterally spreading tumors (LSTs) [29]. Therefore, the change of IGRs methylation level plays an important role in the occurrence and development of tumors, and the specific mechanism of action remains to be further explored.

Aberrant DNA methylation of a single gene can be used as a biomarker for CC prognosis. Compared with single gene
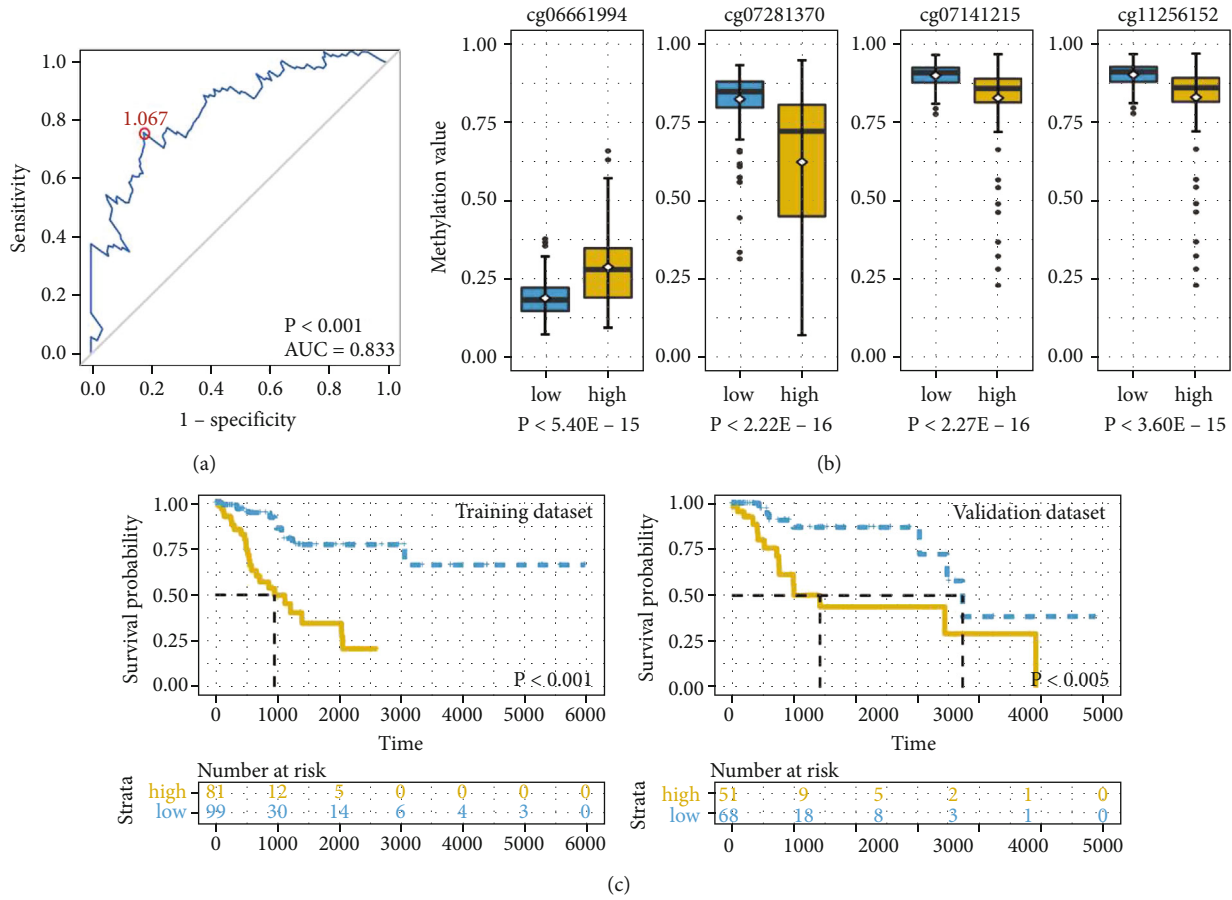
FIGURE 3: Evaluation of the predictive performance of the four-CpG biomarker model. (a) Time-dependent ROC analysis of the predictive accuracy on the OS of patients in the training set. The AUC reaches 0.833 ($P < 0.001$), and the optimization risk score cut-off threshold value is 1.067. (b) The β value of each methylation site in high- and low-risk groups. (c) The Kaplan–Meier curve of the OS between the high- and low-risk groups in the training ($n = 180$) and validation datasets ($n = 119$).

DNA methylation as a predictor, the combined use of DNA methylation sites can obtain higher sensitivity and specificity. In this study, based on DMC analysis, four CpG sites were identified by Cox regression analysis, which were significantly associated with OS in CC patients. Calculating the patient's risk score based on this set of biomarkers can help predict the probability of patient survival time. In survival analysis, the four-CpG biomarker model proved to be reliable independent prognostic factors for CC and superior to other molecular markers, such as RASSF2, LRIG1, and VIM methylation status. Independent group analysis indicated that this model has an independent predictive value for FIGO stage and age and is more suitable for predicting survival time in patients with SCC and G2/G3. Meanwhile, the four-CpG biomarker model showed superior reliability and stability in assessing patient survival outcomes compared with their corresponding gene expression, suggesting that it can be used as an independent risk factor predictor for CC. Another major advantage of epigenetic features is their biological significance. MIR125B1 is a member of the miR-125 family, and hypermethylation of promoter region leads to gene expression downregulation and increases the risk of lymph node metastasis and poor prognosis in breast cancer [30]. And it also showed decreased expression in CC, which is closely related to prognosis [31]. Therefore, cg07281370 acts as a methylation site of the promoter region of MIR125B1, altering methylation level may affect gene expression and might be used as a potential prognostic marker during CC development. Although the mechanism of C20orf195 in cervical cancer is still unclear, the samples were pooled and ranked according to the expression data of C20orf195, which were divided into high expression group ($n = 73$) and low expression group ($n = 73$). Kaplan–Meier survival analysis showed that cervical cancer patients with high expression had significantly worse disease-free survival than those with low expression ($P = 0.026$), and their abnormal hypomethylation status of the promoter region may have a potential role in promoting gene expression. In addition, TFAP4 and TRAPPC9, which correspond to CpG sites in the gene region, are associated with epithelial-mesenchymal transition [32, 33]. TFAP4, highly expressed in non-small-cell lung cancer, is closely related to pathological score, recurrence, and metastasis [34]. While it is worth noting that TRAPPC9 presents low expression in CC, this may not be directly related to the hypermethylated cg11256152 locus. However, hypermethylated loci are the
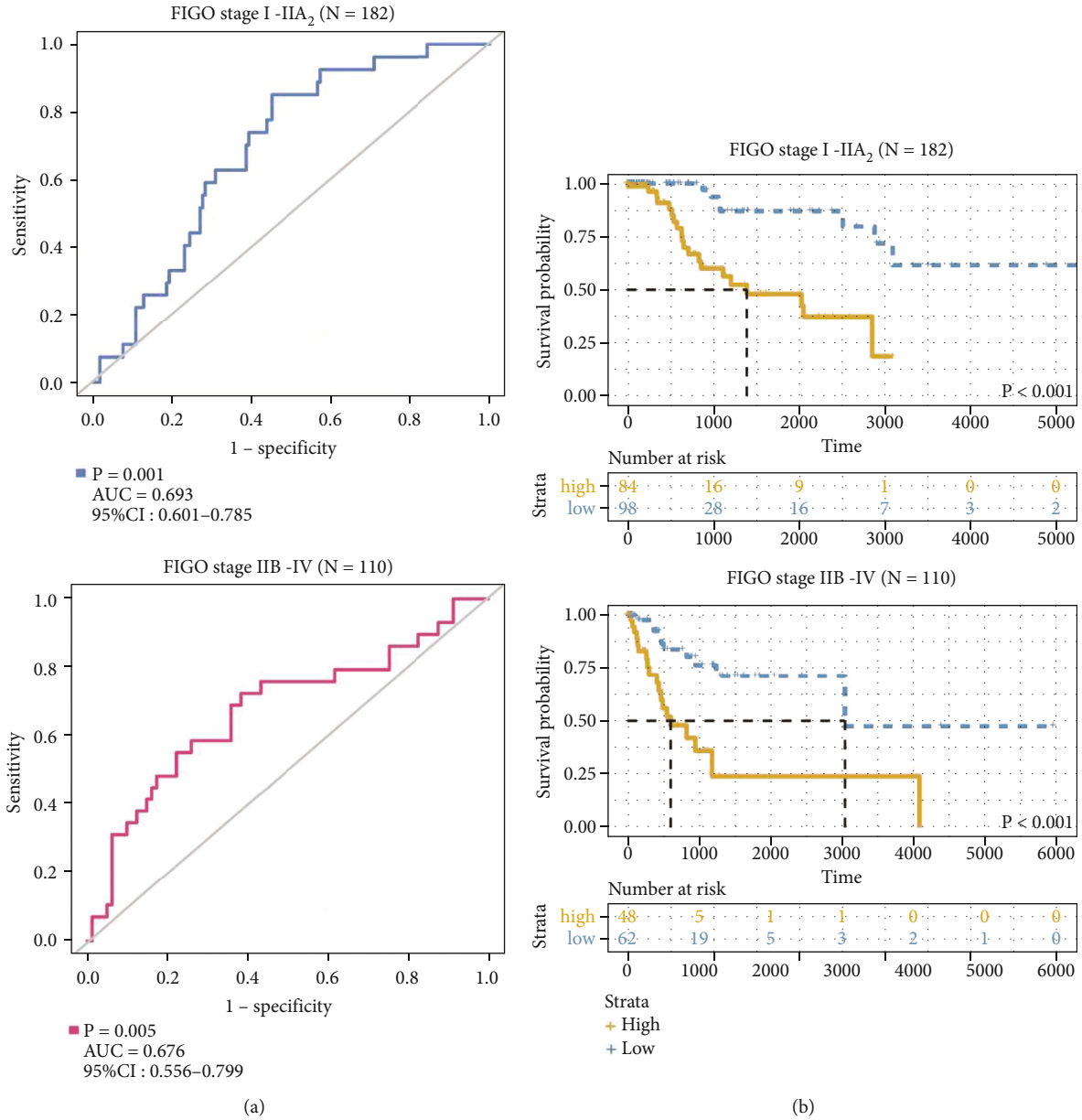
(a)   (b)

FIGURE 4: Kaplan–Meier and ROC analyses of patients in different stage sets. Two sets: FIGO stages I–IIA2 and IIB–IV. (a) ROC curves and AUC statistics of the four-CpG biomarker model were used to evaluate the predictive efficiency in different stage sets. (b) Kaplan–Meier curve presented the survival differences between the high- and low-risk groups.

TABLE 4: Different stratification analyzed on the Kaplan–Meier and ROC.

Stratified factor	Group	Cases N (%)	Kaplan–Meier P value	AUC (95% CI)
Age at initial diagnosis	<50	177 (59.20%)	1.38e-05	0.717 (0.622-0.81)
	≥50	122 (40.80%)	2.37e-04	0.67 (0.555-0.785)
FIGO stage	I-IIA2	182 (60.88%)	1.89e-05	0.693 (0.601-0.785)
	IIB-IV	110 (36.79%)	1.93e-04	0.676 (0.556-0.799)
Histological type	SCC	247 (82.61%)	1.76e-08	0.697 (0.615-0.779)
	Adenocarcinoma	46 (15.38%)	3.68e-03	0.674 (0.417-0.878)
Histologic grade	G2	131 (43.81%)	2.88e-04	0.627 (0.515-0.738)
	G3	120 (40.13%)	13.33e-05	0.762 (0.636-0.888)

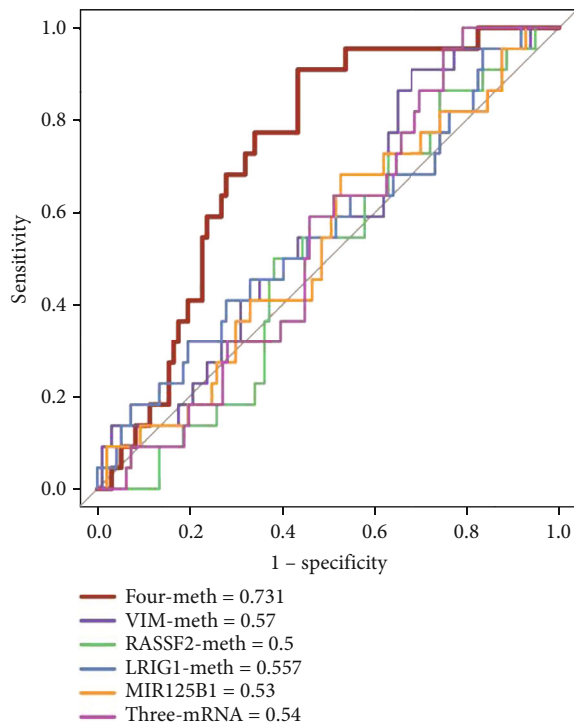95% CI: 95% confidence interval; SCC: squamous cell carcinoma.

Figure 5: The ROC curves compare the sensitivity and specificity of the four-CpG biomarker model with other known biomarkers in predicting the outcome of patients in the validation dataset.

best factors for evaluating prognosis. This is related to the recent proposal that tumor-specific methylation can lead to high variability in expression by disrupting heterochromatin resulting in uncontrolled epigenetic and transcriptional regulation [35].

Nevertheless, there were several limitations to this study. Due to incomplete clinical data, we were unable to carry out comprehensive stratified analysis for some prognostic factors, such as tumor size, lymph node metastasis, and HPV. In addition, the results of the study were generated from a single database; further validation using other CC datasets is necessary.

5. Conclusions

This work identified 34,389 DMCs and their associated distribution features by genome-wide DNA methylation difference analysis of CC using AI algorithm and developed a novel prognostic biomarker model. Comprehensive survival analysis indicated that the risk score of the four-CpG biomarker model could be used as an independent factor for the prognosis of patients with CC. The predictive value was also confirmed with respect to different clinicopathological characteristics, including FIGO stage, age, histological type, and histological grade. Furthermore, the four-CpG biomarker model outperforms other known prognostic signatures and can be more useful in predicting the OS of patients. These results indicate the prediction model could play an important role as a prognostic marker for forecasting

the survival time of patients with CC. In the future, the functional mechanism and potential carcinogenic association of these four CpG sites in CC require further elucidation.

Data Availability

The dataset used in this study is from publicly available datasets. This data can be found here: http://cancergenome.nih.gov and http://firebrowse.org/.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

References

[1] Q. F. Hassan, *Internet of Things A to Z: Technologies and Applications, [M.S. thesis]*, Wiley-IEEE Press, New York, 2018.

[2] C. Lin, G. Han, X. Qi, M. Guizani, and L. Shu, "A distributed mobile fog computing scheme for mobile delay-sensitive applications in SDN-enabled vehicular networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 5, pp. 5481–5493, 2020.

[3] L. The, "Artificial intelligence in health care: within touching distance," *Lancet*, vol. 390, no. 10114, p. 2739, 2017.

[4] J. Andreu-Perez, C. C. Poon, R. D. Merrifield, S. T. Wong, and G. Z. Yang, "Big data for health," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 4, pp. 1193–1208, 2015.

[5] J. Yu, Y. Hu, Y. Xu et al., "LUADpp: an effective prediction model on prognosis of lung adenocarcinomas based on somatic mutational features," *BMC Cancer*, vol. 19, no. 1, p. 263, 2019.

[6] B. H. Kann, A. Hosny, and H. Aerts, "Artificial intelligence for clinical oncology," *Cancer Cell*, vol. 39, no. 7, pp. 916–927, 2021.

[7] C. J. Lynch and C. Liston, "New machine-learning technologies for computer-aided diagnosis," *Nature Medicine*, vol. 24, no. 9, pp. 1304-1305, 2018.

[8] C. Lin, G. Han, X. Qi, J. Du, T. Xu, and M. Martinez-Garcia, "Energy-optimal data collection for unmanned aerial vehicle-aided industrial wireless sensor network-based agricultural monitoring system: a clustering compressed sampling approach," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 6, pp. 4411–4420, 2020.

[9] R. O. Alabi, A. Almangush, M. Elmusrati, and A. A. Makitie, "Deep machine learning for oral cancer: from precise diagnosis to precision medicine," *Frontiers in Oral Health*, vol. 2, article 794248, 2022.

[10] M. Gupta and B. Gupta, "A novel gene expression test method of minimizing breast cancer risk in reduced cost and time by improving SVM-RFE gene selection method combined with LASSO," *Journal of Integrative Bioinformatics*, vol. 18, no. 2, pp. 139–153, 2020.

[11] C. Zhang, Y. Zhao, X. Xu et al., "Cancer diagnosis with DNA molecular computation," *Nature Nanotechnology*, vol. 15, no. 8, pp. 709–715, 2020.

[12] E. Bozgeyik, N. B. Tepe, and Z. Bozdag, "Identification of microRNA expression signature for the diagnosis and prognosis of cervical squamous cell carcinoma," *Pathology, Research and Practice*, vol. 216, no. 11, article 153159, 2020.

[13] H. Xue, Z. Sun, W. Wu, D. Du, and S. Liao, "Identification of hub genes as potential prognostic biomarkers in cervical cancer using comprehensive bioinformatics analysis and validation studies," *Cancer Management and Research*, vol. Volume 13, pp. 117–131, 2021.

[14] A. Abudukadeer, R. Bakry, G. Goebel et al., "Clinical relevance of CDH1 and CDH13 DNA-methylation in serum of cervical cancer patients," *International Journal of Molecular Sciences*, vol. 13, no. 7, pp. 8353–8363, 2012.

[15] B. Phipson, S. Lee, I. J. Majewski, W. S. Alexander, and G. K. Smyth, "Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression," *The Annals of Applied Statistics*, vol. 10, no. 2, pp. 946–963, 2016.

[16] H. Cho, A. Yu, S. Kim, J. Kang, and S.-M. Hong, "Robust likelihood-based survival modeling with microarray data," *Journal of Statistical Software*, vol. 29, no. 1, pp. 1–16, 2009.

[17] M. K. Lee, E. M. Jeong, J. H. Kim, S. B. Rho, and E. J. Lee, "Aberrant methylation of the VIM promoter in uterine cervical squamous cell carcinoma," *Oncology*, vol. 86, no. 5-6, pp. 359–368, 2014.

[18] D. Guerrero-Setas, N. Perez-Janices, L. Blanco-Fernandez et al., "RASSF2 hypermethylation is present and related to shorter survival in squamous cervical cancer," *Modern Pathology*, vol. 26, no. 8, pp. 1111–1122, 2013.

[19] M. Lando, C. S. Fjeldbo, S. M. Wilting et al., "Interplay between promoter methylation and chromosomal loss in gene silencing at 3p11-p14 in cervical cancer," *Epigenetics*, vol. 10, no. 10, pp. 970–980, 2015.

[20] S. Reddy, J. Fox, and M. P. Purohit, "Artificial intelligence-enabled healthcare delivery," *Journal of the Royal Society of Medicine*, vol. 112, no. 1, pp. 22–28, 2019.

[21] P. Wang, R. Yu, N. Gao, C. Lin, and Y. Liu, "Task-driven data offloading for fog-enabled urban IoT services," *IEEE Internet of Things Journal*, vol. 8, no. 9, pp. 7562–7574, 2021.

[22] J. Spicer and A. N. Sanborn, "What does the mind learn? A comparison of human and machine learning representations," *Current Opinion in Neurobiology*, vol. 55, pp. 97–102, 2019.

[23] P. Wang, C. Lin, M. S. Obaidat, Z. Yu, and Q. Zhang, "Contact tracing incentive for COVID-19 and other pandemic diseases from a crowdsourcing perspective," *IEEE Internet of Things Journal*, vol. 8, no. 21, pp. 15863–15874, 2020.

[24] M. Zhao, W. Huang, S. Zou, Q. Shen, and X. Zhu, "A five-genes-based prognostic signature for cervical cancer overall survival prediction," *International Journal of Genomics*, vol. 2020, Article ID 8347639, 13 pages, 2020.

[25] S. Chen, C. Gao, Y. Wu, and Z. Huang, "Identification of prognostic miRNA signature and lymph node metastasis-related key genes in cervical cancer," *Frontiers in Pharmacology*, vol. 11, p. 544, 2020.

[26] X. Li, W. Zhang, J. Song, X. Zhang, L. Ran, and Y. He, "SLCO4C1 promoter methylation is a potential biomarker for prognosis associated with biochemical recurrence-free survival after radical prostatectomy," *Clinical Epigenetics*, vol. 11, no. 1, p. 99, 2019.

[27] M. Attaleb, M. Khyatti, L. Benbacer et al., "Status of p16(INK4a) and E-cadherin gene promoter methylation in Moroccan patients with cervical carcinoma," *Oncology Research*, vol. 18, no. 4, pp. 185–192, 2009.

[28] L. Fan, A. Zhang, and P. Deng, "LMX1B mRNA expression and its gene body CpG methylation are valuable prognostic biomarkers for laryngeal squamous cell carcinoma," *Biomedicine & Pharmacotherapy*, vol. 117, article 109174, 2019.

[29] L. Zhu, F. Yan, Z. Wang et al., "Genome-wide DNA methylation profiling of primary colorectal laterally spreading tumors identifies disease-specific epimutations on common pathways," *International Journal of Cancer*, vol. 143, no. 10, pp. 2488–2498, 2018.

[30] Y. Zhang, L. X. Yan, Q. N. Wu et al., "miR-125b is methylated and functions as a tumor suppressor by regulating the ETS1 proto-oncogene in human invasive breast cancer," *Cancer Research*, vol. 71, no. 10, pp. 3552–3562, 2011.

[31] C. Gao, C. Zhou, J. Zhuang et al., "MicroRNA expression in cervical cancer: novel diagnostic and prognostic biomarkers," *Journal of Cellular Biochemistry*, vol. 119, no. 8, pp. 7080–7090, 2018.

[32] R. Jackstadt, S. Roh, J. Neumann et al., "AP4 is a mediator of epithelial-mesenchymal transition and metastasis in colorectal cancer," *The Journal of Experimental Medicine*, vol. 210, no. 7, pp. 1331–1350, 2013.

[33] Z. H. Fu, S. Q. Liu, M. B. Qin et al., "NIK- and IKKβ-binding protein contributes to gastric cancer chemoresistance by promoting epithelial-mesenchymal transition through the NF-κB signaling pathway," *Oncology Reports*, vol. 39, no. 6, pp. 2721–2730, 2018.

[34] H. Gong, S. Han, H. Yao, H. Zhao, and Y. Wang, "AP4 predicts poor prognosis in nonsmall cell lung cancer," *Molecular Medicine Reports*, vol. 10, no. 1, pp. 336–340, 2014.

[35] W. Timp, H. C. Bravo, O. G. McDonald et al., "Large hypomethylated blocks as a universal defining epigenetic alteration in human solid tumors," *Genome Medicine*, vol. 6, no. 8, p. 61, 2014.