ISSN 2616-5805 Vol. 6, Issue 2: 69-77, DOI: 10.25236/AJMS.2025.060210

Prediction of Early Gastric Cancer Based on miRNAs Using Penalized Logistic Regression

Huiling Yang[1,a,*]

[1]*Chongqing Technology and Business University, Chongqing, China, 400067*
[a]*1964549331@qq.com*
[*]*Corresponding author*

Abstract: *Early diagnosis of gastric cancer is critical for improved patient prognosis. MicroRNAs (miRNAs), a class of non-coding small molecules (20-25 nucleotides) that regulate gene expression by binding to target RNAs, represent promising disease biomarkers due to their inherent stability in bodily fluids. In this study, based on a dataset of 2,834 serum samples sourced from the Gene Expression Omnibus (GEO) database (1,417 early gastric cancer patients and 1,417 healthy controls), four penalized logistic regression models—LASSO, Elastic Net (ENet), Smoothly Clipped Absolute Deviation (SCAD), and Minimax Concave Penalty (MCP)—are employed for feature selection. These models are subsequently integrated with a coordinate descent algorithm to develop a diagnostic model. The results demonstrated that the MCP model achieved a prediction accuracy of 98.59% using only three miRNAs (hsa-miR-1343-3p, hsa-miR-5100, and hsa-miR-6765-5p). Consequently, model complexity was substantially reduced, and the model's generalization capability was improved. Biological validation revealed that these miRNAs were consistently selected across multiple models, furthermore, they are directly implicated in key pathways of gastric carcinogenesis, including the regulation of cell proliferation and apoptosis. This study provides a high-accuracy, cost-effective diagnostic strategy for early gastric cancer detection and identifies potential therapeutic targets.*

Keywords: *Gastric Cancer, Penalized Logistic Regression, SCAD, MCP, Coordinate Descent Algorithm*

1. Introduction

Gastric cancer (GC) remains a leading cause of cancer-related mortality worldwide. Early diagnosis is of paramount importance for significantly improving the five-year survival rate of GC patients, as early-stage disease exhibits a high curative potential. However, the clinical manifestations of early gastric cancer are often subtle and can be easily overlooked, including mild dyspepsia, bloating, or abdominal discomfort. Delays in seeking medical consultation are frequently caused by the non-specific nature of these symptoms. Furthermore, while traditional diagnostic methods such as endoscopy offer high accuracy, their applicability in widespread population screening is limited by their invasiveness, associated costs, and the stringent technical requirements imposed on operators. Therefore, the development of a simple, non-invasive method for early gastric cancer prediction represents an urgent clinical need.

In recent years, microRNAs (miRNAs), a critical class of non-coding RNA molecules, have been positioned as a focal point in early cancer diagnosis research by advances in molecular biology and high-throughput technologies. This positioning is due to their significant roles in gene expression regulation, cell proliferation, apoptosis, and tumorigenesis[1-3]. Serum miRNAs have been extensively investigated as potential biomarkers, owing to their high biological stability in bodily fluids, ease of accessibility, and straightforward detection. For instance, Xi et al. (2012) identified a panel of 10 serum miRNAs for the diagnosis of early-stage non-small cell lung cancer[4]. Similarly, Roberg et al. (2017) demonstrated that circulating miRNA expression profiles could differentiate prostate cancer aggressiveness and predict prognosis, combinations of these identified serum miRNA markers are anticipated to aid in risk stratification and clinical decision-making for prostate cancer[5]. More recently, Chen et al. (2024) identified three serum miRNAs (miR-106b-5p, miR-129-1-3p, and miR-381-3p) as potential diagnostic biomarkers for prostate cancer[6].

Previous studies have demonstrated the potential of miRNAs in the diagnosis of different cancers, but how to systematically screen and validate combinations of miRNA markers with high specificity and sensitivity for disease prediction is still a key issue that needs to be addressed. Machine learning

ISSN 2616-5805 Vol. 6, Issue 2: 69-77, DOI: 10.25236/AJMS.2025.060210

methods can help construct accurate diagnostic models for early cancer detection based on miRNAs. For example, Wang et al. (2011) combined a genetic algorithm and support vector machine (GA-SVM) methods to screen 13 key features from 185 features, the constructed miR-SF classifier achieved 93.97%prediction accuracy for human miRNA precursors, providing a new approach for the early diagnosis of breast cancer[7]. Furthermore,Bo et al. (2021) developed m6A-miRNAs markers for cancer detection using support vector machine algorithms, with an Area Under the Curve (AUC) of 0.936 being achieved in an external validation cohort[8]. However, traditional machine learning methods, when applied to the analysis of high-dimensional genetic data, often suffer from multicollinearity and overfitting, and their ability to discern variable significance is thereby limited. Therefore, penalized logistic regression has been invoked to address these problems. Shen and Tan (2005) proposed a penalized logistic regression method combining two dimensionality reduction methods (singular value decomposition and partial least squares) to solve the problem of cancer classification in microarray data[9]. Similarly, Liang et al. (2013) developed a sparse logistic regression model based on $L_{1/2}$ penalization and used a coordinate descent algorithm for the selection and classification of cancer-related genes in microarray data[10]. In addition, Algamal and Lee (2015) proposed an improved adaptive elastic net regularized logistic regression method (AAEelastic) to improve the stability of gene selection and classification accuracy in high-dimensional cancer classification by adjusting the initial weights[11]. More recently,Lavanya et al. (2023) developed a new fusion logistic regression using weighted L_1 and L_2 penalties (FLR) to achieve sparsity and oracle properties in gene selection for microarray data to improve performance[12].

The aim of this study is to analyze the differences in serum miRNA expression profiles between early gastric cancer patients and healthy controls using high-throughput miRNA microarray technology. Based on this objective, four penalized logistic regression methods—LASSO, ENet, SCAD, and MCP—are applied to screen miRNAs and predict early gastric cancer through penalized logistic regression modeling. These applications are intended to establish an efficient and reliable miRNA prediction model and to provide a novel, non-invasive screening tool for early gastric cancer in the clinic.

This paper is structured as follows: Section 2 introduces the data sources and variable descriptions, Section 3 describes the models and algorithms, Section 4 presents a comparison of the feature screening ability and prediction accuracy of the four models, Section 5 provides a brief analysis of the screened biomarkers, and Section 6 contains the conclusions and outlook.

2. Data Sources and Variable Descriptions

The gene expression data utilized in this paper originate from the Gene Expression Omnibus (GEO) database, accession number GSE164174. This dataset contains serum microRNA expression profiles of 2940 samples, comprising 1423 patients with early gastric cancer, 1417 non-cancer controls, 50 patients with esophageal cancer, and 50 patients with colorectal cancer. The expression levels of 2565 miRNAs are obtained via high-throughput miRNA microarray technology, using the Toray Industries GPL21263 (3D-Gene Human miRNA V21_1.0.0) platform. In this study, 1417 early gastric cancer (EGC) patients and 1417 non-cancer control individuals are screened from the original dataset to serve as samples, and the expression levels of 2565 miRNAs serve as the variables. Samples from EGC patients are labeled as 1, while non-cancer controls are labeled as 0. Subsequently, the dataset is randomly divided into a training set (n=1984) and a test set (n=850) in a 70% to 30% ratio, these sets are respectively used for constructing the classification model and evaluating its prediction performance. Subsequently, we utilize the following four methods—LASSO penalized logistic regression, ENet penalized logistic regression, SCAD penalized logistic regression, and MCP penalized logistic regression—to conduct variable screening and classification prediction.

3. Models and Algorithm

3.1 Four Penalized Models

A logistic regression model is a generalized regression model commonly used in classification problems. First, we introduce two types of logistic regression

ISSN 2616-5805 Vol. 6, Issue 2: 69-77, DOI: 10.25236/AJMS.2025.060210

$$\begin{aligned} P &:= P(Y=1 \mid X;\beta) = \frac{e^{X^{\top}\beta}}{1+e^{X^{\top}\beta}} = h_{\beta}(X) \\ 1-P &= P(Y=0 \mid X;\beta) = \frac{1}{1+e^{X^{\top}\beta}} = 1-h_{\beta}(X), \end{aligned} \tag{1}$$

Where the response variable Y = 1 denotes the early gastric cancer sample, Y = 0 denotes the normal sample, the predictor variable $X=(1,X_1,X_2,...,X_p)^{\top}$ is a real-valued random variable, P(Y|X) denotes the conditional probability, and $\beta=(\beta_0,\beta_1,\beta_2,...,\beta_p)^{\top}$ denotes the effect of the predictor variable on the classification of the cancer, and its log-likelihood function is

$$\ell(\beta) = \sum_{i=1}^{n}\left(Y^{(i)}\ln\left(h_{\beta}(X^{(i)})\right)+(1-Y^{(i)})\ln\left(1-h_{\beta}(X^{(i)})\right)\right) \tag{2}$$

Penalized negative log-likelihood function $-\ell(\beta)$ Introducing penalization function to get penalized negative log-likelihood function

$$S_{\lambda,\gamma}(\boldsymbol{\beta}) = -l(\boldsymbol{\beta}) + \sum_{j=1}^{p} P(\beta_j;\lambda,\gamma), \tag{3}$$

Where $\sum_{j=1}^{p} P(\beta_j;\lambda,\alpha)$ is the four penalty functions listed in Table 1, λ is the tuning parameter, and α is the regularization parameter.

Table 1: Four kinds of penalty functions

Marking	penalty functions
LASSO	$P_{LASSO}(\beta;\lambda) = \lambda\sum_{j=1}^{p}\lvert\beta_j\rvert, \quad \lambda>0\cdot$
ENet	$P_{ENet}(\beta;\lambda) = \lambda_1\sum_{j=1}^{p}\lvert\beta_j\rvert+\lambda_2\sum_{j=1}^{p}\beta_j^2, \quad \lambda_{1,2}>0$
SCAD	$P_{SCAD}(\beta;\lambda,\alpha) = \begin{cases} \lambda\beta, & \beta\le\lambda,\lambda\ge 0, \\ \lambda\alpha\beta-0.5\frac{(\beta^2+\lambda^2)}{\alpha-1}, & \lambda<\beta\le\alpha,\alpha>2, \\ \frac{\lambda^2(\alpha+1)}{2}, & \beta>\lambda. \end{cases}$
MCP	$P_{MCP}(\beta;\lambda,\alpha) = \begin{cases} \lambda\beta-\frac{\beta^2}{2\alpha}, & \beta\le\alpha\lambda,\lambda\ge 0, \\ \frac{1}{2}\alpha\lambda^2, & \beta>\alpha\lambda,\alpha>1. \end{cases}$

3.2 Coordinate Descent Algorithm

In this paper, a coordinate descent algorithm is employed to estimate parameters in penalized logistic regression. This algorithm iteratively optimizes one variable at a time while holding others fixed, proceeding sequentially until convergence. As demonstrated by Patrick and Jian (2011), coordinate descent is effective for LASSO, SCAD, and MCP-penalized logistic regression, providing iterative estimates for these models[13].

$$\hat{\beta}_j^{LASSO}(Z_j,\lambda) = \frac{S(Z_j,\lambda)}{\nu_j}, \tag{4}$$

$$\hat{\beta}_j^{MCP}(Z_j,\lambda,\alpha) = \begin{cases} \dfrac{S(Z_j,\lambda)}{\nu_j-1/\alpha}, & \text{if } \lvert Z_j\rvert\le\nu_j\lambda\alpha, \\ \dfrac{Z_j}{\nu_j}, & \text{if } \lvert Z_j\rvert>\nu_j\lambda\alpha,\alpha>1/\nu_j; \end{cases} \tag{5}$$

ISSN 2616-5805 Vol. 6, Issue 2: 69-77, DOI: 10.25236/AJMS.2025.060210

$$\hat{\beta}_j^{SCAD}(Z_j,\lambda,\alpha)=\begin{cases}\dfrac{S(Z_j,\lambda)}{\nu_j}, & \text{if } |Z_j|\le\lambda(\nu_j+1),\\ \dfrac{S(Z_j,\alpha\,\text{lambda}/(\alpha-1))}{\nu_j-1/(\alpha-1)}, & \text{if } \lambda(\nu_j+1)<|Z_j|\le\nu_j\lambda\alpha,\\ \dfrac{Z_j}{\nu_j}, & \text{if } |Z_j|>\nu_j\lambda\alpha,\alpha>1+1/\nu_j;\end{cases} \tag{6}$$

Where $S(Z_j,\lambda)$ represents the soft-thresholding operator, defined as $S(Z_j,\lambda)=\text{sign}(Z_j)(|Z_j|-\lambda)$, and $\nu_j=n^{-1}X_j^TWX_j$,where $\hat{P}_t=\dfrac{\exp(x_t^T\hat{\beta}^{\lambda,\alpha}(m))}{[1+\exp(x_t^T\hat{\beta}^{\lambda,\alpha}(m))]}$,

$W_t=\hat{P}_t(1-\hat{P}_t),t=1,...,n,W=\text{diag}\{W_1,W_2,...,W_n\}$,

$\tilde{Y}=x^T\hat{\beta}^{\lambda,\alpha}(m)+W^{-1}(Y-\hat{P}),\hat{P}=(\hat{P}_1,...,\hat{P}_n)$,

$x_{\cdot j}=(x_{1j},...,x_{nj})^T,\nu_j=n^{-1}x_{\cdot j}^TWx_{\cdot j},j=1,...,p$,

$Z_j=n^{-1}x_{\cdot j}^TW(\tilde{Y}-x_{-j}\beta_{-j})=n^{-1}x_{\cdot j}^TWr+\nu_j\hat{\beta}_j^{\lambda,\alpha}(m)$,

$x_{-j}=(x_1,...,x_{(j-1)},0,x_{(j+1)},...,x_p)^T,\beta_{-j}=(\beta_1,...,\beta_{j-1},0,\beta_{j+1},...,\beta_p)^T$.

The coordinate descent algorithm is then applied to the four classes of penalized logistic regression to obtain the final parameter estimates, $\hat{\beta}_0^{\lambda,\alpha}$ and $\hat{\beta}^{\lambda,\alpha}$, and finally calculate the probability estimates:

$$\hat{P}(Y_t=1|x_t;\hat{\beta}_0^{\lambda,\alpha},\hat{\beta}^{\lambda,\alpha})=\frac{\exp(\hat{\beta}_0^{\lambda,\alpha}+x_t^T\hat{\beta}^{\lambda,\alpha})}{1+\exp(\hat{\beta}_0^{\lambda,\alpha}+x_t^T\hat{\beta}^{\lambda,\alpha})},$$

$$\hat{P}(Y_t=0|x_t;\hat{\beta}_0^{\lambda,\alpha},\hat{\beta}^{\lambda,\alpha})=\frac{1}{1+\exp(\hat{\beta}_0^{\lambda,\alpha}+x_t^T\hat{\beta}^{\lambda,\alpha})}. \tag{7}$$

Based on the above, the coordinate descent algorithm for MCP logistic regression is as follows:

Algorithmic 1: Coordinate descent for MCP logistic regression

Require: The training set $\{x_t=(x_{t,1},x_{t,2},...,x_{t,p}),y_t\}_{t=1}^n$, a grid of increasing λ values $\Lambda=\{\lambda_1,...,\lambda_L\}$, $\alpha=15$, a given tolerance limit ε and a maximum iteration number M

1: Initialization $\beta(0)=\hat{\beta}(\lambda_{\max}=\lambda_L,\alpha=15)$

2: for m = 0, 1, ..., each $l\in\{L,L-1,...,1\}$ do

3: repeat

4: $\hat{\eta}_t\leftarrow\beta_0+x_t^T\hat{\beta}^{\lambda,\gamma}(m)$

5: $\hat{P}_t\leftarrow\{\dfrac{e^{\eta_t}}{1+e^{\eta_t}}\}_{t=1}^n$

6: $W\leftarrow\text{diag}\{\hat{P}_1(1-\hat{P}_1),...,\hat{P}_n(1-\hat{P}_n)\}$

7: $r\leftarrow W^{-1}(Y-\hat{P})$

8: $\tilde{Y}\leftarrow\eta+r$

9: while not convergent do

10: for each $j\in\{1,2,...,p\}$ do

11: $\nu_j\leftarrow n^{-1}x_{\cdot j}^TWx_{\cdot j}$

12: $Z_j\leftarrow\dfrac{1}{n}x_{\cdot j}^TW(\tilde{Y}-x_{-j}\beta_{-j})=\dfrac{1}{n}x_j^TWr+\nu_j\hat{\beta}_j^{\lambda,\gamma}(m)$

13: where set the λ for the intercept term to 0

14: if $|Z_j|\le\nu_j\gamma\lambda|$ then

15: $\hat{\beta}_j^{\lambda,\gamma}(m+1)\leftarrow\dfrac{S(Z_j,\lambda)}{\nu_j-1/\gamma}$

16: else

ISSN 2616-5805 Vol. 6, Issue 2: 69-77, DOI: 10.25236/AJMS.2025.060210

17: $\hat{\beta}_j^{\lambda,\gamma}(m+1) \leftarrow \frac{Z_j}{\nu_j}$

18: end if

19: $r \leftarrow r - x_{\cdot j}^T(\hat{\beta}_j^{\lambda,\gamma}(m+1) - \hat{\beta}_j^{\lambda,\gamma}(m))$

20: end for

21: end while

22: until $\| \hat{\beta}^{\lambda,\gamma}(m+1) - \hat{\beta}^{\lambda,\gamma}(m) \|_2^2 \leq \varepsilon$ or do a maximum iteration number M

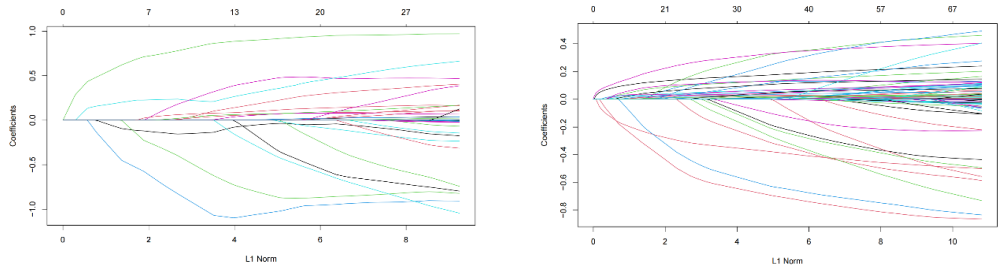
23: end for

Ensure: $\hat{\beta}^{\lambda,\gamma}$

4. Characterization and prediction of performance

A balanced dataset of 2834 samples (1:1 ratio of cancer to normal) with expression data for 2565 miRNAs was used in this study. To mitigate bias and uncertainties introduced by arbitrary threshold optimization, a fixed prediction threshold of 0.5 was chosen. This approach provides a transparent evaluation of the model's classification accuracy on the test set at a pre-defined threshold, enabling objective comparisons with existing literature. The model's classification efficacy is then comprehensively evaluated using accuracy, sensitivity, and specificity metrics at this threshold.

Combining LASSO penalty, ENet penalty and logistic regression, regular paths are obtained by coordinate descent algorithm. The paths are visualized using the R software glmnet package, see Figure 1.



(a) LASSO *(b) ENet(α =0.5)*

Figure 1: Coefficient paths for LASSO and ENet

Figure 1 illustrates the coefficient regularization paths for LASSO and ENet, plotting the coefficient values for each variable against the norm (the sum of the absolute values of all coefficients). This figure demonstrates how the coefficients of different genes evolve during the regularization process, revealing the gene selection behavior of each model. Furthermore, coordinate descent algorithms were also investigated for SCAD/MCP-penalized logistic regression, and the resulting model coefficient paths, generated using the training set and the R package ncvreg, are presented in Figure 2 .

Figure 2 presents the coefficient paths for SCAD with values of 3, 5, 10, and 15, alongside the coefficient paths for MCP with values of 5, 7, 10, and 15. Within the unshaded, locally convex region, the solutions exhibit smoothness and stability. However, in the shaded region on the right, the solutions become discontinuous and noisy. The SCAD path for = 15 in Figure 2 closely resembles and is smoother than the MCP path for = 15, suggesting a more stable model configuration.

ISSN 2616-5805 Vol. 6, Issue 2: 69-77, DOI: 10.25236/AJMS.2025.060210

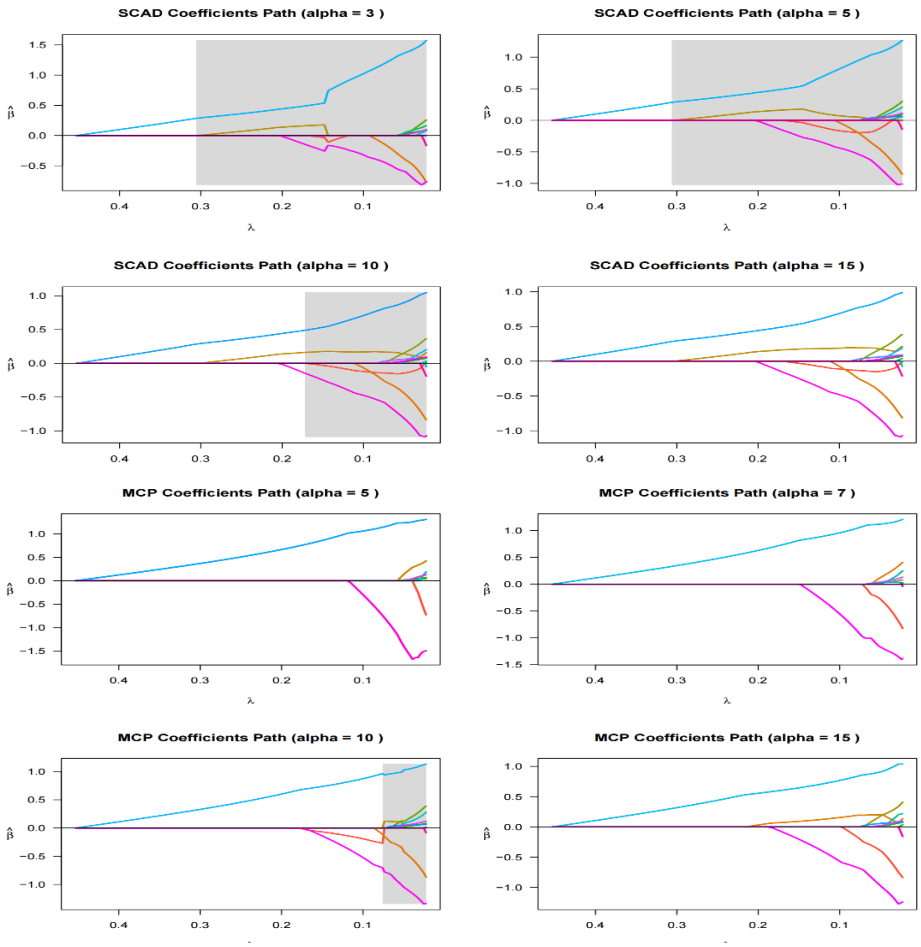
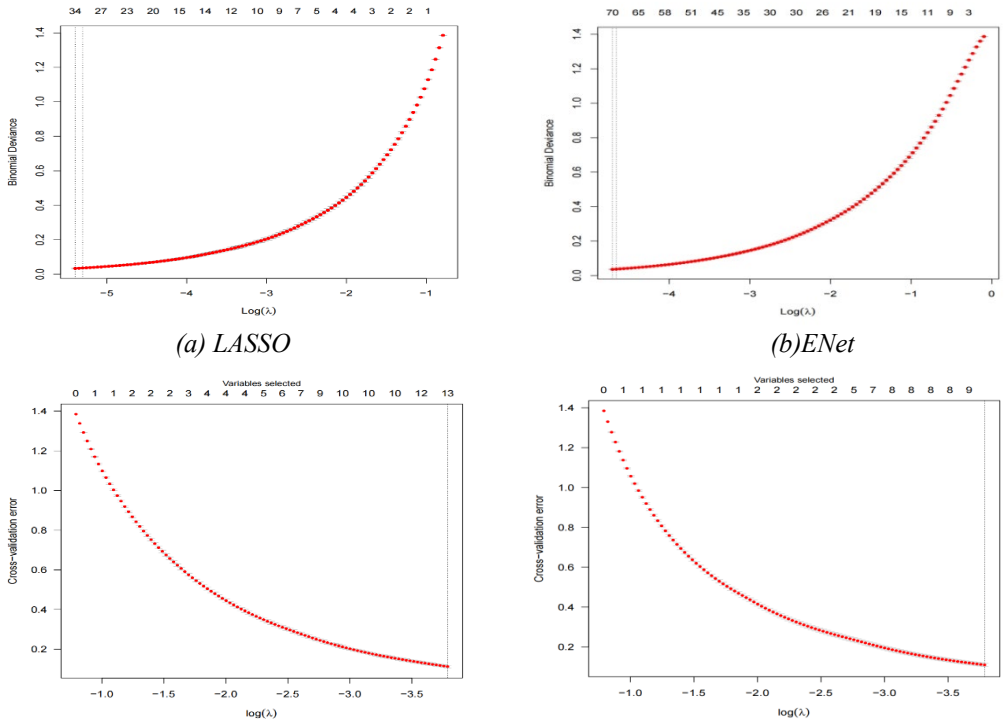


Figure 2: SCAD and MCP coefficient paths for different α



(a) LASSO *(b)ENet*

(c) SCAD *(d) MCP*

Figure 3: Cross-validation error curves for LASSO, ENet, SCAD, and MCP

ISSN 2616-5805 Vol. 6, Issue 2: 69-77, DOI: 10.25236/AJMS.2025.060210

Figure 3(a) and 3(b) present the binomial deviance curves for LASSO and ENet, respectively, generated using the cv function in R. Figures 3(c) and 3(d) display the cross-validation error curves for SCAD and MCP, respectively, plotted using the plot.cv.ncvreg function in R. The number of selected variables is indicated atop each plot. In each curve, the left vertical line denotes the value corresponding to the minimum mean squared error (MSE), while the right vertical line indicates the value within one standard error of the minimum MSE, representing a standard error rule often used for model selection. Given the tendency of cross-validation to select smaller values (leading to over-parameterized models) and considering the characteristics of the data in this study, manual adjustments were applied to the values for all four penalties. Variables were then screened based on performance around the optimal, with the corresponding feature selection capabilities and prediction accuracies summarized in Table 2.

Table 2: Comparison of feature selection ability and prediction accuracy of different models

Model	λ	Number of Features	Sensitivity	Specificity	Accuracy
LASSO(α =1)	0.005	33	0.9976	0.9976	0.9976
	0.01	21	0.9906	0.9976	0.9941
	0.03	12	0.9835	0.9953	0.9894
	0.05	10	0.9765	0.9953	0.9859
ENet(α =0.5)	0.05	35	0.9859	0.9976	0.9918
	0.1	28	0.9812	0.9953	0.9882
	0.3	15	0.9624	0.9953	0.9788
	0.5	9	0.9694	0.9953	0.9824
SCAD(α =15)	0.01	20	0.9859	0.9976	0.9918
	0.02	15	0.9835	0.9953	0.9894
	0.05	10	0.9765	0.9929	0.9847
	0.1	5	0.9765	0.9953	0.9859
MCP(α =15)	0.01	19	0.9859	0.9976	0.9918
	0.02	12	0.9835	0.9953	0.9894
	0.05	8	0.9765	0.9929	0.9847
	0.1	3	0.9765	0.9953	0.9859

Note: $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$, $Sensitivity = \frac{TP}{TP+FN}$, $Specificity = \frac{TN}{TN+FP}$

To optimize the balance between prediction accuracy and feature parsimony, the performance of LASSO, ENet, SCAD, and MCP regularization methods was compared across a range of λ values. With minimal differences in accuracy rates (all exceeding 0.97), the ability to minimize the number of selected variables became the key differentiator for model selection. As shown in Table 2, the LASSO model selected 10 variables at λ = 0.05 (accuracy: 0.9859), ENet selected 9 variables at λ = 0.5 (accuracy: 0.9824), SCAD selected 5 variables at λ = 0.1 (accuracy: 0.9859), and MCP selected only 3 variables at λ = 0.1 (accuracy: 0.9859). Given the importance of cost-effectiveness in screening and diagnostic applications, the MCP (λ = 0.1) model was chosen as the final model due to its minimal feature set and maintained prediction accuracy. This choice significantly reduces model complexity, enhances interpretability and generalization, and mitigates the risk of overfitting. The subsequent analysis focuses on the biological interpretation of the variables selected by the MCP model.

5. Biological analysis of selected traits

A brief analysis of the significant variables screened by each of the four penalized models (LASSO, ENet, SCAD, and MCP) was conducted. Table 3 summarizes the miRNAs identified by each method from the same dataset, with commonly selected miRNAs highlighted in bold. While each model selected a distinct set of miRNAs, all achieved prediction accuracies of 0.98 or higher. Notably, the three miRNAs selected by the MCP method (hsa-miR-1343-3p, hsa-miR-5100, and hsa-miR-6765-5p) were also present in the variable sets identified by the other three methods. This finding underscores the importance and robustness of these three miRNAs as key and reliable biomarkers for early gastric cancer prediction.

ISSN 2616-5805 Vol. 6, Issue 2: 69-77, DOI: 10.25236/AJMS.2025.060210

Table 3: Variable screening for the four penalty models

LASSO	ENet	SCAD	MCP
hsa-miR-1228-5p	hsa-miR-1228-5p	hsa-miR-1228-5p	hsa-miR1343-3p
hsa-miR-1268b	hsa-miR-1290	hsa-miR-1268b	hsa-miR-5100
hsa-miR-1343-3p	hsa-miR-1343-3p	hsa-miR-1343-3p	hsa-miR-6765-5p
hsa-miR-187-5p	hsa-miR-4787-3p	hsa-miR-5100	
hsa-miR-3122	hsa-miR-5100	hsa-miR-6765-5p	
hsa-miR-4787-3p	hsa-miR-6746-5p		
hsa-miR-5100	hsa-miR-6765-5p		
hsa-miR-6511b-5p	hsa-miR-6877-5p		
hsa-miR-668-5p	hsa-miR-8073		
hsa-miR-6765-5p			

Consistent with our findings, Wang et al. (2025) demonstrated the predictive value of hsa-miR-6765-5p in gastric cancer. Furthermore, Yongxin et al[14]. (2024) found that the circular RNA circGLIS3 promotes gastric cancer progression by sponging hsa-miR-1343-3p, suggesting a potential tumor-suppressive role for hsa-miR-1343-3p[15]. Huimin et al.(2022) revealed that hsa-miR-5100 inhibits autophagy in gastric cancer cells by targeting the DEK gene, indicating its potential as a therapeutic target[16]. These independent studies, viewed from different perspectives, provide strong support for the involvement of these three miRNAs in gastric cancer pathogenesis, bolstering the biological plausibility of our study. In conclusion, hsa-miR-1343-3p, hsa-miR-5100, and hsa-miR-6765-5p represent a promising biomarker combination for predictive modeling in early gastric cancer, warranting further investigation of their diagnostic and therapeutic utility in clinical settings.

6. Conclusion and outlook

In this study, a comparative analysis of four regularization methods (LASSO, ENet, SCAD, and MCP) revealed that all achieved high prediction accuracies (≥ 0.97) across various parameter settings. Considering both feature selection efficiency and model complexity, the MCP ($\lambda = 0.1$) model was selected as optimal. This model, utilizing only three variables (hsa-miR-1343-3p, hsa-miR-5100, and hsa-miR-6765-5p), achieved comparable prediction accuracy to other methods while minimizing the risk of overfitting and enhancing model interpretability and generalization. Biological analyses further validated these findings, demonstrating that the selected miRNAs were consistently identified across multiple models and have been implicated in gastric cancer development in previous studies. This suggests their potential as a robust biomarker combination for early gastric cancer prediction, warranting further investigation of their clinical diagnostic and therapeutic value.

References

[1] Shen, J., Liao, J., Guarnera, M. A., Fang, H., Cai, L., Stass, S. A., & Jiang, F.·Analysis of MicroRNAs in Sputum to Improve Computed Tomography for Lung Cancer Diagnosis[J].Journal of Thoracic Oncology, 2014,9(1):33-40.

[2] Zaporozhchenko, I. A., Bryzgunova, O. E., Konoshenko, M. Y., Lekchnov, E. A., Amelina, E. V., Pashkovskaya, O. A., Yarmoschuk, S. V., Zheravin, A. A., Rykova, E. Y., & Laktionov, P. P.·Urine cell-free and extracellular vesicle cargo miRNAs as biomarkers for prostate cancer diagnosis[J]. Annals of Oncology, 2019,30(S5):v22.

[3] Koopaie, M., Manifar, S., Talebi, M. M., Kolahdooz, S., Razavi, A. E., Davoudi, M., & Pourshahidi, S.·Assessment of salivary miRNA, clinical, and demographic characterization in colorectal cancer diagnosis[J]. Translational Oncology, 2024,41:101880.

[4] Xi, C., Zhibin, H., Wenjing, W., Yi, B., Lijia, M., Chunni, Z., Cheng, W., Zhiji, R., Yang, Z., Sijia, W., Rui, Z., Yixin, Z., Heng, H., Chazhen, L., Lin, X., Jun, W., Hongbing, S., Junfeng, Z., Ke, Z., & Chen-Yu, Z.·Identification of ten serum microRNAs from a genome-wide serum microRNA expression profile as novel noninvasive biomarkers for nonsmall cell lung cancer diagnosis[J].·International journal of cancer, 2012, 130(7):1620-1628.

[5] Roberg, S. L. A., Damien, L., Andrea, Z., Mathias, W., Jonathan, W., & Charlotte, B.·A signature of miRNAs in the blood to help prognosticate prostate cancer at the time of diagnosis[J].·Journal of Clinical Oncology, 2017, 35(15):e16558.

[6] Chen, S., Lu, C., Lin, S., Sun, C., Wen, Z., Ge, Z., Chen, W., Li, Y., Zhang, P., Wu, Y., Wang, W., Zhou, H., Li, X., Lai, Y., & Li, H.·A panel based on three-miRNAs as diagnostic biomarker for prostate cancer[J]. Frontiers in Genetics, 2024, 15:1371441.

ISSN 2616-5805 Vol. 6, Issue 2: 69-77, DOI: 10.25236/AJMS.2025.060210

[7] Wang, Y., Chen, X., Jiang, W., Li, L., Li, W., Yang, L., Liao, M., Lian, B., Lv, Y., Wang, S., Wang, S., & Li, X.·Predicting human microRNA precursors based on an optimized feature subset generated by GA–SVM[J].·Genomics,·2011,·98(2):73-78.

[8] Bo, Z., Zhenmei, C., Baorui, T., Chenhe, Y., Zhifei, L., Yitong, L., Weiqing, S., Jing, L., & Jinhong, C.·m6A target microRNAs in serum for cancer detection[J].·Molecular Cancer,·2021,·20(1):170.

[9] Shen, L., & Tan, E. C.·Dimension reduction-based penalized logistic regression for cancer classification using microarray data[J].·Ieee-Acm Transactions on Computational Biology and Bioinformatics,2005,2(2):166-175.

[10] Liang, Y., Liu, C., Luan, X.-Z., Leung, K.-S., Chan, T.-M., Xu, Z.-B., & Zhang, H.·Sparse logistic regression with a L1/2 penalty for gene selection in cancer classification[J].·Bmc Bioinformatics, 2013,14

[11] Algamal, Z. Y., & Lee, M. H.·Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification[J].·Computers in Biology and Medicine, 2015,·67:136-145.

[12] Lavanya, K., Rambabu, P., Suresh, G. V., & Bhandari, R.·Gene expression data classification with robust sparse logistic regression using fused regularisation[J].·International Journal of Ad Hoc and Ubiquitous Computing,·2023,·42(4):281-291.

[13] Breheny, P., & Huang, J.·Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection[J].The annals of applied statistics,·2011,·5(1):232.

[14] Wang, X., Li, Z., & Zhang, C.Integrated Analysis of Serum and Tissue microRNA Transcriptome for Biomarker Discovery in Gastric Cancer[J].Environmental toxicology,·2024,·)

[15] Zhang, Y., Wang, X., Liu, W., Lei, T., Qiao, T., Feng, W., & Song, W.CircGLIS3 promotes gastric cancer progression by regulating the miR-1343-3p/PGK1 pathway and inhibiting vimentin phosphorylation[J].Journal of Translational Medicine,2024,22(1):251.

[16] Zhang, H., Wang, J., Wang, Y., Li, J., Zhao, L., Zhang, T., & Liao, X.·Long non-coding LEF1-AS1 sponge miR-5100 regulates apoptosis and autophagy in gastric cancer cells via the miR-5100/DEK/AMPK-mTOR axis[J].·International Journal of Molecular Sciences,2022,23(9):4787.