Supplementary Material

1. Supplementary Tables

**Table S1. The p values of feature selection**

The detailed results of feature selection (the p values) are listed in Table S1.xlsx

**Table S2. Data after dimensional reduction**

The results of dimensional reduction are listed in Table S2.xlsx.

**Table S3. Pseudo dataset generated by SMOTE**

The results of data augmentation are listed in Table S3.xlsx.

**Table S4. Model performance indicator**

|  |  |  |
| --- | --- | --- |
| Index | Formula | Illustration |
| Accuracy |  | TP: actual positive and is recognized as positive |
| Precision |  | TN: actual negative and is recognized as negative |
| Sensitivity |  | FP: actual negative and is recognized as positive |
| Specificity |  | FN: actual positive and is recognized as negative |

**Table S5. The result of predictive performance comparison**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| AUC | SVM | LR | Naive Bayes | Ensemble |
| Accuracy | 0.9028±0.0337 | 0.9010±0.0391 | 0.8648±0.0450 | 0.9326±0.0301 |
| Precision | 0.9153±0.0596 | 0.8884±0.0678 | 0.7774±0.0827 | 0.9273±0.0487 |
| Sensitivity | 0.8255±0.0800 | 0.8539±0.0754 | 0.9253±0.0522 | 0.8967±0.0702 |
| Specificity | 0.9526±0.0302 | 0.9318±0.0398 | 0.8266±0.0780 | 0.9561±0.0277 |

**Table S6. The p values of T test on accuracy**

|  |  |  |
| --- | --- | --- |
| Model | Accuracy | P value |
| Ensemble | 0.9326±0.0301 | 2.4158e-08 |
| SVM | 0.9028±0.0337 |
| Ensemble | 0.9326±0.0301 | 6.4430e-08 |
| LR | 0.9010±0.0391 |
| Ensemble | 0.9326±0.0301 | 6.1118e-21 |
| NB | 0.8648±0.0450 |

1. Code

Code used for predictive model is available at:

<https://github.com/skytea/CRC>