RESEARCH

Paper 5 in applied machine learning

Aakash Nepal and Cuong Gia Pham

Full list of author information is available at the end of the article
*Equal contributor

Abstract

Goal of the project: Using gene expression data, develop an eXtreme Gradient Boosting classifier for colorectal cancer stages (adenoma to carcinoma). Compare performance and results with machine learning models from the paper by Lacalamita[1].

Main results of the project: We found that the XGBoost's performance is higher than the k - Nearest Neighbors model's performance in all three groups Control, Adenoma, and Cancer. However, it still needs to be refined and tuned to reduce the false negative and false positive rates, which are very important in screening for cancer in specific and in the biomedical field in general.

Personal key learning :

1 Pham Gia Cuong: An understanding of gradient boosting trees.
2 Aakash Nepal: was introduced to LASSO , and learned more about transcriptomics and XGBoost.

Estimation working hours:

1 Pham Gia Cuong: 7 hours per week
2 Aakash Nepal: 7 hours per week.

Project evaluation: 1

Number of words : Approx. 1800(without abstract, captions)

1 Goal of the project

Colorectal cancer (CRC) is a serious worldwide health problem, ranking third in terms of prevalence and second in terms of cancer-related mortality. CRC develops in stages, beginning with normal colonic epithelium and progressing to adenoma, carcinoma, and metastasis. The diagnosis of CRC and adenomatous polyps at an early stage is critical, however, existing screening techniques have limitations. The goal of this project was to develop an eXtreme Gradient Boosting(XGBoost) classifier capable of identifying and classifying distinct stages of colorectal cancer (CRC), with a particular focus on adenoma (a benign tumor) and carcinoma (a malignant tumor). This was accomplished by analyzing gene expression patterns in primary CRC, adenoma, and normal colon epithelial tissues. The generated classifier was evaluated using evaluation metrics such as accuracy, AUC-ROC and compared with models from the Lacalamita[1]. The top 10 features were identified using feature importance analysis and biomedical implications were discussed. By establishing an effective machine learning classifier, we hoped to find certain genes that may serve as reliable indications for the early detection of CRC which may help for further diagnosis.

2 Data

For this project, raw microarray data from four datasets from the Gene Expression Omnibus (GEO) database: GSE100179, GSE117606, GSE4183, and GSE71187 were used. Routine colonoscopies, formalin-fixed paraffin-embedded tissues, and frozen colonic biopsies were used to collect the data. There were 20 samples each in the Healthy Controls, Adenoma, and CRC categories in GSE100179. GSE117606 consisted of 65 Control samples, 59 Adenoma samples, and 74 CRC samples. GSE4183 included 8 Healthy Control samples, 15 Adenoma samples, and 15 CRC samples. GSE71187 included 12 control samples, 58 adenomas, and 99 CRC samples. In total, there were 465 samples, including 105 in the Healthy Controls group, 152 in the Adenoma group, and 208 in the CRC group. The differentially expressed genes (DEGs) were retrieved from each GEO database from all three sets of samples (normal, adenoma, and CRC). GEO2R[2] analysis was used to preprocess the raw data of gene expression profiles. GEO2R uses the limma approach which begins by fitting a linear model to the preprocessed data, followed by the use of an empirical Bayes method to moderate the standard errors of the estimated log-fold changes[4]. To find a statistically significant difference, the DEG analyses with GEO2R used an adjusted P value of 0.05 and $|\text{logFC}| > 0.263$ as a cutoff criterion. After getting significant genes for every dataset, they were only selected from each of the count matrices with the GEOquery[3] package which helped to load the expression data in R. At last, all of the count matrices were merged together using genes to get a single count matrix which contained 465 samples and 2369 significant genes.

3 Methods

3.1 Feature selection

One of the key steps in improving the model performance, reducing overfitting and improving generalization is the feature selection. We performed feature selection using LASSO (Least Absolute Shrinkage and Selection Operator) which is a type of regularization method used in linear regression to select the features and handle multicollinearity. It supports sparsity in coefficient estimates by adding a penalty component to the ordinary least squares objective function which decreases the less relevant features towards zero.

The LASSO method uses the cost function as shown in equation (1) where the a_j is the coefficient of the j−th feature. The final term is called l_1 penalty and the α is a hyperparameter that tunes the intensity of this penalty of the cost function. The

$$\frac{1}{2N_{\text{training}}}\sum_{i=1}^{N_{\text{training}}}(y_{\text{real}}(i)-y_{\text{pred}}(i))^2+\alpha\sum_{j=1}^{n}|a_j| \tag{1}$$

other variables used in the cost function are the number of training instances (Ntraining), the actual and predicted values (yreal and ypred) the number of features (n). Equation 1 computes the sum of absolute coefficient values and the total of prediction errors squared. It establishes a compromise between prediction accuracy and feature sparsity, making feature selection and model regularization straightforward.[4]

When doing feature selection using LASSO, we used a α value of 0.05 as regularization intensity. It was then fitted to the expression data, and the resultant coefficients were computed using the coef_ function. Because non-zero coefficients reflect selected features, their associated indices were gathered. The columns of the selected genes were obtained from the expression data using these indices.

3.2 Machine learning model

We have a dataset consisting of 208 colorectal cancer samples, 152 adenoma samples, and 105 healthy samples. To address the class imbalances, we used the SMOTE algorithm to create new samples based on nearest neighbors, resulting in a balanced dataset of 208 samples in each group. Each sample contains 813 genes. We then employed the XGBoost classifier, which is a scalable machine-learning system for tree boosting. XGBoost iteratively builds an ensemble of decision trees by correcting the mistakes made by previous trees. It combines weak learners, also called decision stumps, and evaluates their performance using a loss function. The loss function measures the discrepancy between the predicted and actual labels. Equation (2) shows us the formula to build the next tree from the previous tree in the XGBoost algorithm. As it is shown, by adding the loss function from the previous tree multiplied by a learning rate to create the next tree. This iterative approach results in an ensemble of boosting trees.

$$F(m) = F(m-1) + \mu * -\frac{\partial(L)}{\partial F(m-1)} \tag{2}$$

where F(m-1) stands for m-1 tree μ is learning rate.
After having an ensemble boosting trees model, equation (3) has been used to predict the output based on a given dataset, where $\hat{y}$ stands for predicted output,$\gamma_t(x)$ is the prediction made by an individual tree t for input x, λ is the regularization term that penalizes the complexity of the model and $f_t(x)$ is the complexity of the tree t for input x.

$$\hat{y} = \sum_{t=1}^{T} \gamma_t(x) + \lambda \sum_{t=1}^{T} f_t(x) \tag{3}$$

For tuning the parameters, such as learning rate, number of iterations ,and maximum depth of a single tree. We used the gridCV function from the sci-kit learn library. Consequently, the gridCV function gave us the best hyperparameters set regarding accuracy score with the learning rate equal to 0.1, the maximum depth of each tree is 3, and the number of iterations is 100.

3.3 Model evaluation

We used accuracy, and area under the ROC curve (AUC-ROC) as criteria for evaluation. The accuracy metric as described by equation (4) evaluates a classification model's overall accuracy by measuring the ratio of truly classified samples to the total number of samples.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.
The AUC-ROC is a summary measure of the model's performance that shows the likelihood that the model would score a randomly chosen positive instance higher than a randomly chosen negative instance. AUC-ROC values range from 0 to 1, with higher values suggesting stronger discrimination and prediction capability of the model.
Besides that, to be able to have a better overview of the performance between the XGBoost and k-NN model from the original paper by Lacalamita[1], we also gathered the precision, recall, and F1-score value of each model. Precision is a crucial evaluation criterion. Precision is the ratio of true positives predicted properly by the model and it is calculated as shown in equation (5).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

Similarly, Recall (also known as sensitivity) is a measure that indicates how well a model can recognize positive results. As outlined in equation (6), recall is obtained by dividing the number of TP by the total number of participants.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

As described by equation (7), the F1-score is a metric that considers both precision and recall. Only when recall and precision both have a value of 1, does the F1 Score become 1. Only when both recall and precision is strong can the F1 score rise.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

4 Results and Discussion

4.1 Model evaluation

From Figure 1 , we can see that the AUC has value of 1 for the normal group which shows that the model can correctly discriminate between normal cases and others. This also means that the model does well in properly detecting normal cases, with no false positives or false negatives. The AUC value for predicting the Adenoma group is thus equal to 0.98 which shows that the model is quite good at recognizing adenomas, with a minimal chance of misclassifying them as normal or malignant. Similarly, the AUC value of 0.97 for the CRC group demonstrates that the model is extremely good in distinguishing CRC cases from others. The model demonstrates significant discriminatory strength in properly identifying CRC occurrences, reducing the possibility of misclassification as normal or adenoma. Furthermore, when

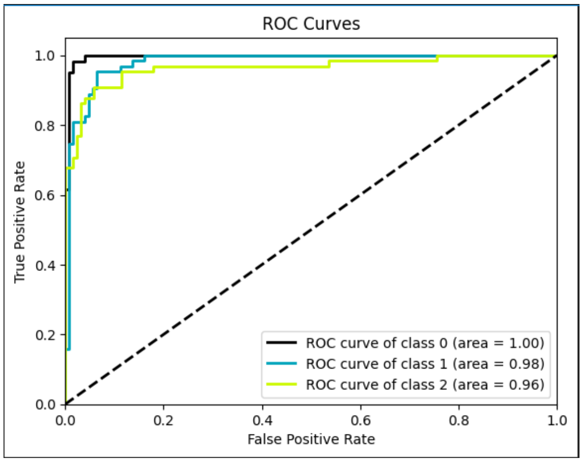

Figure 1 ROC curve of three classes on test set (0 : normal group, 1: Adenoma group, 2: CRC group)

compared to the best model (k-NN) from the paper by Lacalamita[1], the average AUC value of 0.97 from XGBoost is higher than the AUC value of 0.92 from k-NN. To be more specific, we obtained the accuracy, recall, and F1 score of the XGBoost model and compared it to the k-NN from the paper by Lacalamita[1]. As shown in Table 1, the XGBoost model performs better than the k-NN model when compared with precision which has a value of 0.98 for XGBoost vs 0.87 for k-NN in the control group. Similarly, as compared to k-NN with a recall value of 0.93 and F1-score of 0.90, the XGBoost model demonstrated a better recall value of 0.95 and F1-score of 0.97.

In the adenoma group, k-NN performed better with an accuracy of 0.93 and marginally better in the F1-score of 0.91 than the XGBoost model, which scored 0.86 in precision and 0.90 in F1-score. XGBoost, on the other hand, showed higher recall of 0.95 and F1-score of 0.90 than k-NN with a recall of 0.90 and F1-score of 0.91.

For the CRC class, both of the models obtained comparable accuracy, with kNN and XGBoost model achieving scores of 0.93 and 0.93, respectively. However, XG-Boost outperformed kNN in terms of recall which has a score of 0.86 and F1-score

Table 1 Test classification performances: Sensitivity, Precision and F1 score for k-NN and XGBoost on each class

	Precision		Recall		F1-score	
	k-NN	XGBoost	k-NN	XGBoost	k-NN	XGBoost
Control	0.87	0.98	0.93	0.95	0.90	0.97
Adenoma	0.93	0.86	0.90	0.95	0.91	0.90
CRC	0.93	0.93	0.90	0.86	0.91	0.90

of 0.90. In general, the results show that the XGBoost model is better than the k-NN model in terms of accuracy, recall, and F1-score across all classes. Furthermore, This also implies that the XGBoost model can classify better than the kNN model. The accuracy, recall, and F1-score formulas are shown in equations 5, 6, and 7 in methods section.

4.2 Feature importance

Table 2 The 8 most important features identified by XGBoost model

Order	Gene name	Description
1	CA4	carbonic anhydrase 4
2	PPA1	inorganic pyrophosphatase 1
3	MS4A12	membrane spanning 4-domains A12
4	CD177	CD177 molecule
5	PKIB	cAMP-dependent protein kinase inhibitor beta
6	EPB41L2	erythrocyte membrane protein band 4.1 like 2
7	GUCA2A	guanylate cyclase activator 2A
8	CA1	carbonic anhydrase 1

The Table 2 shows the 8 most important genes that were found by feature importance analysis which could be potential bio-markers for early diagnosis of CRC. There was only one gene named ”PKIB” that we found was also reported by the paper from Lacalamita[1]. PKIB stimulates cell proliferation and has been found to be overexpressed in lung cancer. Furthermore, the paper from Lacalamita[1] points out that the expression pattern of this gene diminishes from normal mucosa to adenoma and CRC. However, all the genes listed at the Table 2 are key hub genes found to be related to CRC development. For example, the CA4 gene which is associated with the inhibition of CRC development, and PPA1 which is related to the regulation of CRC development[5][6]. Furthermore, we also discovered that genes such as MS4A12 and GUCA2A are expressed in the apical membrane of the colonic epithelium and that their expression decreases as cancer develops[7][8]. Hence, these results contribute to our understanding of the molecular processes behind CRC and may pave the way for novel biomarkers and treatment targets.

5 Contributions

1 Pham Gia Cuong: Overall 50% report and ran the XGBoost.

2 Nepal Aakash: Overall 50% report and did the preprocessing, plus extracted the feature importance.

References

[1] Andrea Lacalamita et al. “A Gene-Based Machine Learning Classifier Associated to the Colorectal Adenoma-Carcinoma Sequence”. In: *Biomedicines* 9.12 (2021). Published Dec 17, 2021, p. 1937. DOI: 10.3390/biomedicines9121937. URL: https://doi.org/10.3390/biomedicines9121937.

[2] Tanya Barrett et al. “NCBI GEO: archive for functional genomics data sets—update”. In: *Nucleic Acids Research* 41.D1 (2013), pp. D991–D995. DOI: 10.1093/nar/gks1193. URL: https://doi.org/10.1093/nar/gks1193.

[3] Sean Davis and Paul S. Meltzer. “GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor”. In: *Bioinformatics* 23.14 (2007), pp. 1846–1847. DOI: 10.1093/bioinformatics/btm254.

[4] YourDataTeacher. *Feature Selection in Machine Learning Using Lasso Regression.* (Accessed: June 15, 2023). URL: https://www.yourdatateacher.com/2021/05/05/feature-selection-in-machine-learning-using-lasso-regression/.

[5] Jinyun Zhang et al. “Carbonic anhydrase IV inhibits colon cancer development by inhibiting the Wnt signalling pathway through targeting the WTAP-WT1-TBL1 axis”. In: *Gut* 65.9 (2016), pp. 1482–1493. DOI: 10.1136/gutjnl-2014-308614.

[6] Peng Wang et al. “PPA1 regulates tumor malignant potential and clinical outcome of colon adenocarcinoma through JNK pathways”. In: *Oncotarget* 8.35 (2017), pp. 58611–58624. DOI: 10.18632/oncotarget.17381.

[7] JW Han et al. “Plasma Membrane Localized GCaMP-MS4A12 by Orai1 Co-Expression Shows Thapsigargin- and Ca2+-Dependent Fluorescence Increases”. In: *Mol Cells* 44.4 (Apr. 2021), pp. 223–232. DOI: 10.14348/molcells.2021.2031.

[8] Babar Bashir et al. “Silencing the GUCA2A-GUCY2C tumor suppressor axis in CIN, serrated, and MSI colorectal neoplasia”. In: *Hum Pathol* 87 (May 2019), pp. 103–114. DOI: 10.1016/j.humpath.2018.11.032.