Introduction

Colorectal cancer (CRC) is a disease characterized by the abnormal growth of cells in the colon and rectum. CRC is one of the most common cancers and is one of the leading causes of cancer deaths[1]. Oligometastatic CRC (omCRC) describes a severe form of colorectal cancer where a small number of metastatic tumors spread around the body[2]. While in the past, treatments were limited, CRC patients today have many treatments available to them. A treatment of particular interest is stereotactic body radiation therapy (SBRT). SBRT is a treatment characterized by concentrated doses of radiation to a highly specified area over a period of many weeks and is often used for patients with inoperable cancers[3,4]. Despite the success of SBRT, some patients who are treated undergo rapid disease progression, a phenomenon known as early distant progression (EDP)[5]. For these patients, SBRT did not prevent disease progression, yet still induced the side effects of radiation and costed precious time for terminal patients for whom time is limited. As such, it is crucial for oncologists to be able to identify patients likely to undergo EDP prior to SBRT treatment. In recent years machine learning algorithms have been successfully applied to predicting patient outcomes and evaluating the efficacy of treatments[6]. When fed demographic and treatment information, machine learning algorithms can find correlations between patient data and the EDP outcome of a patient. Therefore, the purpose of this study is to create and assess the effectiveness of a machine learning model capable of predicting early distant progression of omCRC patients who are to receive SBRT.

Methods

We examined a dataset of 203 unique lesions taken from 131 patients with omCRC. Each patient had received radiation doses ranging from 30-60 Gy given in 4-6 fractions. Various demographic and treatment data including age, previous surgical treatments and tumor location were obtained. Data was then cleaned by Dr. Lang before being used in data analysis. Python and Google Collaboratory were used for all stages of the project and all code can be found in the appendix. Scikit-learn machine learning libraries were used for building models.

Processing

Null values of the data features were addressed by removing features that were null in more 30% of the training set. This ensures the total number of lesions would not fall below a point where accuracy of the model would be affected. Features removed include the RAS gene, previous adjuvant systemic therapy type and previous definite systemic therapy types.

Splitting the Dataset

Following processing, dataset was split into a 70/30 training and testing dataset. The dataset was stratified based on the whether the lesion in question showed disease progression after SBRT to ensure a consistent event rate across both testing and training groups. Each dataset was then scaled to ensure equal weighting of all features.

Feature Selection

The training dataset was fed into the logistic regression algorithm (LR) and LASSO feature selection was performed to produce the most significant features. A logistic regression model takes the training data, fitting them to the sigmoid curve, and uses the resulting model to predict outcomes for any new data fed to it. LASSO refers to the process of weighting different features based on their relevance to the outcome and selecting features that are most significant. These features were then identified and used to train following models.

Machine Learning Models

In addition to LR, the selected features were used to train a support vector machine (SVM) and random forest classifier (RFC). SVM is a model that finds an optimal dividing plane between clusters of data. A plot of a hypothetical dataset with 2 features and the resulting plane is shown in figure 1. The RFC predicts outcomes through a series of decision trees, where each tree considers different combinations of features and votes on the outcome of the model as shown in figure 2.

Parameters for the models were selected through trial and error based on performance on the testing dataset. Various values were entered and ones that produced the best confusion matrix were chosen. Trained models were evaluated on the testing dataset and results were evaluated using receiver operating characteristic curves (ROC). To assess performance, accuracy, sensitivity, and specificity were calculated as:

Accuracy = (tp + tn) / (tp + tn + fp + fn)

sensitivity = tp / (tp + rn)

Specificity = tn / (fp + tn)

Where tn = true negative, tp = true positive, fp = false positive and fn = false negative.

Results

LASSO Feature Selection

LASSO feature selection produced 6 significant features shown in Table 1 with their corresponding coefficients.

Logistic Regression

Shown in figure 3, LR model produced 51 true negatives and 5 false negatives. This results in an accuracy of 0.911, sensitivity of 0 and specificity of 0. Area under the ROC curve (AUC) is shown in figure 4 and was found to be 0.73.

Support Vector Machine

As shown in figure 5, SVM produced 51 true negatives, 2 false negatives and 3 true positives. This results in an accuracy of 0.964, sensitivity of 0.6 and specificity of 1. AUC was found to be 0.91 as shown in figure 6.

Random Forest Classifier

As shown in figure 7, RFC produces 51 true negatives, 3 false negatives and 2 true positives, leading to an accuracy of 0.946, a sensitivity of 0.4 and specificity of 1. An AUC of 0.969 was found as shown in figure 8.

Discussion

Features selected by LASSO selection followed expectations, with some exceptions. Due to the nature of omCRC, severe disease is characterized lesions spreading around the body. Because of this, a positive correlation between EDP and the number of lesions is expected. Surgery and radiofrequency ablation (RFA) were also positively correlated with EDP. This may be due to the fact that RFA and surgery are both common practices for treating tumors that have metastasized, suggesting a more severe disease stage. Negative value of age is surprising as despite previous literature suggesting a correlation between age and decreased overall survival, LASSO selection found younger age patients to be more susceptible to EDP[8].

Logistic regression performed the worst overall, scoring lower than the other models in all assessment parameters as seen in figure 3 and 4. Much of this performance can be attributed to unoptimized parameters. Because logistic regression is the method through which LASSO selection obtained features, optimizing the model parameters also considered the number of features produced as they will be fed into the other models. As a result, the logistic regression pipeline was expected to have reduced performance as producing the best outcome was not the sole priority.

Shown in figures 5-8, the random forest classifier produced the best AUC, while the SVM produced the best confusion matrix numbers. This suggests that the random forest classifier parameters can be optimized further, surpassing the performance of the SVM. Such optimization may involve a formal optimization process such as grid search and can be an area of future work. However, with the parameters chosen for the models in this study, the SVM performed best.

The dataset suffered from a low event rate. The entire dataset only contains 21 positives among 203 lesions. Because of the lower number of patients that experienced EDP in the dataset, machine learning models had difficulties identifying the features that would determine EDP. Such an issue could be solved in future studies with cross–validation. With this method, the same dataset would be sampled multiple times through different training and testing sets, imitating an increased dataset, and producing more consistent results. Future work can also utilize this method with LASSO selection, where the same dataset is split multiple times and the significant features that are common between the splits are chosen. Such methods would decrease the impact of the low event rate and improve the consistency of the model.

Conclusion

When considering SBRT treatment for a patient, one must consider the possibility of EDP. In this study, we utilized machine learning models to retrospectively identify patients that would undergo EDP. We found especially strong performances from the support vector machine and random forest classifier

which, after further optimization, could prove to be effective tools in assessment patient suitability with SBRT.

References

1. Canadian Cancer Statistics Advisory Committee in collaboration with the Canadian Cancer Society. *Canadian Cancer Statistics 2021.* Updated November 2021. Accessed March 28, 2022. https://cdn.cancer.ca/-/media/files/research/cancer-statistics/2021-statistics/2021-pdf-en-final.pdf?rev=2b9d2be7a2d34c1dab6a01c6b0a6a32d&hash=01DE85401DBF0217F8B64F2B7DF43986
2. Comito T, Cozzi L, Clerici E, et al. Stereotactic Ablative Radiotherapy (SABR) in inoperable oligometastatic disease from colorectal cancer: a safe and effective approach. *BMC Cancer*. 2014;14:619. doi:10.1186/1471-2407-14-619
3. Pan H, Simpson DR, Mell LK, Mundt AJ, Lawson JD. A Survey of Stereotactic Body Radiation Therapy Use in the United States. *Cancer*. 2011;117(19):4566-4572. doi:10.1002/cncr.26067
4. Petrelli F, Comito T, Barni S, et al. Stereotactic body radiotherapy for colorectal cancer liver metastases: A systematic review. *Radiother Oncol*. 2018;129(3):427-434. doi:10.1016/j.radonc.2018.06.035
5. Klement RJ, Abbasi-Senger N, Adebahr S, et al. The impact of local control on overall survival after stereotactic body radiotherapy for liver and lung metastases from colorectal cancer: a combined analysis of 388 patients with 500 metastases. *BMC Cancer*. 2019;19:173. doi:10.1186/s12885-019-5362-5
6. Nartowt BJ, Hart GR, Muhammad W, Liang Y, Stark GF, Deng J. Robust Machine Learning for Colorectal Cancer Risk Prediction and Stratification. *Front Big Data*. 2020;3:6. doi:10.3389/fdata.2020.00006
7. Gandhi R. Support Vector Machine — Introduction to Machine Learning Algorithms. Medium. Published July 5, 2018. Accessed March 28, 2022. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47
8. McKay A, Donaleshen J, Helewa RM, et al. Does young age influence the prognosis of colorectal cancer: a population-based analysis. *World Journal of Surgical Oncology*. 2014;12(1):370. doi:10.1186/1477-7819-12-370

Tables and Figures

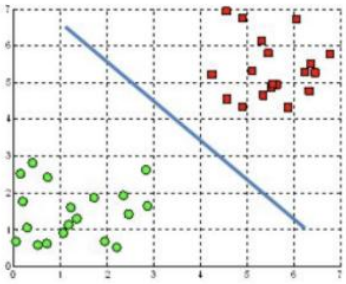

Figure 1. A plot of the plane that a SVM produces to distinguish datasets.[7]

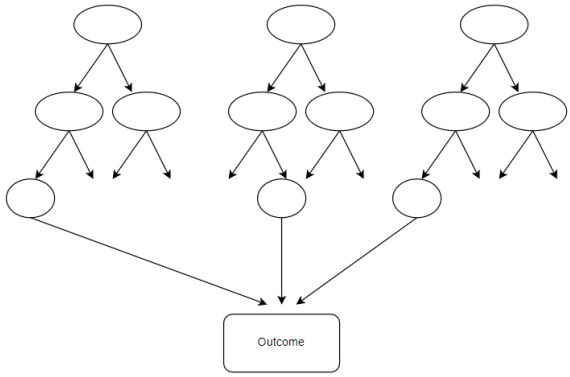

Figure 2. Random forest classifier operates with multiple decisions trees that contribute to an outcome. Shown here are the many decision trees, each coming to their own conclusion, with the outcome of the RFC being dependent on the collective conclusion.

Feature	Coefficient
Number of Lesions	0.069316
Age	-0.023444
Disease Free Interval	-0.092876
Previous Surgery	0.166533
Previous radiofrequency ablation	0.115118
Primary tumor at ascending colon or cecum	0.213037

Table 1. Table of features selected by LASSO. Greater coefficients represent greater relevance and importance to determining outcome. Negative coefficients represent inverse relationships.

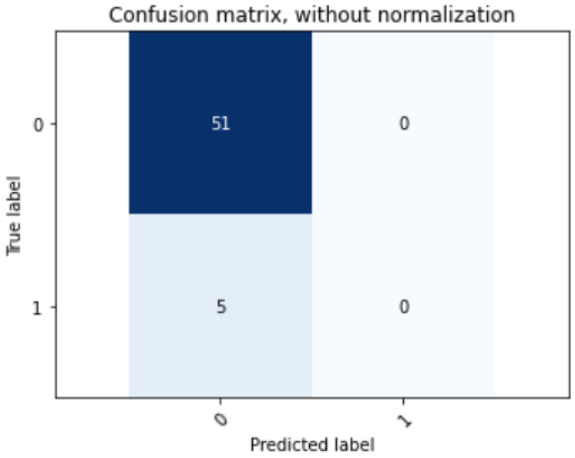

Figure 3. Confusion matrix of the logistic regression model's prediction outcomes.

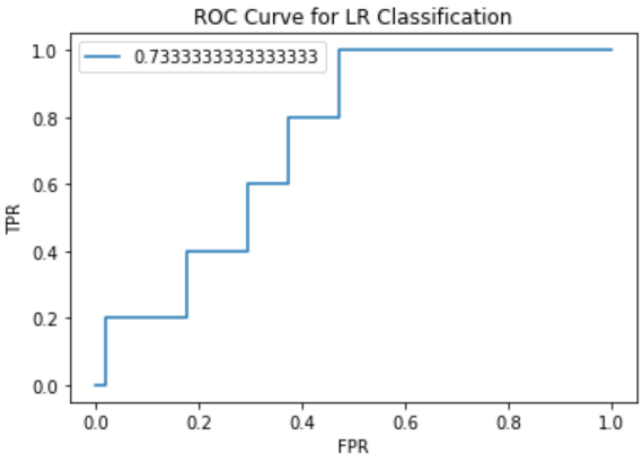

Figure 4. ROC curve for logistic regression. AUC is 0.73

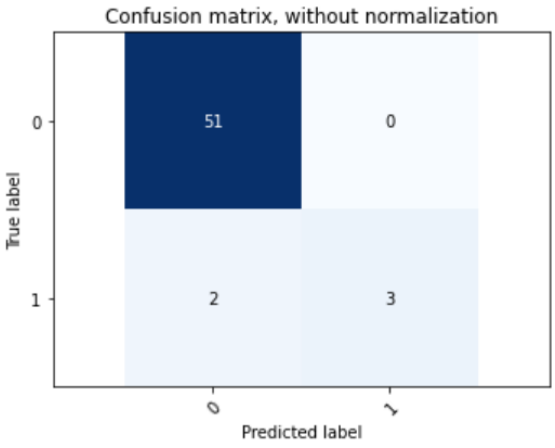

Figure 5. Confusion matrix of SVM model's prediction outcomes.

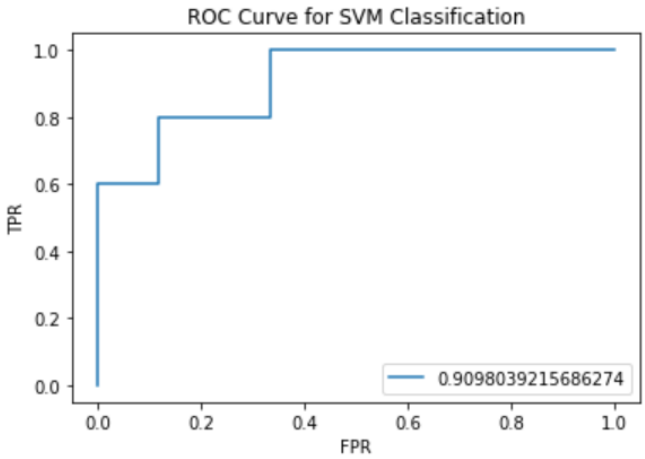

Figure 6. ROC curve for SVM model. AUC is 0.91

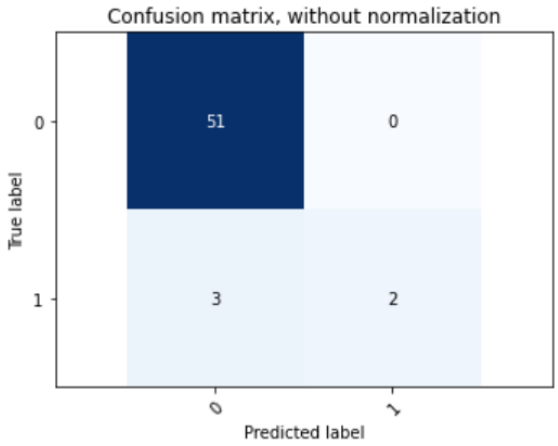

Figure 7. Confusion matrix of RFC model's prediction outcomes.

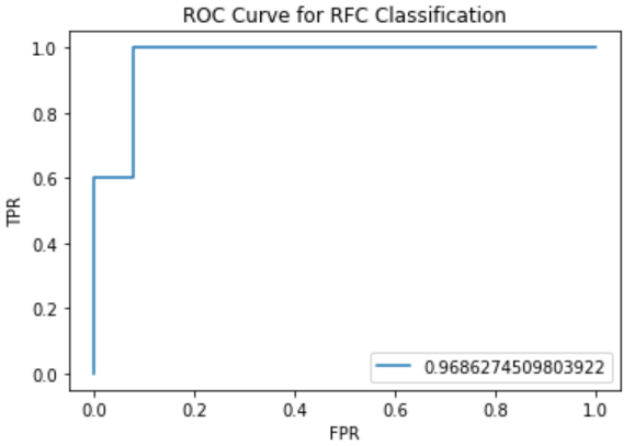

Figure 8. ROC curve for RFC model. AUC is 0.91