Machine learning versus logistic regression methods on predicting pathological features for microsatellite instability in colorectal cancer

Kevin Tsang, Jae Young Kim

Department of Computer Science, University of Southern California, Los Angeles, California 90089, USA

(Dated: May 2, 2021)

Colorectal Cancer, as the third most common cancer diagnosed in US, can be categorized by having high levels of microsatellite instability (MSI-H). Previous research has been done by Greenson et al that used logistic regression and found that Tumor infiltrating lymphocytes (TILS) and Absence of Dirty Necrosis (NoNecrosis) are the crucial predictors of MSI-H. Our team validated the methods of the Greenson et al's study and explore the feature importance of the dataset using twelve total machine learning models. Our results showed consensus in TILS, Moderate Differentiation and NoNecrosis are three of the most important features.

Random Forest, LightGBM and Naive Bayes Classifier model showed the best performance among twelve different models. Especially, Random Forest model showed 0.793 ROC AUC score and it is 4.4% better than the baseline logistic regression model(0.749).

Keywords: Machine Learning, Imbalanced Data, Cancer Biology

I. INTRODUCTION

Colorectal cancer or colorectal carcinoma (CRC) is the third most common cancer diagnosed in both men and women in the United States (Cancer.org) and based on molecular genetic studies, 10 to 15% of CRCs have shown high levels of microsatellite instability (MSI-H). MSI-H is a biological phenomenon when the microsatellite, also known as Short Tandem Repeats, repeated sequences of 1-6 nucleotides, has mutated and is different from what it was when the microsatellite was inherited. This occurs when DNA replication errors do not get repaired by mismatch repair (MMR). MSI-H colorectal cancers have a better prognosis and have shown different responses to chemotherapy [1]. Therefore, it is crucial to know if the colorectal tumors exhibit microsatellite instability. One way to do this is to study the pathological features of tumor cells and to predict whether a tumor is MSI-H. In 2009, Greenson et al. studied the histological features that are most common in MSI-H CRCs by viewing cellular and tissue structure details of the tumor blocks using Hematoxylin and Eosin (H&E) slides under the microscope.[1] These histological features are tumor grade, mucinous differentiation, signet ring cells, histological heterogeneity, growth pattern of tumor at advancing stage, tumor necrosis, prominent Crohn's-like host response, and tumor infiltrating lymphocytes.

A. Previous Logistic Regression Model

From the paper, Pathologic Predictors of Microsatellite Instability in Colorectal Cancer, Greenson2009 model presented a logistic regression model to predict whether the patient's MSI is positive or not. The model resulted in a 0.850 AUC score. Due to the loss in access for the previous script, Dr. Joe Bonner (one of the authors) has made another logistic regression model for new data collected in 2020. However, the AUC score of the pure logistic regression model was 0.747 and the AUC score of the ridge regression model was 0.749, which is a 0.10 difference from the published model.

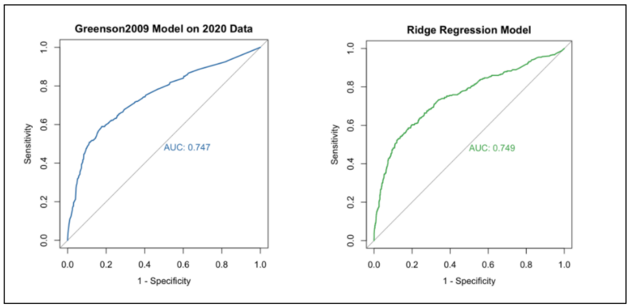

Figure 1. ROC AUC of Previous Models

Our team set out to confirm the model performance of the new set of data in 2020 and explore various kinds of Machine learning models such as k-Nearest Neighbor model(KNN), Decision Tree (DT), Random Forest (RF), Light Gradient Boosting Machine (Light GBM), Naive Bayes Classifier (NB), Support Vector Machine(SVM) and Artificial Neural Network (ANN) in hopes to improve the performance of the model and reveal the important features detecting MSI patients.

B. Feature Explanation

Tumors were categorized by well, moderate or poor differentiation grade defined by Jass et al (Jass, 1986). Mucinous differentiation accounts for tumors with greater than half of the surface area showing extracellular mucin (mucinous), else, it was classified as focal mucinous differentiation. Signet ring cells are cells with a large vacuole and the tumors are graded based on if the vacuole is greater than 50% of the surface area. Histological heterogeneity is a feature when tumors have at least two distinct growth patterns, mucinous tumors excluded (mucinous and nonmucinous areas by definition are two distinct growth patterns). Growth patterns of tumor at advanced stage are studied under low power to determine if the tumor grew with an expansile pattern or an infiltrative pattern (Jass, 1986). Tumor necrosis and prominent Crohn's-like host response are both histological patterns that account for dirty or garand necrosis and inflammatory response. Respectively. Tumor infiltrating lymphocytes (TIL) are small blue mononuclear cells that had a halo around them. These are white blood cells that infiltrated into the tumor cells, which signifies the patient's own body's retaliation to the cancer. Pathologist counted the number of TILs in the areas (5 consecutive 40x fields of an Olympus BX40 microscope with a UPlanF1 objective with the most TILs.

II. METHODS

A. Preprocessing

1. Feature Selection

The features are selected based on the criteria given by Greenson et al. in a R script since they are the domain experts for the pathological features. We have used the same logic to create a clean dataset with some feature engineering, such as creating dummy variables for categorical columns. For various models, forward selection technique was used to optimize the feature importance and model ROC-AUC scores.

2. Synthetic Minority Oversampling Technique

Target feature of data, MSI is highly imbalanced. 307 patients have MSI value 1, and 1603 patients have MSI 0. The ratio of MSI 1 and 0 is about 1:5. This imbalanced data can cause poor performance of the model, especially, for the minority class which is important for our research. Therefore, Synthetic Minority Oversampling Technique(SMOTE) was implemented. SMOTE is an oversampling method which creates data of minority classes. In the preprocessing step, SMOTE was used to make the number of class 0 and 1 to be the same.

III. MODELS

A. Logistic Regression

Logistic Regression (LR) is a kind of linear regression which can be applied to binary classification. Logistic regression is used to explain the relationship between a binary target variable and one or more numerical predictor variables.

It can be regularized to avoid overfitting in three different ways, ridge regression, Lasso and Elastic net. These three methods give penalty to the regression coefficients. Ridge regression gives penalty with l2 norm, Lasso gives penalty with l1 norm and Elastic net gives both l1 and l2 norm penalty. [2]

B. K-Nearest Neighbors

The K-Nearest Neighbors (KNN) algorithm assumes that similar things, such as data points, exist in close proximity, and the similarity of data points are defined by, most commonly, Euclidean distance. It is a simple but effective method for classification as it looks at k number of neighbors around the data point and classifies the data point according to the class most similar to its neighbors. The hyperparameter k is based on the user input and it is non-parametric, which means it makes no assumptions about the data structure. Also the algorithm makes no generalizations.

C. Support Vector Machine

The Support Vector Machine (SVM) algorithm's objective is to find a hyperplane in an N-dimensional space that best classifies the data points, where N is the number of features. There are many possibilities for choosing the best hyperplane. The SVM can either linearly separate the dataset or deal with higher dimensional data with kernel trick. The support vectors are data points close to the hyperplane and will change the orientation of the hyperplane if removed. The goal is to maximize the margin from the support vectors to the hyperplane to achieve best separation of each class.

D. Naive Bayes Classifier

The Naive Bayes (NB) Classifier is a classification algorithm that is based on Bayes' Theorem that finds the probability of an event occurring based on the occurrence of another event. The Bayes Theorem works on conditional probability for an event, which is influenced by the prior knowledge of conditions that might be related to said event. With a complicated set of data, it will be very complicated to calculate the conditional probability of events since many features can influence the class of an event. Therefore, in Naive Bayes classifier, it assumes that all the features are unrelated to each other. This makes the calculation for the hypothesis more tractable. There are several types of NB Classifiers: Gaussian NB, Multinomial NB, and Bernoulli NB. Bernoulli NB fits the dataset the best since it requires value to be binary valued as it assumes each feature to have a binary, Bernoulli distribution whereas Gaussian NB assumes a

Gaussian distribution for each features and the Multinomial NB prefers data that is multinomially distribution, most commonly, data for text classification.

E. Artificial Neural Network

Artificial Neural Networks (ANNs) are machine learning algorithms that mimics the learning process that human brains go through. Neural nets consist of an artificial network of functions, which is presented as layers of neurons and each layer receives inputs from the output of previous layers. A cost function is defined to see how effective the ANN is and each neuron has a weight associated with it and for each iteration of training, also known as, epochs, the cost function is analyzed to see how the weights of each neuron needs to be changed. That process is called back propagation. This process requires a lot of training data for the cost function to be minimized. The architecture of a neural network is very complex and does not provide a clear distinction on how the features of a data set influence the neurons. Therefore, while it is a great model for predicting target variables with an abundance of training data, it lacks the interpretability that other machine learning algorithms possess.

F. Tree models

Decision tree algorithms are an important type of machine learning algorithm that excels at classification analysis. In the research, we used four kinds of tree based algorithms, Decision Tree (DT), Random Forest (RF), XGBoost and Light Gradient Boosting Machine (Light GBM).

1. Decision Tree

Basic decision tree model adopts a binary tree representation which predicts the outcome by asking a set of if-else questions. It finds optimal features by performing data splitting and creating a branching tree. Each tree has a root node that best divides the data and below that are internal nodes (have a parent node, and give two children nodes) and leaf nodes (have a parent node but do not have children nodes). Each branch indicates the values the node can assume, which is oftentimes the boolean answer to the if/else question. Node represents input variables and split points of the dataset while the leaf nodes are output that used to do prediction. Decision tree model determines feature importance scores based on the reduction in the criterion to select split points like Entropy. [3]

2. Random Forest

Random Forest (RF) creates a large number of trees while decision trees build a simple single tree. It creates a large number of training samples and trains decision trees for every sample. After training, based on each trained model, it votes on the predicting results (Bagging). After voting, the most voted result becomes the final prediction result. As it is an Ensemble Learning algorithm of different training samples, it can avoid overfitting.

3. XGBoost

XGBoost is also an Ensemble Model based on decision trees. However, it is a kind of gradient boosting model. Similar to Random forests, xgboost also combines a large number of decision trees. XGboost combines the trees along the way, while Random forest combines the trees at the end of the process. Each of the trees in XGBoost is grown using information from the previously grown trees. (Boosting) As XGBoost has the advantage of both Bagging and Boosting, it is widely used in classification and regression tasks. To approximate tree learning, it proposed a sparsity aware split finding which led the algorithm to learn the sparsity pattern of the data and weighted quantile sketch to propose candidate splitting points. [4]

4. Light Gradient Boosting Machine

Light Gradient Boosting Machine (LightGBM) is also a kind of Gradient boosting ensemble model. LightGBM tackled the way of growing trees of previous Gradient boosting tree models. The efficiency and scalability were not efficient enough since other tree boosting algorithms had to scan all data instances to calculate the information gain of each feature in all splitting points. Therefore, LightGBM proposed Gradient-based One-Side Sampling (GOSS) which excludes a significant amount of data with relatively small gradients. As large gradients have a major effect on information gain, GOSS made LightGBM efficient in terms of time consuming without hurting performance. Another important feature of LightGBM is Exclusive Feature Bundling (EFB). EFB makes it possible to reduce the number of features by using a graph coloring problem. As most data used in real applications have many features and they are sparse (such as one hot encoded features), it can reduce the number of features without hurting performance. [5]

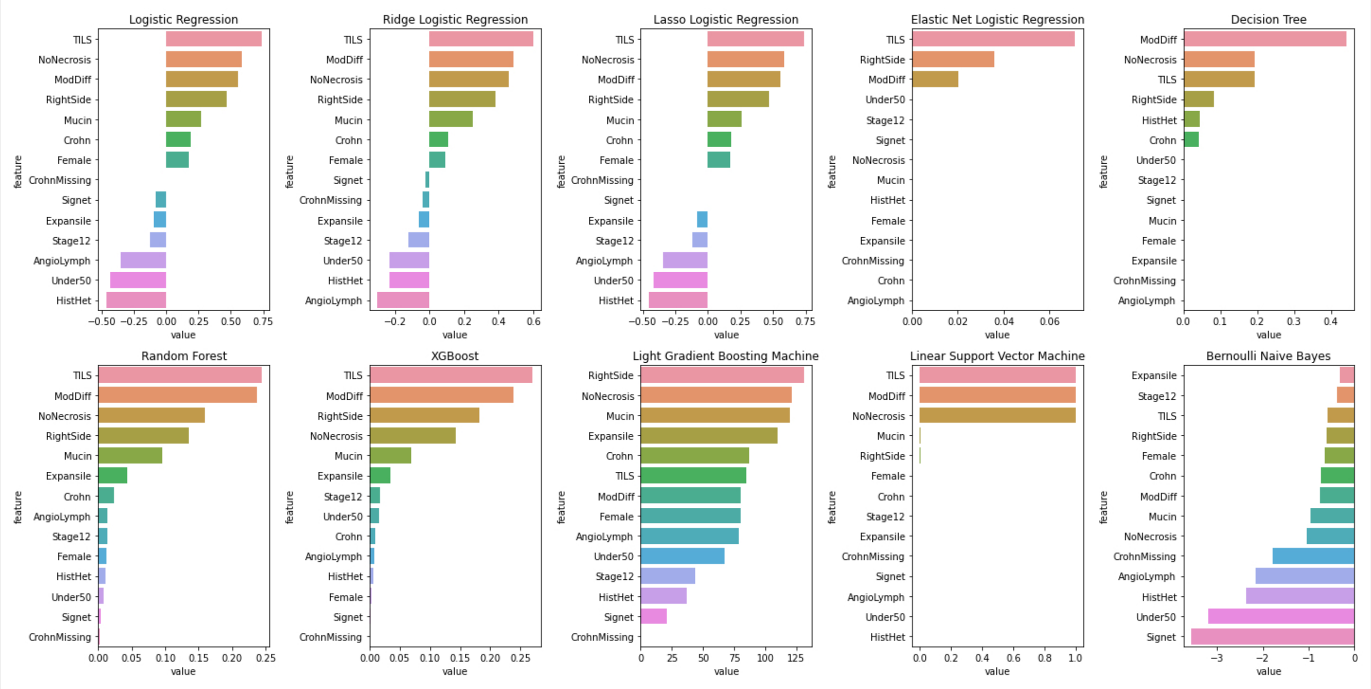

Figure 2. Feature Importance ranking plots for all models

IV. RESULTS

Understanding the importance of features that predict MSI is crucial in assisting clinicians to come up with the best course of action for patients' treatment. We examined the feature importance ranking for each of the models (Figure 2) chosen in this study except for ANN and KNN because those two models did not have inherent feature importance coefficients. We only focused on the models that have that feature, hence we have 10 plots. Out of all the plots, TILS is the most important feature in all four of the logistic regression models and in random forest, XGBoost and Linear SVM (7 out of 10 models). Decision Tree shows ModDiff as first while Light GBM shows RightSide. Bernoulli Naive Bayes Classifier has an interesting feature important plot where it shows Signet ring feature as the one that has the most negative effect as we see all of the features have negative impact for the model predictions. As for the second most important feature, NoNecrosis ranked second in four models (LR, Lasso LR, DT, and Light GBM) and ModDiff ranked second in four models as well (Ridge LR, RF, XGBoost, and Linear SVM) and RightSide ranked second for Elastic Net LR. For the third most important feature, three models (LR, Lasso LR, and Elastic Net LR) agreed on ModDiff, three other models (RF, Linear SVM and Ridge LR) agreed on NoNecrosis while DT showed TILS, XGBoost showed RightSide, and Light GBM showed Mucin.

To measure the performance of the model, ROC AUC score was used since test data was imbalanced. The ROC AUC score of the ridge regression model of Dr. Joe Bonner was 0.749. We tried 12 different machine learning models. Among the 12 models, Random Forest model and LightGBM model showed the best score, 0.793 and 0.790 for each. The single decision tree model showed the lowest performance, 0.738 but generally, tree based models showed high performance.

Model	Roc Auc Score(validation)	Roc Auc(test)
Logistic Regression	0.712	0.755
Ridge Regression	0.709	0.776
Lasso Regression	0.712	0.758
Elastic Net	0.686	0.777
Decision Tree	0.679	0.738
Random Forest	0.736	0.793
XGBoost	0.711	0.773
LightGBM	0.730	0.790
Linear SVM	0.701	0.774
KNN	0.572	0.589
Naive Bayes	0.699	0.789
Neural Network	0.749	0.753

Table I. Table. 1. ROC AUC score of 12 ML models

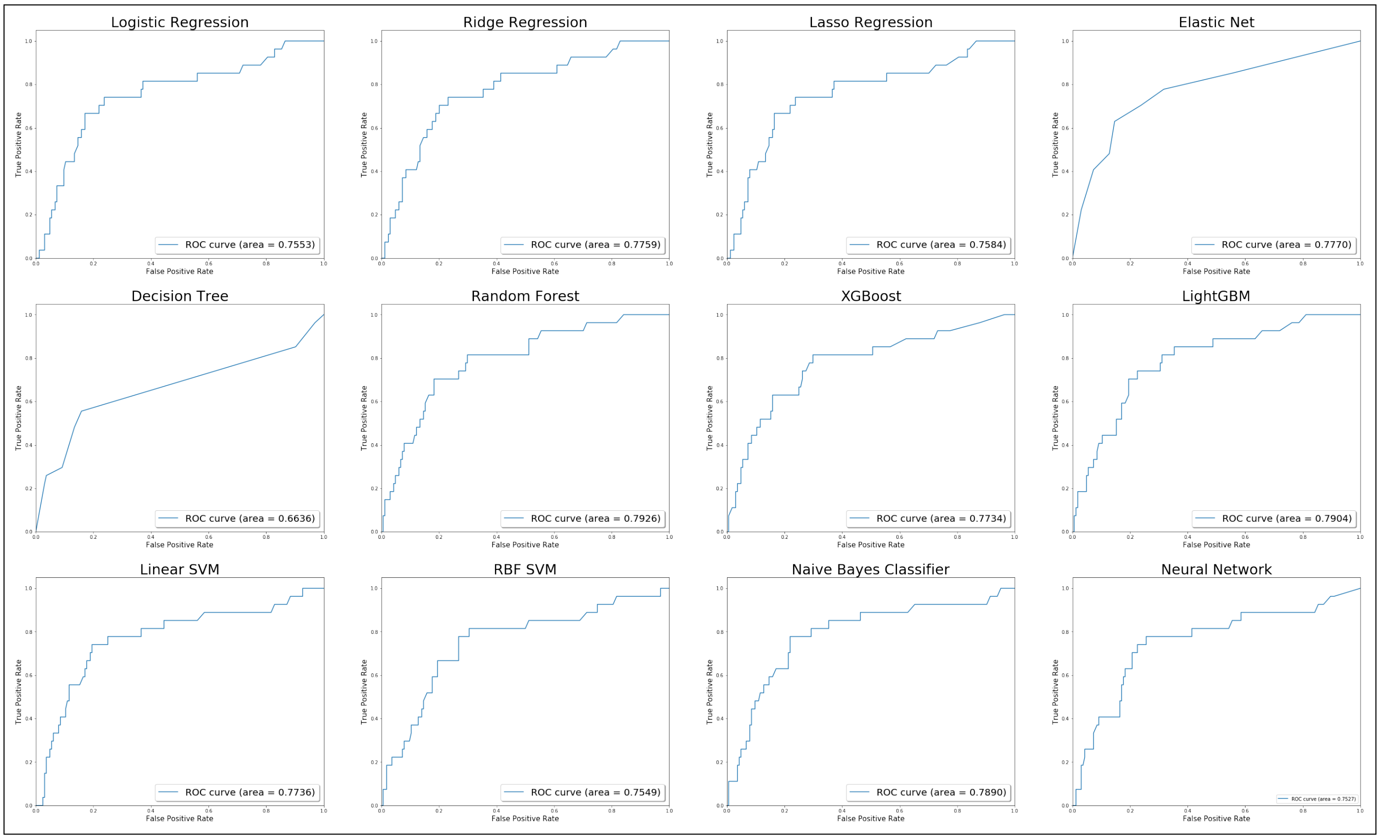

Figure 3. ROC AUC Plots of 12 Machine Learning Models

V. DISCUSSION

Based on the results shown in Figure 2, it is clear that TILS is the most important feature since it is a consensus amongst 7 out of 10 models. While there is a greater split in deciding what the second and the third most important features are, one can observe that NoNecrosis and ModDiff are likely candidates for that. RightSide is also very informative as Light GBM indicates it was the most important, Elastic Net LR ranked it second and XGBoost ranked it third. Therefore, this study shows TILS, NoNecrosis, ModDiff and RightSide are the most important features to predict MSI status for a patient by examining the histology of the tumor slide. Although no formal method is introduced to finalize the ranking, this is good insight for clinicians to narrow down the testing based on these four features. This finding aligns with Greenson et al papers in 2003 and 2009 where they had concluded that TILS and presence of dirty necrosis are the top two features [1]. With this information, clinicians can now use these features to decide how they can filter out patient samples that do not show any presence in all four of the features and this is beyond the scope of the team's research.

In terms of performance, the 12 models showed higher score than 0.73 ROC AUC score. In addition, Random Forest, LightGBM and Naive Bayes Classifier reached 0.793, 0.790 and 0.789 ROC AUC socre. It is higher than the performance of baseline logistic regression model. Generally, tree models showed better performance and we think tree models had strong performance since most features in our selected features were categorical features. However, they could not reach 0.85 ROC AUC score which is the result of Greenson2009 logistic regression model. Even though we used various ML models with optimized hyperparametes, there was limitation. We think the data difference caused this performance difference. Furthermore, we followed the feature selection process used in Greenson model. However, we will try other feature selection methods as our models had limited performance.

VI. CONCLUSIONS

In this study, we have validated the research done in 2009 by Greenson et al and agree with their findings of Tumor infiltrating lymphocytes (TILS) and Absence of Dirty Necrosis (NoNecrosis) are the crucial predictors of MSI-H. To that list, we also add Moderate Differentiation as another important feature as shown in our twelve machine learning models. Throughout 12 models, Random Forest, LightGBM and Naive Bayes Classifier showed the best performance. They achieved 0.793, 0.790 and 0.789

ROC AUC score for each. Instead of logistic regression, our study shows that the future prediction can be done by tree models or Naive Bayes Classifier as they improve the prediction performance overall. Clinicians can then use our feature importance results and select the best model that fits their data for employing MSI-H classification.

DATA AVAILABILITY

The data that support the findings of this study are available from Drs. Steven Gruber and Joe Bonner at City of Hope National Cancer Center. Restrictions apply to the availability of these data, which were used under license for this study. Data are available with the permission of the Gruber et al.

CODE AVAILABILITY

Code is available at `https://github.com/USC-DSCI-552-Spring2021/final-project---codebase-jeayoung114`.

[1] J. K. Greenson, S.-C. Huang, C. Herron, V. Moreno, J. D. Bonner, L. P. Tomsho, O. Ben-Izhak, H. I. Cohen, P. Trougouboff, J. Bejhar, *et al.*, Pathologic predictors of microsatellite instability in colorectal cancer, The American journal of surgical pathology **33**, 126 (2009).

[2] J. O. Ogutu, T. Schulz-Streeck, and H.-P. Piepho, Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions, in *BMC proceedings*, Vol. 6 (Springer, 2012) pp. 1–6.

[3] S. R. Safavian and D. Landgrebe, A survey of decision tree classifier methodology, IEEE transactions on systems, man, and cybernetics **21**, 660 (1991).

[4] T. Chen and C. Guestrin, Xgboost: A scalable tree boosting system, in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (2016) pp. 785–794.

[5] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, Advances in neural information processing systems **30**, 3146 (2017).

[6] C. R. Boland and A. Goel, Microsatellite instability in colorectal cancer, Gastroenterology **138**, 2073 (2010).

[7] J. Jass, W. Atkin, J. Cuzick, H. Bussey, B. Morson, J. Northover, and I. Todd, The grading of rectal cancer: historical perspectives and a multivariate analysis of 447 cases, Histopathology **10**, 437 (1986).