Applying machine learning to understand *Helicobacter pylori* genotype association with gastric cancer

Student ID [redacted]
Name: Vaibhavi Bhardwaj

Introduction

H.pylori selects to reside in the gastric epithelial cells of human beings. For thriving, the bacteria needs to overcome the harsh acidic condition of the stomach. *H.pylori* is classified as a class-I carcinogen. (Wroblewski et al., 2010) have shown the risk of gastric cancer higher in *H.pylori* infected persons and it multiplies in the stomach leading to "gastritis" in humans. Furthermore, the absence of antibacterial treatment for gastritis can culminate in gastric cancer: inflammation, gastric atrophy, intestinal metaplasia and dysplasia. It is found that 70 to 80 percent of gastric complications are caused by *H.pylori*. The main goal of this study is to determine if Random Forest's Feature Importance Measures could be used to identify genomic features of strains associated with gastric cancer development.

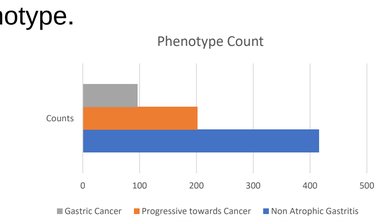
Datasets

Initially, data was collected from Enterobase (https://enterobase.warwick.ac.uk/species/helicobacter/search_strains), transformed and cleaned for consistency of metadata. The collected data represents genomic sequencing of H.pylori in patients diagnosed with gastric disease. Genome assemblies of 714 strains from various countries and continents around the globe with associated Phenotype was downloaded and single nucleotide polymorphisms (SNPs) were called using Snippy (https://github.com/tseemann/snippy) against P12 (Genbank identifier: GCA_000021465.1_ASM2146v1) reference list strain which is ~1.6 megabase pairs (1.6 million characters), leading to 29591 core SNPs per strain with minor allele frequency above or equal to 5%. Associated disease or phenotypes are categorical in nature and are given as 'Non Atrophic Gastritis', 'Progressive towards Cancer' and 'Gastric Cancer'. A table represented below gives an example of how SNPs are found.

Reference Strain	CACAAGCTCAACAA	Progressive towards Cancer
strain X	CTCAAGCTCAAGAA	Non Atrophic Gastritis
strain Y	CACAAGCTCAACAA	Progressive towards Cancer
strain Z	CTCAAGCTCAACAA	Gastric Cancer
SNP	*allele *minor-allele	

The bar graph below represents the counts of the Phenotype.

Phenotype	Counts
Non Atrophic Gastritis	416
Progressive towards Cancer	202
Gastric Cancer	96



Methodology

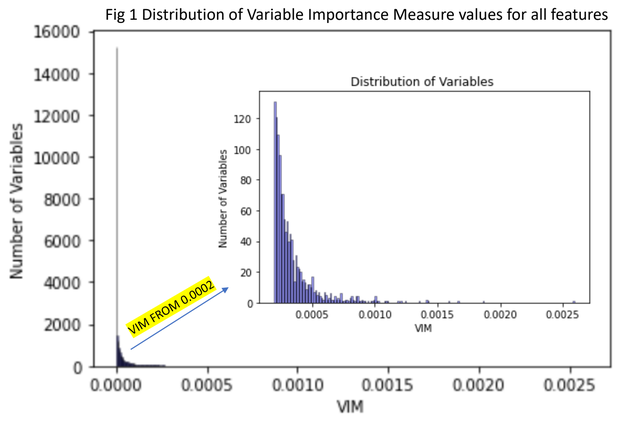
Following steps provide a step-by-step methodology and corresponding results of this project are also provided:

Step 1: Initial Model and Variable selection

Random Forest Classifier from the scikit-learn library was trained with the 75-25 train and test split on the dataset obtained after desired transformations. The random forest was subjected to a large number of features in the dataset which, made it very difficult to get the correct number of features. The initial model showed a test accuracy of 79.4% when all features were taken for prediction.

- **Distribution of Variable Importance Measure values for all features:** VIM of a variable gives the intensity by which a model "utilizes" it for making predictions. Fig 1 shows a distribution of VIM from the base Random Forest. This plot clearly points towards the high count of variables with VIM near 0.0000. Zooming from 0.0002 is also performed for a much clearer distributed view.

Fig 1 Distribution of Variable Importance Measure values for all features



- **Understanding Model Accuracy with respect to VIM or number of features:** Fig 2.a and Fig 2.b shows how accuracy differs when number of features are reduced. For finding the minimum set of features while still maintaining the performance, the lower VIM were removed sequentially and of change in accuracy was recorded at each instance of VIM reduction. Top 1254 variables were selected by this process as the model accuracy was dropping on further variable elimination.

Fig 2.a

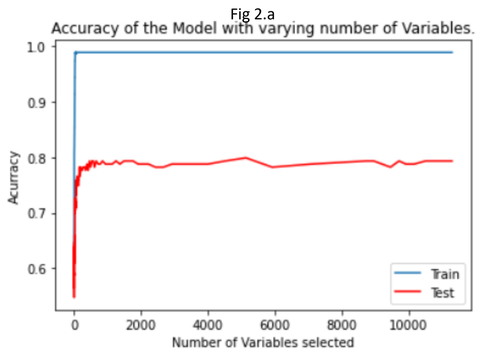
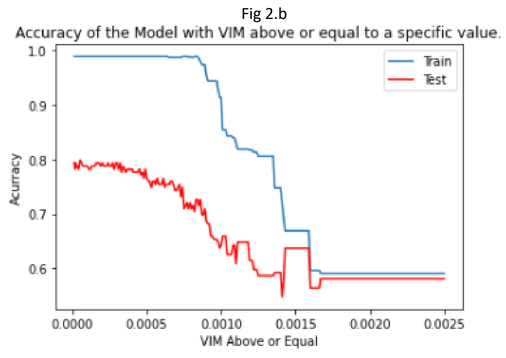


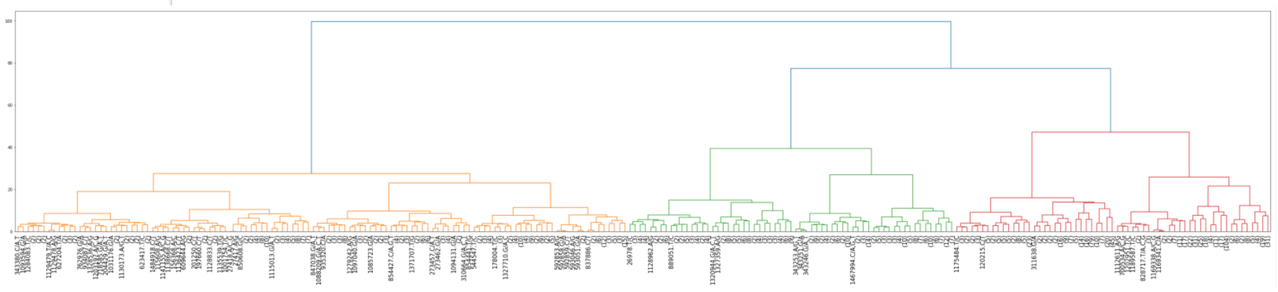
Fig 2.b



Step 2: Correlation Among Selected Features

The next step for us is to take note of any correlation present among the features selected after step 1. Permuting features will not be useful since the same performance can also be achieved by a correlated feature. One way to take note of multicollinearity is to use of dendrogram to explore the correlation linkage. The agglomerative hierarchical clustering is performed on Spearman rank order correlation as shown in Fig 3. When interpreted in terms of Genome Wide Association Studies (GWAS) Fig 3 proves the presence of linkage.

Fig 3: Hierarchical clustering of features



A threshold of 4 with "distance" criterion was picked by visual inspection of dendrogram for grouping features to clusters and further selecting a feature from each cluster. This led to a list of 142 features finally being selected for further modeling and introspection.

Step 3: Re-training on Clustered Features.

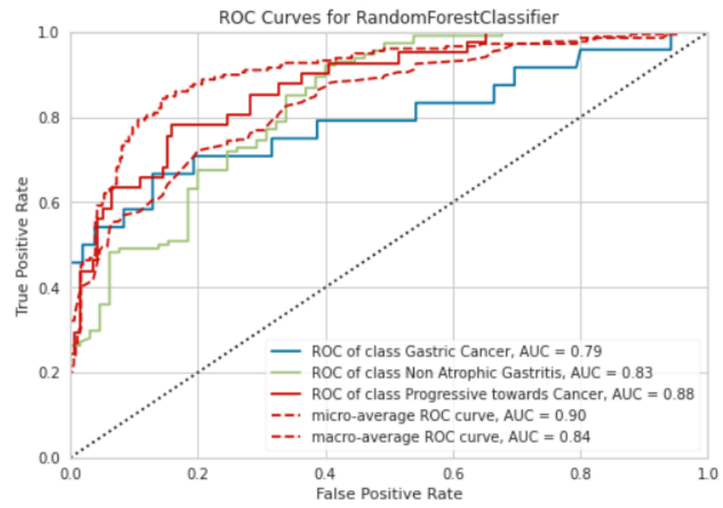
From 29k initial features a reduced feature list of 142 variables was further used for training a new Random Forest Classifier. The reduction of features led to a slight increase in the overall accuracy of the model setting it at 80.44%.

The results presented below give a precise analysis of the model so formed.

	precision	recall	f1-score	support
Non Atrophic Gastritis	1	0.46	0.63	24
Progressive towards Cancer	0.81	0.92	0.86	114
Gastric Cancer	0.74	0.68	0.71	41
accuracy			0.8	179
macro avg	0.85	0.69	0.73	179
weighted avg	0.82	0.8	0.79	179

To further analyze the model ROC curves for the Random Forest Classifier is presented in figure 4. The highest AUC among all the classes is of "Progressive towards Cancer" and the lowest is of "Gastric Cancer".

Fig 4: ROC



Step 4: Comparing Random Forest feature importance to SHAPs

For better understanding a model and it's predictors SHAP (https://github.com/slundberg/shap) is definitely useful. Furthermore, it provides SHAP value for each feature which is based on magnitude of feature attributions. The VIM of features from the new Random Forest were compared towards the SHAP values generated for this model. Top 10 sorted VIM and SHAP features as presented in table 1 and 2 which don't differ much on comparison. Out of top 10 features, 9 are same VIM and SHAP features but the order of importance differs.

Table 1

	Top 10 SHAP Features
0	221331.T/C
1	908373.T/C
2	962858.G/A
3	1009179.C/A,T
4	1181754.C/A,T
5	828717.T/A,C,G
6	310663.A/G,T
7	1115013.G/A,T
8	712033.A/C,G,T
9	1140792.G/A

Table 2

	Top 10 VIM Features
0	712033.A/C,G,T
1	221331.T/C
2	310663.A/G,T
3	828717.T/A,C,G
4	908373.T/C
5	1115013.G/A,T
6	1181754.C/A,T
7	962858.G/A
8	1131910.G/A,C,T
9	1009179.C/A,T

Table 3

Top 10 SHAP Features for detecting Gastric Cancer
221331.T/C
712033.A/C,G,T
119199.C/T
884241.G/A
34261.A/G
332096.C/T
908373.T/C
819317.G/A,C
815966.A/C,T
1320905.T/A,G

Since it was a classification problem, SHAP can clearly segregate features highly responsible for identifying a particular class from others. In Table 3, a list of top 10 SHAP features for detecting gastric cancer are listed. Similarly, this can be done for other classes too.
SHAP also allows one to visualize and interact with a platform to see how a particular feature impacts the classification of each datapoint in our case strain. A snip from the interactive visualization is given in Fig 5.

Fig. 5 : SHAP Interactivity

The following table represents top 17 "Features" from the Random Forest Classifier with "Coding sequence Start chromosomal coordinates", "Coding sequence End chromosomal coordinates", "Gene Identifier", "Product" and "Gene name" sorted on the basis of VIM.

Count	Features	Coding sequence start chromosomal coordinates	Coding sequence end chromosomal coordinates	Gene Identifier	Product	Gene Name
0	712033.A/C,G,T	709540	712278	"HPP12_0668"	"protective surface antigen D15"	
1	221331.T/C	220874	221671	"HPP12_0216"	"cdp-diacylglycerol synthetase"	"cdsA"
2	310663.A/G,T	309196	311679	"HPP12_0294"	"flagellar-hook associated protein 3"	"flgL"
3	828717.T/A,C,G	828445	829518	"HPP12_0779"	"flagellar biosynthetic protein"	"flhB-1"
4	908373.T/C	908318	909700	"HPP12_0858"	"ADP-heptose synthase"	"waaE"
5	1115013.G/A,T	1114613	1115281	"HPP12_1044"	"thiol:disulfide interchange protein"	
6	1181754.C/A,T	1181129	1182427	"HPP12_1109"	"hypothetical protein"	
7	962858.G/A	962720	964273	"HPP12_0910"	"outer membrane protein HopC/AlpA"	"alpA"
8	1131910.G/A,C,T	1131284	1132816	"HPP12_1061"	"outer membrane protein"	
9	1009179.C/A,T	1008993	1009646	"HPP12_0948"	"integral membrane protein"	
10	699442.T/A,C,G	699208	700887	"HPP12_0658"	"soluble lytic murein transglycosylase"	
11	322164.G/A,C	320919	322208	"HPP12_0305"	"glutamate-1-semialdehyde 2,1-aminomutase"	"hemL"
12	979096.G/A,T	976759	983874	"HPP12_0919"	"vacuolating cytotoxin VacA-like protein"	

Note:
Genes so identified are involved in outer membrane lipopolysaccharide biosynthesis, flagella structure and motility, peptidoglycan biosynthesis, and proteins involved in adhesion to the gastric epithelium.
VacA is a well known to play a very important role in H. pylori induced gastritis - VacA is able to enter gastric epithelial cells and induces cell death.

Conclusion and Further Directions

The current project has the potential for further research and requires laboratory testing, which will help confirm the findings. Furthermore, since the results cannot be supported by tangible ground truth, further study is required in the area of simulated data from BacGWASim (https://github.com/Morteza-M-Saber/BacGWASim/tree/master/BacGWASim). The features identified have to be refined by various other statistical methods to reduce and highlight the list of potential features. Since the dataset is imbalanced, modeling with increase recall/precision for minority classes is one of the future priorities of this project.

Referencing

- *Wroblewski, L. E., Peek Jr, R. M., & Wilson, K. T. (2010). Helicobacter pylori and gastric cancer: factors that modulate disease risk. Clinical microbiology reviews, 23(4), 713.*