Prediction of Cancer Risk using DNA Methylation Imbalance Data

Ronak Sumbaly

Department of Computer Science, University of California, Los Angeles
rsumbaly@ucla.edu

1 Abstract

Motivation: Cancer is the most common human genetic disease and the problem of classifying subjects/ patients into disease categories (tumor/normal) is of common occurrence in medical research. Over the past years rapid advances in next-generation sequencing technology has led to the timely advent of The Cancer Genome Atlas (TCGA) project which provides the most comprehensive genomic data for various kinds of cancer. The project publishes both clinical (recurrence, survival & treatment resistance) and molecular profiles (Exon (mRNA) expression, DNA methylation, Copy Number Variations (CNV) & Single Nucleotide Polymorphism (SNP)) for both tumor samples and normal controls. Data from TCGA has improved the ability of oncologists to diagnose, treat, and prevent cancer through a better understanding of the genetic basis of this disease. Previous research indicates classification of patient samples using TCGA Exon expression or SNP datasets. However, recent study show that DNA methylation act as better bio-markers and help in improving the dichotomous outcome (tumor/normal) based on several features. At the same time TCGA data is faced with the problem of class imbalance and high dimensionality of data leading to an increase in the false negative rate. The project aims to tackle these problem and simultaneously lead to the development of a new approach for prognosis of different kinds of cancer using DNA methylation data.

Approach: The project uses the Synthetic Minority Oversampling Technique (SMOTE) algorithm in the pre-processing phase as a method to maintain a balanced class distribution. SMOTE is combined with the T-Link under-sampling technique for data cleaning in order to remove noise. To reduce the feature space of the data only those genes for which mutations have been causally implicated in cancer are considered, these are obtained through resources like The Cancer Gene Census (COSMIC) and Clinical Interpretation of Variants in Cancer (CIVic). Classification of patient samples is then performed utilizing several machine learning algorithms of Logistic Regression, Random Forest and Gaussian Naive Bayes. Each classifier performance is evaluated using appropriate performance measures. The methodology is applied on the TCGA DNA Methylation data for 7 various cancer types.

Results: The SMOTE in combination with T-Link and Random Forest algorithm demonstrated a superior performance in classification of patient samples.

Availability: Existing development and source code in Python is available for contribution and for download by public from GitHub (www.github.com/RonakSumbaly/Cancer-Risk-Prediction).

2 Introduction

Cancer is a disease of genome alterations: DNA sequence changes, copy number variations, chromosomal rearrangements and modification in DNA methylation together drive the development and progression of human malignancies. Accurate cancer prognosis in patients is crucial for timely treatment. Prediction models have been built over the years that provide an important approach to assess the risk and prognosis by identifying individuals at high risk, aiding the design and planning of clinical trials, and enabling better living. As the number and advancement of cancer risk prediction models have grown, so too has the interest in guaranteeing that they are correct applied, modeled and assessed.

The number of models has grown steadily ever since the first risk prediction model constructed for a chronic disease [11]. In recent years, cancer risk prediction models have combined clinical and epidemiological risk factors with new biologic and genetic data to more accurately assess cancer risk. DNA methylation (DM) has been one of the most intensely studied epigenetic modification in humans. There has been large evidence that DM patterns [5] in 'tumor' tissues are aberrant as compared to 'normal' tissues, and hence been associated with a large number of human diseases.

The Cancer Genome Atlas (TCGA) Research Network [18] has profiled and analyzed large numbers of human tumors to discover molecular aberrations at the DNA, RNA, protein and epigenetic levels. The resulting rich data provide a major opportunity to develop classification models based on various tumor lineages. Like most data in the real-world, TCGA data is imbalance in nature. Classification on these highly imbalanced data are affected by the majority calls leading to an increase in false negative rate. Several methods have been proposed to adjust the standard classification process in the presence of imbalance class problem. These methods can be applied both at an algorithmic-level and well as a data level [17, 19]. Studies have shown advanced sampling methods [13] to reduce the chance of information loss and therefore improve the performance of the classification algorithm.

Researchers have previously proposed molecular classification schemes for tumors based on genetic features such as mutations, copy-number variations, gene expression, or combinations of these features. Several studies have already shown that random forest models based on gene expression profiles can be used for successful breast cancer subtype classification [12, 14]. Separate classification of tumor types based on DNA methylation has been studied [20] and comparison of the same is done with the projects' methodology towards the end.

The project focuses on usage of one advanced sampling techniques on medical TCGA DM imbalance data of all tumor types combination and not subclassifications with the goal to improve the classification models sensitivity toward the minority class (normal tissue).

3 Methodology

This project can be divided into three phases as shown in Figure 1.

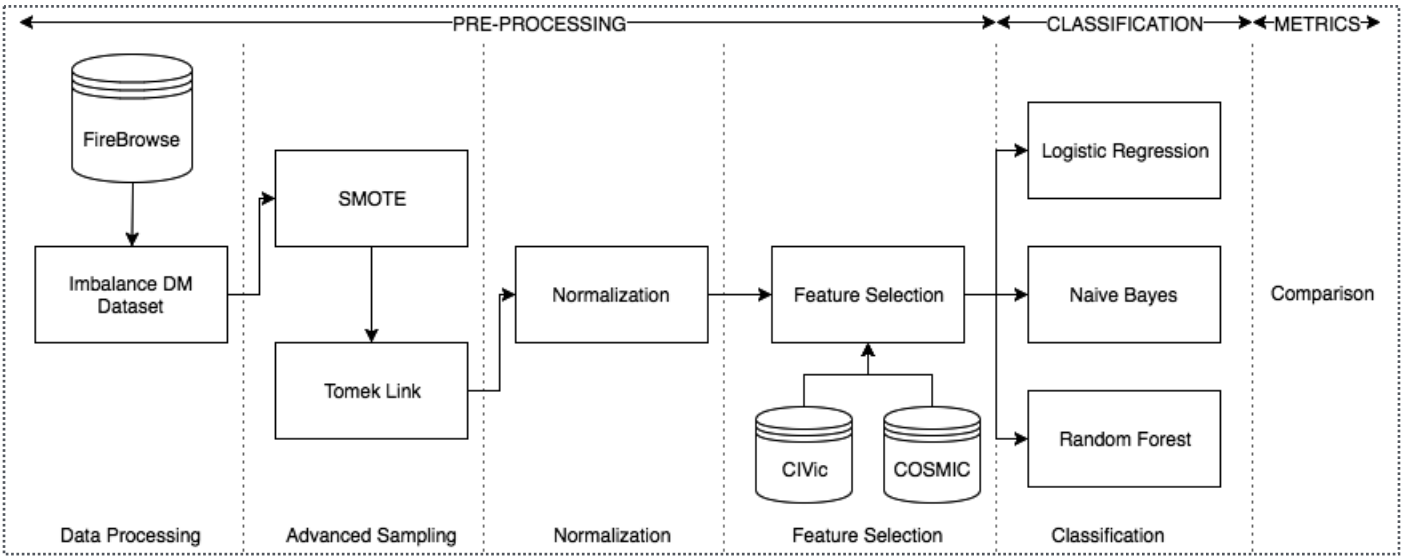

Figure 1. Research Methodology

3.1 Pre-processing

This phase is divided into sub phases as follows:

3.1.1 Data Processing

The TCGA project publishes molecular profile data in the form of DNA methylation data for over 28 cancer types. The raw data $(0 \leq x \leq 1)$ is available online through the TCGA data portal and is mapped to specific position or range on the human chromosome (e.g. chr19:19033575 indicates the position 19033575 on Chromosome 19). Since the raw data from TCGA was complex in nature to process, the project uses a pre-processed version which is made available through Broad Institute's FireBrowse.

Unlike TCGA raw data FireBrowse [6] maps values to a specific human gene [10] which is annotated based on the HGNC nomenclature. The mapping makes it easier for feature selection as discussed in the following section. Each sample file is annotated with a TCGA identifier value [16] which indicates whether the sample is a tumor or a normal case (e.g. $TCGA - 2F - A9KW - \mathbf{01}$ **:** Tumor type ranges from $01 - 09$ (Class 1) and normal type ranges from $10 - 19$ (Class 0). In this project seven tumor typologies are considered and the statistics for each can be seen in Table 1.

Table 1. Statistics of TCGA DM Data for seven cancer types used in the project

Tumor Type	Abbrev.	# Patients	Tumor - 1	Normal - 0
Breast Invasive Carcinoma	BRAC	885	788	97
Lung Adenocarcinoma	LUAD	492	460	32
Urothelial Bladder Carcinoma	BLAC	434	413	21
Prostate Adenocarcinoma	PRAD	549	499	50
Lung Squamous Cell Carcinoma	LUSC	412	370	42
Thyroid Cancer	THCA	567	511	56
Head-Neck Squamous Cell Carcinoma	HNSC	580	530	59

3.1.2 Feature Selection

TCGA DNA Methylation Data for all the various cancer types contains > 20000 protein-coding genes as their feature variables. Feature selection is exceptionally important in this case, though univariate feature selection on the entire set of human genes can make mistakes in terms of incorporating noise and unrelated features. Hence instead of following the traditional method of feature selection the project looks towards only those genes that have been biologically been identified as having implication in cancer mutation.

These genes are obtained using the union of resources - The Cancer Gene Census (COSMIC) [7] and the Clinical Interpretation of Variants in Cancer (CIVic) [8]. These resources have maintained an updated list of experimentally found genes implicating various tumors. Considering only the genes that are obtained from these catalogues gives a reduced set of 624 **'onco -genes'** which act as a finer set of features.

3.1.3 Sampling

Table 1 indicates that the data is imbalance in nature. This is due to the distribution of the target class which is not uniform for the normal cases. Applying classification on this type of data was found to be very challenging as all the classifiers indicated in the following section were modeled to show bias to the majority class which overall would give a very high accuracy but very low sensitivity towards the normal cases (Class 0). Therefore the goal was to build a model that would maximize the sensitivity of the minority class but at the same time give good accuracy. Several works have been done in this field. The project employs two of the advanced sampling methods

1. Synthetic Minority Oversampling Technique (SMOTE)

SMOTE is an advanced oversampling method developed by Chawala [2]. The main objective of the SMOTE algorithm is to create artificial examples of the minority class rather than just replicating the existing examples which most of the oversampling algorithm does. This avoids the problem of overfitting and beneficial in giving a uniform distribution between the different levels of the target class.
The algorithm works sampling by taking each observation from the minority class and finding its k-nearest neighbors, out of which a few neighbors are randomly selected based on the rate of oversampling which then generates artificial observations and spreads it along the line join the observation and the neighbors selected. Several modification have been made to the SMOTE algorithm [1] but this project only employs the original algorithm to balance the distribution of the target class.

2. Tomek Link (T-Link)

Since the number of tumor cases $>>$ number of normal cases, there is a probability that by applying the SMOTE oversampling algorithm we might encounter noise. In order to counter-act this issue the project further applies an under sampling algorithm - Tomek Link (T-Link) [17]. The T-link algorithm works on the concept that two points that are each others closest neighbors but share different class labels are noise in the data. The T-link under sampling algorithm is applied right after applying SMOTE [15] in order to remove any noise created by the latter.

3.2 Classification

The project at its heart is a classification problem for which most of the input features x are continuous-valued random variables and $y \sim Bernoulli(\phi)$. The project employs three different supervised learning methods -

1. Gaussian Naive Bayes (GNB)

Since the data deals with continuous values as the feature space with each belonging to class $k \in \{0, 1\}$, it is assumed that they are distributed according to a Gaussian distribution [9] given as,

$$p(x_j = x|y_j = k) = \frac{1}{\sqrt{2\pi\sigma_{jk}^2}} exp\Big[-\frac{1}{2}\Big(\frac{x - \mu_{jk}}{\sigma_j k}\Big)^2\Big]$$

The model is trained and tested on this assumption.

2. Logistic Regression (LR)

LR [4] finds the best fitting model to describe the dichotomous characteristic of the classification problem and the set of feature variables. Formally, the model logistic regression models is,

$$log\frac{p(x)}{1 - p(x)} = \beta_0 + x.\beta$$

3. Ensemble Random Forest (RF)

Random Forest classifier [3] was trained with $K = 250$ decision trees constructed by random sampling over the reduced feature space. The prediction on a new sample is done by majority voting over every decision tree constructed.

3.3 Comparison

The comparison phase involves the comparison of the models recall / sensitivity towards the minority class using different machine learning algorithms as indicated in the above section. The model performance was also evaluated using measures such as precision and F-1 score. Following is a description of each of the measure -

- **Sensitivity / Recall** - Commonly referred to as the True Positive Rate. It measures the proportion of positives that are correctly identified

$$Sensitivity = \frac{TruePositive}{TruePositive + FalseNegative}$$

- **Precision** - Positive predictive value

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

- **F-1 score** - Measure of test's accuracy.

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

The project also considers the comparison of the area under the ROC Curve (AUC). The area measures discrimination, that is, the ability of the test to correctly classify those with and without the tumor.

---

4 Discussion

TCGA Methylation data of seven cancer typologies was downloaded from FireBrowse. The molecular profile data was parsed to retrieve sample types $y \in \{0, 1\}$. In order to reduce the feature space of the data only the genes present in The Cancer Gene Census and CIVic catalogue were considered while rest of the genes were removed from the final matrix. The dataset was split into training and testing by considering a 70/30 split. Figure 2 shows the PCA 2D plots of the training data which indicates the imbalance distribution within the target variable.

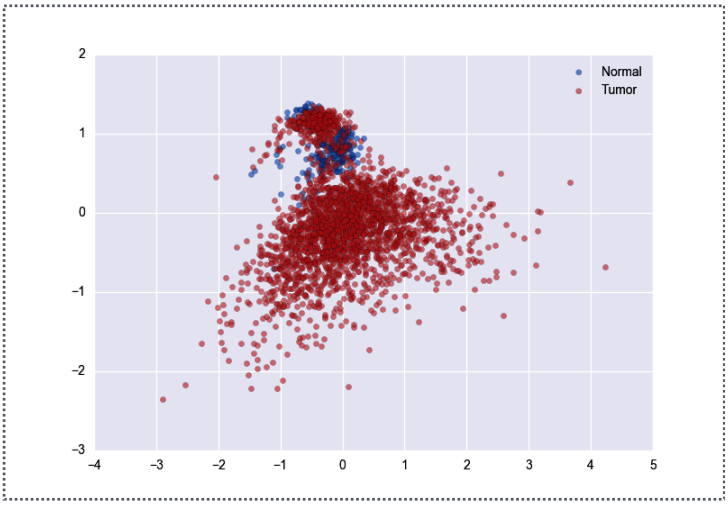

Figure 2. Class distribution before sampling

To demonstrate the issues of imbalance data the classification methods were initially tested and the evaluation metrics calculated were as follows.

Table 2 - Evaluation Metrics before Sampling and Normalization

Classification Methods	Sensitivity / Recall		Precision		F-1 Score	
	0	1	0	1	0	1
Logistic Regression	**0.79**	0.99	0.89	0.98	0.84	0.99
Gaussian Naive Bayes	**0.49**	0.83	0.35	0.99	0.51	0.90
Ensemble Random Forest	**0.77**	1.00	0.87	0.98	0.89	0.99

Table 2 shows that all the three classification methods have a high overall accuracy but the sensitivity of the Class 0 is terrible which is attributed to the imbalance within the data. The ROC curves for the classifiers are presented below.

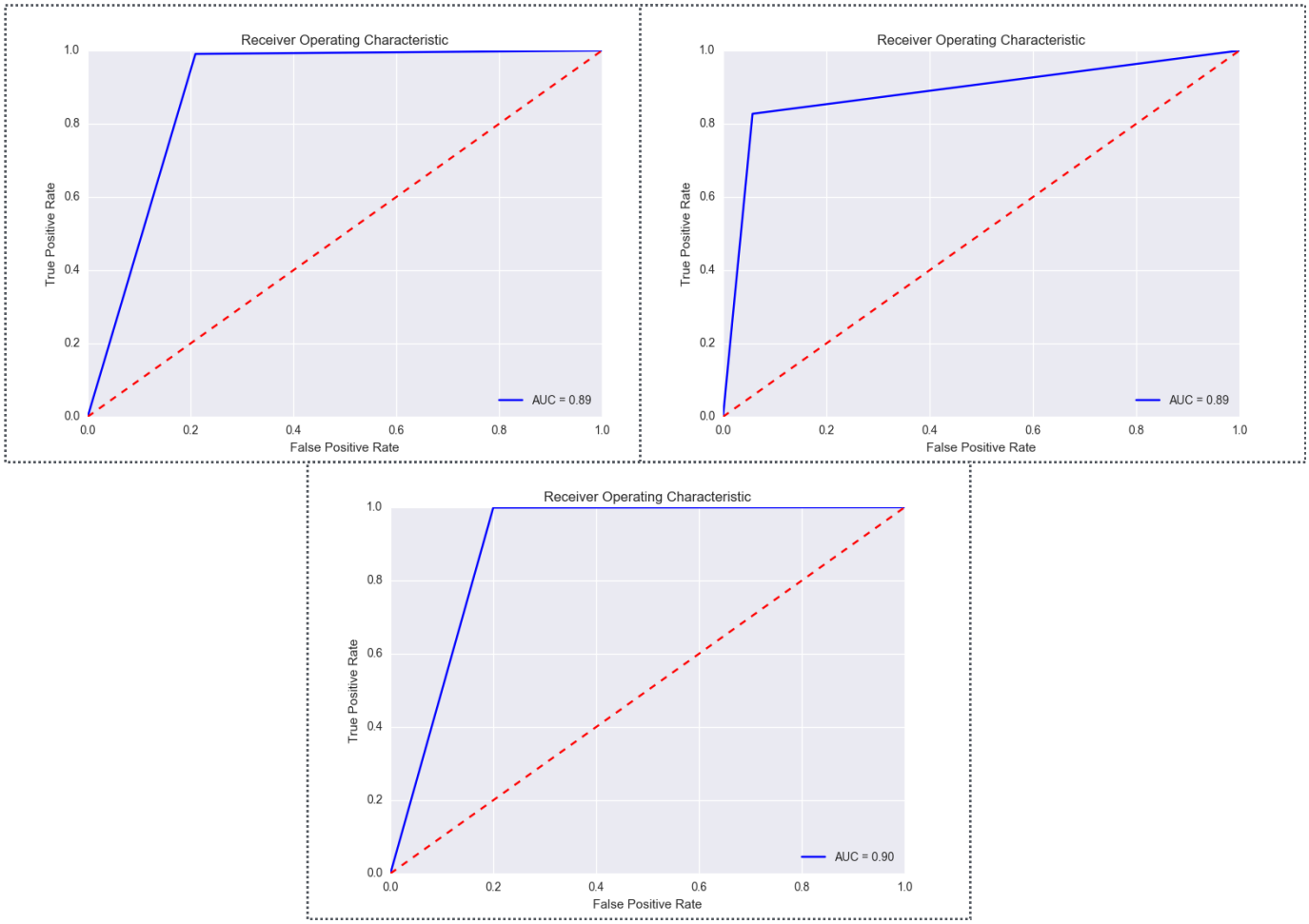

Figure 3. ROC Curves - LR : 0.89 (Top Left) | GNB : 0.89 (Top Right)| RF : 0.90 (Bottom)

In order to overcome this issue the SMOTE and the T-Link algorithms were applied which resulted in a balance within the target class distribution. This can be seen in the PCA 2D plots of the sampled data in Figure 4.
The data was normalized and the classification methods were applied again on the sampled data. The resulting metric values are presented in Table 3.

Table 3. Evaluation Metrics after Sampling and Normalization

Classification Methods	Sensitivity / Recall		Precision		F-1 Score	
	0	1	0	1	0	1
Logistic Regression	0.93	0.98	0.59	1.00	0.74	0.96
Gaussian Naive Bayes	0.90	0.89	0.45	0.99	0.60	0.94
Ensemble Random Forest	**0.92**	**0.98**	0.84	0.99	0.88	0.99

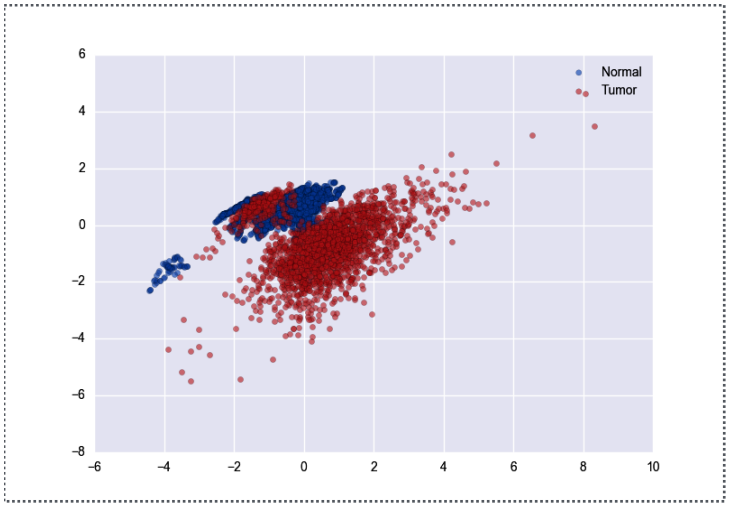

Figure 4. Class distribution after sampling and normalization

It can be clearly seen from Table 3 that feature selection along with SMOTE/T-Link sampling algorithms provided better prediction statistics indicating that prior biomedical knowledge and sampling algorithms are essential to build better classifiers. Comparing area under the curve values of the ROC curve between Figure 3 and Figure 5 we can clearly see that proposed methodology has better determination values for classification.

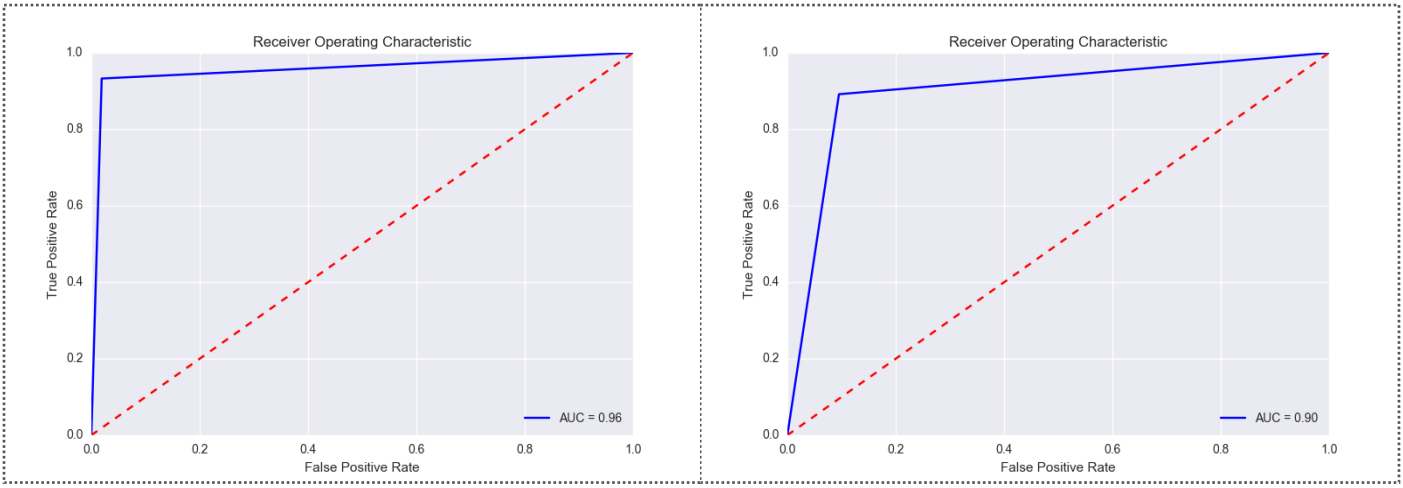

Figure 5. ROC Curves - LR : 0.96 (Left) | GNB : 0.90 (Right)| RF : 0.95 (Below)

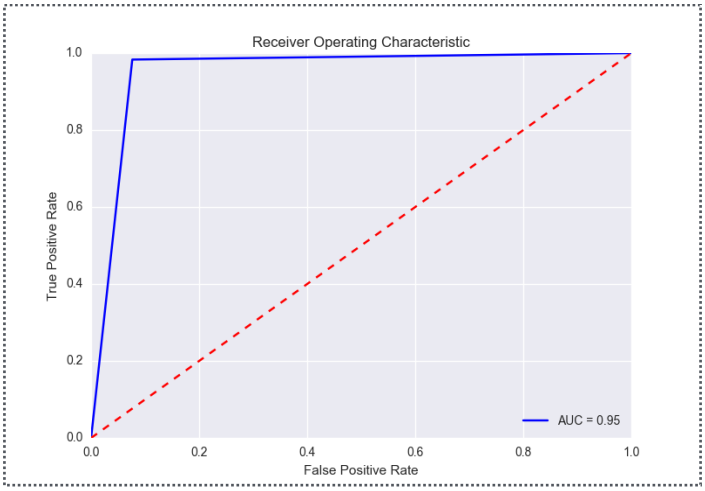

Figure 5. ROC Curves - LR : 0.96 (Top Left) | GNB : 0.90 (Top Right)| RF : 0.95 (Bottom)

As previously indicated prior work has been done in the field of classifying tumors in a dichotomous manner. To provide a basis of comparison the project considers [20] written by Maulik Kamdar. The paper considers classification of each tumor type separately performing feature selection using L1- based feature selection and adding patient demographics (1711 features) and a 4-fold cross validation to train the model. In order to facilitate a comparison between the two methodology a single tumor type **'Lung Adenocarcinoma - LUAD'** was considered and classification was performed. Table 4 shows a comparison between the sensitivity of the two models. Even though the projects methodology gives a slightly better sensitivity further analysis would be required since the simulation grounds might not be the same.

Table 4 Comparison of Sensitivity

Classification Algoritm	Maulik Kamdar Paper	Proposed Methodology
Support Vector Machine	0.96	-
Decision Tree	0.91	-
Random Forest	**1.00**	**1.00**
Logistic Regression	-	1.00
Gaussian Naive Bayes	**0.76**	**0.90**

5 Conclusion and Future Work

This project revolves around building classifiers using imbalance TCGA DM data and comparing the classification results. The results show that the SMOTE in combination with T-Link and Random Forest algorithm demonstrated a superior performance in classification of patient samples. The project contributes to the field by developing a new workflow to tackle imbalance data along with incorporating biomedical knowledge to classify samples as tumorous or normal. The major strength of the workflow is that it strengthens the assumption of using genes that have been statistically found to be implicating in cancer mutations. While there is still work needed in perfecting the model, looking into to other molecular profile data such as SNPs and CNVs in combination with samples clinical data can help extend the diagnostic framework to generate better prognosis. Similarly using advanced classification techniques of neural networks and deep learning methods can help improve the system. Overall the project goal was to build custom classifiers based on prior knowledge and demonstrate how TCGA data can be used for supervised learning.

Acknowledgement

I would like to express my deep gratitude and thank Prof. Sriram Sankararaman for his support and assistance in this project and for providing me with the necessary guidance throughout the quarter.

References

[1] Blagus, Rok and Lara Lusa. "SMOTE For High-Dimensional Class-Imbalanced Data". BMC Bioinformatics 14.1 (2013): 106.

[2] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 16: 321-357.

[3] Breiman, L. Machine Learning (2001) 45: 5. doi:10.1023/A:1010933404324

[4] Cucchiara, Andrew, D. Hosmer, and S. Lemeshow. "Applied Logistic Regression". Technometrics 34.3 (1992): 358.

[5] Das, P.M., Singal, R.: "DNA methylation and cancer". Journal of Clinical Oncology 22(22), 4632–4642 (2004).

[6] "Firebrowse". Firebrowse.org.

[7] Forbes, S. A. et al. "COSMIC: Exploring The World's Knowledge Of Somatic Mutations In Human Cancer". Nucleic Acids Research 43.D1 (2014): D805-D811.

[8] Griffith M, Spies NC, et al. CIViC: A knowledge base for expert-crowdsourcing the clinical interpretation of variants in cancer. (2016) ; doi: http://dx.doi.org/10.1101/072892

[9] Hastie, Trevor, Robert Tibshirani, and J Jerome H Friedman. The Elements of Statistical Learning. Vol.1. N.p., Springer New York, 2001.

[10] "HGNC Database Of Human Gene Names | HUGO Gene Nomenclature Committee". genenames.org.

[11] Kannel WB, McGee D, Gordon T. A general cardiovascular risk profile: the Framingham Study. Am J Cardiol 1976;38:46–51.

[12] M. B. Kursa. Robustness of Random Forest-based gene selection methods. BMC Bioinformatics, 15(1):8, 2014.

[13] McCarthy K, Zabar B, Weiss GM (2005) Does Cost-Sensitive Learning Beat Sampling for Classifying Rare Classes?. Proc. Int'l Workshop Utility-Based Data Mining, pp: 69-77.

[14] R. Dıaz-Uriarte and S. Alvarez de Andres. Gene selection and classification of microarray data using random forest. BMC Bioinformatics, 7:3, 2006.

[15] T, Elhassan and Aljurf M. "Classification Of Imbalance Data Using Tomek Link(T-Link) Combined With Random Under-Sampling (RUS) As A Data Reduction Method". Journal of Informatics and Data Mining 1.2 (2016)

[16] "TCGA Barcode - TCGA - National Cancer Institute - Confluence Wiki". wiki.nci.nih.gov.

[17] Tomek Ivan. "An Experiment With The Edited Nearest-Neighbor Rule". IEEE Transactions on Systems, Man, and Cybernetics 6.6 (1976): 448-452.

[18] "The Cancer Genome Atlas - National Cancer Institute". https://tcga-data.nci.nih.gov

[19] Van Hulse, Jason and Taghi Khoshgoftaar. "Knowledge Discovery From Imbalanced And Noisy Data". Data & Knowledge Engineering 68.12 (2009): 1513-1542. Web.

[20] Maulik Kamdar. "Visualizing Personalized Caner Risk Prediction." Biomedical Informatics Training Program, Stanford University