**DataCan: Data Analysis and Prediction**

**for Cancer Dataset**

**A Minor Project Report**

**Submitted in partial fulfillment of the**

**requirement for the Award of the Degree**

**of**

**BACHELORS OF TECHNOLOGY (B. Tech)**

In

Information Technology

By

**Tushar Ranjan Varun Goel Vishal Jangir**

**159102147 159102156 159103073**

Under the guidance of

**Mr. Venkatesh Gauri Shankar**



**School of Computing and Information Technology**

**MANIPAL UNIVERSITY JAIPUR**

**JAIPUR -303007**

**RAJASTHAN, INDIA**

**May/2018**

**CERTIFICATE**

3rd May 2018

This is to certify that the project titled **DataCan: Data Analysis and Prediction for Cancer Dataset** is a record of the bonafide work done by **VARUN GOEL** **(159102156)** submitted in partial fulfilment of the requirements for the award of the Degree of Bachelor of Technology (B. Tech) in **Information Technology** of Manipal University Jaipur, during the academic year 2017-18.

**Mr. Venkatesh Gauri Shankar**

*Project Guide, Dept of Information Technology*

*Manipal University Jaipur*

**Dr. Sumit Srivastava**

*HOD, Dept of Dept of Information Technology*

*Manipal University Jaipur*

**CERTIFICATE**

3rd May 2018

This is to certify that the project titled **DataCan: Data Analysis and Prediction for Cancer Dataset** is a record of the bonafide work done by **VISHAL JANGIR (159103073)** submitted in partial fulfilment of the requirements for the award of the Degree of Bachelor of Technology (B. Tech) in **Information Technology** of Manipal University Jaipur, during the academic year 2017-18.

**Mr. Venkatesh Gauri Shankar**

*Project Guide, Dept of Information Technology*

*Manipal University Jaipur*

**Dr. Sumit Srivastava**

*HOD, Dept of Dept of Information Technology*

*Manipal University Jaipur*

**CERTIFICATE**

3rd May 2018

This is to certify that the project titled **DataCan: Data Analysis and Prediction for Cancer Dataset** is a record of the bonafide work done by **TUSHAR RANJAN (159102147)** submitted in partial fulfilment of the requirements for the award of the Degree of Bachelor of Technology (B. Tech) in **Information Technology** of Manipal University Jaipur, during the academic year 2017-18.

**Mr. Venkatesh Gauri Shankar**

*Project Guide, Dept of Information Technology*

*Manipal University Jaipur*

**Dr. Sumit Srivastava**

*HOD, Dept of Dept of Information Technology*

*Manipal University Jaipur*

**ABSTRACT**

While tracing the history of the past two decades, it can be seen that high-throughput biology has triggered the active statistical research in high-dimensional data. Recently, to handle the real dimension of genomic data, which is usually beyond the scale of hundreds of variables, methods to handle ultrahigh-dimensional data are emerging. The diversity of cancer data types together with the availability of related studies on similar types of cancers adds another two dimensionalities of complexity. It is of critical clinical and biological interest to understand what subtypes a cancer has, how genomic profiles and survival rates of patients vary among subtypes, whether a patient’s survival can be predicted from his or her genomic profiles, and how one type of genomic profile is correlated with another type of genomic profile.

Once sequenced, a cancer tumor can have thousands of genetic mutations. But the challenge is distinguishing the mutations that contribute to tumor growth (drivers) from the neutral mutations (passengers). Currently this interpretation of genetic mutations is being done manually. This is a highly time-consuming task where a clinical pathologist has to manually review and classify every single genetic mutation based on evidence from text-based clinical literature.

Identiﬁcation of cancer subtypes plays an important role in revealing useful insights into disease pathogenesis and advancing personalized therapy. The recent development of high-throughput sequencing technologies has enabled the rapid collection of multi-platform genomic data (e.g., gene expression, miRNA expression, and DNA methylation) for the same set of tumor samples. Although numerous integrative clustering approaches have been developed to analyze cancer data, few of them are particularly designed to exploit both deep intrinsic statistical properties of each input modality and complex cross-modality correlations among multi-platform input data.

Our project focuses on finding the cancer-causing genes and their specific mutations and classifying the genes on the 9 classes of cancer. This will help in predicting which genetic mutation causes which type of cancer. We have used Sci-kit Learn and NLTK for our project to analyse what every class means by classifying every genetic mutation into 17 major mutation types which we analyzed from the dataset and make predictions using Random Forest Classifier. Text based data was also used as features for the model. GridSearchCV is used to get the best parameters for the Random Forest Classifier Model. 10-fold Cross Validation is applied to get the best accuracy score of the trained Machine Learning model.

i

**List of Figures**

|  |  |  |
| --- | --- | --- |
| **Figure No.** | **Figure Title** | **Page No.** |
| 3.1.1 | Mutation Type Frequency in each Class | 6 |
| 3.2.1 | Process Flow Chart | 7 |
| 3.2.2 | Random Forest Model Structure | 7 |
| 4.1.1 | Distribution of Genetic Mutation Classes | 8 |
| 4.1.2 | Frequency of Mutation Types | 8 |
| 4.1.3 | Frequency of maximum occurring Genes in each Class | 9 |
| 4.1.4 | Number of Words by Class | 10 |
| 4.1.5 | Confusion Matrix (a)Normalized (b)Without Normalization | 10 |
| 5.1.1 | Timeline Chart | 11 |
| 5.1.2 | Month-Wise Plan of Work | 11 |

ii

|  |  |  |  |
| --- | --- | --- | --- |
| **Contents** | | | **Page No.** |
| Abstract | | | i |
| List Of Figures | | | ii |
| **Chapter 1** | | **INTRODUCTION** |  |
|  | **1.1** | Motivation | 1 |
| **1.2** | Project Statement | 2 |
| **Chapter 2** | | **BACKGROUND OVERVIEW** |  |
|  | **2.1** | Conceptual Overview | 3 |
| **2.2** | Technologies Involved | 4 |
| **Chapter 3** | | **METHODOLOGY** |  |
|  | **3.1** | Detailed methodology | 5 |
| **3.2** | Architecture Diagram | 7 |
| **Chapter 4** | | **IMPLEMENTATION AND RESULTS** |  |
|  | **4.1** | Modules | 8 |
| **Chapter 5** | | **FUTURE WORK AND CONCLUSION** |  |
|  | **5.1** | Progress Chart/Time Line Chart | 11 |
| **5.2** | Future Work and Conclusion | 12 |
| **REFERENCES** | | | 13 |

**1. INTRODUCTION**

***1.1 Motivation***

Over the past decades, a continuous evolution related to cancer research has been performed. Scientists applied different methods, such as screening in early stage, in order to find types of cancer before they cause symptoms. Moreover, they have developed new strategies for the early prediction of cancer treatment outcome. With the advent of new technologies in the field of medicine, large amounts of cancer data have been collected and are available to the medical research community. However, the accurate prediction of a disease outcome is one of the most interesting and challenging tasks for physicians. As a result, Machine Learning methods have become a popular tool for medical researchers. These techniques can discover and identify patterns and relationships between them, from complex datasets, while they are able to effectively predict future outcomes of a cancer type. We attempt to aid the professionals in the medical field by describing the different stages of analysis of cancer and discussing our approach, as well as by providing practical advice on how to access and use resources, and how to implement recommendations regarding cancer issues and their related problems.

Once sequenced, a cancer tumor can have thousands of genetic mutations. But the challenge is in distinguishing the mutations that contribute to tumor growth from the neutral mutations (passengers). Currently this interpretation of genetic mutations is being done manually. This is a very time-consuming task where a clinical pathologist has to manually review and classify every single genetic mutation based on evidence from text-based clinical literature. I will be using seaborn and Deep Learning tools to study and visualize the structures in the data. We have been challenged to automatically classify genetic mutations that contribute to cancer tumor growth (also called “drivers”) in the presence of mutations that are don’t affect the tumors (passengers).

The fundamental goals of cancer prediction and prognosis are distinct from the goals of cancer detection and diagnosis. In cancer prediction/prognosis one is concerned with three predictive foci: 1) the prediction of cancer susceptibility (i.e. risk assessment); 2) the prediction of cancer recurrence and 3) the prediction of cancer survivability. In the first case, one is trying to predict the likelihood of developing a type of cancer prior to the occurrence of the disease. In the second case one is trying to predict the likelihood of redeveloping cancer after to the apparent resolution of the disease. In the third case one is trying to predict an outcome (life expectancy, survivability, progression, tumor-drug sensitivity) after the diagnosis of the disease. In the latter two situations the success of the prognostic prediction is obviously dependent, in part, on the success or quality of the diagnosis. However, a disease prognosis can only come after a medical diagnosis and a prognostic prediction must take into account more than just a simple diagnosis.

1

***1.2 Project Statement***

* While tracing the history of the past two decades, it can be seen that high-throughput biology has triggered the active statistical research in high-dimensional data.
* Recently, to handle the real dimension of genomic data, which is usually beyond the scale of hundreds of variables, methods to handle ultrahigh-dimensional data are emerging.
* The diversity of cancer data types together with the availability of related studies on similar types of cancers adds another two dimensionalities of complexity.
* It is of critical clinical and biological interest to understand what subtypes a cancer has, how genomic profiles and survival rates of patients vary among subtypes, whether a patient’s survival can be predicted from his or her genomic profiles, and how one type of genomic profile is correlated with another type of genomic profile.
* Once sequenced, a cancer tumor can have thousands of genetic mutations. But the challenge is distinguishing the mutations that contribute to tumor growth (drivers) from the neutral mutations (passengers).
* Currently this interpretation of genetic mutations is being done manually. This is a highly time-consuming task where a clinical pathologist has to manually review and classify every single genetic mutation based on evidence from text-based clinical literature.

2

**2. BACKGROUND OVERVIEW**

***2.1 Conceptual Overview***

Before beginning with a detailed analysis of what machine learning methods work best for which kinds of situations, it is important to have a good understanding of what machine learning is – and what it isn’t. Machine learning is a branch of artificial intelligence research that employs a variety of statistical, probabilistic and optimization tools to “learn” from past examples and to then use that prior training to classify new data, identify new patterns or predict novel trends.

Machine learning, like statistics, is used to analyze and interpret data. Unlike statistics, though, machine learning methods can employ Boolean logic (AND, OR, NOT), absolute conditionality (IF, THEN, ELSE), conditional probabilities (the probability of X given Y) and unconventional optimization strategies to model data or classify patterns. These latter methods actually resemble the approaches humans typically use to learn and classify. Machine learning still draws heavily from statistics and probability, but it is fundamentally more powerful because it allows inferences or decisions to be made that could not otherwise be made using conventional statistical methodologies. For instance, many statistical methods are based on multivariate regression or correlation analysis. While generally very powerful, these approaches assume that the variables are independent and that data can be modeled using linear combinations of these variables. When the relationships are nonlinear and the variables are interdependent (or conditionally dependent) conventional statistics usually flounders. It is in these situations where machine learning tends to shine. Many biological systems are fundamentally nonlinear and their parameters conditionally dependent. Many simple physical systems are linear and their parameters are essentially independent.

Success in machine learning is not always guaranteed. As with any method, a good understanding of the problem and an appreciation of the limitations of the data is important. So too is an understanding of the assumptions and limitations of the algorithms being applied. If a machine learning experiment is properly designed, the learners correctly implemented and the results robustly validated, then one usually has a good chance at success. Obviously if the data is of poor quality, the result will be of poor quality (garbage in = garbage out). Likewise, if there are more variables than events to predict then it is also possible to create a series of redundant learners. This is a set of learning algorithms that seems to perform at the same (low) level regardless of the choice of input data.

Sometimes conventional statistics proves to be more powerful or more accurate than machine learning. Initial determinations about the interdependence and nonlinearity of the data would have been wrong. This is not necessarily a weakness to machine learning, it is just a matter of choosing the right tool for the right job. Likewise, not all machine learning methods are created equal. Some are better for certain kinds of problems while others are better for other kinds of problems. For instance, some machine learning algorithms scale nicely to the size of the biological domains, others do not. Likewise, some methods may have assumptions or data requirements that render them inapplicable to the problem at hand. Knowing which method is best for a given problem is not inherently obvious. This is why it is critically important to try more than one machine learning method on any given training set. Another common misunderstanding about machine learning is that the patterns a machine learning tool finds or the trends it detects are non-obvious or not intrinsically detectable. On the contrary, many patterns or trends could be detected by a human expert – if they looked hard enough at the data. 3

* 1. ***Technologies Involved***
* **Programming language:**

Python v3.6

* **Frameworks:**

NumPy

Pandas

Matplotlib

Seaborn

Regular Expression

NLTK

ScikitLearn

* **Software:**

Anaconda Navigator v1.6.9

Jupyter Notebook

* **Hardware:**

Processor (two quad core processors) with 32 GB RAM

4

**3. METHODOLOGY**

***3.1 Detailed methodology***

* Our dataset on cancer genes and their mutations is provided by MSKCC and ONCOKB, cancer related organizations based in US.
* Used libraries:

Numpy

Pandas

Matplotlib

Seaborn

Regular Expression

NLTK

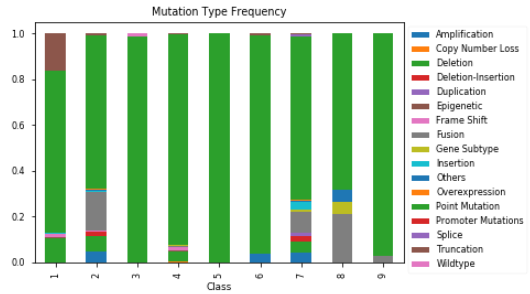
Random Forest Classifier

Linear SVC

* The dataset is of two types: CSV Format and Text Based data. After some preprocessing we discovered that there is a total of 8989 IDs,264 unique genes,2996 unique variations and 9 classes and also found out the maximal and minimal occurring genes.
* We plotted the frequency of each class using countplot and found that Class 7 is the most frequent. Plotted the maximal occurring genes against each class to find out which gene is dominating in which class.
* After learning the HGVS Nomenclature and referring to sites like OncoKB, we discovered that there are around 17 major types of cancer gene mutations in our dataset. So, using python regular expression and string functions we classified all the variants in these types and created a separate feature for it.
* Then we plot a bar graph of classes against the frequency of mutation types in each class. This tells us which class contains which types of mutations and in what frequency. With this knowledge we can determine what each of the 9 classes mean.
* After this we plot a count plot of mutation types of train variant data, to know which mutation type has maximum occurrence and how much they contribute to the occurrence of cancer. We found that point mutations have most frequency.
* Now we merged the text and variant dataframes on ID for test and train dataset. Using this, we plotted a distplot for the frequency of number of words. Resulting graph tells us that that majority ID’s have text length between 2000-15000 for train and test dataset.
* Our text data needs to be preprocessed so that we can use it for Random Forest and SVM model. We wrote functions to convert the data in useful format and extract necessary features from the data.

1. Tokenizing words: This function is used to convert original text to tokenized words and then remove stopwords and numbers. This was done using Stopwords library of NLTK.
2. Mutation Word table: This function is used to create a table for words to describe the mutation types from a list of tokenized words.
3. Gene-like words: This function is used to get Gene-name like words from the list of tokenized words. Regular expression is used for this.
4. Converting Variants to Mutation Types: This function is used to convert the 'Variant' Data into mutation\_type in a new column and returns the new data with a new column. 5

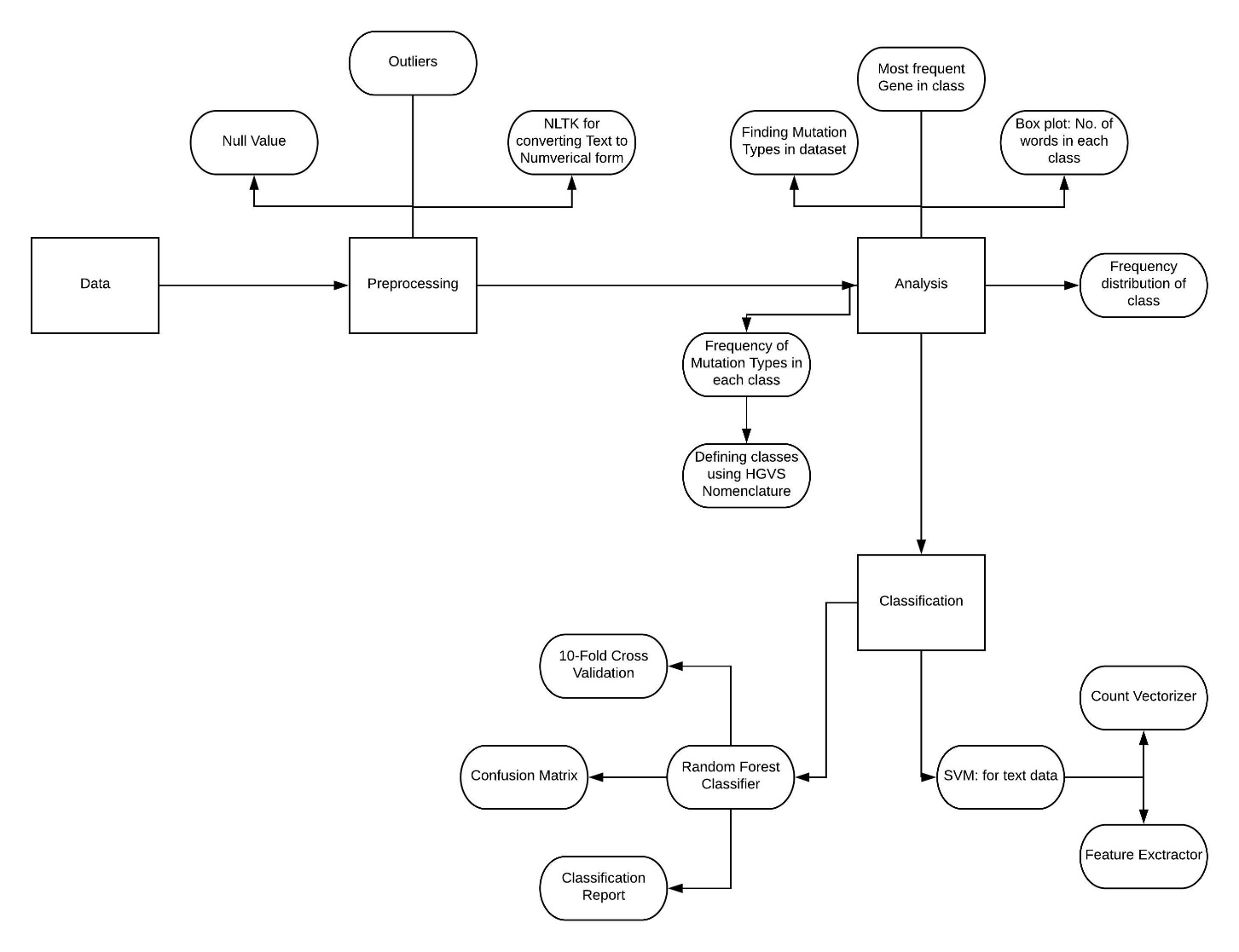
* Now we created a Mutation table, with mutation types as columns and IDs as rows. This table tells the presence of mutation types in each ID.
* Text data contains some very useful words which don’t lie in the mutation types but can be used for our classifications models to increase its accuracy. This process is known as Feature Importance. So, for this we created a table of Gene-like words. The table contains all the words which are like genes, as columns and IDs as rows.
* Next using LabelEncoder and OneHotEncoder, we created a Gene table with unique genes as columns and IDs as rows and a Mutation Type table with Mutation Types as columns and IDs as rows. These tables tell us about the presence of unique gene and mutation type for each ID.
* Now we merged all the above tables on ID column to create a Features table on which we will apply classification models.
* On applying Random forest, we got an accuracy of 65.86% and SVM got us an accuracy of 50.25%.

****

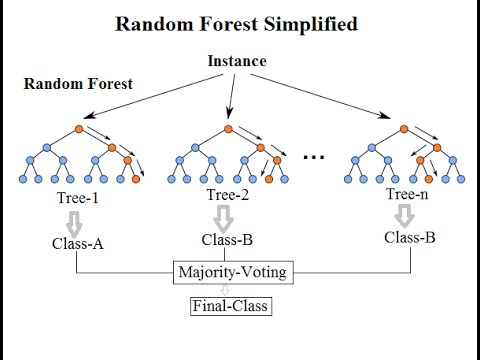
**Fig. 3.1.1 Mutation Type Frequency in each Class**

6

***3.2 Architecture Diagram***

****

**Fig. 3.2.1 Process Flow Chart**

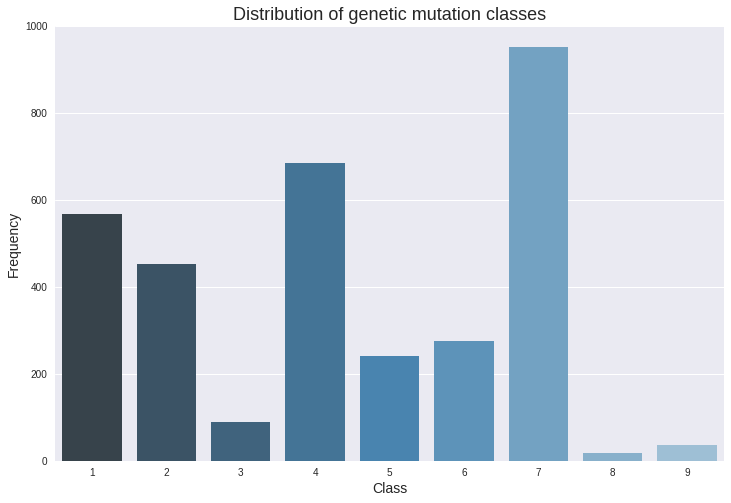
****

**Fig. 3.2.2 Random Forest Model Structure**

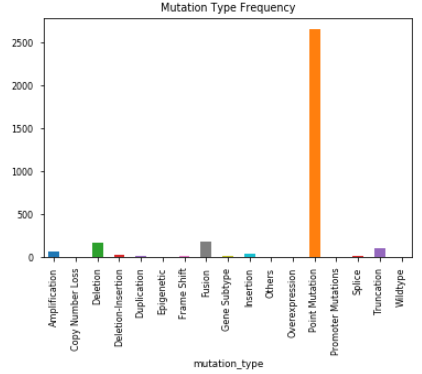
7

**4. IMPLEMENTATION AND RESULTS**

***4.1 Modules***

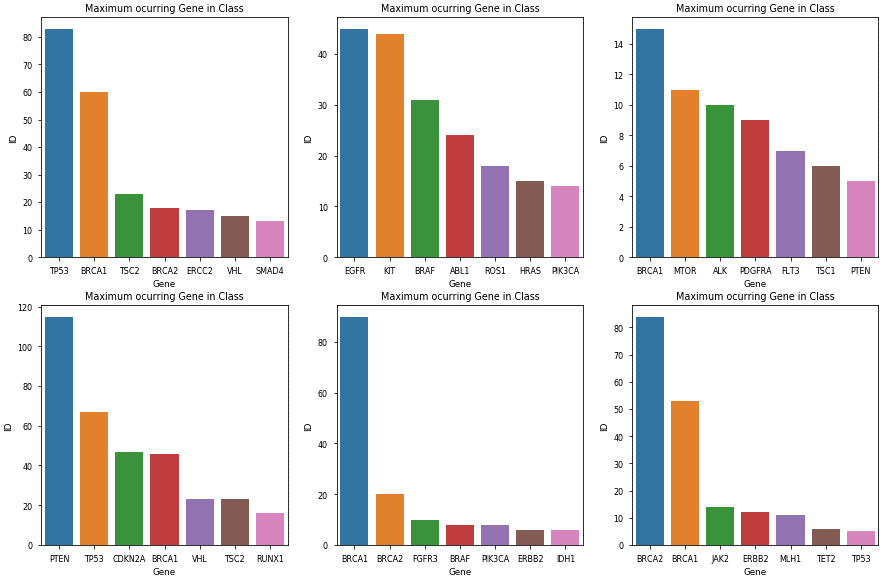
****

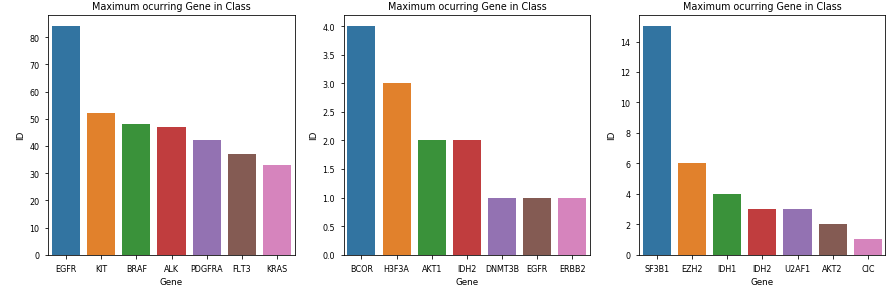
**Fig. 4.1.1 Distribution of Genetic Mutation Classes**

****

**Fig. 4.1.2 Frequency of Mutation Types**

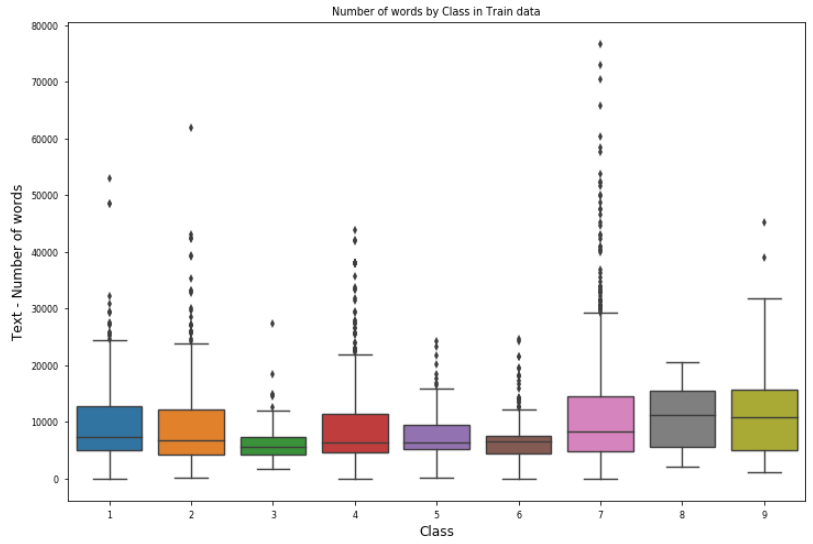
8

****

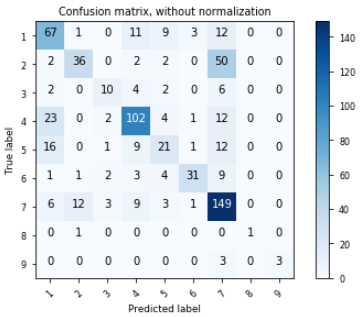
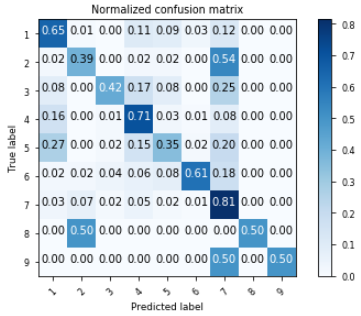
****

**Fig. 4.1.3 Frequency of maximum occurring Genes in each Class**

9

****

**Fig. 4.1.4 Number of Words by Class**

****

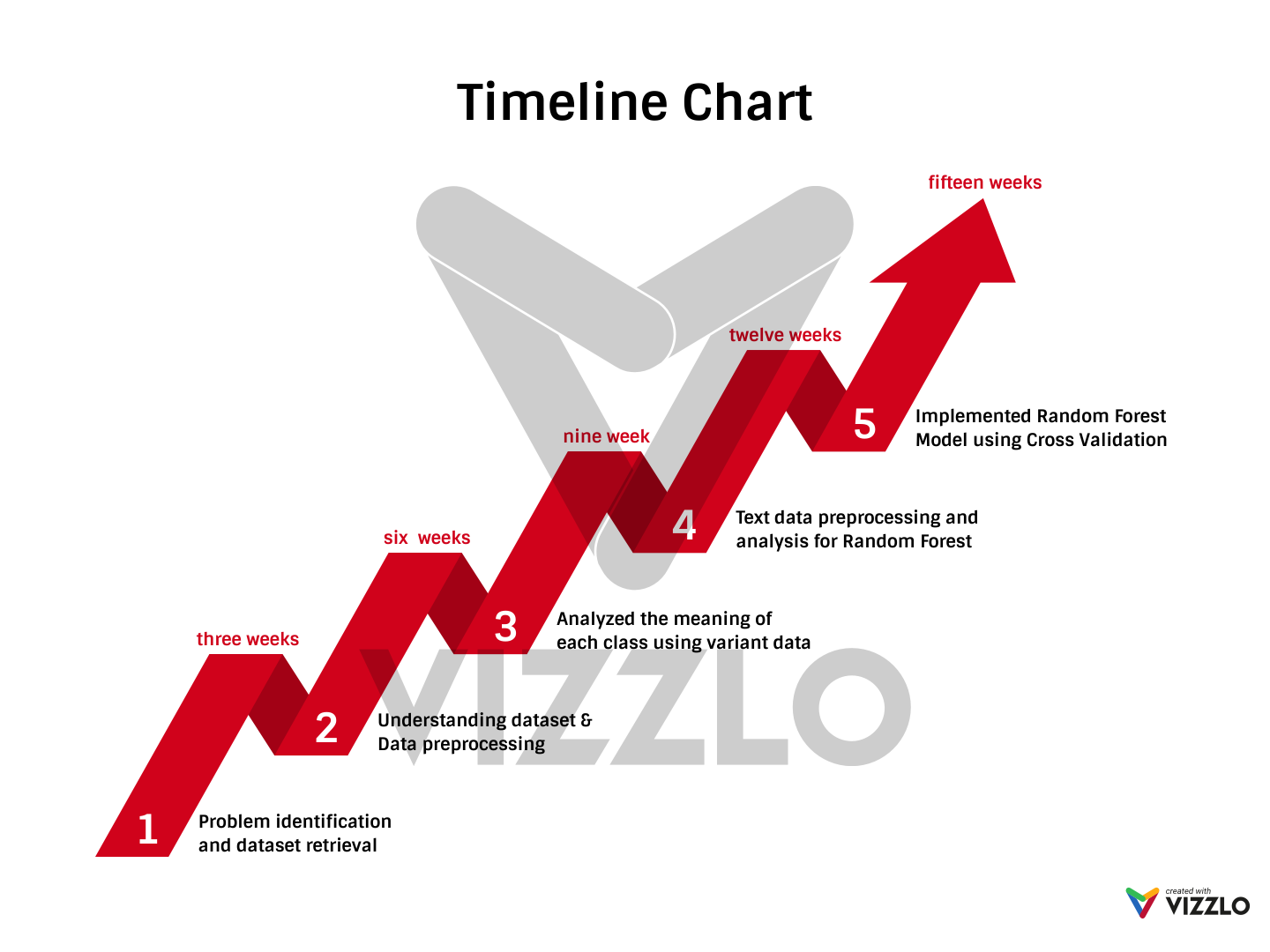
**(a) (b)**

**Fig. 4.1.5 Confusion Matrix (a)Normalized (b)Without Normalization**

10

**5. FUTURE WORK AND CONCLUSION**

***5.1 Progress Chart/Time Line Chart***

****

**Fig. 5.1.1 Timeline Chart**

****

**Fig. 5.1.2 Month-Wise Plan of Work**  11

***5.2 Future Work and Conclusion***

* After the analysis of the dataset and learning about cancer genes and their features, we concluded that the classes appear to be as follows:

1. Likely Loss-of-function
2. Likely Gain-of-function
3. Neutral
4. Loss-of-function
5. Likely Neutral
6. Inconclusive
7. Gain-of-function
8. Likely Switch-of-function
9. Switch-of-function

* The best accuracy is achieved from Random Forest model, which is 66.19% and is not very good as our entire dataset is text based and in converting it to numerical form we lost a lot of important features.
* After reading about the nomenclature of the genes, the strategy for improving model results was refocused on the information carried in the `Variation` variables themselves. Amino acid changes are coded according to a set of guidelines and different types of protein mutations are more likely to result in certain types of outcomes. Different amino acids also carry different properties of phosphorylation and charge which are meaningful to protein function. So, for example, it is probably important if the amino-acid change was to or from a threonine (Thr, T), serine (Ser, S), or tyrosine (Tyr, Y) as these could have more likelihood of loss or gain of function. But some values in our dataset doesn’t follow the nomenclature properly because of which we couldn’t classify more accurately.
* Our dataset contained Genes and their mutations from which we analyzed the mutation types. Using the mutation types, we were able to identify the meaning of each class. Now for future we plan on using Recurrent Neural Network and Natural Language Processing for the preprocessing and classification of Text-Based data. We believe that using Deep Learning in our project will help in better text classification and increase our accuracy score.

12

**REFERENCES**

[1] Muxuan Liang, Zhizhong Li, Ting Chen, and Jianyang Zeng, ‘Integrative Data Analysis of Multi-Platform Cancer Data with a Multimodal Deep Learning Approach’, IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 12, NO. 4, JULY/AUGUST 2015

[2] Amit Bhola and Arvind Kumar Tiwari, ‘MACHINE LEARNING BASED APPROACHES FOR CANCER CLASSIFICATION USING GENE EXPRESSION DATA’, Machine Learning and Applications: An International Journal (MLAIJ) Vol.2, No.3/4, December 2015

[3] Shweta Kharya, ‘USING DATA MINING TECHNIQUES FOR DIAGNOSIS AND PROGNOSIS OF CANCER DISEASE’, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012

[4] A. Priyanga and S. Prakasam, ‘Effectiveness of Data Mining - based Cancer Prediction System (DMBCPS)’, International Journal of Computer Applications (0975 – 8887) Volume 83 – No 10, December 2013

13