Machine Learning for Classification of Cancer Dataset for Gene Mutation Based Treatment

Jai Santosh Mandava
Department of Computer Science
California State University, Fullerton
Fullerton, USA
mandavajsantosh@csu.fullerton.edu

Fulya Kocaman
Department of Computer Science
California State University, Fullerton
Fullerton, USA
fulyakocaman@csu.fullerton.edu

Doina Bein
Department of Computer Science
California State University, Fullerton
Fullerton, USA
dbein@fullerton.edu

Abhishek Verma
Department of Computer Science
California State University, Northridge
Northridge, USA
abhishek.verma@csun.edu

Marian Sorin Nistor
Department of Computer Science
Universität der Bundeswehr München
Neubiberg, Germany
sorin.nistor@unibw.de

Stefan Pickl
Department of Computer Science
Universität der Bundeswehr München
Neubiberg, Germany
stefan.pickl@unibw.de

***Abstract*— The objective of this paper is to develop a Machine learning model that can classify cancer patients. Gene mutation-based treatment has a very good success ratio, but only a few cancer institutes follow it. This research uses natural language processing techniques to remove unwanted text and convert the categorical data into numerical data using response coded and one-hot encoded. Then we apply various classification algorithms to classify the training data. The proposed system has the advantage of reducing the time to analyze and classify clinical data of patients, which translates into less wait time for patients in order to get results from pathologists. The results of our experiment will demonstrate that the Stacking Classifier algorithm with One-Hot encoding and Term Frequency - Inverse Document Frequency (TF-IDF) techniques perform better than other Machine Learning methods with around 67% accuracy on the test data.**

Keywords*— *Gene mutation, Cancer patient, Machine Learning, Natural Language Processing, NLP, Stacking Classifier.

I. Introduction

A sizable set of people suffer from various types of cancer every year, and nearly half of them die without proper diagnosis. According to the National Cancer Institute [7], gene mutations will enhance the treatment of disease by supplying effective medicines and decreasing the number of deaths. A considerable amount of scientific evidence collected over the last decades can be used to enhance cancer detection. Data on the health of past patients with diagnostic medicine are found in DNA or gene mutation. Based on the previous results, this approach can help physicians to treat future patients. Gene mutations are carried out by manual techniques. An experienced health care professional has to manually apply knowledge to track specifics and suit the gene of the patient. By using machine learning algorithms, the clinical data analysis can be automated to improve the diagnosis and treatment in terms of accuracy and time. Artificial Intelligence can predict the gene mutation in the patient groups by implementing machine-learning algorithms. In gene mutations linked to lung cancer, family genetics could play a significant role in analyzing the likelihood of cancer risk [1].

While gene mutations are known for lung cancer diagnosis, they can be used to collect the data and compare them manually. However, the alternative is to replace it with greater precision through the use of machine learning algorithms. This research applies various natural language processing algorithms to the clinical dataset collected from lung cancer patients to perform data cleaning. Furthermore, the paper focuses on applying machine learning classification.

According to Kourou, et al. [4], "Machine Learning (ML) Algorithms can detect and classify trends and interrelationships from complex data sets when effectively predicting future results from cancer." The preparation, testing, and prediction are three phases of ML. Algorithms do software cleaning, replication, redundancy, and data preparation during the preparation, whereas the ML techniques are used in the validation process for test data using specialized expertise. To ensure the correct algorithm is chosen, the accuracy should be determined. Next, the algorithm is used for the prediction on new data.

The proposed system was implemented in the Jupyter notebook using Python language. The problem encountered in the gene mutation treatment is that any manual process takes more time to analyze the data of the patient than an automated one. For the past several years, there has been much improvement in machine learning algorithms, parallel processing, and the ability of algorithms to process data at

scale. We implemented machine learning algorithms on the dataset provided by Memorial Sloan Kettering Cancer Center [3], which has around 3200 data with the clinical text of each patient. We analyze two separate files, described as follows:

- Training variants - the file that includes the definition of genetic variations used for training. Training variants Fields are IDs (including the gene ID for the row of clinical evidence used to link the mutation), gene (gene where the genetic mutation is located), variations, class (includes 1-9 classes).
- Training text – contains the clinical text of the patients with their ID.

The proposed system has the following advantages. First, it reduces the time to analyze and classify the clinical data of patients, so this translates into less waiting time for patients to get results from pathologists. Second, our model can handle a large amount of data at a time.

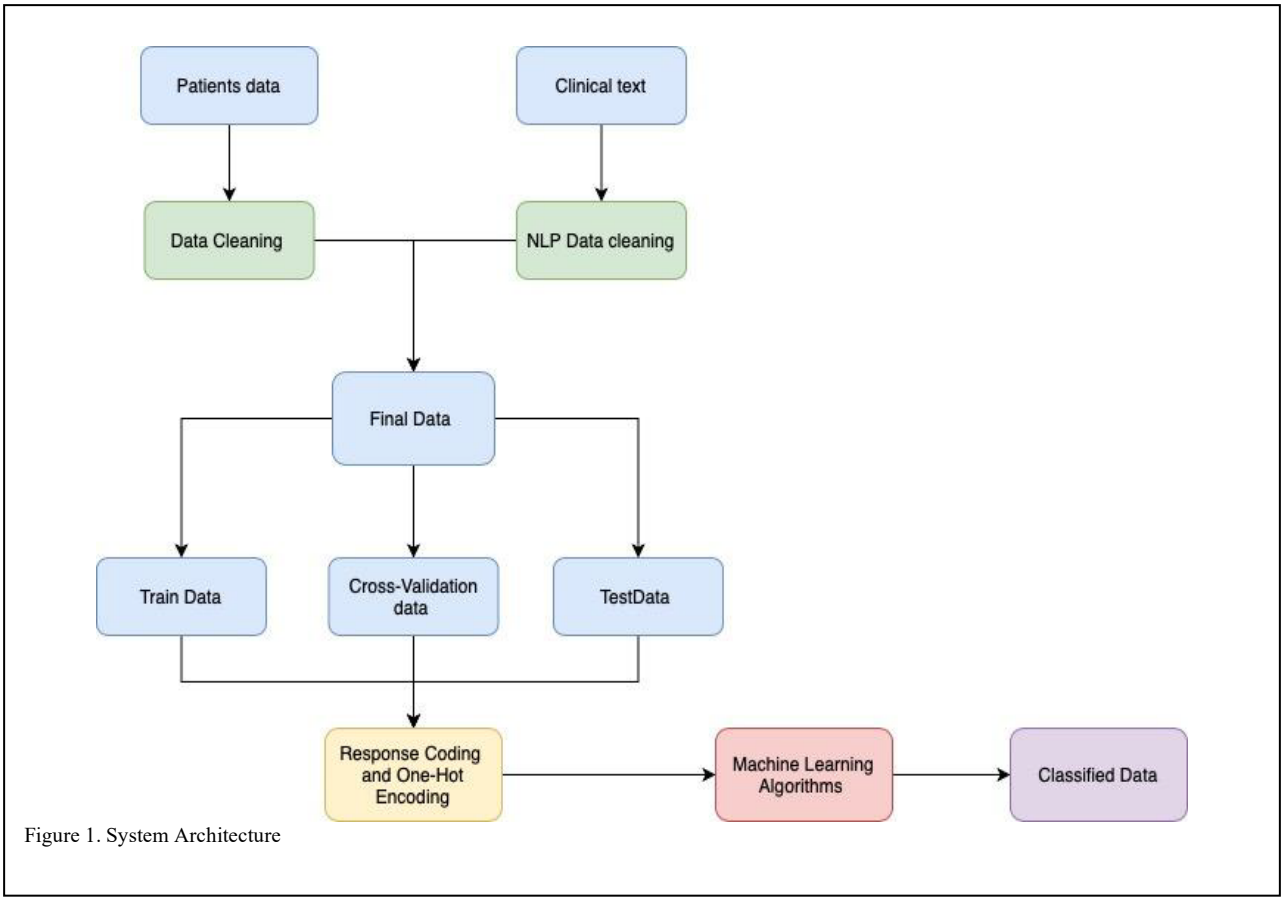

Figure 1. System Architecture

The paper is organized as follows. In Section II, we present background information, followed by the proposed system model in Section III. Simulation results are shown in Section IV, concluding remarks and future work are presented in Section V.

II. Background

In the eighteenth century, cancer was treated using surgery, which was considered then as primary treatment. Post-surgery herbal therapy, castor oil, and arsenic were administered to the patient. Radiation therapy came into existence in 1895 but only cured a few cancer types [2]. In the treatment of lymphoma cancer with the subsequent approval of the FDA, immunotherapy was introduced at the end of 1987. In 1990, gene-based treatment was approved by the FDA for immunodeficiency disorder. Later, gene-based treatment was used on cancer patients and successfully mitigated different types of cancer such as brain tumors, acute lymphocytic, and others. According to Memorial Sloan Kettering Cancer Center [3], genetic tests are conducted in various cancer types and are divided into nine different classes for treatment. Each class is a cluster of gene-related treatments in which patients get treated according to the class. Patient classification happens through three phases, and eventually, the appropriate patient class is determined. The first phase is where a collection of the records of the patient is used, such as PET scan, CT scan, and previous health reports. Then pathologist looks for the clinical evidence from the medical literature and converts it into clinical text for further analysis. In the last phase, pathologists classify the patient into nine different classes by analyzing the clinical text. The final phase usually takes one to three days to classify the patients into a particular category.

Lung cancer is a type of cancer that can be traced from family history. Although, according to Memorial Sloan Kettering Cancer Center [3], many people do not get proper

treatment because the selected medication does not work on them, this is where gene mutation-based treatment can be helpful. However, the clinical data of the patient needs to be analyzed manually, and that requires the time of experienced staff to make the decisions.

The authors in [5] have used the Cancer Genomic Atlas results, including results from the patients who have lung adenocarcinoma (LUAD). In order to acquire genes for LUAD patients, the author applies a machine learning algorithm. Further information on the dataset is given by using the KNN algorithm, decision trees, SVM, and Naïve Bayes. They show that ZNF560 and DRD3 are the positive genes capable of surviving LUAD. In order to assess the patient survival rate, the top six genes are later tested by the research model. The findings have been comparable to the original.

The authors in [6] implemented deep learning algorithms to classify lung cancer into gene mutations. They considered two major lung cancer types: Lung Adenocarcinoma (LUAD) and Lung Squamous Cell Carcinoma (LUSC). Conventional Neural Networks are used to study the images and classify the genes, according to the report. After training the model, they researched in the real world through means of new data of the patient for testing. The results were compared to the results of the pathologist from different sources, which show there is slight misclassification of the model. Their research does not involve other types of lung cancer and focuses on only two major types.

III. Proposed System Architecture

Our proposed system architecture is shown in Fig. 1. We implemented our program in Jupyter Notebook using Python language. Unfortunately, the final phase of the gene mutation-based process takes more time, so we replace that with the machine learning algorithm.

We have approximately 3,200 data points [3] divided into three parts: training data, cross-validation data, and test data. Before dividing the data, we need to clean the data and remove empty rows from the data. Using Natural Language Processing (NLP), we removed unwanted text in the clinical data of the patients. After splitting data, we implemented Response Coding and One-Hot encoding to convert the text data to numerical form for fitting the model in machine learning algorithms. Then the genetic mutations were classified into one of nine classes.

There are two columns of categorical data, which are genes and variations. To categorize the data, we implement one-hot encoding and response coding. One-hot encoding is implemented using the count vectorizer in the scikit-learn library. Response coding is part of the machine learning technique where it is used for categorizing the data. This

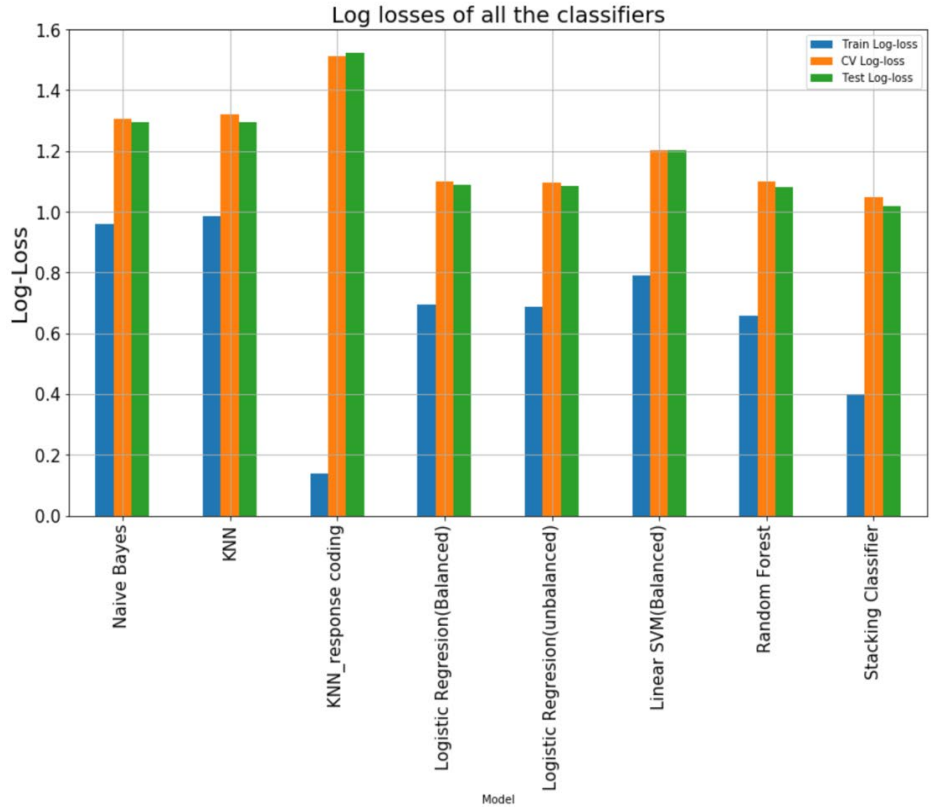

Figure 2. Comparison of log-loss performance for various machine learning classification algorithms.

technique is implemented to represent the data point by calculating the probability for each class by category. For example, suppose we have 233 unique genes for nine classes. We get 233 features by calculating the probability for each class. Suppose we take the three genes as three classes in the data. Table 1 describes how response coding is done for this example.

Table 1. Response Encoded Table

Gene	Class 0	Class 1	Class 2
ALK	4/20	5/20	11/20
CBL	14/35	10/35	11/35
KIT	4/10	3/10	3/10

There are three types of genes such as ALK, CBL, and KIT. There are twenty types of patients with genes as ALK, and twenty have different classes. For example, in Table 1, Class 0 has 4 ALK genes. The probability is calculated by the total number of genes in the class, divided by the total number of the specific gene contained in the data.

We implemented both Response Coding and scikit-learn's TF-IDF vectorizer methods for the text data as well. We then combine the features using One-Hot encoding on gene and variations with TF-IDF on the text data in each data set separately. We also combined all three features, where Response Coding was applied to gene, variation, and text data. The following classifiers with different NLP techniques were tested with different values using the CalibratedClassifierCV from scikit-learn library. The best hyperparameters based on the smallest log-loss for each classifier were found during cross-validation. We then implemented the following classifiers using the best hyperparameters for each one.

We use the following classifiers:

- Naïve Bayes Classifier is based on Bayes theorem for classification technique and assumes that predictors are independent. Thus, the existence of a specific feature in a class is irrelevant to the existence of any other feature in Naive Bayes classification. MultinomialNB is implemented in this research, and it can be imported from scikit-learn library.
- K - Nearest Neighbors (KNN) can be used for statistical problems of classification as well as regression.
- Linear Support Vector Machine (SVM) is a linear model of a supervised machine learning algorithm which can be implemented in regression and classification modes. This is used primarily in the classification model. The value of each data element is the value of each coordinate in the SVM algorithm as a point in the n-dimensional field. Then we define the hyperplane, which differentiates very well between the two groups.
- Random Forest Classifier is a guided algorithm of learning. It can be used both for regression and classification. It is based on the idea of bagging. A large number of weak classifiers named decision trees pooled together is better than a single strong classifier. The random forest generates randomly selected data samples for decision trees, obtains predictions of each tree, and selects by voting the best solution. It also gives a very clear indication of the value of the feature.
- Stacking Classifier is a technique for combining many models of classification with a meta-classifier. The individual classification models are trained based on the full training set, and the meta-classifier is then installed on the output from the individual model classifications within the ensemble-meta-features. The meta-classifier may also be educated on the predicted class labels or ensemble probabilities.

IV. Experimental Results

We implemented various machine learning algorithms on the dataset provided in [3], which has around 3,200 data with the clinical text of each patient. Our results are displayed in Fig. 2. Best accuracy is obtained from the random forest and stacking classifier. Table 2 shows the misclassification accuracy and log-loss of each model on the test data we use in this project.

Table 2. Classification Results

ML Classifier	**Mis-Classifion Test Accuracy**	**Test Log-Loss**
Naïve Bayes with One-Hot Encoding and TF-IDF	38.25%	1.2939
Naïve Bayes with Response Coding	60.09%	1.7293
KNN with One-Hot Encoding and TF-IDF	39.91%	1.2935
KNN with Response Coding	47.74%	1.5251
Logistic Regression (Balanced)	34.64%	1.0891
Logistic Regression (unbalanced)	34.49%	1.0850
Linear SVM(Balanced)	35.84%	1.2016
Random Forest	33.58%	1.0831
Stacking Classifier	33.37%	1.0183

V. Conclusion and Future Work

In this paper, we describe Machine learning models that can classify cancer data of the patients. Our approach focuses on replacing the final phase of the gene mutation treatment for cancer using a machine learning algorithm. This implementation helps in improving the accuracy of the analysis and reduces the wait time for classification. The classification was performed using several popular machine learning

algorithms. The test log-loss of the Stacking classifier was close to 1. The classification accuracy came out to be about 67% on test data. The problem is reasonably challenging, and more work is needed to reduce the error rate of the model so that the model accuracy can match the real-world accuracy of experienced pathologists.

Future work for the project will be mainly focusing on improving the accuracy of the model. In addition, we would expand the dataset beyond 3,200 data points to implement deep learning algorithms that require larger amounts of training data and may produce improved classification accuracy. Furthermore, such a model can also be adapted to a triage process. For example, it might reduce the time to analyze and classify the triage protocols and the exiting clinical data of patients to improve the treatment response.

ACKNOWLEDGMENT

This research was sponsored by the NATO Science for Peace and Security Programme under grant SPS MYP G5700.

REFERENCES

[1] Genetics Home Reference. (2019). Available online at https://ghr.nlm.nih.gov/condition/lung-cancer#sourcesforpage . Last accessed December 10, 2020.

[2] M.H. Amer. Gene therapy for cancer: present status and future perspective. Molecular and cellular therapies, 2, 27. https://doi.org/10.1186/2052-8426-2-27, 2014.

[3] Memorial Sloan Kettering Cancer Center. MSK-IMPACT: A Targeted Test for Mutations in Both Rare and Common Cancers. Available online at https://www.mskcc.org/msk-impact, Last accessed December 10, 2020.

[4] K. Kourou, T.P Exarchos, K.P Exarchos, M.V Karamouzis, and D.I Fotiadis. Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal, Volume 13,Pages 8-17, ISSN 2001-0370, 2015.

[5] H.-J. Cho, S. Lee, Y.G. Ji, and D.H. Lee, D. H. Association of specific gene mutations derived from machine learning with survival in lung adenocarcinoma. PLOS ONE, 13(11), e0207204. doi: 10.1371/journal.pone.0207204, 2018.

[6] N. Coudray et al.. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nature Medicine, 24(10), 1559-1567, 2018.

[7] National Institute of Cancer. (2017). The genetics of cancer. October 12, 2017. Available online at https://www.cancer.gov/about-cancer/causes-prevention/genetics. Last accessed December 10, 2020.