Cancer Tumor Detection using Genetic Mutated Data and Machine Learning Models

1st Aniruddha Mohanty
Computer Science and Engineering
CHRIST (Deemed to be University)
Bangalore, Karnataka, India
aniruddha.mohanty@res.christuniversity.in

2nd Alok Ranjan Prusty
DGT, RDSDE
NSTI(W)
Kolkata, West Bengal, India
dralokrprusty@gmail.com

3rd Ravindranath C. Cherukuri
Computer Science and Engineering
CHRIST (Deemed to be University)
Bangalore, Karnataka, India
cherukuri.ravindranath@christuniversity.in

***Abstract*—Early detection of a disease is a crucial task because of unavailability of proper medical facilities. Cancer is one of the critical diseases that needs early detection for survival. A cancer tumor is caused due to thousands of genetic mutations. Understanding the genetic mutations of cancer tumor is a tedious and time-consuming task. A list of genetic variations is analysed manually by a molecular pathologist. The clinical strips of indication are of nine classes, but the classification is still unknown. The objective of this implementation is to suggest a multiclass classifier which classifies the genetic mutations with respect to the clinical signs. The clinical evidences are text-evidences of gene mutations and analysed by Natural Language Processing (NLP). Various machine learning concepts like Naïve Bayes, Logistic Regression, Linear Support Vector Machine, Random Forest Classifier applied on the collected dataset which contain the evidence based on genetic mutations and other clinical evidences that pathology or specialists used to classify the gene mutations. The performances of the models are analysed to get the best results. The machine learning models are implemented and analyzed with the help of gene, variance and text features. Based on the variants of gene mutation, the risk of the cancer can be detected and the medications can be prescribed accordingly.**

Keywords: Cancer Tumor, Machine leaning models, Natural Language Processing, Datasets.

I. Introduction

Some of the human body cells growing in an uncontrolled way and affecting body parts is called cancer. Human body cells grow and multiply generally, is called cell division. This is the process where a new cell is generated and body requires them. When cells grow and become older or damaged, the new cell replace them. Sometimes this orderly process breaks and multiply the cells in uncertain manner. These cells may form tumors which are the lumps of tissues that can be cancerous or non-cancerous. Cancerous tumors spread into other body parts which is caused by thousands of genetic mutations.

In DNA or RNA, gene is a sequence of nucleotides which translates the synthesis of a gene product. There are five nucleobase adenine, guanine, cytosine, thymine, and uracil, which is represented with the symbols A, G, C, T, and U, respectively. The variation in the DNA sequence in each of our genomes is known as Genetic Variations. Genetic Variations are caused because of multiple ways such as genetic recombination and genetic mutation [1]. Recombination of gene comprises the exchange of genetic substances either between multiple chromosomes or between other areas of the same chromosomes. Genetic mutation is the permanent change in DNA or the alteration of nucleotide sequence of genome of a creature [2].

Genetic testing is one of the innovative area in medicine techniques and the way diseases like cancer are treated. The major hurdle in using gene classification is huge amount of manual work [3]. In order to generate clinical report, the process starts by the molecular pathology laboratory of "reportable" and "non-reportable" variants.

The applications of Machine Learning classification techniques like Logistic Regression classifier and Random Forest classifier, along with deep learning Recurrent Neural Network (RNN) classifiers have been used [4] to classify the genetic mutations on collected text data.

The rest of the paper is as follows: Section II will explain the background details of the experiment. Section III describes the various investigations performed connected to gene mutations. Section IV presents the system design with both training and testing dataset analysis. This section also explains the black box view of the designed approach. Section V explains methodologies with the various NLP and text transformation models. Several other evaluation metrics and research models are also discussed. Section VI explains the experimental results and analysis followed by conclusion and future scope in Section VII.

II. Background study

Cancer is basically a genetic and epigenetic changes in human body [5]. The unusual gene functions and altered patterns of gene expressions are the causes of cancer [6]. Excess body growth also causes cancer [7]. As per World Health Organization (WHO), the cause of cancer is due to genetic changes because of communication between a person's genetic factors and the external agents are of three classes.

- Physical hazards.
- Chemical hazards.
- Biological carcinogens.

Tumorous cancers are of six categories, bone tumor, brain tumor, malignant soft tissue tumors, organ tumors, ovarian germ cell tumors and skin tumors. Other types of cancers are blood cancers, lymphoma and brain cancer having their

978-1-6654-7258-6/22/$31.00

own staging systems [8]. Cancer diagnosis methods are [9] Lab tests like Diagnostic imaging, Endoscopic exams, genetic tests and tumor biopsies.

III. Related work

Gangmin and others [10] classify the genetic mutation of cancer treatment by applying XGBoost and SVM machine learning concepts. In the initial step, the texts in the training datasets are cleaned and useful features are extracted. The text files contain sets of words of gene and its variant of characters. By the help of TF-IDF technique, text fields are converted into numerical vectors which helps to analyse the frequency of words in the literature. Then balance the training data by the help of Synthetic Minority Oversampling Technique (SMOTE) which picks the points from the minority classes randomly and calculates k-Nearest Neighbour for the point and add the selected points among its neighbours. Then apply XGBoost and SVM classification technique as regression and classification.

Sandy and others [11] proposed the cancer diagnosis through image analytics which is known as "radiomics". The conventional radiomics is the claim of deep learning in which convolutional neural networks is used to distinguish the most useful sections and features. With medical imaging, images or subregions of labelled images are used as the input layer of training. Then, convolved in sublayers along with the identified classifiers to identify those image features helping in classification. The source task labels are related to underlying tumor pathophysiology and the training image passed as source task. Then it transfers the parameter to target task by the help of a series of convolutional layer and max pooling layers to classify and identify the tumors.

Tin Ceraj and others [12] applied Word2vec embedding model for deep learning classification of gene mutations which can affect the tumor growth. The Word2vec capturers the context of a word in a document, semantic and syntactic similarity, relation with other words. Initially, preprocessing is used to remove punctuations, symbols, stop words, tokens which consist digits only and finally lemmatization. Word embeddings generated with variations in hyperparameters are fed to the deep model which is used for the final classification and comparison of word encoding models. The model used is bidirectional LSTM and it consists of dense, LSTM and dropout layers which uses Word2vec embeddings as input layer resulting with a probability for each of nine classes for genetic mutations.

IV. System Design

Understanding the genetic mutations which matter in a cancer tumor is a really toughest task. To solve this problem using knowledge of Data Science and Machine Learning, the basic work flow is composed of three parts.

- A list of genetic variations is selected and analysed by a molecular pathologist.
- The relevant information on genetic variations is searched by the molecular pathologists.
- Finally, the molecular pathologist analyses the evidence related to every single variation to classify them.

The aim is to substitute the above steps by the help of machine learning models as multi-class classification problem.

The dataset contains the description of gene mutations with genes and their variations. All these data is separated by commas and the data captures having ID, Gene, Variation and Class field values.

In order to implement the proposed approach, the raw data is collected and analysed with the help of Exploratory Data Analysis technique. Then the texts are pre-processed using Natural Language Toolkit (NLTK) so that the data will be ready to use as input for the machine learning model. The data is tokenised, stop words and extra spaces removed, texts converted to lower cases is done as part of pre-processing. Then modelling is

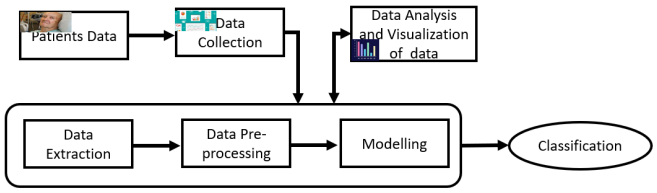

Fig. 1. Cancer Tumor Detection Approach

applied after segregating the data into train, test and cross validation distributions. Below modelling concepts have been applied to classify the genetic variations automatically and performances can be compared. This implementation process is described in Fig. 1.

- Naïve Bayes.
- Logistic Regression.
- Linear Support Vector Machines.
- Random Forest Tree.

To verify the performance of each model, confusion matrix is used to identify correctly and incorrectly classified points. Also, hyper parameter tuning is used to get the best hyperparameters.

V. Methodology

In this implementation, machine learning algorithms are used to help the expert annotated knowledge base as starting point. By this the genetic variations will be classified automatically and helps to detect unique features for each class.

Initially, the collected data need to be visualized and extracted all possible information. It helps to identify the relationship between the gene, it's mutations and the classes. Exploratory Data Analysis helps in data visualization.

Exploratory Data Analysis (EDA) is a tactic to investigate and analyse the datasets to summarize their main characteristics in pictorial mode [13]. Exploratory data analysis is implemented in two ways

- **Univariate Analysis:** In this approach, there is only one dependable variable by which data can be derived, defined, summarized and analysed the patterns. It can accept categorical and numerical variables.

- **Multivariate Analysis:** In this method, more than two variables can be analysed simultaneously and more complex data also be studied.

Naïve Bayes: In the field of text classification, two types of Naïve Bayes classifiers [14] are mostly used. Naïve Bayes is easy and fast to predict the classification, used in multi class classification. So, it is used in this implementation.

Logistic Regression: Logistic regression [15] is used for classifying both linear and non-linear data with binary responses which takes 1 to indicate a success and 0 a failure. For a document (Text data) and class (Class labels), logistic regression directly estimates the parameters of $Pr(Classlabels|Textdata)$.

$$Pr(Classlabels|Textdata) = Pr(Textdata) \tag{1}$$

Where the text data will fall into one of the class categories (r), known or unknown.

$$r = \begin{cases} 0 & Known \\ 1 & Unknown \end{cases} \tag{2}$$

Linear Support Vector Machines: SVM [16] is defined to find the optimized separated data into two classes over a vector space. The decision surface for a linearly separable space is a hyperplane which can be written as

$$wy + b = 0 \tag{3}$$

where y is an arbitrary object which will be classified, the vector w and constant b are obtained from a training set of linearly distinguishable objects. This is more efficient in high dimensional space.

Random Forest Tree: Random Forest [17] is an ensemble classifier which consists of assembly of tree-structured base classifiers and accomplished by bootstrap sample of the data subsets. The RF algorithm associated with a set of training documents D and N_f features.

VI. Experiment And Result Analysis

To determine the genetic variations or mutations, various machine learning techniques are used like Naïve Bayes, Logistic Regression, Linear Support Vector Machines and Random Forest Tree. Intel core i7 powered processor with 16GB RAM are used for execution purpose. Scikit-learn is a python library is also used. Natural Language Toolkit (NLTK) is a standard python library which is used to process the text data and matplotlib, seaborn are used to visualize the data. Apart from this other python libraries like pandas, time, regular expression (re), collections and math python libraries are also used.

A. Dataset

The dataset is collected from the Kaggle competition (https://www.kaggle.com/c/msk-redefining-cancer-treatment/data) and Memorial Sloan Kettering Cancer Centre (MSKCC) [18] [19]. The dataset is having three parameters - genes, variations and clinical text. In training dataset, there are nine classes of mutations. Both training and test datasets are shared via two different files. One $(training/test\_variants)$ encloses the genetic mutations, the other $(training/test\_text)$ helps the clinical evidence (text). Both the files are linked via the ID field. This dataset is used to classify the nine types of gene variants which is treated as multiclass classifications.

B. Result Analysis

Machine learning algorithm helps to classify genetic variations automatically using knowledge base as baseline. The objective is to predict the probability of each data-point fitting to one of the nine classes with the constraints like interpretability, class probabilities, penalize the errors in class probabilities and no latency constraints. In order to verify the penalize errors in class probabilities multi class logloss and confusion matrix are used. In this experiment, dataset has been fragmented randomly into three parts train, cross validation and test with 65%, 15%, 20% of data respectively.

Reading the data: After collecting the dataset, the variant and text data need to read from both the files. Variant file is a comma separated file holding the explanation of the genetic mutations.

Preprocessing: After reading the data, prepossessing is done by the help of NLTK library. As part of preprocess, removing stop words, replacing every special character with space, replacing multiple space with single space is done. Then entire character is converted into lower case.

Performance metrics used: There are two metrices which can measure the performance of the proposed model, namely

- Multi class log loss
- Confusion Matrix

Multi class logloss: This log loss is also known as categorical cross-entropy. The equation for the metric is

$$logloss = -\frac{1}{n}\sum_{i=1}^{P}\sum_{j=1}^{M} y_{ij} log(p_{ij}) \tag{4}$$

where P is the number of instances of data points, M is the number of classes [20]

Confusion Matrix: The performance of a classification model is evaluated by a Confusion Matrix [12] which is of $M * M$, where M denotes the count of target classes. The vital part of confusion matrix is True Positive, True Negative, False Positive and False Negative.

Featurization: As genes and variations are categorical features, they need to convert to numerical vectors. For vectorization one-hot encoding and response coding approach can be followed.

- **One-hot Encoding:** In one-hot encoding for each level of categorical values, a new variable is mapped with a binary value either 0 or 1. Where, 0 symbolizes the absence and 1 for the existence of the category [12].
- **Response coding:** This helps to determine the probability of the data point for a particular class. Based on the value of categorical data with N new features are considered for $N-class$ classification problem. Then the probability of the data point belonging to each class are computed.

C. Analyse the data

Univariate analysis: It is a process to derive the data, define, summarize and analyse the pattern in the data. In this implementation, univariate analysis can be performed for gene, variance and text features. Gene and variance are the categorical features whereas text is the text features with sequence of words. All these features are helping to detect the gene classification.

In the dataset, there exists 235 distinct genes and in variance features, training data contains a total of 2124 data points and from these 1927 are the unique variations. In text feature, every row of data is considered as "TEXT" and split each row into words by space. Then create a dictionary with those words and increment the count when a word is identified. The difference of training and cross validation in variance features are more than gene features. So, gene feature is more stable than variance feature which is shown in Table I. At the end of the univariate analysis, 97.12 % of word of test data appeared in train data and 98.05% of word of Cross Validation appeared in train data.

TABLE I
COMPARISON OF FEATURES

Feature	Train	Cross Validation	Test	Stability
Gene	0.8	1.68	1.73	Stable
Variance	1.04	1.23	1.2	Less Stable
Text	0.76	1.3	1.19	Stable

Data Preparation: Gene, variance and text features are useful and stable by verifying with the help of logistic regression models individually. For data preparation, one-hot encoding is applied to gene, variance and text features, then all the features are merged. Same implementation is also applied for response coding and all the features are merged. So, below are the observations for one-hot encoding and Response encoding shown in Table II and Table III.

TABLE II
ONE HOT ENCODING FEATURES

Data	No. of data points	No. of features
Train	2124	57039
Cross-validate	665	57039
Test	532	57039

TABLE III
RESPONSE ENCODING FEATURES

Data	No. of data points	No. of features
Train	2124	27
Cross-validate	665	27
Test	532	27

In one-hot encoding and response coding 57039 and 27-dimension data are extracted. So, one-hot encoding gives high dimensional data which can be handled by logistic regression, linear SVM, Naïve bayes models. Trees like decision tree, Random Forest models do not handle high dimension data but it can handle low dimension data. Based on the dimension of the data, the models can be implemented.

D. Machine Learning Model implementations

Base line model: For text data the baseline model is Naïve Bayes. High dimension data is available by one-hot encoding featurization. As part of Naïve Bayes implementation, multinomial Naïve Bayes is used. Calibrate classifier is also used here to convert the predicted data to probability values. Here also used alpha Laplace smoothing as hyperparameter and the values of hyperparameter used are like 0.00001, 0.0001, 0.001, 0.1, 1, 10, 100, 1000. The best hyperparameter value is 0.1, presented in Fig. 2. Training, cross validation and test log loss are 0.9033, 1.3509, 1.2784. The performance of the model is also verified from confusion matrix. In Fig. 2, $9 * 9$ confusion

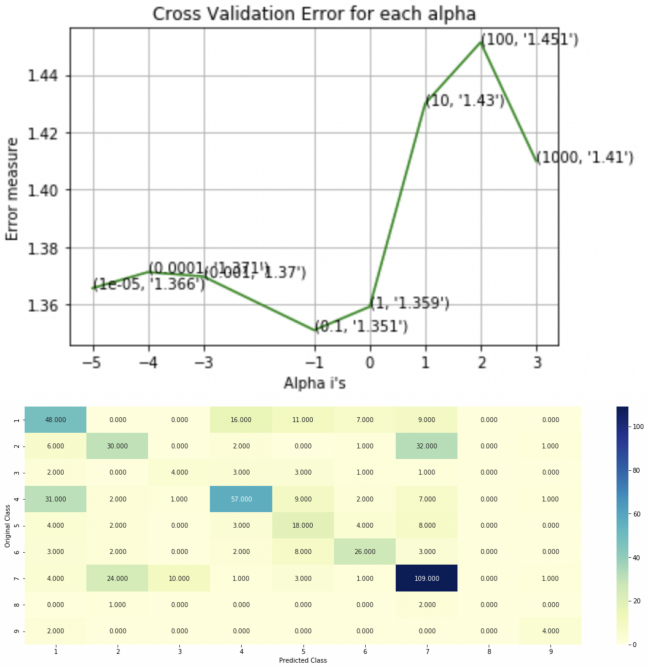

Fig. 2. Confusion matrix for Naïve Bayes

matrix, there are some confusions happening between $class\ 1$ and $class\ 4$. Again, between $class\ 2$ and $class\ 7$.

Logistic Regression: Logistic Regression can use one-hot encoding featurization because it can handle high dimension data and find plains in high dimension data to separate various classes. Two types of logistic regression models are implemented here.

In logistic regression class balancing implementations SGD classifier is used after balancing the data. This model is verified with various hyperparameter values like $1e-06, 1e-05, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100$ to find the best hyperparameter. The best hyperparameter value is 0.001 in which train, cross validation and test log loss are 0.6153177097029675, 1.217324457704446, 1.1047257743181136. The data is imbalanced and class balancing is performed in the implementation shown in Fig. 3. Class 8 and 9 perform very well. More

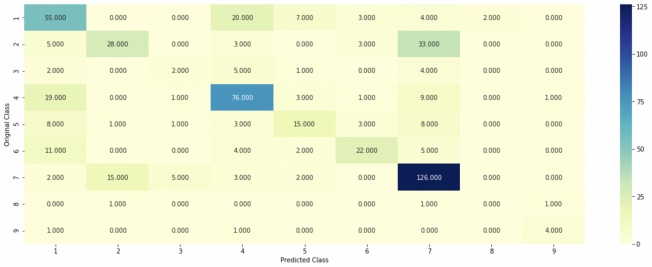

Fig. 3. Confusion matrix for LR with class balancing

Fig. 5. Confusion matrix for Linear SVM

prominent values are shown for class 1, 2, 4 and 7. In Logistic Regression without class balancing implementation, the model can apply large dimension of data. So, one-hot encoding featurization can be used. This model is verified with $1e-06, 1e-05, 0.0001, 0.001, 0.01, 0.1, 1$ hyperparameter values. In this model, the best hyperparameter value is 0.001. Train, cross validation and test log loss are 0.6257422677412771, 1.2521811628755943, 1.1306020069615057 respectively. Because of the class without balancing, minor class values like $class\ 8$ and 9 are giving better result which is represented in Fig. 4.

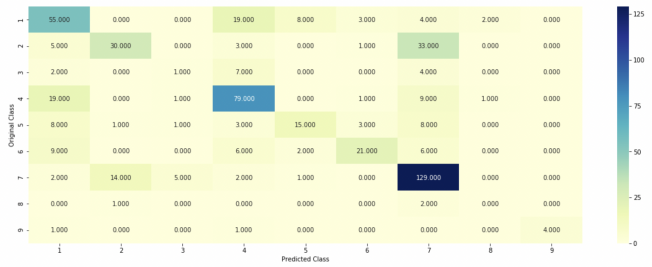

Fig. 4. Confusion matrix for LR without class balancing

Support Vector Machine (SVM): As the data is high dimensional, linear Support Vector Machine (SVM) will give better performance and uses hinge loss to verify the loss in the model. Balancing data is used to train the linear SVM model. Like other models, various hyperparameter values such as $1e-05, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100$ are also evaluated to get the best hyperparameter. The best hyperparameter is 0.01. Train, cross validation and test losses are 0.7628309867716067, 1.2442964974823838, 1.1541891969863685 respectively.

36.46% of data got misclassified in SVM model implementation. In confusion matrix, the diagonal values are prominent which is represented in Fig. 5. Even small classes are also showing better classification.

Random forest: Random forest is an ensemble model which uses decision trees. Decision tree works very well when the dimensionality of data is small and do not work well when the dimensionality is large. So, both the one-hot encoding and response coding featurization are evaluated, shown in Fig. 6. In Random Forest, there exists two hyperparameters. One is number of tree or the number of estimators, second one is the maximum depth of each tree. Various values like $100, 200, 500, 1000, 2000$ are for the number of estimators and values like 5 and 10 used for max depth. In this implementation the best estimator value is 1000 whereas max depth is 5. The train log loss, cross validation log loss and test log loss are 0.7035396732082752, 1.192993079576866, 1.0970923149103904 respectively.

Fig. 6. Confusion matrix for One hot encoding and response coding in Random Forest

In Random Forest with Response coding implementation, response coding is used for featurization. 27-dimensional features are used here. Various values like $10, 50, 100, 200, 500, 1000, 2000$ are for the number of estimators and values like $2, 3, 5$ and 10 used for max depth. In this implementation the best estimator value is 100 whereas max depth is 5. Training, cross validation and test losses are train log loss is 0.051902709444608406, 1.325750118424668 and 1.2116278998341923. There is a huge difference between the train and cross validation loss. This indicates the model overfits and hence the model is useless.

E. Comparison of Machine learning models

Misclassified points of Naïve bayes are very closer but log loss is very different in both cross validation and test data,

shown in Table IV. Confusion between the classes 1 and 4, 2 and 7 still exist as faced in Naïve bayes model.

Random forest with one-hot encoding is giving better performance than logistic regression. All the smaller classes like class $3, 6, 8$ and 9 are giving better result.

Logistic regression with balancing and SVM, both the models are showing almost the same result because in both the cases balancing concept is used with one-hot encoding as featurization.

Compared to other models, misclassified points are having higher value. So, the model can be discarded.

TABLE IV
COMPARISON OF IMPLEMENTED MODELS

Model	Train loss	CV loss	Test loss	Misclass
Naïve Bayes	0.9033	1.3509	1.2784	39.84% (One hot)
LR (Balancing)	0.6153	1.2173	1.1047	38.34% (One hot)
LR (Non Balancing)	0.625	1.252	1.13	37.21% (One hot)
SVM (Balancing)	0.762	1.244	1.154	36.46% (One hot)
Random Forest	0.703	1.192	1.097	36.84% (One hot)
Random Forest	0.052	1.325	1.211	46.80% (Response)
SVM	0.663	1.176	1.081	36.24% (One hot)

VII. CONCLUSION AND FUTURE SCOPE

This work is performed to recommend a multiclass classifier which classifies the genetic mutations based on the clinical evidences. The text information of genetic mutations help to distinguish mutations which results in the improvement of personalized medicine for cancer treatment. To design this multilabel classifier, NLP technique have been implemented. The text transformation models such as Count Vectorizer, TFIDF Vectorizer, and Word2Vec are used to convert text to a matrix of token counts. Machine learning models have been used for the prediction of cancer tumor and its variations. Out of all these models, support vector machine with one-hot encoding featurization is giving better result. The diagnosis of cancer is very expensive as well as onerous in the medical field. The proposed machine learning technique can be useful for medical practitioner in the diagnosis of cancer by detecting cancer tumor and its variants. As part of the future work, the datasets can be evaluated with different deep learning models like Convolutional Neural Network, BERT, Transformer and compared for the better performance.

REFERENCES

[1] Kukovacec, Marin, Kukurin, Toni, Vernier and Marin, "Personalized medicine: Redefining cancer treatment", Text Analysis and Retrieval, 2018, pp. 69.

[2] J. Jackson, Aimee L and Loeb, Lawrence A, "The mutation rate and cancer", Genetics, vol. 148, 4 (1998), pp.1483–1490.

[3] Dienstmann, Rodrigo and Dong, Fei and Borger, Darrell and Dias-Santagata, Dora and Ellisen, Leif W and Le, Long P and Iafrate, A John, "Standardized decision support in next generation sequencing reports of somatic cancer variants", Molecular oncology, vol. 8,5 (2014), pp. 859–873.

[4] Gupta, Meenu and Wu, Hao and Arora, Simrann and Gupta, Akash and Chaudhary, Gopal and Hua, Qiaozhi, "Gene Mutation Classification through Text Evidence Facilitating Cancer Tumour Detection", Journal of Healthcare Engineering, 2021.

[5] Prendergast, George C and Metz, Richard and Muller and Alexander J, "Towards a genetic definition of cancer-associated inflammation: role of the IDO pathway", vol. 176, 5 (2010), The American journal of pathology, pp.2082–2087.

[6] Jones, Peter A and Baylin, Stephen B, "The epigenomics of cancer", vol. 128, 4 (2007), Cell, pp.683–692.

[7] Bergström, Anna and Pisani, Paola and Tenet, Vanessa, Wolk, Alicja and Adami, Hans-Olov, "Overweight as an avoidable cause of cancer in Europe", vol. 91, 3 (2001), International journal of cancer, pp.421–430.

[8] Bookstein, Robert and MacGrogan, Donal and Hilsenbeck, Susan G and Sharkey, Francis and Allred, D Craig, "p53 is mutated in a subset of advanced-stage prostate cancers", vol. 53, 14 (1993), Cancer research, pp.3369–3373.

[9] Ahdoot, Michael and Wilbur, Andrew R and Reese, Sarah E and Lebastchi, Amir H and Mehralivand, Sherif and Gomella, Patrick T and Bloom, Jonathan and Gurram, Sandeep and Siddiqui, Minhaj and Pinsky, Paul and others, "MRI-targeted, systematic, and combined biopsy for prostate cancer diagnosis", vol. 382, 10 (2020), New England Journal of Medicine, pp.917–928.

[10] Li, Gangmin and Yao, Bei, "Classification of Genetic Mutations for Cancer Treatment with Machine Learning Approaches", vol. 7 (2018), International Journal of Design, Analysis and Tools for Integrated Circuits and Systems, pp.63–67.

[11] Napel, Sandy and Mu, Wei and Jardim-Perassi, Bruna V and Aerts, Hugo JWL and Gillies, Robert J, "Quantitative imaging of cancer in the postgenomic era: Radio (geno) mics, deep learning, and habitats", vol. 124, 24 (2018), International Journal of Design, Analysis and Tools for Integrated Circuits and Systems, pp.4633–4649.

[12] Ceraj, Tin and Kliman, Ivan and Kutnjak, Mateo, "Redefining cancer treatment: comparison of Word2vec embeddings using deep BiLSTM classification model", vol. 124, 24 (2019), Text Analysis and Retrieval 2019 Course Project Reports, pp.10.

[13] Sahoo, Kabita and Samal, Abhaya Kumar and Pramanik, Jitendra and Pani, Subhendu Kumar, "Exploratory data analysis using Python", vol. 8, 12 (2019), International Journal of Innovative Technology and Exploring Engineering (IJITEE), pp.2019.

[14] Chen, Jingnian and Huang, Houkuan and Tian, Shengfeng and Qu, Youli, "Feature selection for text classification with Na¨ıve Bayes", vol. 36, 3 (2009), Expert Systems with Applications, pp.5432–5435.

[15] MurtiRawat, Ram and Panchal, Shivam and Singh, Vivek Kumar and Panchal, Yash, "Breast Cancer detection using K-nearest neighbors, logistic regression and ensemble learning", vol. 5, 1 (2020), 2020 international conference on electronics and sustainable communication systems (ICESC), pp.534–540.

[16] Zhang, Wen and Yoshida, Taketoshi and Tang, Xijin, "Text classification based on multi-word with support vector machine", vol. 21, 8 (2008), Knowledge-Based Systems, pp.879–886.

[17] Kakulapti, V and Lalitha Bhavani, P and Swathi Reddy, K and Nissar Ahmed, P, "Comparative Analysis of Genomic Personalized Cancer Diagnosis by Machine Learning Approaches ROC Curve", Communication Software and Networks, 2021, pp.511–528.

[18] Ortiz, Michael V and Kobos, Rachel and Walsh, Michael and Slotkin, Emily K and Roberts, Stephen and Berger, Michael F and Hameed, Meera and Solit, David and Ladanyi, Marc and Shukla, Neerav and others, "Integrating genomics into clinical pediatric oncology using the molecular tumor board at the Memorial Sloan Kettering Cancer Center", vol. 63, 8 (2016), Pediatric blood & cancer, pp.1368–1374.

[19] Hyman, David M and Solit, David B and Arcila, Maria E and Cheng, Donavan T and Sabbatini, Paul and Baselga, Jose and Berger, Michael F and Ladanyi, Marc, "Precision medicine at Memorial Sloan Kettering Cancer Center: clinical next-generation sequencing enabling next-generation targeted therapy trials", vol. 20, 12 (2015), Drug discovery today, pp.1422–1428.

[20] Kabani, AbdulWahab and El-Sakka, Mahmoud R, "Object detection and localization using deep convolutional networks with softmax activation and multi-class log loss", vol. 67 (2016), International Conference on Image Analysis and Recognition, pp.358–366.