­



WHITE PAPER

DATA SCIENCE WITH PYTHON

TEAM – CRAZZY PIES

**Team Members**

Ashish Tele

Aditya Nagori

Kuldeep Kumar

Sriram K.

Yashwanth Vemula

Table of Contents

1. Introduction 3
2. Objective 6
3. Research and Literatures………………………………………………………………. 6
4. Data Exploration 7
5. Data Pre-Processing ……………………………………………………………………11
6. Modeling…………………………………………………………………………….…. 12
7. Model Comparison…………………………………………………….……………… 16
8. Conclusion and Recommendations……………………………………..……………..16
9. References………………………………………………………………………………20
10. **INTRODUCTION**

**The Genetics of Cancer**

Genes are in the DNA of each cell in your body. They control how the cell functions, including:

* How quickly it grows
* How often it divides
* How long it lives

Researchers estimate that each cell contains 30,000 different genes. Within each cell, genes are located on chromosomes.

**About chromosomes**

Chromosomes are the thread-like structures in cells that contain genes. There are 46 chromosomes, arranged in 2 sets of 23.

You inherit one set from your mother and one from your father. One chromosome in each set determines whether you are female or male. The other 22 chromosome pairs determine other physical characteristics. These chromosome pairs are called autosomes.

**How genes work**

Genes control how your cells work by making proteins. The proteins have specific functions and act as messengers for the cell. Each gene must have the correct instructions for making its protein. This allows the protein to perform the correct function for the cell. All cancers begin when one or more genes in a cell mutate. A mutation is a change. It creates an abnormal protein. Or it may prevent a protein’s formation. An abnormal protein provides different information than a normal protein. This can cause cells to multiply uncontrollably and become cancerous.

About genetic mutations: There are 2 basic types of genetic mutations:

**Acquired mutations**. These are the most common cause of cancer. They occur from damage to genes in a particular cell during a person’s life. For example, this could be a breast cell or a colon cell, which then goes on to divide many times and form a tumor. A tumor is an abnormal mass. Cancer that occurs because of acquired mutations is called sporadic cancer. Acquired mutations are not found in every cell in the body and they are not passed from parent to child. Factors that cause these mutations include:

* Tobacco
* Ultraviolet (UV) radiation
* Viruses
* Age

**Germline mutations**. These are less common. A germline mutation occurs in a sperm cell or egg cell. It passes directly from a parent to a child at the time of conception. As the embryo grows into a baby, the mutation from the initial sperm or egg cell is copied into every cell within the body. Because the mutation affects reproductive cells, it can pass from generation to generation. Cancer caused by germline mutations is called inherited cancer. It accounts for about 5% to 20% of all cancers.

**Mutations and cancer**

Mutations happen often. A mutation may be beneficial, harmful, or neutral. This depends where in the gene the change occurs. Typically, the body corrects most mutations.

A single mutation will likely not cause cancer. Usually, cancer occurs from multiple mutations over a lifetime. That is why cancer occurs more often in older people. They have had more opportunities for mutations to build up.

**Classification of genes linked to cancer**

Many of the genes that contribute to cancer development fall into broad categories:

**Tumor suppressor genes**. These are protective genes. Normally, they limit cell growth by:

* Monitoring how quickly cells divide into new cells
* Repairing mismatched DNA
* Controlling when a cell dies

When a tumor suppressor gene mutates, cells grow uncontrollably. And they may eventually form a tumor

**Ongoing problem**

Free-form text remains the primary means by which physicians record their observations and clinical findings. Unfortunately, this rich source of textual information is severely underutilized by predictive models in oncology. Current models rely primarily only on structured data. Most of the cancer research today takes place in biology and medicine. Computer science plays a minor supporting role in this process if at all. So, it is quite clear that NLP as a field has a chance to play a significant role in this battle.

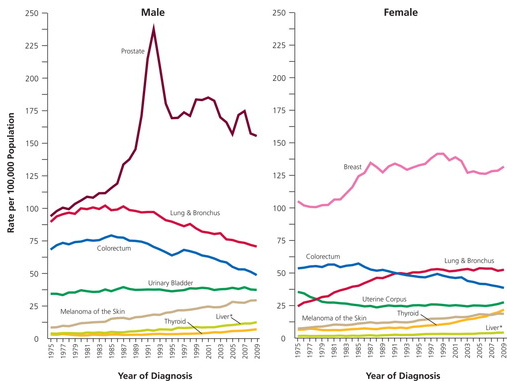
Indeed, free-form text remains the primary means by which physicians record their observations and clinical findings. Unfortunately, this rich source of textual information is severely underutilized by predictive models in oncology. Current models rely primarily only on structured data.

1. **OBJECTIVE**

The tremendous and increasing availability of electronic health records and clinical observations challenging the present cancer treatment facilities to go beyond their potential and it is creating the opportunities for automated extraction of information from clinical text. Our hypothesis is that natural language processing (NLP) could substantially reduce the burden of manual extraction in the available analyzed outcomes, like cancer recurrence, that are documented in unstructured clinical text, such as progress notes, radiology reports, and pathology reports.

1. **RESEARCH AND LITERATURES**

The Trends in Incidence Rates for Selected Cancers by Sex, United States, 1975 to 2009 are given below.



* Over 250 Billion have been invested
* Cancer rate dropped by only 5%

**Literatures**

* NLP algorithm to measure quality prostate cancer care – (term mapping to build a medical vocabulary)
* Using NLP to Improve Efficiency of Manual Chart Abstraction in Research

**Our work**

Evaluation of relations between text, Gene, and Variation columns.

1. **DATA EXPLORATION**

**Data description**

Memorial Sloan Kettering Cancer Center (MSKCC)

**Take Personalized medicine to its full potential!!!**

**Variables**

Training variants & Test variants

* ID – Unique number of the row used to link the row to clinical mutation to the clinical evidence
* Gene – the gene where genetic mutation is located
* Variation – amino acid change for the mutations
* Class – 9 classes to classify genetic mutations (target)

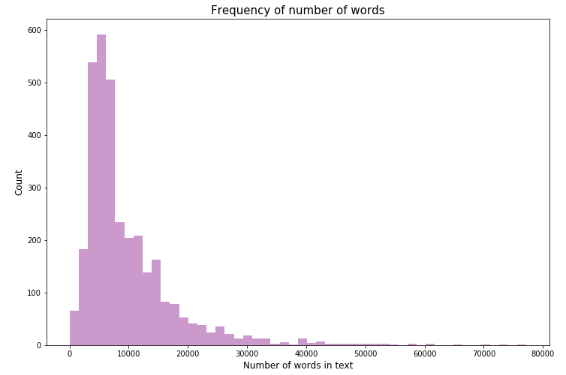
Training text and Test text

* ID
* Text – the clinical evidence used to classify the genetic mutation

**Frequency Distribution of Words:**

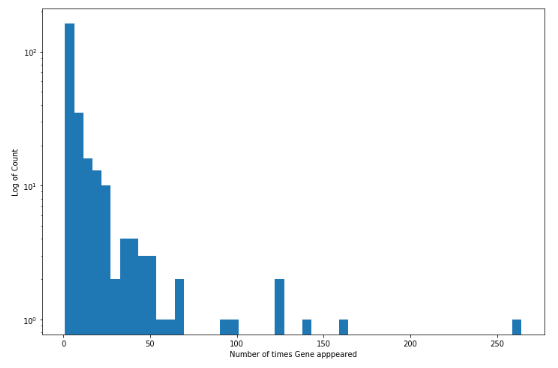
As we can see from the below graph, the text descriptions available in the training dataset have long lengths. The frequency distribution of number of words is right skewed. The distribution has peak around 5000-6000 words on an average. It means that text description has on an average 5500 words in the training data. Also, we have a few cases where the text description is huge, and the word count is more than 70,000.

It becomes very difficult for clinicians to go through such big textual descriptions and understand the gene and the variation. The probability of misclassifying the gene and the class goes up with the number of words in the text description. We can leverage the machine learning techniques to classify the classes, but as the textual data volume is large, it raises the concern of ‘curse of dimensionality’.



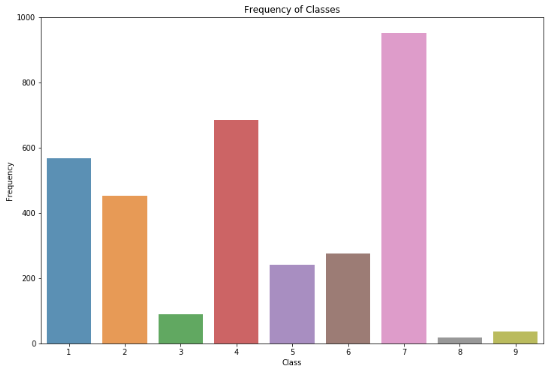
**The gene distribution (log count):**

Below graph shows the distribution of gene occurrence in the log scale (count). As we can see, most of the gene have low occurrence frequency. There are very few cases where the gene occurrence is more than 100 in the training dataset.



**Frequency of Classes:**

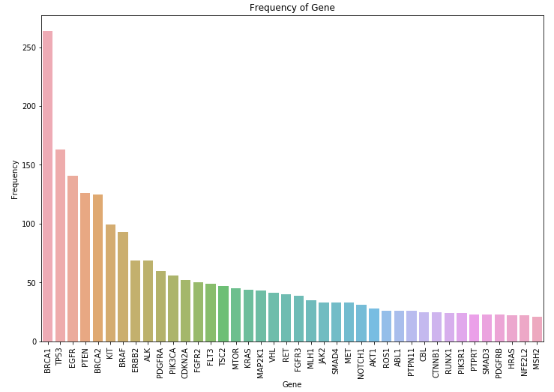
Below graph shows the distribution of the classes in the training dataset. We can see that the distribution of classes is not uniform in the training set. The classes 7 and 4 have higher frequency while classes 8 and 9 have very low frequency. The distribution of classes is important because the performance of the machine learning model depends upon the dependent variable.



The machine learning model becomes biased when the dependent class distribution is not uniform. We need to give extra weight to classes which have less frequency occurrence.

**Frequency Distribution of Gene:**

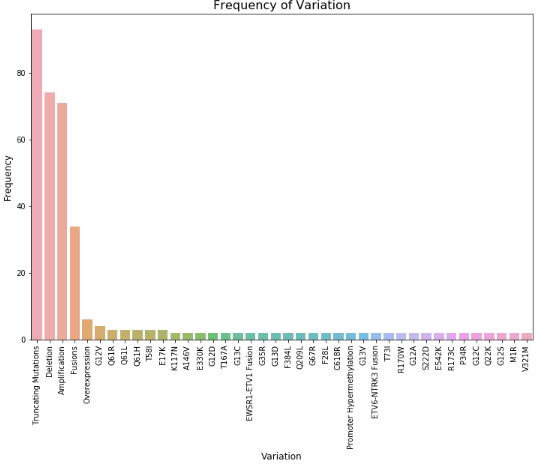
As we can see from the frequency distribution graph of gene, the distribution is right skewed. There are few gene categories having large occurrence in the training dataset while most of the gene types are not frequent. We have 264 distinct gene types in the dataset.



Predicting the gene category using text description is difficult because of the level of granularity in gene column.

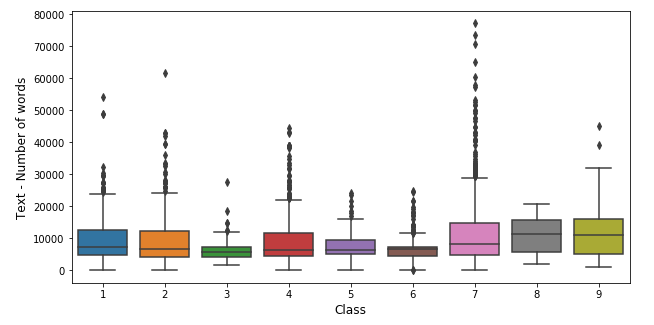
**Frequency Distribution of Variation:**

The frequency distribution of variation is also right skewed as we have seen in case of gene distribution. Also, the main variations in the gene are not because of the amino-acid changes. Deletion, Amplification, Fusions are the important reasons in variation category for mutation.



**Class vs Number of words Distribution:**

Below graph shows the boxplot for each of the classes and the spread is based on the number of words in the text. We can relate the given graph with a couple of previous graphs and understand the spread better. We can see that the class 7 has wide spread and there are many points above the top whisker. It shows that the distribution of count of words is not uniformly distributed.



1. **DATA PRE-PROCESSING**

Before starting with the modeling part, we did a couple of data cleaning steps to reduce dimensionality of our dataset

* **Stop word removal** - When data analysis needs to be data driven at the word level, the commonly occurring words (stop-words) should be removed. One can either create a long list of stop-words or one can use predefined language specific libraries
* **Stemming/Lemmatization** - Stemming describes the process of transforming a word into its root form. It is a rudimentary rule-based process of stripping the suffixes (“ing”, “ly”, “es”, “s” etc) from a word. In contrast to stemming, Lemmatization aims to obtain the canonical (grammatically correct) forms of the words, the so-called lemmas. Lemmatization is computationally more difficult and expensive than stemming

1. **MODELING**

We tried a number of natural language processing techniques to classify genetic mutations:

**Bag of words model**

Text data can also be quantified directly into numbers using several techniques. Two of the most common approaches are word count and tf-idf.

**COUNT BASED**

In word count, after creating the bag of words model which has all the important words as different features in the dataset and text documents as rows, the model counts the number of times a word has occurred in a particular document and substitute the word count as a data point.

**TF-IDF**

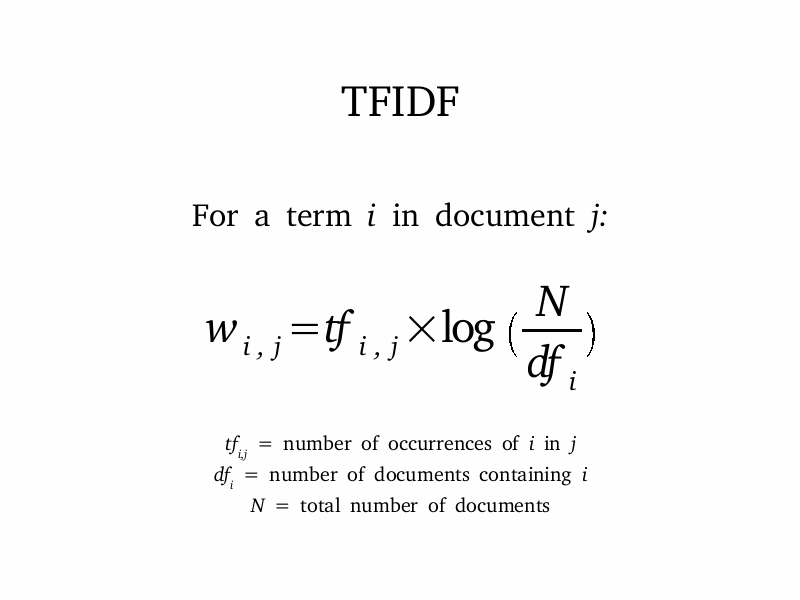
Rather than just counting we can find the tf-idf score of a word to rank its importance. It consists of a combination of two parts

Term Frequency (tf(t, d)) - A alternative approach to characterize text documents rather than binary values is the term frequency The term frequency is typically defined as the number of times a given term t (i.e., word or token) appears in a document d (this approach is sometimes also called raw frequency). In practice, the term frequency is often normalized by dividing the raw term frequency by the document length.

t: Raw term frequency (the count of term t in document d), d: The total number of terms in document d.

tf(t, d) = t/d

Inverse document frequency idf(t) - Another alternative for characterizing text documents. It can be understood as a weighted term frequency, which is especially useful if stop words have not been removed from the text corpus. The idf(t) approach assumes that the importance of a word is inversely proportional to how often it occurs across all documents.

idf(t): log\_e(Total number of documents / Number of documents with term t in it) 

**Word embeddings (Word2Vec)**

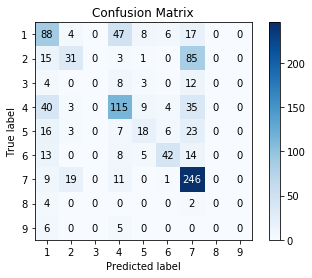
Word2Vec is a Word Embeddings model made by google. Count vectorizer and TF-IDF vectorizer relies on the word count in a sentence but they don’t preserve any contextual similarity between the words such as how the words are related to each other and together what information they can convey about the document.

The aim of word embedding is to redefine the high dimensional word features into low dimensional feature vectors. In word embeddings every word would be assigned a vector and randomly initialized or a one hot vector (only the word would have a bit of 1 rest all is o). After this we take a window size and iterate through our document. There are two models which comes into picture. The continuous bowl of words (we try to fit the neighboring words of the window to the central word) and skip-gram (we try to make the central word closer to the neighboring words). After this process we get trained vectors for each word which preserves syntactical information.

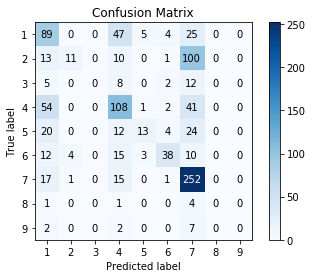
**Classification models**

After creating the bag of words model (count and tf-idf) and word embedding (vectors for each word), we input the output of these models in various classification algorithms such as Logistic regression, Random Forest, SVM (Linear/RBF), XGBoost and compared their performance on validation sets.

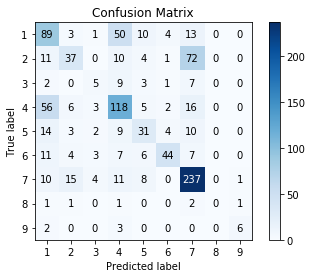
Logistic regression on full data was performed for each combination of count occurrence, tf-idf and word2vec and we noticed that performance of Logistic-word2vec is the best and the following result was observed.



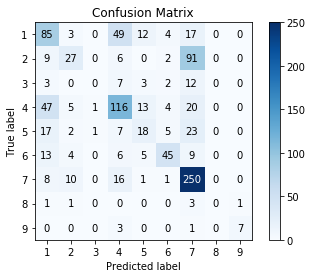
To check how the bootstrap model would work for our data, we proceeded with the following analysis. The Random forest technique creates multiple trees parallelly, using the bootstrapping technique, and the result is a combination of the outputs of all the trees generated. In our case, a classification model, the data is of categorical datatype. Hence the output is the classification which occurs the greatest number of times in all the trees produced by bootstrapping. Performance of Random forest-TF-IDF is the best and the following confusion matrix was observed



Then we created a boosted tree model (XGboost) to classify. Since boosted trees generate trees in a sequential manner, one following the other, based on the residuals of the preceding tree, we were hoping to see an improvement in performance when compared with the Logistic Regression model. We tried creating multiple boosted trees by varying the tuning parameters such as learning rate and number of splits per tree. Performance of XGboost-TF-IDF is the best and the following confusion matrix was observed



SVM with linear kernel was used to classify, as we observed that the cluster of classes of response variable are identified to be linearly separable when observations are plotted against principal components which explains majority of variance after performing PCA. Performance of SVM-TF-IDF is the best and the following confusion matrix was observed



To obtain the above optimal predictions we used two layers of optimization using random and grid search by optimizing the hyperparameters space where expected value of objective function is minimal.

1. **MODEL COMPARISON**

|  |  |  |
| --- | --- | --- |
| **Models** | **Log Loss** | **Accuracy** |
| **Logistic - count occurrence** | **1.65** | **48.2** |
| **Random Forest – count occurrence** | **1.44** | **50.5** |
| **Logistic - TF-IDF** | **1.51** | **46.3** |
| **Random Forest - TF-IDF** | **1.35** | **51.3** |
| **SVM (Linear) - TF-IDF** | **1.21** | **55.1** |
| **XGBoost - TF-IDF** | **1.23** | **56.9** |
| **Logistic - Word2Vec** | **1.32** | **54.2** |
| **SVM (Linear) - Word2Vec** | **1.29** | **54.4** |
| **XGBoost - Word2Vec** | **1.26** | **56.7** |
| **LSTM** | **1.51** | **50.1** |

We generated various models to compare which model gives us the best results. We created Logistic regression, Random Forest, SVM (Linear/RBF) and XGBoost models. By comparing the respective best models of every technique, we observed that log loss and Accuracy was best for XGBoost-TF-IDF model.

1. **CONCLUSION AND RECOMMENDATIONS**

Machine learning is a branch of artificial intelligence that employs a variety of statistical, probabilistic and optimization techniques that allows computers to “learn” from past examples and to detect hard-to-discern patterns from large, noisy or complex data sets. This capability is particularly well-suited to medical applications, especially those that depend on complex proteomic and genomic measurements. As a result, machine learning is frequently used in cancer diagnosis and detection. This latter approach is particularly interesting as it is part of a growing trend towards personalized, predictive medicine. According to a 2015 report, more than 800 medicines and vaccines to treat cancer were in trial.

Among the better designed and validated studies, it is clear that machine learning methods can be used to substantially (15–25%) improve the accuracy of predicting cancer susceptibility, recurrence and mortality. At a more fundamental level, it is also evident that machine learning is also helping to improve our basic understanding of cancer development and progression.

The table below shows how ML techniques have improved the performance.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Cancer Type | Clinical Endpoint | Machine Learning Algorithm | Benchmark | Improvement (%) | Training Data |
| bladder | recurrence | fuzzy logic | statistics | 16 | mixed |
| breast | survivability | decision tree | statistics | 4 | clinical |
| breast | susceptibility | SVM | random | 19 | genomic |
| breast | recurrence | ANN | statistics | 1 | clinical |
| breast | survivability | clustering | statistics | 0 | clinical |
| breast | recurrence | ANN | expert | 5 | mixed |
| breast | survivability | ANN | statistics | 1 | clinical |
| breast | recurrence | ANN | statistics | 23 | mixed |
| breast | survivability | ANN | expert | 5 | clinical |
| breast | recurrence | ANN | expert | 10 | clinical |
| colorectal | recurrence | ANN | statistics | 12 | clinical |
| colorectal | survivability | ANN | statistics | 9 | clinical |
| colorectal | recurrence | ANN | statistics | 9 | mixed |
| colorectal | survivability | ANN | expert | 11 | clinical |
| esophageal | survivability | ANN | statistics | 3 | clinical |
| liver | recurrence | ANN | statistics | 25 | genomic |
| lung | survivability | ANN | statistics | 9 | mixed |
| lymphoma | survivability | ANN | statistics | 22 | genomic |
| lymphoma | survivability | ANN | expert | 10 | mixed |
| head/neck | survivability | ANN | statistics | 11 | clinical |
| prostate | recurrence | ANN | statistics | 0 | clinical |
| prostate | recurrence | ANN | statistics | 16 | mixed |
| prostate | recurrence | ANN | statistics | 11 | mixed |
| prostate | recurrence | SVM | statistics | 6 | clinical |
| prostate | recurrence | ANN | statistics | 0 | clinical |
| prostate | recurrence | ANN | statistics | 0 | clinical |
| prostate | recurrence | ANN | statistics | 13 | clinical |
| prostate | recurrence | naïve Bayes | statistics | 1 | clinical |
| prostate | recurrence | ANN | statistics | 17 | clinical |
| skin | survivability | ANN | expert | 14 | clinical |
| skin | recurrence | ANN | expert | 27 | proteomic |
| skin | survivability | ANN | expert | 0 | clinical |
| stomach | recurrence | ANN | expert | 28 | clinical |

Today machine learning methods are being used in a wide range of applications ranging from detecting and classifying tumors via X-ray and CRT images to the classification of malignancies from proteomic and genomic (microarray) assays.

The use of machine learning in [preliminary (early-stage) drug discovery](http://blogs.royalsociety.org/in-verba/2016/10/05/machine-learning-in-the-pharmaceutical-industry/) has the potential for various uses, from initial screening of drug compounds to predicted success rate based on biological factors. This includes R&D discovery technologies like [next-generation sequencing](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3841808/).

There are many initiatives taken up by various companies in this field.

* In October 2016, IBM Watson Health announced [IBM Watson Genomics,](https://www.mskcc.org/ibm-watson-and-quest-diagnostics-launch-genomic-sequencing-service-using-data-msk) a partnership initiative with Quest Diagnostics, which aims to make strides in precision medicine by integrating cognitive computing and genomic tumor sequencing.
* Boston-based [biopharma company Berg](http://berghealth.com/pipeline/) is using AI to research and develop diagnostics and therapeutic treatments in multiple areas, including oncology. Current research projects underway include dosage trials for intravenous tumor treatment and detection and management of prostate cancer.
* [Microsoft’s Project Hanover](http://hanover.azurewebsites.net/) is using ML technologies in multiple initiatives, including a collaboration with [the Knight Cancer Institute](http://www.ohsu.edu/xd/health/services/cancer/) to develop AI technology for cancer precision treatment, with a current focus on developing an approach to personalize drug combinations for [Acute Myeloid Leukemia (AML)](https://www.lls.org/who-we-are/leading-the-charge-against-aml).
* Over the next decade, increased use of [micro biosensors and devices](http://www.mckinsey.com/industries/pharmaceuticals-and-medical-products/our-insights/how-big-data-can-revolutionize-pharmaceutical-r-and-d), as well as mobile apps with more sophisticated health-measurement and remote monitoring capabilities, will provide another deluge of data that can be used to help facilitate R&D and treatment efficacy.

A futuristic application would be to embed the X-ray and CRT images data with EMR data to detect and classify various kinds of tumors and help in early prediction.

1. **REFRENCES**

<http://gco.iarc.fr/>

https://www.cancer.net/navigating-cancer-care/cancer-basics/genetics/genetics-cancer

https://people.csail.mit.edu/regina/talks/CNLP.pdf

<http://ascopubs.org/doi/abs/10.1200/JCO.2017.35.8_suppl.232>

<https://www.sciencedirect.com/science/article/pii/S2001037014000464>

<https://www.techemergence.com/ai-in-pharma-and-biomedicine/>

[https://www.sciencedirect.com/science/article/pii/S2001037014000464#bb0015](https://www.sciencedirect.com/science/article/pii/S2001037014000464)