# Integration of Data Mining techniques to Identify factors involved in Cancer Therapeutics

Dhruv Kaushik
MT18037

Subhani Shaik
MT18117

Akshat Joshi
MT18064

## Abstract

**A lot of information related to cancer treatment is freely available online in the form of US cancer patents. The aim of this project is to apply text mining techniques on them in order to extract useful information in the form of features like Cancer type, Chemotherapeutic agent used, Oncolytic Virus used, cell lines, etc. The extracted information was put in excel sheet and was cross verified manually.**

## I. Introduction

A lot of data related to Cancer is available online in the form of documents like Cancer related patents [1], Free-text Medical Reports [2] , electronic medical records of patients [3][4], etc. Thus there is a huge scope of using Text mining techniques for Cancer related Knowledge discovery from them. The application of this is for extracting information from these documents, one need not go through all the documents. Rather using text mining, the specific features related information can be directly extracted from them.

In this project, we have also developed a system by combining some existing text mining (n-grams, lemmatization, stemming, etc.) and pattern matching (token matching, Longest common subsequence) that can perform knowledge discovery on US cancer patents. The developed system has been designed to work best on the template followed by the US patents. Most of the developed system uses text mining, though an intermediate part of it - reading data from graphical image can only be done manually. There exists no algorithm which can efficiently extract data from such graphical images present in the patents.

## II. Related work

Previously, the knowledge discovery and information retrieval work has been done related to patents. Almost all of them are retrieving the information using classical NLP (Natural Language Processing) techniques like text summarization, document clustering by topic, text segmentation, topic modelling etc.
Ming Huang et.al.[5] analyzed more than 5 million US patent documents between 1995 and 2017, using summary statistics and dynamic topic modeling. More specifically, it investigated the disease coverage and latent topics in patent documents over time.

Frumkin et.al.[6] worked on the USPTO Cancer Moonshot Patent Data that contains detailed information on published patent applications and granted patents relevant to cancer research and development (R&D). They generate the dataset using USPTO examiner tools to execute a series of queries designed to identify cancer specific patents and patent applications. The final dataset consists of roughly 270,000 patent documents spanning the 1976 to 2016 period.

Martin et.al.[7] used Bibliometric methodologies and technology are used to investigate publications/patents, their contents and relationships on 29237 items published and 16053 patents from 1997 to 2016 including “chemotherapy for breast cancer” were retrieved. Tiwari[8] proposed a content based patent image retrieval system based on Affine-SIFT technique

## III. Dataset

### A. Cancer Patents

The Cancer Patents dataset that we have used in this project is a set of 642 US Cancer patents obtained from Google Scholar by giving query 'Oncolytic Virus' with applying filter like Patent Office : US , Status : Grant, Type : Patent. The web link from which we crawled the documents is referenced [1].

### B. Cell Lines

We collected cell lines related information from multiple standard sources [9], [10] and [11] to create a dictionary of 2514 cell lines, containing information like cell line name, organ and species to which it belongs. This cell line dictionary was used later to identify cell line appearing in text by string matching algo.

### C. Cancers list

We made a list of various types of cancers found in humans and others from multiple online sources [12], [13] and [14]. We made a list of 176 cancer types after combining information from these sources. This list was used later to identify cancer type appearing in the text by string matching algo.

## IV. Methodology

In our dataset, most of the patents are very old and not follow any particular pattern which is difficult in retrieving information. Our dataset also has information that reside in graphs/plots which is very important. Stepwise methods followed are described below:

1. First step is deciding the features that we need to extract from the documents. We selected these features: Cancer name, Experiment related information like: In-vitro or In-vivo, Chemotherapeutic agent used, Oncolytic Virus used, their quantities, cell line used (in-vitro), animal name (in-vivo).
2. Our dataset has information that reside in graphs/plots which is very important for qualitative mining of the information from the patent. To get such files, we used Google Tesseract library which is an OCR(Optical Character Recognition) library to filter what are the files have graphs related information about keywords like Cell survival, Tumor size etc.
3. For the files obtained in step 2, we manually annotated the information from those graphs that contains our key feature information like survival %, tumor size in $mm^3$, associated chemo agent or virus name.
4. By using the graph figure no, we identified that subsection (in the text) of document in which details about this graph is provided. As this graph represents the result of an experiment done, thus this associate subsection is also related to the experiment. So, this subsection is extracted out for further mining.
5. Now, to determine cell line and cancer type, we segment the selected text into unigrams, bigrams and trigrams. Then, we use the cell line dictionary and Cancer type list that we created as described in Dataset section for n-grams matching. Applying frequency threshold on selected tokens frequency, we get the probable cell line used and Cancer type specified in the text.
6. By using the information retrieved from the graph and applying pattern matching (n-grams matching, LCS) on the numerical values coming closer to units like mg, Molar, etc. we get the quantities of Chemo agent/ Virus strain from the selected text.
7. The feature details were put in excel sheet 'final_output.xlsx'.

## V. Results and Analysis

For filter by Google Tesseract we used key-words like "Cell-Survival %", "Tumor size", "Post days after injection", "Survival percent", "Days after injection" etc. It filtered 92

files from our dataset. Some of these files were wrongly selected due to the presence of any of the key-terms in any non-useful graph. These files were directly rejected if no useful graph plot is found in them. The data from graphs: figure no, cell survival % or tumor size mm3, no. of days, virus/chemo agent are noted from it and put in excel file 'annotated_graph_data.xlxs'. This contains 287 entries after annotation. After performing the text mining, the final output sheet 'final_output.xlsx' contains information of 41 documents.

## VI. Individual Contribution

| Steps in the process | Group Member |
|---|---|
| 1. **Literature survey** on similar work | Akshat Joshi |
| 2. **Obtaining Datasets**: Retrieving US Cancer Patents; Combining Cancer types and Cell lines info from multiple sources. | Dhruv Kaushik |
| 3. **Filter out useful patents:** Use Google Tesseract to filter out patents containing useful graphs. | Subhani Shaik |
| 4. **Manually annotated graphs** in the 92 docs obtained above in step 3. | Akshat Joshi |
| 5. **Text mining** on the filtered 92 docs to automatically retrieve the feature values from the filtered 92 patents. | Dhruv Kaushik, Subhani Shaik |

## VII. Conclusion

Our implementation is used to discover knowledge from US cancer patents and retrieve the features like Chemotherapeutic agents, In-Vitro, In-Vivo, Tumor size etc. We performed the qualitative and quantitative analysis on the results and it is retrieving information well. It has some limitations like as extracting information from distorted images is a difficult task, we need annotated data of images that will be fed to our program to retrieve useful feature information.

## VIII. References


[1] US Cancer Patents source link: https://patents.google.com/?q=oncolytic+virus&country=US&status=GRANT&language=ENGLISH&type=PATENT&num=100

[2] Collection of Cancer Stage Data by Classifying Free-text Medical Reports by Iain A. McCowan, Darren C. Moore, Anthony N. Nguyen, Rayleen V. Bowman, Belinda E. Clarke, Edwina E. Duhig, Mary-Jane Fry, American Medical Informatics Association, 2007.

[3] The Health IT Dashboard: https://dashboard.healthit.gov/datadashboard/data.php

[4] ORBDA: An openEHR benchmark dataset for performance assessment of electronic health record servers by Douglas Teodoro, Erik Sundvall, Mario João Junior, Patrick Ruch, Sergio Miranda Freire, PLoS ONE, January 2018.

[5] Technological Innovations in Disease Management: Text Mining US Patent Data From 1995 to 2017., Ming Huang et.al.

[6]. Cancer Moonshot Patent Data., Frumkin et.al. https://www.uspto.gov/learning-and-resources/electronic-data-products/cancer-moonshot-patent-data

[7]. Patents and publications trends in breast cancer chemotherapy.,Martin Perez Santos et.al.

[8]. Tiwari, A. and V. Bansal. Patseek: Content based image retrieval system for patent database. in Proceedings of International Conference on Electronic Business, Beijing, China; 2004.

[9] ATCC Cell Line documents available at: https://www.atcc.org/en/Landing_Pages/Cancer_and_Normal_Cell_Lines_by_Tissue_Type.aspx

[10] NCI-60 Cell Line documents available at: https://dtp.cancer.gov/discovery_development/nci-60/

[11] CCLE Cell Line documents available at: https://portals.broadinstitute.org/ccle/data

[12] A Cancer types source: https://www.cancerresearchuk.org/about-cancer/type

[13] A Cancer types source: https://www.cancer.gov/types

[14] A Cancer types source: https://www.cancer.net/cancer-types

[15] Google Tesseract: https://opensource.google/projects/tesseract