EDA Presentation and Proposed Modeling Technique

Group members and email:

- Disha Lamba: jb.dishalamba@gmail.com
- Kiarash Rastegar: krastegar0@gmail.com
- Somasundaram Palaniappan: plsaran97@gmail.com
- Mercy Oyekanmi: creativeart19@icloud.com

A Growing Problem

There are various cancers that exists today, however, there is a lack in the number of techniques out there that can aid Biomedical Scientists in extracting the vital information that they need. By employing a deep learning NLP model, Biomedical Scientist will be able to get the relevant information of biological markers that are needed in detecting specific cancers.

IntilLP's Main Objective

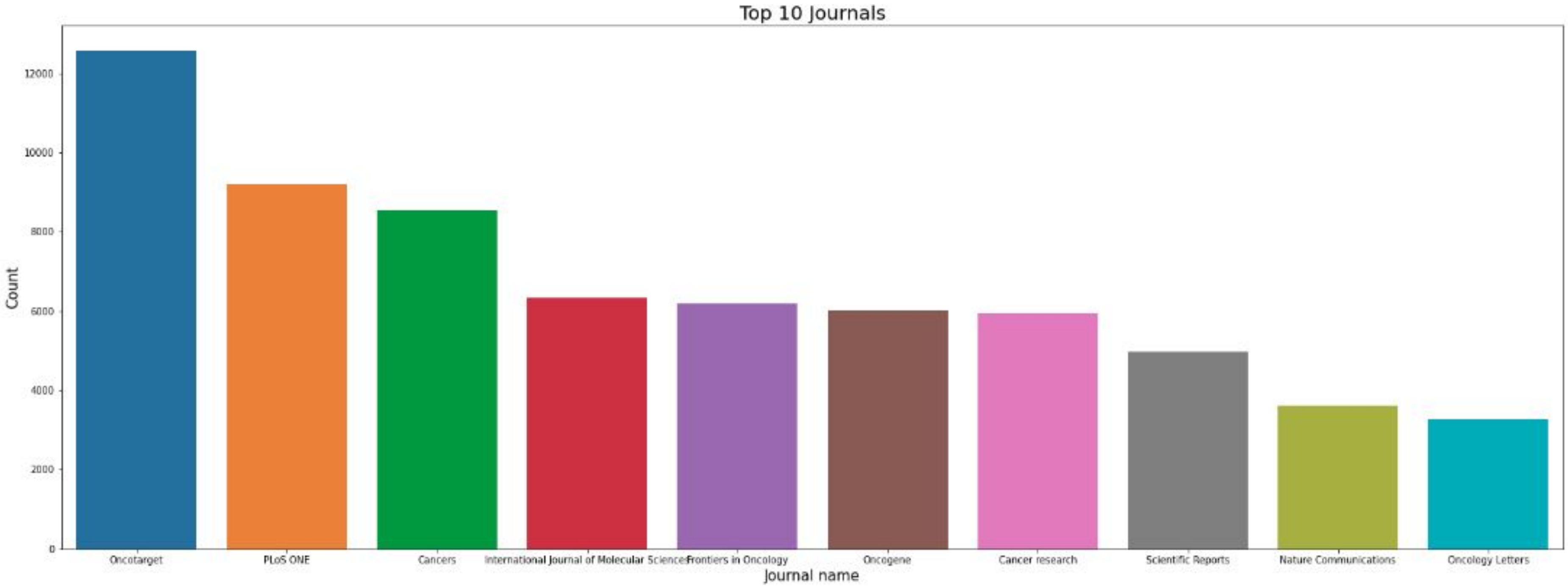
IntiLP aims to utilize deep learning NLP techniques to gain insight into the biological markers that is observed in certain cancers.

Our project CancerMine focused on three different cancers namely: lung, colon and pancreatic cancers.

Dataset

- Data Provider
    - Pubmed
    - The link to the original data used can be found at https://zenodo.org/record/6811941#.YySztexBxb-
- Dataset
    - The main data columns used in the EDA analysis are as follows:
        - Journals
        - Sentence
        - Gene types
        - Cancer Normalized

Top Journals

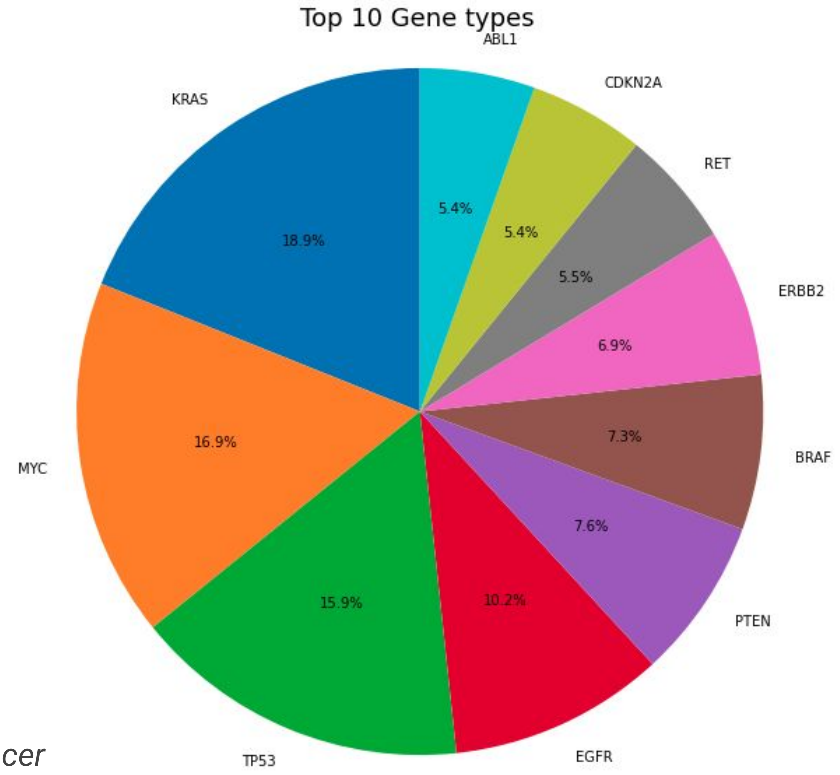


- Looking at the distributions of journals in our data set, we seem to have an even spread of journals that our data is coming from
- Oncotarget seems to be the largest contributor to this specific dataset
    - This is perfect because we want to get text data focusing specifically on cancer diseases

Top Genes

- Investigating the top genes in our data set we see that 3 cancers make up over 50% of our dataset
    - KRAS
    - MYC
    - TP53
- KRAS has been seen in as a drug target in all three cancers
- MYC is an oncogene that is actively expressed during cancers
- TP53 is the gene that encodes P53 which is a tumor suppressor

All three genes seem to be conserved across multiple different cancer types

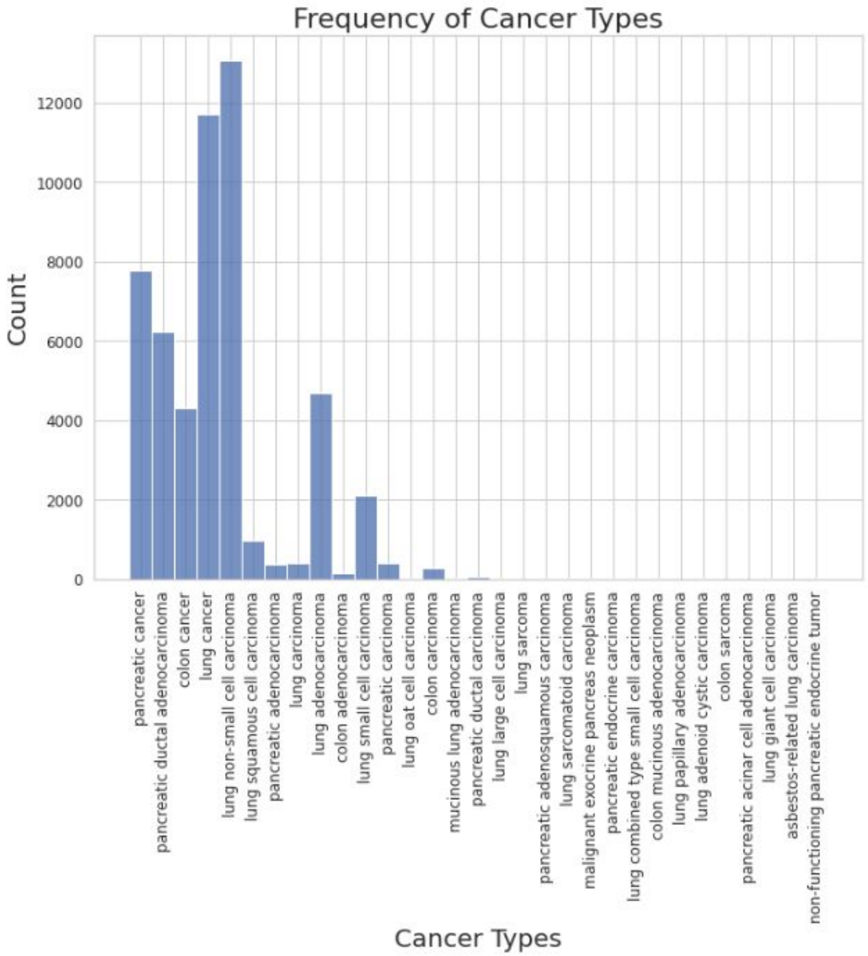


We can clearly see that KRAS, MYC, TP53 are the top 3 citied genes

Frequency of Cancer Types

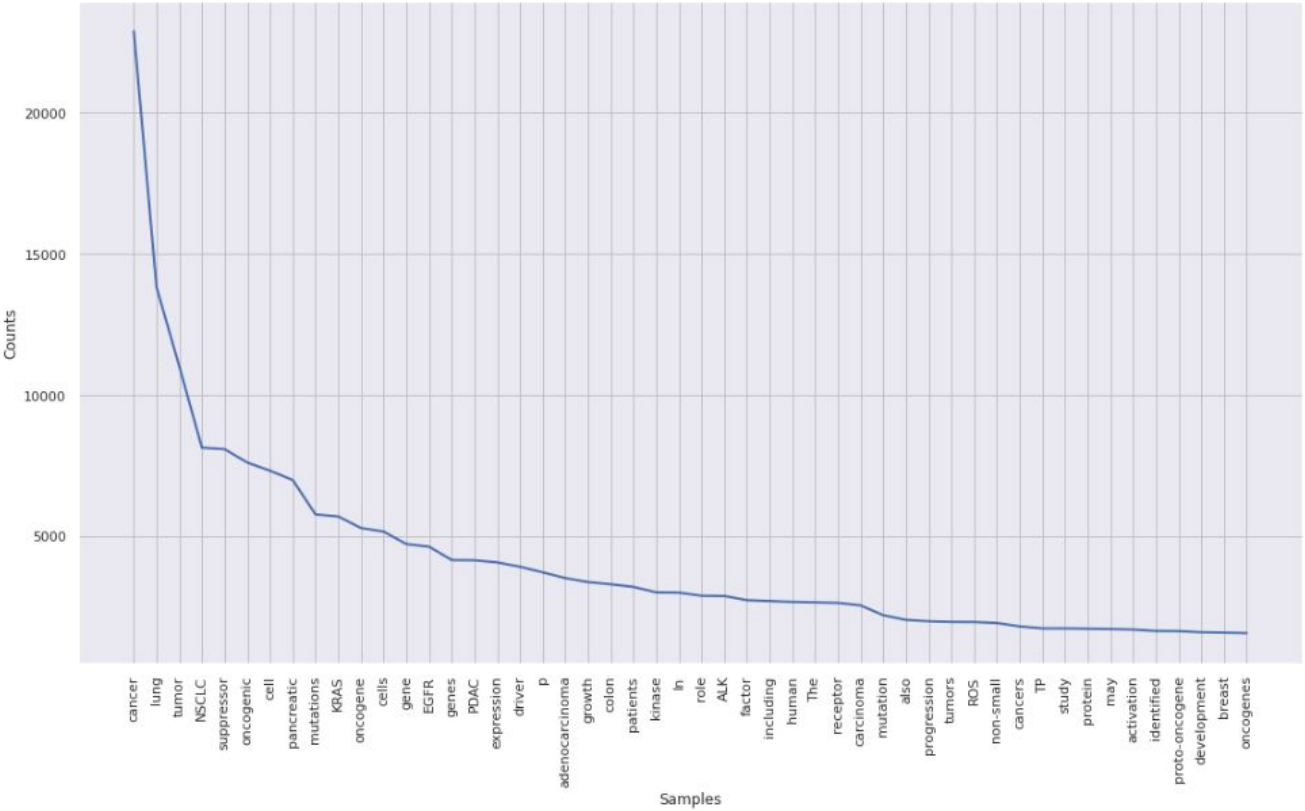
- We filtered the cancer_mine dataset by only keeping entries that contained the “lung”, “panc”, “colon” in the cancer normalized column

- We can see that our dataset seems skewed towards different types of lung cancers
  - Both lung non-small cell carcinoma and lung squamous cell carcinoma are our most frequent types of cancers in this dataset



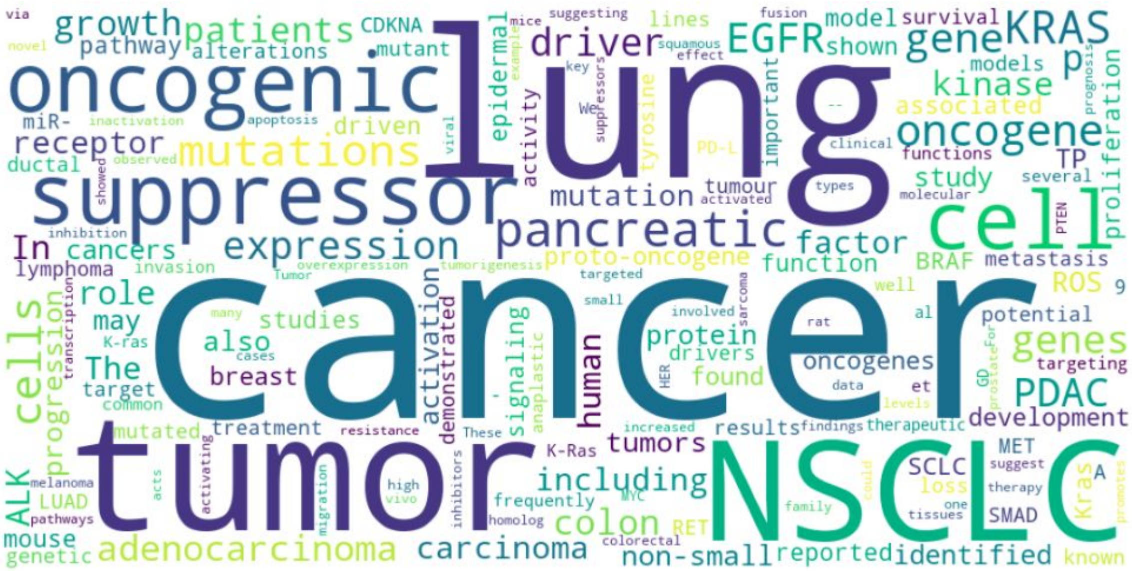
Word Frequency

- After removing all the stop words and creating tokens, we plotted the frequency of the tokens
- Obviously are most common word in this dataset is cancer.
    - Followed by lung, tumor and NSCLC (which stands for non-small cell lung cancer
- This does give us confidence that the stop word removal and tokenization was successful



Word Cloud

- Looking at our word cloud we again see that cancer is our most common word
- All of the extremely large words in the word cloud, seem to be a little too obvious and do not really provide any new insight into specific cancers
- Looking at some of the small words we do see some uncommon words such as SMAD and ROS
- Upon further investigation
  - We see that SMAD is a group of proteins that are used in signal transduction of TGF-B, which are a superfamily of proteins used to regulate cell growth and development
- These are some examples of key terms that we should focus on and will develop a method to highlight in our final model


cancer
lung
tumor
NSCLC
oncogenic
suppressor
cell
pancreatic
oncogene
mutations
expression
growth
patients
driver
EGFR
KRAS
gene
kinase
receptor
mutation
cells
role
adenocarcinoma
carcinoma
colon
including
PDAC
genes
factor
function
proto-oncogene
human
signaling
activation
progression
pathway
alterations
CDKNA
mutant
epidermal
activity
tyrosine
important
survival
model
shown
models
associated
study
proliferation
metastasis
BRAF
ROS
potential
targeting
oncogenes
development
results
identified
reported
non-small
known
tumors
SCLC
SMAD
MET
loss
therapy
suggest
Kras
breast
studies
also
may
The
target
common
mutated
treatment
ALK
LUAD
mouse
genetic
pathways
melanoma
lymphoma
In
cancers
invasion
ductal
miR-
driven
apoptosis
inactivation
observed
inhibition
tumorigenesis
overexpression
Tumor
protein
drivers
found
data
et
al
tumour
activated
mutation
clinical
functions
TP
several
molecular
types
PTEN
fusion
lines
mice
suggesting
via
novel
squamous
effect
example
key
We
suppressors
viral
PD-L
p
K-ras
K-Ras
MYC
HER
anaplastic
demonstrated
frequently
homolog
colorectal
RET
family
tissues
one
A
promotes
prostate
GD
could
levels
well
rat
involved
small
sarcoma
targeted
increased
findings
therapeutic
resistance
These
high
vivo
migration
inhibitors
activating
acts
transcription
showed
many
cases