COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Frederick National Laboratory
for Cancer Research

sponsored by the National Cancer Institute

Master’s Research and Projects:
FNLCR and Columbia University Partnership

Master's Research and Projects:
FNLCR and Columbia University Partnership

2:00 PM: Introduction Prof. Michael Robbins

2:10 PM: Opening remarks Dr. Ethan Dmitrovsky

2:20 PM: Student Presentations

- Cloud Deployment, Optimization Strategies for Teaching, Training and Collaborative Reproducible Research
- Survey to Identify Emerging Infectious Disease Datasets for Machine Learning
- Survey to Identify Cancer Datasets for Machine Learning
- Q & A

3:20 PM: Closing remarks Dr. Eric Stahlberg

Project Team

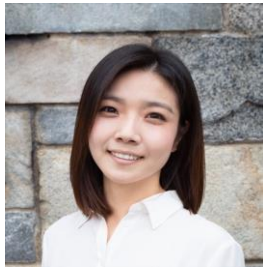
Mahitha Kotipalli

Jim Hu

Niranjana Moleyar

Malin Ortenblad

Kerry Hu

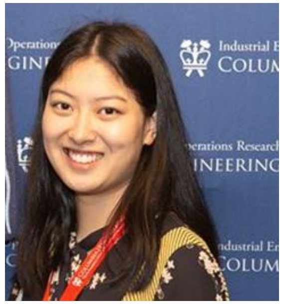


Jie Chen

Mengyao He

Om Vaghasia

Panagiotis Misirlis



Jiaxi Zhou

Xinyao Wang

Qinwei Zhang

Yue Hu

Zihui Zhou

Naomi Ohashi, MPA, PMP, ITIL
Biomedical Informatics and Data Science (BIDS)
Frederick National Laboratory for Cancer Research

Ravichandran Sarangan, PhD, PMP
Biomedical Informatics and Data Science (BIDS)
Frederick National Laboratory for Cancer Research

Michael Robbins
Professor
Columbia University

Nicole Soder
TA, Project Manager
Columbia University

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Frederick National Laboratory
for Cancer Research

sponsored by the National Cancer Institute

Survey to Identify Top-10 Commonly Occurring Cancer Datasets for ML

Team Members: Jim Hu, Kerry Hu, Mahitha Kotipalli, Malin Ortenblad, Niranjana Moleyar
Affiliation: IEOR Department, Columbia University
Sep. 8, 2020

DEPARTMENT OF HEALTH AND HUMAN SERVICES • National Institutes of Health • National Cancer Institute

Frederick National Laboratory is a Federally Funded Research and Development Center operated by Leidos Biomedical Research, Inc., for the National Cancer Institute

Table of content

Project Team

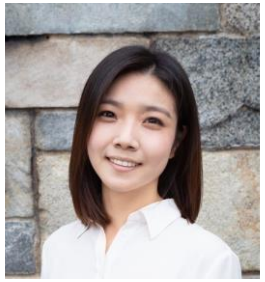
Mahitha Kotipalli
Columbia University, MSMSE
Project 1: Bronchus, Lung and Prostate Cancer
Prior experience in healthcare consulting. Worked as Business Analytics MBA intern with Bayer, Business Insights team this summer

Malin Ortenblad
Columbia University, MSBA
Project 1: Skin and Breast Cancer
Previous work experience in healthcare consulting, spent the summer at Bayer in their Business Insights department as a Business Analytics MBA.

Kerry Hu
Columbia University, MSOR
Project 1: Spinal Cord, Cranial Nerves and Ovary
Prior experience in financial services.Interned at Credit Suisse as quantitative researcher this summer

Jim Hu
Columbia University, MSBA
Project 1: Bronchus, Lung and Prostate Cancer
Previous Data Science Researcher at Point72, Business Analyst at Bayer Women's Health, Business Insights

Niranjana Moleyar
Columbia University, MSMSE
Kidney and Colon Cancer
Prior experience in Recommendation Systems and Convolutional Neural Networks for image comparison.
Worked with Rent the Runway to optimise their inventory



Jie Chen
Columbia University, MSOR
Project 1: Hematopoietic and reticuloendothelial (bone marrow) and Pancreas
Prior experience in financial service.
Interned at Bank of America as GBAM associate intern.

Naomi Ohashi, MPA, PMP, ITIL
Biomedical Informatics and Data Science (BIDS)
Frederick National Laboratory for Cancer Research

Ravichandran Sarangan, PhD, PMP
Biomedical Informatics and Data Science (BIDS)
Frederick National Laboratory for Cancer Research

Michael Robbins
Professor
Columbia University

Nicole Soder
TA, Project Manager
Columbia University

Project Overview: What Human Cancer Datasets (Biomolecular/Drug/Phenotype) are Available for Machine-Learning?

Objective

- Using PubMed to identify available datasets for machine learning in the oncology space, focusing on;
  - Breast, Skin, Lung, Bronchus, Prostate, Kidney, Colon, and Spinal Cord Cancer

Process

- Published work summarized in excel trackers with links to publication, abstract, title, dataset, and methodology

- Results available at the following GitHub repositories:
  - Skin and breast cancer: https://github.com/Ortenblad/ColUniv-FNL-BIDS-Project1_BreastandSkinCancer
  - Bronchus, Lung and Prostate cancer: https://github.com/Jim-Hu/ColUniv-FNL-BIDS-Project1_LungBronchusProstate
  - Kidney and Colon Cancer: https://github.com/NiranjanaMoleyar/FNL-BIDS-Project1-KidneyAndColonCancer
  - Spinal Cord, Cranial Nerves and Ovary: https://github.com/foevee/FNL-BIDS-Project1-SpinalCord_CranialNerves_Ovary

Table of content

In addition to searching for papers, many authors were contacted and bibliographies were used to identify additional research and data

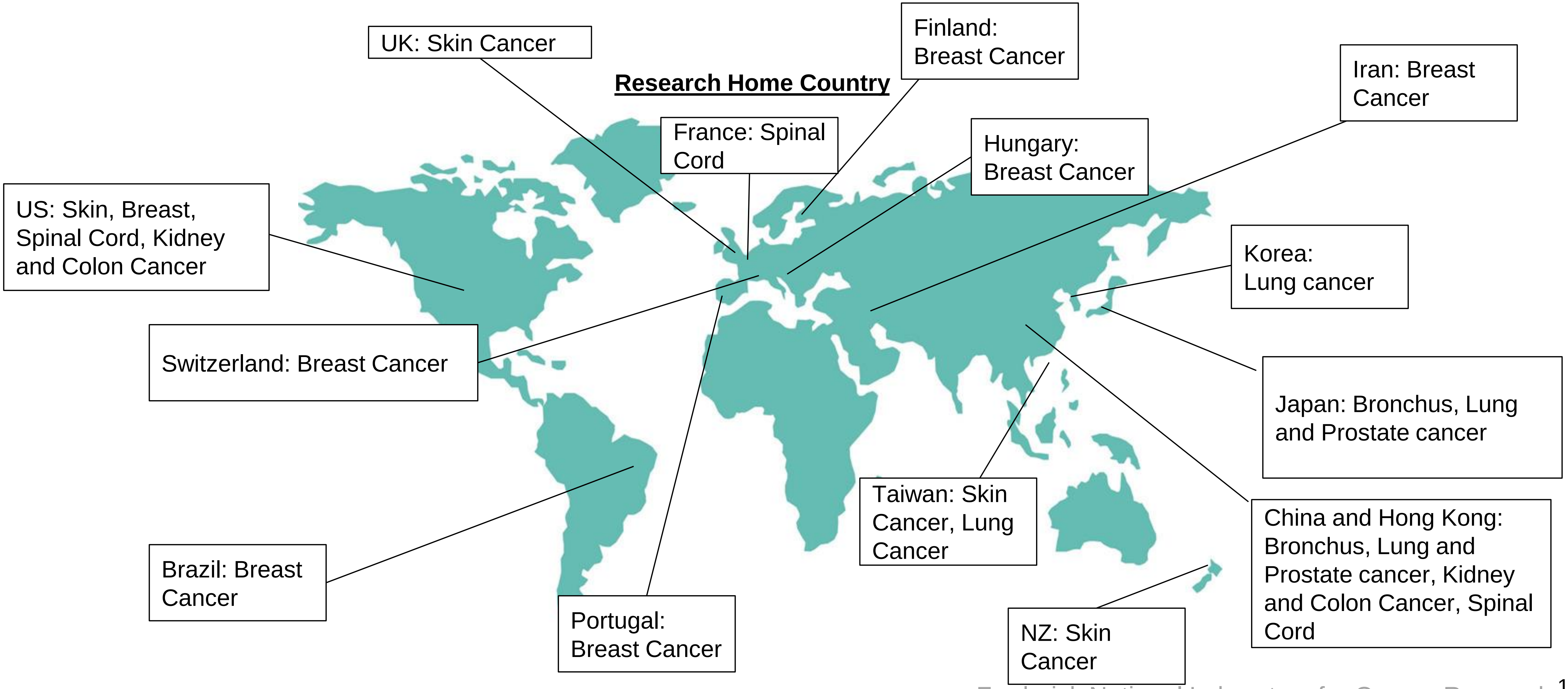
Search Terms

- Pubmed search engine link: https://pubmed.ncbi.nlm.nih.gov/
- Skin and Breast:
  - Machine Learning Skin Cancer
  - Machine Learning Breast Cancer
- Bronchus, Lung and Prostate Cancer:
  - Machine Learning Bronchus Cancer
  - Machine Learning / Deep Learning Lung Cancer
  - Machine Learning Prostate Cancer
- Kidney and Colon Cancer:
  - Kidney Cancer Machine Learning / Artificial Intelligence
  - Colon Cancer Machine Learning / Artificial Intelligence
- Spinal Cord, Cranial Nerves and Ovary:
  - Spinal Cord and Machine Learning / Neural Networks
  - Cranial Nerves and Machine Learning / Neural Networks
  - Ovary and Machine Learning / Neural Networks

Additional Steps

- Identify additional publications from footnotes and bibliography of papers
- Emailed authors requesting dataset (response rate ~5%)
- Looked up the identified papers at the Plos One journal since they provide data sources frequently

Based on the papers we gathered, research is conducted globally



Issues with gathering datasets leveraged in publications

Common Data Sources

- Most of the research is used typically comes from the following sources
  a. Publicly available datasets (e.g. on the website of **National Cancer Institute)**
  b. Datasets published on the website of different **universities' cancer research groups**
  c. Datasets available from competitions
  d. Data gathered by the researcher for the specific paper (hardest to acquire)

Issues faced to access the data

- Data owned by universities and other research groups is often difficult to access because
  a. Regulations in the country where research is carried out
  b. Requirement of **contract** between the parties to share data
  c. Some data has to be **deleted** after five years (limiting data availability to research published in last five years)
- Low author response rates

Project team has summarized 221 publications, and identified 79 datasets across 10 cancer types

Skin

Breast

Kidney

Colon

Spinal Cord

Cranial Nerves

Ovary

Bronchus

Lung

Prostate Gland

221 published ML Papers

79 Available Datasets

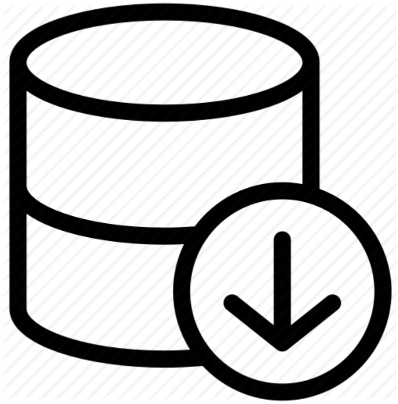


Table of content

For breast and skin cancer the most cost common use of ML was for classification and diagnosis purposes

Breast Cancer

Use cases and objectives for ML research:

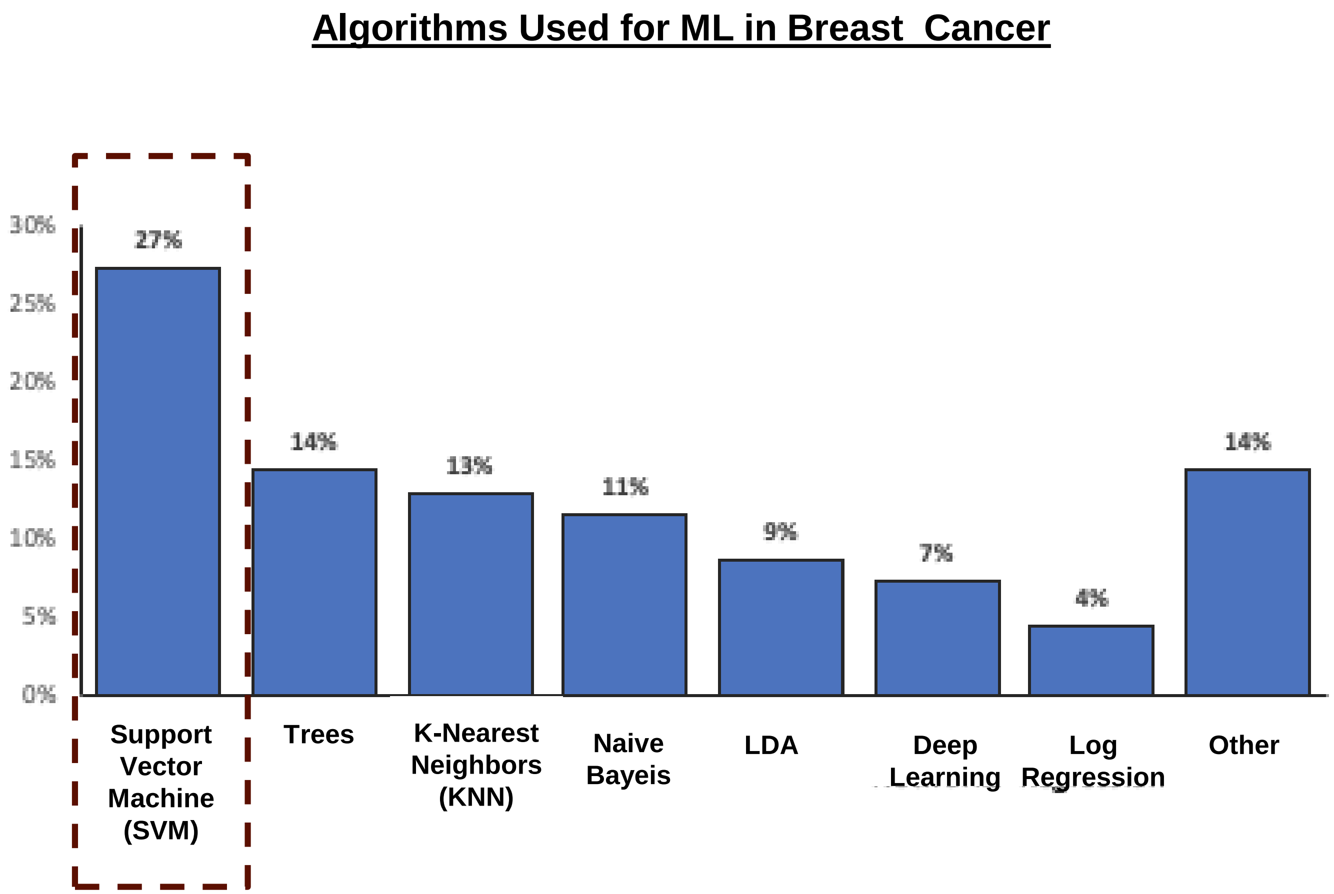
- Classification (most common):
  - Create machine to diagnose breast cancer
  - Outlining machine learning algorithms in use to detect breast cancer in mammogram images
  - Automate the method to detect cancer using ultrasounds
  - Comparison of different algorithms in diagnosing breast cancer
  - Improving diagnosis accuracy
- Patient Journey:
  - Predicting regression patterns in breast cancer patients
  - Predicting Tumor growth
  - Symptom analysis
- Sample: 50

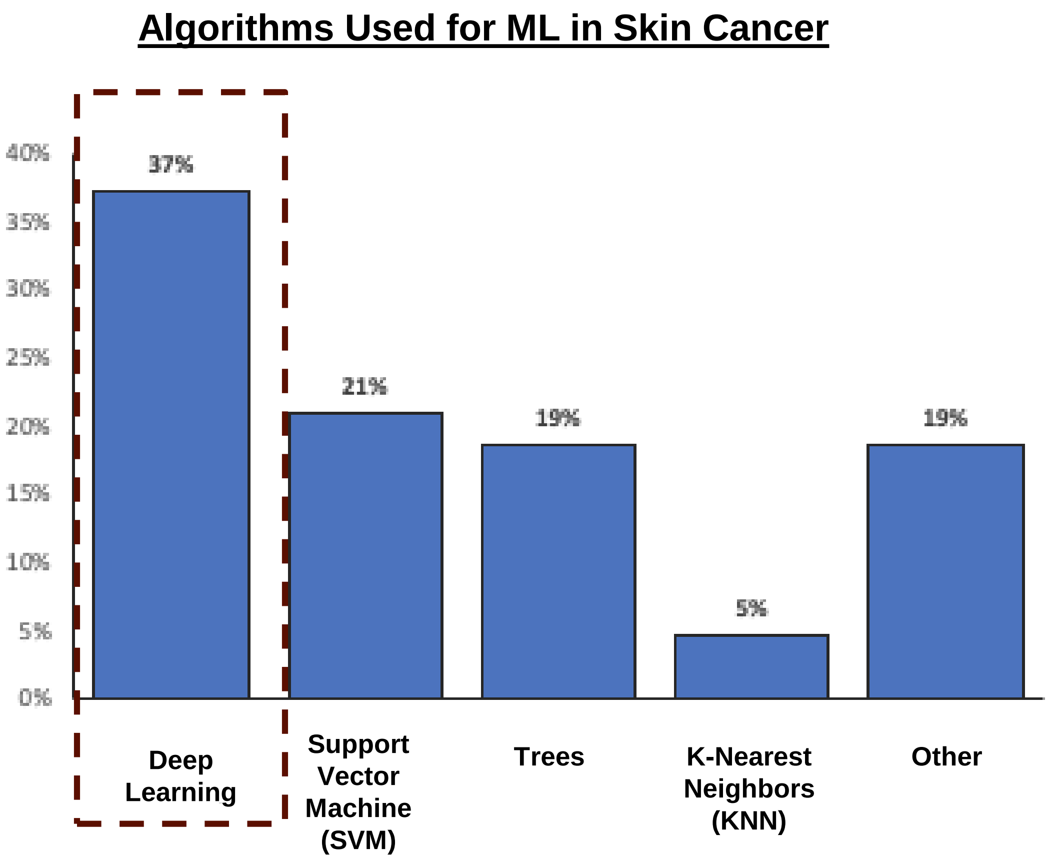
Skin Cancer

Use cases and objectives for ML research:

- Classification:
  - Comparison of AI and dermatologist in identifying skin cancer
  - Development and testing of 3D image scanning
  - Evaluation of image quality on diagnosis accuracy
  - Skin cancer diagnosis: ML, dermatologists, vision enhancing tools
  - Clinical feature selection and discovery of new biomarkers
  - Differentiative skin cancer disease classification
  - Smartphone app classifier accuracy
- Patient Journey:
  - Patient pathway utilization: Survivorship care plans
- Sample:39

The most common algorithms to use in breast and skin cancer research are SVM and Neural Networks, respectively





24 Datasets were identified from Breast Cancer Machine learning Publications

	Dataset	Link		Dataset	Link
1	GSE59198	https://www.omicsdi.org/dataset/geo/GSE59198	13	Mouse-Mammary	https://wiki.cancerimagingarchive.net/display/Public/Mouse-Mammary
2	Research Patient Data Registry	https://rc.partners.org/about/who-we-are-risc/research-patient-data-registry	14	TCGA-BRCA	https://portal.gdc.cancer.gov/projects/TCGA-BRCA
3	MIAS	https://www.kaggle.com/kmader/mias-mammography	15	QIN Breast DCE-MRI	https://wiki.cancerimagingarchive.net/display/Public/QIN+Breast+DCE-MRI
4	DDSM	https://www.kaggle.com/skooch/ddsm-mammography	16	BREAST-DIAGNOSIS	NA
5	CBIS-DDSm	https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM	17	RIDER Breast MRI	https://wiki.cancerimagingarchive.net/display/Public/RIDER+Breast+MRI
6	ISPY1	https://wiki.cancerimagingarchive.net/display/Public/ISPY1	18	BCDR	https://bcdr.eu/information/about
7	Breast-MRI-NACT-Pilot	https://wiki.cancerimagingarchive.net/display/Public/Breast-MRI-NACT-Pilot	19	TCGA-BRCA	https://portal.gdc.cancer.gov/projects/TCGA-BRCA
8	QIN-Breast	https://wiki.cancerimagingarchive.net/display/Public/QIN-Breast	20	BreakHis	https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/
9	Wisconsin Breast Cancer (original) datasets	https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)	21	Inbreast	https://github.com/wentaozhu/deep-mil-for-whole-mammogram-classification/issues/12
10	ICAR 2018	https://iciar2018-challenge.grand-challenge.org/Dataset/	22	Finprog Cancer Dataset	finprog.org/data_set2.asp
11	GWAS	https://cbi.ornl.gov/data	23	Metabolomics data of the 162 metabolites	https://www-sciencedirect-com.ezproxy.cul.columbia.edu/science/article/pii/S1874391913005113?via%3Dihub
12	271 breast cancer samples	https://www-ncbi-nlm-nih-gov.ezproxy.cul.columbia.edu/geo/	24	Mini-MIAS	http://peipa.essex.ac.uk/info/mias.html

20 Datasets were identified from Skin Cancer Machine learning Publications

	Dataset	Link		Dataset	Link
1	Innternational Skin Image Collaboration	https://www.isic-archive.com/#!/topWithHeader/wideContentTop/main	11	COSMIC	https://cancer.sanger.ac.uk/cosmic
2	NHIRD	https://nhird.nhri.org.tw/en/	12	DermNEt	https://dermnetnz.org/
3	SCP Survey Data	http://www.scpwiki.com/2018-survey-results	13	ImageNet	http://www.image-net.org/
4	digital dermoscopic database	https://www.sec.gov/Archives/edgar/data/1051514/000095012305009834/y09078a3sv1za.htm	14	TCGA-SKCM	https://portal.gdc.cancer.gov/projects/TCGA-SKCM
5	ISIC Challenge dataset	https://challenge2020.isic-archive.com/	15	Asan dataset	https://figshare.com/articles/Asan_and_Hallym_Dataset_Thumbnails_/5406136
6	MED-NODE	http://www.cs.rug.nl/~imaging/databases/melanoma_naevi/	16	Edinburgh dataset	(https://licensing.eri.ed.ac.uk/i/software/dermofit-image-library.html)
7	Atlas Derm dataset	https://www.dermatlas.net/	17	Hallym Dataset	https://figshare.com/articles/Asan_and_Hallym_Dataset_Thumbnails_/5406136
8	DermIS	https://www.dermis.net/dermisroot/en/home/index.htm	18	ILSVRC; ISBI 2016 (Gutman	ISBI 2016 (Gutmanet al., 2016).
9	Derm101 (No longer available under this name)	https://www.emailmeform.com/builder/form/Ne0j8da9bb7U4h6t1f	19	HAM1000	https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T
10	GSE122703	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE122703	20	992 dermoscopy images	http://www.bccancer.bc.ca/HPI/SkinCancerAtlas/Melanoma/default.htm http://www.cancer.org/cancer/skincancer-melanoma/detailedguide/melanoma-skin-cancer-key-statistics. http://www.dermnet.org.nz/lesions/img/melanoma/mel-is.html, http://www.meddean.luc.edu/lumen/MedEd/medicine/dermatology/melton/content1.htm

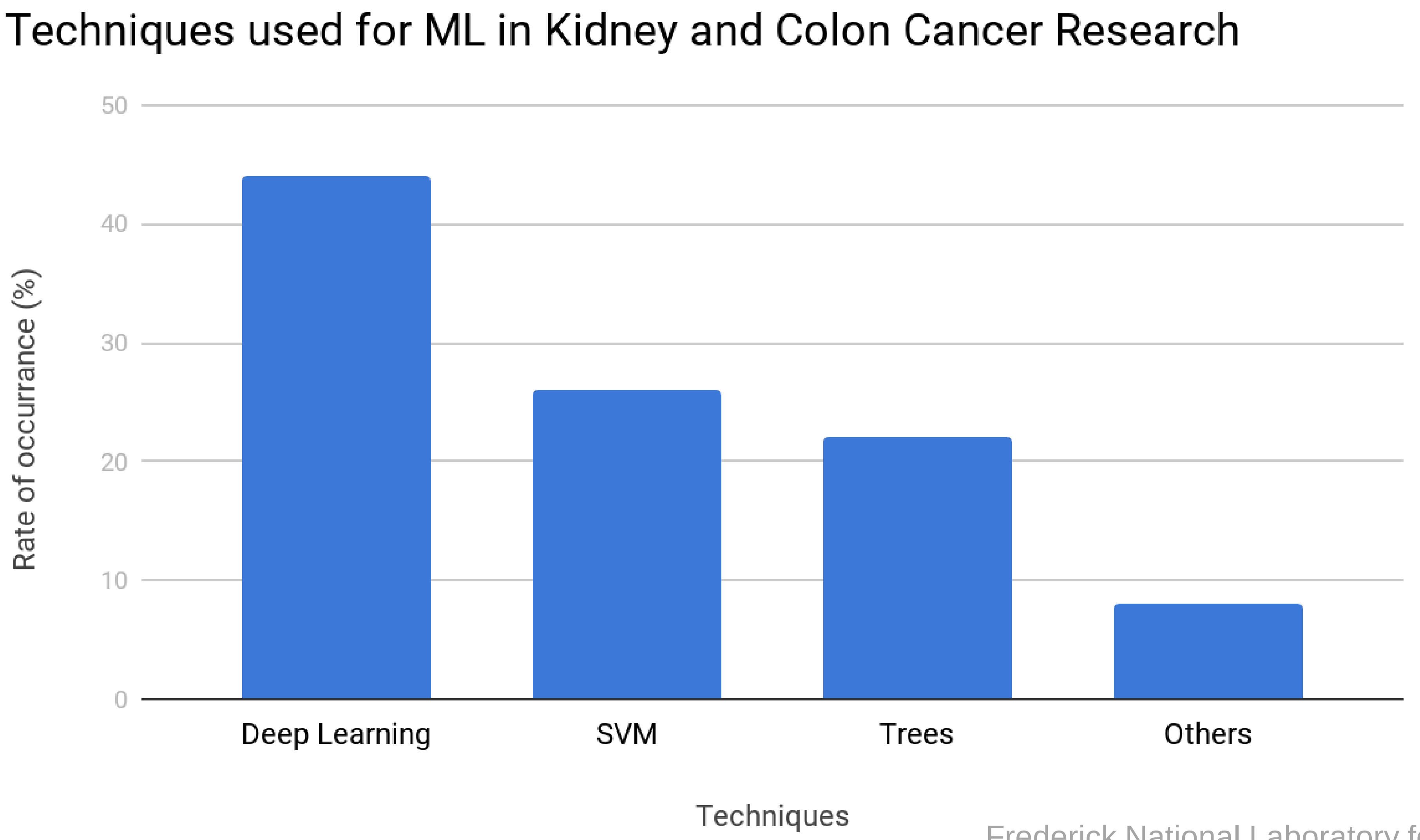
Table of content

- Project Overview
- Methodology
- **Results**
  - Skin and Breast Cancer
  - **Kidney and Colon Cancer**
  - Spinal Cord, Cranial Nerves and Ovary
  - Bronchus, Lung and Prostate Gland
- Next Steps

Most common use of ML in Kidney and Colon cancer is in identification, prognosis and progress prediction

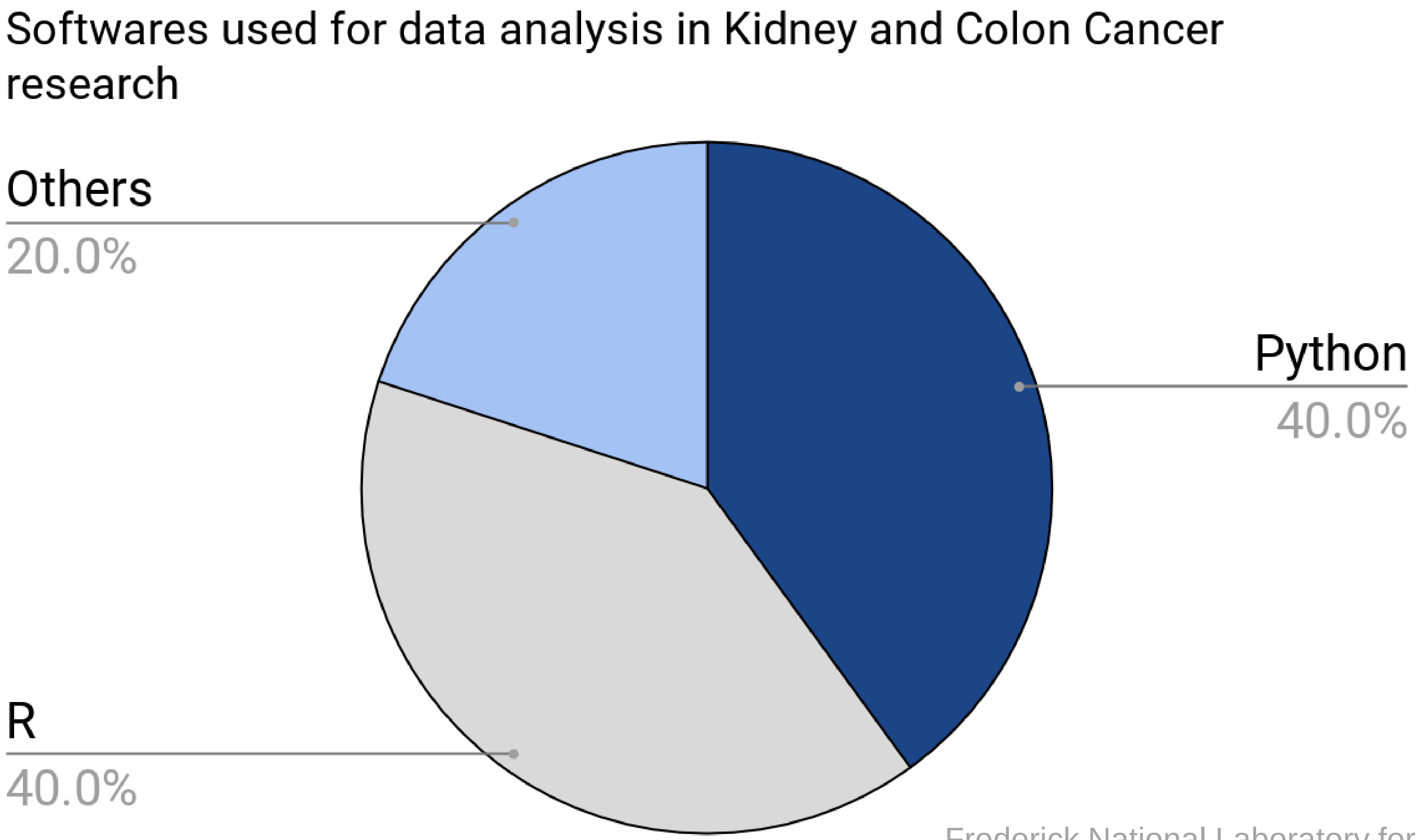
Kidney Cancer	Colon Cancer
• ML based CT analysis • Prognosis and survival prediction • Automated detection • Radiation dose estimation • Prediction of the nuclear grade of renal cell carcinoma	• Prediction of progression of colon cancer and survival • Expanding TNM for cancers • Establishing gene expression signature for prognosis • Automated detection • Classification of cancer types

Deep Learning is the most widely used technique followed by SVM



Python and R are predominantly used in research papers for Kidney and Colon cancer

Python is mainly used for deep learning, whereas R is predominantly used for more traditional ML techniques.



List of common public data repositories and datasets

Dataset	Cancer Type	Link
TCGA (The Cancer Genome Atlas) data repository	Any	https://cancergenome.nih.gov/
Harvard Dataverse	Any	https://dataverse.harvard.edu/
Gene Expression Omnibus	Any	https://www.ncbi.nlm.nih.gov/geo/
National Cancer Database	Any	https://www.facs.org/quality-programs/cancer/ncdb
A public H&E-stained image dataset of colorectal cancer	Colorectal	https://zenodo.org/record/1214456
National Cancer Institute SEER Data	Any	https://seer.cancer.gov/data/
Genomics Data Commons	Any	https://portal.gdc.cancer.gov/
CPTAC Clear Cell Renal Cell Carcinoma (CCRCC) Discovery Study	Kidney cancer	https://wiki.cancerimagingarchive.net/display/Public/CPTAC-CCRCC
The Cancer Imaging Archive	Any	https://www.cancerimagingarchive.net/collections/

Table of content

Only about 37 publications related ML applications in Spinal Cord, Cranial Nerves and Ovary research were found due to rarity

Research Overview

- Most of the data are coming from privately gathered information from different hospitals or medical center
- Despite the rarity of these cancer types, literature reviews show the potentials of machine learning techniques applied for both classification and regression prediction
- Machine learning maintains numerous advantages over conventional regression techniques, such as a reduced requirement for a priori knowledge on predictors and better ability to manage large datasets

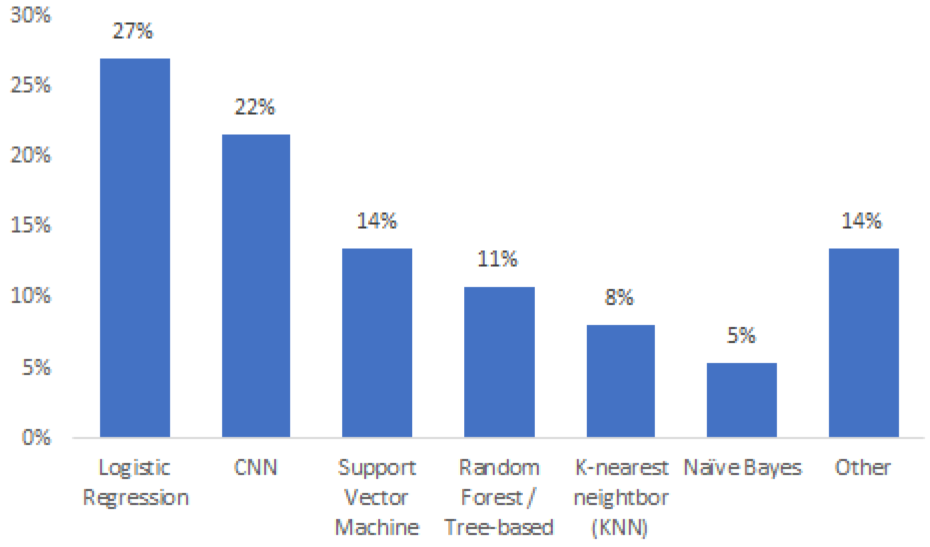
Spinal Cord, Cranial Nerves and Ovary

Use cases and objectives for ML research:

- Predictive Modeling (mostly classification):
  - Supplemented to clinical prognostic models for predicting the outcome of spinal cord injury (SCI)
  - Helped explore the non-linear relationships of patient features and prognostic functions
  - Classification in segmenting spinal cord and detecting ovarian cancer biomarker
  - Overperformed the traditional Logistic regression
- Pros for using ML:
  - Help describe nonlinear relationship
  - Less prior knowledge required
  - Improving diagnosis accuracy and efficiency
- Sample: 37

The most common algorithms to use in Spinal Cord, Cranial Nerves and Ovary research are Logistic Regression and CNN, despite the rarity of datasets

Algorithms Used for ML in Spinal Cord, Cranial Nerves and Ovary



Available Dataset

Dataset	Cancer Type	Link
The Rick Hansen Spinal Cord Injury Registry (RHSCIR)	Spinal Cord	https://www.nature.com/articles/sc2011109
Spinal Cord Injury Registry - NACTN	Spinal Cord	https://clinicaltrials.gov/ct2/show/results/NCT00178724
European Multicentre Study of Human Spinal Cord Injury	Spinal Cord	https://clinicaltrials.gov/ct2/show/study/NCT01571531
Ovary data	Ovary	https://figshare.com/articles/Raw_data/6025748

Problems appeared in the research of Spinal Cord, Cranial Nerves and Ovary

a. No public data source - Most of cases are privately collected from hospitals
b. Ambiguous in describing the data samples / features
c. Not robust performance due to the imbalance class distribution (less positive or diagnosed patients sample)

a. Instability of algorithm feasibility when applying to different datasets

a. Lack of automated tools for implementing or scaling the algorithm in real life

Table of content

- Project Overview
- Methodology
- Results
  - Skin and Breast Cancer
  - Kidney and Colon Cancer
  - Spinal Cord, Cranial Nerves and Ovary
  - **Bronchus, Lung and Prostate Gland**
- Next Steps

About 60 publications related machine learning applications in Lung, Bronchus and Prostate Cancer research were summarized

Use cases and objectives for ML in cancer research

Lung Cancer

- Predictive Modeling (mostly classification):
  - Patient survival probability
  - High risk-low risk classification of patients
  - Lung Cancer treatment response
- Deep Learning:
  - Image classification for lung cancer prognosis
  - Effectiveness of therapeutic antibody targeting in the body

Bronchus Cancer

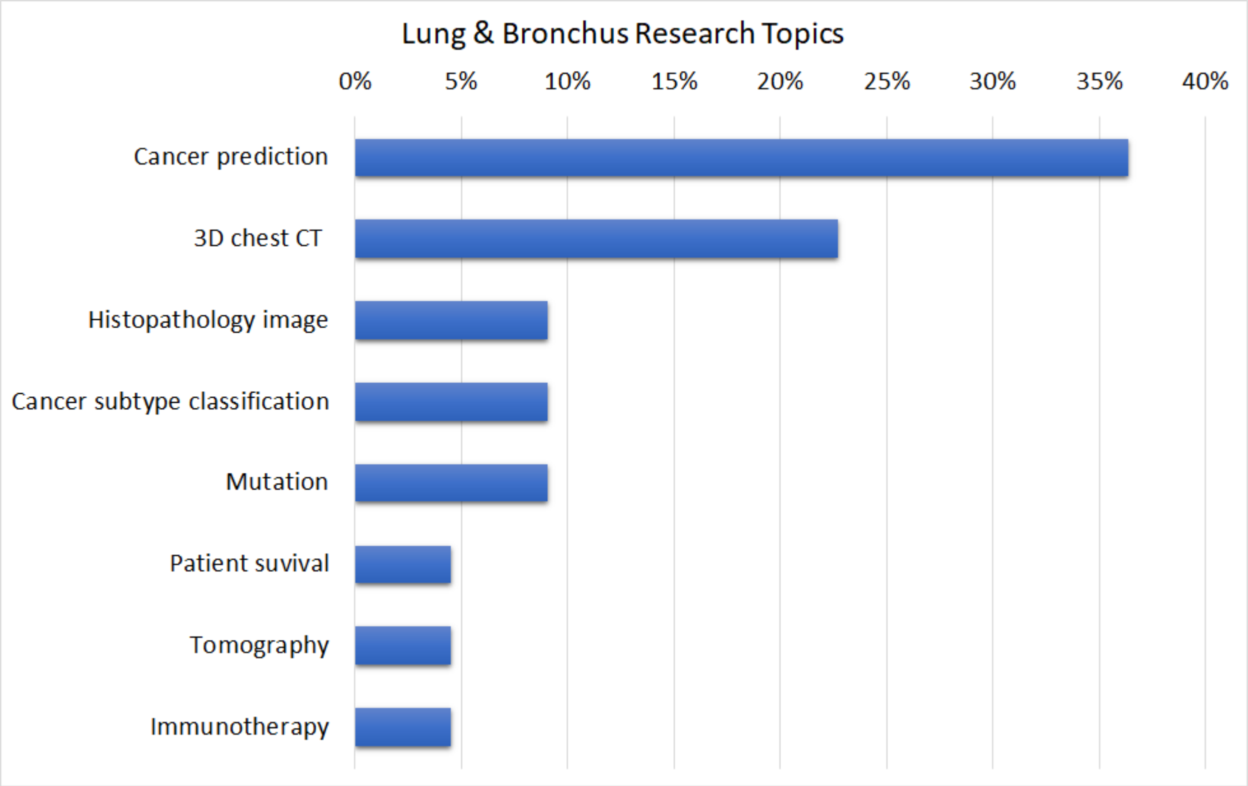
- Predictive Modeling (mostly classification):
  - Predict treatment outcomes of chemotherapy
- Deep Learning:
  - Automated anatomical labeling of bronchial branches extracted from 3D CT images
- Limited research as compared to lung and prostate cancer

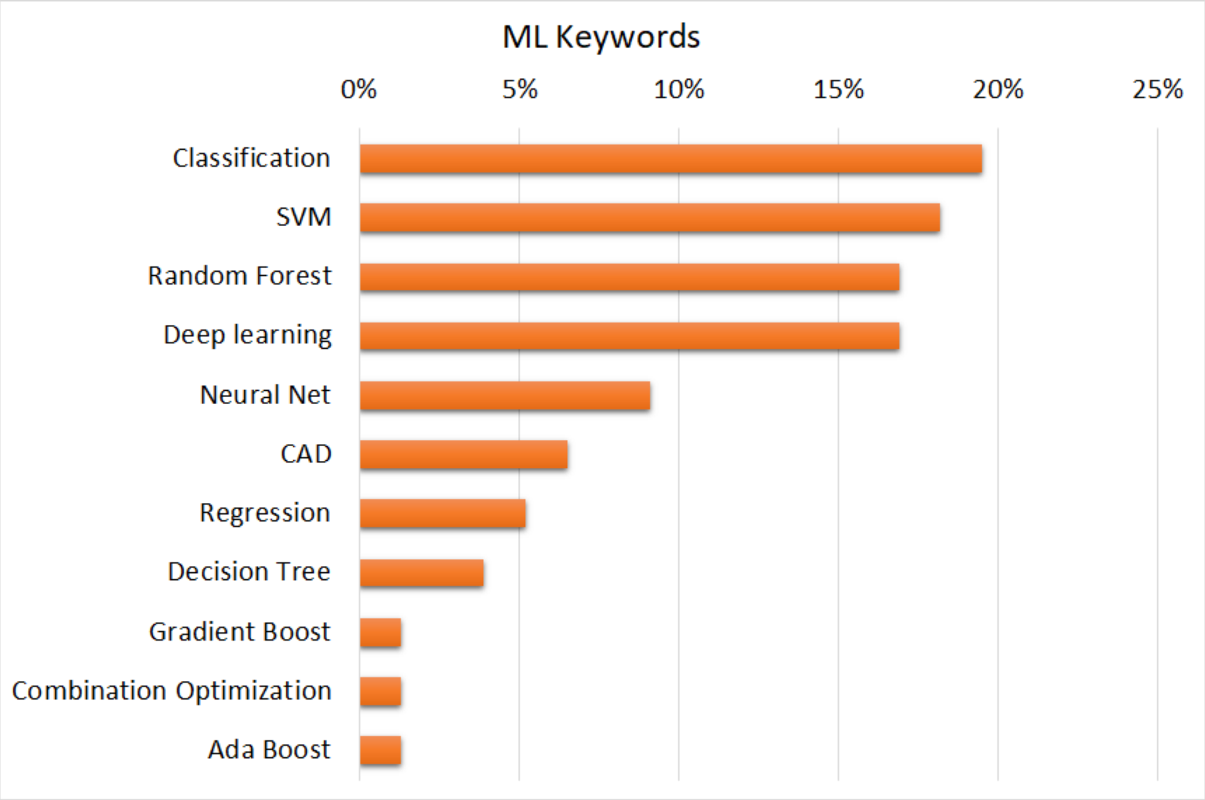
Prostate Cancer

- Predictive Modeling (mostly classification):
  - Prostate cancer probability prediction for early diagnosis
  - Survival probability
- Deep Learning:
  - Automated gleason grading of prostate cancer tissue
  - Prostate cancer magnetic resonance imaging

Cancer Prediction is the most popular research topic and Classification is the most frequently used machine learning technique

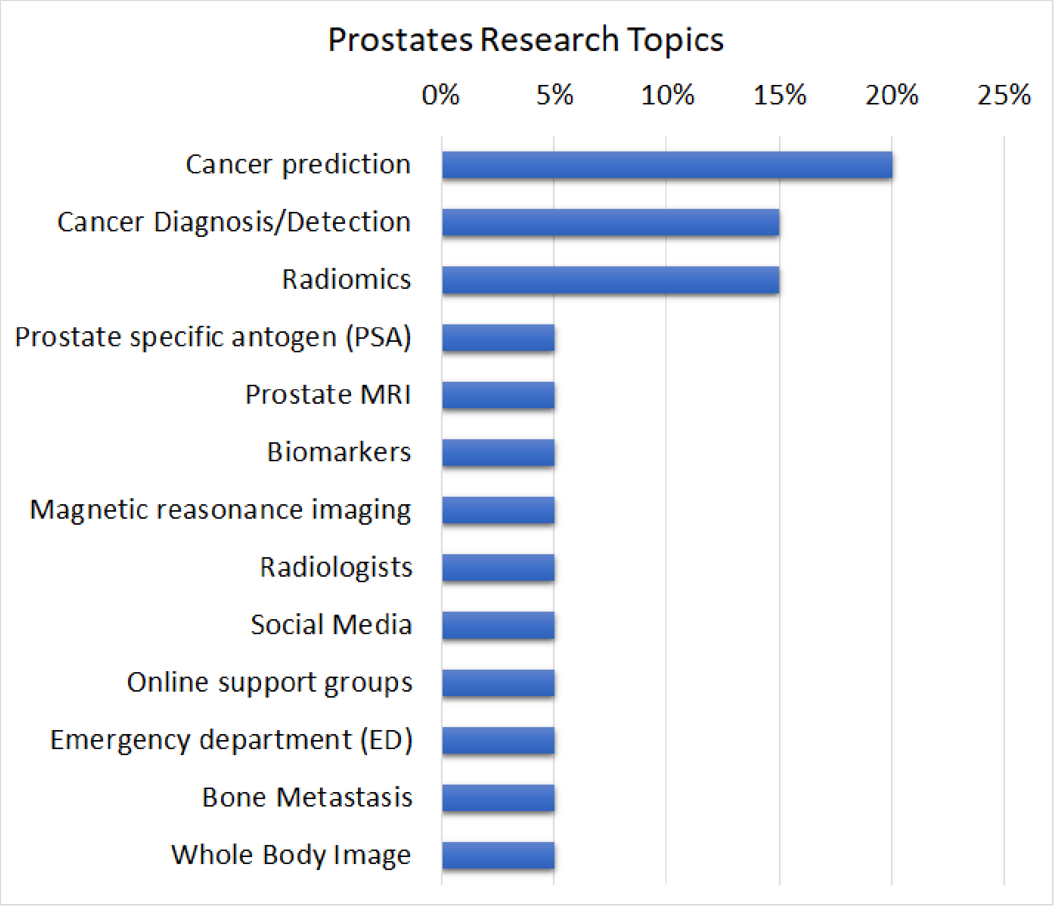
Top research topics and machine learning key words..

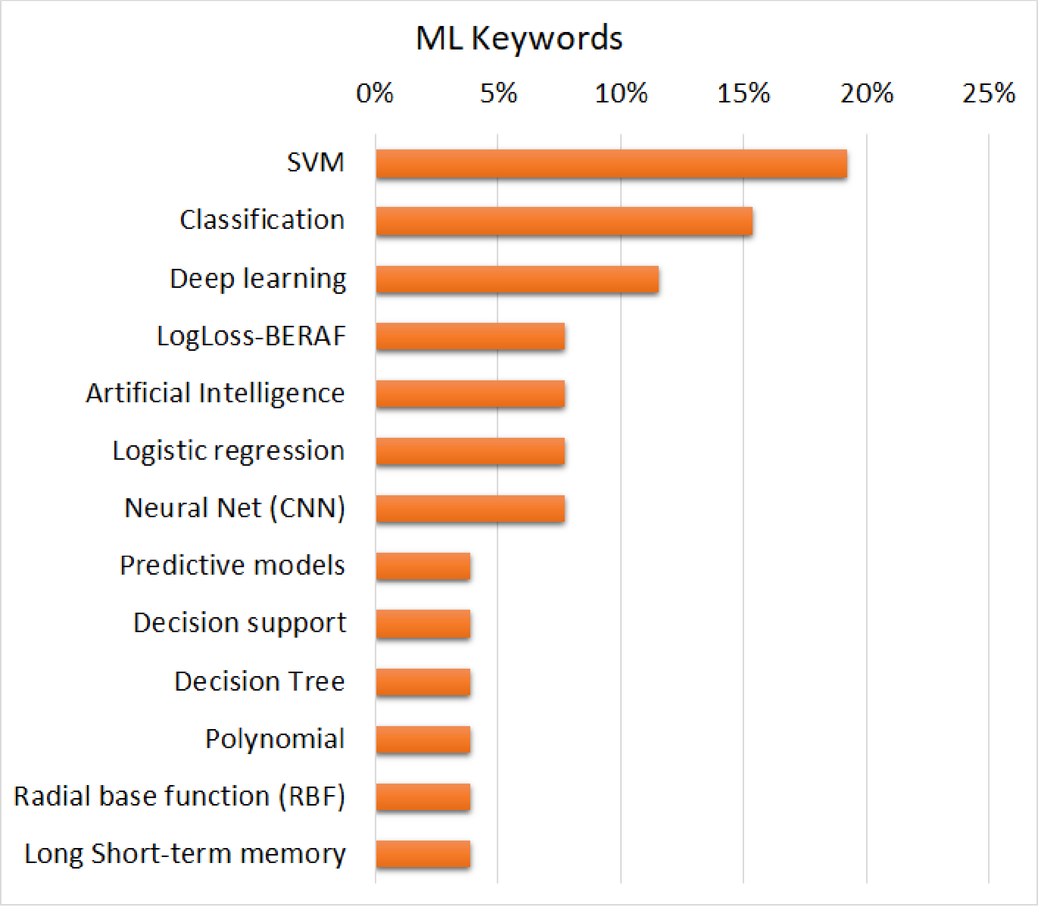




Cancer Prediction is the most popular research topic and SVM is the most frequently used machine learning technique

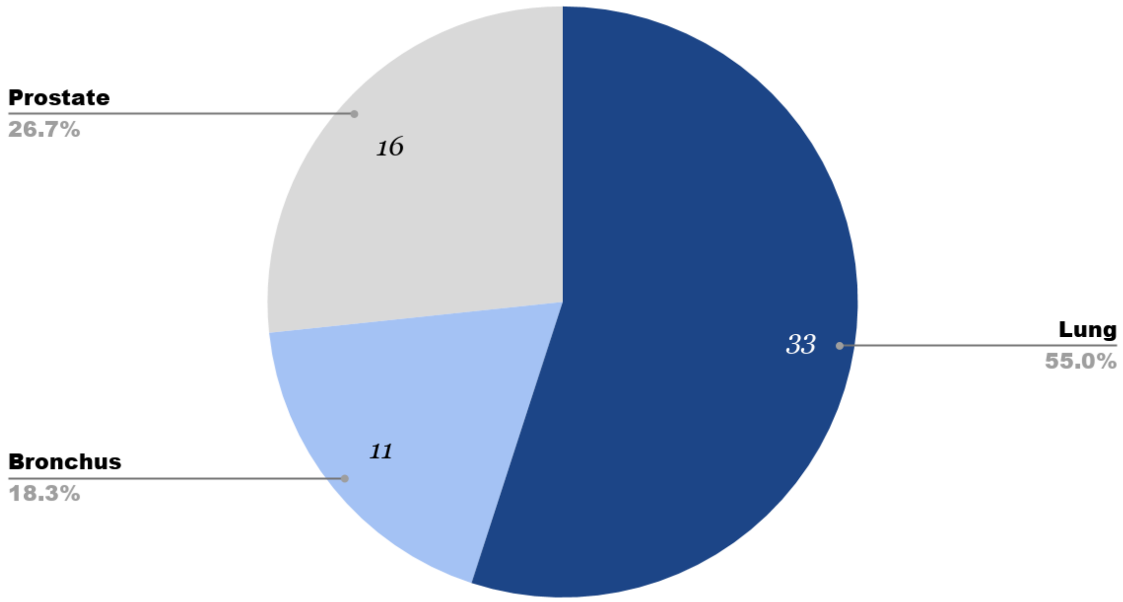
Top research topics and machine learning key words..



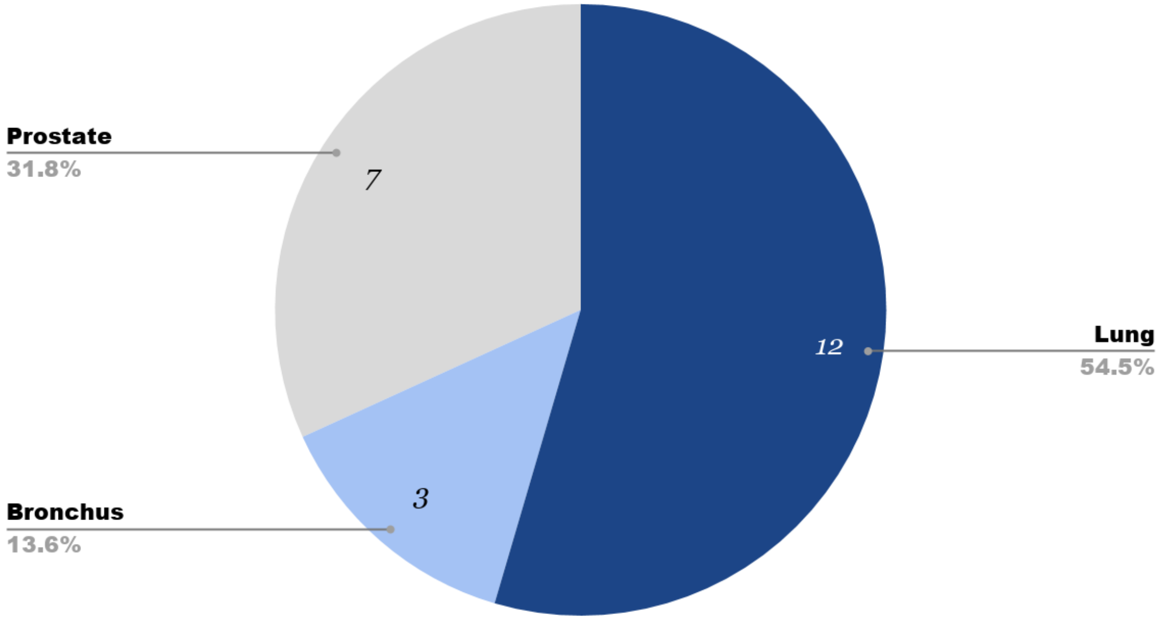


Gathered data sources for ~37% publications summarized and most of them are publicly available

Number of publications summarized by type of cancer



Number of datasets identified by type of cancer



In total 22 datasets were collected in total out of 60 papers summarized for Prostate, Bronchus and Lung cancer

20 Datasets were identified from Lung and Prostate cancer research publications

S.No.	Dataset (L - Lung, P - Prostate)	Link	S.No.	Dataset (L - Lung, P - Prostate)	Link
1	SEER Incidence Data - L	https://seer.cancer.gov/data/	11	Biomarker Data - L	https://www.iprox.org/page/project.html?id=IPX0001153000
2	Genomic Data Commons - L	https://portal.gdc.cancer.gov/	12	Immunotherapy Data - L	https://wiki.cancerimagingarchive.net/display/Public/
3	LIDC-IDRI - L	https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI	13	Biomarker Data - P	https://luisrueda.myweb.cs.uwindsor.ca/datasets/Hierarchical-Prostate-Cancer-Gleason.rar
4	NLST Pathology Images - L	https://cdas.cancer.gov/learn/nlst/images/	14	OSG Data - P	https://github.com/tharindurb/PRIME.
5	NSCLC-Radiomics - L	https://wiki.cancerimagingarchive.net/display/Public/NSCLC-Radiomics	15	GEO Dataset - P	https://www.ncbi.nlm.nih.gov/gds
6	NSCLC-Radiomics-Genomics - L	https://wiki.cancerimagingarchive.net/display/Public/NSCLC-Radiomics-Genomics	16	Cancer Genome Atlas - P	https://docs.google.com/spreadsheets/d/12CXisExMlbP_11GJNZp0Ut7-1VX92_eK0GfWysbmxPk/edit#gid=201540152
7	Gene-expression Data - L	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE58661	17	Radiographs - P	https://figshare.com/articles/Data_from_An_Investigation_of_Machine_Learning_Methods_in_Delta-radiomics_Feature_Analysis/9943334
8	RIDER Collections - L	https://wiki.cancerimagingarchive.net/display/Public/RIDER+Collections	18	UTI Diagnosis Data - P	https://doi.org/10.1371/journal.pone.0194085.s001
9	TCGA Dataset - L	https://gdc.cancer.gov/about-data/publications/pancanatlas	19	CPC-GENE Data - P	http://www.cbioportal.org/study/summary?id=prad_cpcg_2017
10	NGC Data - L	https://stm.sciencemag.org/content/10/457/eaar7939/tab-figures-data	20	PRAD-FR - P	https://dcc.icgc.org/projects/PRAD-FR

Table of content

Image processing and predictive modeling are common machine learning applications in cancer research

There are some commonalities in the research carried out across different types of cancer

Image processing for cancer detection

- Machine learning based image processing techniques to identify the cancer type and further classify patients as high risk and low risk groups.
- Deep Learning and Convolutional Neural Networks (CNNs) are most common techniques followed by SVM

ML applications in cancer prognosis

- Predictive modeling to improve accuracy in predicting cancer susceptibility, recurrence and survival prediction.
- However, there is lack of external validation or testing regarding the performance of these predictive models

Future Work

- The datasets gathered can be further used to conduct research at FNL
- Building on the datasets identified, one can dive into those datasets to validate the integrity of the research publications
- Continue to gather more datasets available, as are a still lot research papers available for some cancer areas that are not summarized during this project
- In future, loop in more senior people in emails to authors to improve response rate
- NLP techniques can be applied on research summary table (e.g. to identify popular softwares used, ML packages etc.)
- Conduct a trend analysis, which ML techniques are gaining popularity in the recent years

THANK YOU!

Thank you for this opportunity, especially to Ravi and Naomi. We learned about FNL and about collaborating to solve real machine learning problems in medical science. We've learned a great deal and hope to be able to work with FNL again.

We have identified ***79 publicly available datasets*** for machine learning publications in the oncology space.

Any questions?