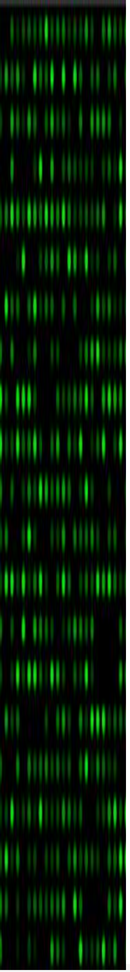


# Patient Stratification

## Team: Personalized Medicine

**Kenny Chou, Shashi Sharma, Cesar Ramirez Ibanez, Xiaoping Zhu**

*Classifying cancer mutations for personalized treatments*

https://github.com/kfchou/PersonalizedMedicine/

# The Problem

Personalized Medicine holds true potential to bring precise and effective treatments for the patients.

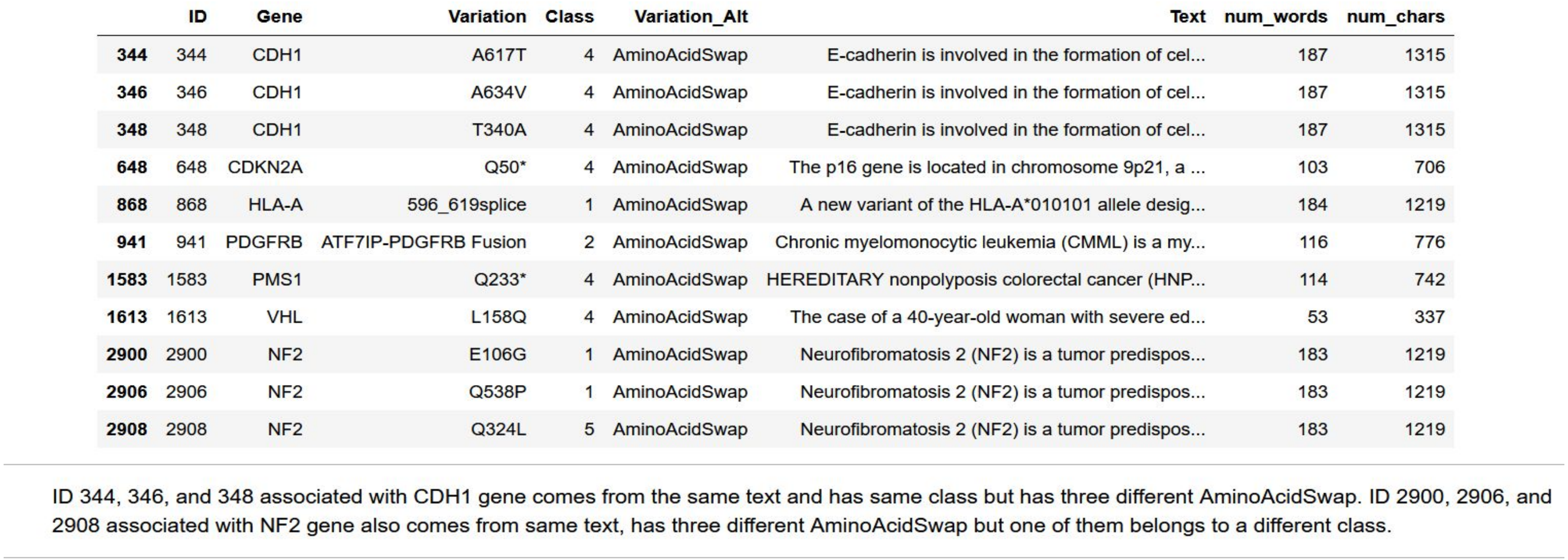
This is one of the key questions in cancer genetics and precision medicine that can improve lives of uncountable patients every year.

**Goal:** Our goal is to create an algorithm that classifies mutated genes from cancerous tumors by analyzing text based clinical literature in order to match patients with the best treatment.

[1] Cancer.gov
[2] American Cancer Society

# Data Source

Data Sources: Memorial Sloan Kettering Cancer Center (Available on Kaggle)

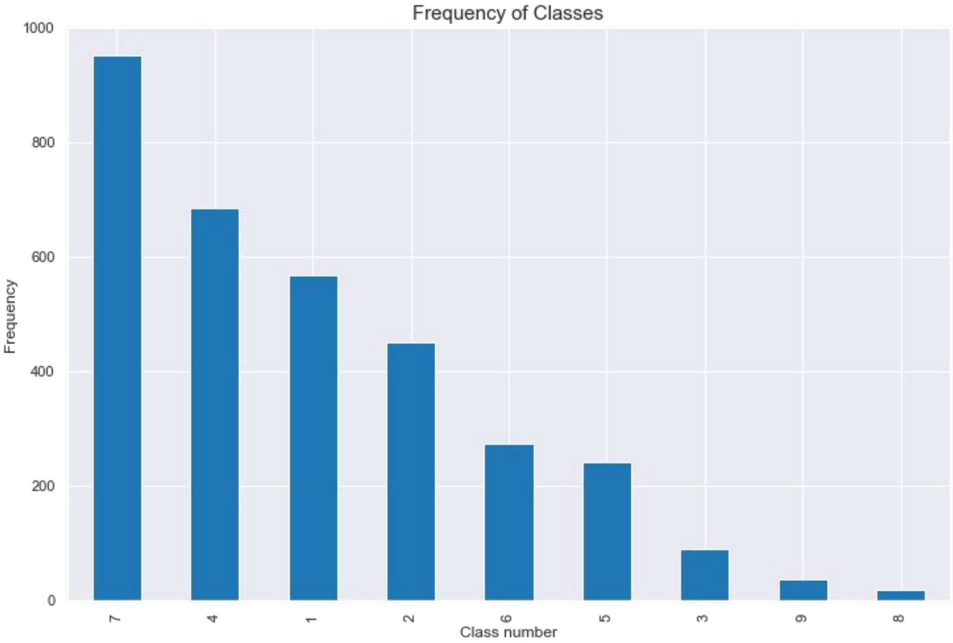


| | ID | Gene | Variation | Class | Variation_Alt | Text | num_words | num_chars |
|---|---|---|---|---|---|---|---|---|
| **344** | 344 | CDH1 | A617T | 4 | AminoAcidSwap | E-cadherin is involved in the formation of cel... | 187 | 1315 |
| **346** | 346 | CDH1 | A634V | 4 | AminoAcidSwap | E-cadherin is involved in the formation of cel... | 187 | 1315 |
| **348** | 348 | CDH1 | T340A | 4 | AminoAcidSwap | E-cadherin is involved in the formation of cel... | 187 | 1315 |
| **648** | 648 | CDKN2A | Q50* | 4 | AminoAcidSwap | The p16 gene is located in chromosome 9p21, a ... | 103 | 706 |
| **868** | 868 | HLA-A | 596_619splice | 1 | AminoAcidSwap | A new variant of the HLA-A*010101 allele desig... | 184 | 1219 |
| **941** | 941 | PDGFRB | ATF7IP-PDGFRB Fusion | 2 | AminoAcidSwap | Chronic myelomonocytic leukemia (CMML) is a my... | 116 | 776 |
| **1583** | 1583 | PMS1 | Q233* | 4 | AminoAcidSwap | HEREDITARY nonpolyposis colorectal cancer (HNP... | 114 | 742 |
| **1613** | 1613 | VHL | L158Q | 4 | AminoAcidSwap | The case of a 40-year-old woman with severe ed... | 53 | 337 |
| **2900** | 2900 | NF2 | E106G | 1 | AminoAcidSwap | Neurofibromatosis 2 (NF2) is a tumor predispos... | 183 | 1219 |
| **2906** | 2906 | NF2 | Q538P | 1 | AminoAcidSwap | Neurofibromatosis 2 (NF2) is a tumor predispos... | 183 | 1219 |
| **2908** | 2908 | NF2 | Q324L | 5 | AminoAcidSwap | Neurofibromatosis 2 (NF2) is a tumor predispos... | 183 | 1219 |

ID 344, 346, and 348 associated with CDH1 gene comes from the same text and has same class but has three different AminoAcidSwap. ID 2900, 2906, and 2908 associated with NF2 gene also comes from same text, has three different AminoAcidSwap but one of them belongs to a different class.

Useful packages: pandas

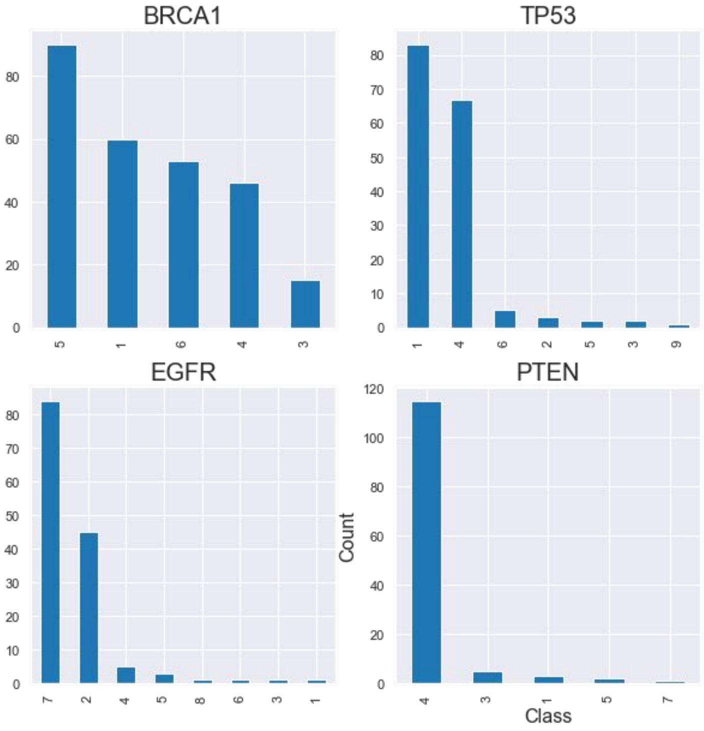
# Highly non-uniform distribution of mutations per class



Challenge: Skewed data may make ML models biased toward certain classes

# Gene & mutation type have some predictive power

Distribution of frequency occuring genes into various classes

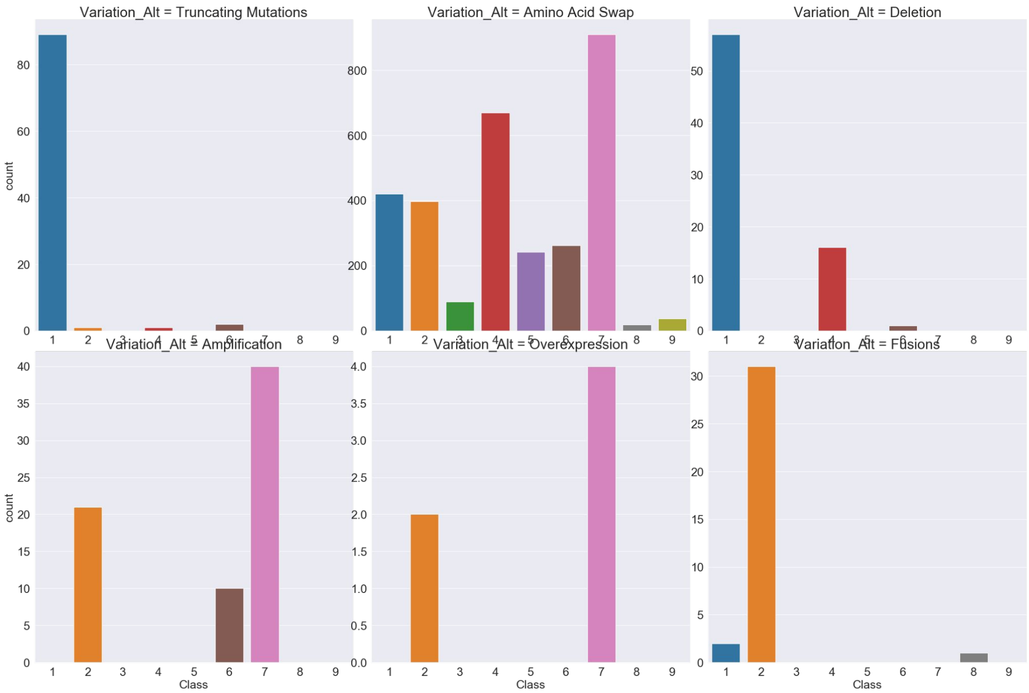


BRCA1

TP53

EGFR

PTEN



Variation_Alt = Truncating Mutations

Variation_Alt = Amino Acid Swap

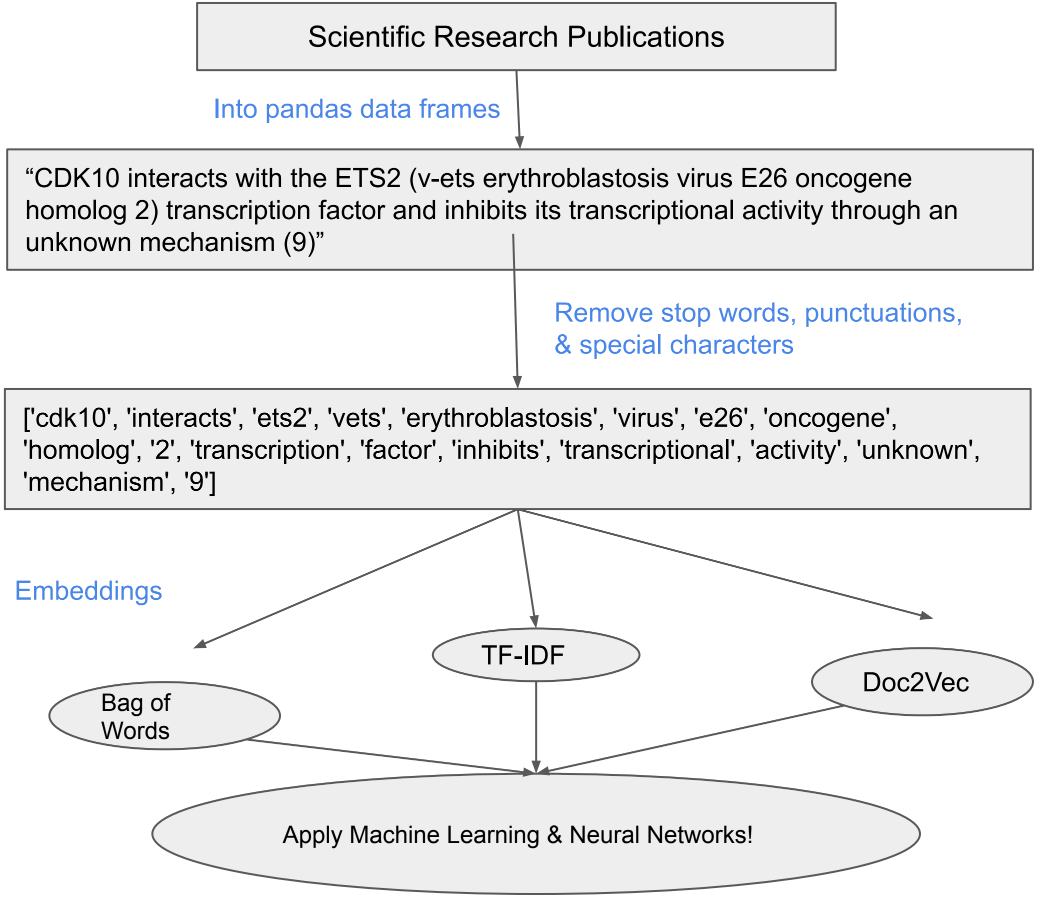
Variation_Alt = Deletion

Variation_Alt = Amplification

Variation_Alt = Overexpression

Variation_Alt = Fusions

# Data Cleaning



Scientific Research Publications

Into pandas data frames

“CDK10 interacts with the ETS2 (v-ets erythroblastosis virus E26 oncogene homolog 2) transcription factor and inhibits its transcriptional activity through an unknown mechanism (9)”

Remove stop words, punctuations, & special characters

['cdk10', 'interacts', 'ets2', 'vets', 'erythroblastosis', 'virus', 'e26', 'oncogene', 'homolog', '2', 'transcription', 'factor', 'inhibits', 'transcriptional', 'activity', 'unknown', 'mechanism', '9']

Embeddings

Bag of Words

TF-IDF

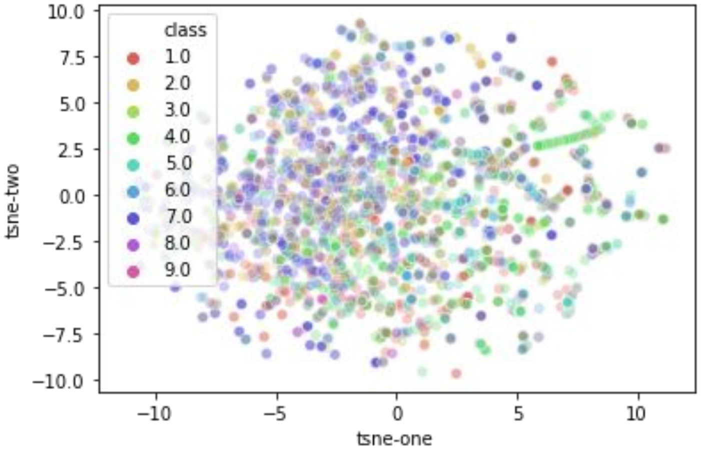
Doc2Vec

Apply Machine Learning & Neural Networks!

Useful packages: nltk, re, sklearn, pandas

# Feature Encoding

Dov2Vec Visualization w/ t-SNE



Bag of Words - counts the occurrence of each word in a document (SS)

Doc2Vec - calculates an embedding for entire documents, based on the word embeddings of each word in the document (KC)

TF-IDF - calculates how relevant a word is to a document, in a collection of documents (CRI, XZ)

Useful packages: sklearn, gensim

# Classification & Assessment

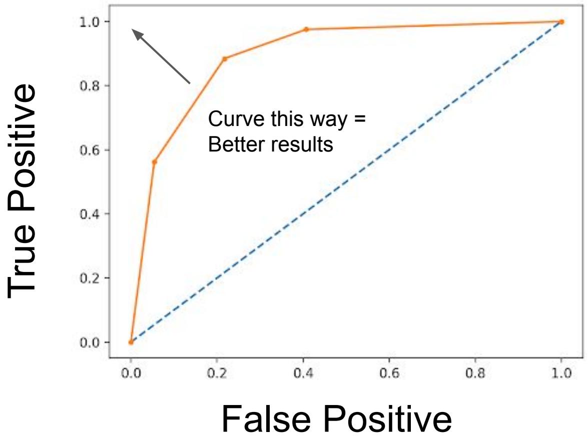
## Classification

- Logistic Regression (SS)
- Random Forest (SS)
- SVC (CRI)
- Feed Forward Network (KC)
- LSTM (KC)

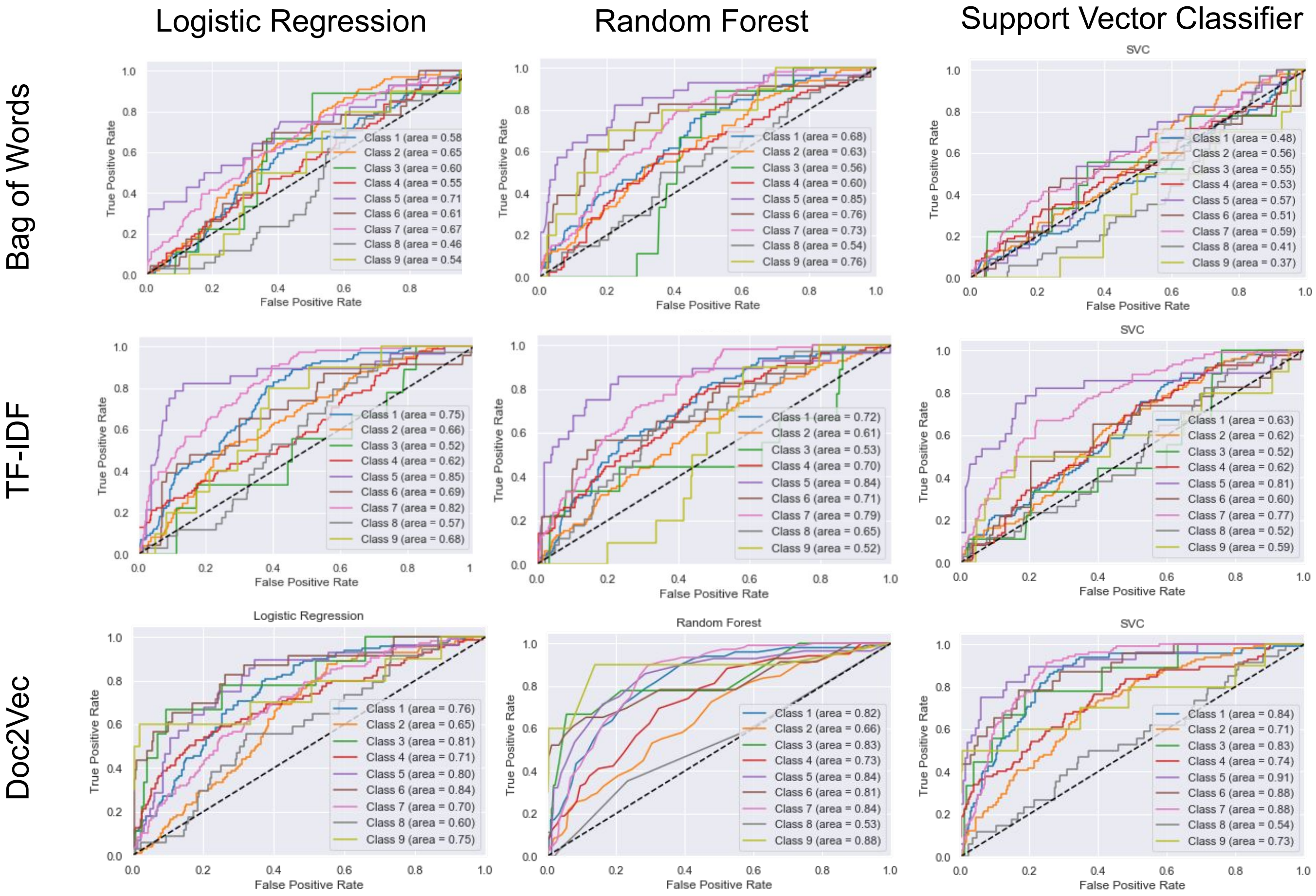
Useful packages: sklearn, keras

## Assessment

ROC curves & Area under ROC

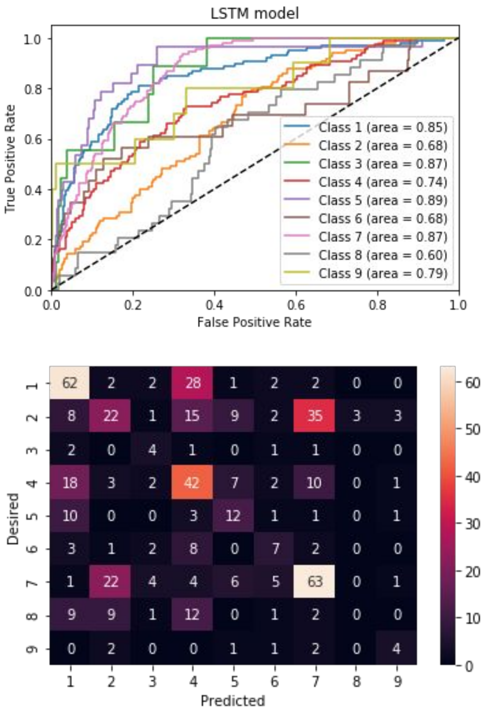


# Classical ML Results



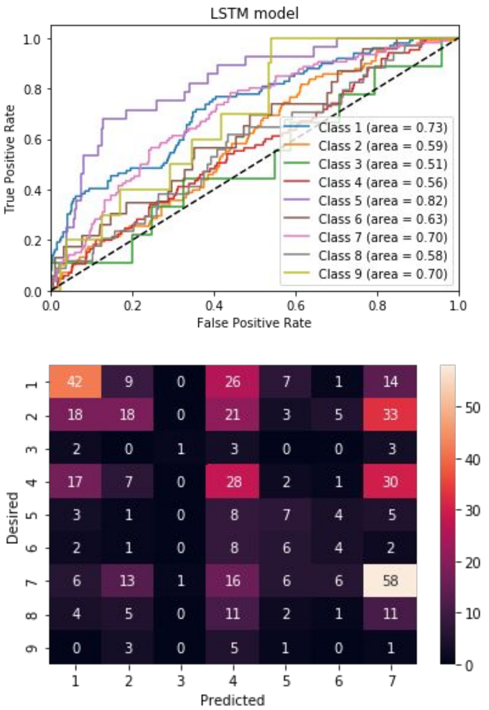
# Deep Learning Results

Doc2Vec



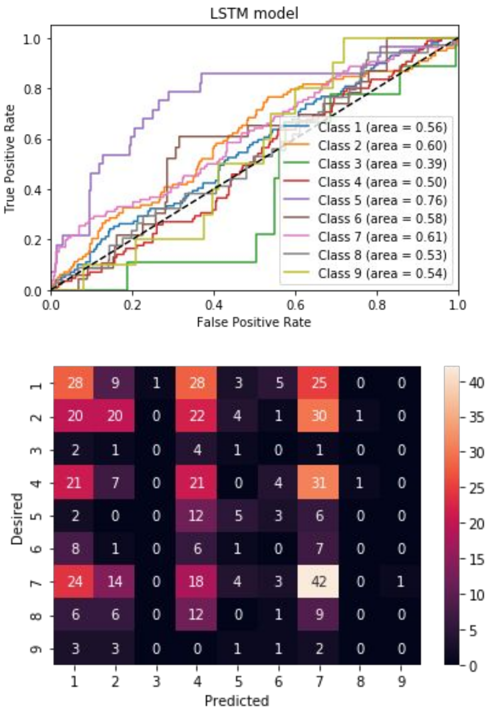
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 62 | 2 | 2 | 28 | 1 | 2 | 2 | 0 | 0 |
| 2 | 8 | 22 | 1 | 15 | 9 | 2 | 35 | 3 | 3 |
| 3 | 2 | 0 | 4 | 1 | 0 | 1 | 1 | 0 | 0 |
| 4 | 18 | 3 | 2 | 42 | 7 | 2 | 10 | 0 | 1 |
| 5 | 10 | 0 | 0 | 3 | 12 | 1 | 1 | 0 | 1 |
| 6 | 3 | 1 | 2 | 8 | 0 | 7 | 2 | 0 | 0 |
| 7 | 1 | 22 | 4 | 4 | 6 | 5 | 63 | 0 | 1 |
| 8 | 9 | 9 | 1 | 12 | 0 | 1 | 2 | 0 | 0 |
| 9 | 0 | 2 | 0 | 0 | 1 | 1 | 2 | 0 | 4 |

TF-IDF



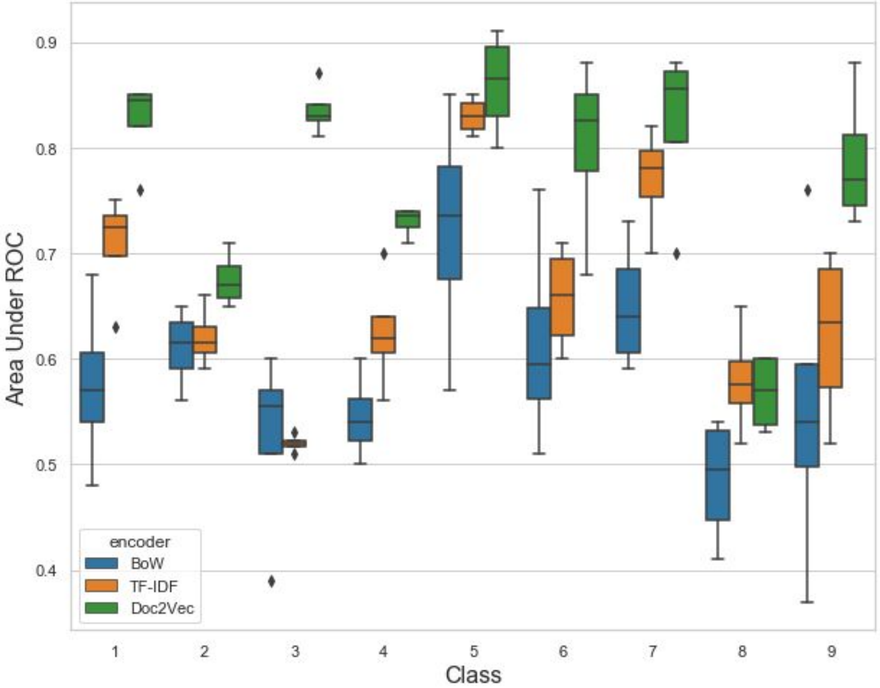
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 1 | 42 | 9 | 0 | 26 | 7 | 1 | 14 |
| 2 | 18 | 18 | 0 | 21 | 3 | 5 | 33 |
| 3 | 2 | 0 | 1 | 3 | 0 | 0 | 3 |
| 4 | 17 | 7 | 0 | 28 | 2 | 1 | 30 |
| 5 | 3 | 1 | 0 | 8 | 7 | 4 | 5 |
| 6 | 2 | 1 | 0 | 8 | 6 | 4 | 2 |
| 7 | 6 | 13 | 1 | 16 | 6 | 6 | 58 |
| 8 | 4 | 5 | 0 | 11 | 2 | 1 | 11 |
| 9 | 0 | 3 | 0 | 5 | 1 | 0 | 1 |

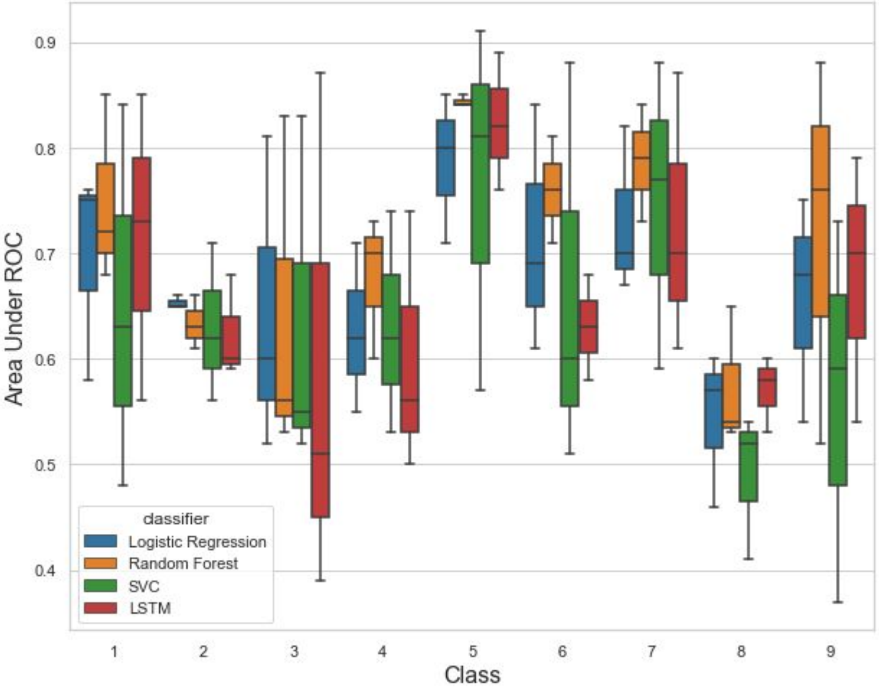
Bag of Words



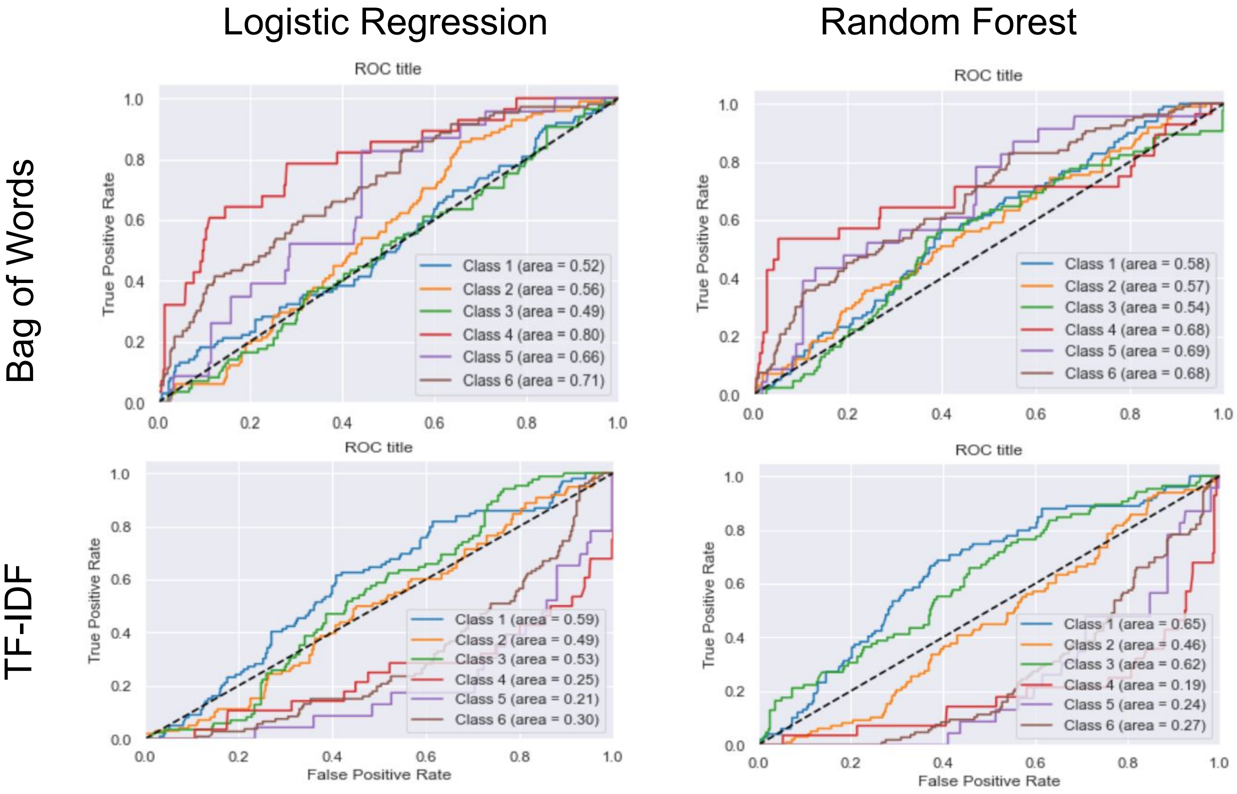
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 28 | 9 | 1 | 28 | 3 | 5 | 25 | 0 | 0 |
| 2 | 20 | 20 | 0 | 22 | 4 | 1 | 30 | 1 | 0 |
| 3 | 2 | 1 | 0 | 4 | 1 | 0 | 1 | 0 | 0 |
| 4 | 21 | 7 | 0 | 21 | 0 | 4 | 31 | 1 | 0 |
| 5 | 2 | 0 | 0 | 12 | 5 | 3 | 6 | 0 | 0 |
| 6 | 8 | 1 | 0 | 6 | 1 | 0 | 7 | 0 | 0 |
| 7 | 24 | 14 | 0 | 18 | 4 | 3 | 42 | 0 | 1 |
| 8 | 6 | 6 | 0 | 12 | 0 | 1 | 9 | 0 | 0 |
| 9 | 3 | 3 | 0 | 0 | 1 | 1 | 2 | 0 | 0 |

# Text encoder plays a larger role than the classifier





# Remove Rare Classes



# NEXT STEPS

1. Augment data by scraping from pubmed
2. Try to decorrelated easily confused classes - 1 & 4, 2 & 7
3. Try modern deep learning models - attention & transformers

# *Team: Personalized Medicine*

**https://github.com/kfchou/PersonalizedMedicine**

| | | |
|---|---|---|
| **Kenny Chou** | kfchou@bu.edu | Boston University |
| **Shashi Sharma** | Shashi.Sharma@rutgers.edu | Rutgers University |
| **Cesar Ramirez Ibanez** | cr718@scarletmail.rutgers.edu | Rutgers University |
| **Xiaoping Zhu** | xz349@math.rutgers.edu | Rutgers University |

# We thank Erdős Institute for this career forwarding opportunity!

Gratitudes to:

Matt Osborne

Roman Holowinsky

Lindsay Warrenburg

Nirav Patel

All fellow teams today