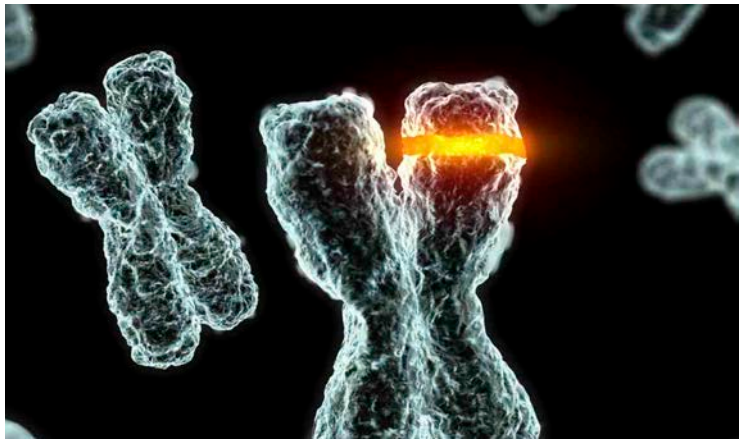
# Personalized Medicine: Redefining Cancer Treatment

Team Karkinos

# Introduction



| Conventional Workflow |
| --- |
| **Step 1:** Select genetic mutations |
| **Step 2:** Search in literature on selected mutations |
| **Step 3:** **Manually** analyze text, and classify mutations<br>**tremendous time/efforts/expertise** |

| Our Task |
| --- |
| **Task:**<br>**Automate Step 3** |
| **Dataset:**<br>expert-annotated knowledge base on genetic mutations |
| **Classification** problem:<br>Classify entries of **genes + mutations + text** into 9 labeled classes |

# Pipeline

## Data & Preprocessing

### Academic Articles

Shared Genetic Factors Involved in Celiac Disease, Type 2 Diabetes and Anorexia Nervosa Suggest Common Molecular Pathways for Chronic Diseases.

### Gene + Mutation Names


## Feature Extraction

- TF-IDF Text
- CountVectorizer Gene & Mutation n-grams

→ Latent Semantic Analysis Truncated SVD

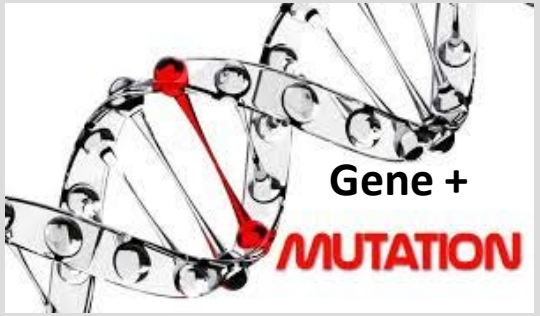
- Genes/mutation families encoded
- Word counts extracted from text/genetic data

## Modeling

→ XGBoost modeling + cross validation

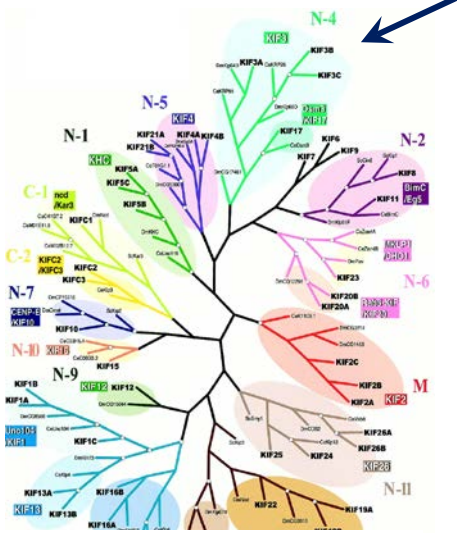
→ Ensemble of Top 5 XGBoost Models

## Input Data

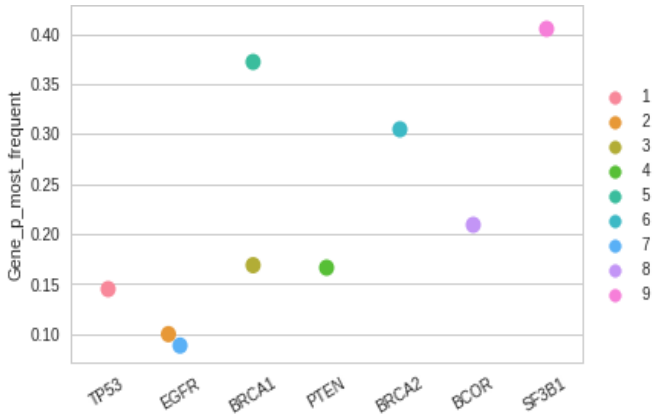



## Academic Articles

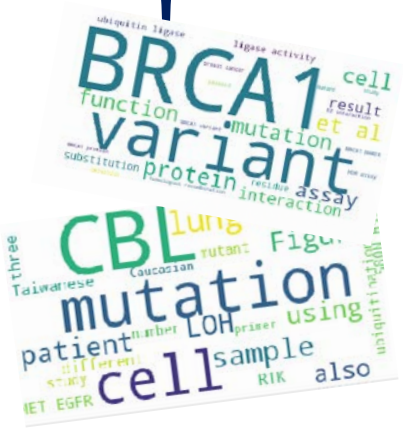

## Features Extracted & Unionized




Gene Family Root & Mutation Types

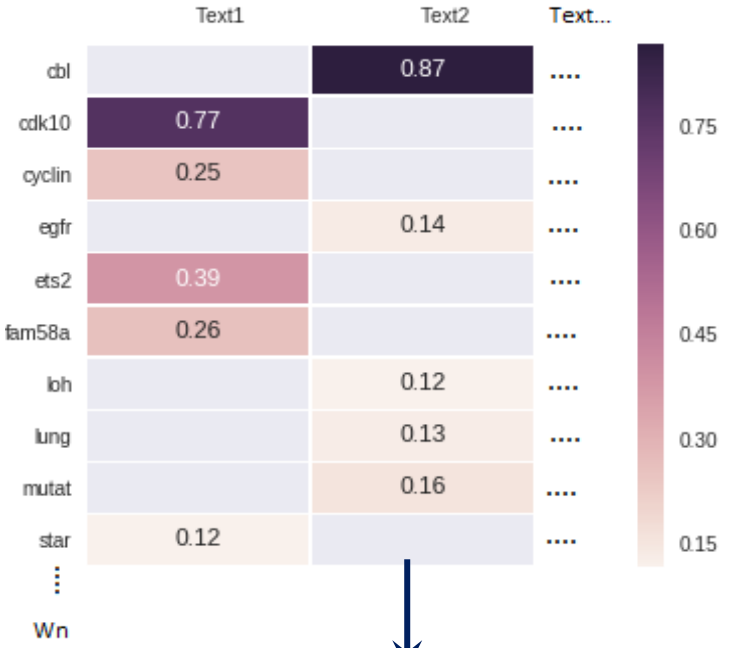



Dimensions reduced with Truncated SVD




Frequency of gene/mutation in text & word count

## Tf-idf

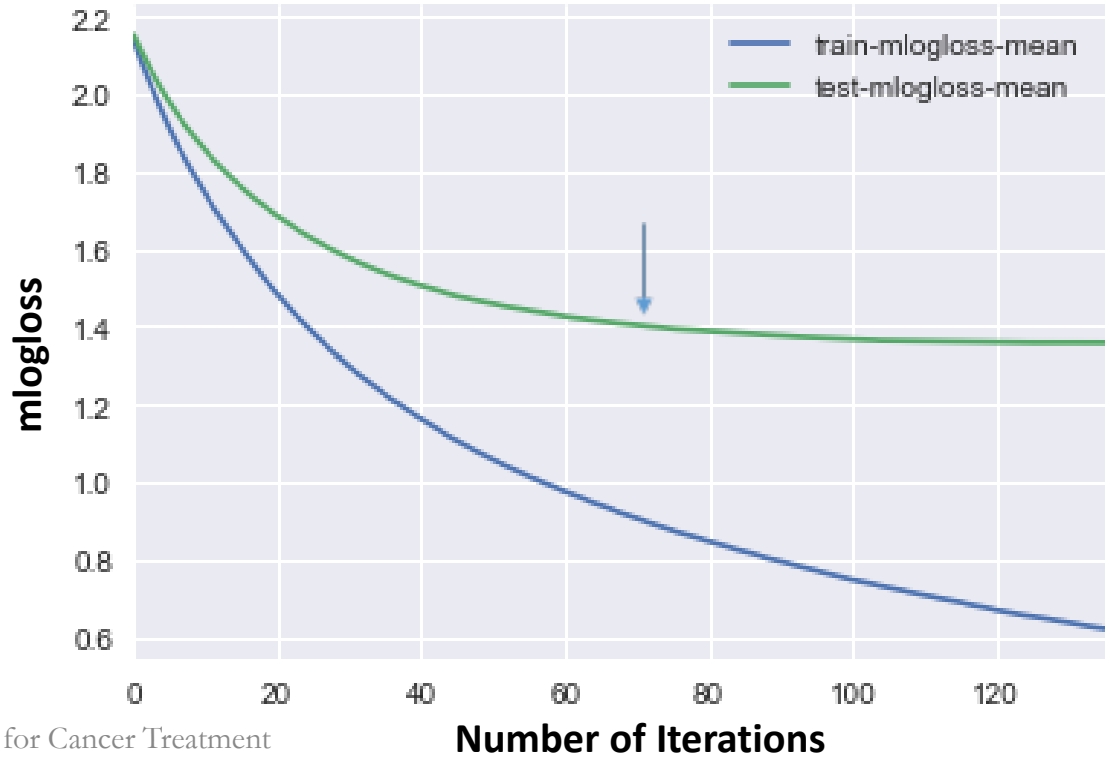



LSA (Truncated SVD)

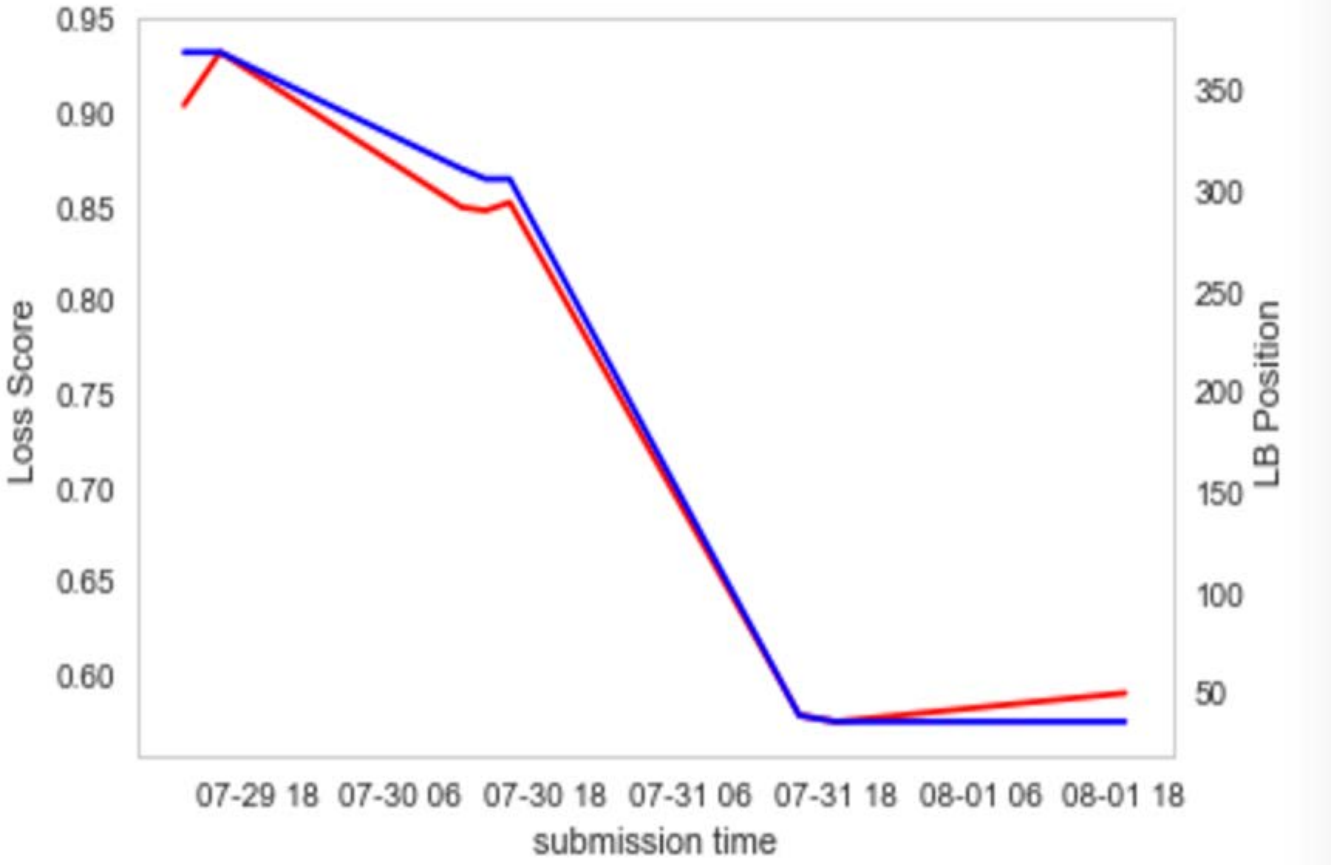
# Modeling

Model selection: **XGBoost**


| Cross Validation |
| --- |
| Evaluation metric: mlogloss |
| Hyperparameter tuning: ensemble of 5 XGBoosts with top 5 sets of hyperparameter |
| Number of iterations: problem of overfitting |



# Results & Future Steps




Final Log Score on Training: 0.29693

Final Log Score on Validation: 0.94718

Final Log Score on Testing: 0.58404

Final Leader-Board Position: Top 10%

## Future Steps

1. Word2Vec model pre-trained on larger corpus of Bio texts for better text mining

2. Better ensembling of unrelated models and Stack generalization

# Thank you!