Data Science Institute
Columbia University

# Cutaneous T-Cell Lymphoma: Data Science in Medicine

Data Science Capstone Project with
Prof. Itsik Pe'er

George Bingham Reynolds, Haoyang Shen, Adrian Garcia Hernandez, Zhaoyu Wu, Sung Jun Won
Celine M. Schreidah, Dr. Larisa Geskin
Prof. Itsik Pe'er

## Introduction

Our project topic is Cutaneous T-Cell Lymphoma (CTCL), a very rare form of skin cancer. As data scientists, our core goal was to use genomic, clinical and textual data to predict diagnostic and survival outcomes. Further, we aimed to find key drivers of these outcomes in the hopes of guiding future research and clinical approaches.

## Methodology and Results

### Diagnostic Modeling

We tried several classifiers for diagnosing two subtypes of CTCL (Sezary Syndrome and Mycosis Fungoides) including linear models, tree-based models, support vector machine classifiers and boosted classifiers. Adaboost classification and XGBoost are the two most performant. We then applied Principal Component Analysis to the inputs, improving performance for most classifiers, although AdaBoost has the best performance - precision scores above .92 - without it. Precision and Accuracy results are shown below.

To find gene level drivers, we first found the 25 inputs with the highest Adaboost feature importance through 10k runs. We then used permutation testing for all genes located in the associated chromosomes. Gene results are still proprietary; region results are shown.

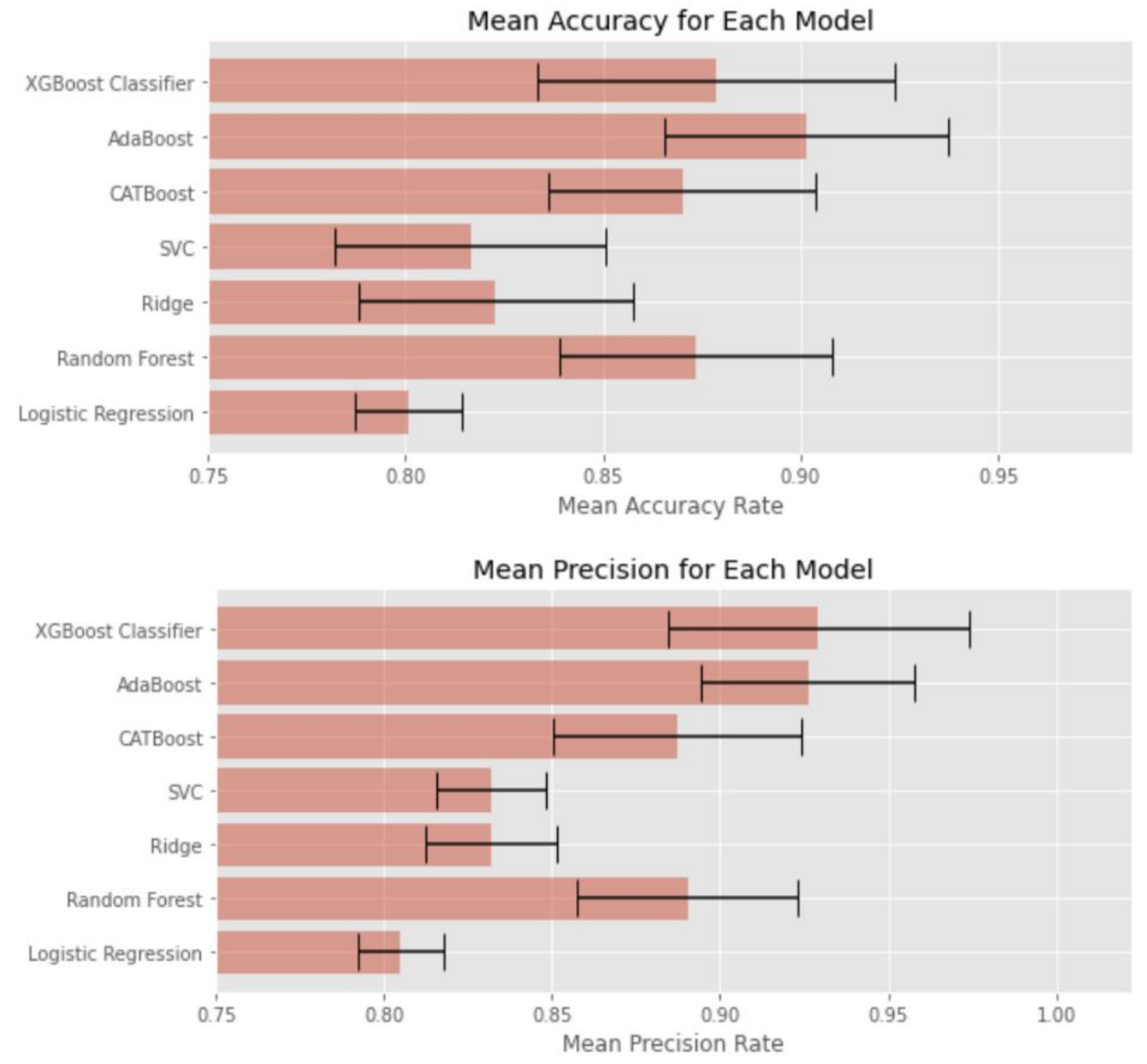

Figure 1: Model Performance

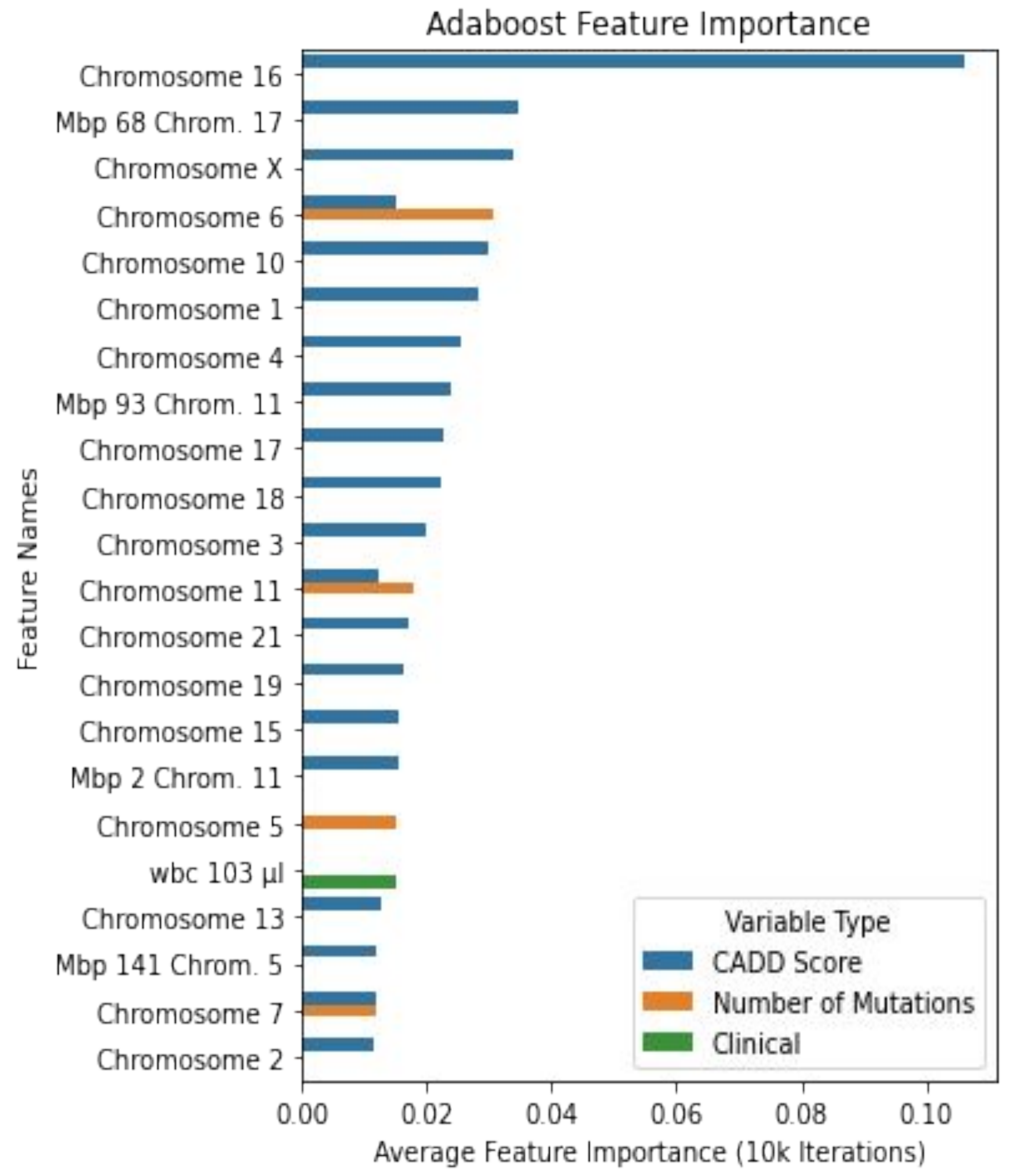

Figure 2 : Genetic Region Feature Importance

### Survival Analysis

Nine stages at diagnosis were assigned to patients in order to find patterns of overall survival. The eight stages were IA, IB, IIA, IIB, IIIA, IIIB, IVA1, IVA2, and IVB. These stages were further grouped into three levels where IA - IB was considered an early stage while IIA-IIB was considered a mid-stage and IIIA - IVB a late stage. The Cox Proportional Hazards regression with LASSO regularization was used for the analysis.

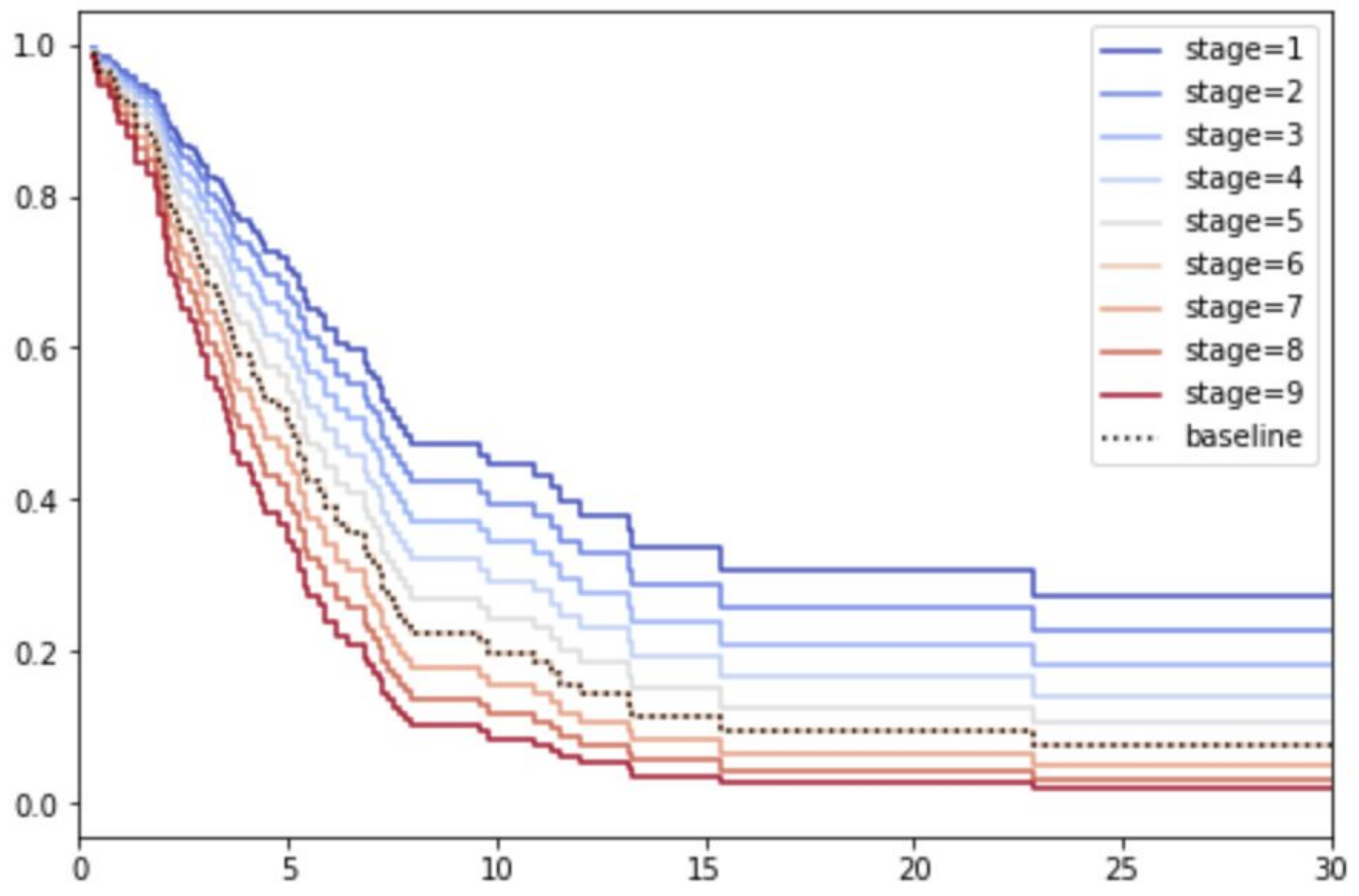

Figure 3: Survival analysis curve for each survival stage

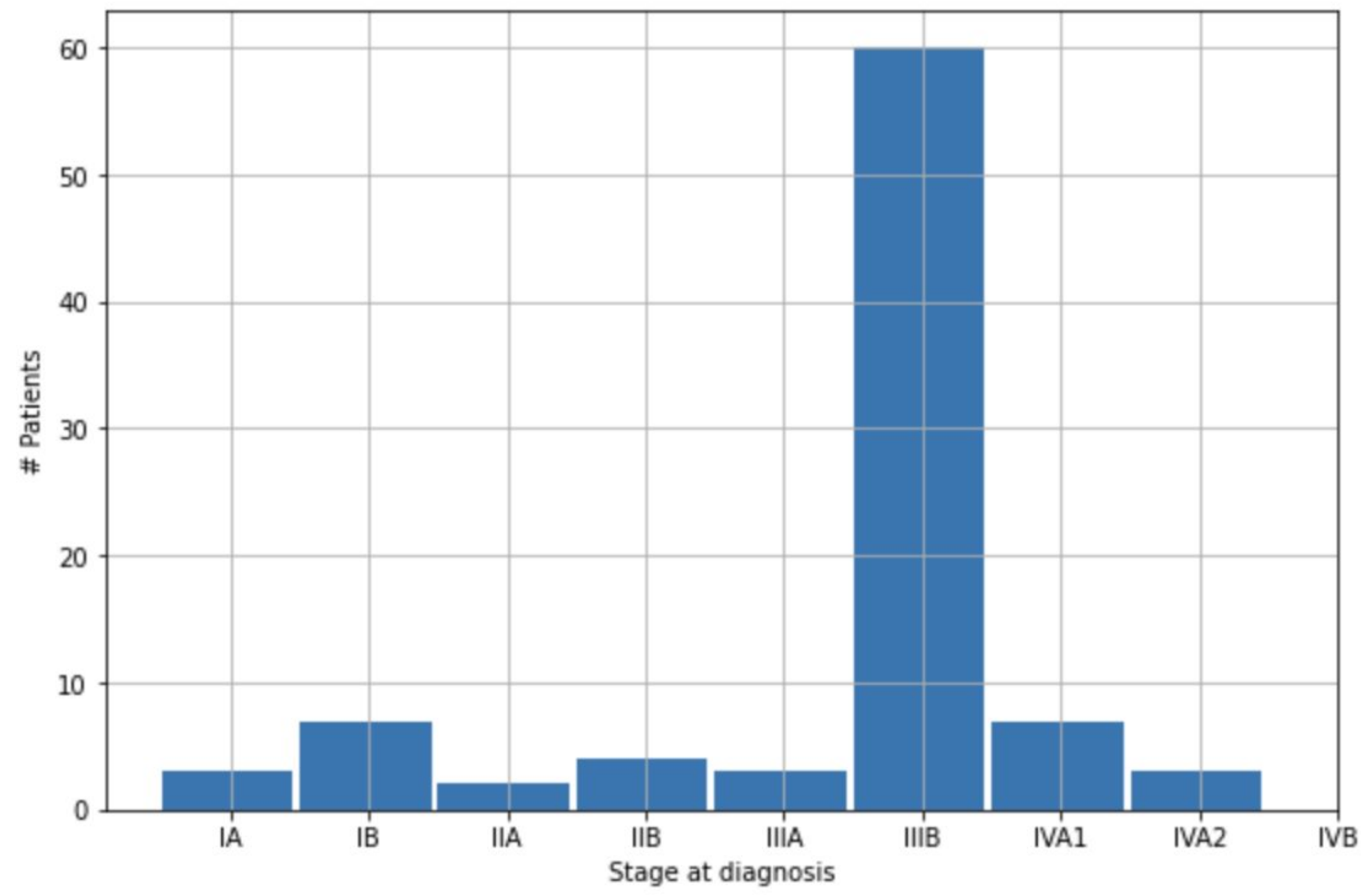

Figure 4: Patient distribution for each survival stage

### NLP

A sub-task of this project was the use of natural language tools to distinguish CTCL patients from non-CTCL patients based on clinical notes vectorized using a TF-IDF vectorizer as well as a BERT-base-uncased embedding model.

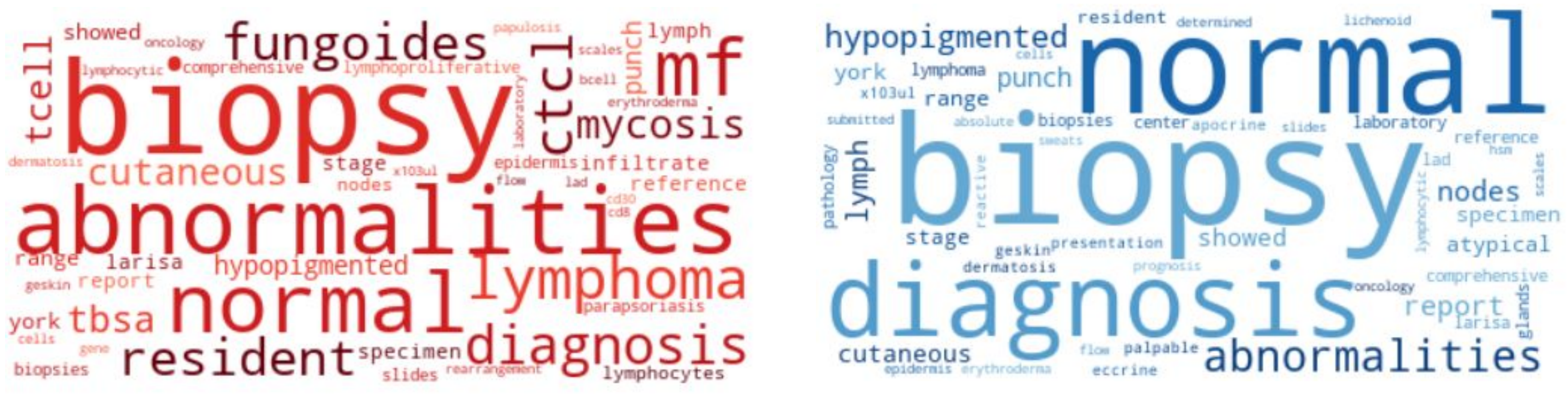

Figure 5: Most Common terms by TF-IDF Frequency. Left - CTCL cases. Right - Controls.

| | # of Patients | | # of Notes | |
|---|---|---|---|---|
| | Filtered | Out of: | Filtered | Out of: |
| Cases | 201 (59%) | 338 | 571 (21%) | 2,676 |
| Controls | 2,300 (74%) | 3,092 | 6,796 (20%) | 33,833 |
| Total | 2,501 (73%) | 3,430 | 7,367 (20%) | 36,509 |

Figure 6: Clinical Notes Dataset Summary

| TF-IDF vectorizer data | |
|---|---|
| Classifier | CTCL F1-Score |
| Logistic Regression | 0.77 |
| **Random Forest** | **0.84** |
| K-Neighbors | 0.70 |
| AdaBoost | 0.80 |
| BERT embeddings | |
| Logistic Regression | 0.46 |
| Random Forest | 0.10 |
| K-Neighbors | 0.28 |
| AdaBoost | 0.28 |

Figure 7: NLP Model Performance on TF-IDF and Bert Embeddings.

## Conclusions & Recommendations

We believe that there is a clear proof of concept for Machine Learning applications for patient-level CTCL data. Regarding next steps, diagnostic modeling may be nearing completion without more data. The greatest added knowledge likely comes from further development of survival analysis and modeling, as well as greater cleaning of clinical notes to make NLP more feasible.

## Acknowledgments

We would like to acknowledge MD Anderson Cancer Center, King's College London, Yale University, Leiden University, Stanford University, and Institut Hospital del Mar, who contributed data to this project.

## References

Chang L-W, Patrone CC, Yang W, et al. An integrated data resource for genomic analysis of cutaneous T-cell lymphoma. *Journal of Investigative Dermatology*. 2018;138(12):2681-2683. doi:10.1016/j.jid.2018.06.176