Leverage OncoKB's Curated Literature Database to Build an NLP Biomarker Identifier

Kritik Seth kls8193@nyu.edu
Saahil Jain sbj7913@nyu.edu
Umair Ayub ua2057@nyu.edu
Abhilash Anand aa9798@nyu.edu

Abstract

The landscape of cancer research has undergone a transformative shift with the advent of technology, causing a large influx of research in the domain. As genomic sequencing becomes an integral component of cancer diagnostics, the need for accurate and up-to-date databases containing information about the oncogenic potential of various genes is paramount. The OncoKB database, a widely utilized resource in the field of oncology, serves as a repository for curated information on the clinical significance of genomic alterations in cancer-related genes. The process of updating this database as new research surfaces is a manual process and time consuming process. This project aims to leverage the power of Machine Learning and Large Language Models to automate the process of annotating genes in research articles for updating the database.

1 Introduction

OncoKB [1] is an oncology database and a widely utilized resource in the field of oncology. It curates a list of human genes along with their variants and oncogenicity - oncogenic, tumor suppressant, both or neither. This information is used by physicians while diagnosing and treating patients based on their genomic sequencing reports. The need to keep this database accurate and updated is of paramount importance.

However, the exponential growth of scientific literature, coupled with the complexity of genomic data, poses a challenge in keeping such databases current. The identification and annotation of genes as oncogenic, tumor suppressant, both, or neutral require meticulous curation, often performed manually by experts in the field. This process is not only time-consuming but also susceptible to human error, limiting the scalability and efficiency of database maintenance.

In response to these challenges, this research attempts to bridge the gap between the burgeoning literature and the imperative need for precision in gene annotation for oncology. We formulate this problem as a Named Entity Recognition problem and leverage large language models such as Biomed-BERT [2] to automatically retrieve genes mentioned in research papers and, subsequently, categorize them based on their oncogenic attributes. The primary goal is to streamline and enhance the curation process, ensuring that the OncoKB database remains comprehensive, accurate, and reflective of the dynamic landscape of cancer genomics.

2 Related Work

Given the extreme specificity of the problem, we observe that there has been very limited work done in the field. PubTator Central [2] tries to annotate entities such as Genes, Disease, Chemicals, Mutations, Species, and CellLine within PubMed articles. However it does not provide any more information about each entity. Given our task is to identify the genes and predict their oncogenicity, PubTator proved to be helpful in our initial phase of identifying the right genes from research articles.

Preprint. Under review.

3 Problem Definition and Algorithm

3.1 Task

The primary objective of our project is to train deep learning models, specifically focusing on identifying genetic mutations and the associated cancer types. This task is anchored in the fine-tuning of advanced language models, such as Biomed-BERT. By fine-tuning these models, we aim to create specialized embeddings that are crucial for the precise identification of genetic alterations. These embeddings then form the foundation of a supervised learning model dedicated to detecting and categorizing genetic mutations.

A critical feature of our approach is ensuring the model's capability in both few-shot and zero-shot learning. This dual proficiency is vital for the model to recognize genes that are rarely mentioned in existing literature as well as to identify completely new genes.

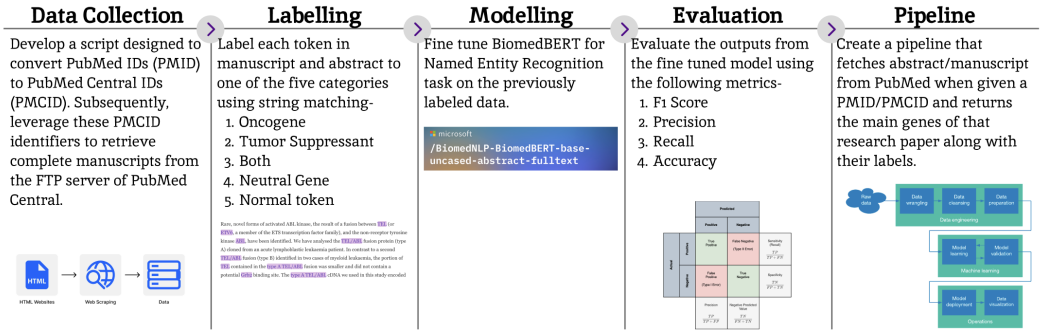

Figure 1: Task Flow

3.2 Algorithm

While building our model we had to optimize for both accuracy and speed. First, we need to have a baseline model which can provide us with a reference point for improvements.

3.2.1 Baseline Model

Our baseline model encompassed two distinct steps. First, the model is trained to discern the classification of each token as either a gene or non-gene. To accomplish this, we utilized a pre-trained PubmedBert, also known as Biomed-Bert (Microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract) [2]. This pre-trained model had been initially trained on abstracts from Pubmed, and we further fine-tuned it to suit our entity recognition task, distinguishing between GENE and NON-GENE entities. Next, the tokens classified as genes are fed as input into a deep neural network that classifies the genes based on their oncogenicity. While this method provides accurate results, its redundancy in the prediction process, involving two steps, leads to decreased computational efficiency.

3.2.2 Final Model

Our final model is a refined one-for-all model which is an entity recognition model designed to predict each label as a different entity. To enhance the efficiency of the model's predictions, we incorporated additional preprocessing steps. Specifically, sentences containing genetic information were retained, contributing to the mitigation of class imbalance to a certain extent. Notably, oncogenic genes were assigned higher loss weights compared to other classes, aligning with their central role in our research focus thereby reducing type II error. The chosen model for this task is BioMed-Bert [2] [3], fine-tuned on our annotated manuscript data. Noteworthy improvements were observed, with this model exhibiting higher F1-score values compared to our baseline. Furthermore, the inference process demonstrated increased speed, attributed to the streamlined use of a singular model for the entirety of the prediction task. This consolidation marks a notable enhancement in both predictive performance and computational efficiency.

3.2.3 Initial Alternate Experiments

While the following experiments were not implemented in the deployment phase, they served as a valuable foundation, offering insightful perspectives into the intricacies of gene modeling.

Naive gene filtering To remove class imbalance we also removed sentences in the abstract that did not contain GENE. This was done using a simple NLTK POS Tagger which kept sentences that only had Proper Nouns (since all genes are tagged as Proper Nouns). However, this does not reduce the imbalance completely but is fairly useful for the learning process.

Representation To represent gene embeddings we started by training two models, the Fasttext model and Spacy en-core-web-sm (English language multi-task Convolutional Neural Network(CNN)) on all the abstracts available to understand medical data embeddings. Both the Fasttext and Spacy model can handle out-of-vocabulary words.

Similarity For inference, each word embedding in the manuscript was compared with a list of gene embeddings using the above models (about 40000 genes collected from OncoKB), and if they exceeded a certain similarity threshold they were classified as GENE else NON-GENE.

Clustering Another method that was used for inference was clustering. Here we used K-means clustering to create two gene clusters based on embeddings obtained from the model and then assigned each token in the manuscript.

These alternate experiments had 1 major drawback, the prediction for each manuscript took about 1-3 minutes which is infeasible when deployed as a production model.

4 Experimental Evaluation

4.1 Data

Our dataset is a comprehensive compilation:

- **PubMed Abstracts (12,278):** Python-scraped abstracts from OncoKB, providing a diverse pool of cancer research insights.
- **PubMed Central Manuscripts (1,545):** Full papers sourced via Python, offering in-depth exploration of methodologies and findings.
- **PMID-Indexed Gene Focus CSV:** Links unique PMIDs to main genes and alterations under study, enhancing precision in research connections.
- **Human Genes and Aliases CSV:** A complete gene reference aiding in standardizing nomenclature across studies.
- **Labeled Gene Classification CSV:** Manually labeled gene classifications (oncogenic, tumor suppressant, both, or none), adding contextual layers to our analysis.

This multifaceted dataset forms the core of our project, facilitating nuanced exploration of cancer genetics and paving the way for impactful insights in our presentation at the conference.

4.2 Methodology

In evaluating our model, our primary focus was on reducing Type II error while maximizing recall. The metrics we emphasized included F1 Score, Precision, and Recall.

- **Type II Error Reduction:** Balancing precision and recall is essential. We adjusted the model's threshold and fine-tuned it to achieve the desired trade-off, considering the project's specific requirements.
- **F1 Score Interpretation:** We provided a detailed interpretation of the F1 score, emphasizing its significance in capturing both precision and recall. Improvements in the F1 score directly contributed to the model's overall effectiveness.

- **Precision and Recall Analysis:** We broke down precision and recall metrics for each gene classification, enabling a granular understanding of the model's performance across different categories. This helped identify areas of strength and areas that required improvement.
- **Cross-Validation:** Implementing cross-validation techniques ensured the model's robustness and generalization across different subsets of the dataset. This approach provided more reliable performance estimates and highlighted potential overfitting concerns.

4.3 Results

To assess the performance of all our approaches, we employed core evaluation metrics applicable to each model developed for Gene Identification and Gene Classification. The metrics used include Precision, Recall, F1 Score, and Accuracy.

For the Baseline Gene Identification task, we presented results from the pre-trained PubMed BERT model fine-tuned on our data. This model identified genes within the abstracts of research papers. Subsequently, we developed a Deep Neural Network for multi-class classification, predicting gene classes (Oncogenic, Tumor Suppressant Gene, Both, Neutral). Our tasks were bifurcated in the baseline while utilizing Name Entity Recognition, we designed the BioMed-BERT model to perform both tasks simultaneously. This approach yielded superior results and was approximately four times faster than our baseline.

Given our project's objective of identifying and classifying genes for prescription decisions, we prioritized Recall over other metrics to address False Negatives. Consequently, our Recall is slightly higher than Precision. Due to imbalances in some classes in our data, overall accuracy remains fairly high.

In addition to these evaluation metrics, we predict confidence levels for each task, providing an extra layer of assurance for researchers. OncoKB utilizes our BioMed-BERT model in its daily operations, appreciating the added confidence in its deployment and usage.

	Task	Precision	Recall	F1 Score	Accuracy
Baseline	Gene Identification	0.570	0.780	0.717	0.823
Baseline	Gene Classification	0.620	0.640	0.610	0.620
Final	Gene Identification & Classification	**0.891**	**0.915**	**0.903**	**0.992**

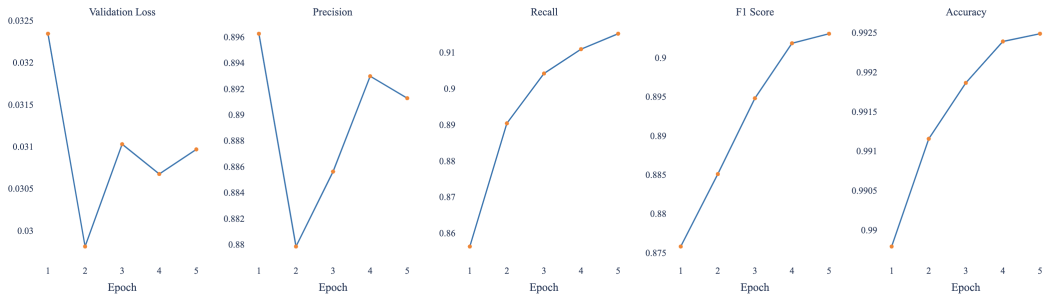

Figure 2: Evaluation Metrics vs Epochs

4.4 Discussion

As we embarked on implementing the baseline models, our intuition led us to explore the potential for significantly improved performance through the utilization of Language Model Models (LLMs) pre-trained on medical data. As we scaled up the size of our models, we observed a clear correlation between model size and enhanced evaluation metrics. Notably, the Biomed-BERT model emerged as the top performer, garnering praise from OncoKB researchers. This success can be attributed to the model's pre-training on PubMed data, specifically tailored for our use case.

Through the implementation of fine-tuning over just a few epochs, our model demonstrated impressive results. It became apparent, however, that the model started to exhibit signs of overfitting beyond 8

epochs. Despite this, we are confident that our model is robust enough for deployment in production. To ensure its readiness, the model will undergo further scrutiny and internal testing within OncoKB, ensuring its suitability for integration into their daily workflow.

Moreover, we recognize the importance of continuous refinement and evaluation. The model's performance will be closely monitored, and any necessary adjustments will be made to maintain its efficacy in real-world applications. This iterative process underscores our commitment to delivering a reliable and impactful solution to OncoKB.

5 Conclusions

At the end of this capstone project, we achieved the development of a refined Named Entity Recognition (NER) model tailored for gene identification and classification. The model extends its capabilities beyond mere gene identification, providing classifications such as Gene, Oncogene, Tumor Suppressor, or Both.

A notable outcome of our efforts is the creation of an end-to-end pipeline, a cohesive system that fetches and processes abstracts when provided with PubMed IDs (PMID). This pipeline incorporates preprocessing steps before leveraging our NER model to extract and classify genes from the abstracts. The imminent deployment of this pipeline for internal use marks a practical application of our model's capabilities and potentially makes efficient the OncoKB workflow.

The process of updating the database involves fetching the latest articles from PubMed, inputting them into our model, and extracting a curated list of genes with their respective oncogenic classifications. This list undergoes meticulous verification by researchers before being employed to update the OncoKB database.

While the Biomed-BERT model showcased commendable performance on existing labeled data, the validation process with new data from OncoKB researchers highlighted specific areas for enhancement. We are actively addressing these identified shortcomings, striving to augment the model's precision and versatility.

In summation, this project delivers not only a sophisticated NER model but also an efficient end-to-end pipeline. The imminent deployment of this solution holds the promise of enhancing the efficiency and accuracy of database updates, ensuring OncoKB remains at the forefront of incorporating the latest research findings.

Areas of Future Work In the case of the continuation of this project, our next step includes predicting gene alteration to effectively aid Clinicians in assigning the right treatment. This is an extension of our current pipeline where we can try and generate a graph to map genes and their alterations. We would also like to delve deeper into the entity recognition models in the LLM space for medical data. Experimenting with various architectures such as T5 and Llama 2 could aid in utilizing a singular model to produce a natural language output. Since the current model, Biomed-Bert is going to be deployed into production, speed is an added parameter to be considered. Finding the optimal tradeoff between model speed and performance is a crucial next step. To do this we have to compare models of different sizes (Example: T5 base (0.2B) versus T5 (3B)).

Shortcomings and Solution One of the main shortcomings in our data process is Data Labeling. With fuzzy string matching, we were able to maneuver around this to some extent. This however doesn't change the fact that we are still working with noisy data. Though it is infeasible to manually annotate every token from scratch, manual assistance in the verification of labels might help improve the quality of the labeled data. Another limitation of our model which it currently is incapable of doing is creating an analysis report by searching for all the manuscripts given a query gene. This requires thorough contextual understanding across sparse data which is difficult and resource expensive.

6 Lessons learned

Initially, we had limited knowledge about the domain and we had to research quite a lot about the different aspects of genomics to identify the correct approach to solve this problem. This helped us

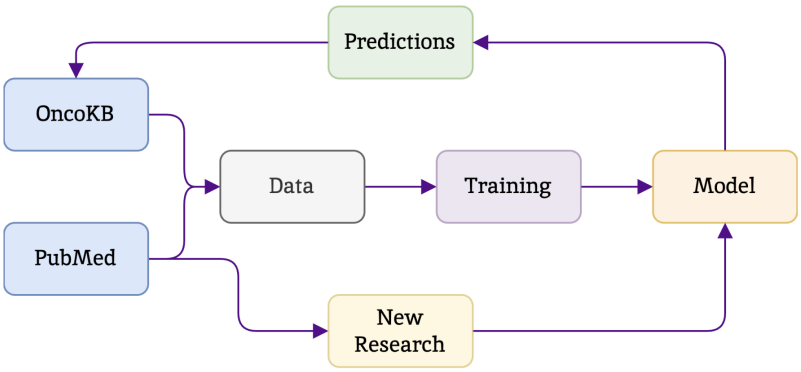

Figure 3: End-to-end workflow

acquire skills in applying Natural Language Processing (NLP) techniques, including Large Language Models (LLMs), to address challenges in the healthcare domain.

One more problem that we faced and had to maneuver around was scraping the large amounts of manuscripts. We had to extract thousands of manuscripts using web scraping but each of those files was in tar.gz. files which were usually 10 GB in size or more. We learned how to deal with files of such large sizes and store them efficiently.

We learned how to deal effectively with sparse data. When formulating the gene identification problem using the word similarity approach. When doing this we had to calculate cosine similarities of embeddings of each word in our abstract to every gene in our database and doing this was very slow. We solved this problem by POS Tagging and only finding similarities between a word and a gene if the word was a proper noun.

7 Student Contributions

1. Kritik Seth - Data ETL, Baseline (Gene Identification, Gene Classification), Final Model, End-to-end Pipeline
2. Saahil Jain - Data ETL, Baseline (Gene Identification, Gene Classification), Final Model, End-to-end Pipeline
3. Abhilash Anand - Initial Model Exploration, Baseline (Gene Identification, Gene Classification), Final Model
4. Umair Ayub - Initial Model Exploration, Baseline (Gene Identification, Gene Classification), Final Model

References

[1] et. al. Debyani Chakravarty. Oncokb: A precision oncology knowledge base. *JCO Precision Oncology*, 2017.

[2] Hao Cheng Michael Lucas Naoto Usuyama Xiaodong Liu Tristan Naumann Jianfeng Gao Yu Gu, Robert Tinn and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *arXiv:2007.15779*, 2020.

[3] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, September 2019.