GENOME MUTATION & CANCER CLASSIFIER

Team Detail

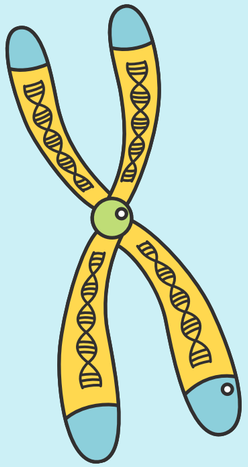
Abhishek Sahay	S20220010003
Prajwal Kumar	S20220010178
Sahil Goyat	S20220010190
M Mamatha	S20220010143

ABOUT OUR PROJECT

CarciTrack🔗: **Gene Mutation & Cancer Classifier**

Problem Statement

Early detection of cancer is difficult due to the complexity of genetic data. Identifying mutations in cancer genes manually is time-consuming and error-prone, necessitating an automated, accurate approach to improve diagnosis and treatment planning.



Motivation

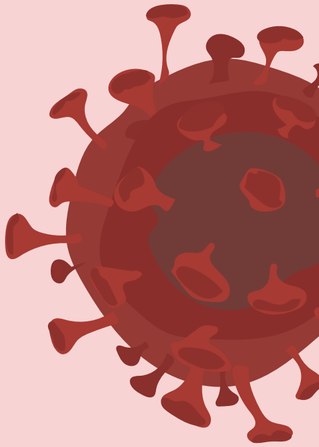
With rising cancer cases globally, there’s a critical need for efficient diagnostic tools. Leveraging deep learning to identify gene mutations can accelerate early detection, reduce human error, and support personalized treatment, ultimately improving patient survival and care quality.

INTRODUCTION

This project focuses on three major cancer types:

- BLCA (Bladder Urothelial Carcinoma)
- LUSC (Lung Squamous Cell Carcinoma)
- KIRC (Kidney Renal Clear Cell Carcinoma)

These cancers are linked to specific genetic mutations such as *Missense, Nonsense, Silent*. By analyzing mutation patterns using deep learning, the project aims to support early diagnosis and personalized treatment strategies for better patient outcomes.



RESEARCH PAPERS

BASE PAPER

Paper 1

A hybrid machine learning model for classifying gene mutations in cancer using LSTM, BiLSTM, CNN, GRU, and GloVe

REFERENCE PAPER

Paper 3

Identification of 12 cancer types through genome deep learning

REFERENCE PAPER

Paper 5

Cancer gene mutation frequencies for the U.S. population

REFERENCE PAPER

Paper 2

Signatures of mutational processes in human cancer

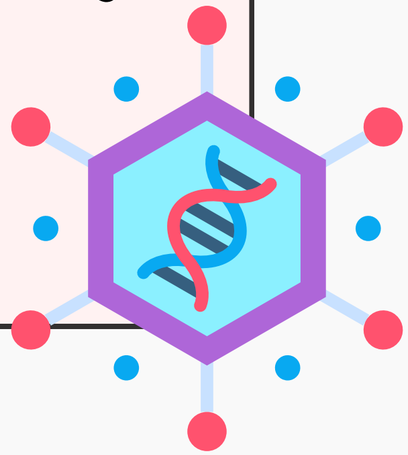
REFERENCE PAPER

Paper 4

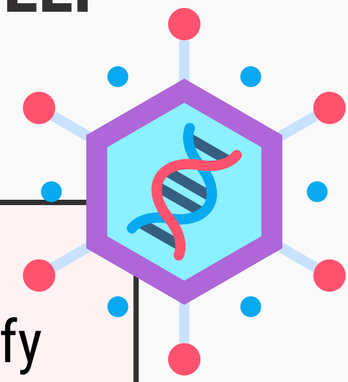
Mutation Detection in Genes Sequence Using Machine Learning

A HYBRID MACHINE LEARNING MODEL FOR CLASSIFYING GENE MUTATIONS IN CANCER USING LSTM, BILSTM, CNN, GRU, AND GLOVE

- General classification of gene mutations in cancer, rather than targeting a specific type of cancer. It uses the "Personalized Medicine: Redefining Cancer Treatment" dataset from Kaggle.The hybrid model's performance underscores its potential in enhancing the precision of cancer diagnoses and treatments, contributing significantly to the advancement of personalized healthcare.
- Accuracy and Results: The model achieved a training accuracy of 80.6%, precision of 81.6%, recall of 80.6%, and an F1 score of 83.1%, outperforming several transformer-based models.
- Models Used: A combination of LSTM, BiLSTM, CNN, GRU, and GloVe embeddings.



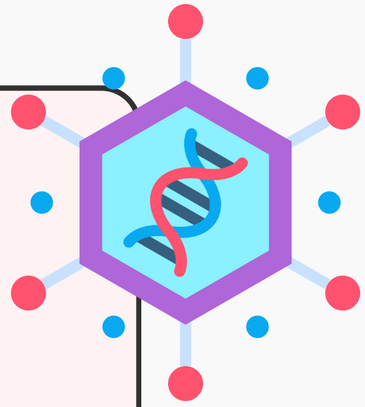
IDENTIFICATION OF 12 CANCER TYPES THROUGH GENOME DEEP LEARNING



Overview: This research proposes a Genome Deep Learning (GDL) method to identify cancer types based on genomic variations. By analyzing whole-exome sequencing (WES) data, the study aims to improve cancer diagnosis through genomic information.

- Accuracy and Results: The specific models achieved an accuracy of 97.47%, the total-specific model 94.70%, and the mixture model 70.08%. These high accuracies demonstrate the effectiveness of the GDL approach in cancer identification.
- Models Used: The study employs deep neural networks tailored to distinguish between specific cancer types and healthy tissues. It constructs 12 specific models for individual cancer types, a total-specific model for cancer vs. healthy identification, and a mixture model for distinguishing among all 12 cancer types.

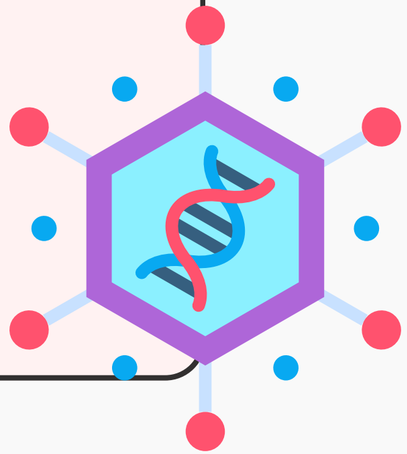
CANCER GENE MUTATION FREQUENCIES FOR THE U.S. POPULATION



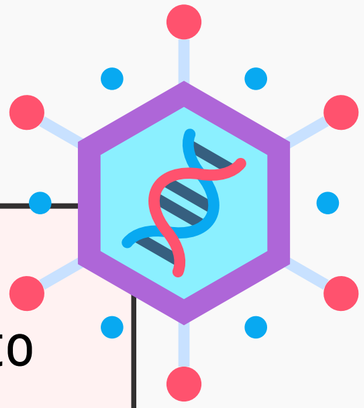
- Overview: This study provides estimates of the prevalence of specific gene mutations across various cancer types in the U.S. population, combining genomic and epidemiological data to inform public health and research.
- Analysis: The research utilizes statistical analysis to estimate mutation frequencies, focusing on the proportion of cancer cases with specific gene mutations.
- Accuracy and Results: Findings indicate that TP53 is the most commonly mutated gene (35%), followed by PIK3CA (13%), KRAS (11%), and BRAF (8%). The study also notes that epigenetic regulators like KMT2C, KMT2D, and ARID1A are among the top ten most commonly mutated driver genes.

SIGNATURES OF MUTATIONAL PROCESSES IN HUMAN CANCER

- Overview: This paper investigates the patterns of somatic mutations across various cancer types to identify distinct mutational signatures associated with different mutational processes.
- Models Used: The study employs computational methods to analyze somatic mutation data, extracting mutational signatures through statistical modeling.
- Accuracy and Results: The analysis reveals over 20 distinct mutational signatures, each associated with specific mutational processes such as aging, exposure to carcinogens, or defects in DNA repair mechanisms.
- Key Insights: Understanding these mutational signatures provides insights into the etiology of cancers & can inform prevention strategies and therapeutic interventions.

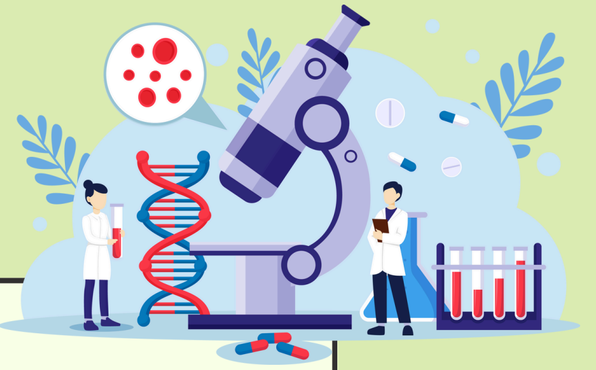


MUTATION DETECTION IN GENE SEQUENCES USING MACHINE LEARNING



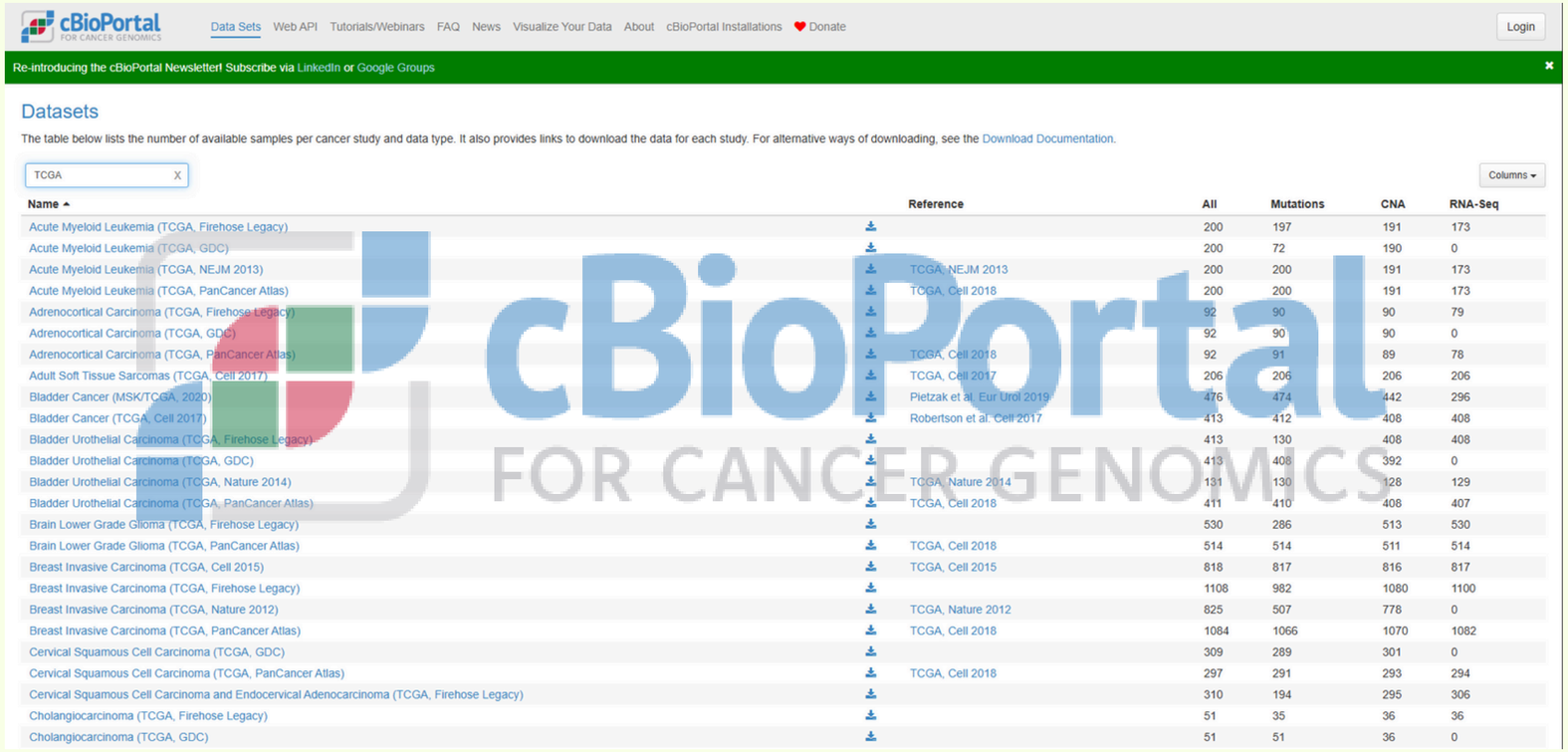
- Overview: This study explores the application of machine learning techniques to detect mutations in gene sequences, aiming to improve the accuracy and efficiency of mutation detection.
- Models Used: The research utilizes various machine learning algorithms, including Support Vector Machines (SVM), Decision Trees, and Neural Networks, to classify gene sequences and detect mutations.
- Accuracy and Results: The models demonstrated high accuracy in detecting mutations, with performance varying based on the algorithm and dataset used.
- Key Insights: The study highlights the potential of machine learning approaches in automating and enhancing the detection of gene mutations, which is crucial for genetic research and clinical diagnostics.

DATASET



Dataset source: [cBioPortal for Cancer Genomics]

URL🔗: https://www.cbioportal.org/datasets



cBioPortal FOR CANCER GENOMICS Data Sets Web API Tutorials/Webinars FAQ News Visualize Your Data About cBioPortal Installations Donate Login

Re-introducing the cBioPortal Newsletter! Subscribe via LinkedIn or Google Groups

Datasets

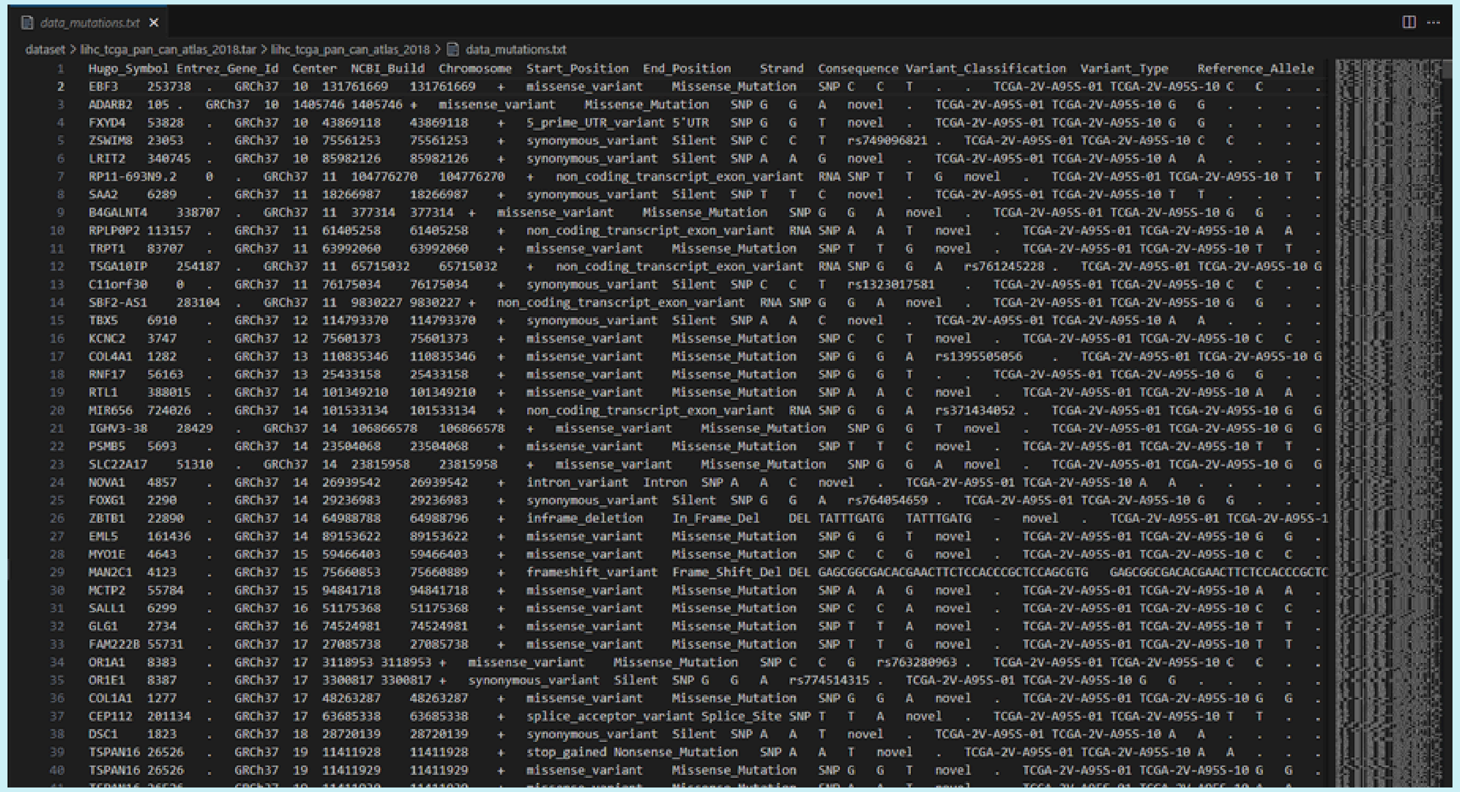
The table below lists the number of available samples per cancer study and data type. It also provides links to download the data for each study. For alternative ways of downloading, see the Download Documentation.

TCGA X Columns

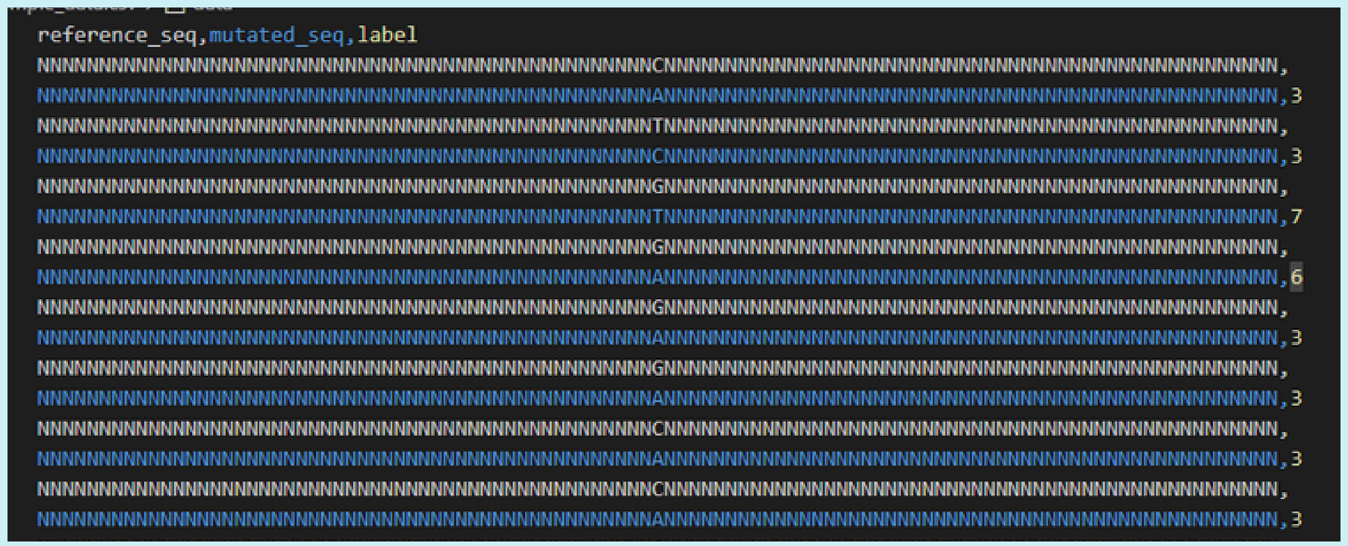
Name	Reference	All	Mutations	CNA	RNA-Seq
Acute Myeloid Leukemia (TCGA, Firehose Legacy)		200	197	191	173
Acute Myeloid Leukemia (TCGA, GDC)		200	72	190	0
Acute Myeloid Leukemia (TCGA, NEJM 2013)	TCGA, NEJM 2013	200	200	191	173
Acute Myeloid Leukemia (TCGA, PanCancer Atlas)	TCGA, Cell 2018	200	200	191	173
Adrenocortical Carcinoma (TCGA, Firehose Legacy)		92	90	90	79
Adrenocortical Carcinoma (TCGA, GDC)		92	90	90	0
Adrenocortical Carcinoma (TCGA, PanCancer Atlas)	TCGA, Cell 2018	92	91	89	78
Adult Soft Tissue Sarcomas (TCGA, Cell 2017)	TCGA, Cell 2017	206	206	206	206
Bladder Cancer (MSK/TCGA, 2020)	Pietzak et al. Eur Urol 2019	476	474	442	296
Bladder Cancer (TCGA, Cell 2017)	Robertson et al. Cell 2017	413	412	408	408
Bladder Urothelial Carcinoma (TCGA, Firehose Legacy)		413	130	408	408
Bladder Urothelial Carcinoma (TCGA, GDC)		413	408	392	0
Bladder Urothelial Carcinoma (TCGA, Nature 2014)	TCGA, Nature 2014	131	130	128	129
Bladder Urothelial Carcinoma (TCGA, PanCancer Atlas)	TCGA, Cell 2018	411	410	408	407
Brain Lower Grade Glioma (TCGA, Firehose Legacy)		530	286	513	530
Brain Lower Grade Glioma (TCGA, PanCancer Atlas)	TCGA, Cell 2018	514	514	511	514
Breast Invasive Carcinoma (TCGA, Cell 2015)	TCGA, Cell 2015	818	817	816	817
Breast Invasive Carcinoma (TCGA, Firehose Legacy)		1108	982	1080	1100
Breast Invasive Carcinoma (TCGA, Nature 2012)	TCGA, Nature 2012	825	507	778	0
Breast Invasive Carcinoma (TCGA, PanCancer Atlas)	TCGA, Cell 2018	1084	1066	1070	1082
Cervical Squamous Cell Carcinoma (TCGA, GDC)		309	289	301	0
Cervical Squamous Cell Carcinoma (TCGA, PanCancer Atlas)	TCGA, Cell 2018	297	291	293	294
Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (TCGA, Firehose Legacy)		310	194	295	306
Cholangiocarcinoma (TCGA, Firehose Legacy)		51	35	36	36
Cholangiocarcinoma (TCGA, GDC)		51	51	36	0

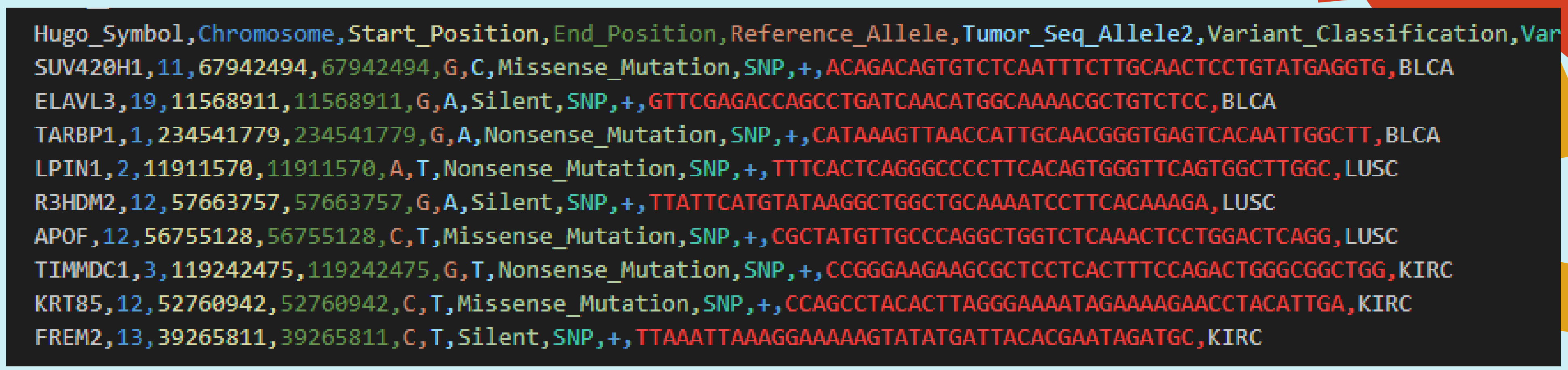
PREPROCESSING

Mutation data from individual cancer folder downloaded from cBioPortal (data_mutations.txt) is processed by querying the Ensembl API to retrieve authentic genomic sequences surrounding each mutation site. The resulting data is further cleaned by removing incomplete entries and saved into a CSV format (mutations.csv) for downstream analysis and model training.



data_mutations.txt





mutations.csv

MODEL ARCHITECTURE

```
Mutation type distribution:
Variant_Classification
Missense_Mutation    22512
Silent               20348
Nonsense_Mutation    18499
Name: count, dtype: int64
Cancer type distribution:
Cancer_Type
BLCA    20667
LUSC    20645
KIRC    12499

Name: count, dtype: int64
Mutation classes: ['Missense_Mutation' 'Nonsense_Mutation' 'Silent']
Cancer classes: ['BLCA' 'KIRC'         'LUSC']
Number of mutation classes: 3
Number of cancer classes: 4
```

Mutation Classes

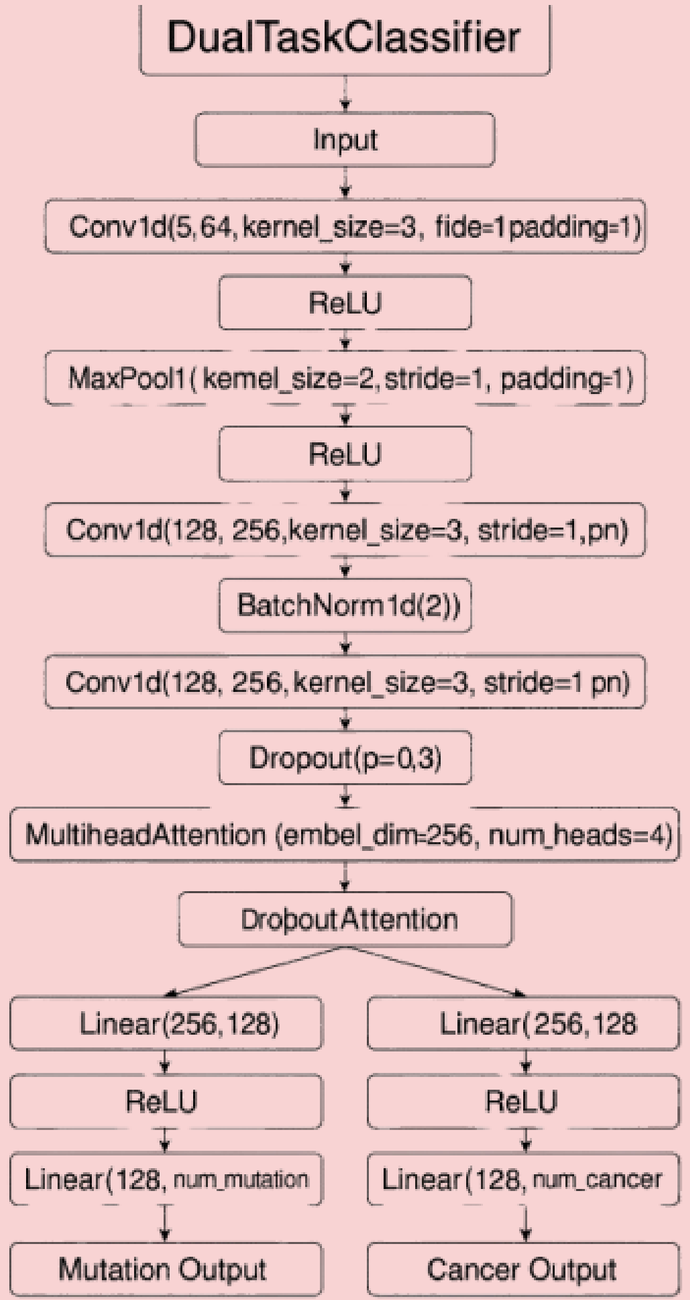
Missense Mutations - 22512
Silent - 20348
Nonsense Mutations - 18499

Cancer Classes

BLCA - 20667
LUSC - 20645
KIRC - 12499

Extracted Features

- **Hugo_Symbol**
- **Chromosome**
- **Start_Position**
- **End_Position**
- **Reference_Allele**
- **Tumor_Seq_Allele2**
- **Variant_Classification**
- **Variant_Type**
- **Strand**
- **Genomic_Context_Sequence**
- **Cancer_Type**

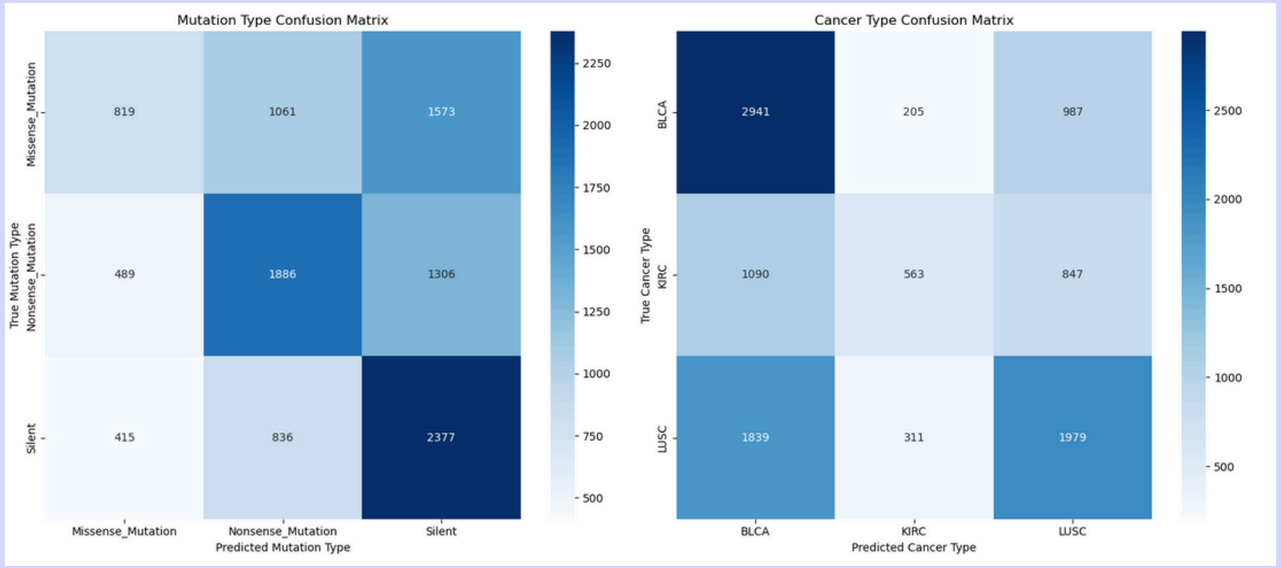


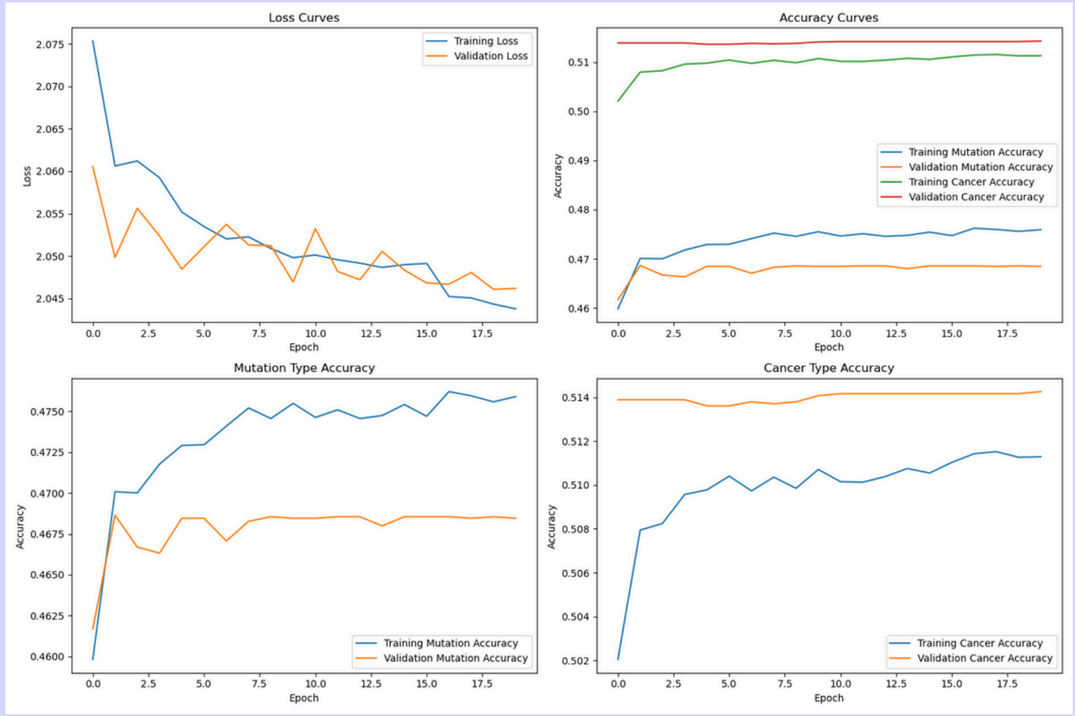
TRAINING RESULTS

After cleaning and preprocessing the dataset from cBioPortal, our best model achieved

Cancer Accuracy: 0.4746 ± 0.0073
Mutation Accuracy: 0.5120 ± 0.0020

```
Cross-Validation Results:
Fold 1: Mutation Accuracy = 0.4613, Cancer Accuracy = 0.5109
Fold 2: Mutation Accuracy = 0.4817, Cancer Accuracy = 0.5136
Fold 3: Mutation Accuracy = 0.4789, Cancer Accuracy = 0.5149
Fold 4: Mutation Accuracy = 0.4786, Cancer Accuracy = 0.5111
Fold 5: Mutation Accuracy = 0.4722, Cancer Accuracy = 0.5095
```





REPORT

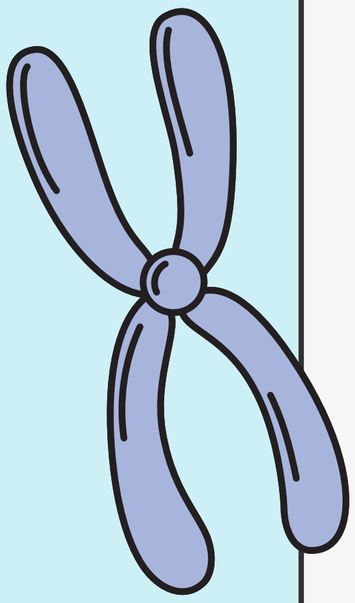
Merits

Our model provides valuable insight into mutation patterns across multiple cancer types—BLCA, LUSC, and KIRC—using real patient data from cBioPortal. It incorporates relevant biological features and demonstrates decent accuracy, as shown in training curves and confusion matrices. The use of genomic sequence context and variant classification enriches prediction quality, making it potentially useful for cancer research and early diagnostics.

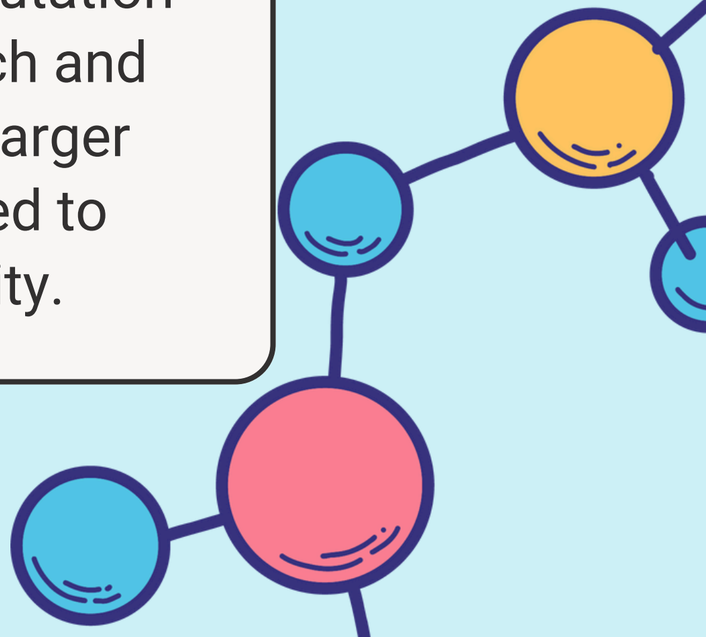
De-Merits

Despite promising results, the dataset suffers from class imbalance, which may bias predictions. Additionally, the model's generalizability is limited to the selected cancers and mutation types. Real-world application would require more extensive validation and inclusion of additional biological factors like gene expression or epigenetic markers.

CONCLUSION



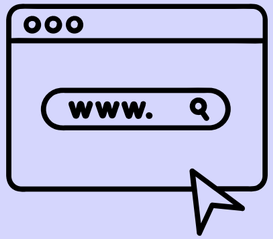
In this project, we successfully analyzed mutation data for three major cancer types: BLCA, LUSC, and KIRC. By extracting key genomic features and applying deep learning models, we achieved meaningful predictions about mutation classifications. Our results show that understanding mutation patterns can significantly contribute to cancer research and early diagnosis. However, further improvements like larger datasets and deeper biological integration are needed to enhance model accuracy and real-world applicability.



REFERENCES

[1] https://www.sciencedirect.com/science/article/pii/S2772941924000395
[2] https://pubmed.ncbi.nlm.nih.gov/31754222/
[3] https://pubmed.ncbi.nlm.nih.gov/34645806/
[4] https://pubmed.ncbi.nlm.nih.gov/23945592/
[5] https://ieeexplore.ieee.org/document/9693088
[6] https://www.cbioportal.org/datasets
[7] https://rest.ensembl.org/

WEBSITE



[CarciTrack: Genome Mutation & Cancer Classifier](#)

[Github](#)

Click Here

Share

Gene Mutation & Cancer Classifier

Predict mutation type and associated cancer type from genomic data

Input Data

Genomic Context Sequence

Enter DNA sequence (e.g., ATGCGTACGTAGCTAGCTAGCT...)

Reference Allele

e.g., A

Tumor Allele

e.g., G

Predict Mutation & Cancer Type

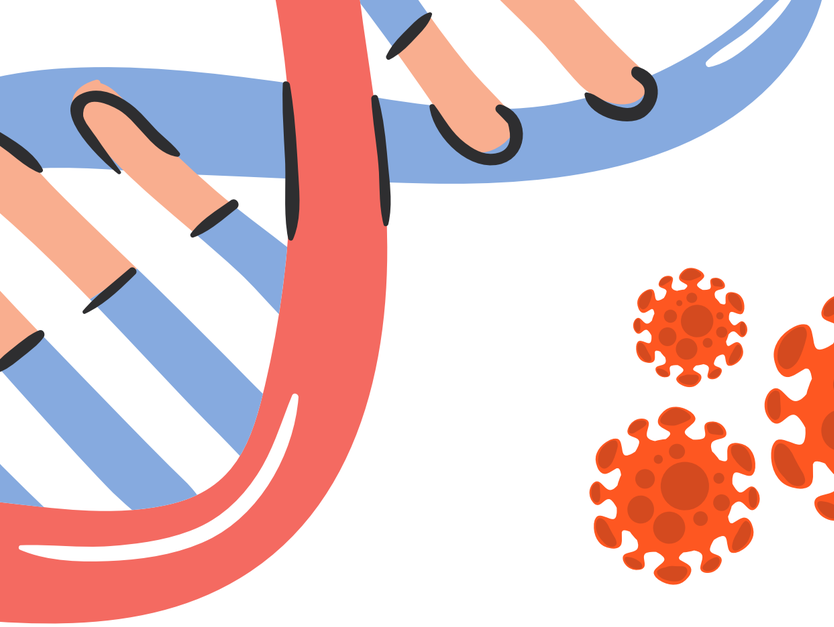
Enter a genomic sequence and alleles, then click 'Predict Mutation & Cancer Type' to see results.

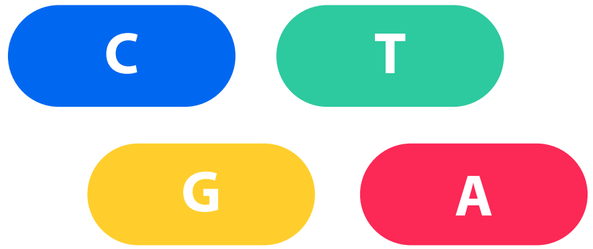
See Example Input

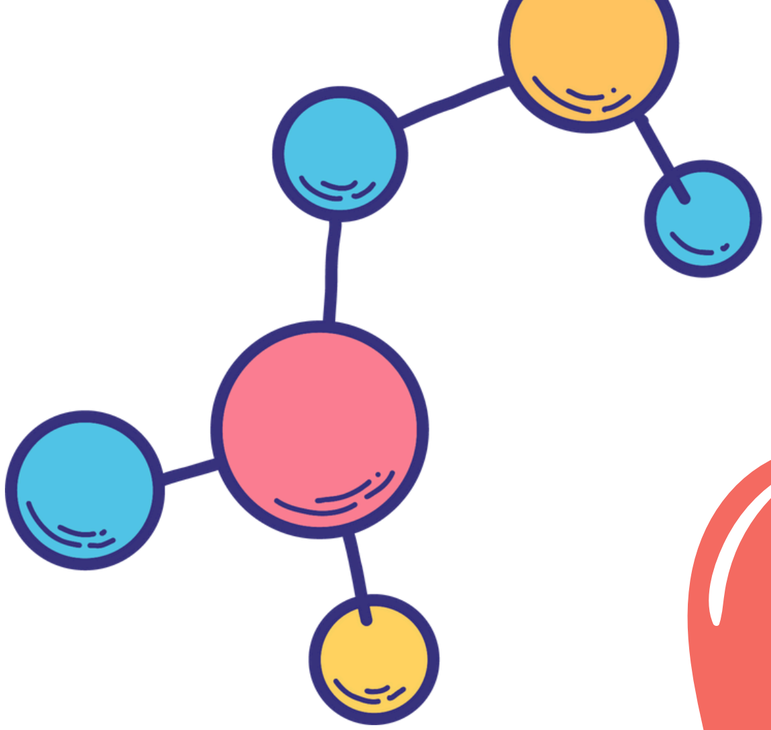
About Mutation & Cancer Types

Gene Mutation & Cancer Type Classifier | Created with PyTorch and Streamlit

Manage app

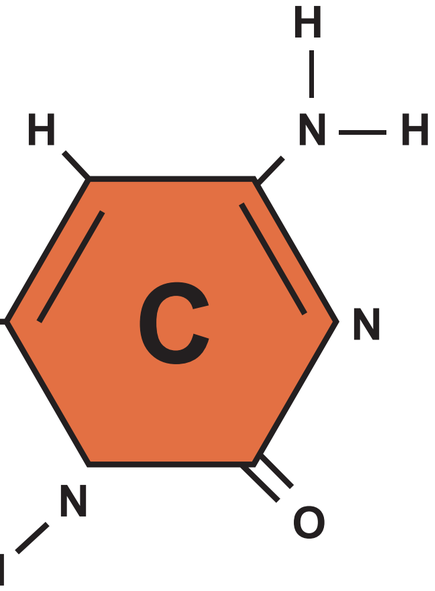


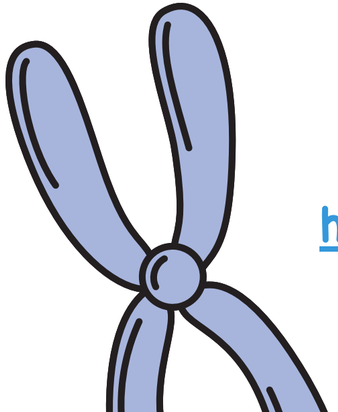


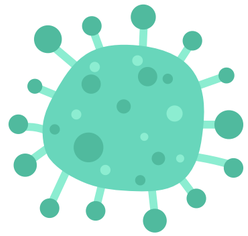


THANK YOU

CarciTrack







https://carcitrack.streamlit.app