

GENOME MUTATION & CANCER CLASSIFIER

Team Detail

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

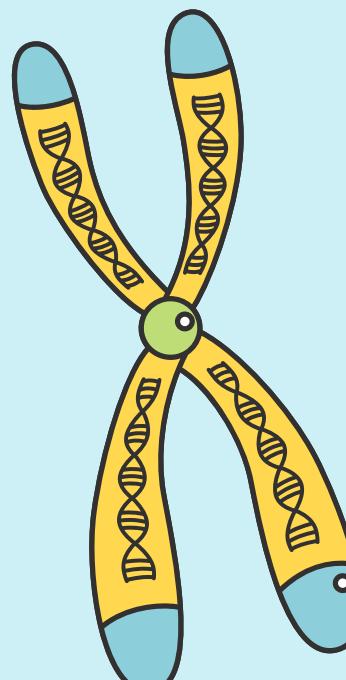
ABOUT OUR PROJECT

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

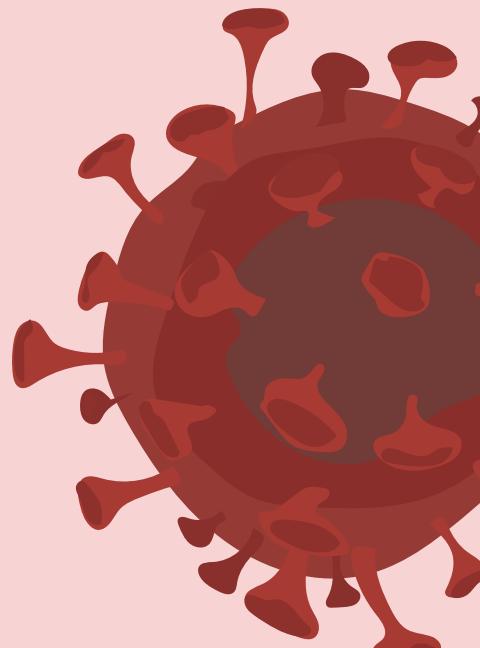
With rising cancer cases globally, there's a critical need for efficient diagnostic tools. Leveraging deep learning to identify gene mutations can accelerate early detection, reduce human error, and support personalized treatment, ultimately improving patient survival and care quality.

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



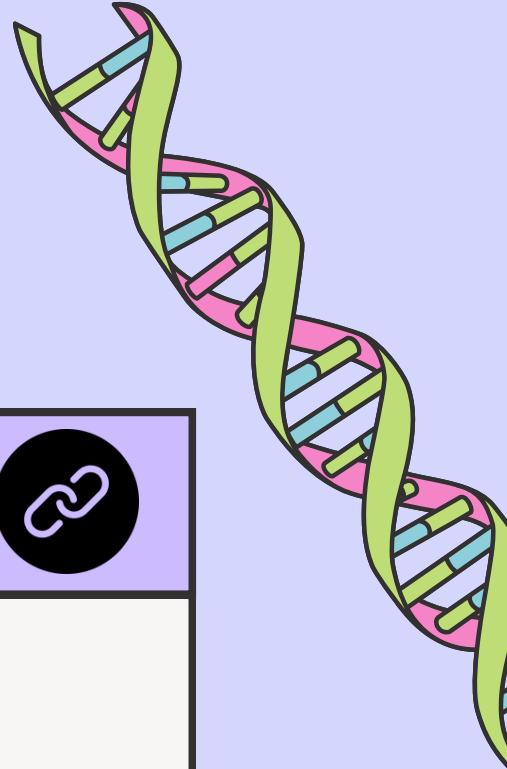
Signatures of mutational processes in human cancer

REFERENCE PAPER

Paper 4

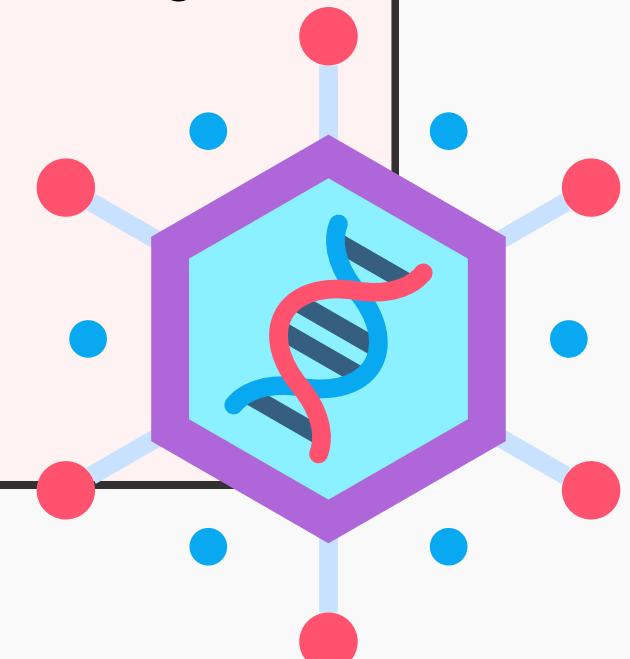


Mutation Detection in Genes Sequence Using Machine Learning



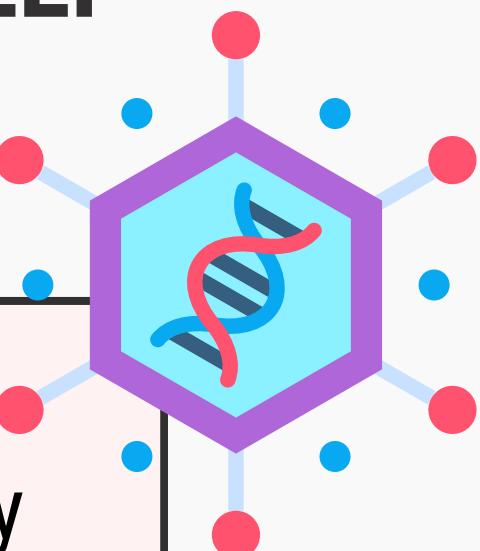
A HYBRID MACHINE LEARNING MODEL FOR CLASSIFYING GENE MUTATIONS IN CANCER USING LSTM, BILSTM, CNN, GRU, AND GLOVE

- 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.



RESEARCH

IDENTIFICATION OF 12 CANCER TYPES THROUGH GENOME DEEP LEARNING



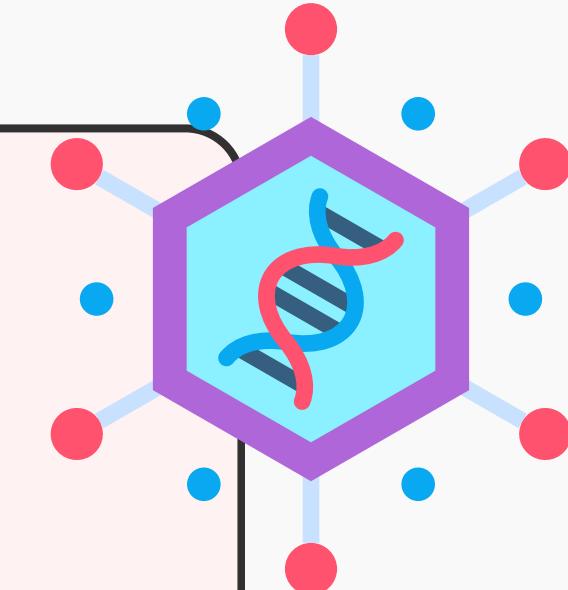
Overview: This research proposes a Genome Deep Learning (GDL) method to identify cancer types based on genomic variations. By analyzing whole-exome sequencing (WES) data, the study aims to improve cancer diagnosis through genomic information.

- **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.

RESEARCH

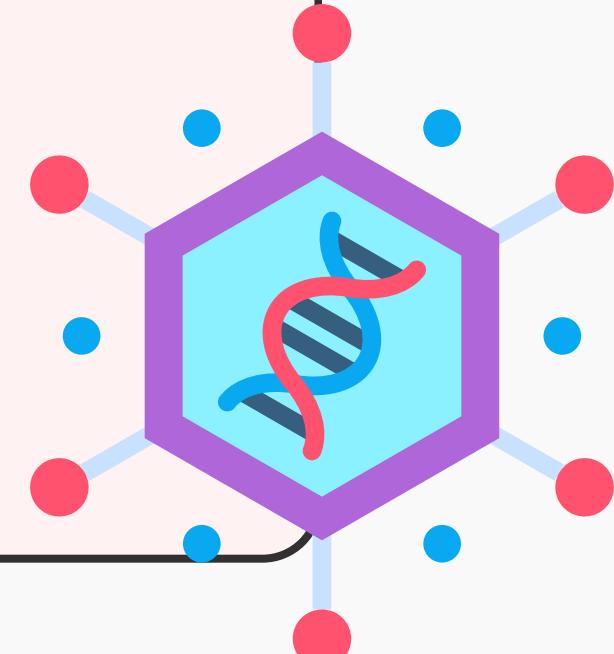
CANCER GENE MUTATION FREQUENCIES FOR THE U.S. POPULATION

- 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.



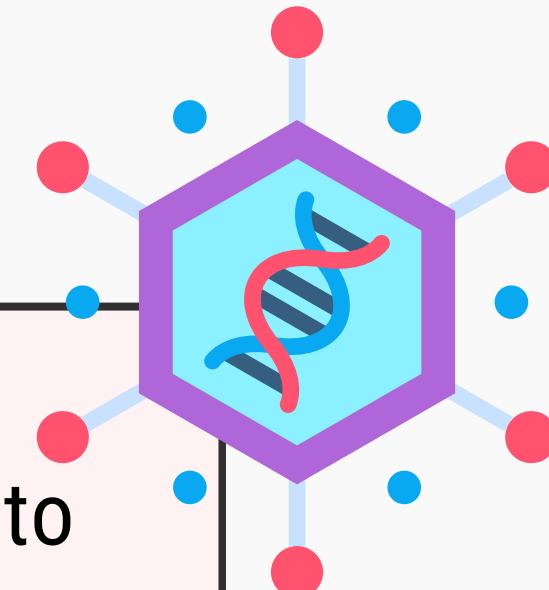
SIGNATURES OF MUTATIONAL PROCESSES IN HUMAN CANCER

- 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.



RESEARCH

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](https://www.cbioportal.org/datasets)
URL : <https://www.cbioportal.org/datasets>

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](#).

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

PREPROCESSING



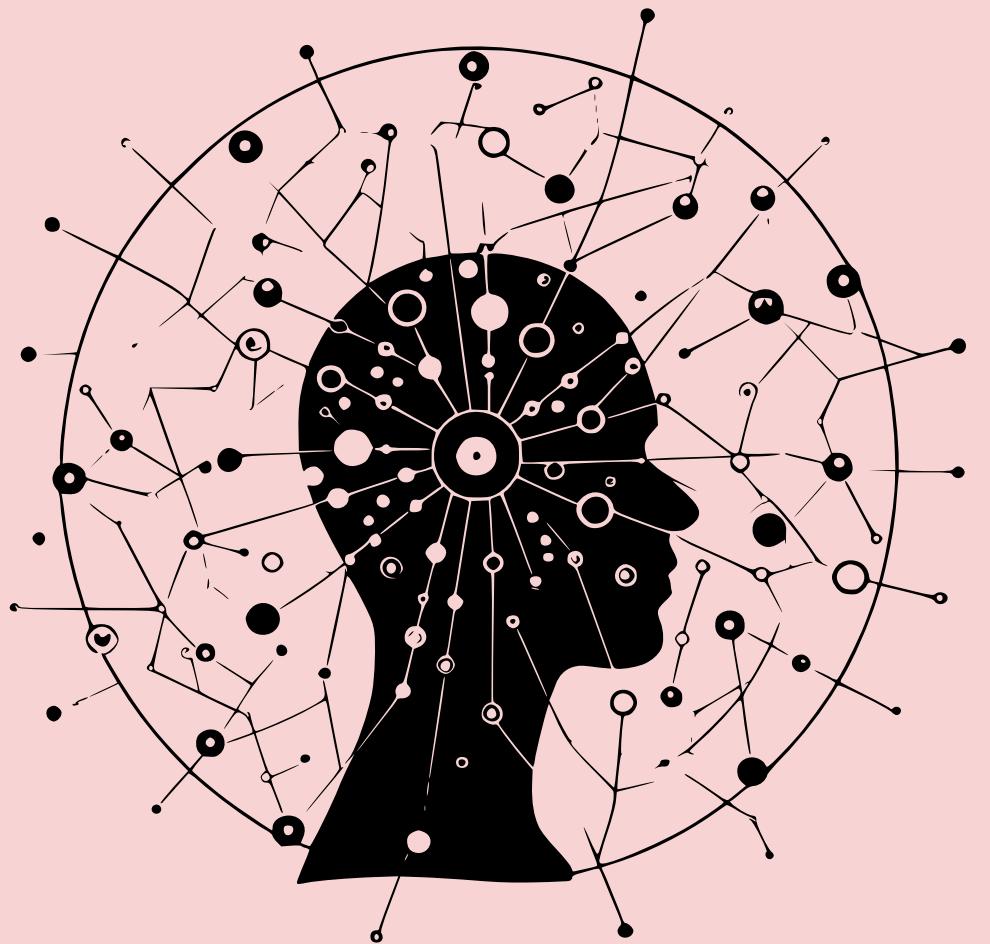
data_mutations.txt																	
dataset > lihc_tcga_pan_can_atlas_2018.tar > lihc_tcga_pan_can_atlas_2018 > data_mutations.txt	Hugo_Symbol	Entrez_Gene_Id	Center	NCBI_Build	Chromosome	Start_Position	End_Position	Strand	Consequence	Variant_Classification	Variant_Type	Reference_Allele	Alt	Filter			
1	EBF3	253738	.	GRCh37	10	131761669	131761669	+	missense_variant	Missense_Mutation	SNP	C	T	.	TCGA-2V-A955-01 TCGA-2V-A955-10 C C . . .		
2	ADARB2	105	.	GRCh37	10	1405746	1405746	+	missense_variant	Missense_Mutation	SNP	G	A	novel	.	TCGA-2V-A955-01 TCGA-2V-A955-10 G G . . .	
3	FXYD4	53828	.	GRCh37	10	43869118	43869118	+	5_prime_UTR_variant	5'UTR	SNP	G	T	novel	.	TCGA-2V-A955-01 TCGA-2V-A955-10 G G . . .	
5	ZSMW8	23053	.	GRCh37	10	75561253	75561253	+	synonymous_variant	Silent	SNP	C	T	r5749096821	.	TCGA-2V-A955-01 TCGA-2V-A955-10 C C . . .	
6	LRT12	348745	.	GRCh37	10	85982126	85982126	+	synonymous_variant	Silent	SNP	A	G	novel	.	TCGA-2V-A955-01 TCGA-2V-A955-10 A A . . .	
7	RP11-69N9.2	0	.	GRCh37	11	104776270	104776270	+	non_coding_transcript_exon_variant	RNA_SNP	T	T	G	novel	.	TCGA-2V-A955-01 TCGA-2V-A955-10 T T . . .	
8	SAA2	6289	.	GRCh37	11	18266987	18266987	+	synonymous_variant	Silent	SNP	T	T	C	novel	.	TCGA-2V-A955-01 TCGA-2V-A955-10 T T . . .
9	B4GALNT4	338707	.	GRCh37	11	377314	377314	+	missense_variant	Missense_Mutation	SNP	G	G	novel	.	TCGA-2V-A955-01 TCGA-2V-A955-10 G G . . .	
10	RPLP02	113157	.	GRCh37	11	61465258	61465258	+	non_coding_transcript_exon_variant	RNA_SNP	A	T	T	novel	.	TCGA-2V-A955-01 TCGA-2V-A955-10 A A . . .	
11	TRPT1	83707	.	GRCh37	11	63992060	63992060	+	missense_variant	Missense_Mutation	SNP	T	T	G	novel	.	TCGA-2V-A955-01 TCGA-2V-A955-10 T T . . .
12	TSGA10T	254187	.	GRCh37	11	65715032	65715032	+	non_coding_transcript_exon_variant	RNA_SNP	G	G	A	rs761245228	.	TCGA-2V-A955-01 TCGA-2V-A955-10 G G . . .	
13	C11orf30	0	.	GRCh37	11	76175034	76175034	+	synonymous_variant	Silent	SNP	C	T	r5320817581	.	TCGA-2V-A955-01 TCGA-2V-A955-10 C C . . .	
14	SBF2-AS1	283104	.	GRCh37	11	9830227	9830227	+	non_coding_transcript_exon_variant	RNA_SNP	G	G	A	novel	.	TCGA-2V-A955-01 TCGA-2V-A955-10 G G . . .	
15	TBX5	6918	.	GRCh37	12	114793370	114793370	+	synonymous_variant	Silent	SNP	A	A	novel	.	TCGA-2V-A955-01 TCGA-2V-A955-10 A A . . .	
16	KCNK2	3747	.	GRCh37	12	75661373	75661373	+	missense_variant	Missense_Mutation	SNP	C	C	T	novel	.	TCGA-2V-A955-01 TCGA-2V-A955-10 C C . . .
17	COL4A1	1282	.	GRCh37	13	110835346	110835346	+	missense_variant	Missense_Mutation	SNP	G	G	A	r53195505056	.	TCGA-2V-A955-01 TCGA-2V-A955-10 G G . . .
18	RNF17	56163	.	GRCh37	13	25433158	25433158	+	missense_variant	Missense_Mutation	SNP	G	G	T	.	TCGA-2V-A955-01 TCGA-2V-A955-10 G G . . .	
19	RTL1	388015	.	GRCh37	14	101349210	101349210	+	missense_variant	Missense_Mutation	SNP	A	A	C	novel	.	TCGA-2V-A955-01 TCGA-2V-A955-10 A A . . .
20	MIR656	724026	.	GRCh37	14	181533134	181533134	+	non_coding_transcript_exon_variant	RNA_SNP	G	G	A	r5371434052	.	TCGA-2V-A955-01 TCGA-2V-A955-10 G G . . .	
21	IGHM3-38	28429	.	GRCh37	14	106866578	106866578	+	missense_variant	Missense_Mutation	SNP	G	G	T	novel	.	TCGA-2V-A955-01 TCGA-2V-A955-10 G G . . .
22	PSMB5	5693	.	GRCh37	14	235040608	235040608	+	missense_variant	Missense_Mutation	SNP	T	T	C	novel	.	TCGA-2V-A955-01 TCGA-2V-A955-10 T T . . .
23	SLC22A17	51310	.	GRCh37	14	23815958	23815958	+	missense_variant	Missense_Mutation	SNP	G	G	A	novel	.	TCGA-2V-A955-01 TCGA-2V-A955-10 G G . . .
24	NOVA1	4857	.	GRCh37	14	26939542	26939542	+	intron_variant	Intron	SNP	A	A	C	novel	.	TCGA-2V-A955-01 TCGA-2V-A955-10 A A . . .
25	FOXP1	2298	.	GRCh37	14	29236983	29236983	+	synonymous_variant	Silent	SNP	G	G	A	rs764854659	.	TCGA-2V-A955-01 TCGA-2V-A955-10 G G . . .
26	ZBTB1	22890	.	GRCh37	14	64988788	64988786	+	inframe_deletion	In_Frame_Del	DEL	TATTTGATG	TATTTGATG	.	novel	.	TCGA-2V-A955-01 TCGA-2V-A955-10 G G . . .
27	EMLS	161436	.	GRCh37	14	89153622	89153622	+	missense_variant	Missense_Mutation	SNP	G	G	T	novel	.	TCGA-2V-A955-01 TCGA-2V-A955-10 G G . . .
28	MYO1E	4643	.	GRCh37	15	59466403	59466403	+	missense_variant	Missense_Mutation	SNP	C	C	G	novel	.	TCGA-2V-A955-01 TCGA-2V-A955-10 C C . . .
29	MAN2C1	4123	.	GRCh37	15	75660853	75660889	+	frameshift_variant	Frame_Shift_Del	DEL	GAGCCGGCACAGCAACTCTCCACCGCTCAGCGTC	GAGCCGGCACAGCAACTCTCCACCGCTCAGCGTC
30	MCTP2	55784	.	GRCh37	15	94841718	94841718	+	missense_variant	Missense_Mutation	SNP	A	A	G	novel	.	TCGA-2V-A955-01 TCGA-2V-A955-10 A A . . .
31	SALL1	6299	.	GRCh37	16	51175368	51175368	+	missense_variant	Missense_Mutation	SNP	C	C	A	novel	.	TCGA-2V-A955-01 TCGA-2V-A955-10 C C . . .
32	GLG1	2734	.	GRCh37	16	74524981	74524981	+	missense_variant	Missense_Mutation	SNP	T	T	A	novel	.	TCGA-2V-A955-01 TCGA-2V-A955-10 T T . . .
33	FAM22B	55731	.	GRCh37	17	27085738	27085738	+	missense_variant	Missense_Mutation	SNP	T	T	G	novel	.	TCGA-2V-A955-01 TCGA-2V-A955-10 T T . . .
34	OR1A1	8383	.	GRCh37	17	3118953	3118953	+	missense_variant	Missense_Mutation	SNP	C	G	C	rs763288963	.	TCGA-2V-A955-01 TCGA-2V-A955-10 C C . . .
35	OR1E1	8387	.	GRCh37	17	3308017	3308017	+	synonymous_variant	Silent	SNP	G	G	A	rs774514315	.	TCGA-2V-A955-01 TCGA-2V-A955-10 G G . . .
36	COL1A1	1277	.	GRCh37	17	48263287	48263287	+	missense_variant	Missense_Mutation	SNP	G	G	A	novel	.	TCGA-2V-A955-01 TCGA-2V-A955-10 G G . . .
37	CEP119	201134	.	GRCh37	17	63685338	63685338	+	splice_acceptor_variant	Splice_Site	SNP	T	T	A	novel	.	TCGA-2V-A955-01 TCGA-2V-A955-10 T T . . .
38	DSC1	1823	.	GRCh37	18	28720139	28720139	+	synonymous_variant	Silent	SNP	A	A	T	novel	.	TCGA-2V-A955-01 TCGA-2V-A955-10 A A . . .
39	TSPAN16	26526	.	GRCh37	19	11411928	11411928	+	stop_gained	Nonsense_Mutation	SNP	A	A	T	novel	.	TCGA-2V-A955-01 TCGA-2V-A955-10 A A . . .
40	TSPAN16	26526	.	GRCh37	19	11411929	11411929	+	missense_variant	Missense_Mutation	SNP	G	G	T	novel	.	TCGA-2V-A955-01 TCGA-2V-A955-10 G G . . .

data_mutations.txt

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.

Hugo_Symbol	Chromosome	Start_Position	End_Position	Reference_Allele	Tumor_Seq_Allele2	Variant_Classification	Var
SUV420H1	11	67942494	67942494	G,C	Missense_Mutation	SNP,+	ACAGACAGTGTCTCAATTCTGCAACTCCTGTATGAGGTG,BLCA
ELAVL3	19	11568911	11568911	G,A	Silent	SNP,+	GTTCGAGACCAGCCTGATCAACATGGCAAACGCTGTCTCC,BLCA
TARBP1	1	234541779	234541779	G,A	Nonsense_Mutation	SNP,+	CATAAAGTTAACCATGCAACGGGTGAGTCACAATTGGCTT,BLCA
LPIN1	2	11911570	11911570	A,T	Nonsense_Mutation	SNP,+	TTTCACTCAGGGCCCCTCACAGTGGTTAGTGGCTTGGC,LUSC
R3HDM2	12	57663757	57663757	G,A	Silent	SNP,+	TTATTCTATGTATAAGGCTGGCTGAAAATCCTCACAAAGA,LUSC
APOF	12	56755128	56755128	C,T	Missense_Mutation	SNP,+	CGCTATGTTGCCAGGCTGGCTCAAACCTGGACTCAGG,LUSC
TIMMDC1	3	119242475	119242475	G,T	Nonsense_Mutation	SNP,+	CCGGGAAGAAGCGCTCCTCACTTCCAGACTGGCGGCTGG,KIRC
KRT85	12	52760942	52760942	C,T	Missense_Mutation	SNP,+	CCAGCCTACACTTAGGGAAAATAGAAAAGAACCTACATTGA,KIRC
FREM2	13	39265811	39265811	C,T	Silent	SNP,+	TTAAATTAAAGGAAAAGTATGATTACACGAATAGATGC,KIRC

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

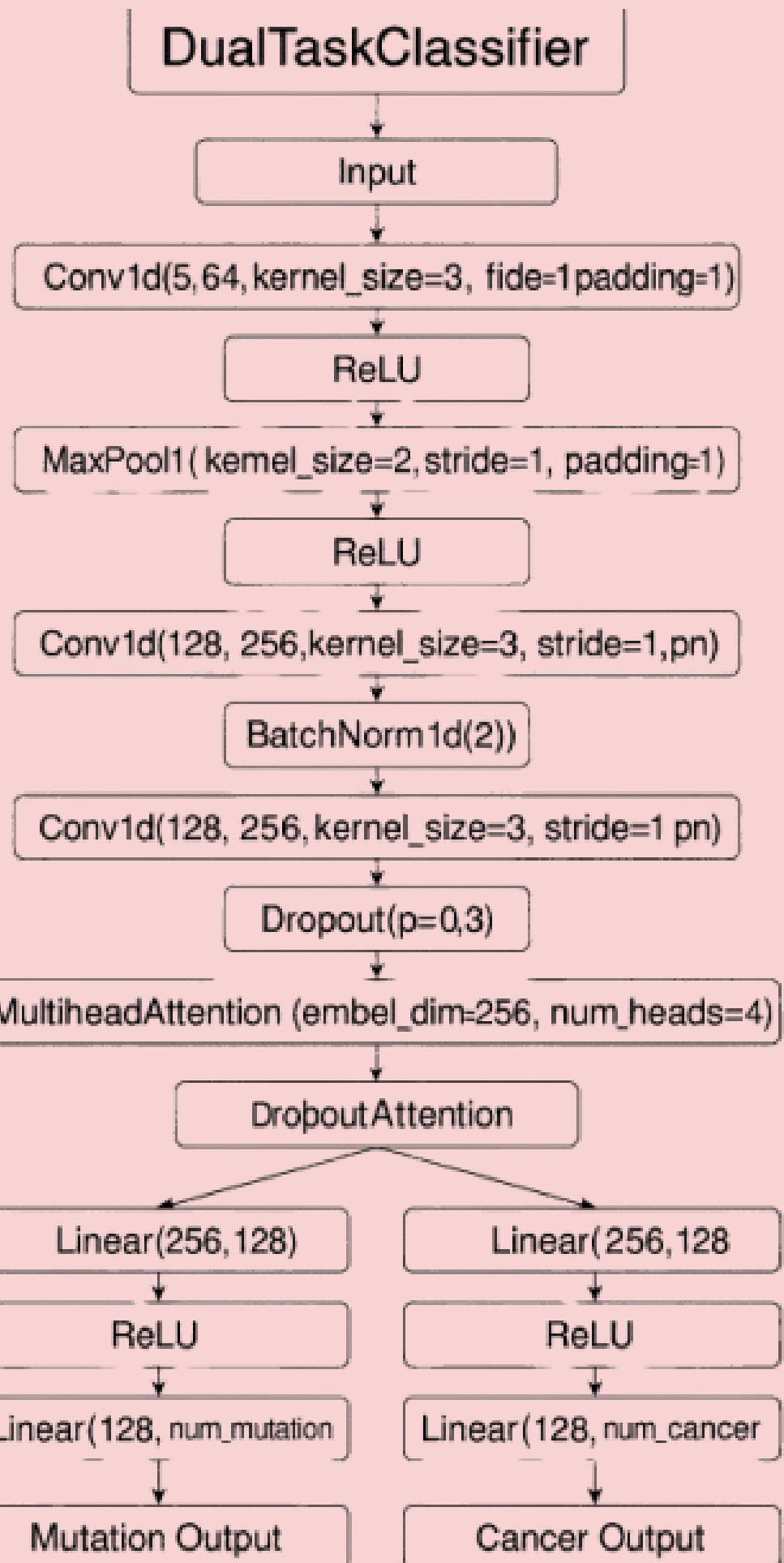
Missense Mutations - 22512
Silent - 20348
Nonsense Mutations - 18499

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

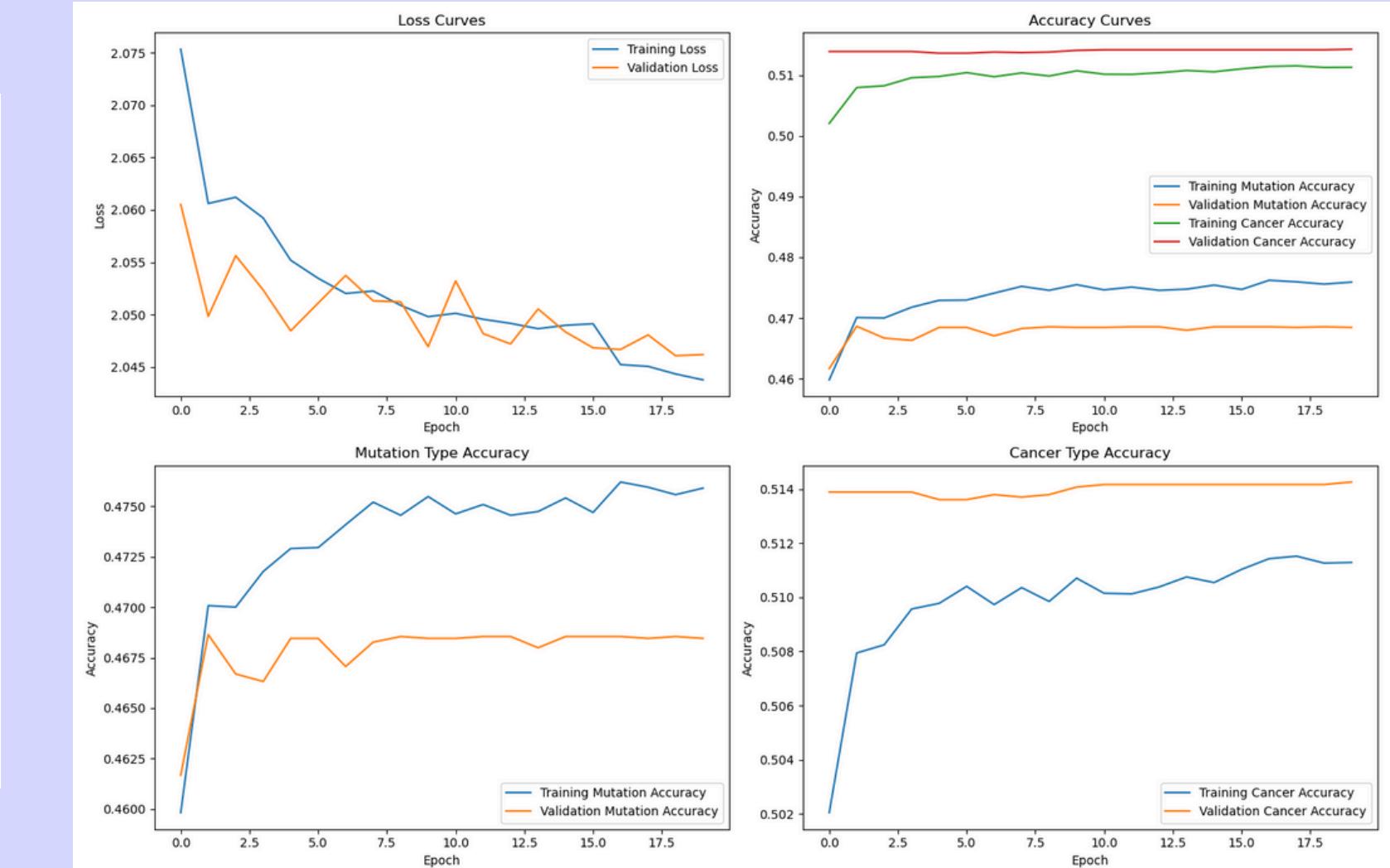
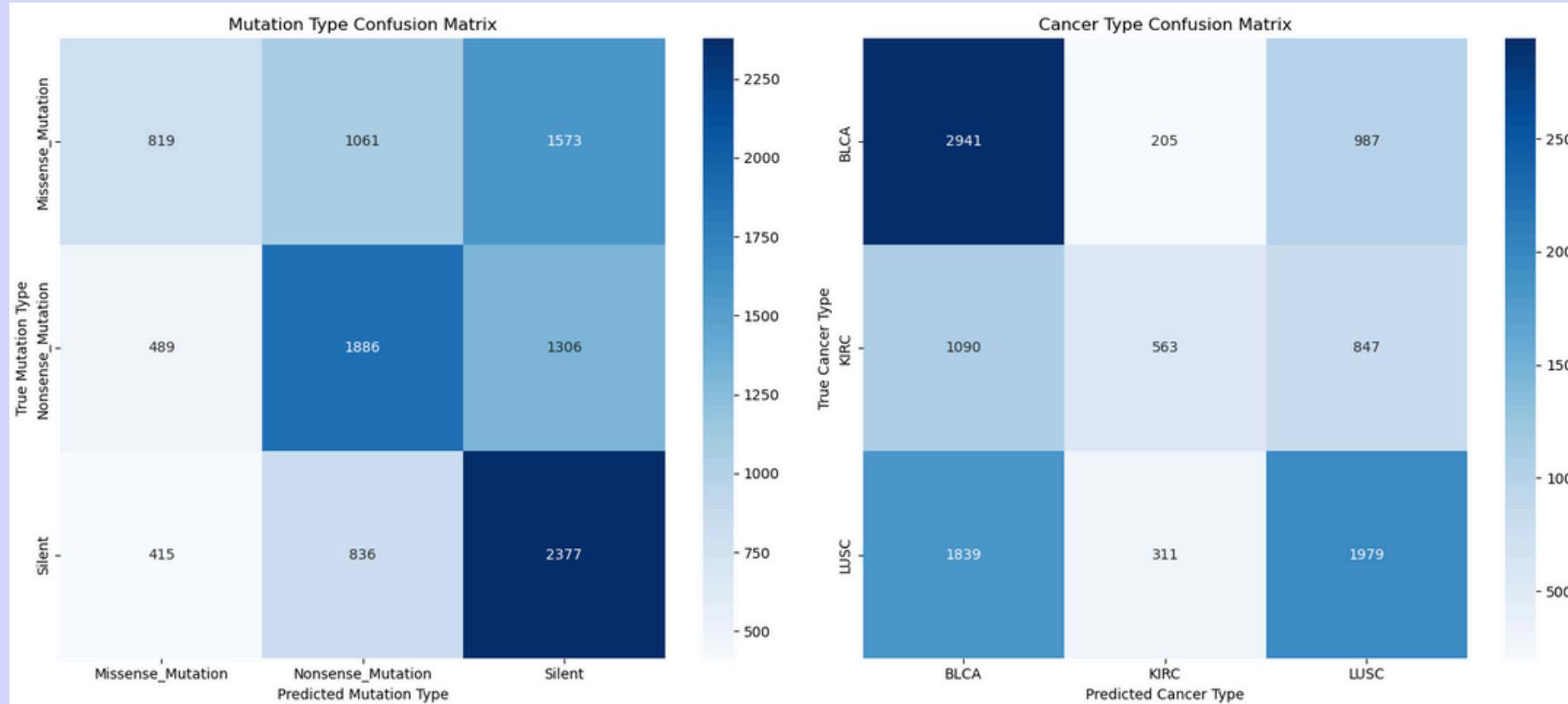


After cleaning and preprocessing the dataset from cBioPortal, our best model achieved

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

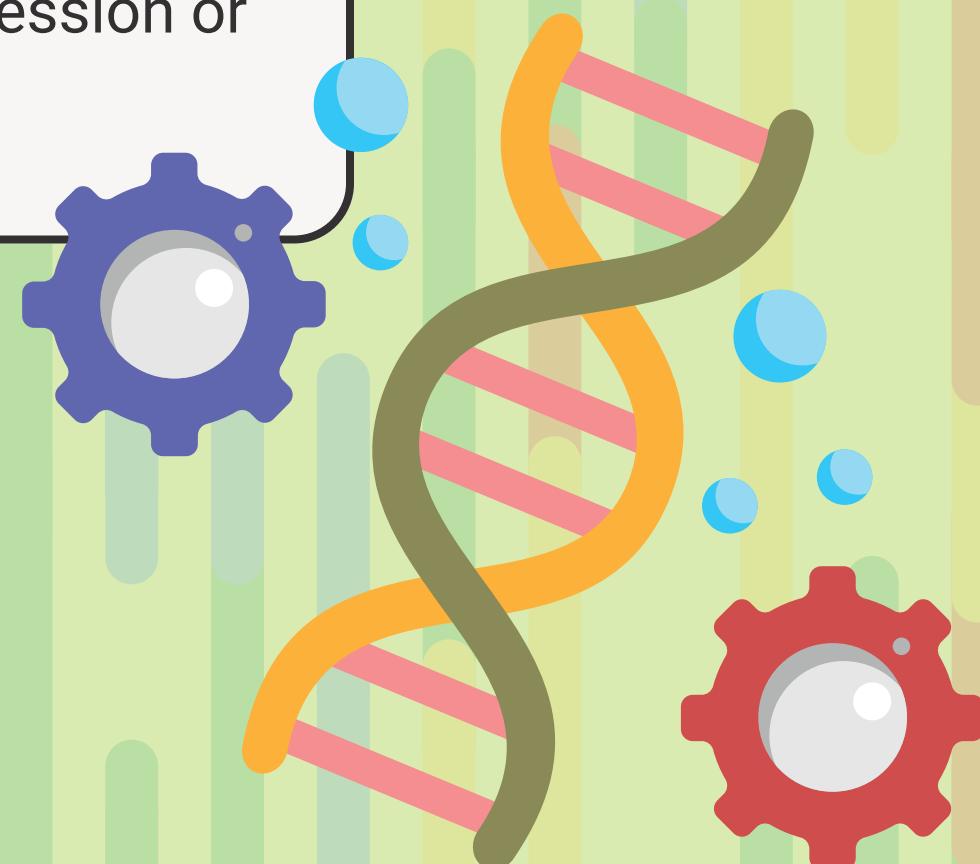


Merits

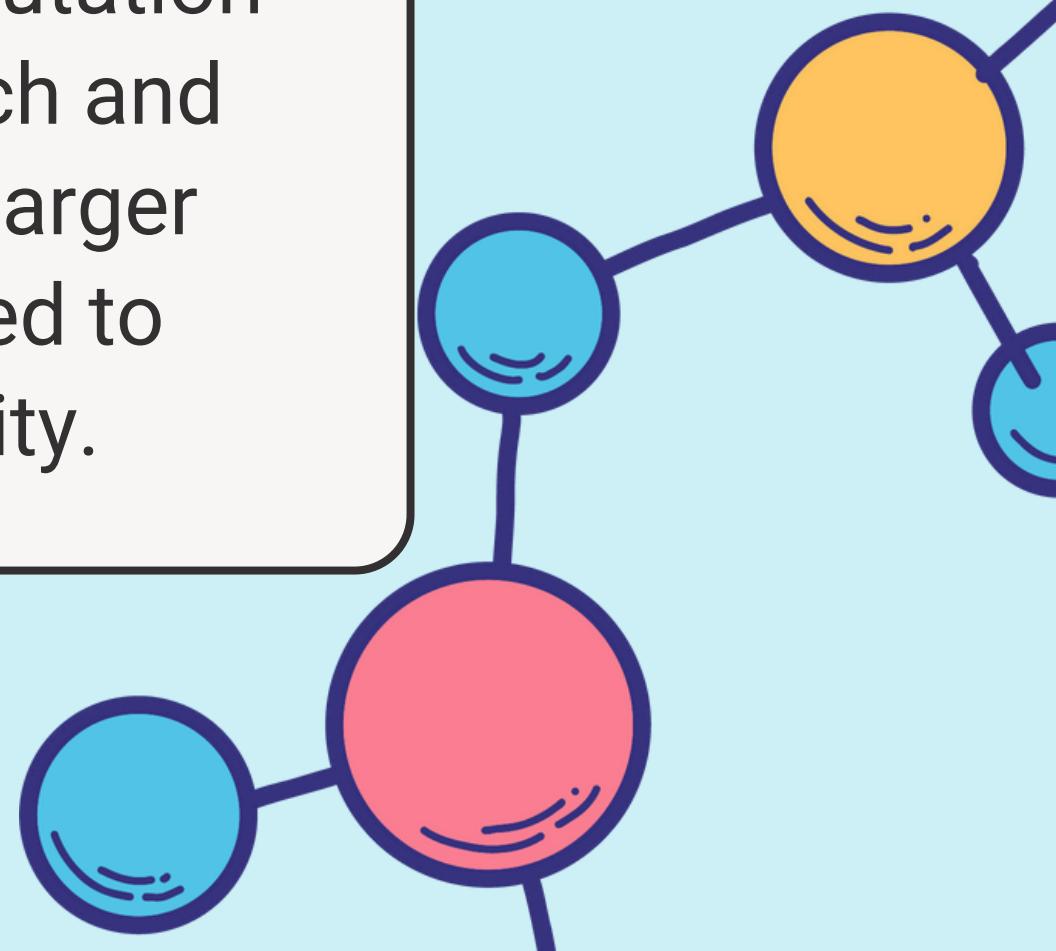
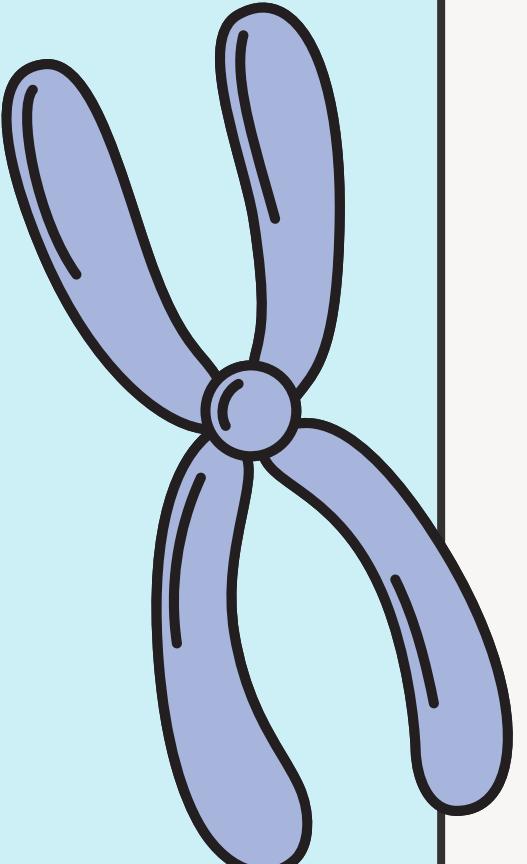
Our model provides valuable insight into mutation patterns across multiple cancer types—BLCA, LUSC, and KIRC—using real patient data from cBioPortal. It incorporates relevant biological features and demonstrates decent accuracy, as shown in training curves and confusion matrices. The use of genomic sequence 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

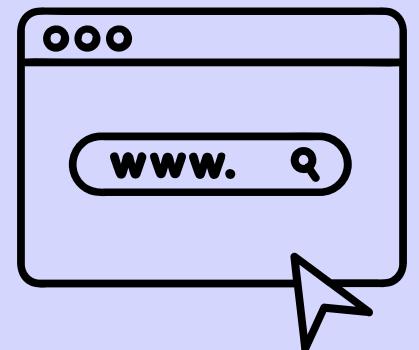


In this project, we successfully analyzed mutation data for three major cancer types: BLCA, LUSC, and KIRC. By extracting key genomic features and applying deep learning models, we achieved meaningful predictions about mutation classifications. Our results show that understanding mutation patterns can 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.ncbi.nlm.nih.gov/datasets>
- [7] <https://rest.ensembl.org/>

WEBSITE



[CarcitTrack: Genome Mutation & Cancer Classifier](#)
[Github](#) 

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

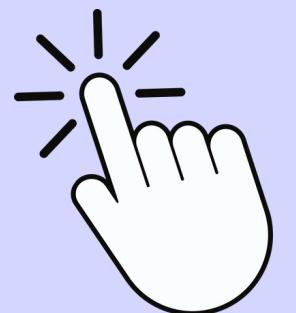
Enter a genomic sequence and alleles, then click 'Predict Mutation & Cancer Type' to see results.

See Example Input

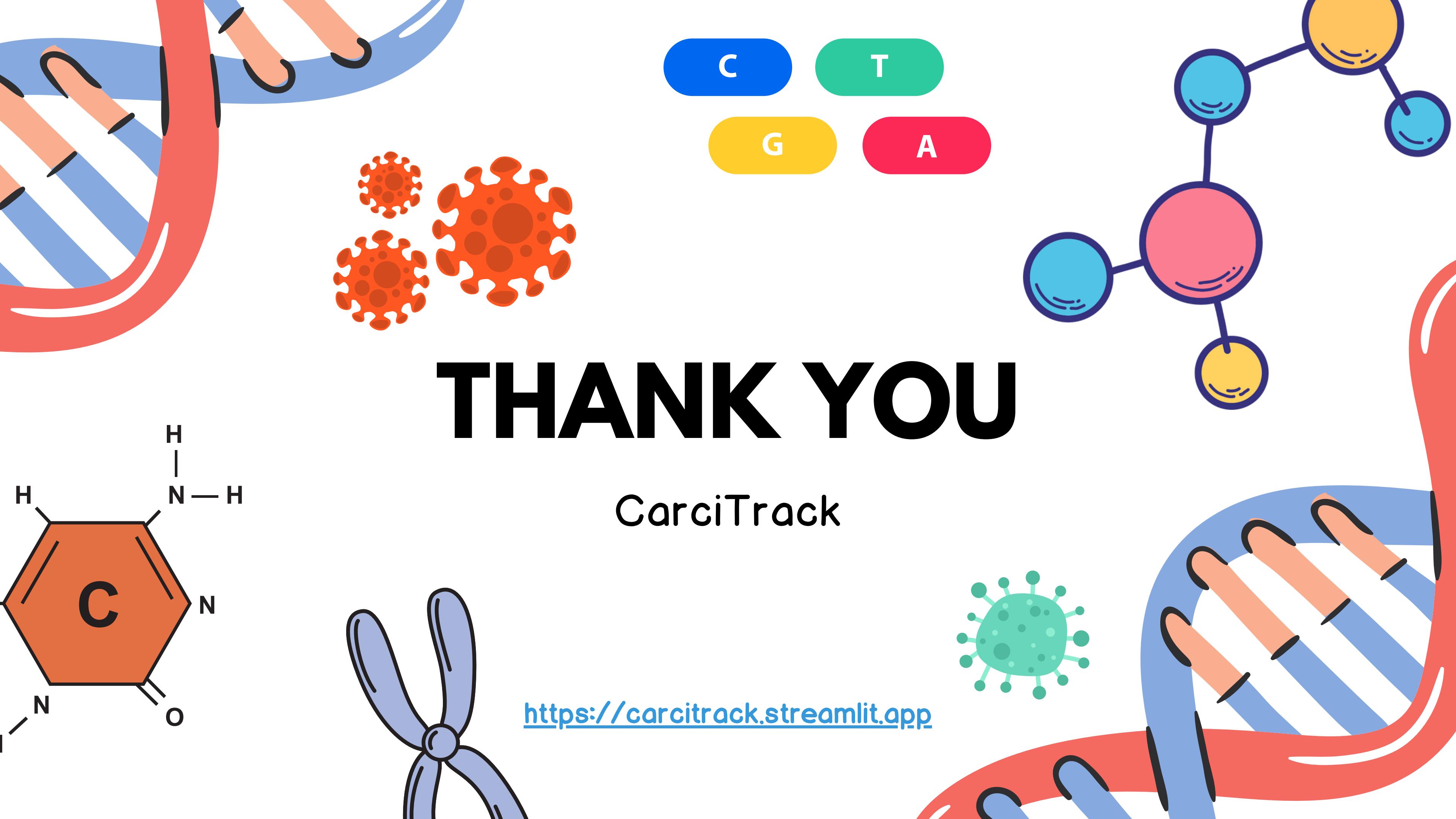
About Mutation & Cancer Types

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

< Manage app



Click
Here



THANK YOU

CarciTack

<https://carcittrack.streamlit.app>