CS4.409. Data Foundation Systems

AI MODEL DEPLOYMENT (GENOME SEQUENCING)

Smruti Biswal - 2020112011
Aakash Reddy Gorla - 2020102034

INTRODUCTION

The general use of AI is restricted to or at least promoted to tasks in which human intelligence is required, but in the case of genome sequencing the main purpose for the deployment of AI is to reduce the number of errors as it is a task extremely prone to human error considering the type of data being processed. Facial image recognition can help molecular diagnosis and help identify unusual genetic illnesses. In this project we aim to provide a platform with services: AI models which is user friendly and scalable. The user will be able to mix-n-match different preprocessing stages in the pipeline as well as the AI model(restricted to a select few).

REQUIREMENTS

FUNCTIONAL

01

AI - MODEL

A deep learning model based on multinomial base classifier. Train on a dataset of Genome Sequences and be optimized for high accuracy

02

WEB APP

A simple online application that lets users submit their genomic data as a file, chose a model and then obtain results based on the model. The proper preprocessing steps will be available in the same manner as above for the data. Time for processing is around 10 seconds.

03

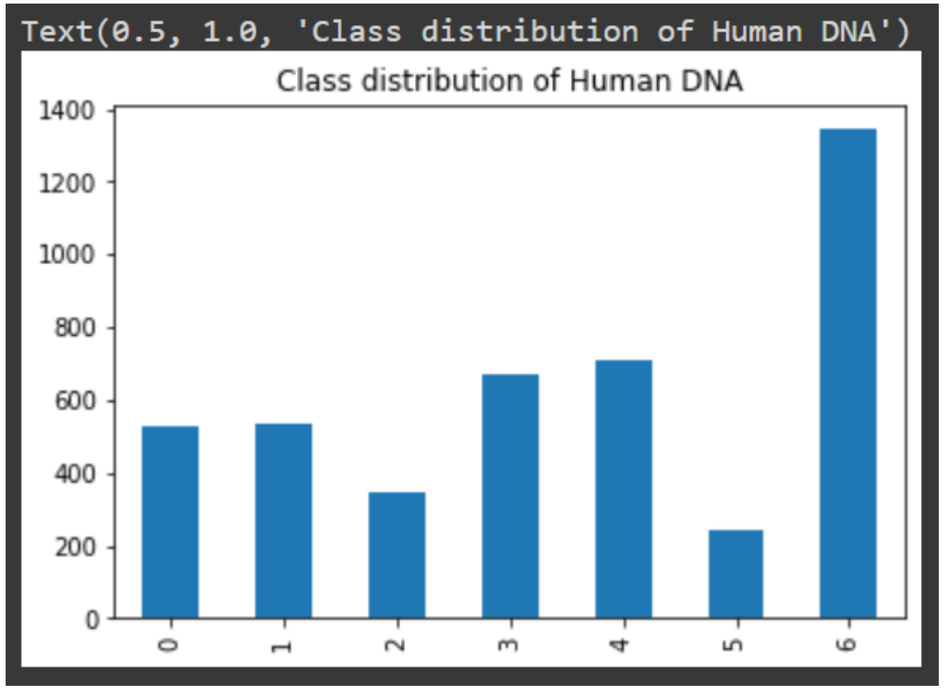
DATA

Datasets were not provided. Kaggle datasets were taken to train the model.

ML MODEL FOR GENOME SEQUENCING

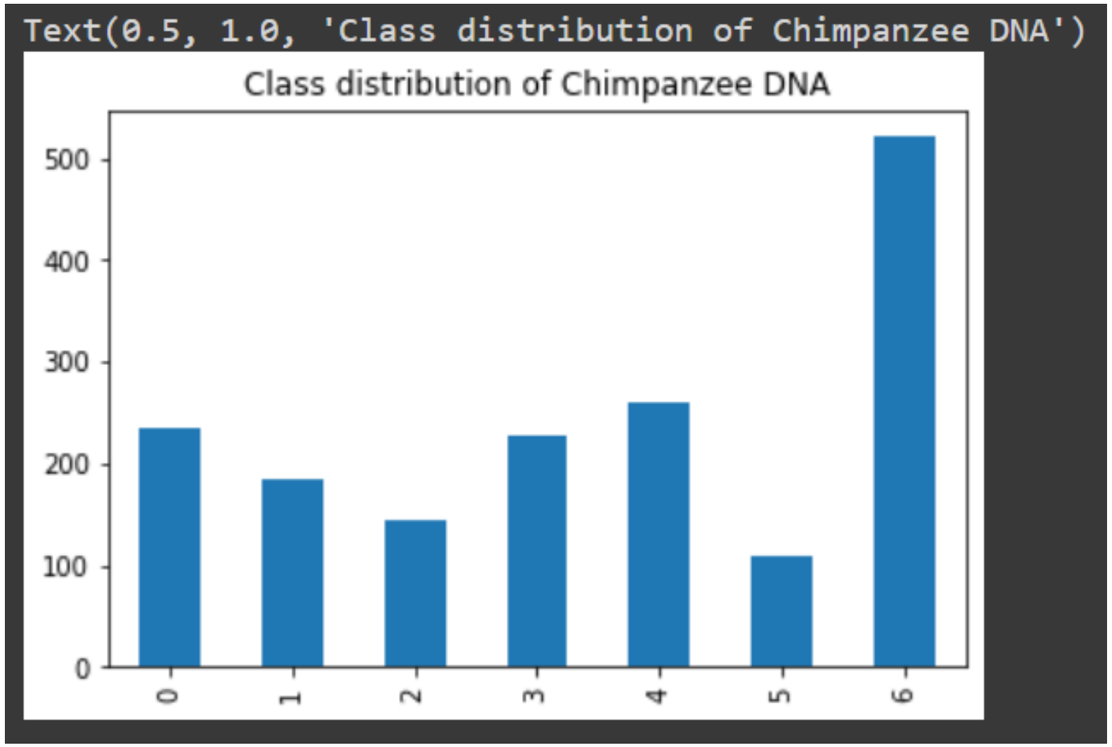
Build a classification model that is trained on the human DNA sequence and can predict a gene family based on the DNA sequence of the coding sequence. To test the model, we will use the DNA sequence of humans, dogs, and chimpanzees and compare the accuracies.

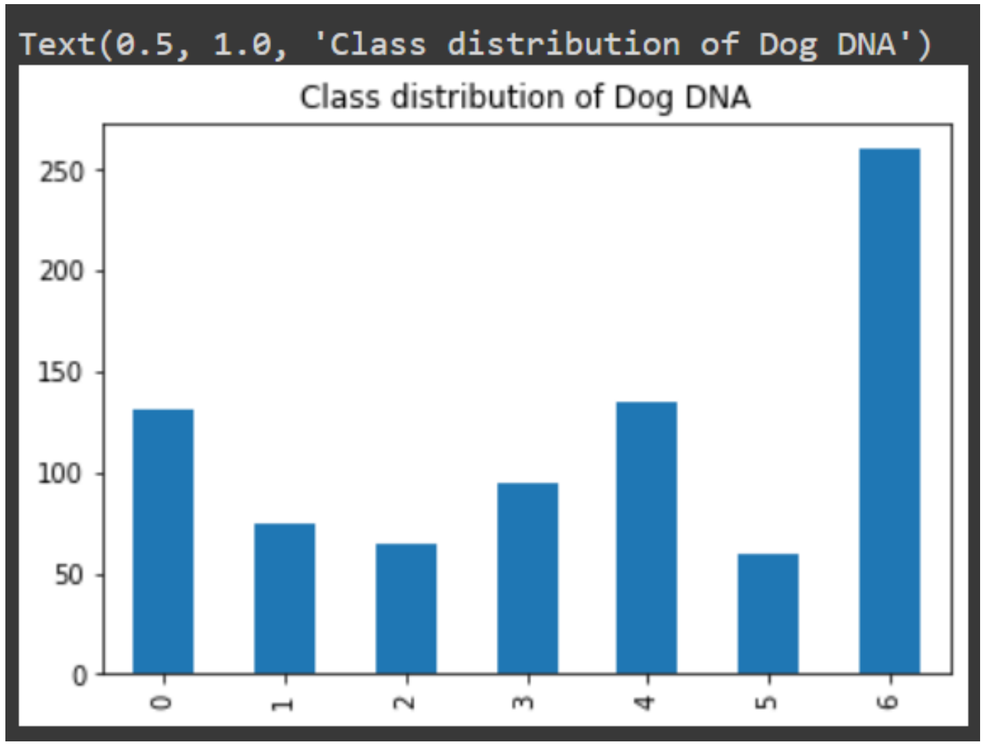
	sequence	class
0	ATGCCCCAACTAAATACTACCGTATGGCCCACCATAATTACCCCCA...	4
1	ATGAACGAAAATCTGTTCGCTTCATTCATTGCCCCCACAATCCTAG...	4
2	ATGTGTGGCATTTGGGCGCTGTTTGGCAGTGATGATTGCCTTTCTG...	3
3	ATGTGTGGCATTTGGGCGCTGTTTGGCAGTGATGATTGCCTTTCTG...	3
4	ATGCAACAGCATTTTGAATTTGAATACCAGACCAAAGTGGATGGTG...	3



ML MODEL FOR GENOME SEQUENCING

Gene families are groups of related genes that share a common ancestor. Gene paralogs are genes with similar sequences from within the same species while gene orthologs are genes with similar sequences in different species. The dataset contains human DNA sequence, Dog DNA sequence, and Chimpanzee DNA sequence.





DATASET

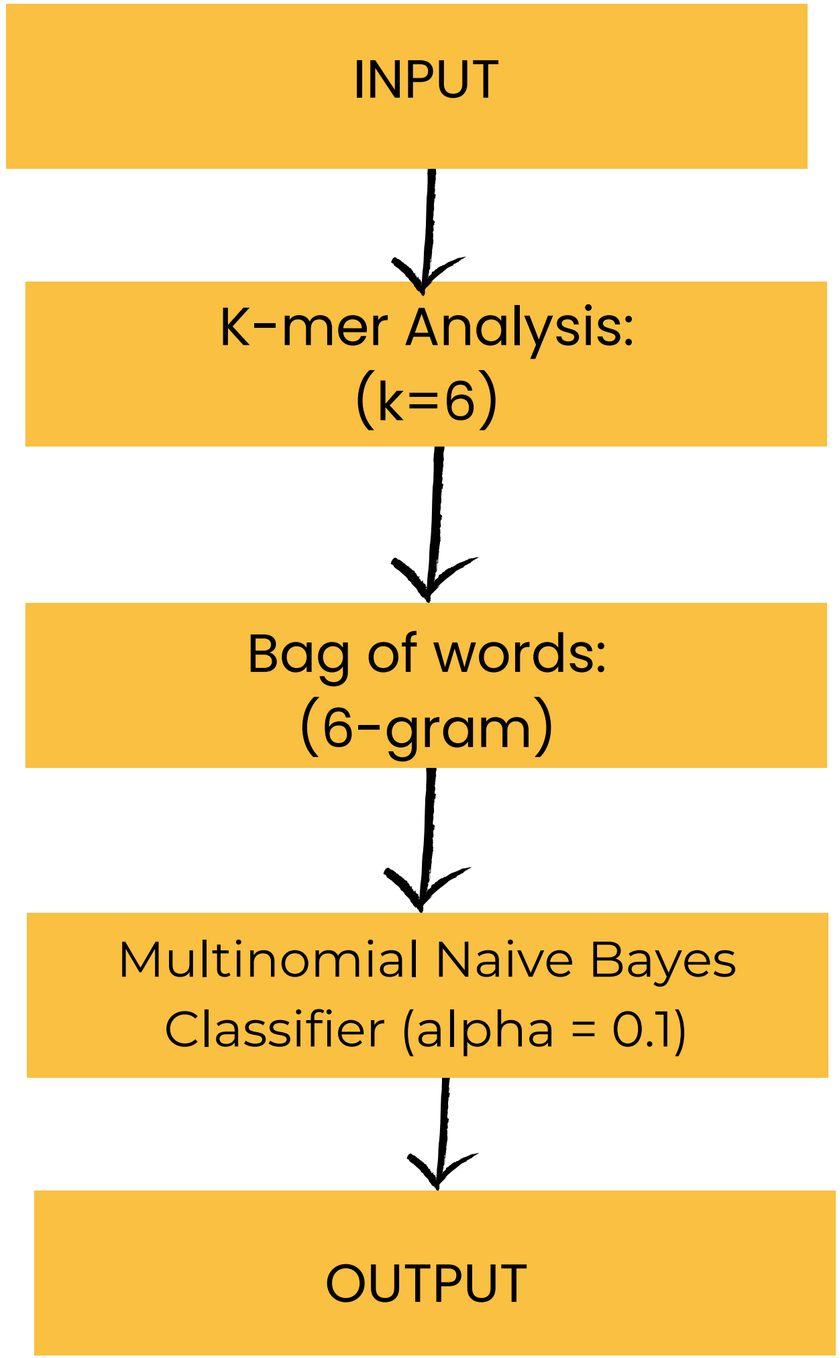
The Datasets used are obtained from kaggle. We use chimpanzee, dog and human datasets.

As of now the model requires text file with the genomic sequence in it.

Input formats: .txt

Other forms such as .csc and .xlsx can also be used with minor changes.

MODEL ARCHITECTURE

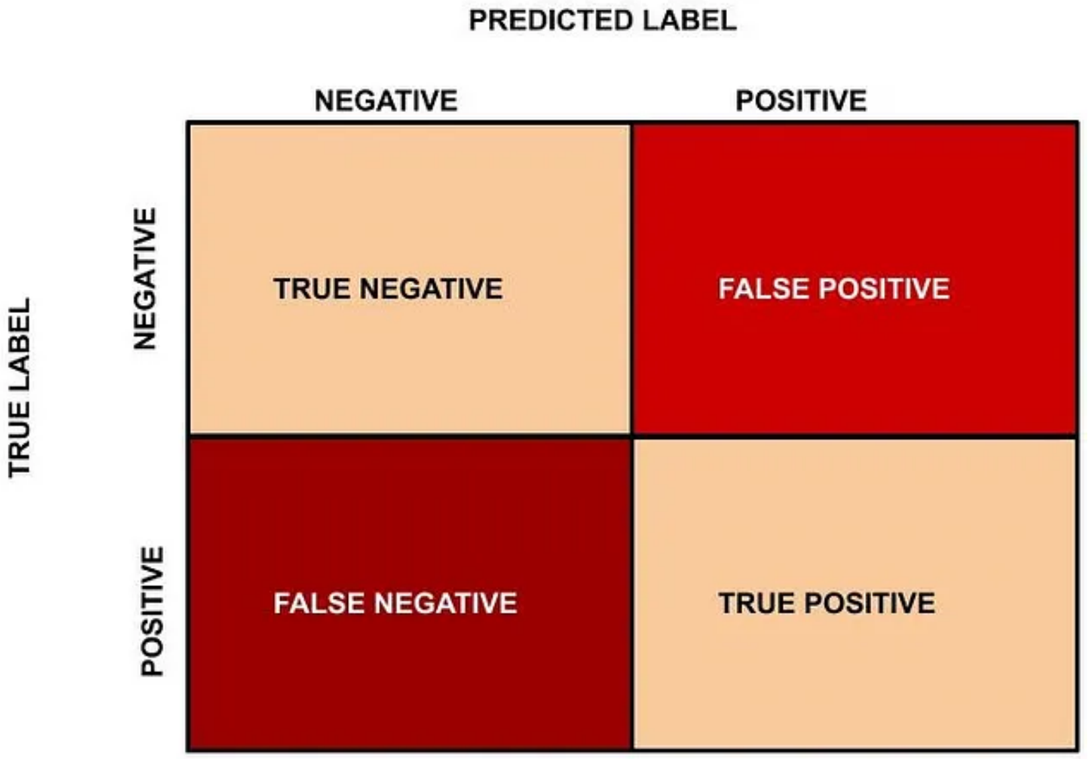


OUTPUT FORMAT

```
accuracy = 0.993
precision = 0.994
recall = 0.993
f1 = 0.993
```

```
accuracy = 0.926
precision = 0.934
recall = 0.926
f1 = 0.925
```

OUTPUT FORMAT



$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

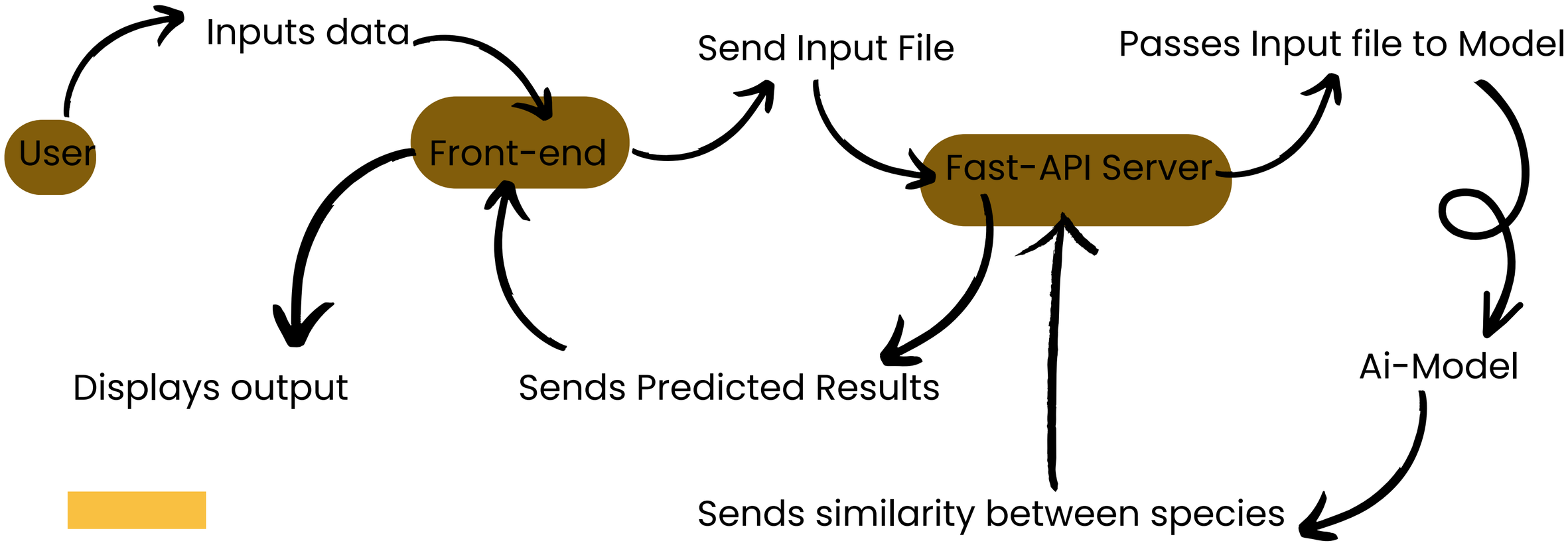
TECH STACK

We developed a user-friendly interface for users to input files and then display the similarities using REACT JS as the frontend.

As a backend for our model, we used Fast API. The user input is handled by Fast API, which then passes it to the model for prediction before returning the output to the front end.

FRONT-END	REACT JS
BACK-END	FAST API

CODE FLOW


User
Inputs data
Front-end
Send Input File
Fast-API Server
Passes Input file to Model
Ai-Model
Displays output
Sends Predicted Results
Sends similarity between species

DEMO

THANKS