SOFTWARE REQUIREMENTS SPECIFICATION

AI Model Training & Deployment Genome Sequencing

CS4.409 Data Foundation Systems

Jewel Benny (2020102057) & Srujana Vanka (2020102005)

Contents

1 Overview

1.1 Description

The main aim of the project is to develop, train and deploy an AI/ ML model that deals with genomic sequences. Since this is a vastly open-ended problem, there are multiple problems which can be taken. However, the lack of high quality publically available genomic sequence datasets poses a large issue.

Machine learning models have become indispensable in the context of the COVID-19 pandemic, serving a multitude of vital purposes. These models enable early detection and prediction of the virus's spread, optimizing resource allocation in healthcare, aiding in diagnostics and risk assessment, expediting drug discovery and vaccine development, facilitating contact tracing efforts, and providing genomic insights into the virus. Moreover, they help evaluate the effectiveness of public health interventions, analyze scientific literature, and model epidemic trajectories.

Gauging the origin of a COVID-19 sample is important for epidemiological tracking, monitoring variants, investigating outbreaks, shaping public health policies, aiding research and vaccine development, facilitating global surveillance, and managing supply chains. This information helps control the spread of the virus, tailor interventions, and inform global responses. However, it must be handled responsibly, respecting privacy and ethical considerations.

RCoV19 is a comprehensive collection of data related to the COVID-19 pandemic caused by the SARS-CoV-2 virus. This dataset was created to facilitate research and analysis of the virus, its transmission, and its impact on public health. It includes a wide range of information, such as epidemiological data (e.g., cases, deaths, recoveries), clinical data, genomic data, and more. Using the genomic data, we believe it is possible to develop an ML model that can accurately predict which location a particular sample is from, by training using genomic data and associated metadata.

1.2 Understanding of the Project

Our project's core objective is to leverage the genomic data from the RCoV19 dataset, which encompasses a wide range of COVID-19-related information, to develop a sophisticated ML model. This model will be capable of accurately determining the geographic origin of COVID-19 samples. The significance lies in its potential applications, including epidemiological tracking, monitoring virus variants, investigating outbreaks, shaping public health policies, aiding research and vaccine development, facilitating global surveillance, and managing supply chains. By harnessing the power of genomics and associated metadata, the project aims to provide valuable insights for controlling the virus's spread, customizing interventions, and informing global responses, all while maintaining a strong commitment to privacy and ethics.

The system's operation begins with the collection of COVID-19 genomic sequences and associated metadata, such as geographic location information. These inputs undergo thorough data preprocessing to ensure quality and completeness. A subset of the dataset is selected for training an AI/ML model that establishes correlations between genomic patterns and locations.

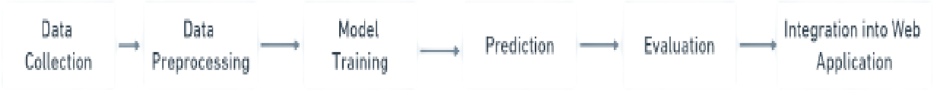

Figure 1.1: High-level Flowchart

The model processes this input and delivers precise predictions of the sample's geographic origin. Outputs include the predicted location and model evaluation results. This comprehensive approach not only facilitates accurate predictions but also enhances accessibility and engagement by incorporating a web application for user interaction and visualization of results.

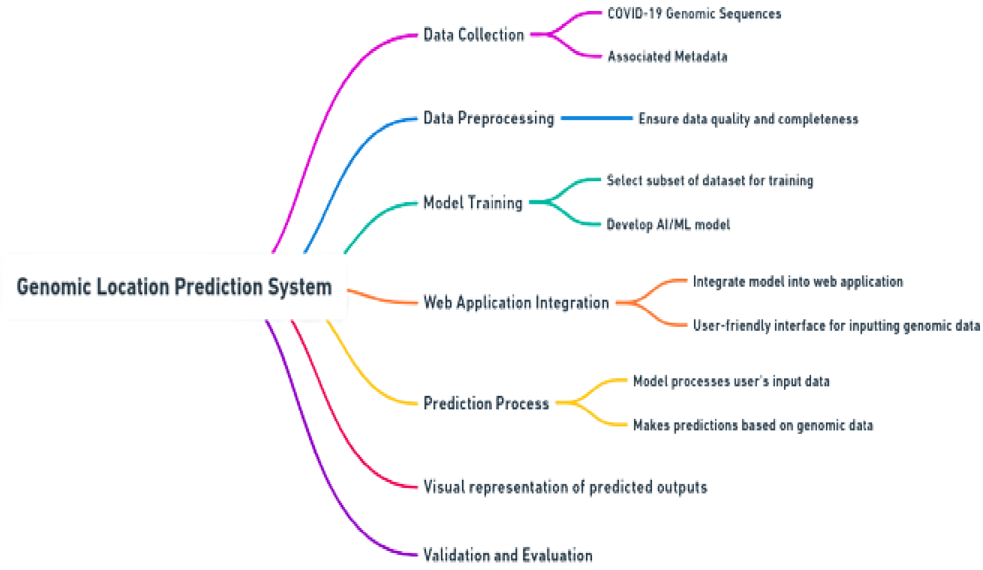
Genomic Location Prediction System
Data Collection
COVID-19 Genomic Sequences
Associated Metadata
Data Preprocessing
Ensure data quality and completeness
Model Training
Select subset of dataset for training
Develop AI/ML model
Web Application Integration
Integrate model into web application
User-friendly interface for inputting genomic data
Prediction Process
Model processes user's input data
Makes predictions based on genomic data
Visual representation of predicted outputs
Validation and Evaluation

Figure 1.2: Mindmap

2 Proof of Concept

2.1 Idea and Problem Statement

The idea behind this PoC is to harness the power of genomic data from the RCoV19 dataset to develop a machine learning model capable of accurately determining the geographic origin of COVID-19 samples. The problem we aim to address is the need for precise and efficient methods to trace the spread of the virus at a regional level. This capability has significant implications for epidemiological tracking, variant monitoring, outbreak investigations etc.

2.2 Objectives

The main objectives of this PoC are listed below,

- Develop a data preprocessing pipeline to clean and prepare genomic sequences and associated metadata.
- Select and preprocess a representative subset of the RCoV19 dataset for model training.
- Train and fine-tune a machine learning model to correlate genomic patterns with geographic locations.
- Evaluate the model's performance using appropriate metrics, including accuracy, precision, recall, and F1-score.
- Create a user-friendly interface for users to input genomic data and obtain predictions of geographic origin.
- Deploy the machine learning model as a web service or API, ensuring scalability and accessibility to users.
- Provide thorough documentation and user guides for the deployed system.

2.3 Resources Needed

The main resources needed for this PoC are listed below,

- Data: Access to the RCoV19 dataset or relevant genomic data with associated geographic metadata.

- Computing Resources: Sufficient computational power to handle data preprocessing and model training. Based on our estimated, the model is small enough to be trained on a local computer.
- Software and Tools: Machine learning frameworks (e.g., TensorFlow, PyTorch), data preprocessing tools (e.g., Python libraries), and web development tools (e.g., HTML/CSS, JavaScript) for creating the user interface, other tools for the database, serverside logic, API and backend.

2.4 Scope of the PoC

This project's scope involves creating a machine learning model using genomic data from the RCoV19 dataset to pinpoint where COVID-19 samples originate. It includes data prep, model training, and building a user-friendly interface. We'll also deploy the model as a secure web service with a database. The goal is to enhance epidemiological tracking, variant monitoring, outbreak investigation etc.

3 System Requirements

3.1 Functional Requirements

3.1.1 Dataset

Genomic data and associated metadata will be sourced from RCoV19, primarily those sampled from India. This dataset will be used for training, testing and validation of our ML model. Preprocessing and filtering of the data is also crucial to ensure better results.

3.1.2 AI/ ML Model

We plan to develop an ML model using the current trends (NLP etc.) with the bulk of the code written in Python. The model will be able to take genomic string data as input and predict the location which that sample likely originated from.

3.1.3 User Interface

We plan to develop an elegant UI that features intuitive navigation, clear instructions, and an aesthetically pleasing layout. This will primarily be done using HTML, CSS and JavaScript and Frameworks such as ReactJS, AngularJS, Bootstrap etc. It will support operations such as file upload and a dashboard that displays the results.

3.1.4 API

The API will be primarily be written using frameworks such as Flask and NodeJS and will be developed keeping in mind RESTful practices.

3.1.5 Backend

The backend will take care of all the preprocessing of the data input by the user. This involved conversion of the input file to the required format, cleaning etc. It will also host the ML model which will predict the geographic location of the sample. Other useful features will also be written for the backend.

3.1.6 Database

We plan to use a MySQL or PostgreSQL database that will be used to store data that is uploaded by the user. The database can be used to provide additional data, which will be helpful in improving the ML model in the future.

3.2 Non Functional Requirements

3.2.1 Usability Requirements

Usability is a core non-functional requirement driving our system's design. We prioritize delivering a user-friendly, intuitive, and responsive interface. Our user interface will feature intuitive navigation, clear instructions, and an aesthetically pleasing layout. We'll ensure compatibility across web browsers, devices, and screen sizes, providing a seamless user experience on various platforms.

3.2.2 Performance Requirements

Another non-functional requirement is to optimize system performance. We aim to minimize response times even during peak traffic periods and when handling a substantial volume of queries. This commitment to performance ensures that users receive swift and efficient service, enhancing their overall experience.

3.2.3 Scalability Requirements

As our system grows and encounters increased traffic and database loads, it must be able to adapt and expand effortlessly. Therefore, we will design the system architecture with scalability in mind, allowing for both horizontal and vertical scaling as needed. This ensures that the system can accommodate future growth without sacrificing performance or user experience.

3.2.4 Security Requirements

Security is paramount in safeguarding sensitive data. Our system will implement robust security measures, including user authentication and access controls, to protect against unauthorized access and data breaches, ensuring user data's confidentiality and integrity.

4 Project Deliverables

4.1 ML Model - Highest Priority

The development and training of the ML model described in the previous sections will be of utmost priority.

4.2 Data Collection - High Priority

Collection of high quality genomic data will be of high priority since the model will need a large amount of data to produce accurate results.

4.3 Web App and User Interface - High Priority

The frontend and UI will hold high priority. It will be a responsive design that can be opened by any browser, be it desktop or mobile phone.

4.4 API, Backend and Database - High Priority

The development of API, backend and database will have high priority. If time permits, all of these will be optimised further to decrease fetching times, execution times etc., and to reduce storage overhead.

4.5 Deployment - High Priority

Deployment of the project on computers other than the local host will also hold high priority.

4.6 Clean, Readable Code - High Priority

All of the code written for the project will be properly commented, formatted and well documented. This will ensure that it easy for a third person to understand the code for purposes such as improvement etc.