Genomic Sequence Analysis Using Machine Learning

An Engineering Project in Community Service

Phase – II Report

Submitted by

1. **19BCE10054 Shreyash Mall**
2. **19BCE10177 Ananya Singh**
3. **19BCE10061 Mudit Sorikh**
4. **19BCE10336 K Aniket Prusty**
5. **19BOE10081 Chetna Bisen**
6. **19BAI10112 Abhishek J Nair**
7. **19MIM10048 Pranay Pratap Singh**
8. **19MIM10113 Varnika Singh**

in partial fulfillment of the requirements for the degree of

Bachlore of Engineering and Technology

VIT Bhopal University
Bhopal
Madhya Pradesh

April, 2022

VIT®
BHOPAL
www.vitbhopal.ac.in

Bonafide Certificate

Certified that this project report titled **“Genomic Sequence Analysis Using Machine Learning”** is the bonafide work of “**19BCE10054 Shreyash Mall, 19BCE10177 Ananya Singh, 19BCE10061 Mudit Sorikh, 19BCE10336 K Aniket Prusty, 19BOE10081 Chetna Bisen, 19BAI10112 Abhishek J Nair, 19MIM10048 Pranay Pratap Singh, 19MIM10113 Varnika Singh”** who carried out the project work under my supervision.

This project report (Phase II) is submitted for the Project Viva-Voce examination held on 01/03/22

Supervisor

Comments & Signature (Reviewer 1)

Comments & Signature (Reviewer 2)

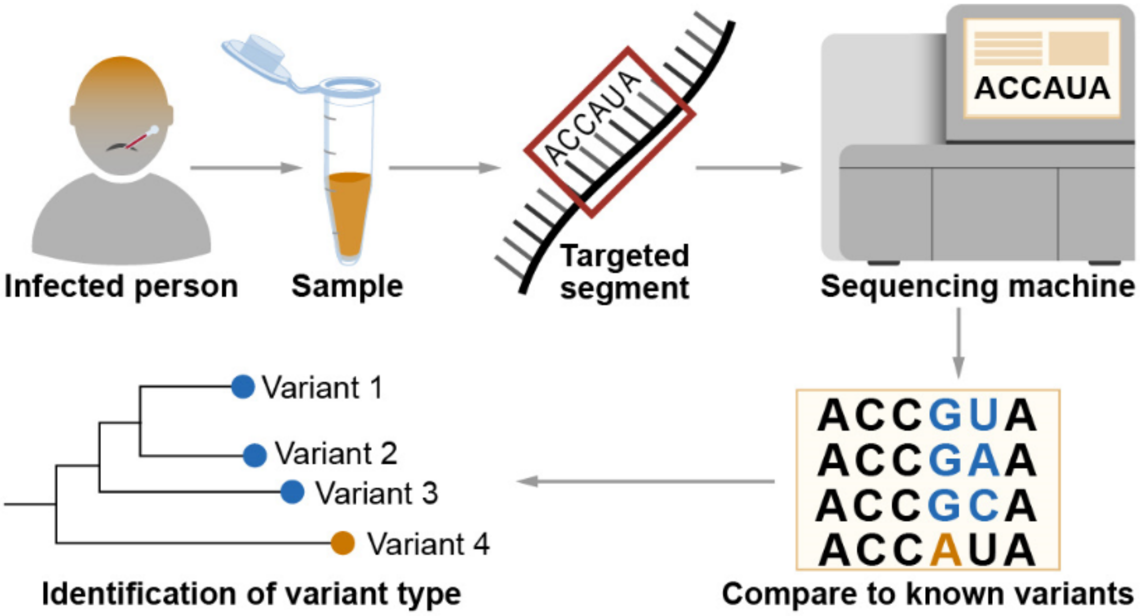
INTRODUCTION

Bioinformatics is an interdisciplinary field that merges biological science and genetics with applied science, mathematics, and statistics. to unravel data-intensive, large-scale biological concerns, computational techniques are applied. the foremost common problems are molecular modeling of biological processes and drawing inferences from data. Typically, a bioinformatics solution comprises the subsequent steps: Statistical information could also be derived from biological data. Construct a computational model. Resolve a controversy through computational modeling. Put a computer algorithm through its paces and see how it fared.

Machine learning has grown in popularity in bioinformatics and computational biology. Recently, its popularity has grown, and it's now employed in an exceedingly form of sectors like genetic sequencing, categorization, and lots of others.

A genome could be a full collection of genetic instructions for an organism. Each genome includes all of the knowledge required to construct the organism and permit it to grow and develop.

A laboratory technique for determining the entire genetic make-up of an organism or cell type. This approach is also accustomed detect changes in specific regions of the genome. These modifications may aid scientists in understanding how certain illnesses, like cancer, develop. Genomic sequencing results may potentially be utilized to diagnose and treat illness.



Machine learning (ML) is that the study of the autonomous learning of computers that don't seem to be explicitly programmed. it's employed in bioinformatics and focuses on producing data-driven predictions. this system enables algorithms to form complex predictions on massive datasets.

1.1 Motivation

Prior to the introduction of machine learning, bioinformatics algorithms had to be constructed by hand, which proved difficult for problems like protein structure prediction. Deep learning and other machine learning algorithms may be able to identify elements of data sets without requiring the programmer to define them individually.

Artificial Intelligence (ML)

ML is a subfield of AI that uses statistical learning techniques to develop intelligent systems. ML systems may learn and evolve on their own without being explicitly written. ML is used in music and movie streaming businesses' recommendation systems. Machine learning algorithms are classified into three types: supervised, unsupervised, and reinforcement learning.

Intensive Learning (DL)

This AI subset relies on an approach influenced by how the human brain processes information. it's associated with learning by example. DL systems help a computer model forecast and categorize data by filtering incoming data across layers. Deep Learning handles data within the same manner because the human brain. it's employed in technologies like self-driving cars. Deep learning network designs are classified into three types: convolutional neural networks, recurrent neural networks, and recursive neural networks.

The algorithm may then learn to mix low-level characteristics to provide more abstract features, and so on. This multi-layered approach, when correctly taught, enables such computers to form sophisticated predictions. These methods vary from earlier computational biology approaches in this, while they create use of existing datasets, they are doing not allow the info to be analyzed and appraised in novel ways. the amount and type of biological datasets accessible has grown substantially in recent years.

1.2 Objective

Gene sequencing using machine learning that can help identify various genetic defects and mutations in an organism.

Existing Work / Literature Review

PlasClass (PLOS ONE, 2020) uses logistic regression of kmer frequency vectors to determine whether they are derived from plasmid sequences or chromosomal regions. This is a binary classification tool. PlasFlow (released in 2018, Nucleic Acid Research)

This tool determines strain classification and whether a particular contig is a plasmid or chromosome. Neural networks are used in addition to the kmer frequency vector. MetaBCCLR (Bioinformatics, 2020) uses tdistributed stochastic neighbor embedding (tSNE) to size long readings of the genome and perform trimer-vector binning of metagenomic data. Whole Genome Genotyping Human Leukocyte Antigen (HLA) Typing with HTNGS: A three-step procedure for HLA typing has been established. In the first step, HLAA, B, C, DRB1 and DQB1 were amplified using long-range PCR. In the second step, we used the 454GSFLX platform to sequence the amplicon. In the third step, we used AssignNG software to analyze the sequence data. Lind etc. 2010

HTNGS in Prenatal Testing: A comprehensive review of the effects of HTNGS on prenatal testing. Raymond etc. 2010

Intrauterine Disease Screening: New Studies Show the Feasibility of Genome-Wide Fetal Genotyping Using Non-Invasive Next Generation Maternal Blood Sequencing Burgess 2011

DeNovo assembly and human genome reassembly. Genome rearrangement via pool of DNA: This study proposed a new statistical approach, CRISP (Comprehensive Read Analysis for Identification of SNPs from Pooled Sequencing).). Banal 2010

Genome Resequencing and HTNGS Platform: This study evaluated a performance comparison between the Illumina Genome Analyzer and Roche 454GSFLX for resequencing 16 genes associated with hypertrophic cardiomyopathy (HCM). This study reveals the possibility of combining LRPCR with the NGS platform for targeted resequencing of HC-related genes. Dames et al. 2010

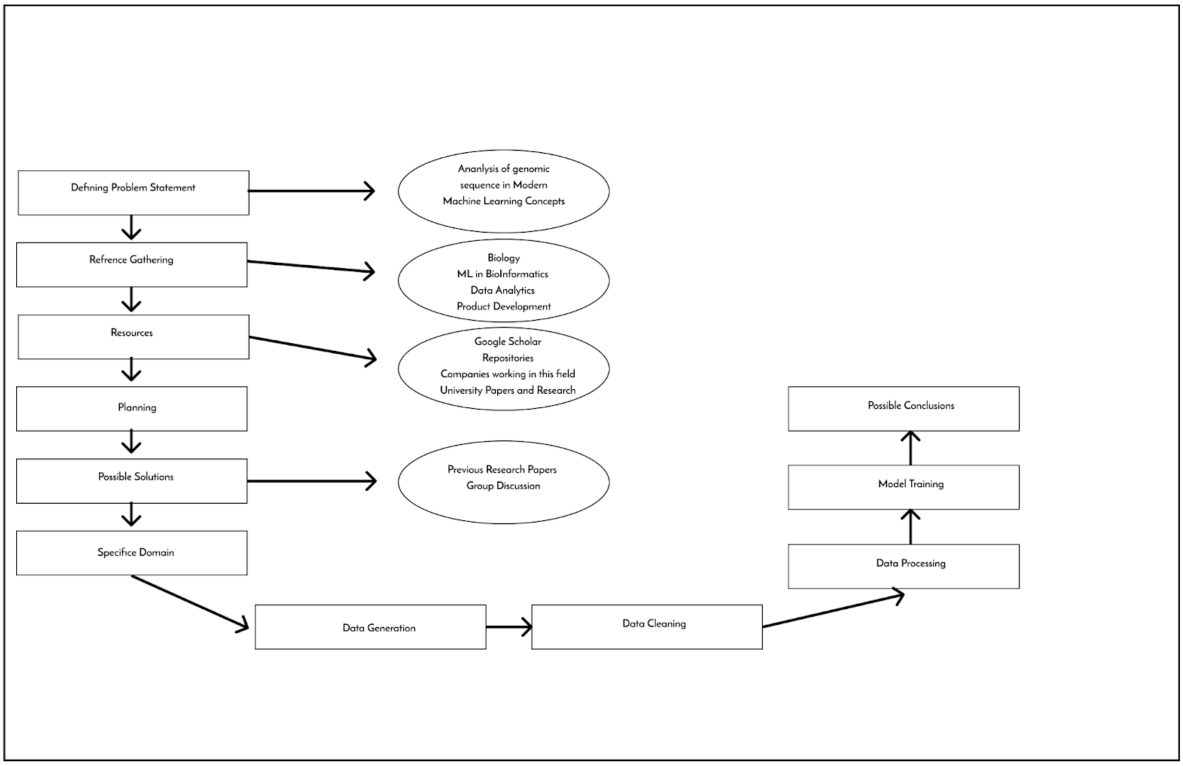
Human Genome de novo Assembly: The study has proposed a new method for the de novo assembly of the human genome from a short read sequence. This method successfully constructed frameworks of N50 contig sizes of 7.4 and 5.9 kilobases (kb) and 446.3 and 61.9 kb from the Asian and African human genomes. Please read al. 2010a

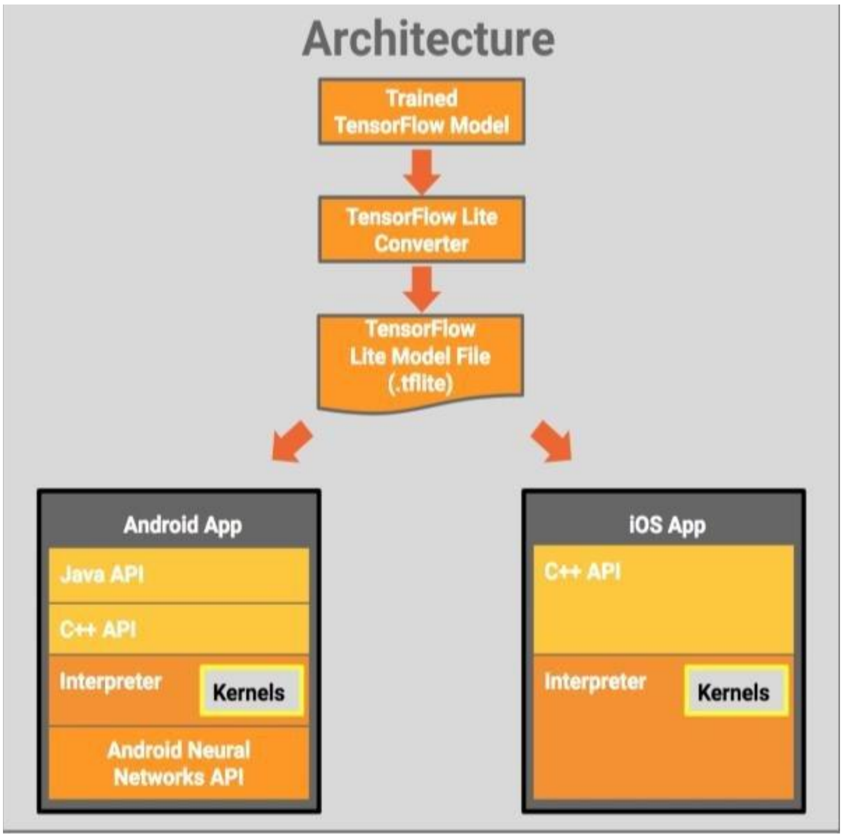
Assembling the Human Genome: A comprehensive review of recent developments in software packages for the analysis of new generation sequence data. Nagarajan and Pop 2010

HTNGS in Ancient Genome Research: The study sequenced the entire genome of a 4,000-year-old human with 20x coverage, providing a new perspective on the history of the human population. Shapiro and Hofreiter 2010

Topic of the work

a. System Design / Root Map:





b. Working Principle

- Defining this model for a given genomics dataset is a critical decision for bioinformaticians, biologists, and clinicians. This selection is defined in numerous use cases, for example, based on the consistency of the DNA sequence length across the dataset.
- Traditional ML methods (Linear and Logistics Regressions, Decision Trees, Support Vector Machines, Random Forest, Boosting Algorithms, Bayesian Networks, and so on) may be used to any length of DNA sequence.
- Modern ANN methods, like as CNN and RNN, need uniform DNA sequence length over the whole dataset column. The Py DNA library includes a simple method for determining whether or not the specified DNA sequence string has a consistent length.
- Let's start with the first use case. Assume we need to create a classification model that can predict a gene family based on a dataset of human DNA sequences.
- Genes are classified into families based on common nucleotide or protein sequences. A gene family is a group of multiple related genes produced by the duplication of a single original gene and often performing similar biochemical tasks.

c. Expected Results:

The predicted outcomes are determined by six key computed metrics used in classification model validation: accuracy score, precision, recall, f1 score, confusion matrix, and classification report. To identify an effective algorithm for training the model to achieve the best outcomes.

INDIVIDUAL CONTRIBUTION:

Ananya Singh (19BCE10177)

Bioinformatics is a multidisciplinary field that incorporates molecular biology, genetics, computer science, mathematics, and statistics. To solve data-intensive, large-scale biological concerns, computational techniques are applied. The most common concerns are biological process molecular modelling and data inference. Typically, a bioinformatics solution comprises of the following steps: Biological information may be utilised to produce statistical data. Construct a computational model. Use computational modelling to solve a problem. Put a computer algorithm through its paces and see how it fared.

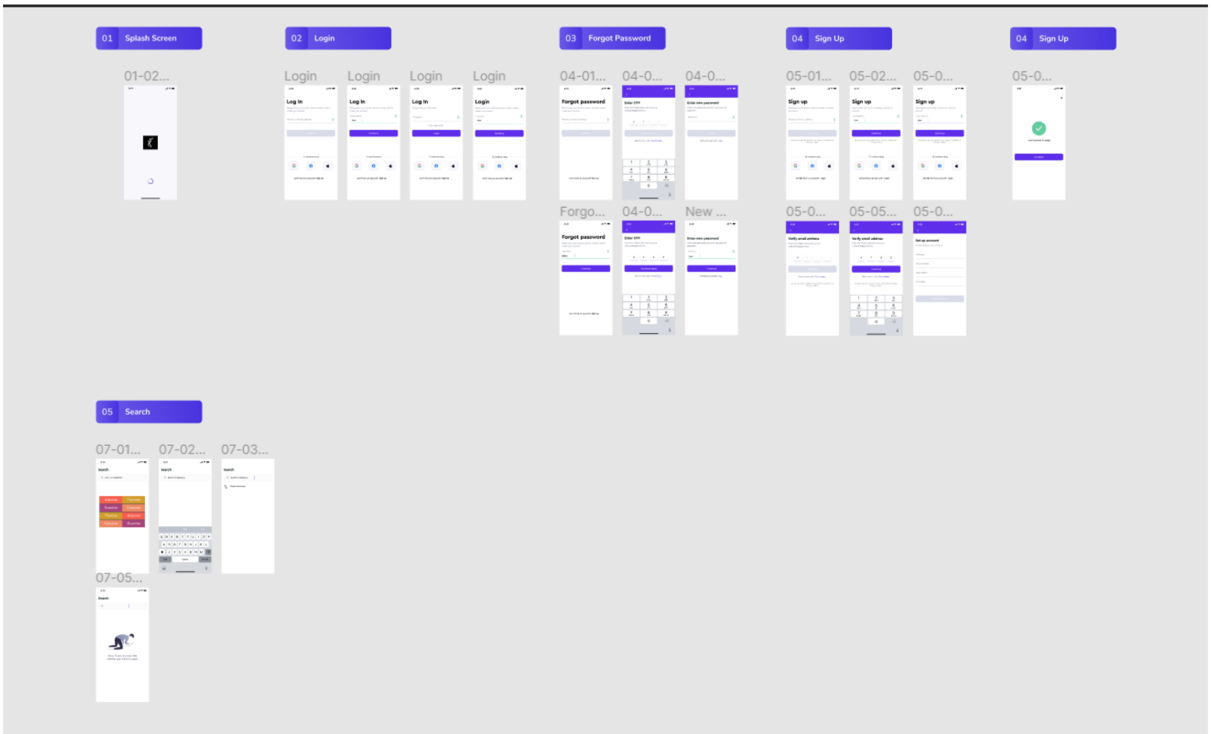
This project required me to work as a mobile application developer and front-end designer. I successfully finished all of the UI/UX design campaigns with all of the required beauty, sturdiness, and user-friendliness. As a member of the mobile app development team, my responsibility was to ensure that the deployment system ran as smoothly as the predicting system; in the end, users will interact with the front end rather than the model directly, which is important because the user may not be as technically savvy to understand the model predictions.

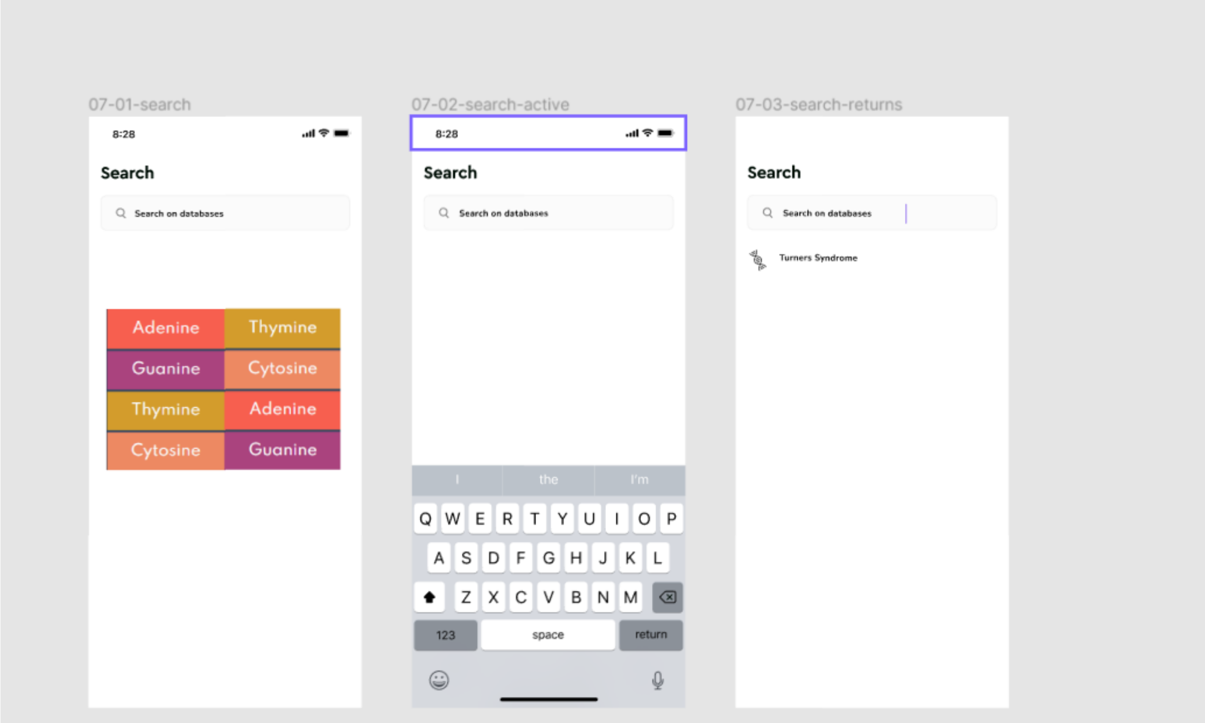
In the previous phases I have **designed** the **UI** of the Flutter using Figma which later fetches the files from the user as input.

a. **UX/UI Development:**
   Technology Used:
    1. Figma
    2. Adobe XD
    3. Flutter

Figma is a vector graphics editor and prototype tool that is mostly web-based, with desktop apps for macOS and Windows enabling extra offline functionality. The Figma mobile app for Android and iOS allows users to view and interact with Figma prototypes on their mobile devices in real time. Figma's feature set is geared at usage in user interface and user experience design, with a focus on real-time collaboration.

Results and Screenshots:

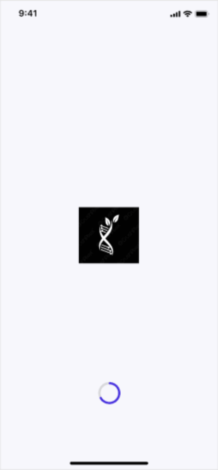




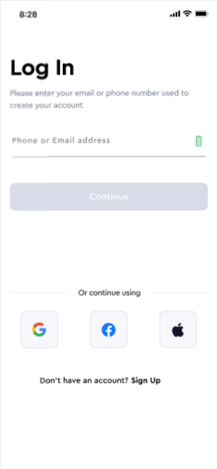
01 Splash Screen

02 Login

01-02-splash-screen-loading



Login


8:28

Log In

Please enter your email or phone number used to create your account

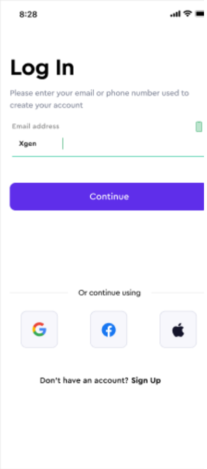
Phone or Email address

Continue

Or continue using

Don't have an account? **Sign Up**

Login


8:28

Log In

Please enter your email or phone number used to create your account

Email address

Xgen

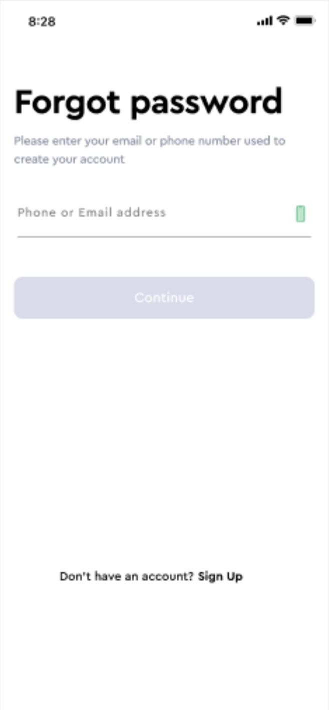
Continue

Or continue using

Don't have an account? **Sign Up**

03 Forgot Password

04-01-forgot-password-blank


8:28

Forgot password

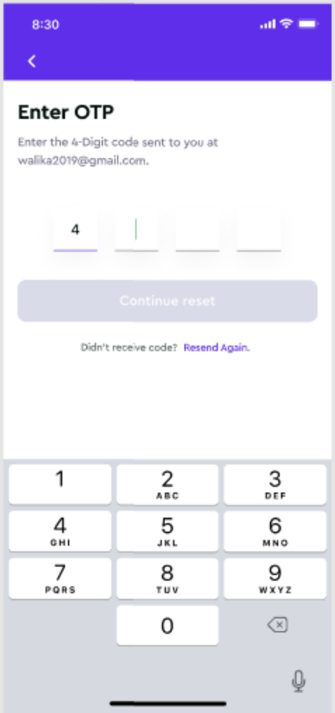
Please enter your email or phone number used to create your account

Phone or Email address

Continue

Don't have an account? **Sign Up**

04-02-forgot-password-otp


8:30

Enter OTP

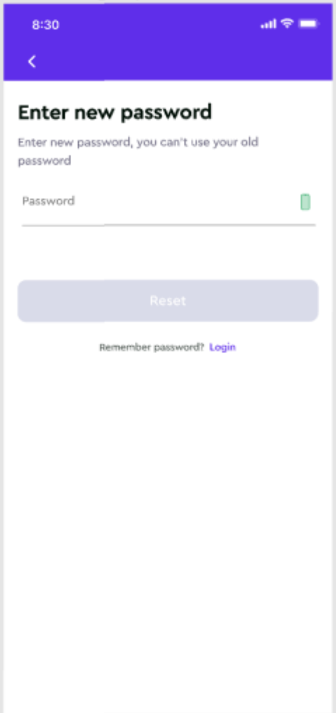
Enter the 4-Digit code sent to you at walika2019@gmail.com.

4

Continue reset

Didn't receive code? Resend Again.

04-03-forgot-password-new


8:30

Enter new password

Enter new password, you can't use your old password

Password

Reset

Remember password? Login

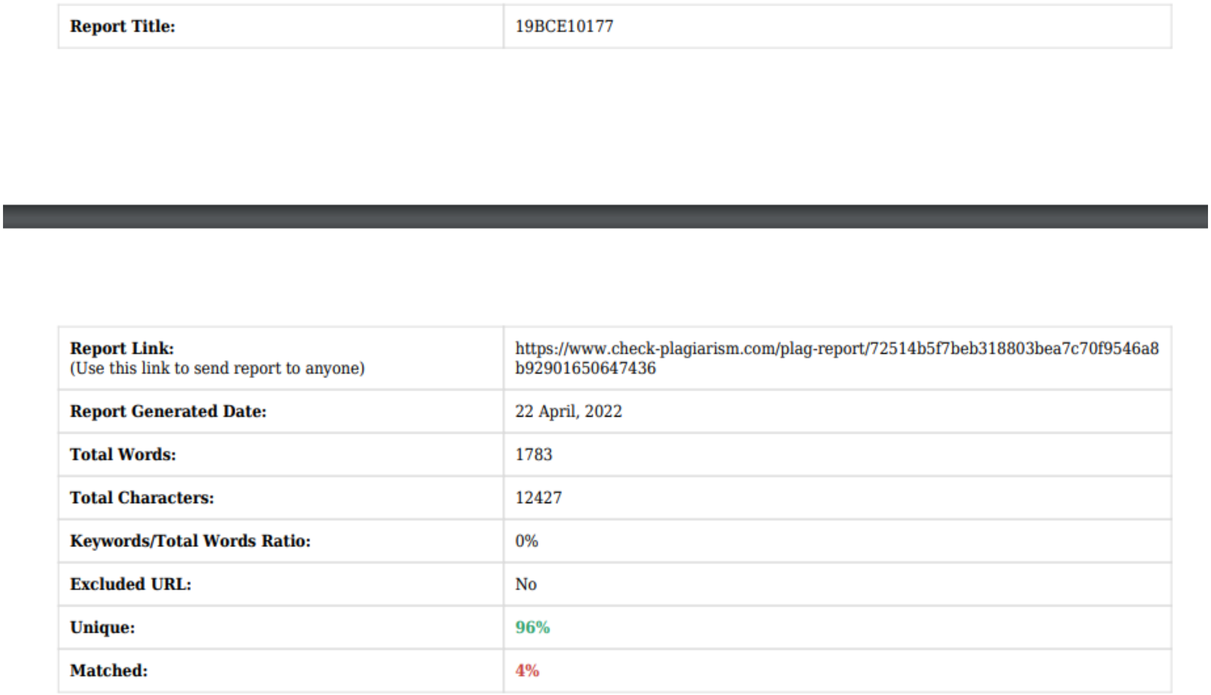
Conclusion

We acquired information about numerous bioinformatics approaches in phase 1 and found that 'Multinomial Naive Bayes and Multi-layer Perceptron classifier models' is the best algorithm for constructing a model based on that algorithm. Throughout phase 2 development, we worked on all imaginable components of the prediction and deployment application. In this final phase, we have deployed the models on the web as well as the app. To ensure the model's longevity, we set out to make it as strong and modular as possible, so that the most important characteristic we bring to society is the flexibility and agility to go forward in this fast changing technological world. We are all aware that the pandemic has caused significant hardship for the vast majority of us. This project will help us get a solid start because of its speedier clinical techniques, such as ease of sample and results display, as well as the integration of Machine Learning, which will make the process faster and more accurate. There might be a lot more reports and a lot of faster sequencing to find new mutations, especially dangerous ones.

Reference

1. https://towardsdatascience.com/machine-learning-for-genomics-c02270a51795?gi=e4e0c90e2e7b
2. https://www.researchgate.net/publication/353291346_Analysis_of_DNA_Sequence_Classification_Using_CNN_and_Hybrid_Models
3. https://firebase.google.com/?gclid=Cj0KCQjwpImTBhCmARIsAKr58cyTCm4RNP54H-sLGxaytQT1ADQltN6-36Btf4PrkMPkRZ47DSNAe7caAgomEALw_wcB&gclsrc=aw.ds
4. https://www.w3schools.com/
5. https://developer.mozilla.org/en-US/
6. https://www.ecma-international.org

Plagiarism Report:



Report Title:	19BCE10177

Report Link: (Use this link to send report to anyone)	https://www.check-plagiarism.com/plag-report/72514b5f7beb318803bea7c70f9546a8b92901650647436
Report Generated Date:	22 April, 2022
Total Words:	1783
Total Characters:	12427
Keywords/Total Words Ratio:	0%
Excluded URL:	No
Unique:	**96%**
Matched:	**4%**