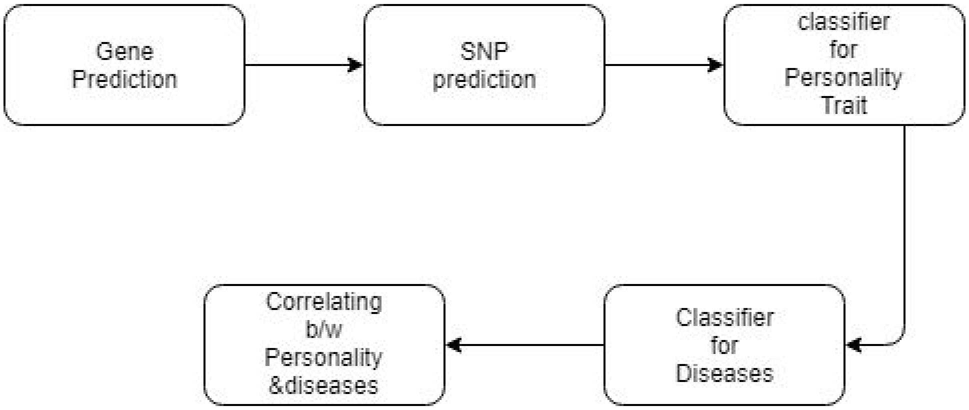
ARCHITECTURE OF PROJECT

OVERVIEW

Introduction

These Modules are the over all phases and abstract view of the project going to be done and the detailed view of the project will be explained below.


Gene Prediction
SNP prediction
classifier for Personality Trait
Classifier for Diseases
Correlating b/w Personality &diseases

Gene Prediction

Responsible for the gene prediction for the sample only for predetermined genes in the research papers (primarily mentioned in the document) while all other neglected using threshold.

SNP Prediction

Responsible for the single nucleotide polymorphism prediction which is also responsible for the mutations in the genes and forms a keyhole in disease and personality prediction

Personality Trait Classification

Responsible for the classification of the personality trait based on the genotype and snp identified from the test sequence using above two phases.

Disease Classification

Responsible for the disease classification from the genotype and SNP using the first two phases

Disease vs Personality

Responsible for finding the correlation between the diseases and personality traits from results obtained from the third and fourth phase of the project. Thus forms a meaning to our project.

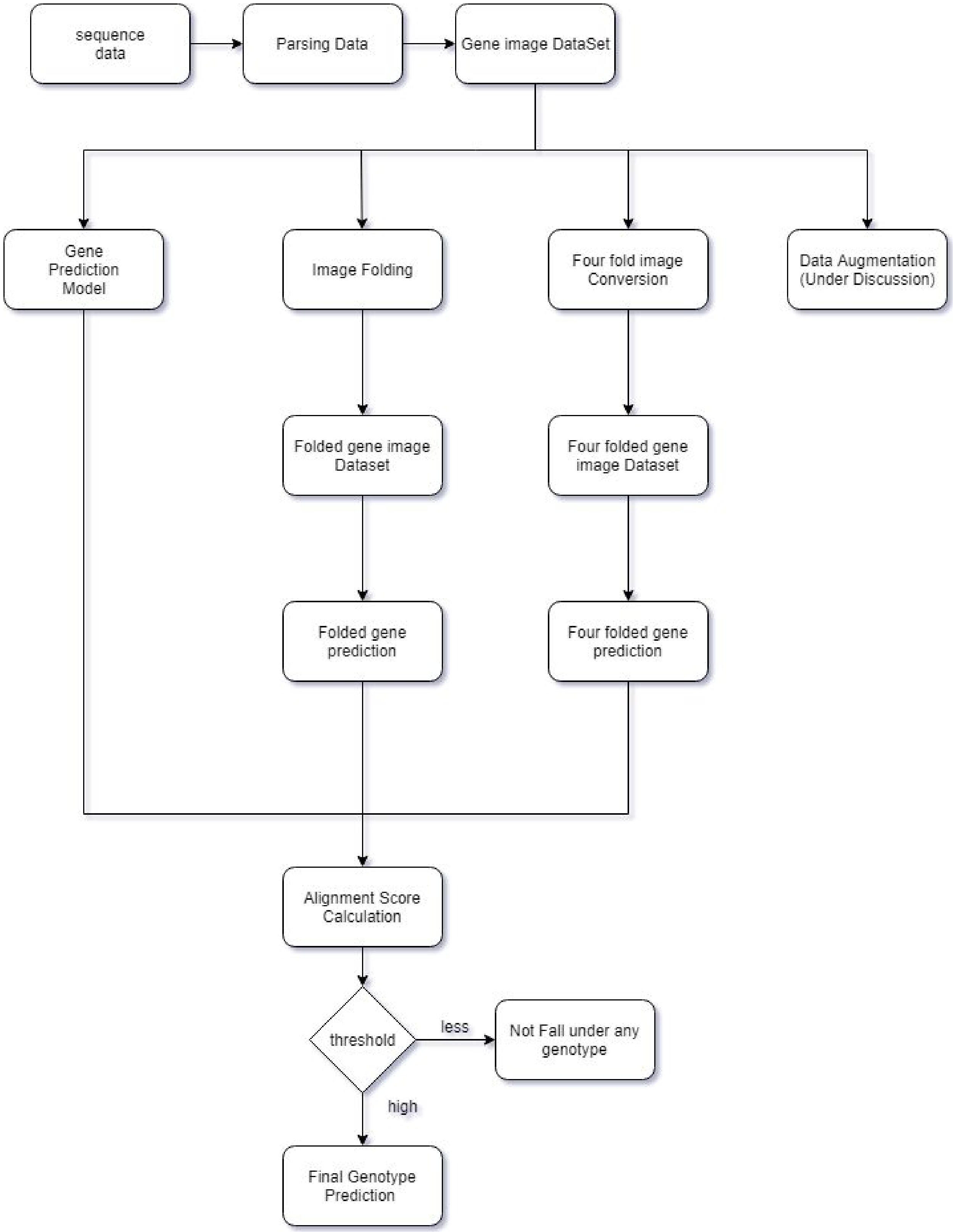
DETAILED VIEW

(Completed Phase → Gene Prediction)

Gene Prediction Architecture

The Flow goes as follows according to the figure provided below. It reads the sequence data from the sequence dataset which had been downloaded from the NCBI from the gene database. As there are large number of gene are there in the database we limited to our dataset collection to the preceded genotypes mentioned as per the researches done so far. Then Parsing the data comes into the picture as we cannot load and store the data each time as it was too costly in both time and space so Parsing data is essential. So we move on in converting the gene sequence into a image using the opencv with representation of four colors (green → G, Red → A, Blue → T, Black → C). In this way all the gene sequences are converted to gene sequence to image representation of genes which eventually lead to the gene image dataset as this going to be the root data for the further processing of data. Then this data is split up into folded images, one set stores the images in the vertical fold and the other will have the horizontal fold and then another way of four fold is also done with some slight preprocessing of data to improve the accuracy in the prediction of the genotypes. Two fold method holds the dataset with two divisions such as horizontal fold and vertical fold. The Four fold method holds dataset of one quarter of the original image believing that would help in prediction. while Data Augmentation Phase will form the 20 different ways of image distortion (still in discussion to use or not). Then Convolution neural network comes into the picture (Architecture of the ConvNet is given in the first review document) extract the feature and hidden pattern over the images from the dataset so there would three models each predicts the genotype. Now Upon getting the three gene predictions traditional PWA - Pairwise Sequence Alignment is done to get the final genotype from the test sample by fixing the threshold to avoid the sequence that does not hold any of the mentioned genotypes.

Architecture of Gene Prediction


sequence data
Parsing Data
Gene image DataSet
Gene Prediction Model
Image Folding
Four fold image Conversion
Data Augmentation (Under Discussion)
Folded gene image Dataset
Four folded gene image Dataset
Folded gene prediction
Four folded gene prediction
Alignment Score Calculation
threshold
less
Not Fall under any genotype
high
Final Genotype Prediction