Machine Learning Genetic Predictor: Deltect

CMPT 310: Introduction to AI Course Project

Brandon Tang, Tanvir Samra, Kaira Martinez, Jenny Wei

Fall 2025

1 AI System (1 page)

1.1 Introduction

The goal of our project is to develop a machine learning model that can predict the pathogenicity of genomic deletions. Genomic deletions are a type of structural variant where a segment of DNA is missing. Depending on where they occur, deletions can be associated with benign, to pathogenic variants associated with diseases.

In clinical genetics, manually classifying the pathogenicity of a deletion is challenging and time consuming. Fortunately, there are databases such as ClinVar that provide curated labels of pathogenicity, but for most, deletions remain unclassified, especially when extracted from sequencing data such as BAM/SAM files that are commonly used in practice. Our model aims to leverage the curated deletions to train a classifier than can estimate a patient's genomic data through BAM/CRAM files, and classify if a deletion is pathogenic.

1.2 Methodology

Our methodology for our AI model is based on a system with two main components:

1. **Deletion Extractor**
   Identifies deletions from analyzing a BAM's CIGAR strings. (should this be included? probs not right because there is inference)

2. **Pathogenic Predictor**
   A supervised learning model trained on ClinVar deletion variants to predict the pathogenicity of these deletions extracted from the Deletion Extractor.

1.2.1 Feature Engineering

For each deletion variant, we construct a set of features based on its position and annotations:

- Chromosome Number
- Start and end points ..... continue this

- talk about categorical features and standardized to adhere to numeric feature vector consistency. - mention LabelEncoder and StandardScaler

1.2.2 ML Model Choice and Training

We used Sci-kit Learn's Random Forest Classifier to predict pathogenicity.

The classifier produces a probability that each variant is pathogenic (?) Talk about the data with unknown and if we gave it a probability like MS 2

1.2.3 Testing

We split 80% of the dataset for training and 20% for testing. We perform 10-fold cross validation on the training set to esitmate the performance of our model. We evaluate the model using mean squared error (MSE) on the predictions, and as well as precision, recall, and specificity.

1.3 AI Pipeline

Our system can be viewed as an end-to-end AI pipeline with two main stages:

1. **Training**
   Using Random Forest to learn pathogenicity from ClinVar
2. **Inference**
   The trained predictor, is reused during inference on deletions extracted from BAM files.

Training Pipeline
Inference Pipeline

1.4 Limitations

2 Features Table (1-2 pages)

Description	Platform	Completeness	Code	Authors(s)	Notes
Random Forest Regressor	Python				
BAM extraction					

3 External Tools & Libraries (1/2 page)

3.1 Datasets

3.2 Frameworks

3.3 External Open Source