Deltect: Genomic Deletion Pathogenicity Classifier

CMPT 310: Introduction to Artificial Intelligence

Brandon Tang, Tanvir Samra, Kaira Martinez, Jenny Wei

Fall 2025

1 AI System

1.1 Introduction

The goal of our project is to develop a machine learning system, **Deltect**, that predicts the pathogenicity of genomic deletions. Genomic deletions are structural variants where a segment of DNA is missing. Depending on their location and size, deletions range from benign polymorphisms to pathogenic variants associated with genetic diseases like cancer predisposition syndromes (e.g., BRCA1/2 deletions) and hereditary disorders.

In clinical genetics, manually classifying deletion pathogenicity is time-consuming and requires expert knowledge. While databases like ClinVar provide curated pathogenicity labels, the vast majority of deletions found in patient sequencing data (BAM/CRAM files) remain unclassified. Deltect addresses this gap by training on clinically annotated deletions to predict pathogenicity for novel variants extracted from patient genomes.

1.2 Methodology

Our methodology involves supervised learning trained on ClinVar deletion variants and HG002/1000 Genomes to predict the pathogenicity. We use Sci-kit learn's Random Forest Classifier, with Gradient Boosting and optional XGBoost

1.3 AI Methods

Ensemble Classification: Random Forest (200 trees) + Gradient Boosting (200 estimators) + XGBoost (optional). Soft voting aggregates probability predictions. Balanced class weights address imbalanced training data.

Features: 18 biological attributes: genomic location (chromosome, position, deletion size), sequence composition (GC content, complexity, repeats), gene context (known disease genes, gene presence).

1.4 Data Acquisition

ClinVar (NCBI): 11,218 deletion variants on chr17 via Entrez API (9,559 pathogenic, 1,659 benign/likely benign). Chromosome 17 contains high-value genes (BRCA1, BRCA2, TP53).

Reference Genome Sampling (hs37d5): 7,852 benign regions sampled from reference genome to balance dataset. Matches deletion length distribution (1bp–15Mb). Final ratio: 1.01:1 (pathogenic:benign).

Supporting Resources:
- hs37d5 reference genome (GRCh37, 3GB FASTA)
- GENCODE v19 GTF (gene annotations, 1GB)
- NIST GIAB HG002 benchmark VCF (validation, 200MB)

Mitigating Class Imbalance

To address this issue of a lower recall performance, we balanced the training set of 10,000 deletions with:
- 5,000 pathogenic deletions (from ClinVar)
- 5,000 benign or uncertain deletions (from HG002)

This balance ensures that the model avoids majority class bias, and improves sensitivity to pathogenic variants.

Rationale for Balancing: ClinVar alone is 85% pathogenic, causing over-prediction. Reference sampling adds benign variants, improving specificity from 62% to 88.6% while maintaining 96% recall.

1.5 Feature Engineering

Genomic (7): chr, deletion_length, log_deletion_length, normalized_chr_position, is_small/medium/large_del

Sequence (6): gc_content, at_content, cpg_islands, repeat_content, homopolymer_run, complexity_score (Shannon entropy)

Gene Context (5): has_gene, is_known_disease_gene (30 curated genes), gene_length, is_ensembl_id, gene_encoded

Architecture:
- Random Forest: 200 trees, max_depth=15, balanced class weights
- Gradient Boosting: 200 estimators, learning_rate=0.05
- XGBoost: scale_pos_weight auto-tuned (0.99)

Performance (10-fold CV): Precision 89.54%, Recall 96.89%, Specificity 88.62%, AUC-ROC 97.44%

Test Set: Precision 89.42%, Recall 96.02%, Specificity 88.60%, AUC-ROC 97.54% (TP=1,835, TN=1,686, FP=217, FN=76)

1.6 AI Pipeline

Training: Fetch ClinVar variants → Preprocess coordinates → Extract sequences → Sample reference genome → Engineer 18 features → Train ensemble (10-fold CV) → Save model

Inference: Parse BAM CIGAR strings → Extract deletions → Annotate genes (GTF) → Extract sequences → Predict probabilities → Output ranked JSON (threshold$\geq$0.6)

1.7 Limitations & Mitigations

1. Single-chromosome training (chr17): May not generalize to all chromosomes. *Mitigation:* Chr17 contains diverse gene types; future work to expand to all autosomes.

2. Gene annotation dependency: Accuracy drops 5–10% without gene context. *Mitigation:* Require GTF input and warn when missing.

3. Assembly conflicts: ClinVar uses GRCh37/38; BAM files may differ. *Mitigation:* Document hs37d5 requirement.

4. Specificity gap: 88.6% vs recall 96.0%. *Mitigation:* Adjustable threshold (default=0.6); clinically prioritizes pathogenic detection.

5. Sequence extraction: Training has sequences; BAM variants need fetching. *Mitigation:* Inference pipeline auto-extracts via pysam.

2 Features Table (1-2 pages)

Description	Plat	Comp	Code	Author(s)	Notes
ClinVar API Client	Local	5	Python	Brandon, Kaira	Entrez API, 200/batch
Variant Preprocessing	Local	5	Python	Brandon, Jenny	Extract chr:start-end, seqs
Ref Genome Sampler	Local	5	Python	Brandon	Match size dist, ratio=0.7
Feature Engineering	Local	5	Python	Brandon, Jenny	18 features, no leakage
Random Forest	Local	5	Python	Brandon	200 trees, balanced
Gradient Boosting	Local	5	Python	Brandon	200 est, lr=0.05
XGBoost	Local	5	Python	Brandon	Auto scale_pos_weight
Ensemble Voting	Local	5	Python	Brandon	Soft vote, thresh=0.6
Cross-Validation	Local	5	Python	Brandon, Jenny	Stratified 10-fold
BAM Deletion Extract	Local	5	Python	Brandon	CIGAR parse, MAPQ$\geq$20
Gene Annotation	Local	5	Python	Brandon, Tanvir	GTF overlap, GENCODE v19
Pathogenicity Pred	Local	5	Python	Brandon, Jenny	Main inference, JSON out
Visualization	Local	5	Python	Tanvir, Kaira	Conf matrix, ROC, resid
Training Pipeline	Local	5	Python	All	End-to-end auto
Inference Pipeline	Local	5	Python	Brandon, Tanvir	BAM $\rightarrow$ predictions
Validation Pipeline	Local	3	Python	Brandon	GIAB comparison
Jupyter Demo	Local	5	Python	All	Step-by-step tutorial
CLI (main.py)	Local	5	Python	Brandon	train/infer/validate

3 External Tools & Libraries

ML Frameworks: scikit-learn (v1.7.2): RandomForest, GradientBoosting, train_test_split, StratifiedKFold, RobustScaler, LabelEncoder, metrics; XGBoost (v2.1.3); pandas (v2.3.3); numpy (v2.3.4)

Genomics: pysam (v0.23.3): BAM/FASTA parsing, CIGAR analysis; biopython (v1.84): Entrez API

Visualization: matplotlib (v3.10.0), seaborn (v0.13.2)

Utilities: python-dotenv, pathlib, logging

Datasets:
- ClinVar (NCBI): 11,218 deletions (chr17); Entrez API access
- hs37d5 (1000 Genomes): 3GB FASTA; sequence extraction (we only use a subset of this)
- GENCODE v19: 1GB GTF; gene annotations (GRCh37)

Open-Source Code: None reused. All original implementation. Followed sklearn/pysam documentation for ClinVar batching, reference sampling, CIGAR parsing, feature engineering, ensemble training.

Dependencies: Python 3.9+, 8GB RAM, 5GB storage

License: Educational use (CMPT 310). Authors retain copyright; SFU has academic sharing permission.