



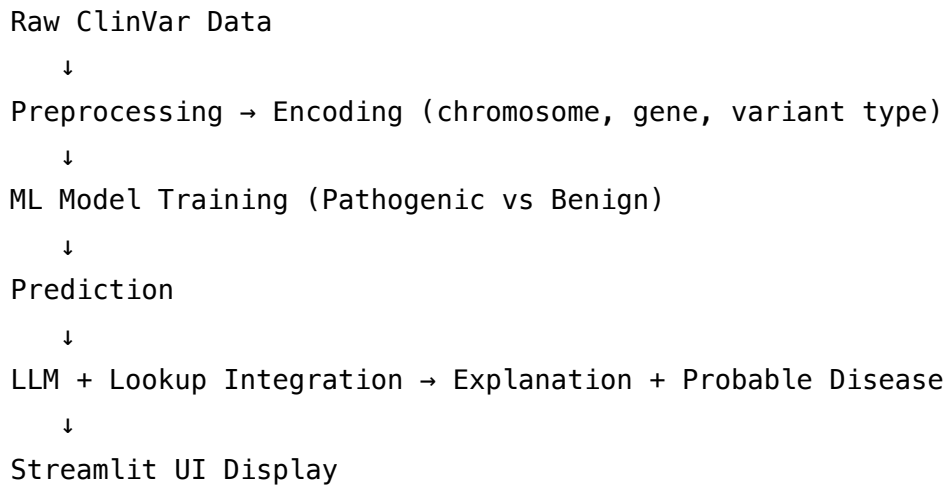
AI-powered Disease Prediction from Gene Mutations

Genome-Dx is a genomics-driven machine learning project that predicts the **pathogenicity** (benign / pathogenic / uncertain) of gene mutations and provides **probable disease associations** with **LLM-generated explanations** for better interpretability.






Project Overview

Component	Description
Objective	Predict whether a gene mutation is <i>pathogenic</i> or <i>benign</i> and infer the most probable disease association.
ML Model	Supervised classification model trained on encoded genomic features: Chromosome_Encoded , Gene_Symbol_Encoded , IS_SNP , IS_INDEL .
LLM Integration	Uses Gemini LLM through DSPy to generate natural-language reasoning and human-readable explanation text.
Data Source	Processed ClinVar-derived dataset (gene mutations, clinical significance, and disease annotations).
Interface	Streamlit app for interactive prediction, explanation, and variant lookup.

Workflow



Features

-  **ML-based Pathogenicity Prediction** — trained on numeric genomic features.
-  **Probable Disease Lookup** — gene-level and variant-level mapping using ClinVar disease names.
-  **LLM-Generated Explanations** — contextual text via Gemini + DSPy for interpretability.
-  **Hybrid Results** — combines deterministic lookup with AI reasoning.
-  **Streamlit Interface** — user-friendly input/output layout for experimentation.

Tools & Technologies

Category	Stack
Language	Python 3.10
Frameworks	scikit-learn, pandas, numpy
LLM / AI	Google Gemini LLM + DSPy
Visualization / UI	Streamlit
Data Handling	joblib, csv, numpy

Category	Stack
Version Control	Git / GitHub

Future Extensions

- Integrate **Qdrant** / **FAISS** for semantic disease retrieval.
- Add **multi-task model** to jointly predict disease + pathogenicity.
- Include **phenotype and clinical notes** for richer context.
- Expand dataset with **OMIM** / **HPO mappings** for rare diseases.

Example Output

Input: Gene = BRCA1, IS_INDEL = 1

Model Prediction: Pathogenic

Probable Disease: Hereditary breast and ovarian cancer

Explanation:

Based on ClinVar and model inference, this mutation in *BRCA1* is classified as pathogenic and is most frequently associated with hereditary breast and ovarian cancer (HBOC).

License

This project is released for educational and research use.

© 2025 Vishal Saxena

Genome-Dx — *bridging genomics and AI to explain disease risk from gene mutations.*