DNA Dataset Creation and Fusion Methodology

1. Data Sources Used

1.1 Sequences Dataset (`sequences.csv`):
- Source: Custom generated synthetic sequences mimicking random 100-nucleotide DNA sequences.
- Contribution: Provided the genomic base sequences.
- Reference: [Synthetic DNA Generator Tools](https://www.bioinformatics.org/sms2/random_dna.html)

1.2 Parameters Dataset (`parameters.csv`):
- Source: Synthetic parameters generated to mimic mutation rates, GC content bias, and sequence variability.
- Contribution: Provided mutation rate, GC content bias, and variation rates per sequence.
- Reference: [Nucleotide Composition & Mutations Studies](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4030807/)

1.3 Region/City Dataset (Manually defined synthetic regions like Asia, Europe, Africa, etc.):
- Contribution: Provided contextual regional tags to sequences.
- Simulated known regional disease risks and mutation prevalence.
- Reference: [Global DNA Variation Studies](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7408841/)

2. Technical Methodology for Combining Datasets

2.1 Objective:
To create a region-aware DNA dataset that embeds region-specific variations and mutation rates, supporting models that predict disease risks with awareness of genomic and regional diversity.

2.2 Combination Strategy

Component	Purpose	How It Was Integrated
Sequences (`sequences.csv`)	Base DNA sequences containing motifs and	Each row provided a unique 100-nucleotide sequence

	random regions.	forming the genomic base.
Parameters (`parameters.csv`)	Provided mutation rates, GC content bias, and variation rates per sequence.	Parameters were applied to each sequence to introduce synthetic but controlled mutations.
Cities (Regions)	To simulate region-specific disease risk and mutation patterns.	Each sequence was tagged with a synthetic region and adjusted based on corresponding parameter variations.

2.3 Technical Logic of Fusion

- Controlled Variability Per City:
Different cities/regions were linked to specific mutation and GC content patterns (from `parameters.csv`). This models the real-world observation that environmental and genetic backgrounds vary by region.

- Synthetic Diversity Injection:
Based on the mutation rates from `parameters.csv`, sequences from `sequences.csv` were mutated per city. Regions with higher simulated mutation rates had more nucleotide variations injected. GC-rich or AT-rich biases were enforced as per the parameter bias.

- Epidemiological Mimicking:
Diseases were associated probabilistically to regions, ensuring that some diseases are overrepresented or underrepresented based on known epidemiological patterns.

2.4 Example Workflow

For each city:
  For each sequence:
    Apply mutations as per 'parameters.csv' for that city
    Adjust GC/AT content based on parameter bias
    Assign disease label based on city-specific disease prevalence
    Add city, mutated sequence, parameters, and label to final 'dna_dataset.csv'

2.5 Technical Benefits

- Simulation of real-world data complexity.
- Model testing on region-specific bias handling.

- Controlled and reproducible generation process.
- Supports advanced models like region-aware classifiers or federated learning scenarios.