A Hybrid Machine Learning, Deep Learning, And Graph Neural Network Framework for HIV Drug Resistance Prediction and Mutation Interaction Modeling

Jeevietha
Department of Computer Science and Engineering
Amrita School of Computing, Bangalore
Amrita Vishwa Vidyapeetham, India
jeevietha11@gmail.com

Rishi Kumar
Department of Computer Science and Engineering
Amrita School of Computing, Bangalore
Amrita Vishwa Vidyapeetham, India
rk.rishi.kumar.2006@gmail.com

Satvika S
Department of Computer Science and Engineering
Amrita School of Computing, Bangalore
Amrita Vishwa Vidyapeetham, India
satvikas2020@gmail.com

1. Introduction

HIV drug resistance remains a major clinical challenge due to the virus's high mutation rate . These mutations or changes occur in the key proteins like protease,reverse transcriptase, integrase, if these proteins are mutated it will lead to treatment failures. Therefore the prediction of drug resistance from genomic sequences using computer models (machine learning or deep learning etc)is very helpful and important for treatment optimization..

These predictive computer models require datasets that are:

- Biologically coherent
- Non redundant
- Statistically diverse
- Free from near duplicate bias

Repeating or redundant sequences can lead to inflated model accuracy which means high false accuracy.Therefore, the dataset must be cleaned before building any model which includes the two main steps- alignment-guided dataset validation and redundancy filtering ,before implementing hybrid ML, DL, and GNN models..

This project focuses about building a clean dataset by alignment validation, removing the duplicates and building a clean set of HIV sequences from four major classes , so that the computer model can very accurately predict whether the drug will resist the mutation or not .

2. Dataset Used and Biological Relevance

2.1 Source

The dataset was taken from the Stanford HIV Drug Resistance Database (HIVDB), which provides curated genotype–phenotype relationships between HIV mutations and drug resistance which means it maps HIV mutations with how well the drug will work on them .

Each record includes:

- Sequence ID
- Patient ID
- Positional amino acid fields (P1–Pn)
- Mutation list (CompMutList)
- Drug susceptibility measurements

2.2 Drug Classes Included

Drug Class	Target Enzyme
PI	Protease
NRTI	Reverse Transcriptase
NNRTI	Reverse Transcriptase
INI	Integrase

These classes are biologically meaningful as each drug class targets a specific viral enzyme which has a unique evolutionary pattern and different resistance mechanisms.

2.3 Dataset - Topic Relevance

The dataset supports and is valid to our research objective and topic.

All the sequences in the dataset belong to HIV-1 genome. In each of the 4 drug classes chosen, sequences are of the same functional protein. Positional columns show the aligned amino acid sequences. The mutation also shows the resistance phenotype. So this dataset is very valid for finding drug resistance and mutation analysis .

3. Methodology

3.1 Sequence Construction

Each dataset contained aligned amino acid columns (P1–Pn). These columns were combined to reconstruct a complete protein sequence for every record.

This step enabled proper sequence comparison and facilitated redundancy detection across samples.

3.2 Removal of Exact Duplicate Sequences

Sequences with identical mutation patterns were removed from the dataset.

This ensured that:

- Each row represented a unique viral genotype.
- Repeated clinical samples did not bias the analysis.

3.3 Phenotype Filtering

Rows that lacked drug-resistance phenotype values were excluded.

This ensured that:

- Each retained sequence had biological relevance.

- The dataset remained suitable for supervised learning applications.

3.4 Identity-Based Redundancy Removal

Redundancy was defined as sequences that were highly similar to each other.

Sequence identity was calculated as:

Identity = (number of matching positions) / (sequence length)

For each drug class:

- PI, NRTI, and NNRTI:
  Sequences with ≥95% identity were considered redundant, and only one representative sequence was retained.

- INI:
  Identity-based filtering was not applied because integrase sequences were already highly conserved. Applying a strict identity threshold significantly reduced the dataset size below the required minimum. Therefore, only exact duplicate sequences were removed.

This strategy ensured sufficient dataset diversity while maintaining a minimum of 500 sequences for analysis.

4. Results

4.1 Dataset Size

Drug Class	Initial Size	Final Size
PI	4350	3173
INI	1986	1542
NRTI	5355	1573
NNRTI	4911	1673

All final datasets contain more than 500 sequences, satisfying the assignment requirements.

4.2 Interpretation

The results indicate clear differences in sequence diversity across drug classes.

- The protease (PI) dataset retains a large number of sequences after filtering, suggesting high genetic diversity.

- The integrase (INI) dataset shows relatively smaller reduction, indicating that integrase sequences are more conserved.

- The reverse transcriptase datasets (NRTI and NNRTI) show moderate variation, with significant redundancy removed during filtering.

- Overall, redundancy removal successfully preserved mutation diversity while eliminating highly similar sequences.

The cleaned datasets are therefore suitable for downstream machine learning analysis and drug-resistance prediction.

5. Code

```
import pandas as pd

DATASETS = ["PI.csv", "INI.csv", "NRTI.csv", "NNRTI.csv"]

def sequence_identity(seq1, seq2):
    length = min(len(seq1), len(seq2))
    matches = sum(seq1[i] == seq2[i] for i in range(length))
    return matches / length


for file_name in DATASETS:

    print("\nProcessing:", file_name)

    df = pd.read_csv(file_name, low_memory=False)
    print("Initial size:", len(df))

        # Build sequence
        seq_cols = [c for c in df.columns if c.startswith("P") and c[1:].isdigit()]
    seq_cols = sorted(seq_cols, key=lambda x: int(x[1:]))

    df[seq_cols] = df[seq_cols].astype(str)
    df["FullSeq"] = df[seq_cols].agg("".join, axis=1)

      # Remove exact duplicates
        df = df.drop_duplicates(subset=seq_cols).reset_index(drop=True)
    print("After exact dedup:", len(df))

        # Remove rows with no resistance values
        numeric_cols = df.select_dtypes(include="number").columns
    df = df.dropna(subset=numeric_cols, how="all")
    print("After phenotype filtering:", len(df))

      # Identity filtering

    if file_name != "INI.csv":   # Apply only to PI, NRTI, NNRTI

        threshold = 0.95
        sequences = df["FullSeq"].tolist()
        kept = []
        reps = []

        for i, seq in enumerate(sequences):
            redundant = False
            for r in reps:
                if sequence_identity(seq, r) >= threshold:
                    redundant = True
                    break
            if not redundant:
                reps.append(seq)
                kept.append(i)

        df_final = df.iloc[kept].reset_index(drop=True)
        # If <500 → try 90%
        if len(df_final) < 500:
            print("Re-running at 90% identity")
            threshold = 0.90
            kept = []
            reps = []

            for i, seq in enumerate(sequences):
                redundant = False
                for r in reps:
                    if sequence_identity(seq, r) >= threshold:
                        redundant = True
                        break
                if not redundant:
                    reps.append(seq)
                    kept.append(i)

            df_final = df.iloc[kept].reset_index(drop=True)

    else:
        print("Skipping identity filtering for INI (retain ≥500)")
        df_final = df.copy()

    print("Final size:", len(df_final))

    output = file_name.replace(".csv", "_FINAL.csv")
    df_final.to_csv(output, index=False)
    print("Saved:", output)

print("\nAll datasets processed.")
```

6. Conclusion

This work produced non-redundant HIV drug-resistance datasets across four drug classes.

By removing duplicate and highly similar sequences while maintaining at least 500 sequences per class, the final datasets are:

- biologically valid
- statistically diverse
- suitable for ML/DL/GNN modeling

These cleaned datasets can now be used for predicting HIV drug resistance and analyzing mutation interactions.