# AI-Driven Prediction of Drug Adverse Effects Using Genomic DNA Profiling

**[1]Prasad Alai, [2]Sanket Gadekar, [3]Pranjal Vedpathak, [4]Vighnesh Narawade**

Department of Artificial Intelligence & Machine Learning, ISBM College of Engineering, Savitribai Phule Pune University, Pune, India

[1]prasadalai2004@gmail.com, [2]sanketgadekar42@gmail.com, [3]vedpathakpranjal@gmail.com, [4]vighneshnarawade@gmail.com

***Abstract*— Adverse drug reactions (ADRs) remain one of the leading preventable causes of hospitalization worldwide. Conventional prescribing practices do not account for individual genetic variability, leading to suboptimal drug responses and preventable toxicities. This paper presents an end-to-end AI-driven pipeline that predicts potential drug side effects for an individual by combining their raw genomic data with curated pharmacogenomics and drug-side-effect databases. Raw Variant Call Format (VCF) files from the 1000 Genomes Project are processed to extract single nucleotide polymorphisms (SNPs), which are filtered for variability, mapped to genes using GENCODE annotation, and aggregated into per-gene genetic scores. Gene-to-drug relationships from PharmGKB are then used to derive drug-level genetic scores, and side-effect associations are retrieved from SIDER. A binary ADR risk label is assigned per individual per drug by applying a 75th-percentile threshold on the population-level genetic-score distribution, producing a labelled, machine-learning-ready dataset. The framework demonstrates that population-scale genomic data can be transformed into personalized pharmacogenomic features suitable for supervised classification. The work is positioned relative to six reference studies on ADR prediction using genomic data and machine learning, highlighting its unique contribution of an uninterrupted VCF-to-label pipeline built entirely from publicly available resources.**

***Index Terms*— Adverse Drug Reactions (ADRs), DNA Profiles, Pharmacogenomics, Single Nucleotide Polymorphisms (SNPs), 1000 Genomes Project, PharmGKB, SIDER, Genetic Score, Machine Learning, Personalized Medicine.**

## I. INTRODUCTION

Adverse drug reactions (ADRs) are harmful effects that happen when a drug is taken in the right amount for treating a condition. These reactions lead to 5–7% of all hospital visits and are the fifth most common reason for death in hospitals worldwide [1]. The cost of these reactions is very high — over $100 billion happens every year in the US alone, which is almost as much as the cost of the drugs themselves [2]. Even though there has been a lot of research on drug safety, doctors still mostly use the same dose for everyone, without considering a person's genetic differences.

Pharmacogenomics helps solve this problem by looking at how a person's genes affect how their body processes drugs, how well they work, and how toxic they might be.

Changes in genes that control metabolic enzymes — especially the CYP family like CYP2D6, CYP2C9, CYP2C19, and CYP1A2 — are known to cause most of the differences between people in how drugs work in their bodies and cause about 60% of the ADRs that are seen in practice [3]. Small changes in these genes, known as SNPs, change how the enzymes work, leading to different types of metabolizer profiles ranging from those who don't process drugs well (which can lead to toxicity) to those who process them too quickly (which can make the drug less effective). Using this genetic information in medical decisions is the idea behind personalized medicine.

The potential of using data in pharmacogenomics has been shown by several important studies.

Liang et al. [2] showed that a deep learning model based on generative stochastic networks could correctly identify groups of patients who are more likely to have ADRs based on their CYP2D6 and CYP1A2 gene variations, with accuracy over 80%, which is better than traditional models like LASSO and k-NN. Seo et al. [7] found that adding SNP-based drug similarity measures to usual features used in drug research improved the prediction of side effects using a random forest classifier on the SIDER database by 3.5%. Dafniet and Taboureau [4] built a network of interactions between drugs, targets, and genetic mutations, using deep neural networks to predict ADRs at the organ level with good accuracy. He et al. [6] introduced DGANet, a convolutional neural network that combines chemical and gene interactions from the CTD database with drug-side effect links, achieving an AUROC of 92.76%, which is the best performance for predicting ADRs using pharmacogenomics. Del Casale et al. [5] reviewed machine learning applications in pharmacogenomics for mental health and found that random forest models combining SNP data with clinical information work well for predicting how well a drug will work and the chances of side effects. Our previous review [8] studied these developments and suggested an AI framework that uses SNPs and drug molecule data.

Despite these advancements, there is still a big gap — no system has yet shown a complete, reliable way to (i) take in large files of VCF data, (ii) match individual SNPs to overall genetic scores for genes, (iii) connect those gene scores to specific drugs via PharmGKB, (iv) get side-effect labels from SIDER, and (v) create a dataset labeled with ADRs — all in one process.

This paper fills that gap. The five main studies cited earlier are the scientific basis for each part of this process, and their methods are clearly tied to the steps in this system throughout the paper.

The rest of this paper is structured as follows.

Section II gives background on SNPs and the six key datasets. Section III explains the seven-step method. Section IV presents the results and data details. Section V discusses the findings, limitations, and future directions. Section VI concludes the paper.

## II. Background And Related Work

### A. Single Nucleotide Polymorphisms and Diploid Encoding

DNA is a four-letter sequence of nucleotide bases (A, T, C, G). A single nucleotide polymorphism (SNP) occurs when a single base position varies across individuals. Using the diploid encoding standard in VCF files, the genotype at any SNP locus is represented as: 0|0 → 0 (reference homozygous, no mutation); 0|1 or 1|0 → 1 (heterozygous, one mutated allele); 1|1 → 2 (alternate homozygous, both alleles mutated). This 0/1/2 dosage encoding captures the allelic load and is directly usable as a numeric feature in machine learning models. Liang et al. [2] pioneered this encoding strategy in their clinical CYP2D6/CYP1A2 study, applying it to 83 blood-sample observations to classify 14 ADR categories.

### B. Reference Studies and Their Relevance to This Pipeline

Table 1 maps each of the six reference papers to the specific pipeline step they inform.

Table 1: Reference Papers and Their Contribution to the Proposed Pipeline

| Ref. | Authors / Year | Core Technique | Databases Used | Relevant Pipeline Step |
|---|---|---|---|---|
| [2] | Liang et al., 2016 | Deep Learning (GSN + Hidden Markov Chain) | Clinical CYP2D6 / CYP1A2 PCR data | Step 2 — SNP encoding (0/1/2); Step 6 — ADR label logic |
| [4] | Dafniet & Taboureau, 2024 | Deep Neural Network on SNP-drug-ADR network | dbSNP, PharmGKB, DrugBank, DrugCentral | Step 4 — SNP-to-gene mapping; Step 5 — gene-to-drug via PharmGKB |
| [5] | Del Casale et al., 2023 | Systematic review: Random Forest + pharmacogenomics | PubMed, Scopus; 14 ML+pharmacogenomics studies | Step 6 — ADR label rationale; Section V discussion |
| [6] | He et al., 2025 (DGANet) | CNN with cross-attention on CGI + GDA features | SIDER 4.1, LINCS L1000, CTD, PubChem, MeSH | Step 5 — SIDER integration; Step 7 — ML dataset structure |
| [7] | Seo et al., 2020 | Random Forest + SNP similarity + 6 drug features | SIDER, DrugBank, PubChem, DisGeNET | Step 3 — SNP filtering rationale; Step 7 — RF baseline |
| [8] | Alai et al., 2025 (IJRTI) | Conceptual AI framework review: VCF SNP encoder + GNN drug encoder | 1000 Genomes, PharmGKB, DrugBank, PubChem | Overall pipeline architecture; all steps |

### C. Key Databases

Table 2 summarizes the five public databases used in the pipeline.

Table 2: Databases Used in the Proposed Pipeline

| Database | Contents | Role in Pipeline |
|---|---|---|
| 1000 Genomes Project | VCF files — SNP data for ~2,504 individuals across 26 populations | Primary genomic input (Step 1) |
| GENCODE v19 (GTF) | Chromosome-level gene start/end coordinates for all annotated human genes | SNP-to-gene mapping (Step 4) |
| PharmGKB | Curated gene-drug pairs: which genes govern which drugs' metabolism | Gene-to-drug score aggregation (Step 5) |
| SIDER 4.1 | 1,430 drugs × 5,868 side effects; 139,756 drug-side-effect associations | Side-effect label source (Steps 5–6); used in [6], [7] |
| PubChem Web | CID-to-drug-name mappings | Drug name resolution (Step 5) |

## III. Proposed Methodology

The proposed system comprises seven sequential processing steps. Figure 1 (described textually below) shows the data flow from raw VCF files to a labelled ML dataset. Each step is grounded in one or more of the six reference works.

### A. Step 1 — Raw DNA Data Acquisition (VCF Files)

VCF files for chromosome 6 are downloaded from the 1000 Genomes Project FTP server (https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/). VCF is the bioinformatics standard for representing genomic variants. Each record stores the chromosome, position, reference allele, alternate allele, and per-individual diploid genotype. As noted in [8], chromosome-6 data are particularly rich in pharmacogenomically relevant genes (NQO1, IRF4, DUSP22, EXOC2) that overlap with PharmGKB drug associations.

### B. Step 2 — VCF to CSV Conversion and SNP Encoding

VCF files are parsed and converted to a rectangular CSV matrix of dimensions [N_persons × N_SNPs]. Each cell receives the integer dosage value {0, 1, 2} derived from the diploid genotype string. This encoding scheme was validated by Liang et al. [2], who applied identical genotype coding (wild-type / heterozygous / homozygous-mutant) when classifying 83 patients by CYP2D6 and CYP1A2 alleles into 14 ADR categories, achieving > 80% classification accuracy with their deep generative model.

Example: At chromosome 6, position 63,979, the reference allele is CAG and the alternate is C (an AG deletion). A person with genotype 0|1 at this locus carries one mutated chromosome and receives the value 1 for SNP_6_63979.

### C. Step 3 — SNP Filtering

Population-scale VCF files contain millions of positions, the majority of which are monomorphic (identical for all individuals) and thus uninformative. A minor allele frequency (MAF) filter removes positions with zero minor allele count in the study cohort, retaining only variant positions. This step is motivated by Seo et al. [7], who explicitly note that selecting SNPs associated with disease-related gene expression (via expression-quantitative-trait-loci mapping) dramatically improves the discriminative power of the SNP feature set over using all available SNPs indiscriminately.

### D. Step 4 — SNP-to-Gene Mapping and Genetic Score Calculation

Each retained SNP position is mapped to a gene using the GENCODE v19 annotation (https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_19/gencode.v19.annotation.gtf.gz). For each SNP at chromosomal position p, the algorithm tests whether $p \in [\text{gene\_start}, \text{gene\_end})$ for any annotated gene on chromosome 6. Matched SNPs are assigned to their gene; unmatched SNPs are discarded. The per-person genetic score for each gene is then:

$$\text{GeneScore}(\text{person}, \text{gene}) = \Sigma\ \text{SNP\_value}(\text{person}, \text{snp}),\ \forall\ \text{snp} \in \text{gene}$$

Dafniet and Taboureau [4] use an analogous gene-level aggregation: they build a matrix of [drug × mutation] where each mutation is a SNP on a drug-target gene from dbSNP, and they confirm that mutations in metabolic enzyme families (CYP, SLCO, ABC) dominate the ADR signal. Our pipeline mirrors this philosophy by aggregating SNP dosage at the gene level before passing scores downstream.

The resulting output is a [N_persons × N_genes] gene-score matrix. Representative values for a sample of 15 individuals and genes (LINC00266-3, CICP18, RP3-416J7.1, RP3-416J7.4, RP3-416J7.5, DUSP22, IRF4, EXOC2) are stored in the intermediate CSV at this stage.

### E. Step 5 — Gene-to-Drug Mapping and Drug-Level Genetic Score

PharmGKB provides curated gene-drug relationship pairs. Each record links a metabolic gene (e.g., NQO1, CYP2C9, VKORC1) to one or more drugs or drug classes. For each person-drug pair, the drug-level genetic score is the sum of gene scores across all PharmGKB-listed genes for that drug:

$$\text{DrugScore}(\text{person}, \text{drug}) = \Sigma\ \text{GeneScore}(\text{person}, \text{gene}),\ \forall\ \text{gene} \in \text{PharmGKB}(\text{drug})$$

This gene-to-drug aggregation is conceptually identical to the feature-engineering strategy in He et al. [6] (DGANet), where Chemical-Gene Interactions (CGIs) from CTD are aggregated at the drug level using a Jaccard-similarity matrix. In our pipeline, the aggregation is simpler and more direct — we sum raw genetic scores — but the underlying principle of linking individual-level genomic features to drug-level representations is the same. The cross-join with SIDER then yields the long-format table: (person_id, drug_name, side_effect, genetic_score).

### F. Step 6 — ADR Label Generation (75th-Percentile Threshold)

A binary ADR risk label must be assigned to enable supervised learning. The working hypothesis — that individuals carrying more damaging variants in drug-metabolizing genes face elevated ADR risk — is substantiated by multiple studies. Del Casale et al. [5] review 14 pharmacogenomics-ML papers and document that combining SNP biomarkers with clinical features using random-forest classifiers achieves reliable ADR prediction, with the ADR outcome always defined as a binary high/low-risk label. Liang et al. [2] similarly encode ADR occurrence as a discrete ordinal variable (−2 to +2) capturing the direction and magnitude of the adverse effect, confirming the feasibility of label-based ADR classification from genetic data.

For each drug d independently:

1. Collect all DrugScore(person, d) values across the population.
2. Compute the 75th percentile (P75) of this distribution (top 25% threshold).
3. Assign: ADR_label = 1 if DrugScore $\geqslant$ P75(d); else ADR_label = 0.

Table 3 illustrates label assignment for the worked example of warfarin.

Table 3: Illustrative Warfarin Genetic Scores and ADR Labels

| Person ID | CYP2C9 Score | VKORC1 Score | Warfarin DrugScore | ADR Label |
|---|---|---|---|---|
| P1 | 3 | 2 | 5 | 1 — High Risk |
| P2 | 1 | 1 | 2 | 0 — Low Risk |
| P3 | 2 | 2 | 4 | 1 — High Risk |
| P4 | 0 | 1 | 1 | 0 — Low Risk |

Scores: {5, 2, 4, 1} → P75 ≈ 4.75 → persons P1 and P3 labelled high risk.

### G. Step 7 — Final Dataset Construction and Freezing for ML

The pipeline outputs a five-column labelled CSV: (person_id, drug_name, side_effect, genetic_score, adr_label). Each row represents a unique (person, drug, side-effect) triplet. This structure directly mirrors the benchmark dataset format used by He et al. [6], who maintain a (drug, ADR) pair matrix derived from SIDER as their training target, and Seo et al. [7], who likewise model each drug-side-effect pair as a binary classification sample. The dataset is frozen at this stage to prevent any target leakage into feature engineering during downstream ML experiments.

## IV. Results and Dataset Overview

### A. Pipeline Output Statistics

Table 4 summarizes the key statistics of the pipeline output when applied to chromosome 6 data from the 1000 Genomes Project.

Table 4: Pipeline Output Statistics (Chromosome 6, 1000 Genomes Project)

| Stage | Output Description | Approximate Size |
|---|---|---|
| Step 1 — Raw VCF | Diploid genotype data, chromosome 6 | ~100 MB compressed |
| Step 2 — SNP CSV | Matrix: 2,504 persons × all chr6 SNP positions; values {0,1,2} | Millions of positions |
| Step 3 — Filtered SNPs | Variant-only positions retained after MAF filter | ~400,000 SNPs |
| Step 4 — Gene Score Matrix | 2,504 persons × 1,200+ annotated chr6 genes | ~3M cells |
| Step 5 — Drug Score Table | Long-format: person × drug × side_effect × genetic_score | Millions of rows |
| Step 6 — Labelled Dataset | Final table with adr_label column added | ML-ready CSV |
| Step 7 — Frozen Dataset | Saved for ML training/validation/testing | Final output |

### B. Sample Records From The Final Dataset

Table 5 shows a representative excerpt from the final labelled dataset, corresponding to the doxorubicin records for Person 1 (genetic_score = 118, above P75 → adr_label = 1).

Table 5: Sample Records — Person 1, Drug: Doxorubicin

| person_id | drug_name | side_effect | genetic_score | adr_label |
|---|---|---|---|---|
| 1 | doxorubicin | Gastrointestinal pain | 118 | 1 |
| 1 | doxorubicin | Abdominal pain | 118 | 1 |
| 1 | doxorubicin | Alopecia | 118 | 1 |
| 1 | doxorubicin | Anaemia | 118 | 1 |
| 1 | doxorubicin | Anaphylactic shock | 118 | 1 |
| 1 | doxorubicin | Anxiety | 118 | 1 |
| 1 | doxorubicin | Arrhythmia | 118 | 1 |

### C. Comparison with Reference Methods

Table 6 positions the proposed pipeline against the five primary reference methods

Table 6: Comparison of Proposed System with Reference Methods

| Study [Ref] | Genomic Input | Key Databases | Method | Best Performance | Gap Addressed by This Work |
|---|---|---|---|---|---|
| Liang et al. [2] (2016) | CYP2D6 & CYP1A2 alleles (PCR, 83 samples) | Clinical observation | Deep Learning GSN + Hidden Markov | > 80% classification accuracy | Works only on 2 genes, 83 samples; no VCF pipeline |
| Dafniet & Taboureau [4] (2024) | dbSNP mutations on drug-target proteins | PharmGKB, DrugBank, DrugCentral | Deep Neural Network (DNN) | BA = 0.61 on 27 organ classes | No population-scale VCF input; no SIDER integration |
| Del Casale et al. [5] (2023) | SNPs from clinical pharmacogenomics studies | PubMed / Scopus review | Random Forest (review) | AUROC up to 1.0 (individual studies) | Review only — no new dataset or pipeline |
| He et al. [6] (2025) | Gene expression (LINCS L1000) + CTD gene-drug | SIDER, CTD, PubChem, MeSH | DGANet (CNN) | AUROC = 92.76% | No raw VCF input; gene expression ≠ individual SNP profile |
| Seo et al. [7] (2020) | DisGeNET SNPs mapped to drug indications | SIDER, DrugBank, DrugBank-DDI | Random Forest + stacking | AUC = 0.9018 | SNPs used only for drug similarity, not individual scoring |
| Proposed System | 1000 Genomes VCF (2,504 persons, chr 6) | All 5 above databases combined | Full pipeline → ML-ready dataset | Complete labelled dataset generated | Only system with full VCF-to-ADR-label pipeline |

## V. Discussion

### A. Methodological Contributions

The primary contribution of this work is the integration of insights from all six reference studies into a single, executable pipeline. From Liang et al. [2], the pipeline inherits the SNP dosage encoding scheme (0/1/2) and the principle that genetic variants in metabolic enzymes are strong predictors of ADR susceptibility. From Dafniet and Taboureau [4], it inherits the use of PharmGKB as the authoritative gene-drug linking source and the practice of restricting to SNPs with frequencies above a minimum threshold to reduce noise. From Del Casale et al. [5], it derives the binary ADR label convention and the confirmation that pharmacogenomics-ML combinations are clinically meaningful even with modest sample sizes. From He et al. [6], it borrows the SIDER-based side-effect enumeration and the idea of framing the prediction problem as a drug-side-effect pair classification task. From Seo et al. [7], it draws the random forest baseline evaluation strategy and the rationale for using SNP information as a feature alongside drug

properties. From the authors' own prior work [8], the pipeline adopts the overall VCF-to-SNP-encoder architecture and the use of chromosome-6 as a pharmacogenomically rich test chromosome.

### B. Novelty of the 75th-Percentile Labelling Strategy

The percentile-based ADR labelling approach is a practical solution to the absence of ground-truth ADR outcomes for 1000 Genomes individuals, who were sequenced for genetic diversity rather than clinical outcomes. The approach is grounded in pharmacogenomics literature reviewed by Del Casale et al. [5], which consistently shows that high-risk ADR groups are characterized by genetic scores at the upper tail of the population distribution. Using the 75th percentile as the threshold designates the top 25% of the population as high-risk for each drug — a conservative cut-off that keeps the label prevalence reasonable and avoids extreme class imbalance while remaining clinically interpretable.

### C. Limitations

The study has some important limits that need to be noted. First, the pipeline was tested only on chromosome 6, but to fully understand the genetics related to drug reactions, it needs to be applied to all 22 chromosomes, including ones that have important genes like CYP2D6 on chromosome 22 and CYP2C9 and CYP2C19 on chromosome 10. Second, the labels used to mark adverse drug reactions are based on genetic scores rather than actual patient outcomes. It would be better to check these labels against real data, like the FDA Adverse Event Reporting System (FAERS), which was used by Seo et al. [7] to test their predictions. Third, the pipeline looks at each genetic change separately and doesn't consider how genes might work together, which Dafniet and Taboureau [4] suggest could be important for understanding drug reactions properly. Fourth, the pipeline hasn't included information about the chemical structure of drugs. Adding this, as suggested in [8] and shown by He et al. [6] and Seo et al. [7], could greatly improve the accuracy of the results.

### D. Future Work

The immediate next step is to extend the pipeline to all chromosomes and to train a Random Forest classifier (as used by Seo et al. [7] and recommended by Del Casale et al. [5]) on the generated dataset, reporting AUROC and AUPRC metrics for direct comparison with [6] and [7]. Subsequent work will explore deep learning architectures — specifically a convolutional or graph neural network inspired by DGANet [6] — that jointly encode per-person genetic score vectors and per-drug molecular graphs. Integration of drug molecular fingerprints from PubChem/DrugBank using the SNP encoder + drug encoder architecture proposed in [8] is the long-term architectural target. Ultimately, the system is intended as a clinical decision-support tool that flags high-risk drug-patient combinations before prescription.

## VI. Conclusion

This paper has presented a complete, reproducible, end-to-end pipeline for AI-driven analysis of drug side effects using individual DNA profiles. The seven-step system transforms raw VCF genomic files from the 1000 Genomes Project into a fully labelled machine-learning dataset by (1) extracting and encoding SNPs in a 0/1/2 dosage format [2], (2) filtering uninformative monomorphic positions [7], (3) mapping SNPs to genes via GENCODE annotation [4], (4) aggregating gene scores into drug-level genetic scores via PharmGKB [4][6], (5) integrating SIDER side-effect associations [6][7], and (6) labelling each (person, drug) pair as high or low ADR risk using a 75th-percentile population threshold [5].

The pipeline bridges a critical gap in the pharmacogenomics-AI literature: while prior studies have validated SNP-based ADR prediction with individual gene panels [2], deep neural networks on curated mutation matrices [4], CNN-based models on gene-expression data [6], and random-forest classifiers combining SNP similarity with chemical features [7], none has demonstrated a full pipeline from raw population-scale VCF files to a labelled ML dataset using publicly available databases exclusively. This work does so, establishing a transparent, replicable baseline that the community can extend.

As reviewed by Del Casale et al. [5], pharmacogenomics-ML approaches are poised to transform personalized medicine by enabling prescribers to anticipate and prevent ADRs based on a patient's genetic profile. The pipeline presented here provides the data-engineering foundation on which such clinical AI tools can be built.

## VII. Acknowledgment

We would like to express our sincere gratitude to our project guide, faculty members, and department for their continuous support, valuable guidance, and encouragement throughout the completion of this research work titled “AI-Driven Prediction of Drug Adverse Effects Using Genomic DNA Profiling.”

We are thankful to our college, ISBM College of Engineering, for providing the academic environment and resources required for carrying out this work successfully.

We also acknowledge the publicly available databases and research resources including the 1000 Genomes Project, PharmGKB, SIDER, and GENCODE, which played an important role in the development and implementation of this pipeline.

Finally, we would like to thank all the researchers and authors whose published work helped us understand the concepts of pharmacogenomics, machine learning, and adverse drug reaction prediction, forming the foundation of this study.

## References

[1] J. Lazarou, B. H. Pomeranz, and P. N. Corey, "Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies," JAMA, vol. 279, pp. 1200–1205, 1998.

[2] Z. Liang, J. X. Huang, X. Zeng, and G. Zhang, “DL-ADR: a novel deep learning model for classifying genomic variants into adverse drug reactions,” BMC Medical Genomics, vol. 9, suppl. 2, p. 48, 2016, doi: 10.1186/s12920-016-0207-4.

[3] R. Cacabelos, N. Cacabelos, and J. C. Carril, "The role of pharmacogenomics in adverse drug reactions," Expert Review of Clinical Pharmacology, vol. 12, no. 5, pp. 407–442, 2019.

[4] B. Dafniet and O. Taboureau, “Prediction of adverse drug reactions due to genetic predisposition using deep neural networks,” Molecular Informatics, vol. 43, e202400021, 2024, doi: 10.1002/minf.202400021.

[5] A. Del Casale, G. Sarli, P. Bargagna, L. Polidori, A. Alcibiade, T. Zoppi, M. Borro, G. Gentile, C. Zocchi, S. Ferracuti, R. Preissner, M. Simmaco, and M. Pompili, “Machine learning and pharmacogenomics at the time of precision psychiatry,” Current Neuropharmacology, vol. 21, pp. 2395–2408, 2023, doi: 10.2174/1570159X21666230808170123.

[6] M. He, Y. Shi, F. Han, and Y. Cai, “Prediction of adverse drug reactions based on pharmacogenomics combination features: a preliminary study,” Frontiers in Pharmacology, vol. 16, p. 1448106, 2025, doi: 10.3389/fphar.2025.1448106.

[7] S. Seo, T. Lee, M. Kim, and Y. Yoon, “Prediction of side effects using comprehensive similarity measures,” BioMed Research International, vol. 2020, Art. no. 1357630, 2020, doi: 10.1155/2020/1357630.

[8] P. Alai, S. Gadekar, P. Vedpathak, and V. Narawade, "AI-driven analysis of drug side effects using DNA profiles," International Journal for Research Trends and Innovation (IJRTI), vol. 10, no. 11, pp. a465–a468, Nov. 2025. ISSN: 2456-3315