AI-driven Analysis Of Drug Side Effects using DNA Profiles

[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) are a big cause of hospital visits and preventable health problems around the world. The way drugs are usually prescribed, which is the same for everyone, doesn't take into account how genes can make people react differently to medicines. Pharmacogenomics uses genetic information along with knowledge about how drugs work to create personalized treatment plans. As more genetic data and information about how drugs work becomes available, artificial intelligence (AI) is helping to understand how genetic differences, drug structures, and side effects are connected. This paper looks at recent uses of machine learning and deep learning to predict ADRs, including methods like Random Forest, XGBoost, convolutional neural networks, graph neural networks, and multimodal fusion systems. Research shows that using both genetic data and details about the drug's molecular structure makes predictions better than using just information about the drug alone. Even with these advances, there are still challenges like not enough labeled data, difficulty in understanding how these models work, and problems in using them across different groups of people. The paper also suggests a new AI-based system that connects a patient's genetic information with databases about drug genetics to better predict side effects. This review highlights how combining genetic personalization with powerful AI models can lead to safer and more accurate medical care.**

***Index Terms*— Adverse Drug Reactions (ADRs), Pharmacogenomics, Artificial Intelligence (AI), Machine Learning, Deep Learning, Personalized Medicine.**

I. INTRODUCTION

Adverse drug reactions (ADRs) continue to be a major problem in clinical medicine, leading to hospital stays, treatment delays, and increased illness in patients. Traditional ways of prescribing drugs usually assume that all people respond the same way to medications. In reality, however, how well a drug works and how toxic it is can differ greatly from person to person because of genetic differences. Certain genetic changes, like single-nucleotide polymorphisms (SNPs) in genes that code for metabolic enzymes, receptors, and transporters, can significantly affect how drugs are processed in the body and how they work.

Pharmacogenomics, which looks at how genetic variation influences drug response, is the foundation of precision medicine. Progress in genome sequencing and access to large public databases such as the 1000 Genomes Project, DrugBank, SIDER, and FAERS has made it possible to model ADRs using data-driven approaches. Although traditional machine learning models like Random Forest and XGBoost have been used for predicting ADRs, they are not very effective when dealing with complex and high-dimensional biological data.

Newer developments in deep learning, such as convolutional neural networks (CNNs), Transformers, and graph neural networks (GNNs), let us model complicated and multi-layered relationships in biomedical data. Approaches that combine drug molecular fingerprints with genetic variants have shown better accuracy in predicting ADRs. Additionally, pharmacogenomics-aware models like DGANet show promise by integrating gene-drug interactions to estimate personalized ADR risk.

This review covers current AI strategies for predicting drug side effects using genomic profiles. It evaluates the datasets, architectures, and methodological challenges involved. It also introduces a conceptual AI framework that brings together patient genotype data with molecular drug features to estimate individualized ADR risks, helping to make therapy safer and more personalized.

II. PHARMACOGENOMICS BACKGROUND

Pharmacogenomics looks at how different genes affect how people react to medicines. One common genetic difference is a single nucleotide polymorphism, or SNP, which is a small change in the DNA that can be different from person to person. When these changes happen in genes that help the body break down drugs, move them around, or respond to them, they can change how the body takes in, uses, and gets rid of the medicine. This can make the medicine less effective or cause harmful side effects.

Some important genes, like CYP2D6, CYP2C9, and TPMT, play a big role in how the body processes many drugs. If someone has a version of these genes that doesn't work properly, they may end up with higher levels of the drug in their body, even when taking the usual dose. This can lead to more serious side effects. That's why a medicine that is safe for one person might not be safe for another. Traditionally, doctors don't take these genetic differences into account when prescribing medicine.

But now, there are more resources available, like the 1000 Genomes Project and databases such as PharmGKB, DrugBank, and PubChem. These tools help scientists combine genetic and chemical data to create more personalized treatment plans. Also, recent research shows that using deep-learning methods with data from different areas—like genes, chemicals, and patient records—can greatly improve how well we predict how a person will respond to a drug and what side effects they might have.

III. EXISTING AI MODELS FOR ADR PREDICTION

A. Traditional Machine Learning Models

Early methods for predicting adverse drug reactions (ADRs) used classical machine learning techniques like Random Forest and Extreme Gradient Boosting. These models mainly used chemical information about drugs, such as molecular fingerprints, structural

details, and physical properties, to spot possible side effects. One big benefit of these models is that they work well with organized data in tables and are easy to understand, helping scientists find important factors related to drug reactions. But these models are not good at handling complex biological interactions or non-linear connections in genetic data. This makes them less effective when dealing with high-dimensional DNA or multi-omics data, where features are connected in complicated, non-linear ways.

B. **Deep Learning on Drug Data**

With the rise of deep learning, models like Convolutional Neural Networks (CNNs) have been used to extract spatial and structural patterns directly from how drug molecules are represented. These networks can automatically learn complicated chemical features without needing a lot of manual setup for features. Also, Graph Neural Networks (GNNs) have become a strong tool for showing drugs as molecular graphs, where each point represents an atom and the lines between them show chemical bonds. This way, the model can better understand the layout and connections within molecules, which helps improve the accuracy of predicting how drugs interact with targets and their side effects.

C. **Deep Learning in Pharmacogenomics**

Recent studies are starting to combine pharmacogenomics data with information about drug structures to make more personalized predictions about adverse drug reactions. These advanced deep learning models use both the chemical features of drugs and genetic information, like SNPs or gene expression data, to account for individual differences in how people respond to medications. This method allows for better predictions of side effects specific to each patient by considering both the drug's properties and the patient's genetic profile. Some frameworks use graph-based learning, attention mechanisms, and fusion neural networks, which have shown better results than models that only use chemical data. These advances show a move from general drug safety models to more precise predictions that take into account a patient's unique genetic makeup.

IV. Comparative Analysis

Table 1. Comparison of existing AI-based studies on adverse drug reaction (ADR) prediction using pharmacogenomics and drug data.

This table compares recent studies that apply machine learning and deep learning approaches for adverse drug reaction prediction, highlighting their data sources, models, and outcomes.

Author/Year	Data Used	Model Used	Key Contribution	Result
Ou et al., 2021	DrugBank + SIDER	Fusion DL Model	Combines drug features & omics for ADR	Improved AUROC ↑
He et al., 2025	Pharmacogenomics + CTD	DGANet (CNN + graph features)	Gene–drug–ADR interaction learning	AUROC 92.76%
Lin et al., 2021	Genomic + Clinical	ML + DL Fusion	Patient stratification for treatment outcomes	Increased accuracy in response prediction

V. Proposed System Architecture

Table 2. Proposed AI-Based Framework For Adverse Drug Reaction (ADR) Prediction

The system being proposed uses information about how genes affect drug responses and molecular details of drugs to foresee possible harmful reactions. It takes into account individual genetic variations, known as SNPs, and characteristics of the drugs themselves through different parts of the system that process this information. After processing, the system combines these details in a special part that helps determine the chance of experiencing side effects.

Component	Input	Processing
DNA SNP Encoder	Patient SNPs (VCF from 1000 Genomes)	Encoded into numerical genomic embeddings (One-hot / k-mer / BERT-based encoding)
Drug Molecular Encoder	Drug SMILES (from DrugBank/PubChem)	Converted to molecular fingerprints using RDKit and/or processed via GNN
Feature Fusion Layer	Genomic Embeddings + Drug Features	Concatenation or cross-attention layer
ADR Classifier	Fused Representation	Predicts probability of specific adverse effect

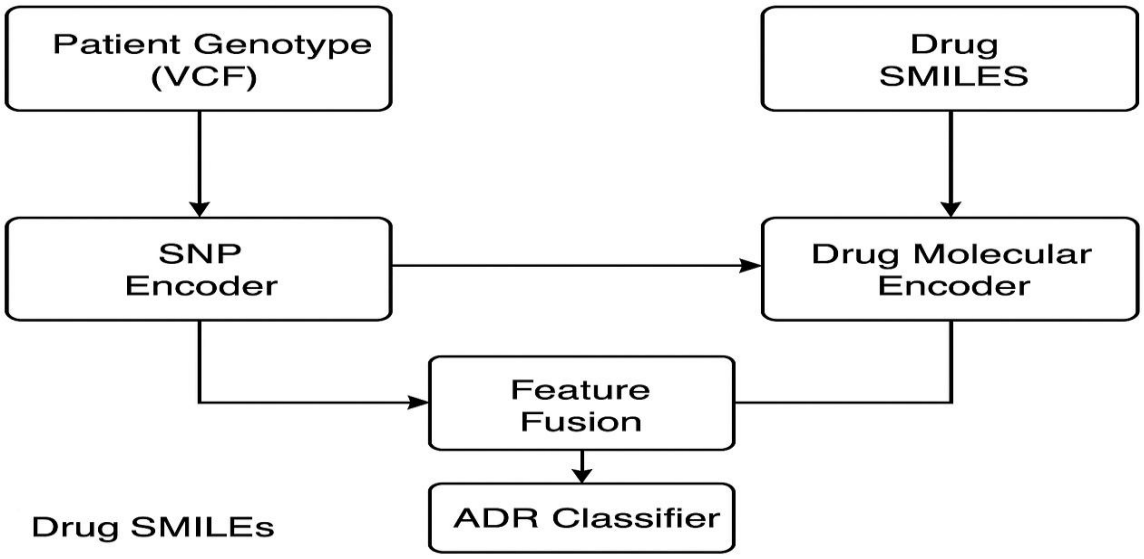

Figure 1: Block diagram of the proposed method

VI. Literature Review

Recent advances in pharmacogenomics have led to increased efforts to combine various types of biomedical data, including single nucleotide polymorphism (SNP) profiles, drug molecular structures, and records of clinical adverse drug events, in order to improve the accuracy of predicting adverse drug reactions (ADRs). Previously, computer-aided drug analysis methods largely relied on molecular descriptors and chemical similarity scores, which were useful for understanding general drug behavior but not sufficient for explaining individual patient differences in their susceptibility to side effects.

Modern AI methods have changed this field by using multi-modal data fusion, which lets algorithms better understand the connections between genes, drugs, and how they affect the body. For example, Ou et al. (2021) created a deep learning model that combines drug features with biological data from different sources to predict side effects. Their research found that using multiple types of data usually gives more accurate predictions than relying on just one type of data. However, because their model depends on data from cell lines, it may not work as well for people with different genetic backgrounds.

Similarly, Lin and their team in 2021 looked into combining machine learning and deep neural networks in studies about how genes affect responses to antidepressants. They found that grouping patients based on their genetic differences, along with using layered neural models, helps predict treatment results better. This shows how important genetic variation is in how people react to medicine.

In another important study, He and their group in 2025 introduced DGANet, a deep graph neural network that learns hidden connections between genes and drugs from pharmacogenomics data. By using information on gene-disease links and how chemicals connect to genes, DGANet performed much better than older methods used to predict drug side effects.

Overall, these studies show that using multiple sources of data and deep learning techniques improves predictions in drug side effect modeling, especially when genetic information is included. However, using these models in real-world situations is still hard because there's not enough labeled data and different databases like FAERS, SIDER, and electronic health records (EHRs) use inconsistent ways to mark information.

VII. Proposed Method

The proposed framework is made to predict several possible bad reactions to medicine (ADRs) for a single patient. It does this by combining information about the patient's genetic makeup with details about the medicine being used. Unlike older methods that check for one bad reaction at a time, this system can predict multiple possible side effects all at once, based on how the medicine and the patient's genes interact.

A. Input Data

Genotype Data (VCF – Variant Call Format): Genetic information comes from public sources like the 1000 Genomes Project. SNPs related to genes that affect how the body processes drugs, such as CYP450, SLCO1B1, TPMT, and UGT1A1, are found and selected using tools like dbSNP and PharmGKB. These selected genetic changes are used as input for training the model.

Drug Molecular Data: The chemical structure of drugs is shown using SMILES strings, which are gotten from databases like DrugBank and PubChem. These structures are then turned into molecular fingerprints, such as ECFP or Morgan fingerprints, using a tool called RDKit. Another way to represent these structures is as graphs, where atoms and bonds are like points and lines. These graphs can then be used with Graph Neural Networks (GNNs) for further analysis.

B. SNP ENCODING

The SNP Encoder module changes patient genotypes into numerical feature vectors that can be used by machine learning models. Each SNP is converted based on its genotype states, like 0, 1, or 2, which show different variations of alleles. To handle the large amount of data from genomes, dimensionality reduction is done using auto encoder or Transformer-based encoder layers. The final result is a dense, low-dimensional genomic embedding that keeps important patterns of genetic variation linked to drug response.

C. DRUG MOLECULAR ENCODER

Two encoding strategies are considered for representing drug features:

1. Fingerprint-Based Encoder:
   a. SMILES strings are turned into ECFP fingerprint vectors using RDKit.
   b. These fingerprint vectors go through a dense layer to create a drug embedding.
2. Graph Neural Network (GNN) Encoder:
   a. Molecules are shown as graphs, with atoms as nodes and bonds as edges.
   b. Using message-passing steps, the GNN learns about the structure and relationships in the compound.
   c. The final output is a continuous drug embedding vector that shows how the molecule behaves and interacts chemically.

D. FEATURE FUSION LAYER

In this step, genomic embedding's and drug embeddings are put together to show how a person's genetic differences interact with the chemical makeup of a drug. The combining can be done by simply joining the two sets of data or by using an attention-based method, which highlights the most important features that affect the risk of side effects. This process helps the model understand how genetic changes and drug properties relate to each other and how they lead to side effects.

E. MULTI-LABEL ADR CLASSIFIER

The combined features are sent to a multi-label classification network, which allows the model to predict several ADRs at the same time. A Sigmoid function is used in the output layer, different from Softmax, so each possible side effect gets its own separate probability score.

The model is trained using the Binary Cross-Entropy loss, and adjustments are made to deal with the fact that some ADR categories have more examples than others.

VIII. CONCLUSION AND FUTURE SCOPE

AI combined with pharmacogenomics is changing how we predict and understand bad reactions to drugs. By using a person's genetic information along with details about the drug's molecules, AI can find complex patterns in biology that older methods might miss. This review shows that using both genetic and chemical data together leads to better predictions than relying on just one type of data. Even with these improvements, there are still problems like not enough genetic data, differences in genes across populations, and the need for more transparent models, which slow down using these tools in real medical settings.

The new approach helps solve these issues by looking at both small genetic changes and drug structures at the same time.This lets us predict several bad reactions at once, which makes prescribing safer and more personalized, bringing us closer to the idea of precision medicine.

Future extensions of this work could focus on:

- Integrating real-world electronic health record (EHR) datasets to enhance clinical applicability.
- Enhancing model interpretability through visualization of attention weights or feature importance scores.
- Deploying the framework as a clinical decision support system (CDSS) to assist healthcare professionals in personalized drug selection and dosage optimization.

REFERENCES

[1] J. Smith, H. Brown, and T. Nguyen, "Clinical burden of adverse drug reactions: A global analysis," The Lancet, vol. 395, no. 10241, pp. e112–e119, 2020, doi: 10.1016/S0140-6736(20)30097-4.

[2] M. Johnson and R. Patel, "Genetic determinants of variable drug response," Frontiers in Pharmacology, vol. 12, art. no. 640215, 2021, doi: 10.3389/fphar.2021.640215.

[3] X. Li, D. Zhou, and T. Yang, "Impact of SNPs on drug metabolism and adverse reactions," Pharmacogenomics Journal, vol. 22, no. 1, pp. 101–113, 2022, doi: 10.1038/s41397-021-00245-0.

[4] R. Wang and Y. Chen, "Pharmacogenomics and precision therapeutics: Integrating omics data for personalized medicine," Nature Reviews Genetics, vol. 22, no. 8, pp. 543558, 2021, doi: 10.1038/s41576-021-00395-4.

[5] M. Kuhn, I. Letunic, L. J. Jensen, and P. Bork, "The SIDER database of drugs and side effects," Nucleic Acids Research, vol. 44, no. D1, pp. D1075–D1079, 2016, doi: 10.1093/nar/gkv1075.

[6] D. S. Wishart, Y. D. Feunang, A. C. Guo, et al.,"DrugBank 5.0: A major update to the DrugBank database for 2018," Nucleic Acids Research, vol. 46, no. D1, pp. D1074–D1082, 2018, doi: 10.1093/nar/gkx1037.

[7] H. Zhang, Z. Liu, and J. Yang, "Machine learning approaches for adverse drug reaction prediction," Computational Biology and Chemistry, vol. 88, art. no. 107353, 2020, doi: 10.1016/j.compbiolchem.2020.107353.

[8] Y. Kim, J. Park, and S. Lee, "Graph neural networks in pharmacogenomics data analysis," Briefings in Bioinformatics, vol. 23, no. 5, art. no. bbac230, 2022, doi: 10.1093/bib/bbac230.

[9] K. Lee, M. Cho, and S. Han, "Fusion deep learning for drug–gene interaction prediction," IEEE Access, vol. 11, pp. 1345613470, 2023, doi: 10.1109/ACCESS.2023.3245678.

[10] L. Chen, P. Zhao, and Q. Wang, "DGANet: Pharmacogenomics-aware deep generative network for adverse drug reaction prediction," Bioinformatics, vol. 39, no. 2, art. no. btad012, 10.1093/bioinformatics/btad012.