International Journal of Pharmaceutics and Drug Analysis

Content Available at www.ijpda.org ISSN: 2348:8948

ADVERSE DRUG REACTION PREDICTION USING GENERATIVE ADVERSARIAL NETWORK

SHIKSHA ALOK DUBEY[1], SONAL KANUNGO SHARMA[2] AND ASHWINI ATUL RENAVIKAR[3]

[1]Author Affiliation1 Master of Computer Application Department, Thakur Institute of Management Studies, Career Development & Research Thakur Educational Campus, Shyamnarayan Thakur Marg, Thakur Village, Kandivali (E), Mumbai – 400 101.

[2]Author Affiliation2 Master of Computer Application Department, Thakur Institute of Management Studies, Career Development & Research Thakur Educational Campus, Shyamnarayan Thakur Marg, Thakur Village, Kandivali (E), Mumbai – 400 101

[3]Author Affiliation2 Master of Computer Application Department, Thakur Institute of Management Studies, Career Development & Research Thakur Educational Campus, Shyamnarayan Thakur Marg, Thakur Village, Kandivali (E), Mumbai – 400 101

Received: 26.04.2026 Revised: 19.05.2026 Accepted: 12.06.2026

ABSTRACT

Adverse Drug Reactions (ADRs) pose significant challenges to patient safety and the effectiveness of drug therapies. This study focuses on improving drug safety by anticipating and reducing ADRs through advanced predictive modelling. By leveraging a hybrid approach that combines machine learning and deep learning techniques, along with the generation of synthetic data using Generative Adversarial Networks (GANs), the study enhances the accuracy of ADR prediction. Among the evaluated models, Deep Learning Regression demonstrated superior performance with the highest R^2 score (0.8222) and lowest RMSE (0.05411), indicating robust predictive capabilities. Polynomial Regression also showed promising results with the lowest MSE (0.0021). In contrast, Lasso and ElasticNet Regression models exhibited poor performance due to overfitting and negative R^2 values. While the results validate the efficacy of AI-based models in surpassing the limitations of traditional clinical trials and pharmacovigilance systems, further improvements in algorithm optimization and model interpretability are necessary. These advancements are essential to support the widespread adoption of AI tools in healthcare and to ensure a safer drug development lifecycle.

Keyword: *Adverse Drug Reactions (ADRs); Drug Safety; Deep Learning Regression; Data Predictive Modeling; Generative Adversarial Networks (GANs); Machine Learning.*

***CORRESPONDING AUTHOR**

Dr. Shiksha Alok Dubey

DOI: https://doi.org/10.47957/ijpda.v14i2.748

PRODUCED AND PUBLISHED BY

South Asian Academic Publications

INTRODUCTION

The process of drug development, from discovery to market, is both lengthy and expensive. Clinical trials play a crucial role in ensuring the safety and effectiveness of newly developed drugs. The primary objective of these trials is to confirm that a drug is both safe and effective for treating a specific disease before it becomes available to the general public [1]. Clinical trials are conducted in three phases, each involving a progressively larger group of participants. Phase I typically includes 1–50 participants, Phase 2 involves 100–500, and Phase 3 expands to 1,000–5,000 participants [2]. This process is illustrated in Figure 01.

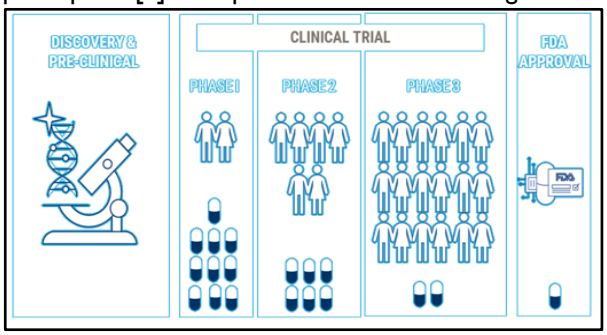

Figure 01: Demonstrating clinical trials of drugs on study patients

Since drug development is a long process and involves about 10-12 years for the development of a single drug molecule getting it approved through clinical trials is again a cumbersome issue. Therefore, there is always pressure to limit the size of the study population on whom the medicine is to be evaluated. Moreover, in this process, the sample size derived based on the population does not reflect the original set of people in terms of the genetic and physiological makeup. Hence, there are chances that adverse reactions to drugs will be experienced by the population even after clinical trials are over. However, the real goal of the pharmaceutical and healthcare sectors is to guarantee overall drug safety and reduce the incidence of adverse drug responses. In the present scenario, all drugs carry inherent risks [3]. As a result, only those whose therapeutic benefits exceed the hazards involved are authorized for use in medicine. Adverse Drug Reactions (ADRs) are any unexpected or detrimental consequences of a medication on human health. "A response to a drug that is noxious and unintended and occurs at doses normally used in man for the prophylaxis, diagnosis, or therapy of disease, or for the modification of physiological function"[4] represents an adverse drug reaction (ADR), according to the World Health Organization (WHO) [5]. ADRs are, in essence, unintended and unexpected medication reactions. ADRs are acknowledged as a major contributor to human morbidity and mortality. Studies show that adverse drug reactions (ADRs) contribute to approximately 5% of all hospital admissions and rank as the fifth leading cause of death in hospitals [6]. Due to their severe and harmful effects, numerous drugs have been withdrawn from the market. During the span of 20years, around 40 drugs were globally recalled due to serious adverse reactions [6].

The serious of the impact of these drugs are visualized in the below Figure 02.

Figure 02: Countries banning drugs based on their reactions

The most common drug-induced toxicities included in the below Figure 03.

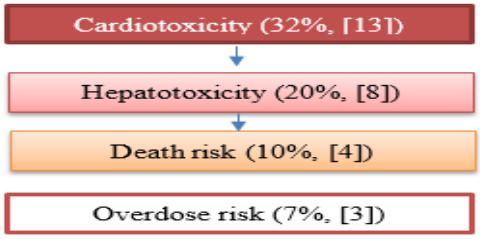

Figure 03: Drug Toxicities

A notable example is Sibutramine (Meridia), which was initially approved by the FDA as an appetite suppressant but was withdrawn in 2010 due to an increased risk of heart disease and stroke [7]. The severity of ADRs is also reflected in healthcare costs and extended hospital stays [8]. Several factors contribute to ADR development in humans [9] are grouped as follows in Figure 04.

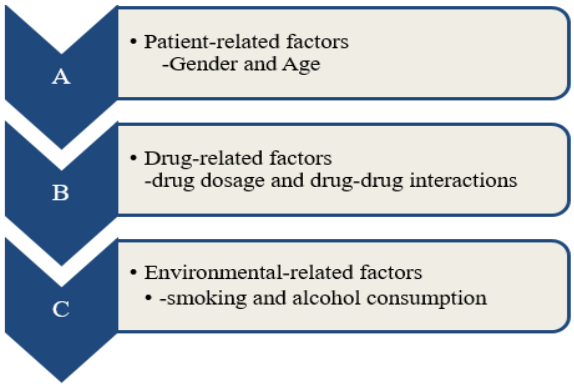

Figure 04: ADR related factors

Early detection and precise ADR prediction are crucial for improving patient safety and healthcare outcomes. Even at standard dosages, adverse drug reactions are a subset of the larger category of adverse drug events. Figure 5 is an illustration of this viewpoint.

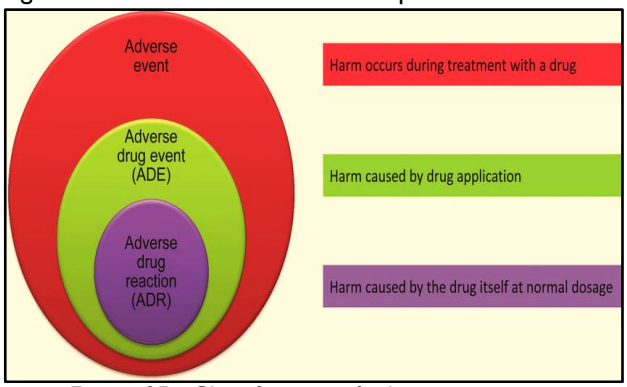

Figure 05: Classification of adverse events.

Patients and medical personnel conduct ADR reporting for adverse events that occur after clinical trials. The nation's regulatory agencies are notified of these ADR occurrences [10]. In order to take further action against the reported substance, they are analysed and kept in databases. These databases can also be used for research and review purposes. A framework has been designed based on the steps stated in the Figure 06.

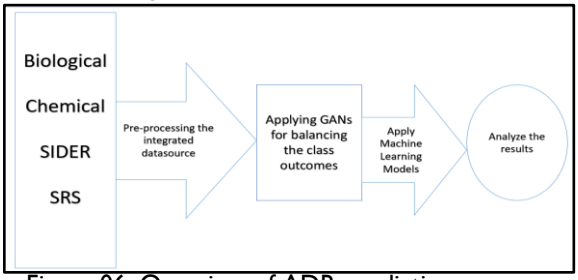

Figure 06: Overview of ADR prediction process

The block diagram shows the major steps performed for ADR prediction. Let's discuss each step in detail in the remaining part of the report.

Aim & Objectives

The basic aim of research study is to develop a model using artificial intelligence approach that will try to minimize the occurrence of severe ADR and its subsequent consequence on human health. This goal is broken down into number of sub-objectives that is stated as follows: -

To identify and predict Adverse Drug Reaction Prediction at an early stage of drug development lifecycle.

Generate synthetic data for balancing the class outcomes.

Evaluation the performance of ML models based on different performance metrics

Datasource Selection

Different data sources [11-14] selected for ADR prediction is identified and reflected in Figure 07.

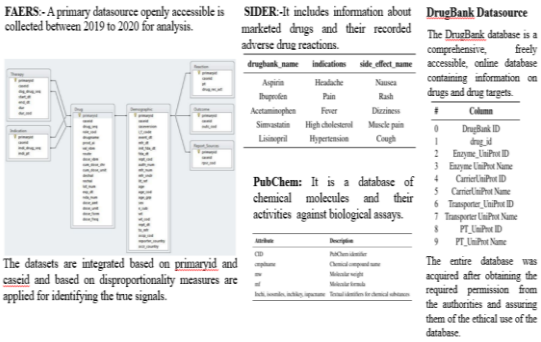

Figure 07: Selection of data sources

The selected dataset is integrated based on the identifier-based integration. The integration results are shown in the below Figure 08.

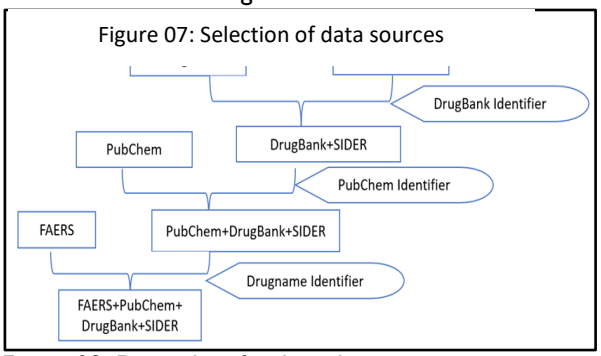

Figure 08: Drug identifier based integration

The resulting integrated dataset from both integration techniques includes a number of redundant columns, null values, and categorical feature variables that need to be pre-processed before further analysis. The steps involved in the pre-processing of identifiers integrated datasets are described in the following Figure 09.

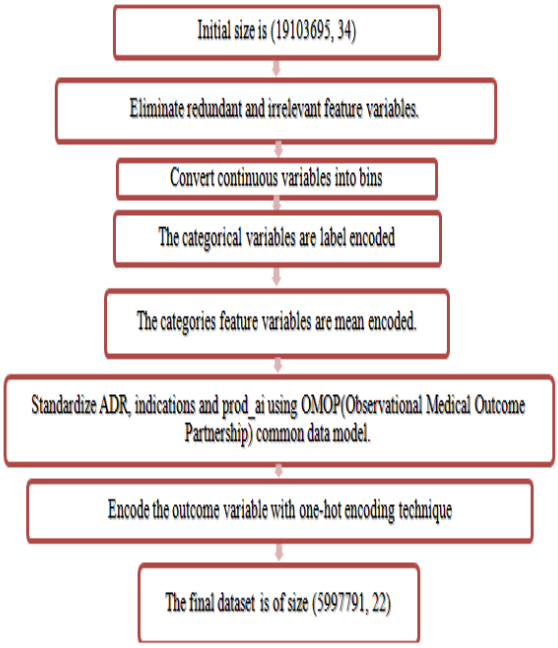

Figure 09: Pre-processing steps for integrated dataset

The integrated dataset was further reviewed with input from a domain expert, whose valuable feedback helped refine its features. Based on the expert's recommendations, some features were removed while others were added. The target type and target sequence are two of the recently added features. Four types of targets are distinguished: ion channels, receptors, enzymes, and carrier molecules. The genetic variations that a particular medicinal molecule interacts with are referred to as the target sequence. Table 01 lists the additional features that were added to the ADR dataset in addition to these.

Table 01: Feature variable description

LogP	A useful metric that influences a drug's action in the human body is lipophilicity. The compound's permeability to the body's target tissue is indicated by its Log P value [15].
LogS	The absorption and distribution properties of a chemical are greatly influenced by its water solubility. Since poor absorption usually accompanies low solubility, the general goal is to steer clear of weakly soluble substances. The unit stripped logarithm (base 10) of the solubility expressed in mol/liter is our estimated logS value [16].
CYP inhibitors	The inhibitors are responsible for delaying the action of target proteins and that lead to large amount of drug disposition in human body which is harmful and severe.
Toxicity	The toxic nature of the drug molecule on the human body.

After reviewing the dataset based on the features dropped and added to the existing dataset, the columns of the final integrated dataset specified are as follows:

```
Index(['drug_id', 'mw', 'route', 'age', 'sex', 'wt', 'occr_country',
       'role_cod', 'outc_cod', 'pt_id', 'indi_pt_id', 'prod_ai_id',
       'Type of target', 'Sequence', 'LOGP', 'Log S', 'CYP1A2 inhibitor',
       'CYP2C19 inhibitor', 'CYP2C9 inhibitor', 'CYP2D6 inhibitor',
       'CYP3A4 inhibitor', 'Hepatotoxicity'],
      dtype='object')
```

METHODOLOGY

Currently, the size of the integrated dataset is about 8284189 rows and 22 feature variables that are very huge and complex to handle for machine learning algorithms. So, as we analyze the different adverse drug reactions in the dataset and only the two most common reactions are selected for further processing. The adverse drug reaction selected for further processing is stated as follows: -

Febrile Neutropenia (development of fever)

Pulmonary Embolism (A disorder in which a blood clots blocks one of the pulmonary arteries in the lungs.)

Based on the selected adverse drug reaction the size of the new dataset is 4417120 rows and 22 columns. The dataset is prepared based on the target outcome i.e., for ADR prediction. But for ADR prediction, this dataset includes only positive ADR samples, and to apply machine learning model it should also include negative ADR samples. Therefore, we must create synthetic data samples and add them to the original dataset in order to address the problem of class imbalance. The author suggests using GANs (Generative Adversarial Networks) designs to create synthetic data samples [17]. Figure 10 below illustrates how GANs are used to create synthetic data.

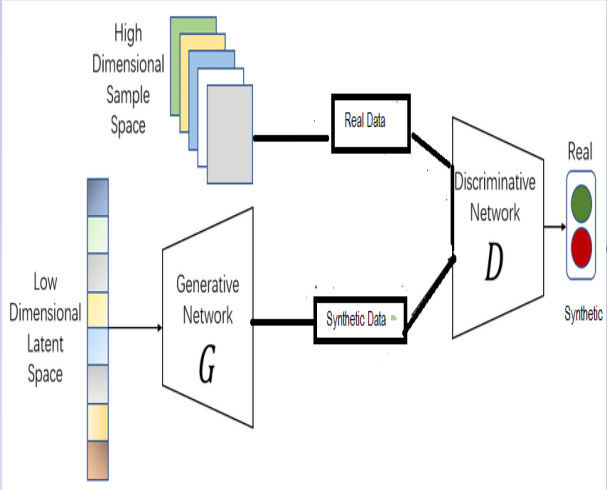

Figure 10: Generative Adversarial Network for synthetic data creation

Ten lakh records in all were produced for ADRs presence and absence. Additionally, ADR prediction techniques were used to this dataset. Before using machine learning models, the datasets are examined and shown from various angles in subsequent sections. Data exploration is performed to analyze the dataset from different perspectives and get greater insights in the data for further processing.

The description of the resultant dataset is shown in Table 02:

Table 02: ADR prediction dataset description

Statistical Summary

	Type of target	Sequence	LogP	Log S	route	age	sex	CYP2D6 inhibitor	Hepatotoxicity	ind_id_pk	pt_id
count	1000000.000000	1000000.000000	1000000.000000	1000000.000000	1000000.000000	1000000.000000	1000000.000000	1000000.000000	1000000.000000	1000000.000000	1000000.000000
mean	0.836895	0.955941	0.592996	0.885900	0.282530	0.962699	0.772491	-0.006668	0.390452	0.250288	0.852861
std	0.115827	0.063190	0.062786	0.133856	0.436534	0.072080	0.089686	0.040303	0.112389	0.427120	0.222544
min	0.546189	0.509011	0.367807	0.471215	-0.037324	0.484561	0.428713	-0.087940	0.231427	-0.213199	0.421488
25%	0.742906	0.933504	0.568465	0.713377	0.002985	0.915885	0.719729	-0.011919	0.289911	-0.026959	0.524055
50%	0.800974	0.953861	0.591627	0.932002	0.004256	0.937020	0.784673	-0.006641	0.414485	-0.027562	0.955212
75%	0.967171	0.986061	0.611250	0.979712	0.917882	0.981033	0.816898	-0.006095	0.452769	0.873800	1.033786
max	2.790111	3.239545	3.675369	2.369404	3.492590	1.701662	2.728523	0.005534	3.905976	2.663648	1.390002

The dataset has been described in terms of multiple perspectives. The count denotes the total rows in the dataset while mean and standard deviation are the average value and variance in each column. The min and max signifies the distribution of data in various quartiles.

Correlation Matrix

The correlation matrix demonstrates the relationship between feature variables [18]. It can be direct or indirect or neutral. For ADR prediction, the correlation matrix plotted is shown in Figure 11.

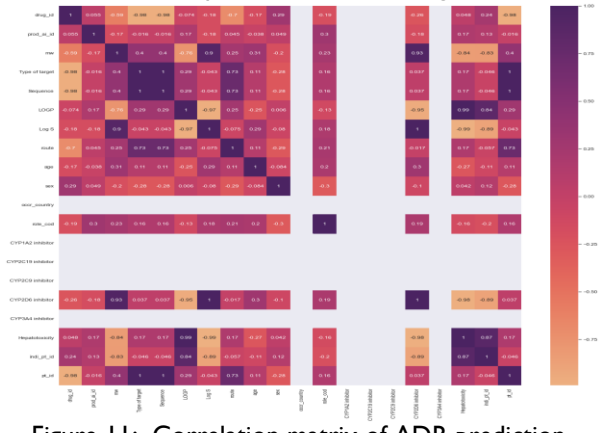

Figure 11: Correlation matrix of ADR prediction dataset

The correlation matrix above depicts the relationship of feature variables with respect to target variable. The matrix shows zero correlation with respect to occurrence country, CYP inhibitor except CYP2D6. Drug, type of target, sequence and route has strong

correlation with respect to target outcome as compared to other feature variables.

Feature Scaling and Selection

The dataset has feature variables with different ranges. In binary classification model if the features are not scaled there is possibility of biasness towards feature variables with similar ranges as compared to those feature variables with different range. An overview of feature variables in the dataset with different ranges in Table 03.

Table 03: Data frame showing different feature variables with different ranges.

prod_ai	mw	Type of target	Sequence	LOGP	Log S	route	age	sex	...	role_cod	CYP1A2 inhibitor	CYP2C19 inhibitor	CYP2C9 inhibitor	CYP2D6 inhibitor	CYP3A4 inhibitor	Hepatotoxicity
45040773.0	252.27	0.01105	0.01105	2.26	-3.55	0.0000	4	1	...	0.043222	0	0	0	0	0	0
45315416.0	252.27	0.01105	0.01105	2.26	-3.55	0.1862	9	1	...	0.043222	0	0	0	0	0	0
45361669.0	252.27	0.01105	0.01105	2.26	-3.55	0.1862	9	1	...	0.179673	0	0	0	0	0	0
45315416.0	252.27	0.01105	0.01105	2.26	-3.55	0.1862	9	1	...	0.043222	0	0	0	0	0	0
45361669.0	252.27	0.01105	0.01105	2.26	-3.55	0.1862	9	1	...	0.179673	0	0	0	0	0	0

Feature scaling is required because the medicine and product active ingredient feature in the previous table has a larger range of values than the other feature variables. One technique for normalizing the range of independent variables or data characteristics is feature scaling [19]. The following is a discussion of the feature scaling technique.

MINMAX Scaler

The MinMaxscaler is one of the simplest scalars to understand. It scales values between zero and one[20]. The formula used for scaling independent feature variables in the dataset is stated as follows

$$x\_scaled = (x - x\_min)/(x\_max - x\_min)$$

---formula for min-max scaler

The following QQ plot in Figure 12 shows the outcomes of applying min-max scaling to the provided dataset.

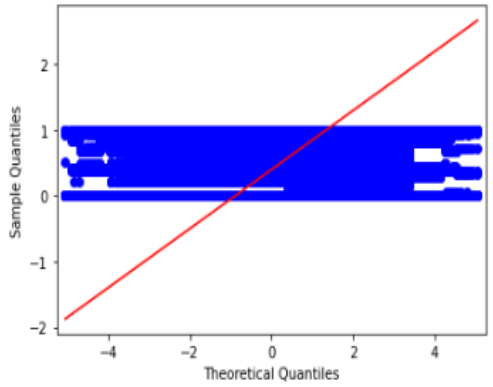

Figure 12: QQ plot showing the results of min-max scaling.

The result clearly shows that the data is scaled between zero and one. In our case our dataset does not have Gaussian distribution therefore we cannot apply standard scaling technique to it.

So the best scaling technique which does not considers the distribution of the dataset is Min-Max scaling technique. Therefore, on both the dataset we have proceeded with the application of Min-max scaling technique and showed the results in the following Table 04.

Table 04: Results derived after application of Min-Max scaler technique

mw	Type of target	Sequence	LOGP	Log S	route	age	sex	...	occr_country	role_cod	CYP1A2 inhibitor	CYP2C19 inhibitor	CYP2C9 inhibitor	CYP2D6 inhibitor	CYP3A4 inhibitor	Hepatotoxicity	indi_pt_
0.0	0.0	0.0	1.0	0.0	0.863461	0.000000	1.0	...	0.0	0.238823	0.0	0.0	0.0	0.0	0.0	1.0	0.5056
0.0	0.0	0.0	1.0	0.0	1.000000	0.833333	1.0	...	0.0	0.238823	0.0	0.0	0.0	0.0	0.0	1.0	0.9685
0.0	0.0	0.0	1.0	0.0	1.000000	0.833333	1.0	...	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	1.0	0.9685
0.0	0.0	0.0	1.0	0.0	1.000000	0.833333	1.0	...	0.0	0.238823	0.0	0.0	0.0	0.0	0.0	1.0	0.9685
0.0	0.0	0.0	1.0	0.0	1.000000	0.833333	1.0	...	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	1.0	0.9685

Feature Selection

There are several low variance feature variables in the current dataset that won't have a significant impact on the desired result. The model's performance will increase if these feature variables are removed. There are three general groups into which feature selection strategies can be divided [21] in the below Figure 13.

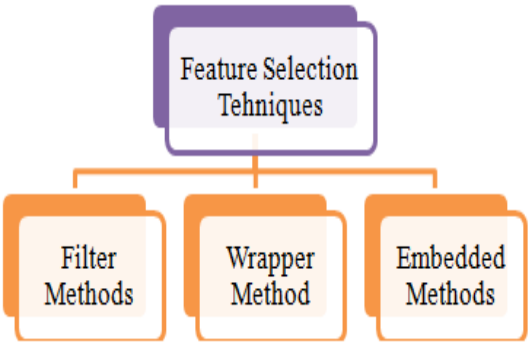

Figure 13: Classification of feature selection methods

Each technique is discussed in detail in the following section.

Filter Methods: These methods analyse the intrinsic properties of features and use statistical techniques for selection. They are computationally efficient and faster compared to other methods.

Wrapper Methods: These methods identify different subsets of feature variables by evaluating their performance within a classifier model. They employ a greedy search strategy to determine the most relevant subset of features based on their relationship with the target variable.

Embedded Methods: These methods follow an iterative process, selecting the most relevant subset of features at each step. This approach ensures that all selected features contribute significantly to the outcome variable. The Extremely Randomized Trees Classifier

(Extra Trees Classifier) [21] is an ensemble learning technique that aggregates the results of multiple de-correlated decision trees within a "forest" to produce its classification output.

The embedded method was chosen to identify the most important features in the given dataset because it combines the advantages of both filter and wrapper methods by considering feature interactions while maintaining a reasonable computational cost. The following results were obtained by applying this technique to the dataset for ADR prediction in Figure 14.

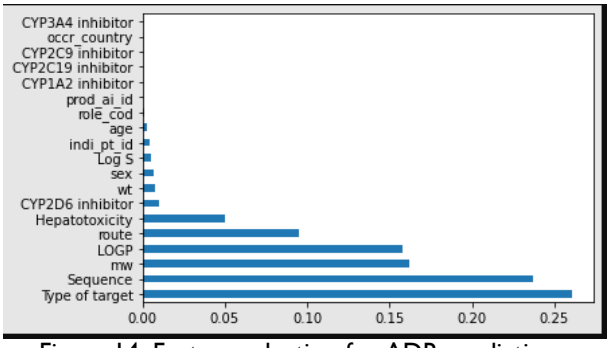

Figure 14: Feature selection for ADR prediction

In the above Figure 14 of feature selection process signifies that target type has the highest selection measure with the ADR output followed by sequence, mw, logP, hepatotoxicity, drug, and route. The remaining features have low feature selection value and no value at all. These features are eliminated in the further process.

RESULTS

Outcome of Gan in the Creation of ADR Prediction Dataset.

As previously discussed in Section III, GANs are responsible for generating synthetic dataset using real time dataset feature variables. This entire process can be summarized as follows:

Step 1: Generate random noise in the latent space and reshape it to match the input dimensions of the generator model.

Step 2: Define the `generate_fake_samples` function to produce synthetic data. The generator takes latent points (random noise) as input and outputs a NumPy array of fake samples. Since these are artificially generated, their label is set to **0**.

Step 3: Create another function to generate real samples by randomly selecting data from the actual dataset. The label for real samples is 1.

Step 4: Build a **generator model** using a simple sequential architecture with the Keras module. The input dimension matches the shape of the input samples, and the kernel is initialized using `'he_uniform'`. The model consists of three layers: two hidden layers activated by the `'relu'` function and an output layer activated by the `'linear'` function. The output layer's dimension matches that of the dataset.

Step 5: Define the **discriminator model**, another sequential model with three dense layers. The first two layers use the 'relu' activation function, while the output layer applies the 'sigmoid' function to distinguish between real (True) and fake (False) samples.

Step 6: Finally, create the **GAN model**, a sequential model that integrates both the generator and discriminator.

The above stated steps can be visualized in the following Figure15.

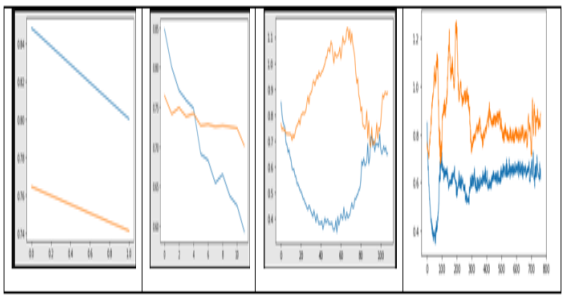

Figure 15: The training process of GAN model

While the discriminator and generator are being trained. We will compute the average loss for each epoch after combining half batches of actual and fake data. The train_on_batch method will be used to update the combined model. For later use, the trained generator will be stored. The following Figure 16 illustrates the average loss.

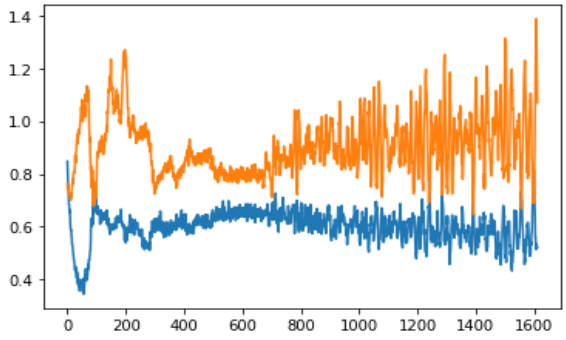

Figure 16: The Loss of generator and discriminator changes are plotted as followed: blue-loss of discriminator; orange-loss of generator

ADR Prediction Results

The results based on different prediction models are discussed and shown in the following Table 5 & 6.

Table 05: Machine Learning Algorithm Description

Model	Description
Linear Regression	Predicts the target variable based on a linear relationship with input features.
Ridge Regression	By reducing the coefficients, linear regression with L2 regularization avoids overfitting.
Lasso Regression	Linear regression with L1 regularization that can shrink some coefficients to zero, performing feature selection.
Polynomial Regression	Extends linear regression by adding polynomial terms, allowing it to model non-linear relationships.
ElasticNet Regression	Combines L1 and L2 regularization (Ridge + Lasso), balancing feature selection and regularization.
Bayesian Ridge Regression	Applies Bayesian methods to linear regression, estimating the distribution of coefficients rather than single values.

Table 06: ADR dataset results

Sr. No	Algorithm Name	MSE	RMSE	R2 Squared
1	Linear Regression	0.006	0.077	0.68
2	Ridge Regression	0.0064	0.08	0.6688
3	Lasso Regression	0.058	0.2428	-8.97E-06
4	Polynomial Regression	0.0021	0.0467	0.807
5	ElasticNet Regression	0.05	0.242	-8.97E-06
6	BayesianRidge Regression	0.006	0.077	0.6804
7	Huber Regressor	0.007	0.089	6.31E-01
8	Deep Learning Regression	0.0411	0.05411	0.8222

The results can be diagrammatically visualized in Figure 17.

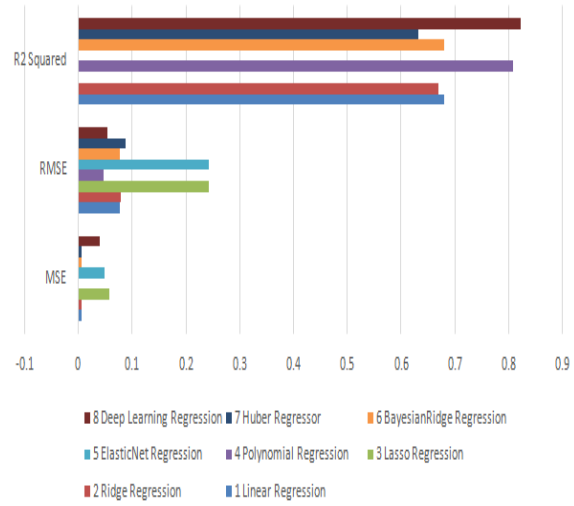

Figure 17: Comparison of various evaluation metrics applied by different prediction algorithms

Deep Learning Regression outperformed all other models with the highest R^2 (0.8222) and lowest RMSE (0.05411), indicating strong predictive accuracy and generalization. Polynomial Regression showed the lowest MSE (0.0021) and a high R^2 (0.807), making it another strong performer. Linear, BayesianRidge, and Huber Regressors delivered moderate performance with R^2 values around 0.68. Lasso and ElasticNet Regression performed poorly with negative R^2 values and high error rates, indicating underperformance or overfitting.

CONCLUSION

This study emphasizes the critical importance of predicting and minimizing adverse drug reactions (ADRs) to enhance drug safety. By integrating advanced machine learning and deep learning techniques supported by synthetic data generation using GANs the study demonstrates a more accurate and reliable forecast of ADR occurrences. The findings showcase the potential of AI-driven models in overcoming the limitations of traditional clinical trials and pharmacovigilance systems.

Among the models evaluated, Deep Learning Regression achieved the highest predictive performance with an R^2 of 0.8222 and the lowest RMSE of 0.05411, indicating superior accuracy and generalization. Polynomial Regression also performed well with the lowest MSE (0.0021) and a strong R^2 of 0.807. In contrast, Lasso and ElasticNet Regression underperformed, showing negative R^2 values and high errors. While these advancements demonstrate significant potential in enhancing drug safety, further improvements in model optimization and interpretability are necessary. These refinements will be crucial for broader adoption across healthcare environments and for ensuring a safer and more efficient drug development lifecycle.

CONFLICT OF INTEREST

The authors have no conflicts of interest regarding this investigation.

ACKNOWLEDGMENTS

In this section, I would like to extend my heartfelt gratitude to my guide Late Dr. Anala A. Pandit for her immense support and vision throughout my research work. She has been a pillar of support and confidence in my research journey.

Funding

NA

CONFLICT OF INTEREST

Not Applicable

INFORMED CONSENT

Open-source data source is considered

ETHICAL STATEMENT

NA

AUTHOR CONTRIBUTION

Idea Conceptualization and Model creation

REFERENCES

1. Patrick DL, Burke LB, Powers JH, Scott JA, Rock EP, Dawisha S, et al. Title of article. **Value Health.** 2007;10(Suppl 2):S125-S137. doi:10.1111/j.1524-4733.2007.00275.x.
2. Pushparajah DS, Geissler J, Westergaard N. EUPATI: Collaboration between patients, academia, and industry to champion the informed patient in the research and development of

medicines. **J Med Dev Sci.** 2016;1(1):74-80. doi:10.18063/jmds.v1i1.122.
3. Hazell L, Shakir SA. Under-reporting of adverse drug reactions. **Drug Saf.** 2006;29(5):385-396. doi:10.2165/00002018-200629050-00003.
4. ICH Expert Working Group. Post-approval safety data management: Definitions and standards for expedited reporting. ICH Harmonised Tripartite Guideline. 2003. Available from: http://www.fda.gov/cber/gdlns/ichexrep.htm
5. Lazarou J, Pomeranz BH, Corey PN. Incidence of adverse drug reactions in hospitalized patients: A meta-analysis of prospective studies. **JAMA.** 1998;279(15):1200-1205. doi:10.1001/jama.279.15.1200.
6. Fung M, Thornton A, Mybeck K, Wu JHH, Hornbuckle K, Muniz E. Evaluation of the characteristics of safety withdrawal of prescription drugs from worldwide pharmaceutical markets: 1960 to 1999. **Ther Innov Regul Sci.** 2001;35(1):293-317. doi:10.1177/009286150103500115.
7. Wysowski DK, Swartz L. Adverse drug event surveillance and drug withdrawals in the United States, 1969-2002: The importance of reporting suspected reactions. **Arch Intern Med.** 2005;165(12):1363-1369. doi:10.1001/archinte.165.12.1363.
8. Mulchandani R, Kakkar AK. Reporting of adverse drug reactions in India: A review of the current scenario, obstacles, and possible solutions. **Int J Risk Saf Med.** 2019;30(1):33-44. doi:10.3233/JRS-180228.
9. Alomar M. Factors affecting the development of adverse drug reactions. **Saudi Pharm J.** 2014;22(2):83-94. doi:10.1016/j.jsps.2013.02.003.
10. Biswas P. Pharmacovigilance in Asia. **J Pharmacol Pharmacother.** 2013;4(Suppl 1):S7-S19. doi:10.4103/0976-500X.120957.
11. Dal Pan GJ, Arlett PR. The US Food and Drug Administration-European Medicines Agency collaboration in pharmacovigilance: Common objectives and common challenges. **Drug Saf.** 2015;38:13-15. doi:10.1007/s40264-014-0259-4.
12. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. **Nucleic Acids Res.** 2016;44(D1):D1075-D1079. doi:10.1093/nar/gkv1075.
13. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, et al. DrugBank: A knowledgebase for drugs, drug actions, and drug targets. **Nucleic Acids Res.** 2008;36(Database issue):D901-D906. doi:10.1093/nar/gkm958.
14. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH. PubChem: A public information system for analyzing bioactivities of small molecules. **Nucleic Acids Res.** 2009;37(Suppl 2):W623-W633. doi:10.1093/nar/gkp456.
15. Van De Waterbeemd H, Smith DA, Beaumont K, Walker DK. Property-based design: Optimization of drug absorption and pharmacokinetics. **J Med Chem.** 2001;44(9):1313-1333. doi:10.1021/jm0100883.
16. Yu J, Li N, Lin P, Li Y, Mao X, Bao G, et al. Intestinal transportations of main chemical compositions of *Polygoni multiflori* radix in the Caco-2 cell model. **Evid Based Complement Alternat Med.** 2014;2014:483641. doi:10.1155/2014/483641.
17. Park N, Mohammadi M, Gorde K, Jajodia S, Park H, Kim Y. Data synthesis based on generative adversarial networks. **arXiv** [Preprint]. 2018. doi:10.48550/arXiv.1806.03384.
18. Alin A. Multicollinearity. **Wiley Interdiscip Rev Comput Stat.** 2010;2(3):370-374. doi:10.1002/wics.84.
19. Singh D, Singh B. Feature-wise normalization: An effective way of normalizing data. **Pattern Recognit.** 2022;122:108307. doi:10.1016/j.patcog.2021.108307.
20. Saeed VA, Ahmed NS, Sadiq BH. Comparative analysis of preprocessing techniques for KNN classification on the diabetes dataset. In: **Proceedings of the International Conference on Innovations in Computing Research**. Cham: Springer; 2024. p. 213-221.
21. Stańczyk U. Feature evaluation by filter, wrapper, and embedded approaches. In: **Feature Selection for Data and Pattern Recognition**. Cham: Springer; 2015. p. 29-44. doi:10.1007/978-3-319-10221-8_3.