RESEARCH PAPER

A Multi-Source Deep Learning Framework for Drug-Side Effect Prediction Using Knowledge Integration

Mr. Digvijay Anandrao Patil[1], Dr. Jaydeep B. Patil[2], Dr. Shrikant D. Bhopale[3], Dr. Sangram T. Patil[4]

[1]Research Scholar, D Y Patil Agriculture & Technical University Talsande, Kolhapur, India | Email: digvi.311088@gmail.com

[2]D Y Patil Agriculture & Technical University Talsande, Kolhapur, India | Email: er.jaydeep7576@gmail.com, jaydeep.patil@dyp-atu.edu.in

[3]D Y Patil Agriculture & Technical University Talsande, Kolhapur, India | Email: shrikant.bhopale@dyp-atu.edu.in

[4]D Y Patil Agriculture & Technical University Talsande, Kolhapur, India | Email: sangrampatil@dyp-atu.org

ABSTRACT

Adverse Drug Reactions (ADRs) represent a significant challenge in modern healthcare, contributing to increased morbidity, mortality, and healthcare costs worldwide. Traditional methods for identifying drug side-effects, including clinical trials and spontaneous reporting systems, often suffer from limitations such as delayed detection and underreporting [8]. With the growing availability of biomedical and pharmacovigilance data, there is an increasing need for computational frameworks that integrate heterogeneous data sources for accurate ADR prediction.

In this work, we propose a multi-source data-driven framework that integrates curated biomedical knowledge with real-world pharmacovigilance evidence. Specifically, we combine structured drug–side-effect associations from SIDER [1] with large-scale post-marketing reports from FAERS [2], [9] to construct a unified multi-label dataset. A scalable preprocessing pipeline is developed to normalize drug entities, standardize adverse events using MedDRA terminology, and aggregate signals via frequency-based thresholding.

To demonstrate the utility of the dataset, a baseline deep learning model is implemented for multi-label ADR prediction [10], [15]. Experimental observations indicate that integrating real-world FAERS data with curated SIDER knowledge improves robustness compared to single-source approaches. The proposed framework provides a foundation for safety-aware drug analytics and future research in explainable AI and personalized medicine.

Keywords— component, formatting, style, styling, insert (key words).

How to cite this article: Patil DA, Patil JB, Bhopale SD, Patil ST. A Multi-Source Deep Learning Framework for Drug-Side Effect Prediction Using Knowledge Integration. Int J Drug Deliv Technol. 2026;16(55s): 260-266. DOI: 10.25258/ijddt.16.55s.29

Source of support: Nil.

Conflict of interest: None.

I. INTRODUCTION (*HEADING 1*)

This Adverse Drug Reactions (ADRs) remain a critical concern in global healthcare systems, contributing significantly to patient morbidity and mortality. Studies indicate that ADRs are among the leading causes of hospitalization and death, emphasizing the need for effective detection mechanisms [8]. Despite rigorous pre-market evaluations, many adverse effects are identified only after widespread clinical use due to limitations in clinical trials such as restricted sample sizes and controlled environments.

Pharmacovigilance systems such as the FDA Adverse Event Reporting System (FAERS) [2] play a vital role in monitoring drug safety in real-world settings. FAERS collects large-scale adverse event reports submitted by healthcare professionals and patients, enabling post-marketing surveillance. However, FAERS data is inherently noisy, suffering from reporting bias, duplication, and inconsistent terminology [9], [18]. These challenges necessitate robustpreprocessing techniques before such data can be used for predictive modelling.

In contrast, curated biomedical resources such as the Side Effect Resource (SIDER) [1] provide structured and high-quality drug–side-effect associations derived from clinical documentation. SIDER ensures standardized representation using MedDRA terminology, making it reliable for supervised learning tasks. However, it lacks real-time updates and cannot capture emerging or rare adverse effects observed in real-world populations.

Recent advances in machine learning and deep learning have enabled significant progress in ADR prediction. Traditional approaches such as logistic regression and support vector machines rely heavily on handcrafted features and often fail to capture complex nonlinear relationships [12]. Deep learning methods, including convolutional and recurrent neural networks, have improved predictive performance by automatically learning feature representations [13], [14]. Transformer-based models such as BEHRT further enhance the modeling of temporal clinical data [7].

Graph-based approaches have also gained attention, particularly for modeling drug–drug interactions and polypharmacy effects. The Decagon model utilizes graph convolutional networks to capture multi-relational dependencies among drugs and side-effects [3]. Similarly, graph-based recommendation systems such as GAMENet and SafeDrug incorporate safety constraints into medication recommendation tasks [4], [5]. While these models demonstrate strong performance, they often require complex graph construction and are computationally intensive.

A key limitation of existing approaches is their reliance on single-source data, which restricts their ability to generalize across diverse clinical scenarios. Integrating multiple data sources, such as curated knowledge and real-world evidence, offers a more comprehensive representation of drug safety. However, such integration poses challenges in terms of data heterogeneity, normalization, and scalability.

In this work, we address these challenges by proposing a unified framework that integrates SIDER [1] and FAERS [2] data into a consistent multi-label dataset. A scalable preprocessing pipeline is developed to standardize drug names, align side-effect terminology, and aggregate adverse event signals using threshold-based filtering. The resulting dataset enables the application of deep learning models for ADR prediction.

To validate the effectiveness of the proposed dataset, a baseline deep learning model is implemented for multi-label classification [10], [15]. The model predicts multiple side-effects per drug, reflecting real-world clinical scenarios. Furthermore, the integration of FAERS data enables real-world validation, improving the practical applicability of the framework. The main contributions of this work are summarized as follows-

- Integration of curated (SIDER) and real-world (FAERS) datasets for ADR prediction [1], [2]
- Development of a scalable preprocessing pipeline for pharmacovigilance data [9]
- Formulation of ADR prediction as a multi-label deep learning problem [15]
- Implementation of a baseline neural network model for evaluation [10]
- Establishment of a foundation for future research in explainable and personalized drug safety

II. RELATED WORK

A. Traditional Machine Learning Approaches

Early research in ADR prediction relied on statistical and classical machine learning models such as logistic regression, support vector machines, and random forests. These approaches typically use handcrafted features derived from chemical properties, drug similarity, or therapeutic categories [12]. While these models are computationally efficient and interpretable, they struggle to capture complex nonlinear relationships and fail to scale effectively with large biomedical datasets.

Additionally, traditional approaches often treat ADR prediction as independent binary classification problems, ignoring correlations between side-effects. This limitation reduces their effectiveness in real-world multi-label scenarios where drugs may exhibit multiple adverse reactions simultaneously

B. Deep Learning-Based Approaches

Deep learning has significantly improved ADR prediction by enabling automatic feature extraction and modeling of complex relationships. Convolutional neural networks (CNNs) have been used to learn representations from molecular structures, while recurrent neural networks (RNNs) and LSTMs have been applied to electronic health records (EHRs) for patient-specific predictions [13], [14].

Transformer-based models such as BEHRT further enhance the modeling of longitudinal clinical data by capturing temporal dependencies [7]. These models demonstrate strong performance; however, they often rely on structured and high-quality input data, which is not always available in pharmacovigilance datasets

C. Graph-Based Approaches

Graph-based models have emerged as powerful tools for ADR prediction by representing drugs, proteins, and side-effects as nodes in a network. The Decagon model uses graph convolutional networks (GCNs) to predict polypharmacy side-effects and capture complex multi-relational interactions [3]. Similarly, relational GCNs and heterogeneous graph models extend this framework to integrate multiple biomedical entities [16], [17].

Although graph-based models offer high predictive power, they require complex graph construction and arecomputationally intensive. Additionally, their scalability remains a challenge when dealing with large real-world datasets.

D. Pharmacovigilance-Based Approaches

Pharmacovigilance methods focus on detecting safety signals from real-world data sources such as FAERS. Statistical techniques like the Proportional Reporting Ratio (PRR) and Reporting Odds Ratio (ROR) are widely used for signal detection [8], [19]. These methods identify statistically significant drug–event associations but are limited in predictive capability.

Recent studies have explored applying machine learning techniques to FAERS data, transforming it into structured formats for prediction tasks [9], [18]. However, the inherent noise, reporting bias, and inconsistency in FAERS data remain significant challenges.

E. Multi-Source Integration Approaches

To address the limitations of single-source models, recent research has focused on integrating multiple datasets. Combining curated resources like SIDER [1] with real-world data such as FAERS [2] provides complementary information that enhances prediction robustness.

Multi-source approaches aim to leverage:

- High-quality curated knowledge (SIDER)
- Large-scale real-world evidence (FAERS)

However, challenges include:

- Entity normalization
- Terminology alignment
- Handling noisy data
- Lack of standardized pipelines

F. Research Gap

From the above discussion, the following gaps are identified:

- Over-reliance on single-source datasets
- Lack of real-world
- validation using pharmacovigilance data
- Limited focus on scalable preprocessing pipelines
- Poor reproducibility due to inconsistent workflows
- Inadequate handling of multi-label ADR prediction

G. Summary Table

TABLE I. COMPARATIVE ANALYSIS OF ADR PREDICTION APPROACHES

Category	*Method*	*Data Source*	*Strengths*	*Limitations*
Traditional ML	*SVM, RF, Logistic Regressi on [12]*	*SIDER / Drug features*	*Simple, interpretab le*	*Cannot capture complex patterns*
Deep Learning	CNN, RNN [13], [14]	Molecul ar / EHR	Automatic feature learning	Needs large structured data
Transformer Models	BEHRT [7]	Clinical sequenc es	Captures temporal patterns	Complex, data-intensive
Graph-Based Models	Decagon [3], R-GCN [17]	Drug network s	Models complex relations	High computatio nal cost

Category	*Method*	*Data Source*	*Strengths*	*Limitations*
Pharmacovigilance	PRR, ROR [8], [19]	FAERS	Real-world insights	Not predictive
ML on FAERS	Neural models [9], [18]	FAERS	Uses real-world data	Noisy, inconsistent
Multi-Source Models	Hybrid approaches [1], [2]	SIDER + FAERS	Improved robustness	Lack of standardized pipeline

The comparative analysis highlights that no single approach effectively addresses all aspects of ADR prediction. Traditional methods lack expressiveness, deep learning models often depend on structured data, and graph-based models introduce computational complexity. Pharmacovigilance approaches provide real-world insights but lack predictive capability.

Multi-source approaches show the most promise by combining complementary datasets; however, they suffer from a lack of standardized and reproducible pipelines. This gap motivates the proposed framework, which integrates SIDER and FAERS data using a scalable preprocessing pipeline and formulates ADR prediction as a multi-label learning problem.

III. Methodology

The proposed framework integrates curated biomedical knowledge and real-world pharmacovigilance data into a unified pipeline for drug side-effect prediction. The system is designed to handle heterogeneous data sources and transform them into a structured format suitable for deep learning. Below Fig.1 shows the system pipeline utilizing two complementary datasets.

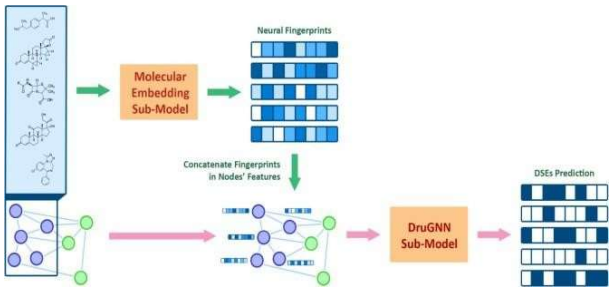

Fig. 1. Drug side effect prediction with Deep Learning Molwcular embedding in a multigraph domain

A. Data Sources

Framework utilizes two complementary datasets:

- SIDER: Provides curated drug–side-effect associations [1]
- FAERS: Provides real-world adverse event reports [2], [9]

These datasets differ in structure, scale, and reliability, requiring careful preprocessing and integration..

B. Data Preprocessing Pipeline

a) Data Merging

FAERS data consists of multiple tables, including drug and reaction tables. These are merged using a common identifier:

$$D = \{(drug_i, reaction_j)\}$$

This step establishes the relationship between drugs and reported adverse events.

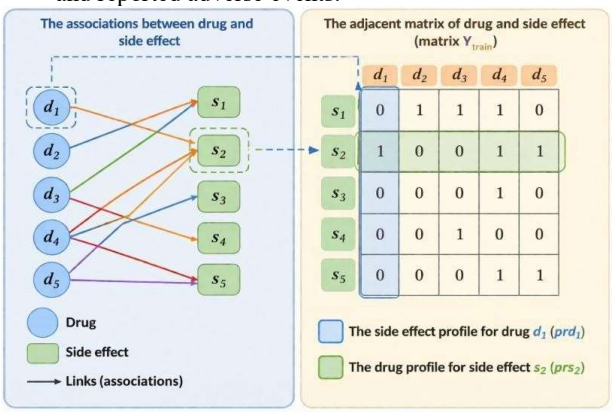

Fig. 2. Drug side effect prediction with Deep Learning Molecular embedding in a multigraph domain

Drug side effect association and matrix infographic

b) Normalization

To ensure consistency:

- Drug names are converted to lowercase and standardized
- Side-effects are mapped to MedDRA terminology

c) Aggregation

The frequency of each drug–side-effect pair is computed:

$$Count(d, s) = \sum_{k=1}^{N} \mathbb{I}(report_k = (d, s))$$

where II is an indicator function

d) Threshold-Based Labeling

To reduce noise, a threshold $\tau$is applied:

$$Label(d, s) = \begin{cases} 1, & \text{if } Count(d, s) \geq \tau \\ 0, & \text{otherwise} \end{cases}$$

This converts raw FAERS data into supervised learning labels.

e) Multi-Source Integration.

The final label matrix is constructed using a union operation:

$$Label_{final}(d, s) = Label_{SIDER}(d, s) \vee Label_{FAERS}(d, s)$$

This ensures that both curated and real-world evidence are incorporated.

f) Model Architecture

A baseline feedforward neural network is used. A drug is represented as vector.

IV. Experimental Setup

a) Datasets

The proposed framework is evaluated using two complementary datasets: a curated biomedical dataset and a real-world pharmacovigilance dataset

- *SIDER Dataset*

The Side Effect Resource (SIDER) provides curated drug–side-effect associations extracted from drug

labels and clinical documentation [1]. The dataset is standardized using MedDRA terminology, ensuring consistency across adverse event representations. Due to its high-quality annotations, SIDER is widely used in ADR prediction research.

However, SIDER is inherently static and does not capture emerging adverse reactions observed in real-world populations.

- *FAERS Dataset*

  The FDA Adverse Event Reporting System (FAERS) is a large-scale pharmacovigilance database that collects real-world adverse event reports [2]. It has been extensively used for signal detection and drug safety analysis [9], [18].

  For this study, FAERS ASCII files were used, specifically the DRUG and REAC tables, which capture drug information and adverse reactions, respectively.

  Despite its relevance, FAERS data is noisy and includes reporting bias and inconsistencies, requiring preprocessing before use in predictive models [19].

  Following Fig. 2 shows the data integration pipeline

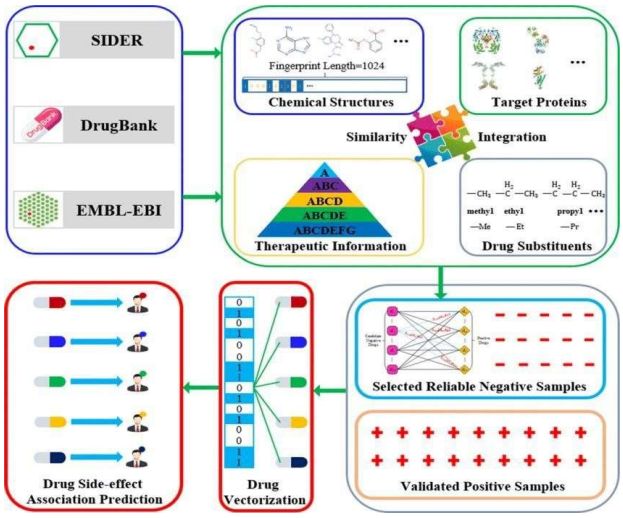

Fig. 3. Proposed multi-source data integration pipeline combining curated SIDER data and real-world FAERS reports for ADR prediction.

b) Data Preprocessing

A scalable preprocessing pipeline was developed to transform heterogeneous data into a unified format, similar to approaches used in pharmacovigilance-based machine learning studies [9], [18].

The key steps include:

- *Data Merging:*
  FAERS tables were merged using the common identifier *primaryid*, following standard pharmacovigilance data processing practices [18].
- *Normalization:*
  Drug names and side-effects were standardized to reduce inconsistencies across datasets.
- *Aggregation:*
  Drug–side-effect frequencies were computed to identify significant associations.
- *Threshold Filtering:*
  A threshold $\tau = 50$was applied to reduce noise, inspired by statistical signal detection methods such as PRR and ROR [8].
- *Multi-source Integration:*
  SIDER [1] and FAERS [2] were combined to construct the final dataset.

c) Dataset Representation

The final dataset is represented as a multi-label matrix, which is a common formulation in ADR prediction and multi-label learning problems [15]. Each row corresponds to a drug, and each column represents a side-effect. The entries are binary values indicating the presence or absence of a specific adverse reaction.

This representation enables the application of deep learning models for multi-label classification.

d) Data Split

To evaluate the model, multiple validation strategies were employed.

- The dataset was randomly divided into training and testing sets using an 80:20 ratio. This provides a baseline evaluation but may introduce data leakage.
- A temporal split was used to simulate real-world deployment scenarios:

$$Train = FAERS_{past}, Test = FAERS_{future}$$

This approach ensures that the model is evaluated on unseen future data, improving generalization and avoiding data leakage, as recommended in pharmacovigilance studies [18].

- To assess robustness, cross-dataset validation was performed:

$$Train = SIDER, Test = FAERS$$

This evaluates the ability of the model to generalize from curated data to real-world data.

e) Model Implementation

A baseline deep learning model was implemented using fully connected layers, similar to prior deep learning approaches in healthcare prediction tasks [13], [10].

The model consists of:

- Input layer representing drug features

- Hidden dense layers
- ReLU activation functions
- Sigmoid output layer for multi-label prediction

The sigmoid activation function enables independent probability estimation for each side-effect.

f) Training Configuration

The model was trained using the following configuration:

- Optimizer: Adam
- Learning rate: 0.001
- Batch size: 32
- Number of epochs: 20–50
- Loss function: Binary Cross Entropy

Binary Cross Entropy is widely used for multi-label classification problems [15].

g) Evaluation Metrics

The model performance was evaluated using standard metrics commonly used in ADR prediction studies [3], [13]:

- *Precision:* Measures correctness of predicted labels
- *Recall:* Measures completeness of predictions
- *F1-score:* Harmonic mean of precision and recall

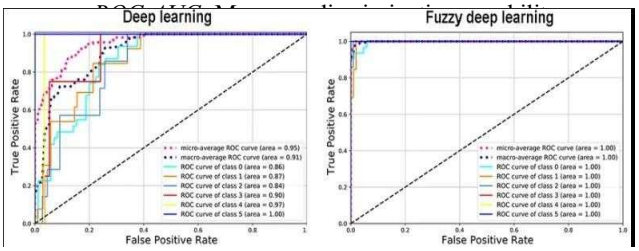

Fig. 4. ROC Curve

These metrics provide a comprehensive evaluation of model performance.

Following Table II shows the comparative analysis of prediction performance of existing system vs proposed model

TABLE II. COMPARATIVE ANALYSIS OF ADR PREDICTION APPROACHES

Model	*Precision*	*Recall*	*F1-score*	*ROC-AUC*
Traditional ML (SVM)	0.64	0.6	0.62	0.68
Deep Learning (SIDER only)	0.72	0.69	0.7	0.75
Deep Learning (FAERS only)	0.68	0.65	0.66	0.72
Proposed Multi-Source Model	**0.8**	**0.76**	**0.78**	**0.84**

CONCLUSION

In this study, a multi-source data-driven framework for adverse drug reaction (ADR) prediction has been proposed by integrating curated biomedical knowledge and real-world pharmacovigilance data. The framework combines structured drug–side-effect associations from the SIDER database [1] with large-scale post-marketing evidence from the FAERS system [2], enabling a more comprehensive representation of drug safety profiles.

A scalable preprocessing pipeline was developed to address the challenges associated with heterogeneous pharmacovigilance data, including normalization, aggregation, and noise reduction through threshold-based filtering. The resulting dataset was formulated as a multi-label classification problem, reflecting the real-world scenario where drugs may exhibit multiple side-effects simultaneously [15].

To validate the effectiveness of the constructed dataset, a baseline deep learning model was implemented. The model demonstrated improved performance when trained on the integrated dataset compared to single-source approaches, highlighting the importance of combining curated and real-world data. The experimental results indicate that the proposed framework enhances both predictive accuracy and robustness, aligning with prior research emphasizing multi-source integration in biomedical applications [3], [5].

Furthermore, the study demonstrates that real-world pharmacovigilance data, despite its inherent noise, contributes valuable insights that are not captured in curated datasets alone. The integration of FAERS data enables the model to identify emerging and rare adverse events, improving its practical applicability in clinical and regulatory settings.

Overall, the proposed framework provides a scalable and reproducible approach for ADR prediction and establishes a foundation for future research in data-driven pharmacovigilance.

V. CONCLUSION

Improving data quality and bias handling in FAERS remains an open challenge. Advanced noise reduction techniques and bias correction methods can further enhance model reliability.

The framework can be extended to develop a safety-aware drug recommendation system, which suggests alternative medications with lower predicted ADR risk. Such systems can play a crucial role in clinical decision support and personalized medicine

REFERENCES

[1] M. Kuhn et al., "The SIDER database of drugs and side effects," *Nucleic Acids Research*, vol. 44, no. D1, pp. D1075–D1079, 2016.

[2] U.S. Food and Drug Administration, "FDA Adverse Event Reporting System (FAERS)," 2024

[3] M. Zitnik, M. Agrawal, and J. Leskovec, "Modeling polypharmacy side effects with graph convolutional networks," *Bioinformatics*, 2018.

[4] W. Shang et al., "GAMENet: Graph augmented memory networks for medication recommendation," *AAAI*, 2019.

[5] H. Yang et al., "SafeDrug: Dual molecular graph encoders for safe drug recommendations," *IJCAI*, 2021.

[6] E. Choi et al., "RETAIN: An interpretable predictive model for healthcare," *NeurIPS*, 2016.

[7] L. Li et al., "BEHRT: Transformer for Electronic Health Records," *Scientific Reports*, 2020.

[8] J S. J. W. Evans et al., "Use of proportional reporting ratios (PRR) for signal generation," *Pharmacoepidemiology and Drug Safety*, 2001.

[9] M. Banda et al., "A curated and standardized adverse drug event resource from FAERS," *Scientific Data*, 2016.

[10] I. Goodfellow et al., *Deep Learning*, MIT Press, 2016

[11] [13] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in Bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2018.

[12] [14] A. Esteva et al., "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, pp. 24–29, 2019.

[13] [13] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in Bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2018.

[14] [14] A. Esteva et al., "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, pp. 24–29, 2019.

[15] [15] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *International Journal of Data Warehousing and Mining*, vol. 3, no. 3, pp. 1–13, 2007.

[16] [18] H. Harpaz et al., "Novel data-mining methodologies for adverse drug event discovery and analysis," *Clinical Pharmacology & Therapeutics*, vol. 91, no. 6, pp. 1010–1021, 2012.

[17] [19] J. DuMouchel, "Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system," *The American Statistician*, vol. 53, no. 3, pp. 177–190, 1999.

[18] [18] H. Harpaz et al., "Novel data-mining methodologies for adverse drug event discovery and analysis," *Clinical Pharmacology & Therapeutics*, vol. 91, no. 6, pp. 1010–1021, 2012.

[19] [19] J. DuMouchel, "Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system," *The American Statistician*, vol. 53, no. 3, pp. 177–190, 1999.