Vol. 8, No. 1, February 2026

Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

# Small Language Models for Drug-Drug Interaction Extraction from Biomedical Text: A Systematic Literature Review

**Fortunate Mutanda[1], Belinda Ndlovu[2]**

[1,2]Informatics and Analytics Department, National University of Science and Technology, Bulawayo, Zimbabwe

**Received:**
December 1, 2025
**Revised:**
January 15, 2026
**Accepted:**
January 28, 2026
**Published:**
February 26, 2026

Corresponding Author:
**Author Name*:**
Belinda Ndlovu
**Email*:**
Belinda.ndlovu@nust.ac.zw

DOI:
10.63158/journalisi.v8i1.1430



**Abstract.** Drug–drug interaction (DDI) extraction from biomedical text is central to pharmacovigilance but remains challenging in resource-constrained clinical environments. While large language models have shown promise, their computational cost and deployment complexity limit practical adoption. This study systematically reviews the role of small language models (SLMs) for DDI extraction and examines their effectiveness, efficiency, and deployability. A systematic literature review was conducted following PRISMA guidelines, covering empirical studies published between 2020 and 2025 in PubMed, IEEE Xplore, ACM and SpringerLink. Eligible studies were analysed with respect to model architectures, datasets, evaluation metrics, and deployment considerations. Quality assessment was applied to ensure methodological robustness. The synthesis indicates that SLM-based approaches, including CNN-, LSTM-, and lightweight transformer models, can achieve competitive F1-scores on benchmark DDI datasets while requiring substantially fewer computational resources than large language models. However, performance varies across datasets, and real-world clinical evaluations remain limited. These findings support the feasibility of deploying SLM-based DDI extraction systems in resource-constrained clinical and pharmacovigilance settings and provide a baseline for future benchmarking and comparative research in clinical natural language processing.

**Keywords**: Drug-Drug Interaction, Biomedical Text Mining, Small Language Models, Pharmacovigilance, Clinical NLP

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

## 1. INTRODUCTION

Excessive polypharmacy in clinical practice has increased the risk of Drug–Drug Interactions (DDIs), with effects of treatment inefficacy, adverse reactions, and avoidable hospitalizations [1], [2]. Detecting DDIs at scale remains a significant challenge in pharmacovigilance. Traditional rule-based and database-driven systems rely on static sources that fail to record novel or context-specific interactions in real time [3], [4], [5], which is crucial for the clinical relevance of an interaction [2], [6], [7]. The rapid expansion of biomedical text sources, PubMed abstracts, clinical trial reports, and DrugBank entries, has created new opportunities to extract and predict DDIs directly from unstructured text through natural language processing (NLP) [3], [5], [6]. Building on this, [3], [5], [6], [8]has shown that a substantial body of DDI evidence resides in unstructured biomedical narratives and is particularly well-suited to automatic DDI extraction.

Early computational efforts in DDI prediction employed machine-learning models that used molecular fingerprints, chemical similarities, and pharmacological properties. While such models improved efficiency, their ability to capture semantic context from biomedical literature was limited [3], [5]. The emergence of deep learning (DL) introduced more expressive representational capabilities, and recent surveys consistently organize DDI prediction methods into neural-network-based, graph-neural-network (GNN) based, knowledge-graph (KG) based, and multimodal families, with GNNs and KG embeddings emerging as powerful frameworks for modelling complex drug–drug relationships [3], [4]. At the same time, advances in NLP, particularly transformer-based architectures, have revolutionized text-driven DDI extraction by learning contextual dependencies between drug entities within sentences and documents [3], [9]. As noted by [3], [5], Biomedical variants of BERT have become the dominant backbone for text-based DDI extraction, largely because they consistently perform well on corpora derived from DrugBank and MEDLINE. This trend is reinforced by [4], [6], who identify the DDIExtraction 2013 dataset as the field's primary benchmark for relation extraction across literature-based DDI tasks.

According to [5], Deep Learning (DL) approaches for DDI prediction can be grouped into pure DL, KG, and hybrid methods, with hybrids performing best when textual and molecular evidence are fused. Similarly, [3], [4] indicate that transformer-based models do better than conventional ML models, though they are computationally expensive.

Moreover, [10] observes that models capable of multimodal fusion, integrating text representations and molecular graphs, typically surpass text-only models, though at the cost of an extra computational burden. In parallel, [6] demonstrates that BioBERT and PubMedBERT outperform CNN and RNN architectures for DDI extraction from biomedical abstracts, yet their deployment remains limited by model size and memory constraints. These findings suggest that while transformer-based models currently offer the highest predictive performance, they are not easily deployable in clinical or low-resource environments due to their heavy computational footprint [11], [12]. The high computational cost of large transformer architectures has prompted growing interest in SLMs, which are compact transformer variants designed for deployment in resource-constrained clinical settings. Compared with Large Language Models (LLMs) such as GPT-based systems that depend on extensive GPU resources, SLMs rely on parameter-efficient fine-tuning strategies (PEFT) and knowledge distillation to reduce computational demands while retaining competitive predictive performance. Despite these advantages, the application of SLMs to biomedical DDI extraction has received limited empirical attention.

Despite growing research interest, several limitations persist across previous studies. Firstly, as noted in recent surveys, most prior work emphasizes large and complex deep-learning architectures, particularly GNN-based, KG-based, and transformer-based models, or else surveys the wider AI and ML landscape for DDI prediction, with little attention given to SLMs specifically designed for text-driven DDI extraction [3], [4], [13]. Secondly, benchmark datasets such as DDIExtraction 2013 are widely reused, a pattern noted by [4], [5], which increases the risk of overfitting to narrow biomedical domains and contributes to the under-representation of rare drug pairs [3], [8]. Moreover, curated literature corpora differ markedly from real-world clinical narratives, resulting in persistent domain-shift challenges [6]. Third, [6], [14] both highlight that most DDI systems continue to frame interaction detection as a static classification task and seldom integrate mechanisms for explainability or causal reasoning features essential for clinician trust and regulatory acceptability in medical AI. Finally, recent reviews show that PEFT and other light optimization methods remain lacking in the DDI extraction task despite their promise to make transformer-based models more compact, faster, and interpretable, thereby improving sustainability in real-world clinical settings [3], [4], [6].

Literature identifies another important gap where most models for DDI use combined literature-based data, not considering the influence of individual patient-specific factors such as age, gender, or comorbidities, despite their major contribution to clinical DDI risk [2], [3]. Prediction models, including clinical decision support systems (CDSS), provide uniform alerts once an interaction is detected, without risk stratification by patient context or personalized practical recommendations such as pharmacological alternatives with reduced risks, dose alterations, and improved monitoring methods, among others [5], [7]. This consistent absence of patient-centered reasoning and outcome-oriented mitigation emphasizes the need for adaptive and interpretative models capable of going beyond binary interaction detection towards broader context-driven recommendations suitable for specific patient subgroups [6], [14].

Existing reviews acknowledge the potential of SLMs and parameter-efficient approaches, yet they only introduce these models in passing within broader discussions of BERT-related architectures, distillation, and low-rank adaptation, without providing a dedicated synthesis of their application in DDI prediction [3], [4], [13]. Moreover, ongoing AI-oriented DDI research continues to prioritize large deep-learning, GNN-based, and full-scale transformer models, rarely distinguishing SLMs as a separate methodological class or evaluating their comparative advantages in efficiency, interpretability, or deployment in clinical settings [7], [8]. As a result, despite rising interest in resource-efficient biomedical NLP, there is still no comprehensive synthesis focused solely on SLMs for DDI extraction from biomedical text data, a gap that is increasingly significant given their potential to deliver high-quality pharmacovigilance insights under realistic computational constraints [14], [15].

Transformer-based architectures have advanced biomedical relation extraction substantially; however, much of the existing literature concentrates on LLMs and computationally demanding multimodal systems. Although these approaches achieve strong benchmark performance, their high memory requirements, high training costs, and limited interpretability make them unsuitable for deployment in clinical settings. In contrast, SLMs offer an alternative approach, achieving competitive predictive accuracy with markedly lower computational requirements. This study bridges this gap by providing a systematic literature review focused specifically on the use of SLMs for DDI extraction from biomedical text. It examines model architectures and fine-tuning

strategies, relates architectural choices to recurring challenges such as class imbalance, semantic ambiguity, computational overhead, and limited explainability, and evaluates performance–efficiency trade-offs to clarify the practical relevance of SLMs for scalable pharmacovigilance applications.

By consolidating architectural, performance, and deployment trade-offs across small language models, this review provides a reference baseline for future research and benchmarking in clinical DDI extraction. The review aims to address the following questions:

1) RQ1: Which SLMs, transformer-based architectures, and fine-tuning strategies have been applied to biomedical text for DDI extraction and prediction?
2) RQ2: Which biomedical text sources and datasets are commonly used for training and evaluating DDI extraction models?
3) RQ3: How does the performance of SLMs compare with that of LLMs and graph-based approaches?
4) RQ4: What are the challenges reported in implementing SLMs for scalable and potentially personalized DDI prediction?

## 2. METHODS

This review was conducted following the PRISMA framework [16]. Study selection proceeded through identification, screening, eligibility assessment, and inclusion, using predefined criteria and systematic evaluation of the results.

### 2.1. Search Strategy

An initial exploratory search using Google Scholar identified key terms and indexing trends. This was followed by a systematic search of PubMed, ACM Digital Library, SpringerLink, and IEEE Xplore, limited to publications from 2020 to 2025. These databases were chosen to reflect the intersection of biomedical informatics, clinical natural language processing, and artificial intelligence research. Scopus and Web of Science were excluded because institutional access was limited and because much of their indexed content overlaps with material accessible through SpringerLink and IEEE Xplore. Together, the selected sources span both biomedical and computational research domains while reducing duplication during retrieval. The strategy used a set of keywords

and their synonyms to formulate a search query, which was adjusted as needed to conform to the database syntax. The core search expression was:

*("drug-drug interaction" OR "DDI") AND ("small language model" OR "transformer" OR "BERT" OR "BioBERT" OR "PubMedBERT" OR "machine learning" OR "deep learning" OR "artificial intelligence") AND ("biomedical text" OR "biomedical literature" OR "clinical text" OR "electronic health records" OR "EHR").*

To ensure completeness, forward and backward citation tracing was performed using Connected Papers, facilitating the identification of additional empirical studies not retrieved through the initial database search.

### 2.2. Inclusion and Exclusion Criteria

Only English-language publications were included in order to avoid translation-related ambiguity in biomedical terminology, model descriptions, and evaluation metrics, and enabled consistent extraction of methodological details. Table 1 shows the Inclusion and exclusion criteria used.

**Table 1.** Inclusion and exclusion criteria

| Criteria | Inclusion | Exclusion |
|---|---|---|
| Study Topic | Studies focusing on the extraction or prediction of drug-drug interactions from biomedical text sources. | Studies focusing on unrelated tasks (drug-target, drug-disease prediction) or molecular simulations (PPI, ADR prediction) without text mining. |
| Data Type | Uses biomedical text as either primary input or in multimodal combinations. | Studies relying exclusively on non-textual data without an NLP component. |
| Model Type | Implements SLMs, Transformer-based architectures, or hybrid Deep Learning models used as comparative baselines. | Approaches utilizing strictly traditional machine learning or rule-based systems without neural embeddings. |

| Criteria | Inclusion | Exclusion |
|---|---|---|
| Task Output | Produces specific DDI classification or relation extraction outputs. | Does not produce specific DDI-specific outputs |
| Empirical Evidence | Reports empirical evaluation metrics on standard benchmarks. | Theoretical frameworks lacking experimental validation. |
| Publication Type | Peer-reviewed journal articles or full conference papers. | Reviews, commentaries, editorials, letters, book chapters, theses, posters, and preprints without peer review. |
| Accessibility | Full text available through institutional access or open access. | Unavailable or pay-walled papers without retrievable content. |
| Timeframe | Studies published between 2020 and 2025 | Studies published before 2020 |
| Language | Strictly English | Non-English publications |

### 2.3. Screening

The initial search yielded 536 records. The distribution across databases included SpringerLink (n = 355), ACM Digital Library (n = 108), IEEE Xplore (n = 33), PubMed (n = 29), and other sources (n = 11). After the removal of 12 duplicate papers, there were 524 papers left to screen. During initial screening based on title and abstract, 487 papers were removed because they were review papers or not in scope in certain domains, such as computational chemistry, molecular biology, and pure pharmacology. The studies left for full-text review were 37.

### 2.4. Eligibility

The review included peer-reviewed empirical papers published between 2020 and 2025. The review timeframe (2020–2025) was selected to capture the period of widespread adoption of transformer-based architectures and the emergence of SLMs in biomedical NLP. The criteria included only studies focusing on DDI prediction using SLMs or biomedical text data, or hybrid methods combining both. Studies were excluded if they:

(1) did not apply language or transformer-based models; (2) were solely based on predictions or extractions other than DDI; (3) were based only on non-textual data sources; or (4) were not peer-reviewed studies. During full-text reading, studies were excluded if they applied transformer models to adverse drug reaction detection or drug–target prediction without performing DDI extraction, relied mainly on molecular graph reasoning without textual input, or employed only traditional machine learning approaches, such as CNNs or SVMs, without transformer components.

### 2.5. Included

Only 28 studies met the predefined inclusion criteria. Some papers, such as Systematic Literature Reviews, were excluded because they provided relevant foundational literature for understanding DDI prediction. The selection process is shown in a PRISMA flow diagram in Figure 1.

### 2.6. Data Extraction and Synthesis

After screening, all qualified studies were then imported into Mendeley Desktop Reference Manager. Full-text papers were retrieved from databases and systematically categorized by title and publication year. Data extraction was done using a Microsoft Word table. Extracted variables include author, publication year, fine-tuning strategy, model architecture, main metrics, data source, task type, key challenges, and future directions. All data was cross-checked for accuracy by having FM complete the data extraction first before verifying it with BN.

### 2.7. Quality Assessment

A systematic quality assessment was conducted to evaluate the methodological quality and validity of the 28 included studies. The assessment was done utilizing the Kitchenham Quality Assessment Framework, [17] which was tailored to account for the characteristics of computational biomedical NLP tasks and transformer-based models. A customised evaluation form included 12 criteria research questions covering data adequacy, model transparency, reproducibility of fine-tuning, and practical relevance to pharmacovigilance.

Each criterion was scored on a three-level scale (1 = fully addressed, 0.5 = partially addressed, 0 = not addressed), yielding a maximum score of 12. Studies were classified as

high (≥9), medium (6–8.5), or low quality (<6). The mean quality score across the 28 included studies was 8.0 out of 12 (66.7%). Five studies (17.9%) were rated as high quality (≥9), while the remaining 23 studies (82.1%) were rated as medium quality (6–8.5). Studies receiving higher scores consistently reported reproducibility-related details, including model transparency and empirical evaluation. In contrast, medium-quality studies often omitted elements such as full hyperparameter disclosure, ablation analyses, or validation beyond standard benchmark datasets. Quality scores were used to inform weighting and interpretation during synthesis rather than as exclusion thresholds. The study delimitation process is illustrated in Figure 1, the PRISMA flowchart by [18]:

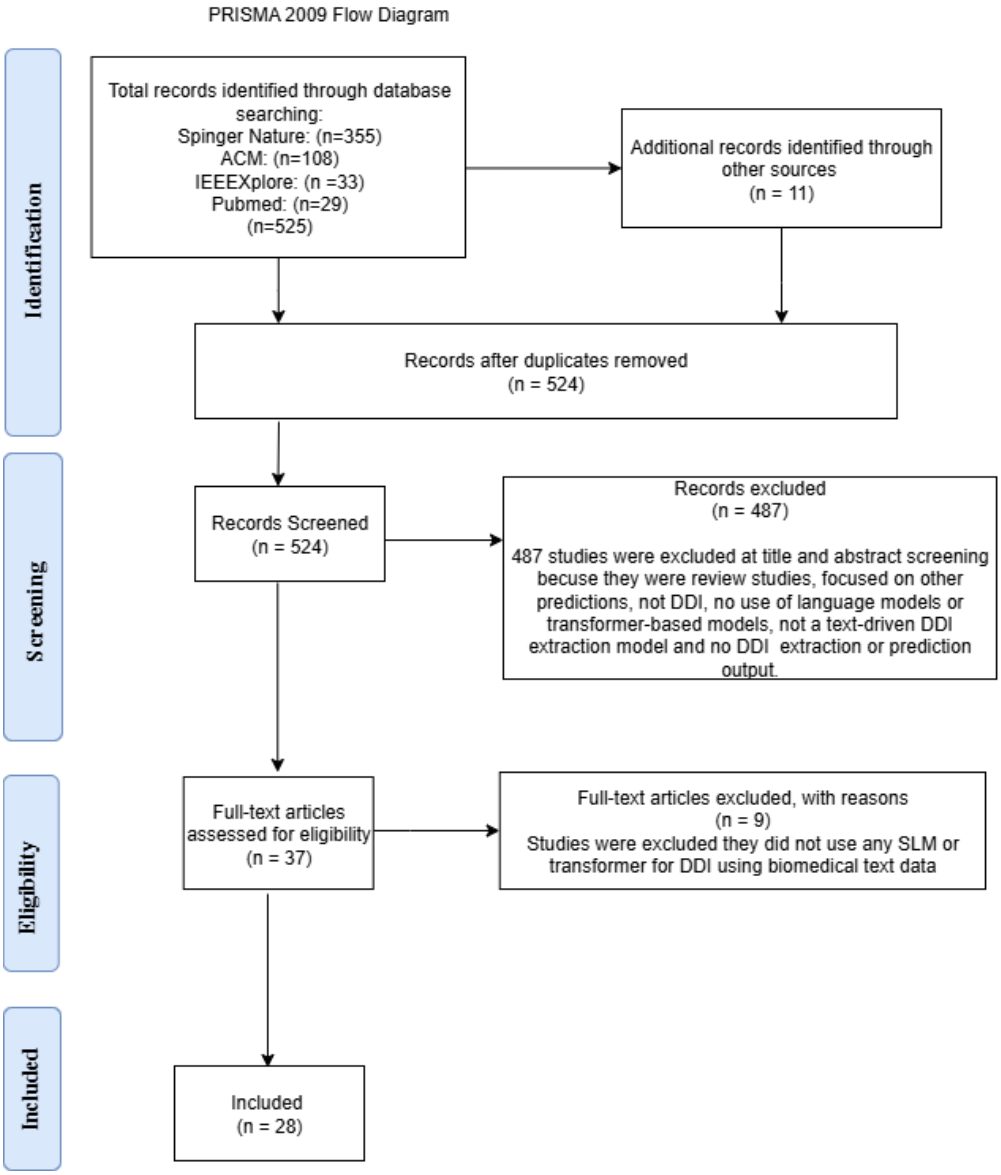


**Figure 1:** Prisma Flow Diagram

## 3. RESULTS AND DISCUSSION

Table 2 provides a consolidated, study-by-study overview of the 28 included papers, organized to directly answer the four review questions: model architecture (RQ1), benchmark datasets (RQ2), predictive performance (RQ3), and reported technical challenges and opportunities (RQ4). By bringing these dimensions into a single matrix, Table 2 enables rapid cross-study comparison (e.g., transformer vs. CNN/LSTM vs. graph/multimodal vs. small/decoder-only LLM approaches), while still preserving the implementation details that are often lost in high-level summaries (fine-tuning strategy, task formulation, and dataset choice). Because the table is intentionally comprehensive, it is designed primarily as a reference point for the reader; the interpretation and thematic synthesis of patterns observed across studies are emphasized in the subsequent narrative analysis.

Reading across Table 2 reveals clear methodological clustering around a small number of dominant paradigms. Earlier or lightweight baselines rely on traditional machine learning with engineered features, which typically show competitive but capped performance and limited expressivity for complex relation cues. A second cluster uses CNN/LSTM-style architectures (e.g., CNN-DDI and hierarchical ConvBLSTM) that frequently perform well on sentence-level classification benchmarks (notably SemEval-2013 and DDIExtraction 2013), but repeatedly report weaknesses in rare classes and limited contextual modeling—issues that are explicitly noted in the "Key Challenges" column of Table 2. More recent studies increasingly adopt transformer-based encoders (BioBERT, PubMedBERT, Bio-RoBERTa variants) and enhanced entity-aware adaptations, which generally improve representation quality yet remain sensitive to data imbalance, type confusion (particularly Int vs. Effect), and the constraints of single-sentence relation extraction, as documented across multiple rows in Table 2. Finally, a growing subset explores graph-based and multimodal fusion (e.g., knowledge-graph integration, molecular structure encoders, formula/image fusion), which can boost predictive strength but often introduces practical barriers such as computational overhead, pipeline complexity, and interpretability gaps, again consistently recorded in Table 2.

The dataset landscape summarized in Table 2 also explains why some performance numbers are not directly comparable across studies. A large portion of the included work

benchmarks on SemEval-2013 Task 9 and DDIExtraction 2013, which support standardized evaluation but also concentrate research on sentence-level signals and a fixed set of interaction categories. In contrast, DrugBank-centric studies—particularly those framed as DDI prediction rather than strict sentence-level extraction—often depend on the completeness and coverage of structured records, with limitations such as underrepresentation of rare drug combinations and single-source dependence explicitly highlighted in Table 2. Studies using additional resources (e.g., TWOSIDES, Hetionet, PubMed-derived corpora, TAC datasets, or EHR contexts) indicate a push toward broader generalization; however, Table 2 shows that cross-dataset and out-of-distribution (OOD) evaluation remains uneven, and several papers note limited external validation as an unresolved weakness.

Importantly, the "Key Challenges" and "Future Opportunities" columns in Table 2 converge on a small set of recurring technical bottlenecks: (i) severe class imbalance and minority-class underperformance, (ii) semantic overlap between relation types (especially Int vs. Effect), (iii) limitations in handling cross-sentence or document-level interactions, (iv) reliance on incomplete knowledge graphs or noisy dependency parses for graph-based methods, and (v) the trade-off between performance and computational feasibility for clinical deployment. At the same time, Table 2 indicates shared directions for improvement, including targeted data augmentation for sparse types, interpretability techniques (attention visualizations, heatmaps, explanation frameworks), stronger knowledge integration (UMLS/MeSH, richer KGs, retrieval-augmented pipelines), and more robust OOD evaluation designs.

**Table 2.** Summary of Models, Architectures, Performance, and Key Challenges in Biomedical Text-Based DDI Extraction

| Study | Model/ Architecture | Fine-Tuning Strategy | Metrics | Dataset | Task | Key Challenges | Future Opportunities |
|---|---|---|---|---|---|---|---|
| [19] | D3, Decoder-only transformer (LLaMA/Mistral) | Supervised Fine-Tuning (SFT) with AdamW | F1: 85.8%<br>Acc: 85.8%<br>Prec: 85.3%<br>Rec: 86.4% | DrugBank | DDI prediction (multi-class classification : Minor, Moderate, Major) | • Limited to drugs with complete DrugBank records<br>• Rare drug combinations are underrepresented. | • Integration of multi-modal features.<br>• Improved personalization and contextual modelling |

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

| Study | Model/ Architecture | Fine-Tuning Strategy | Metrics | Dataset | Task | Key Challenges | Future Opportunities |
|---|---|---|---|---|---|---|---|
| | | | | | | • single source dependence. | |
| [20] | Traditional ML classifiers (Logistic Regression, SVM, Random Forest, Naïve Bayes, KNN) | Traditional Feature Engineering + Classical ML Training | F1: 82.7%<br>Acc: 79.5%<br>Prec: 86.9%<br>Rec: 81.0% | SemEval 2013. | Binary DDI detection (interaction vs no interaction) | • Only two drug names can be entered at a time.<br>• Predictions are binary and do not distinguish interaction types. | • Adoption of transformer and deep learning models.<br>• Transition towards multi-class DDI classification. |
| [21] | CNN-DDI (CNN + Word2Vec) | Supervised CNN Training with Pre-Trained Word2Vec Embeddings | F1 = 83.81%<br>Acc: 86.81%<br>Prec: 84.87%<br>Rec: 86.81% | SemEval-2013 | DDI classification (5 classes: false, int, effect, mechanism, advice) | • weak performances on rare classes.<br>• Dependence on pre-trained word2Vec.<br>• no cross-dataset generalization beyond Semval-2013.<br>• limited contextual modelling | • Incorporate Interpretability techniques such as heatmaps<br>• Hybridize CNN with lightweight transformers |
| [22] | Bio-RoBERTa + tree-transformers + H-GAT + BiLSTM. | Supervised Multi-Branch Transformer Fine-Tuning + H-GAT Update Mechanism | F1: 95.2%<br>Prec: 95.5%<br>Rec: 94.9%. | DDIExtraction 2013 | DDI extraction (5 classes) | • Computationally heavy.<br>• Limited to single sentence interactions.<br>• High training time. | • Inclusion of knowledge graphs to improve the model performance. |
| [23] | BioMedBERT + GNN + CNN + formula encoder + molecular images (Intermediate fusion architecture) | Intermediate Multimodal Fusion Training (Text + Structure + Formula + Image) Supervised Fine-Tuning with AdamW | F1: 84.45%<br>Acc: 86.81%<br>Prec: 85.15%<br>Rec: 83.76% | SemEval-2013 | Multimodal DDI classification (Mechanism, Effect, Advice, Int) | • Conflicts between heterogeneous modalities<br>• Model functions as a black box, limiting clinical trust.<br>• Data imbalance, negative data significantly outnumber positive data | • Explore Mixture of Experts frameworks for modality conflict resolution.<br>• Enhanced interpretability in multimodal fusion |

| Study | Model/ Architecture | Fine-Tuning Strategy | Metrics | Dataset | Task | Key Challenges | Future Opportunities |
|---|---|---|---|---|---|---|---|
| [24] | DyNAS-DDI: Galactica-1.3B LLM + GNN + NAS + LoRA PEFT | Instruction Tuning + LoRA-Based PEFT + Neural Architecture Search (NAS) | DrugBank: AUC-ROC: 88.86% Acc: 64.86%. F1: 61.27%<br>ChChMiner AUC-ROC: 87.22% Acc: 79.44% F1: 79.35% | ChChMiner, DeepDDI, ZhangDDI, DrugBank. | Out Of Distribution DDI prediction (binary + multi-class) | • Difficult OOD generalization<br>• Requires external biomedical knowledge retrieval.<br>• NAS is computationally intensive. | • Expanded OOD evaluation.<br>• Explainable NAS-LLM pipelines |
| [25] | MOF-DDI: Multi-modality Feature Optimal Fusion for Drug-Drug Interaction Prediction (BioBERT + KG encoder + molecular GNN) | Cross-Modality Alignment Training + Optimal Transport Fusion Supervised Fine-Tuning | DrugBank F1: 89.04% Acc: 91.15%<br>TWOSIDES: Acc: 85.28% ROC-AUC: 92.50% | • DrugBank<br>• TWOSIDES<br>• Hetionet<br>• PubMed | Multi-class DDI prediction with multimodal fusion (text + KG + molecular structure) | • KG incompleteness affects generalization.<br>• Optimal transport is computation-intensive<br>• Missing entities in KG limit coverage. | • Expand KG entity coverage.<br>• Improved fusion and real-time integration. |
| [26] | Hierarchical ConvBLSTM: CNN + BiLSTM | Hierarchical Supervised Training (CNN → BLSTM) Pre-Trained Embedding Integration | Russian: F1: 96.39% English: F1: 98.37% | DDI Extraction 2013 | Hierarchical DDI classification (binary then 4-class: advise, effect, mechanism, int) | • Isolated features exhibit reduced performance.<br>• Binary detection Is Insufficient for clinicians<br>• Limited to DDI2013 dataset | • Incorporate contextual information and expand the dataset<br>• Evaluate on Electronic Health Records |
| [27] | BBL-GAT: BioBERT + BiLSTM + Graph Attention Network | Supervised Fine-Tuning with Graph Attention + BiLSTM Integration | F1: 82.47% Prec: 81.76% Rec: 84.38% | DDI Extraction 2013 | DDI relation extraction (5 classes: mechanism, effect, advice, int, negative) | • GAT requires graph construction.<br>• Dependency parsing errors.<br>• Computationally expensive attention mechanisms | • Clinical decision support integration and advanced NLP algorithms. |
| [28] | BERT (BioBERT) | Supervised Transformer Fine-Tuning (BioBERT) | Acc: 90.69% F1: 81.97% | DDIExtraction 2013 | DDI classification into 5 types: Mechanism, | • Dataset imbalance.<br>• Int class has limited | • Automatic DDI extraction system for |

| Study | Model/ Architecture | Fine-Tuning Strategy | Metrics | Dataset | Task | Key Challenges | Future Opportunities |
|---|---|---|---|---|---|---|---|
| | | | | | Effect, Advice, Int, Negative | performance due to data sparsity • Model struggles with minority classes. | scanned medical documents. • Extended BERT-D2 for image-based medical text processing. |
| [29] | BIOGAN-BERT (Fine-tuned BioGPT-2 + GAN-BERT) | GAN-Based Semi-Supervised Learning PLM-Driven Data Augmentation | Micro F1-Score (minor classes): 85.0% | • DDIExtraction 2013 • TAC 2019 | DDI extraction and classification (5 classes: Negative, Advice, Effect, Mechanism, Int) | • Class imbalance despite mitigation efforts. • High computational requirements for GAN. • Dependence on pseudo-label quality. | • Explore semi-supervised methods like SALTClass for sparse data. |
| [30] | R-BioBERT + BLSTM | Supervised Fine-Tuning (R-BioBERT) + BLSTM Encoding | SemEval 2013: F1-Macro: 83.32% Prec: 86.20% Rec: 86.29%<br>TAC 2019: F1-Macro: 80.23%<br>TAC 2018: F1-Macro: 60.53% | • SemEval 2013 • TAC 2018 • TAC 2019 | DDI classification (Advice, Effect, Int, Mechanism, Negative) | • Semantic similarity between int and effect types causes errors. • Small Int type sample size leading to low recall. • Dataset imbalance. | • Data augmentation strategies for minority classes. |
| [31] | R-BERT/ R-BioBERT variants | Entity-Aware Supervised Fine-Tuning (R-BERT/R-BioBERT) | $R\text{-}BioBERT_1$: F1-Macro: 80.89%<br>$R\text{-}BioBERT_2$ F1-Macro: 80.52%<br>R-BERT: F1-Macro: 79.08% | DDIExtraction 2013 corpus | DDI relation classification into 5 types: Mechanism, Effect, Advice, Int, Negative | • Class imbalance. • Int type has the lowest recall due to sparsity. • Some mislabeled samples exist in the dataset • Imbalance between positive and negative. | • Investigate the multi-labelled drug pairs and expand to other biomedical relation extraction tasks |
| [32] | Bio-ER-BERT | Enhanced Entity Representatio | $Bio\text{-}ER\text{-}BERT_2$: Prec: 84.14% Rec: 83.62% | DDIExtraction 2013 | DDI classification into 5 types: | • Poor performance in the Int class. | • Develop methods to address |

| Study | Model/ Architecture | Fine-Tuning Strategy | Metrics | Dataset | Task | Key Challenges | Future Opportunities |
|---|---|---|---|---|---|---|---|
| | (BioBERT + R-BERT + LSTM enhanced entity embeddings) | n Fine-Tuning (LSTM-Entity Layer) + Supervised Learning | F1: 83.88%<br>Bio-ER-BERT$_1$:<br>Prec: 81.11%<br>Rec: 85.78%<br>F1: 83.38 | corpus (filtered). | Negative, Mechanism, Effect, Advice, Int | • Severe class imbalance.<br>• High error rate for minority classes. | severe class imbalance. |
| [33] | Hybrid neural-symbolic model: PubMedBERT + CNN + Hierarchical Attention + Knowledge Graph Embeddings | Knowledge-Graph-Integrated Training + Multi-Focal Loss (Imbalance-Sensitive Training) | Micro F-score: 86.24%<br>Prec: 86.20%<br>Rec: 86.29% | • DDIExtraction 2013<br>• ChemProt<br>• DrugBank KG | DDI classification (Mechanism, Effect, Advice, Int, Negative) | • Int and effect misclassification<br>• Ambiguous or non-specific formulations (Drug A may affect Drug B") lack clear directional cues.<br>• Mislabeled samples in the dataset. | • Increase the number of "Effect" and "Int" type training examples by data augmentation methods. |
| [34] | BioBERT + multiple entity-aware attention mechanisms + BiGRU | Entity-Aware Multi-Attention Supervised Training + BiGRU Sequence Encoding | F1: 80.9%<br>Prec: 81.0%<br>Reca80.9% | DDIExtraction 2013 | DDI classification into 5 types: Negative, Advice, Effect, Mechanism, Int | • Misclassification of Int as Effect type.<br>• Misclassification of positive instances as negative.<br>• Severe class Imbalance<br>• Multi-labelled drug pairs | • Fine-tune BioBERT.<br>• Address multi-labelled pairs. |
| [35] | LoRA-BiomedBERT | LoRA-Based PEFT + Pseudo-Labelling (Semi-Supervised) | Acc: 79.96%<br>F1: 79.75%<br>Prec: 81.09%<br>Rec: 79.94% | • DrugBank<br>• DrugComb | Binary polarity classification: Synergistic vs Antagonistic | • Classifier struggled with ambiguous or non-specific formulations<br>• Sensitivity to polarity class definition.<br>• Class imbalance sensitivities. | • Integration with real-time EHR systems |
| [36] | BioSentVec + BiLSTM + Hierarchical Attention Network | Supervised BiLSTM + Hierarchical Attention Training | Acc: 90.74%<br>AUROC: 99.20<br>AUPR: 95.66 | DrugBank | Multi-label DDI (164 types) | • Types with similar descriptions are prone to misclassification. | • Address data imbalance through external database integration or |

| Study | Model/ Architecture | Fine-Tuning Strategy | Metrics | Dataset | Task | Key Challenges | Future Opportunities |
|---|---|---|---|---|---|---|---|
| | | | | | | • Fixed input structure and length.<br>• Cannot predict unseen DDI types: Limited to 164 predefined classes.<br>• Limited to pairwise predictions (no 3+ combinations)<br>• Data imbalance | SMOTE augmentation.<br>• Explore meta-learning for unseen DDI types. |
| [37] | BERT (BioBERT BioBERT+LSTM BioBERT+Attention) | Revised Transformer Fine-Tuning (Full Last-Layer Utilisation) Supervised Training | Prec: 81.3%<br>Rec: 80.1%<br>F1: 80.7% | DDI Extraction -2013 | • DDI extraction (binary and multi-class).<br>• ChemProt relation extraction | • Original BERT models only utilize the CLS token, discarding the other last-layer outputs.<br>• Sample imbalance in datasets. | Apply full-layer fine-tuning to other tasks. |
| [38] | • BioBERT<br>• PubMedBERT<br>• BioBERT+BiLSTM<br>• BioBERT+Attention | Sequence-Level Layering (SLL) + Domain-Adaptive Fine-Tuning | • DDI Extraction-2013: BioBERT+SLL_Att+D: F1: 81.9% PubMedBERT+SLL: F1: 84.0%<br>• ChemProt: F1=78.7% BioBERT+SLL_Att+CP: F1: 77.0% | • DDI Extraction-2013<br>• ChemProt | • DDI detection (binary)<br>• PPI extraction<br>• ChemProt relation extraction | • Models struggle with minority classes.<br>• Limited by computational resources.<br>• Sub-domain data quality affects results | • Utilize all last-layer information via SLL fine-tuning. |
| [39] | BioFocal-DDI: (BioGPT + BioBert + ReGCN + focal-loss attention) | PLM-Based Data Augmentation (BioGPT) + Multi-Focal Loss Training | Prec: 86.75%<br>Rec: 86.53%<br>F1:86.64% | DDI Extraction -2013 | DDI extraction (Effect, Mechanism, Advice, Int, False) | • Severe class imbalance.<br>• Effect and mechanism misclassification<br>• Underperformance on rare or | • Incorporate biomedical ontologies such as MeSH or UMLS.<br>• Contextual Augmentation may improve |

| Study | Model/ Architecture | Fine-Tuning Strategy | Metrics | Dataset | Task | Key Challenges | Future Opportunities |
|---|---|---|---|---|---|---|---|
| | | | | | | low-frequency classes. | classification accuracy for overlapping categories. |
| [40] | • Llama 3.2<br>• Microsoft Phi 3.5<br>• Google Gemma 2 | Zero-shot + few-shot in-context learning. | ---- | Electronic Health Records | Medical entity recognition, relation extraction (medication-dosage, drug-drug interaction) | • limited domain grounding.<br>• context understanding gaps in SLMs. | • Combine SLMs with structured KMs. |
| [41] | SCAT (BioBERT + Doc2Vec + Graph Convolutional Network + BiGRU) | Cross-Attention Multimodal Fusion Training Supervised Fine-Tuning | | DDI Extraction -2013. | DDI prediction (Negative, Mechanism, Effect, Advise, Int) | • Class imbalance<br>• Multimodal feature fusion increases training latency.<br>• Filtering may remove valid Instances. | Explore more multimodal data and new embedding approaches to improve interpretability. |
| [42] | SubGE-DDI (PubMedBERT SubAGCN + CNN) | Knowledge-Subgraph-Enhanced Training + Multi-Focal Loss Fine-Tuning | F1_micro: 83.91%, F1_macro: 84.75%. | SemEval-2013 task 9 | DDI classification (Mechanism, Effect, Advise, Interact, Negative) | • Class imbalance<br>• Int class often misclassified as Effect.<br>• The model cannot handle cross-sentence relations. | • Explore meta-paths for graph learning.<br>• Apply data augmentation. |
| [43] | TL-BERT: BioBERT/PubMedBERT + Triplet Loss | Triplet-Loss Metric Learning + Supervised Transformer Fine-Tuning | Prec 89.1% Rec: 88.5% F1: 88.8%. | DDI Extraction -2013 | DDI detection (binary classification . | • Same origin, different class confusion.<br>• Ambiguity caused by shared context.<br>• Triplet Loss is complex and only suitable for binary classification. | • Extend triplet-loss to handle multi-class relation problems. |
| [44] | TP-DDI: BioBERT (NER + RC pipeline) | Pipeline Fine-Tuning: Supervised NER | DDI: Prec: 86.4% Rec: 78.8% F1: 82.4% | DDI Extraction -2013 | DDI classification (4 positive + negative) | • Error propagation from DNER to DDI. | • Improve multi-token entity handling and |

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

| Study | Model/ Architecture | Fine-Tuning Strategy | Metrics | Dataset | Task | Key Challenges | Future Opportunities |
|---|---|---|---|---|---|---|---|
| | | (BioBERT) + Supervised RC (BioBERT) | End-to-end: F1:82.4% | | | • Cannot identify initials/acronyms<br>• Cannot handle cross-sentence relations<br>• Dataset imbalance | end-to-end modelling. |
| [45] | TranformDDI: BioBERT Transformer Encoder + Shared Parameter Layer + Dynamic Pair Attention | Joint Multi-Task Learning (NER + RC) with Dynamic Loss Balancing | F1: 82.4% | DDI Extraction -2013 | DDI (Advice, Effect, Mechanism, Int, no_rel) | • Class imbalance.<br>• Int class is often misclassified as Effect.<br>• Cannot handle cross-sentence relations | • Improve Int class detection via dynamic loss balancing. |
| [46] | Multiple transformers (BERT, BioBERT, SciBERT, BART, RoBERTa, DeBERTa, DistilBERT, Electra, XLNet) | Supervised Transformer Fine-Tuning Across Multiple Architectures | DNER: F1: 98.2%<br>DDIC: F1: 82.4% | DDIExtraction 2013 | DNER, DDI classification. | • Large models are prone to overfitting.<br>• limited cross-dataset testing. | End-to-end systems require error mitigation between tasks. |

To complement the granular, per-study view in Table 2, Table 3 distills the evidence into a higher-level comparative summary that links model family to typical F1 ranges, computational cost, and clinical deployability. This second table makes the practical trade-offs explicit: classical ML methods tend to be low-cost and moderately deployable but sit within a narrower performance band; CNN/LSTM models offer a middle ground; transformers often achieve stronger results but at higher computational and operational cost; and graph/multimodal systems may deliver gains while further raising complexity and deployment friction. Notably, Table 3 also highlights why small/decoder-only LLM approaches (SLMs) can be attractive in applied settings: they may offer competitive performance with lower compute and higher deployability, even when their results are not uniformly superior across all benchmarks. Taken together, Table 2 supports detailed cross-study inspection, while Table 3 provides a decision-oriented synthesis of performance-versus-cost considerations for real-world biomedical DDI extraction and prediction systems.

**Table 3.** Comparative Summary of Model Type, Performance, and Computational Cost

| Model Type | Representative Examples | F1 Range (%) | Computational Cost | Clinical Deployability |
|---|---|---|---|---|
| Classical ML | SVM, Random Forest | 80- 83 | Low | Moderate |
| CNN/LSTM | CNN-DDI, ConvBLSTM | 84- 98 | Medium | Moderate |
| Transformers | BioBERT, PubMedBERT | 79 - 95 | High | Limited |
| Graph/Multimodal | MOF-DDI, KG-enhanced models | 84- 90 | Very High | Limited |
| SLMs | D3 (LLaMA/Mistral) | 80- 87 | Low | High |

### 3.1. Annual Publication Trend

A clear Increasing trend in DDI research is apparent in Figure 2 above. From 2020 to 2023, there was a steady, though low, rate of research, averaging 3 papers per year. 2025 clearly indicates a point of divergence, with a sharp increase to 10 papers, accounting for 36% of the entire corpus. This growth aligns with the emergence of small language models, signaling a shift in the field from traditional classification tasks to generative DDI prediction tasks. Figure 2 illustrates the number of publications per year from 2020 to 2025.

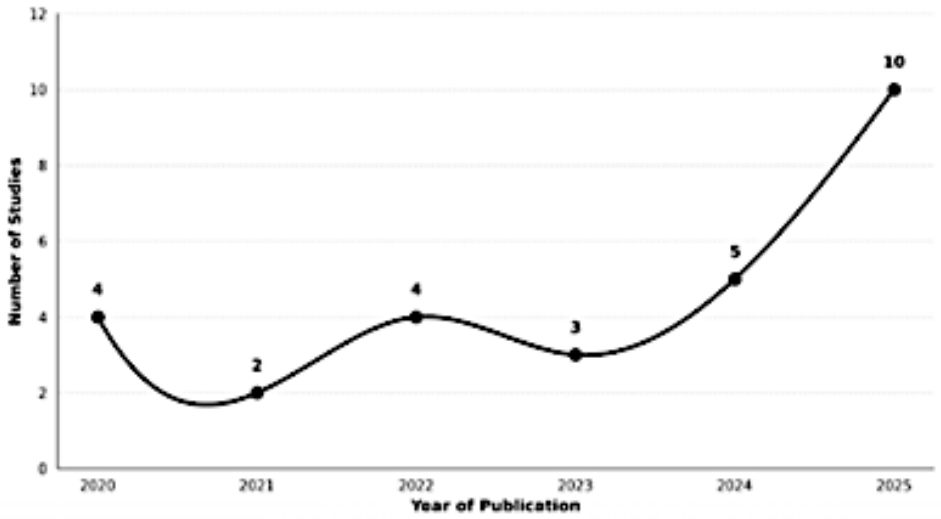


**Figure 2.** Annual Publication Trend

### 3.2. Geographic Distribution

Studies are concentrated in Asia (n=17, 61%), with nine publications from China. North America contributes 21% of the sources (n=6), including five from the United States, where electronic health record privacy is commonly noted. There are also contributions from the European continent (n=5), with a specific group from Greece (n=3) focused on pipeline optimization. In contrast, there is a lack of representation from the Global South, with no sources originating in Africa or South America, suggesting a geographic bias in the model's training database towards populations in the Global North and East Asia. Figure 3 summarizes the institutional affiliations of the 28 included studies, grouped by continental region. Figure 4 shows the categorization of the methodological architectures used in the 28 included studies.

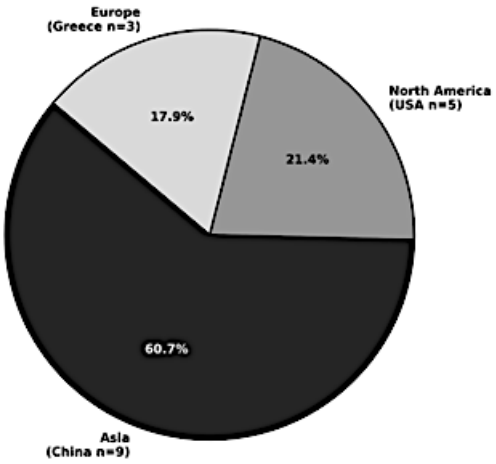


**Figure 3.** Geographic Distribution

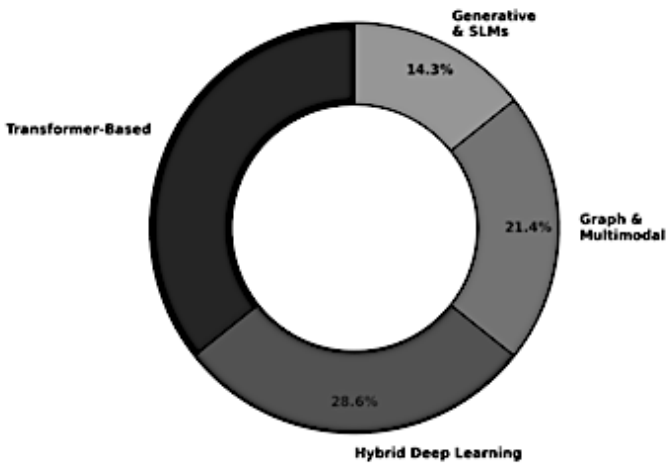


**Figure 4.** Distribution of Algorithms Used

### 3.3. Key Challenges Identified in Literature

Figure 5 shows that severe class imbalance dominates the challenge landscape (n=12; 42.9%), confirming it as the primary bottleneck driving instability in the minority class. High computational cost (n=6; 21.4%) and Int-class-deficient performance (n=6; 21.4%) form the second tier, highlighting the double burden of expensive architecture coupled with the under-representation of generic interaction cues. Lack of clarity on semantics is apparent through many Int–Effect misclassifications (n=5; 17.9%), while poor generalization, KG incompleteness, and cross-sentence limitations (each n=4; 14.3%) reveal representational gaps detrimental to model robustness. Relatively limited patient context (n=3; 10.7%) and limited explainability (n=3; 10.7%) highlight the disconnect between benchmark-optimized modelling and clinically valid decision support. Figure 5 quantifies the technical and clinical barriers explicitly Identified In the reviewed literature.

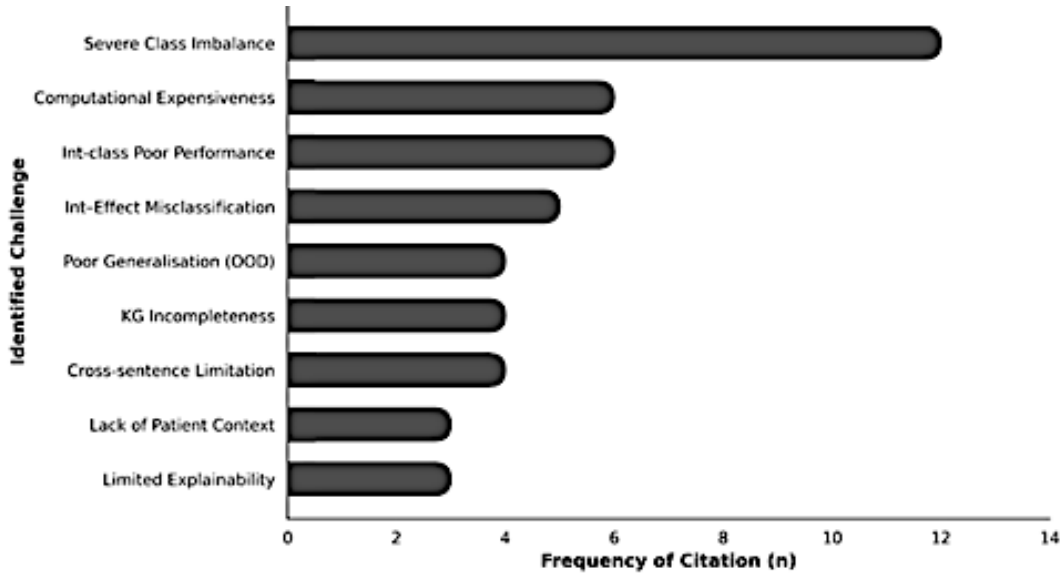


**Figure 5.** Key Challenges Identified in Literature

### 3.4. Evolution Of Model Architectures

The literature shows an overall dominance of hierarchically structured model families: transformer-based architecture (n=10), hybrid models (n=8), graph and multimodal systems (n=6), and generative or SLMs (n=4). The timeline can be divided into 3. Studies published between 2020 and 2021 primarily examined transformer fine-tuning. Work from 2022 to 2023 included hybrid, graph-based, and multimodal systems. Studies from 2024–2025 focus on parameter-efficient generative and SLM architectures, reflecting a focus

on lower computational cost. Figure 6 shows the evolution of algorithms from 2020 to 2025.

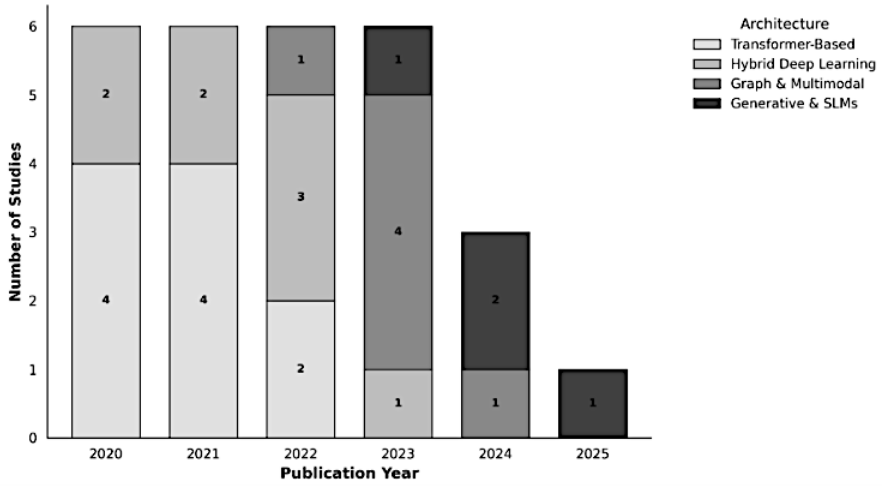


**Figure 6.** Evolution of Model Architectures

### 3.5. Dataset Utilization

Figure 7 shows that SemEval-2013 is the most frequently used dataset (n=19) in biomedical text–based DDI extraction studies. DrugBank (n=5) is the main alternative and is typically used in multi-class DDI prediction or multimodal fusion settings that require pharmacological attributes. TAC 2019 (n=2), TAC 2018 (n=1), TWOSIDES (n=1), and one multilingual corpus appear infrequently, suggesting limited use beyond established benchmarks. Figure 7 shows the distribution of datasets used for training and benchmarking models.

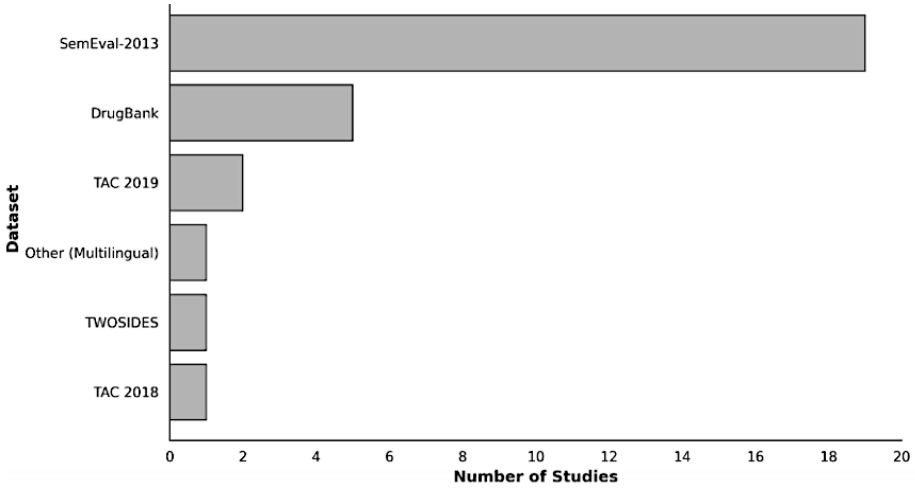


**Figure 7.** Dataset Utilization

### 3.6. Mapping Challenges to Technical Solutions

Transformer fine-tuning targets class imbalance (n=4) through reweighting and focal-loss variants. Hybrid Deep Learning models address both imbalance and syntactic complexity (n=3 each) by incorporating BLSTM/CNN layers that enhance long-range dependency tracking. Graph and multimodal systems dominate solutions for context and knowledge sparsity (n=4), leveraging KGs and molecular structure to compensate for missing biomedical signals. Generative AI and SLMs focus on minority-class augmentation and computational efficiency (n=5 and n=3), enabling performance gains under resource constraints. Figure 8 maps how the 28 studies align architectural solutions with DDI extraction challenges, revealing tightly coupled patterns between limitations and methodological choices.

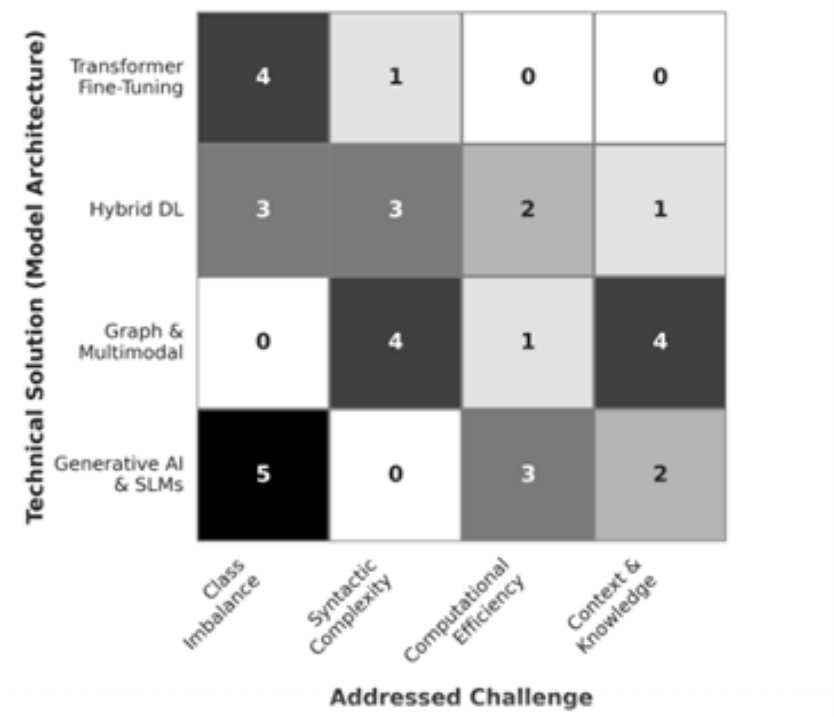


**Figure 8:** Mapping Challenges to Technical Solutions

### 3.7. Discussion

This discussion addresses the four research questions formulated in the Introduction. Taken together, the findings indicate that the central bottleneck in biomedical text–based DDI extraction is not model expressiveness, but representational adequacy under structural constraints such as class imbalance, semantic overlap, and limited contextual

grounding. Across architectures, performance gains plateau unless these foundational issues are addressed, suggesting a paradigm shift from scaling models to refining task-aligned representations.

**RQ1: Which small or transformer-based language models and fine-tuning strategies have been used for biomedical text–driven drug–drug interaction extraction and prediction?**

This review sought to identify and analyse small or transformer-based language models and fine-tuning strategies used to support biomedical text–driven DDI extraction and prediction. Table 2 summarizes model architectures and fine-tuning approaches across the 28 included studies, while Table 3 groups these methods into five architectural categories. Across the literature, transformer architectures, particularly BERT variants, remain the most widely adopted.

1) Transformer-Based Models

Transformer-based models form the primary backbone for DDI extraction, with BioBERT, PubMedBERT, and SciBERT widely adopted (Table 2) due to their multi-head self-attention mechanisms [22], [28], [31], [35], [37], [38], [43], [46]. These models are typically fine-tuned using supervised learning with AdamW optimization and learning rates ranging from 2e-5 to 5e-5, yielding F1-scores between 79% and 95% (Tables 2 and 3). However, [31], [32] show that transformers perform poorly in distinguishing semantically similar DDI types (Int–Effect) due to limited entity-specific representation, and reduced recall on minority classes despite balanced optimisation [28], [38]. These weaknesses further encourage the study of hybrid, graph-enhanced, or generative models.

2) Hybrid Deep Learning Architectures

Table 2 lists hybrid architectures that combine transformers with complementary neural components to address contextual and structural limitations of standalone models. CNN layers can improve local pattern detection [21], [26], BiLSTM/LSTM components model sequential dependencies and temporal ordering [22], [27], [30], [32], and Graph Attention Networks emphasize relation salience through syntactic dependency weighting [27]. These configurations improve confusion between interaction classes, particularly Int and Effect, and often report higher macro-F1 scores than transformer-only baselines [26].

However, they incur higher computational demands (Table 3, medium cost) and show sensitivity to parsing errors [34], [44].

3) Graph-Based and Multimodal Models

Graph-enhanced and multimodal architectures extend DDI extraction by incorporating external biomedical knowledge, including KG embeddings, molecular graphs, drug descriptions, pathway knowledge, and document-level context, forming richer interaction outputs [23], [24], [25], [33], [41], [42]. These approaches combine transformer encoders with graph convolutional networks [25], [41], [42], subgraph-based reasoning [42], or optimal transport fusion mechanisms [25] to model interactions beyond sentence-level textual evidence. Knowledge graph embeddings derived from DrugBank, ChemProt, and Hetionet are used alongside textual representations [25], [33], with reported F1-scores ranging from 84% to 90% (Table 3), including 89.04% for MOF-DDI [25]. At the same time, these models incur very high computational costs (Table 3), largely due to graph construction, multi-source integration, and complex fusion pipelines. Reported limitations also include incomplete knowledge graphs and missing entity coverage [25], [33].

4) Generative and Small Language Models (SLMs)

Study by [19], [24], [29], [35], [39], [47] identify generative and SLMs as the emerging evolution (2024-2025) methods in DDI extraction. SLMs leverage decoder-only architectures (D3 with LLaMA/Mistral [19]), GAN-based augmentation for minority-class synthesis (BIOGAN-BERT [29], BioFocal-DDI [39]), instruction-tuned models (Galactica-1.3B [24]), parameter-efficient fine-tuning methods (LoRA-BiomedBERT [35]), and few-shot learning frameworks (LLaMA 3.2, Phi-3.5, Gemma 2 [40]). As shown in Table 3, SLM-based methods report F1 scores of 80–87% with low computational cost, supporting their use in resource-constrained clinical settings. GAN-based models increase recall for minority interaction types, particularly Int and Advice, through synthetic data generation [29], [39], while parameter-efficient approaches substantially reduce training overhead [24], [35]. However, reliance on prompt formulation, instruction design, and pseudo-label quality contributes to performance variability [24], [29], [40], and generalization beyond training distributions remains limited [24]. Overall, these findings indicate a shift toward lighter and more scalable DDI modelling approaches making them promising candidates for further clinical validation and deployment research.

Transformer architectures remain the most prevalent approach in DDI extraction research, while more recent SLM-based methods report comparable predictive performance with substantially lower computational cost. As shown in Table 3, SLMs combine F1 scores in the range of 80–87% with low computational overhead, distinguishing them from other architectural families in terms of efficiency and deployment suitability for pharmacovigilance applications.

**RQ2: Which biomedical text sources and benchmark datasets are used to train and evaluate these models?**

Across the 28 included studies, biomedical text sources can be grouped into curated benchmark datasets used for supervised evaluation and unstructured biomedical corpora used as supplementary sources of linguistic and pharmacological information. As shown in Table 2, SemEval-2013/ DDIExtraction-2013 is the most commonly used dataset, followed by DrugBank.

1) The SemEval DDIExtraction 2013

Table 2 indicates that 21 studies use SemEval-2013 as the primary evaluation benchmark[20], [21], [22], [23], [27], [28], [29], [30], [31], [32], [33], [34], [37], [38], [39], [41], [42], [43], [44], [45], [46]. The corpus contains 18,491 annotated entity pairs across five relation types and exhibits pronounced class imbalance (Int: 188 samples; Negative: 14,471 samples). Reported results show higher performance for the majority classes (Mechanism, Effect, Advice), alongside recurrent degradation for the Int class [28], [32], [34]. In addition, the sentence-level annotation limits assessment of cross-sentence interactions, long-range dependencies, and document-level structure relevant to clinical narratives.

2) TAC 2018/2019

Two studies evaluated model performance on TAC corpora [29], [30]. The TAC datasets contain document-level clinical narratives with extended contextual dependencies and co-referential mentions. As reported in Table 2, the study [30] observed a decline in performance when moving from SemEval (F1: 83.32%) to TAC 2019 (F1: 80.23%) and TAC 2018 (F1: 60.53%). These results indicate reduced robustness of transformer-based models when applied to longer and less constrained clinical contexts. In this respect, TAC datasets provide a useful benchmark for assessing generalization beyond sentence-level corpora.

3) DrugBank

[19], [24], [25], [33], [35], [36] use DrugBank as both a text corpus and a source of structured knowledge. In studies [19], [35] and [36], textual drug descriptions are used for classification, whereas [25] and [33] integrate DrugBank's knowledge graph through graph-based embeddings. As shown in Table 2, models incorporating KG information report improved performance for Mechanism classification, but also exhibit sensitivity to knowledge graph incompleteness and sparse entity coverage [25], [33]. In addition, the formal terminology used in DrugBank differs from that in clinical narratives, limiting transferability to unstructured EHR text.

4) TWOSIDES and Auxiliary Corpora

Study [25] incorporates TWOSIDES pharmacovigilance data. As indicated in Table 2, [25] combines DrugBank with TWOSIDES signals and PubMed abstracts to increase sensitivity to rare interactions, but at the cost of reduced precision due to noise and inconsistent causal directionality. No other included studies draw on real-world pharmacovigilance databases, despite their relevance to clinical practice.

5) Electronic Health Records Gap

Table 2 indicates that only one study [40] evaluates models using EHR data, with a focus on entity recognition rather than DDI classification. Study [26] similarly notes EHR-based evaluation as an important direction for future work. Across the remaining studies, patient-specific factors such as age, comorbidities, and medication history are not incorporated. As a result, models developed on SemEval and DrugBank remain limited in their ability to support patient-contextualized reasoning needed for clinical deployment.

SemEval-2013 is the most frequently used benchmark, but its scope is limited to sentence-level interactions. Evaluation on TAC datasets highlights reduced generalization in document-level contexts, while DrugBank contributes structured pharmacological knowledge that reflects formal documentation rather than clinical narrative text. The lack of patient-level EHR evaluation remains a major limitation, hindering progress toward context-aware, personalized DDI systems suitable for real-world clinical use.

**RQ3: How do small-language-model-based approaches perform, in terms of predictive accuracy and efficiency, relative to larger transformer or graph-based models?**

Table 3 presents a performance comparison across the five architectural families, contrasting predictive accuracy with computational efficiency. Large transformer models attain the highest reported performance (F1 up to 95%), while small language models achieve results comparable to theirs with substantially fewer parameters.

1) Classical and Non-Transformer Deep Learning Models

Table 3 reports F1 scores for classical machine learning and CNN/LSTM architectures, ranging from 80–98%. The higher values (96–98%) are associated with language-specific hierarchical features reported in [26], rather than typical model performance. Standard implementations generally achieve F1 scores between 80% and 87% [20], [21], with the CNN-DDI model in [21] reaching 83.81%. As indicated in Table 2, these approaches frequently misclassify minority interaction classes, particularly Int and Effect, due to limited sample sizes and constrained semantic modeling [32], [33], [39]. In comparison, SLM-based methods report both higher accuracy and lower computational cost; for example, D3 [19] achieved an F1 score of 85.8% with minimal parameter overhead, exceeding traditional baselines without extensive feature engineering.

2) Relative to Large Transformer Models

Table 3 reports transformer-based models achieving F1 scores between 79% and 95%, with BioBERT and PubMedBERT among the most frequently used architectures [28], [31], [34], [38]. Bio-RoBERTa with tree-transformers attained the highest reported performance (F1: 95.2%) [22]. This level of performance is associated with high computational cost, classified as “High” in Table 3, reflecting GPU memory demands, training requirements, and overall computational overhead [22]. In contrast, SLM-based approaches report comparable performance at substantially lower cost. As summarized in Table 2, D3 (LLaMA-based) [19] achieved an F1 score of 85.8%, BioFocal-DDI [39]reached 86.64%, and BIOGAN-BERT [29] achieved 85.0%, all with “Low” computational overhead (Table 3). LoRA-BiomedBERT [35] reported an F1 score of 79.75% using parameter-efficient fine-tuning with reduced training requirements. Relative to peak transformer performance, these results correspond to approximately 5–10% lower performance, alongside markedly lower computational demands.

3) Graph-Enhanced and Multimodal Models

Table 3 reports that graph-based and multimodal models achieve F1 scores between 84% and 90%, with MOF-DDI [25] reporting an F1 score of 89.04%. These approaches integrate structured knowledge graphs and molecular features alongside textual representations [25], [33], [42]. As indicated in Table 3, such models incur very high computational costs, reflecting the requirements of knowledge graph construction, multi-source integration, and complex fusion pipelines. In addition, Table 2 notes knowledge graph incompleteness and missing entity coverage as recurring factors that limit generalization [25], [33]. By comparison, SLM-based approaches [19], [39] report F1 scores in the range of 85–87% using labelled text alone, without reliance on dedicated knowledge graph infrastructure or molecular feature engineering.

4) Performance-Efficiency Trade-Off Analysis

As shown in Table 3, there is a non-linear association between computational cost and predictive performance. Graph-based and multimodal models are characterized by very high computational cost, with reported F1 gains of approximately 1–5% relative to transformer baselines. In contrast, SLM-based approaches report F1 scores that are approximately 5–10% lower than peak transformer performance while operating at substantially reduced computational cost. As summarized in Table 3, SLMs achieve F1 scores of 80–87% with low computational overhead, distinguishing them from other architectural families in terms of efficiency and deployment suitability. Transformer models achieve higher peak performance but require GPU resources that may limit use in resource-constrained clinical settings. SLMs, by comparison, support inference on more modest hardware while maintaining performance levels relevant to clinical decision support.

SLMs report predictive accuracy comparable to large transformer models (with an approximate 5–10% F1 difference) and to graph-based systems, while operating with substantially lower computational overhead. As summarized in Table 3, these characteristics make SLMs well-suited for scalable DDI extraction in resource-constrained clinical settings, where limited GPU availability and high infrastructure costs restrict the use of large-scale models.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

**RQ4: What are the challenges identified in implementing small language models for scalable and potentially personalized DDI prediction?**

Table 2 summarizes the technical challenges reported across the 28 included studies. Class imbalance is the most frequently cited issue (n=16 studies), followed by underperformance on the Int class (n=8).

1) INT Class Poor Performance

Several studies report Int-class collapse beyond general class imbalance [28], [30], [31], [32], [33], [34], [42], [45]. Int interactions show substantial semantic overlap with Effect types, differing largely in specificity rather than in core relational meaning. Studies [32], [33], [45] describe consistent Int-to-Effect misclassification, with Int-class recall frequently falling below 50% despite overall F1 scores above 80%. This pattern reflects both limited sample availability and ambiguity in sentences such as "Drug A may interact with Drug B," which lack mechanistic or outcome information needed to distinguish Mechanism and Effect relations. Methods including entity-aware attention [34] and triplet-loss–based metric learning [43] provide modest improvements, but reported results indicate that annotation inconsistency and linguistic vagueness remain limiting factors [31], [33].

2) Int–Effect Misclassification

Systematic confusion between classes Int and Effect is observed in [30], [33], [34], [42], [45], due to semantic overlap and lack of minority class examples, causing models to overgeneralise towards effect-dominant predictions. Confusion-matrices, as represented in [29], [33] validate the frequent occurrences of misclassifications, thus obstructing interpretations. Consequently, the mislabelled interactions can underestimate or overestimate the degree of perceived clinical severity, reducing the strength of clinical decision support. Countermeasures were proposed using focal-loss sensitivity tuning [33], synthetic minority enrichment, and multi-view entity sentence alignment [23], which have shown reduced cross-class drift and improved fine-grained discrimination.

3) Class Imbalance

Class imbalance is a dominant structural challenge across DDI research, [19], [23], [28], [29], [30], [31], [32], [33], [34], [35], [36], [39], [41], [42], [43], [45]. The class distribution in SemEval-2013 (Negative: 14,471 samples vs Int: 188 samples) introduces a bias toward majority

classes, with studies reporting high precision for Mechanism, Effect, and Advice relations but substantially reduced recall for Int interactions [28], [30], [32]. Mitigation strategies such as GAN-based data augmentation [29] and focal-loss reweighting [39] yield modest performance gains (approximately 2–5% F1) but do not fully address representational challenges associated with sparse semantic contexts. As summarized in Table 2, class imbalance is observed across all architectural families, indicating a limitation of the underlying data rather than of specific modelling approaches.

4) Computational Cost

[22], [27], [29] identify computational demands as barriers to deployment. Transformer-based models require substantial GPU resources and extended training times, which may limit suitability for clinical environments [22]. Graph-enhanced and multimodal architectures further increase computational requirements through knowledge graph construction, multi-source alignment, and complex fusion mechanisms [25], [41]. As summarized in Table 3, these approaches are classified as having very high computational cost, which confines their use largely to resource-intensive research settings. In contrast, SLM-based approaches reduce computational burden through parameter-efficient fine-tuning [35] and decoder-only architectures [19], while maintaining competitive performance levels compatible with clinical deployment.

5) Knowledge-Graph Incompleteness

Graph-augmented systems frequently confront incomplete or sparsely connected biomedical knowledge graphs. [25] reports inconsistent KG coverage that weakens inference for poorly represented drugs, while [42] shows that sparsely connected subgraphs provide insufficient relational evidence, producing unstable predictions for rare interaction pairs. [21] similarly confirms that missing or weak pharmacological links restrict embedding propagation and degrade mechanistic reasoning. These structural gaps suppress recall for under-documented interactions and undermine mechanistic inference, limiting clinical interpretability. Mitigation strategies include multi-source KG enrichment and subgraph-level evidence propagation [42] and hybrid text–KG fusion [25] improve robustness under sparse conditions, but cannot fully compensate for structurally incomplete biomedical graphs..

6) Poor Generalisation / Domain Shift

Poor generalization under distributional shift appears in only a small subset of studies, but remains a meaningful barrier to real-world deployment. [40] shows accuracy degradation when benchmark-trained models are applied to heterogeneous biomedical corpora, attributing failures to terminology drift and fragmented contextual structures. [25] similarly reports instability across datasets with divergent linguistic profiles, noting that text-only encoders fail to transfer without auxiliary structural information. Evidence from [37] shows that abbreviations and atypical syntax in clinical text impair model performance. This variability reduces the recovery of context-dependent interactions and weakens pharmacovigilance reliability. Approaches reported in [40] apply domain-adaptive fine-tuning and multi-corpus training, while [25] uses structural fusion that combines textual input with molecular knowledge graph representations.

7) Limited Explainability

Table 2 indicates limited attention to explainability across the included studies, despite its relevance to clinical adoption. Study [23] explicitly discusses model opacity, reporting that multimodal fusion architectures operate as "black boxes," which may limit clinician trust. Most DDI systems report categorical predictions without providing mechanistic justification, uncertainty estimates, or actionable mitigation guidance. Such limitations are misaligned with regulatory expectations for medical AI and with clinical requirements for interpretable decision support. Although attention-based architectures [34], [36], [41] model syntactic dependencies, they do not provide systematic methods for deriving human-interpretable explanations of causal reasoning or underlying pharmacological mechanisms.

8) Patient-Specific Context

EHR integration and patient-specific evaluation are identified as important directions for future research [26], [35], [40]. Existing models typically predict DDIs at the pair level, without accounting for patient factors such as age, renal function, hepatic metabolism, comorbidities, or broader polypharmacy context that influence clinical severity. Study [26] highlights the need for EHR-based evaluation to support context-aware prediction, while [35] notes real-time EHR integration as a prerequisite for deployment. Datasets such as DrugBank and SemEval provide interaction labels without patient stratification, limiting their applicability to individualized risk assessment. In clinical practice, interaction

severity may vary across patient groups, with nominally minor interactions becoming clinically significant in vulnerable populations. As summarized in Table 2, no reviewed studies incorporate patient-level temporal reasoning, medication history information, or evolving polypharmacy patterns that are central to real-world pharmacovigilance.

Class imbalance is the most frequently reported challenge in the literature, with underperformance in the Int class representing its most pronounced effect. High computational cost limits the deployment of top-performing architectures, whereas SLM-based approaches remain compatible with resource-constrained environments. At the same time, the lack of explainability mechanisms and patient-specific contextualization continues to limit clinical applicability. Addressing these issues will require not only architectural advances but also changes in dataset design, evaluation practices, and closer integration with EHR-based pharmacovigilance systems.

The findings of this review indicate a shift in biomedical text–based DDI extraction away from increasing model scale and toward improved representational efficiency and task alignment. Although large transformer and multimodal systems achieve strong benchmark performance, their use is limited by computational demands, interpretability constraints, and infrastructure requirements. Small language models, by comparison, exhibit a more balanced trade-off between predictive accuracy and computational efficiency, supporting their suitability for scalable pharmacovigilance applications. Nonetheless, unresolved challenges, including class imbalance, semantic ambiguity, and limited patient-specific context, continue to restrict translation to real-world clinical settings.

In interpreting these findings, it is important to acknowledge several methodological and empirical limitations of the reviewed studies, which are discussed in the subsequent section. These limitations directly inform the practical and theoretical implications of SLM-based DDI extraction and motivate the future research directions outlined thereafter

This systematic literature review was limited to four databases (PubMed, IEEE Xplore, ACM Digital Library, and SpringerLink), which may have excluded relevant studies indexed in other sources. Only English-language publications were considered, thereby omitting

research reported in other languages. The selected time window (2020–2025) restricts insight into longer-term methodological developments. In addition, the review focused on text-based DDI extraction and did not comprehensively analyze multimodal approaches that integrate molecular or pharmacokinetic information. Finally, grey literature, including preprints, technical reports, and theses, was excluded to maintain peer-review standards, which may have resulted in the omission of emerging methods or unpublished negative findings.

This study carries important theoretical, practical, and methodological implications for clinical natural language processing and pharmacovigilance research. From a theoretical perspective, the synthesis extends existing DDI extraction literature by shifting attention from accuracy-centric model evaluation toward a broader consideration of efficiency, deployability, and system-level trade-offs, particularly in the context of small language models. In practice, the findings indicate that SLM-based architectures can offer a viable balance between predictive performance and computational cost, supporting their consideration in resource-constrained clinical and pharmacovigilance settings where large language models may be impractical. Methodologically, the review highlights the need for more consistent evaluation practices, including transparent reporting of computational requirements and cross-dataset generalizability, to enable meaningful comparison across studies. Importantly, the relevance of SLM-based approaches is not confined to current infrastructure limitations but also extends to scenarios that require greater transparency, controllability, and domain-specific optimization, even as large language models continue to evolve. Collectively, these implications position SLM-focused DDI extraction research as a sustainable and adaptable direction for future clinical NLP development rather than a transient response to present-day resource constraints.

Future research should extend this review by advancing evaluation practices for DDI extraction beyond performance-centric metrics, incorporating standardized reporting of computational cost, deployment constraints, and cross-dataset generalizability. In particular, systematic investigation of SLM robustness across clinical subdomains, languages, and institutional corpora would provide deeper insight into real-world applicability. Further work is also needed to examine architectural design choices that shape the performance–efficiency trade-off in SLM-based models, independent of model

scale. Beyond technical evaluation, future studies should explore human-in-the-loop validation frameworks involving clinicians or pharmacovigilance experts to assess trust, interpretability, and decision-support value. Finally, research into explainability techniques tailored to SLM-based DDI extraction remains essential to support transparent, accountable, and clinically meaningful deployment.

## 4. CONCLUSION

This systematic review examined the emerging role of small language models in extracting drug–drug interactions from biomedical text, with a particular focus on performance, efficiency, and deployability in clinical and pharmacovigilance contexts. The synthesis indicates that SLM-based approaches can achieve competitive predictive performance on benchmark datasets while offering substantially lower computational overhead than large language models, highlighting their relevance for settings with non-trivial infrastructure, cost, and transparency constraints. At the same time, the review reveals considerable variability across datasets, architectures, and evaluation practices, underscoring the need for more consistent and deployment-aware assessment frameworks.

Beyond summarising existing evidence, this study positions SLM-focused DDI extraction research as a sustainable methodological direction rather than a transient response to current resource limitations. By consolidating architectural trends, evaluation trade-offs, and deployment considerations, the review provides a reference baseline for future comparative studies and systematic benchmarking efforts in clinical natural language processing. As the field continues to evolve, the findings emphasise that efficiency, controllability, and domain-specific optimisation will remain central considerations alongside raw predictive performance.

## REFERENCES

[1] T. N. G. de Andrade Santos *et al*, "Prevalence of clinically manifested drug interactions in hospitalized patients: A systematic review and meta-analysis," *PLoS One*, vol. 15, no. 7, 2020, doi: 10.1371/journal.pone.0235353.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[2] J. E. Hughes, C. Waldron, K. E. Bennett, and C. Cahir, "Prevalence of Drug–Drug Interactions in Older Community-Dwelling Individuals: A Systematic Review and Meta-analysis," *Drugs and Aging*, vol. 40, no. 2, pp. 117–134, 2023, doi: 10.1007/s40266-022-01001-5.

[3] X. Li, Z. Xiong, W. Zhang, and S. Liu, "Deep learning for drug-drug interaction prediction: A comprehensive review," *Quant. Biol.*, vol. 12, no. 1, pp. 30–52, Mar. 2024, doi: 10.1002/qub2.32.

[4] Y. Xia, A. Xiong, Z. Zhang, Q. Zou, and F. Cui, "A comprehensive review of deep learning-based approaches for drug-drug interaction prediction," *Brief. Funct. Genomics*, vol. 24, 2025, doi: 10.1093/bfgp/elae052.

[5] H. Luo *et al.*, "Drug-drug interactions prediction based on deep learning and knowledge graph: A review," *iScience*, vol. 27, no. 3, p. 109148, 2024, doi: 10.1016/j.isci.2024.109148.

[6] F.-I. Gheorghita, V.-I. Bocanet, and L. B. Iantovics, "Machine learning-based drug-drug interaction prediction: a critical review of models, limitations, and data challenges," *Front. Pharmacol.*, vol. 16, no. July, pp. 1–25, 2025, doi: 10.3389/fphar.2025.1632775.

[7] G. L. Swathi Mirthika and B. Sivakumar, "A Systematic Review on Drug Interaction Prediction Using Various Methods to Reduce Adverse Effects," *Int. J. Inf. Syst. Soc. Chang.*, vol. 14, no. 1, pp. 1–8, 2023, doi: 10.4018/ijissc.329233.

[8] K. Soni Sharmila, S. Thanga Revathi, and P. Kiran Sree, "A Systematic Review on Drug-to-Drug Interaction Prediction and Cryptographic Mechanism for Secure Drug Discovery Using AI Techniques," *Int. J. Artif. Intell. Tools*, vol. 33, no. 08, p. 2450003, Jul. 2024, doi: 10.1142/S0218213024500039.

[9] R. Huang, W. Xu, X. Qi, and Y. Wang, "Protein-Protein Interaction Extraction based on Multi-Feature Fusion and Entity Enhancement," in *Proceedings of the 4th International Conference on Biomedical and Intelligent Systems*, in IC-BIS '25. New York, NY, USA: Association for Computing Machinery, 2025, pp. 229–234. doi: 10.1145/3745034.3745070.

[10] C. Yu *et al.*, "Drug–drug interaction extraction based on multimodal feature fusion by Transformer and BiGRU," *Front. Drug Discov.*, vol. 4, no. October, pp. 1–12, 2024, doi: 10.3389/fddsv.2024.1460672.

[11] H. Mondal *et al.*, "A systematic mapping review on the capability of large language models in drug-drug interaction analysis," *Expert Rev. Clin. Pharmacol.*, vol. 18, no. 9, pp. 683–690, Sep. 2025, doi: 10.1080/17512433.2025.2568090.

[12] F. Busch *et al.*, "Current applications and challenges in large language models for patient care: a systematic review," *Commun. Med.*, vol. 5, no. 1, Dec. 2025, doi: 10.1038/s43856-024-00717-2.

[13] X. H. Liu, Z. H. Lu, T. Wang, and F. Liu, "Large language models facilitating modern molecular biology and novel drug development," 2024, *Frontiers Media SA*. doi: 10.3389/fphar.2024.1458739.

[14] N. H. M. S. E. S. M. Mysara, "A comprehensive landscape of AI applications in broad-spectrum drug interaction prediction: a systematic review," *J. Cheminform.*, 2025, doi: 10.1186/s13321-025-01093-2.

[15] A. Z. Al Meslamani and A. Abou Hajal, "Language models for drug-drug interactions: current applications, pitfalls, and future directions.," *Expert Opin. Drug Metab. Toxicol.*, vol. 21, no. 9, pp. 1083–1102, Sep. 2025, doi: 10.1080/17425255.2025.2551724.

[16] A. Liberati *et al.*, "The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration.," *BMJ*, vol. 339, 2009, doi: 10.1136/bmj.b2700.

[17] S. E. Group, "Guidelines for performing Systematic Literature Reviews in Software Engineering," 2007.

[18] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *BMJ*, vol. 339, no. 7716, pp. 332–336, 2009, doi: 10.1136/bmj.b2535.

[19] A. Ibrahim, A. Hosseini, S. Ibrahim, A. Sattar, and A. Serag, "D3: A Small Language Model for Drug-Drug Interaction prediction and comparison with Large Language Models," *Mach. Learn. with Appl.*, vol. 20, no. April, p. 100658, 2025, doi: 10.1016/j.mlwa.2025.100658.

[20] J. Machado, C. Rodrigues, R. Sousa, and L. M. Gomes, "Drug–drug interaction extraction-based system: An natural language processing approach," *Expert Syst.*, vol. 42, no. 1, pp. 1–12, 2025, doi: 10.1111/exsy.13303.

[21] M. T. Tahir, M. Ibrahim, N. Sarwar, A. Irshad, and G. Atteia, "Enhanced drug-drug interaction extraction from biomedical text using deep learning-based sentence representations," *Sci. Rep.*, vol. 15, no. 1, pp. 1–15, 2025, doi: 10.1038/s41598-025-21782-0.

[22] S. S. Roy and R. E. Mercer, "Extracting Drug-Drug and Protein-Protein Interactions from Text Using a Continuous Update of Tree-Transformers," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, pp. 280–291, 2023, doi: 10.18653/v1/2023.bionlp-1.25.

[23] B. H. Tran, H. M. Hoang, B. N. Nguyen, D. C. Can, and H. Q. Le, "A multifaceted approach to drug–drug interaction extraction with fusion strategies," *J. Biomed. Inform.*, vol. 169, no. May, p. 104874, 2025, doi: 10.1016/j.jbi.2025.104874.

[24] L. Xiao, X. Wang, Z. Zhang, Y. Yao, and W. Zhu, "DyNAS-DDI: Dynamic Pairwise Architecture Search for Generalizable Drug-Drug Interaction LLM," in *Proceedings of the 33rd ACM International Conference on Multimedia*, in MM '25. New York, NY, USA: Association for Computing Machinery, 2025, pp. 2216–2225. doi: 10.1145/3746027.3755791.

[25] Q. Wen, J. Li, C. Zhang, and Y. Ye, "A Multi-Modality Framework for Drug-Drug Interaction Prediction by Harnessing Multi-source Data," in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, in CIKM '23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 2696–2705. doi: 10.1145/3583780.3614765.

[26] M. S. Malik, S. Jawad, S. A. Moqurrab, and G. Srivastava, "DeepMedFeature: An Accurate Feature Extraction and Drug-Drug Interaction Model for Clinical Text in Medical Informatics," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, Mar. 2024, doi: 10.1145/3651159.

[27] Y. Jia, Z. Yuan, H. Wang, Y. Gong, H. Yang, and Z.-L. Xiang, "BBL-GAT: A Novel Method for Drug-Drug Interaction Extraction From Biomedical Literature," *IEEE Access*, vol. 12, pp. 134167–134184, 2024, doi: 10.1109/ACCESS.2024.3462101.

[28] T. T. Datta, P. C. Shill, and Z. Al Nazi, "BERT-D2: Drug-Drug Interaction Extraction using BERT," in *2022 International Conference for Advancement in Technology (ICONAT)*, 2022, pp. 1–6. doi: 10.1109/ICONAT53423.2022.9725979.

[29] M. A. Parameswara and R. Mandala, ""BIOGAN-BERT: BioGPT-2 Fine Tuned and GAN-BERT for Extracting Drug Interaction Based on Biomedical Texts.," in *2024 11th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA)*, 2024, pp. 1–6. doi: 10.1109/ICAICTA63815.2024.10762980.

[30] M. Kafikang and A. Hendawi, "Drug-Drug Interaction Extraction from Biomedical Text using Relation BioBERT with BLSTM," *MAKE*, vol. 5, no. 2, Jun. 2023, doi: 10.1101/2022.08.31.506076.

[31] D. P. Nguyen and T. Ho, "Drug-Drug Interaction Extraction from Biomedical Texts via Relation BERT," *2020 RIVF Int. Conf. Comput. Commun. Technol.*, vol. null, pp. 1–7, 2020, doi: 10.1109/RIVF48685.2020.9140783.

[32] A. Wen, X. Sun, K. Yu, Y. Wu, J. Zhang, and Z. Yuan, "Drug-Drug Interaction Extraction using Pre-training Model of Enhanced Entity Information," in *2020 International Conferences on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics)*, 2020, pp. 527–532. doi: 10.1109/iThings-GreenCom-CPSCom-SmartData-Cybermatics50389.2020.00094.

[33] X. Jin, X. Sun, J. Chen, and R. Sutcliffe, "Extracting Drug-drug Interactions from Biomedical Texts using Knowledge Graph Embeddings and Multi-focal Loss," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, in CIKM '22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 884–893. doi: 10.1145/3511808.3557318.

[34] Y. Zhu, L. Li, H. Lu, A. Zhou, and X. Qin, "Extracting drug-drug interactions from texts with BioBERT and multiple entity-aware attentions," *J. Biomed. Inform.*, vol. null, p. 103451, 2020, doi: 10.1016/j.jbi.2020.103451.

[35] I.-F. Gheorghita, V. Bocăneț, and L. B. Iantovics, "Fine-Tuning BiomedBERT with LoRA and Pseudo-Labeling for Accurate Drug–Drug Interactions Classification," *Appl. Sci.*, vol. null, p. null, 2025, doi: 10.3390/app15158653.

[36] S. Jung and S. Yoo, "Interpretable prediction of drug-drug interactions via text embedding in biomedical literature.," *Comput. Biol. Med.*, vol. 185, p. 109496, Feb. 2025, doi: 10.1016/j.compbiomed.2024.109496.

[37] P. Su and K. Vijay-Shanker, "Investigation of BERT Model on Biomedical Relation Extraction Based on Revised Fine-tuning Mechanism," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2020, pp. 2522–2529. doi: 10.1109/BIBM49941.2020.9313160.

[38] P. S. Vijay-Shanker, "Investigation of improving the pre-training and fine-tuning of BERT model for biomedical relation extraction," *BMC Bioinformatics*, 2022, doi: 10.1186/s12859-022-04642-w.

[39] Zhu Yuan, Shuailiang Zhang, Huiyun Zhang, A. Ping Xie, and Yaxun Jia, "Optimized Drug-Drug Interaction Extraction With BioGPT and Focal Loss-Based Attention," *IEEE J. Biomed. Heal. Informatics*, vol. null, no. 6, p. null, Jun. 2025, doi: 10.1109/JBHI.2025.3540861.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[40] J. Podichetty, S. C. Ma, and K. Romero, "Optimizing Drug Development : Assessing Small Language Models for Efficient Drug-Drug Interaction Data Extraction," p. 881, 2024.

[41] S. Z. Y. Zhang, "SCATrans: semantic cross-attention transformer for drug–drug interaction predication through multimodal biomedical data," *BMC Bioinformatics*, 2025, doi: 10.1186/s12859-025-06165-6.

[42] Y. Shi, M. He, J. Chen, F. Han, Y. C. Id, and Y. Cai, "SubGE-DDI: A new prediction model for drug-drug interaction established through biomedical texts and drug-pairs knowledge subgraph enhancement," *PLOS Comput. Biol.*, vol. 20, p. null, 2024, doi: 10.1371/journal.pcbi.1011989.

[43] Z. Zhao *et al.*, "TL-BERT: A Novel Biomedical Relation Extraction Approach," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2021, pp. 797–800. doi: 10.1109/BIBM52615.2021.9669869.

[44] D. Zaikis and I. Vlahavas, "TP-DDI: Transformer-based pipeline for the extraction of Drug-Drug Interactions.," *Artif. Intell. Med.*, vol. 119, p. 102153, Sep. 2021, doi: 10.1016/j.artmed.2021.102153.

[45] D. Zaikis and I. Vlahavas, "TransformDDI: The Transformer-Based Joint Multi-Task Model for End-to-End Drug-Drug Interaction Extraction," *IEEE J. Biomed. Heal. Informatics*, vol. 29, no. 4, pp. 3045–3056, 2025, doi: 10.1109/JBHI.2024.3507738.

[46] D. Zaikis, S. Kokkas, and I. Vlahavas, "Transforming Drug-Drug Interaction Extraction from Biomedical Literature," in *Proceedings of the 12th Hellenic Conference on Artificial Intelligence*, in SETN '22. New York, NY, USA: Association for Computing Machinery, 2022. doi: 10.1145/3549737.3549753.

[47] A. K and J. T. Podichetty, "Optimizing Drug Development: Assessing Small Language Models for Efficient Drug-Drug Interaction Data Extraction," *Proc. Am. Conf. Pharmacometrics 2024*, vol. null, p. null, doi: 10.70534/gwsf3516.