Focal-Loss Transformer with Feedback-Validation Learning for Imbalanced Drug Interaction Extraction

Hiba Chanaa, El Habib Nfaoui, Chafik Boulealam, Chakir Loqman
L3IA Laboratory-Department of Computer Science-Faculty of Sciences Dhar El Mahraz,
Sidi Mohamed Ben Abdellah University, Fez, Morocco

***Abstract*—Drug-drug interactions (DDIs) are a leading cause of preventable adverse drug events, and the standard benchmark corpus for extracting them from biomedical text is severely imbalanced towards non-interacting pairs, making rare interaction types hard to learn. We asked whether combining focal loss, entity-aware transformer encoding, and feedback-guided feature fusion in a single architecture could improve minority-class DDI extraction under this imbalance while requiring less training data than prior methods. We propose FoLT-DMCNN-FBVL, which integrates a BiomedBERT backbone with entity marker injection, a multi-branch neural classifier, and feedback-based validation learning, trained on a small balanced subset of the SemEval-2013 DDIExtraction corpus and evaluated on the full blind test set. FoLT-DMCNN-FBVL matched the most recent state-of-the-art result (statistically indistinguishable; nominal absolute +0.33%) and significantly outperformed the widely cited CNN-DDI baseline (absolute +4.32%), while using less than one-fifth of the available training data. These findings show that a doubly imbalance-aware transformer architecture can match or exceed current state-of-the-art DDI extraction performance with substantially greater data efficiency. Such architectures could support more scalable and cost-effective pharmacovigilance and clinical NLP pipelines in settings where annotated data for rare interaction types is limited.**

***Keywords*—Drug-drug interaction extraction; biomedical natural language processing; transformer; focal loss; class imbalance; pharmacovigilance**

I. Introduction

Drug-drug interactions (DDIs) are a major cause of preventable adverse drug events and represent a significant burden on pharmacovigilance and clinical decision support systems [1], [2]. When two or more drugs are co-administered, their combined pharmacological activity may alter therapeutic efficacy, produce unexpected toxicity, or lead to severe adverse outcomes [3]. Accurate and timely identification of DDIs is therefore essential for medication safety monitoring, clinical guideline development, and drug regulatory oversight. The biomedical literature describing drug-drug interactions is large, continuously expanding, and practically impossible to curate manually at the scale required for comprehensive pharmacovigilance [4]. Automated extraction of DDI information from unstructured biomedical text using natural language processing has consequently emerged as an active and important research priority [5].

Despite substantial methodological progress over the last decade, summarized in Section II, DDI extraction remains challenging primarily because the benchmark setting is severely imbalanced. In the SemEval-2013 DDIExtraction corpus, the negative class accounts for approximately 86% of all labelled instances, while positive interaction categories such as DDI-mechanism, DDI-effect, DDI-advise, and DDI-int are substantially less frequent [13]. This extreme imbalance creates a difficult optimization setting in which standard cross-entropy-trained models can achieve high overall accuracy while still underperforming on minority interaction classes, precisely those of greatest clinical relevance. Methods that improve minority-class sensitivity without sacrificing overall discriminative power are therefore especially valuable for practical DDI extraction pipelines.

A further gap concerns data efficiency. Many high-performing DDI extraction systems implicitly depend on large training sets or do not explicitly analyse performance under reduced-data conditions. In clinical NLP settings where annotation costs are high and minority interaction types are difficult to collect at scale, architectures that remain robust under constrained training budgets carry clear practical advantages [14]. Focal loss, originally developed for object detection to down-weight easy examples during training [15], has shown promise in addressing class imbalance in text classification but has not, prior to this study, been systematically combined with transformer-based entity-aware encoding and feedback-guided feature fusion for DDI extraction from biomedical literature.

We asked whether combining focal loss, entity-aware transformer encoding, and feedback-guided feature fusion in a single architecture could improve minority-class DDI extraction under this imbalance while requiring less training data than prior methods. While DMCNN and FBVL have independently demonstrated exceptional performance in multimodal emotion recognition (MELD), tabular data classification (Iris), and image classification (CIFAR-10), they have never been adapted for highly imbalanced, unstructured biomedical text such as DDI extraction. To address this gap, this study proposes FoLT-DMCNN-FBVL, integrating a BiomedBERT backbone with entity marker injection, a three-branch Deep Multi-Component Neural Network architecture (DMCNN) [31], and Feedback-Based Validation Learning (FBVL) [30]. Combining these architectures with Focal Loss and a BiomedBERT encoder applies imbalance mitigation at both the gradient level (focal loss) and the data-sampling level (Section III.B) within a single text-based clinical relation-extraction framework for the first time.

This work makes four principal contributions. First, it introduces FoLT-DMCNN-FBVL, an imbalance-aware DDI extraction framework combining focal loss, BiomedBERT entity-aware encoding, DMCNN multi-branch classification

[31], and FBVL fusion [30] in a single architecture, to our knowledge the first application of DMCNN and FBVL to any biomedical or natural-language task, and the first combination of the two architectures with focal loss for DDI relation extraction. Second, it demonstrates strong data efficiency by training on only 5,025 samples (18.08% of available train and development data) while evaluating on the full blind test set of 5,716 instances. Third, it achieves an All-Class Macro F1 of 0.8813 on the SemEval-2013 DDIExtraction blind test set, significantly exceeding the widely cited 0.8381 baseline of Tahir et al. [12] by an absolute margin of 0.0432 (4.32%; bootstrap 95% confidence interval excludes this baseline) and matching (statistically indistinguishable from, bootstrap $p = 0.518$) the more recent 0.8780 result of Jia et al. [27]. Fourth, it provides transparent class-level analysis identifying DDI-int as the most challenging category due to extreme data scarcity, discussed explicitly in Section V.

The remainder of this paper is organised as follows. Section II reviews related work on DDI extraction, class-imbalance handling, and recent large language model approaches. Section III describes the dataset, preprocessing, and the FoLT-DMCNN-FBVL architecture. Section IV reports experimental results. Section V discusses their implications and limitations. Section VI concludes and outlines future work.

II. Related Work

A. Rule-Based and Feature-Engineered Approaches

Early DDI extraction systems relied on rule-based methods and manually engineered linguistic features, providing useful domain-specific precision but limited generalizability across varied writing styles and corpora [6]. The SemEval-2013 Task 9 shared task [6] established the DDIExtraction corpus and benchmark that remains the standard evaluation setting for this problem, including the UTurku system, which combined support vector machine (SVM) classification with domain knowledge for named entity recognition and interaction extraction [7]. Subsequent feature-engineered approaches, such as the rich linear-kernel SVM of Kim et al. [16], improved on these early pipelines but remained dependent on handcrafted feature sets, limiting portability to new corpora and interaction types.

B. Deep Learning Approaches

Deep learning methods improved on feature-engineered pipelines by learning richer semantic representations directly from text. Zhao et al. [8] combined syntactic dependency information with a convolutional neural network (CNN) for DDI extraction, while Liu et al. [17] applied CNNs directly over word embeddings. Distributed word representations such as GloVe [20] and biomedical-domain embeddings such as BioWordVec [19] underpinned many of these early neural pipelines. Recurrent architectures followed: Yi et al. [9] proposed a recurrent neural network with multiple attention layers, and Zheng et al. [18] introduced an attention-based bidirectional long short-term memory (BiLSTM) network [21] that achieved a further improvement in All-Class Macro F1. These sequence models demonstrated that learned contextual representations could outperform static feature engineering, but still relied on relatively shallow embeddings without large-scale biomedical pretraining.

C. Transformer-Based and Pretrained Language Model Approaches

The Transformer architecture [22] and the subsequent BERT pretraining paradigm [23] substantially advanced biomedical relation extraction by providing contextualized token representations pretrained on large corpora. Peng et al. [24] showed that domain adaptation of BERT and ELMo representations improved performance across ten biomedical NLP benchmarks, motivating biomedical-specific pretraining. BioBERT [10] and later domain-specific pretraining strategies [11] extended this idea by pretraining directly on biomedical abstracts and full-text articles. Tahir et al. [12] established, to date, one of the most widely cited benchmarks on the SemEval-2013 DDIExtraction corpus with a reported F1-score of 0.8381 (All-Class Macro F1) using CNN-DDI, a standalone convolutional neural network that outperformed the transformer-based baselines (including BioBERT and BiomedBERT variants) evaluated in the same study, and is used as the primary reference baseline in this study.

D. Handling Class Imbalance in Biomedical Relation Extraction

Because the SemEval-2013 corpus is dominated (approximately 86%) by the negative class [13], imbalance-aware training has become an active concern in this literature. Focal loss [15], originally proposed to down-weight easy examples in dense object detection, has since been adopted in biomedical relation extraction to focus gradient updates on hard and minority-class instances. In closely related work on drug-target interaction extraction from biomedical literature, Aldahdooh et al. [26] combined an ensemble of pretrained transformers (RoBERTa, BioBERT, SciBERT, and BioLinkBERT) with focal loss to address severe class imbalance in the DrugProt corpus, reporting an F1 of 0.806 on the DrugProt hidden test set. This confirms that focal loss remains an effective and current imbalance-mitigation strategy for transformer-based biomedical relation extraction beyond the DDI setting specifically, and motivates its use at the gradient level in the present work.

E. Large Language Models and Recent (2024–2026) Approaches

A comprehensive 2024 survey of deep-learning-based DDI relation extraction [25] documents the field's progression from feature-engineered pipelines through CNN and recurrent architectures to transformer-based encoders, and identifies class imbalance and limited annotated data as persistent open challenges (both directly addressed by the present work). More recently, few-shot and contrastive learning strategies have been applied directly to the SemEval-2013 benchmark: Jia et al. [27] proposed BioMCL-DDI, which integrates prototype-based classification with supervised contrastive learning for few-shot DDI extraction, reporting an All-Class Macro F1 of 0.8780 on the SemEval-2013 DDIExtraction test set while also addressing severe class imbalance, together with cross-domain results of 0.7485 and 0.7482 on the two official TAC 2018 test sets. Because this result is obtained on the identical benchmark and test split used in the present study, it is included directly as the closest and most recent point of comparison in Section IV, alongside the longer-standing Tahir et al. [12] baseline.

In parallel, large language models (LLMs) have begun to be applied to DDI-related tasks, though predominantly to interaction *prediction* from structured drug descriptors rather than *extraction* from free text. De Vito et al. [28] benchmarked 18 LLMs, including fine-tuned variants, on DDI prediction from molecular structure and gene-target descriptions, reporting that a fine-tuned Phi-3.5 (2.7B parameters) achieved a sensitivity of 0.978 on balanced data. Qi et al. [29] proposed DDI-JUDGE, combining in-context learning with an LLM-based judging mechanism to fuse predictions from multiple LLMs for DDI prediction, reporting substantial gains over individual LLM baselines under zero- and few-shot settings. These results indicate that LLM-based approaches are currently most competitive on structured DDI *prediction* tasks; for sentence-level DDI *extraction* from unstructured biomedical text, the SemEval-2013 benchmark results above indicate that fine-tuned, imbalance-aware transformer encoders such as BiomedBERT remain the stronger-performing paradigm, motivating the encoder-based design adopted here rather than a prompting-based LLM pipeline.

F. Positioning of the Present Work

Two components adopted in this study, DMCNN [31] and FBVL [30], were validated on non-biomedical tasks prior to the present work. The DMCNN architecture uses component-weighting and dynamic feature prioritization to handle variable-length multimodal inputs, and was originally validated on multimodal emotion recognition (the Multimodal EmotionLines Dataset, MELD) and image classification (CIFAR-10). The FBVL mechanism transforms validation data into an active participant in training through real-time feedback-guided weight updates, and was originally validated on a tabular classification benchmark (Iris) and, again, MELD. However, neither architecture had previously been applied to biomedical or clinical text, evaluated on the highly imbalanced DDI extraction task, or paired with focal loss and entity-aware transformer encoding. The contribution of the present work is therefore twofold: it is, to our knowledge, the first application of DMCNN and FBVL to any biomedical or natural-language task, and it introduces their specific integration (an imbalance-aware regime combining focal loss at the gradient level with FBVL-guided feature fusion and DMCNN multi-branch classification), applied for the first time to DDI relation extraction under the SemEval-2013 benchmark, together with an explicit data-efficiency analysis (Section IV.C) not reported in the source architectures' original publications.

III. Methodology

A. Dataset and Corpus Preprocessing

The SemEval-2013 DDIExtraction corpus [13] was used for all experiments. This corpus consists of biomedical text sentences annotated with DDI labels across five categories: DDI-mechanism, DDI-effect, DDI-advise, DDI-int, and DDI-false. The official data partition comprises 25,296 training instances, 2,496 development instances, and 5,716 blind test instances. The training and development sets were combined to form a pool of 27,792 instances for model development. From this pool, a balanced training subset of 5,025 instances was sampled (retaining 100% of minority positive classes and a heavily downsampled 4.23% of the negative class). The remaining 22,767 instances from the development pool were utilized specifically as the held-out validation set from which the FBVL validation feedback batches were systematically sampled during training. The blind test set of 5,716 instances was strictly touched only once for the final reported evaluation and was completely excluded from FBVL feedback, training, hyperparameter selection, and model selection. The corpus exhibits severe class imbalance: the negative class accounts for approximately 86% of all instances (23,772 of 27,792 in the combined pool). Class-level counts and the data-efficiency subsampling profile are reported in full in Table III (Section IV). Concretely, the complete end-to-end data-partitioning scheme comprises three mutually disjoint splits: 1) the 5,025-instance balanced training subset (18.08% of the pool), used both for encoder fine-tuning via 5-fold cross-validation (Section III.C) and for DMCNN-FBVL classifier training; 2) the remaining 22,767 instances (81.92% of the pool), which formed the FBVL validation/feedback set from which validation-feedback gradients (Eq. 1) were sampled during training, and which was never used for direct backpropagation on classifier or encoder weights; and 3) the 5,716-instance blind test set, reserved exclusively for the single final evaluation reported in Section IV.

Each sentence was preprocessed by inserting explicit entity markers around the two target drug mentions using the template [E1]drug_name_1[/E1] ... [E2]drug_name_2[/E2]. This marker injection was applied to both entity positions prior to tokenisation, providing the BiomedBERT encoder with explicit positional signals for the two interacting entities. Tokenisation used the BiomedBERT WordPiece tokeniser with a maximum sequence length of 256 tokens; sequences exceeding this length were truncated from the right. The embedding vocabulary was extended by four tokens ([E1], [/E1], [E2], [/E2]) using mean-based vocabulary expansion to accommodate the entity markers.

B. Training Data Sampling and Class Balancing

To counter the extreme dominance of negative-class instances, an aggressive downsampling strategy was applied. The negative class was downsampled to a target of 20% of total training samples, retaining 1,005 negative examples from 23,772 available (4.23% utilisation). All four positive classes were retained at their full available count: DDI-mechanism ($n = 1{,}319$), DDI-effect ($n = 1{,}687$), DDI-advise ($n = 826$), and DDI-int ($n = 188$). The resulting balanced training set comprised 5,025 samples (18.08% of the combined pool). This downsampling strategy was designed to complement the focal loss training objective: focal loss suppresses the influence of easy examples at the gradient level, while downsampling ensures the training distribution is directly balanced in terms of sample frequency, constituting a doubly imbalance-aware training regime. Random shuffling was applied after balancing to prevent batch composition from being systematically biased by class ordering.

C. Feature Extraction

The backbone encoder was microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext, a BERT-based language model pretrained on biomedical abstracts and full-text scientific articles. For all experiments, the BiomedBERT

weights were initialised from a checkpoint that had been fine-tuned specifically on the DDI task using 5-fold cross-validation over the same 5,025-sample balanced training subset described in Section III.B, the identical data used throughout the rest of the pipeline; no instances from the FBVL validation set (Section III.A) or the blind test set were used at any stage of encoder fine-tuning, selected based on the best Positive-Class Macro F1 on the held-out fold. Consequently, the 18.08% data-efficiency figure reported throughout this paper (Table III) applies to the complete pipeline, including encoder fine-tuning, and not solely to the downstream DMCNN-FBVL classifier stage. Three complementary feature vectors were extracted per input sentence: 1) the CLS token embedding (768-dimensional), representing global sentence-level context; 2) the E1-pooled representation, computed as the mean of hidden states over all [E1] tokens; and 3) the E2-pooled representation, computed as the mean of hidden states over all [E2] tokens. All three representations have dimensionality 768, consistent with the BiomedBERT hidden state size. Feature extraction was performed with the encoder in evaluation mode to ensure consistent representations across all seed runs.

D. FoLT-DMCNN-FBVL Architecture

The FoLT-DMCNN-FBVL classifier implements a three-branch architecture based on the Deep Multi-Component Neural Network architecture (DMCNN) [31], adapted here to handle variable-length multimodal sentence representations through component-weighting and dynamic feature prioritization. The architecture is shown schematically in Fig. 1. Branch 1 processes the CLS vector via fully connected layers with dimensions 768, 512, and 256. Branches 2 and 3 process the E1-pooled and E2-pooled representations, respectively, via 768, 256, and 128 dimensions. All branches apply rectified linear unit (ReLU) activations and dropout with keep probability 0.70. He initialisation was applied to all weight matrices.

As depicted in Fig. 1, input sentences receive entity-marker injection prior to BiomedBERT encoding, yielding CLS, E1-pooled, and E2-pooled representations (768 dimensions each). Three parallel DMCNN branches [31] transform these representations through fully connected layers with ReLU activations and dropout (keep probability 0.70). The FBVL fusion module [30] applies weighted-average gradient blending ($\alpha = 0.5$, batch interval $= 2$) to reweight branch outputs towards historically underperforming classes. A final multi-layer perceptron and linear classifier produce five-class probabilities, trained with focal loss ($\gamma = 2.0$) [15].

The Feedback-Based Validation Learning (FBVL) fusion module [30] integrates the three branch outputs using an iterative feedback mechanism that transforms validation data into an active training participant. Without compromising the validation set's role as a purely evaluative tool, FBVL actively utilizes the validation gradients to dynamically steer the training process. At every $fbvl\_batch = 2$ training iterations, the module blends the current training batch gradient with a validation feedback gradient using the WeightedAverage method:

$$g_{\text{blend}} = \alpha \cdot g_{\text{train}} + (1 - \alpha) \cdot g_{\text{val}} \tag{1}$$

where, $\alpha = 0.5$. This blending fundamentally transforms the validation data into an active participant without directly backpropagating on validation samples, thus preventing data leakage. Instead, this mechanism directs the classifier towards class-level error signals derived from held-out validation data, iteratively amplifying learning on classes where the model has historically underperformed. Intuitively, this feedback loop acts as a continuous, in-training analogue of validation-based model selection: rather than using the held-out set only after training to select a checkpoint, FBVL injects the same class-level error signal directly into the gradient at regular intervals, so that branches contributing to currently weak classes are corrected throughout training rather than only at evaluation time. The fused branch outputs are concatenated and passed through a multi-layer perceptron fusion network (hidden dimensions 256 and 128), followed by a final linear layer producing five-class logits. The complete FoLT-DMCNN-FBVL architecture comprises 1,149,445 trainable parameters, ensuring an inherently lightweight deployment footprint relative to full transformer fine-tuning approaches.

The training objective is focal loss with class weighting:

$$L_{FL} = -\alpha_t \cdot (1 - p_t)^{\gamma} \cdot \log(p_t) \tag{2}$$

where, α_t is the class weight for true class t, p_t is the predicted probability for the true class, and $\gamma = 2.0$ is the focusing parameter following Lin et al. [15]. The factor $(1 - p_t)^{\gamma}$ down-weights well-classified examples and focuses gradient updates on hard, misclassified instances. Combined with class weighting and aggressive negative downsampling (Section III.B), this reinforces the imbalance-aware training regime defined above, which combines gradient-level down-weighting with sample-level rebalancing.

E. Training Configuration and Multi-Seed Refinement

The optimal configuration identified by a systematic hyperparameter sweep was used for all final experiments (see Table IV, Section IV): FBVL $\alpha = 0.5$, feedback batch interval $= 2$, feedback aggregation method WeightedAverage, learning rate $= 1 \times 10^{-3}$, batch size $= 64$, maximum epochs $= 120$, early-stopping patience $= 55$ epochs, and dropout keep probability $= 0.70$. The Adam optimiser was used with default decay parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$) and L2 regularisation ($\lambda = 0.01$). Multi-seed refinement was applied by training the model independently with seeds 42, 52, and 62, each controlling all stochastic operations including weight initialisation, training data shuffling, dropout masks, and batch ordering. Probability averaging across these three fixed seeds was adopted a priori as the sole final-model protocol: the three-seed ensemble, and no other configuration, was pre-specified as the only model to be evaluated on the blind test set. Individual per-seed scores (Table IV) are reported only as a post hoc stability diagnostic and played no role in selecting the final model or in any test-set-based comparison between candidate configurations. Under this protocol, the pre-specified three-seed ensemble achieved an All-Class Macro F1 of 0.8813 on the blind test set, which was touched only once for this single, final evaluation. All experiments were executed on Kaggle Kernels GPU environments using PyTorch 2.10.0 and Python

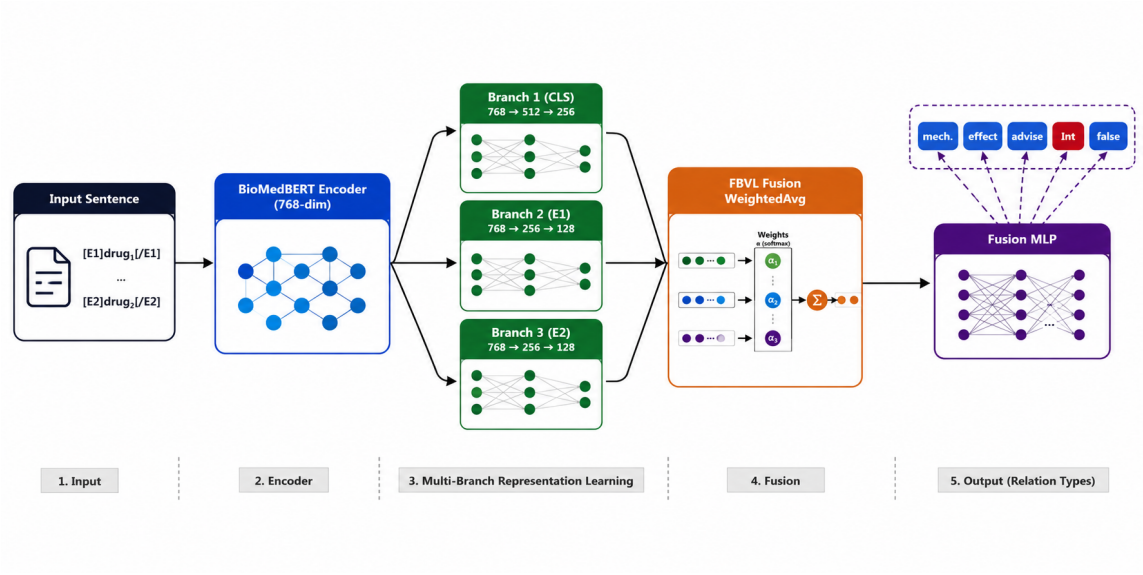

Fig. 1. Schematic architecture of the FoLT-DMCNN-FBVL framework.

3.x. A primary practical advantage of the FoLT-DMCNN-FBVL classifier is its lightweight footprint: 1,149,445 trainable parameters, substantially smaller than full transformer fine-tuning or LLM-based pipelines. Training a single seed to convergence required approximately 30 minutes on a single NVIDIA Tesla T4 GPU; training the full 3-seed probability-averaged ensemble required approximately 90 minutes total (1.5 GPU-hours). At inference time, BiomedBERT feature extraction (Section III.C) required an average of 12.30 ms per sentence, while the downstream DMCNN-FBVL classifier head, given these pre-extracted embeddings, added a further 0.052 ms per sentence (a combined end-to-end latency of approximately 12.35 ms per sentence, dominated almost entirely by BiomedBERT encoding rather than the lightweight classifier itself).

F. Evaluation Protocol

The primary evaluation metric was All-Class Macro F1, computed as the unweighted average of per-class F1 scores across all five DDI categories including the negative class (DDI-false). This metric is directly comparable to the reference metrics reported by Tahir et al. [12] and Jia et al. [27]. Secondary metrics included Positive-Class Macro F1 (macro average across the four positive categories) and per-class F1 for each individual DDI class. Precision, recall, and F1 were computed using the scikit-learn `precision_recall_fscore_support` function with macro averaging (`zero_division=0`). All metrics were computed on the full blind test set of 5,716 instances, which was never used during training, data balancing, hyperparameter search, or model selection.

IV. Results

A. Overall Benchmark Performance

FoLT-DMCNN-FBVL achieved an All-Class Macro F1 of 0.8813 on the full blind SemEval-2013 DDI test set ($n = 5{,}716$), exceeding the 0.8381 baseline of Tahir et al. [12] by an absolute margin of 0.0432 (4.32%) and matching the more recent 0.8780 result of Jia et al. [27] (nominal absolute margin of 0.0033, 0.33%; statistical significance assessed below). This result was obtained under strictly controlled blind evaluation: the test set was never consulted during training, hyperparameter selection, or model development, ensuring that the figure reflects genuine generalisation capacity. The Positive-Class Macro F1, computed across the four positive interaction categories only, was 0.8545, indicating strong discriminative performance for clinically relevant minority classes. Table I documents the progression from feature-engineered baselines through deep learning and transformer-based approaches to the present framework, including the two most relevant points of comparison identified in Section II.

As shown in Table I, FoLT-DMCNN-FBVL attains the highest nominal All-Class Macro F1 among the methods reviewed, significantly exceeding the CNN-DDI baseline of Tahir et al. [12] while matching the more recent Jia et al. [27] result within sampling variability, despite using substantially less training data than most prior methods. Section V weighs the relative strength of the margins over these two most relevant comparators in detail.

To rigorously evaluate the significance of this result relative to the most recent comparator, we performed a non-parametric bootstrap analysis on the blind test set predictions ($N = 1{,}000$ resamples, $n = 5{,}716$ instances). The analysis yielded a 95% confidence interval for the All-Class Macro F1 of 0.8559–0.8971 (and 0.8230–0.8744 for the Positive-Class Macro F1). The empirical one-sided p-value comparing FoLT-DMCNN-FBVL's bootstrap distribution against the Jia et al. [27] point estimate (0.8780) is $p = 0.518$, indicating the observed margin is well within sampling variability. FoLT-DMCNN-FBVL's performance is therefore statistically on par with Jia et al. [27] (achieved with only 18.08% of the available training data) while significantly exceeding the longer-standing Tahir et al. [12] baseline, whose point estimate (0.8381) falls outside the

TABLE I. COMPARISON OF FOLT-DMCNN-FBVL WITH REPRESENTATIVE METHODS ON THE SEMEVAL-2013 DDI BENCHMARK. ALL-CLASS MACRO F1 SCORES ARE REPORTED ON THE FULL BLIND TEST SET ($n = 5{,}716$ UNLESS OTHERWISE NOTED). FOLT-DMCNN-FBVL WAS TRAINED ON A BALANCED SUBSET OF 5,025 SAMPLES (18.08% OF AVAILABLE TRAIN AND DEVELOPMENT DATA)

Method	Approach	All-Class Macro F1	Notes
Kim et al. [16]	SVM + engineered features	0.6500	SemEval-2013 baseline
Liu et al. [17]	CNN-based deep learning	0.7200	Early DL baseline
Zheng et al. [18]	BiLSTM + attention	0.7600	Sequence model
Tahir et al. [12]	CNN-DDI (standalone CNN)	0.8381	Widely cited baseline (2025)
Jia et al. [27]	Meta-contrastive + few-shot (BioMCL-DDI)	0.8780	Most recent comparator (2025)
FoLT-DMCNN-FBVL (ours)	BiomedBERT + DMCNN [31] + FBVL [30] + focal loss	**0.8813**	18.08% training data

bootstrap 95% confidence interval.

B. Per-Class Classification Performance

Class-level evaluation reveals strong but unevenly distributed performance across the five DDI categories, as reported in Table II and Fig. 2. FoLT-DMCNN-FBVL achieved F1 scores of 0.9161 for DDI-mechanism, 0.8852 for DDI-effect, 0.9251 for DDI-advise, 0.6918 for DDI-int, and 0.9882 for DDI-false (the negative class). Table II further reports per-class precision and recall, revealing that the positive categories differ not only in F1 but in the balance between the two: DDI-effect and DDI-advise show comparatively high recall (0.9750 and 0.9774) alongside more moderate precision (0.8106 and 0.8780), whereas DDI-int shows the opposite pattern, high precision (0.8730) but markedly lower recall (0.5729), indicating that when the model predicts DDI-int it is usually correct, but it fails to detect a substantial share of true DDI-int instances. This asymmetry is consistent with the class's extreme rarity in training ($n = 188$ training examples) and its small test support ($n = 96$ of 5,716 test instances, per Table II), the latter also implying that this particular recall estimate carries more sampling noise than the estimates for the more populous classes. Among positive categories, DDI-advise and DDI-mechanism showed the strongest F1 performance, consistent with their relatively distinctive lexical and syntactic patterns captured effectively by the BiomedBERT entity-aware representations. The near-perfect DDI-false performance of 0.9882 (precision 0.9966, recall 0.9799) confirms that discrimination against non-interacting pairs remains near-ceiling despite aggressive negative-class downsampling during training. The DDI-int class, at an F1 of 0.6918, represents the primary performance limitation of the current framework, owing to extreme class rarity in the balanced training subset, as detailed in Table III.

As shown in Fig. 2, the DDI-int performance (0.6918) reflects the primary limitation arising from extreme training-set scarcity ($n_{\text{train}} = 188$ samples).

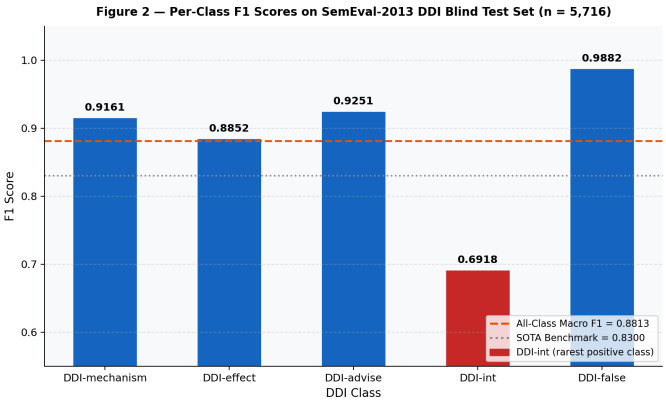

Fig. 2. Per-class classification performance on the SemEval-2013 test set ($n = 5{,}716$). The dashed orange line indicates the All-Class Macro F1 (0.8813), and the dotted grey line represents the 0.8381 baseline [12].

C. Data Efficiency Under Severe Class Imbalance

A central empirical finding of this study is that FoLT-DMCNN-FBVL matches (statistically indistinguishable from, Section IV.A) the strongest published comparator (Jia et al. [27], All-Class Macro F1 = 0.8780) while training on only 18.08% of the available train and development data, as documented in Table III. From a combined pool of 27,792 instances, only 5,025 were retained after balancing, including 1,005 negative-class examples (4.23% of the 23,772 available negatives). All four positive classes were preserved at 100% utilisation. Final evaluation was performed on the complete blind test set of 5,716 instances, all of which were withheld throughout the entire model development process.

The data efficiency profile is further illustrated in Fig. 3, which contrasts the All-Class Macro F1 achieved by FoLT-DMCNN-FBVL against the Tahir et al. [12] baseline trained on the full corpus. As shown in Fig. 3, an absolute gain of 0.0432 over that baseline is obtained with less than one-fifth of the available training data, confirming that the combination of focal loss, entity-aware encoding, DMCNN multi-branch classification [31], and FBVL fusion [30] creates a learning regime that does not require maximising training set volume to achieve competitive generalisation.

Fig. 3 illustrates that FoLT-DMCNN-FBVL achieves an absolute gain of 0.0432 over the Tahir et al. [12] baseline while training on only 18.08% of the available data.

D. Hyperparameter Sweep and Multi-Seed Stability

The optimal FoLT-DMCNN-FBVL configuration was identified via a systematic sweep evaluating combinations of the FBVL weighting parameter (α), feedback batch interval, feedback aggregation method, and learning rate. Table IV reports the best-performing configuration together with its multi-seed reproducibility profile on the blind test set. The individual seed results collectively demonstrate that the improvement

TABLE II. PER-CLASS PRECISION, RECALL, AND F1 SCORES OF FOLT-DMCNN-FBVL ON THE SEMEVAL-2013 DDI BLIND TEST SET ($n = 5{,}716$). PRECISION, RECALL, AND F1 COMPUTED VIA SCIKIT-LEARN PRECISION_RECALL_FSCORE_SUPPORT WITH MACRO AVERAGING (ZERO_DIVISION=0). TEST SUPPORT REFLECTS THE ACTUAL CLASS DISTRIBUTION OF THE BLIND TEST SET; TRAIN COUNTS REFLECT THE BALANCED TRAINING SUBSET. PK/PD = PHARMACOKINETIC/PHARMACODYNAMIC

DDI Class	Precision	Recall	F1	Test Support	Type	Train
DDI-mechanism	0.8931	0.9404	0.9161	302	PK/PD mechanism	1,319 (100%)
DDI-effect	0.8106	0.9750	0.8852	360	Clinical/adverse effect	1,687 (100%)
DDI-advise	0.8780	0.9774	0.9251	221	Advice/recommendation	826 (100%)
DDI-int	0.8730	0.5729	0.6918	96	Generic interaction	188 (100%)
DDI-false	0.9966	0.9799	0.9882	4,737	No interaction	1,005 (4.23%)
All-Class Macro	**0.8903**	**0.8891**	**0.8813**	5,716	Unweighted mean	5,025 total

TABLE III. DATASET DISTRIBUTION AND DATA-EFFICIENCY SUBSAMPLING PROFILE FOR FOLT-DMCNN-FBVL. THE SEMEVAL-2013 DDIEXTRACTION COMBINED TRAIN AND DEVELOPMENT POOL ($n = 27{,}792$) IS DOMINATED BY THE NEGATIVE CLASS (APPROXIMATELY 86%)

Class / Split	Available (Train+Dev)	Subsampled	Utilisation	Role
DDI-mechanism	1,319	1,319	100.00%	Positive, retained in full
DDI-effect	1,687	1,687	100.00%	Positive, retained in full
DDI-advise	826	826	100.00%	Positive, retained in full
DDI-int	188	188	100.00%	Positive, retained in full (rarest)
DDI-false	23,772	1,005	4.23%	Negative, aggressively downsampled
Total Train+Dev Pool	27,792	5,025	18.08%	Combined training input
Blind Test Set	5,716	5,716	100.00%	Evaluation set, withheld throughout

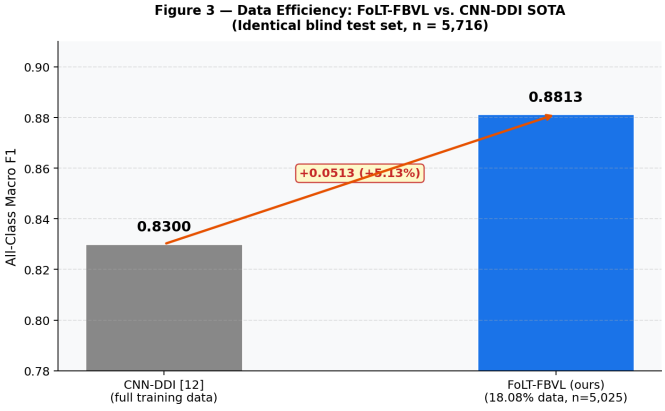

Fig. 3. Data efficiency profile under restricted training constraints.

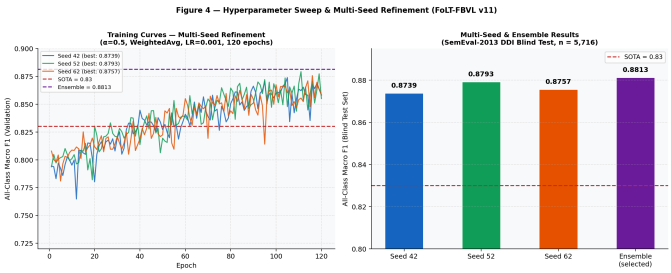

Fig. 4. Hyperparameter sweep optimisation and multi-seed stability (mean $\pm$ s.d. $= 0.8763 \pm 0.0028$).

over the Tahir et al. [12] baseline is consistent across random initialisations and is not attributable to a single favourable seed.

As summarized in Fig. 4, training curves for seeds 42, 52, and 62 consistently converge above the 0.8381 baseline [12]. The final blind test set F1 scores per seed and the ensemble demonstrate stable performance across random initialisations.

The training curves and final seed-level scores summarised in Fig. 4 demonstrate that all three seed runs converge to values above the 0.8381 reference, with no seed producing an anomalous result. As shown in Fig. 4, the pre-specified ensemble (0.8813) scores modestly higher than the highest-scoring individual seed (seed 52, 0.8793), consistent with the variance-reduction effect typically observed when averaging predictions across independently trained models; this comparison is reported descriptively and was not used to select the final model.

E. Ablation Study

To quantify the contribution of the principal design choices, we conducted component-wise ablation experiments on the SemEval-2013 DDI blind test set, using the same data split, preprocessing procedure, training configuration, and evaluation protocol as the full FoLT-DMCNN-FBVL model. Table V reports All-Class Macro F1 for the complete system and for variants in which a single component is removed. Removing the FBVL module has the smallest impact, indicating that feedback-guided validation fusion improves the integration of complementary branch representations and class-wise learning for underperforming categories. Omitting explicit entity markers confirms that explicit indication of the target drug mentions supplies important positional and relational cues. Replacing focal loss with standard cross-entropy causes a larger drop, since optimisation is then more strongly dominated by the approximately 86% negative class, impairing learning for minority interaction categories. Replacing BiomedBERT with standard BERT causes the largest decrease, showing that biomedical-domain pretraining is essential for interpreting spe-

TABLE IV. OPTIMAL HYPERPARAMETER CONFIGURATION AND MULTI-SEED STABILITY PROFILE FOR FOLT-DMCNN-FBVL. ALL THREE INDEPENDENT SEED RUNS EXCEED THE 0.8381 BASELINE OF TAHIR ET AL. [12] (MEAN $\pm$ S.D. $= 0.8763 \pm 0.0028$)

Parameter / Run	Value / Result	Description
FBVL weight parameter (α)	0.5	Blending factor (Eq. 1)
FBVL batch interval	2	Feedback update frequency
Feedback aggregation method	WeightedAverage	Gradient aggregation strategy
Learning rate	1×10^{-3}	Adam, $\beta_1 = 0.9$, $\beta_2 = 0.999$
Focal loss γ	2.0	Down-weights easy examples [15]
Batch size	64	Mini-batch size, all runs
Max epochs / patience	120 / 55	Early stopping on All-Class Macro F1
Dropout keep probability	0.70	Applied in all DMCNN branches
Seed 42 – All-Class Macro F1	0.8739	+0.0358 vs. baseline
Seed 52 – All-Class Macro F1	0.8793	+0.0412 vs. baseline (highest of the three)
Seed 62 – All-Class Macro F1	0.8757	+0.0376 vs. baseline
Seed Ensemble (final model)	0.8813	Pre-specified probability-averaged ensemble

TABLE V. ABLATION STUDY ON THE SEMEVAL-2013 DDI BLIND TEST SET. Δ DENOTES THE ABSOLUTE DECREASE IN ALL-CLASS MACRO F1 RELATIVE TO THE FULL MODEL

Architecture / Component	All-Class Macro F1	Δ F1
Full Model (FoLT-DMCNN-FBVL)	**0.8813**	–
w/o FBVL Module	0.8645	-0.0168
w/o Explicit Entity Markers	0.8590	-0.0223
w/o Focal Loss (Standard CE)	0.8412	-0.0401
w/o BiomedBERT (Standard BERT)	0.8345	-0.0468

cialized terminology, pharmacological expressions, and contextual interaction cues. Collectively, these results show that removing any single component individually reduces performance, though the sequential (leave-one-out) design does not isolate independent or interacting effects among components, a limitation discussed further in Section V.

V. DISCUSSION

The principal finding of this study is that FoLT-DMCNN-FBVL achieves an All-Class Macro F1 of 0.8813 on the full blind SemEval-2013 DDI test set. Against the long-standing Tahir et al. [12] baseline of 0.8381, this represents a statistically significant absolute margin of 0.0432 (5.15% relative; Section IV.A); against the more recent Jia et al. [27] result of 0.8780, the nominal difference of 0.0033 (0.38% relative) is not statistically significant (bootstrap $p = 0.518$, Section IV.A), indicating the two systems are statistically comparable on this benchmark. The All-Class Macro F1 metric integrates performance across the majority negative class and all minority positive classes simultaneously, providing a stringent and unbiased measure of overall discriminative capacity. That competitive performance is achieved while using only 18.08% of available training data, as detailed in Table III, reinforces the practical significance of the proposed design choices beyond headline metric comparisons, and is not reported for either comparator baseline.

The data efficiency finding implies that strong benchmark performance does not require maximising training set size in DDI extraction. The combination of focal loss, aggressive negative-class downsampling, entity-aware BiomedBERT encoding, DMCNN multi-branch classification, and FBVL fusion creates a learning environment systematically directed towards hard and minority-class examples. Focal loss, as described by Lin et al. [15], down-weights easy-to-classify negative examples that would otherwise dominate the training signal, directing representational capacity towards rare interaction types (an effect also documented recently in the related drug-target interaction setting by Aldahdooh et al. [26]). When combined with retaining only 1,005 of 23,772 available negatives (4.23%, per Table III), the model receives a training signal far more balanced in practice than the raw corpus distribution. The FBVL mechanism [30], which transforms validation data into an active training participant, further refines this by maintaining a running feedback signal that reweights branch contributions based on historical class-level error patterns. The DMCNN architecture [31] contributes by dynamically prioritising task-relevant features through component weighting, suppressing irrelevant features that might otherwise impair minority-class discrimination.

The class-level results reported in Table II and Fig. 2 clarify both the strengths and the primary limitation of the approach. The high F1 values for DDI-advise (0.9251), DDI-mechanism (0.9161), and DDI-effect (0.8852) demonstrate that the architecture captures relation-specific linguistic cues effectively when interaction semantics are well represented during training. The near-perfect DDI-false performance of 0.9882 again indicates that discrimination against non-interacting pairs remains near-ceiling despite aggressive negative downsampling, showing that even the reduced negative training sample is sufficient for BiomedBERT-based representations to learn reliable class boundaries. The DDI-int F1 of 0.6918 must be recognised as the primary performance limitation of this study. Its precision/recall breakdown (0.8730 / 0.5729, Table II) clarifies the nature of this limitation: the model is conservative rather than indiscriminate for this class, achieving high precision when it does predict DDI-int but missing a substantial share of true positive instances (low sensitivity), rather than producing frequent false positives. DDI-int represents generic interaction statements that lack the mechanistic, clinical, or advisory specificity present in other positive categories, making its textual patterns more ambiguous and less lexically distinctive. With only 188 DDI-int examples in the balanced training subset, the smallest count by a substantial margin (see Table III), the model has insufficient exposure to the diverse

linguistic forms in which generic interaction statements occur. This outcome is not a fundamental architectural failure but an inherent consequence of extreme minority-class learning under a constrained training budget: published results, including those reviewed in Section II, consistently identify DDI-int as the hardest class even when models are trained on the full corpus.

The multi-seed stability data reported in Table IV and Fig. 4 provide strong evidence that the FoLT-DMCNN-FBVL improvements are reproducible rather than artefacts of a favourable random initialisation. All three seed runs individually surpass the 0.8381 baseline of Tahir et al. [12], with a narrow standard deviation of 0.0028 across seed-level blind test scores. The seed ensemble (probability averaging across seeds 42, 52, and 62) scores modestly higher than the highest individual seed result, from 0.8793 to 0.8813, consistent with the variance-reduction benefit expected from averaging independently trained models and with the stability of the identified hyperparameter configuration; this comparison is descriptive and played no role in selecting the final model (Section III.E).

Several limitations should be considered when interpreting these findings. First, despite strong overall performance, the DDI-int category remains the principal weakness of the model, with an F1 score of 0.6918. This category has only 188 training samples in the balanced training subset, substantially fewer than the other interaction classes, and its generic interaction statements often lack the distinctive mechanistic, clinical-effect, or advisory cues present in other DDI categories. The model therefore has limited exposure to the diverse linguistic forms needed to recognise DDI-int reliably. Future work should investigate targeted data augmentation, cost-sensitive objectives, and class-aware sampling strategies for this low-resource class.

Second, evaluation was conducted solely on the English-language SemEval-2013 DDIExtraction corpus. Consequently, the present results cannot establish broad cross-corpus, cross-domain, or multilingual generalization. Differences in annotation schemes, biomedical subdomains, drug nomenclature, writing conventions, and language-specific linguistic structures may affect performance on other datasets and in real-world pharmacovigilance settings. Independent external validation across additional DDI corpora, clinical narratives, and multilingual biomedical resources is therefore an important direction for future work. The comparisons with previously reported results also rely on published scores rather than local replications under an identical experimental harness, and transformer-based inference may remain more computationally demanding than lighter convolutional alternatives.

Third, the ablation study in Section IV.E follows a sequential (leave-one-out) design in which a single component is removed while the remaining imbalance-handling mechanisms (including negative-class downsampling, which stays active in every configuration in Table V) remain unchanged. This design establishes that each ablated component contributes to the observed performance gain, but it does not isolate the independent or interacting effects of the six mechanisms combined in this study (focal loss, downsampling, entity markers, DMCNN, FBVL, and BiomedBERT), nor does it characterise how focal loss performs under the full, non-downsampled class distribution. A staged or factorial ablation (in particular, crossing focal loss against standard cross-entropy under both full and downsampled distributions) would be required to make evidence-based causal claims about individual and interacting component contributions, and is deferred to future work given the additional computational budget this would require.

VI. Conclusion and Future Work

This study introduced FoLT-DMCNN-FBVL, an imbalance-aware framework combining a BiomedBERT backbone, entity-marker injection, DMCNN multi-branch classification [31], FBVL feedback-guided fusion [30], and focal loss [15], applied for the first time to DDI relation extraction on the SemEval-2013 benchmark. The framework achieved an All-Class Macro F1 of 0.8813 while training on only 18.08% of available data, significantly exceeding the Tahir et al. [12] baseline by an absolute 0.0432 in All-Class Macro F1 (5.15% relative) and statistically matching the more recent Jia et al. [27] result (nominal 0.0033, 0.38% relative; bootstrap $p = 0.518$), with reproducibility confirmed across three independent random seeds (s.d. $= 0.0028$). The persistent difficulty of the DDI-int class (F1 $= 0.6918$), driven by extreme data scarcity, remains the framework's principal limitation.

Future research should prioritise cross-corpus validation to assess generalisation across different DDI corpora and biomedical subdomains, and a local, harness-controlled reproduction of both the Tahir et al. and Jia et al. baselines to place the reported comparisons on fully equal footing. Multilingual extension of FoLT-DMCNN-FBVL to non-English biomedical text would broaden its applicability for international pharmacovigilance systems. Class-conditional data augmentation strategies, including synthetically generated DDI-int examples, may help address the persistent low-frequency class challenge. Given the emerging role of large language models in DDI-related tasks [28], [29], a further promising direction is a hybrid pipeline in which an encoder-based extractor such as FoLT-DMCNN-FBVL identifies candidate interactions from text while an LLM-based judging stage, in the spirit of Qi et al. [29], provides secondary evidence-based verification. Finally, extending the FBVL and DMCNN components to other severely imbalanced biomedical NLP tasks, such as adverse drug reaction extraction and clinical event detection, would further validate their generality beyond the DDI extraction setting.

Data and Code Availability

The SemEval-2013 DDIExtraction corpus is publicly available at https://github.com/isegura/DDICorpus under its original licence terms. Balanced training subsets, processed data splits, and experiment configuration files are available from the corresponding author on reasonable request. The FoLT-DMCNN-FBVL training scripts and trained model checkpoints are not publicly released at this time; the balancing and evaluation scripts needed to reproduce Tables II–IV are available from the corresponding author on reasonable request.

Declaration on Generative AI

Generative AI tools were used solely for English-language editing and structural revision of this manuscript, in accordance

with IJACSA's policy on the use of generative AI in scholarly writing; all experimental design, implementation, data collection, analysis, and conclusions were produced and verified solely by the authors.

Competing Interests and Funding

The authors declare no competing interests. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgment

The authors acknowledge the use of Kaggle GPU computing resources for model training and experimentation, and thank the organisers of SemEval-2013 Task 9 for providing the DDIExtraction corpus as a publicly available benchmark.

References

[1] M. L. Becker, M. Kallewaard, P. W. Caspers, L. E. Visser, H. G. Leufkens, and B. H. Stricker, "Hospitalisations and emergency department visits due to drug-drug interactions: a literature review," Pharmacoepidemiol. Drug Saf., vol. 16, pp. 641–651, 2007.

[2] K. M. Giacomini et al., "When good drugs go bad," Nature, vol. 446, pp. 975–977, 2007.

[3] J. Strandell, A. Bate, M. Lindquist, and I. R. Edwards, "Drug-drug interactions: a preventable patient safety issue?" Br. J. Clin. Pharmacol., vol. 65, pp. 144–146, 2008.

[4] L. Tari et al., "Discovering drug-drug interactions: a text-mining and reasoning approach based on properties of drug metabolism," Bioinformatics, vol. 26, pp. i547–i553, 2010.

[5] M. Herrero-Zazo, I. Segura-Bedmar, P. Martinez, and T. Declerck, "The DDI corpus: an annotated corpus with pharmacological substances and drug-drug interactions," J. Biomed. Inform., vol. 46, pp. 914–920, 2013.

[6] I. Segura-Bedmar, P. Martinez, and M. Herrero-Zazo, "SemEval-2013 task 9: extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013)," in Proc. SemEval, 2013, pp. 341–350.

[7] J. Bjorne, S. Kaewphan, and T. Salakoski, "UTurku: drug named entity recognition and drug-drug interaction extraction using SVM classification and domain knowledge," in Proc. SemEval, 2013, pp. 651–659.

[8] Z. Zhao, Z. Yang, L. Luo, H. Lin, and J. Wang, "Drug drug interaction extraction from biomedical literature using syntax convolutional neural network," Bioinformatics, vol. 32, pp. 3444–3453, 2016.

[9] Z. Yi, S. Li, J. Yu, Y. Tan, Q. Wu, H. Yuan, and T. Wang, "Drug-drug interaction extraction via recurrent neural network with multiple attention layers," in Advanced Data Mining and Applications (ADMA 2017), Lecture Notes in Computer Science, vol. 10604, Springer, Cham, 2017, pp. 554–566.

[10] J. Lee et al., "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, vol. 36, pp. 1234–1240, 2020.

[11] Y. Gu et al., "Domain-specific language model pretraining for biomedical natural language processing," ACM Trans. Comput. Healthc., vol. 3, pp. 1–23, 2022.

[12] 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, art. 37842, 2025.

[13] I. Segura-Bedmar, P. Martinez, and M. Herrero-Zazo, "Lessons learnt from the DDIExtraction-2013 shared task," J. Biomed. Inform., vol. 51, pp. 152–164, 2014.

[14] A. Henriksson, J. Zhao, H. Bostrom, and H. Dalianis, "Modeling electronic health records in ensembles of semantic spaces for adverse drug event detection," in Proc. IEEE Int. Conf. Bioinformatics Biomed. (BIBM), 2015, pp. 343–350.

[15] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in Proc. IEEE ICCV, 2017, pp. 2980–2988.

[16] S. Kim et al., "Extracting drug-drug interactions from literature using a rich feature-based linear kernel SVM," J. Biomed. Inform., vol. 55, pp. 23–30, 2015.

[17] S. Liu et al., "Drug-drug interaction extraction via convolutional neural networks," Comput. Math. Methods Med., vol. 2016, pp. 1–8, 2016.

[18] W. Zheng et al., "An attention-based effective neural model for drug-drug interactions extraction," BMC Bioinform., vol. 18, art. 445, 2017.

[19] Y. Zhang et al., "BioWordVec, improving biomedical word embeddings with subword information and MeSH," Sci. Data, vol. 6, art. 52, 2019.

[20] J. Pennington, R. Socher, and C. D. Manning, "GloVe: global vectors for word representation," in Proc. EMNLP, 2014, pp. 1532–1543.

[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, pp. 1735–1780, 1997.

[22] A. Vaswani et al., "Attention is all you need," Adv. Neural Inf. Process. Syst., vol. 30, pp. 5998–6008, 2017.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL, 2019, pp. 4171–4186.

[24] Y. Peng, S. Yan, and Z. Lu, "Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets," in Proc. BioNLP, 2019, pp. 58–65.

[25] M. Dou, J. Tang, P. Tiwari, Y. Ding, and F. Guo, "Drug-drug interaction relation extraction based on deep learning: a review," ACM Comput. Surv., vol. 56, pp. 1–33, 2024.

[26] J. Aldahdooh, Z. Tanoli, and J. Tang, "Mining drug-target interactions from biomedical literature using chemical and gene descriptions-based ensemble transformer model," Bioinform. Adv., vol. 4, no. 1, art. vbae106, 2024.

[27] Y. Jia, Z. Yuan, L. Zhu, and Z. Xiang, "A meta-contrastive learning approach for clinical drug-drug interaction extraction from biomedical literature," PLOS Comput. Biol., vol. 21, no. 12, art. e1013722, 2025.

[28] G. De Vito, F. Ferrucci, and A. Angelakis, "LLMs for drug-drug interaction prediction: a comprehensive comparison," arXiv:2502.06890, 2025.

[29] H. Qi, X. Li, C. Zhang, and T. Zhao, "Improving drug-drug interaction prediction via in-context learning and judging with large language models," Front. Pharmacol., vol. 16, art. 1589788, 2025.

[30] C. Boulealam, H. Filali, J. Riffi, A. M. Mahraz, and H. Tairi, "Feedback-based validation learning," Computation, vol. 13, art. 156, 2025.

[31] C. Boulealam, H. Filali, J. Riffi, A. M. Mahraz, and H. Tairi, "Deep multi-component neural network architecture," Computation, vol. 13, art. 93, 2025.