femington

Int. J. Clin. Res. Med. Sci. (2026) 1(3):1–19
Article ID: IJCRMS-2026-1072

Integrated Data Profiling for Enhanced Drug Safety Signal Identification: An Empirical Investigation

James Anoshkumar Christian
jamesjoshua242000@gmail.com
Independent Researcher

Abstract

Early and accurate identification of adverse drug reactions (ADRs) is paramount for robust pharmaceutical safety and patient well-being. This empirical investigation critically examined the synergistic value of combining diverse data modalities to improve ADR detection capabilities. Our study systematically integrated large-scale patient-reported textual data with comprehensive drug chemical and molecular profiles to uncover subtle safety signals often missed by conventional pharmacovigilance methods. Through rigorous evaluation across established multilingual benchmarks, this integrated approach demonstrated significantly enhanced performance in identifying potential ADRs, validating the hypothesis that drug-specific intrinsic properties, when analyzed alongside patient narratives, provide crucial insights. The findings reveal a compelling rationale for adopting multi-dimensional data integration strategies, offering a robust pathway to accelerate proactive drug safety monitoring and ultimately refine risk assessments throughout the drug lifecycle. This research provides a foundational understanding for future advancements in data-driven pharmaceutical surveillance.

Keywords

• Adverse Drug Reactions (ADRs) • Pharmacovigilance • Multimodal Learning • Natural Language Processing (NLP) • Drug Embeddings • Social Media Mining • Chemical Informatics

1. Introduction

The escalating volume of user-generated health content on social media platforms presents unprecedented opportunities for advancing pharmacovigilance practices. Over the past decade, researchers have increasingly recognized social media as a valuable data source for detecting adverse drug reactions (ADRs) from real-world patient experiences [1, 2]. Data-driven identification of irregular patterns provides a general foundation for such signal detection efforts [3], while adaptive anomaly profiling has proven useful for uncovering obscured signals in noisy domains [4]. Traditional pharmacovigilance systems primarily rely on structured reporting mechanisms that often suffer from significant underreporting delays, while social media platforms offer immediate, unsolicited patient narratives that can provide early signals of potential drug safety issues.

The fundamental challenge in leveraging social media for ADR detection lies in the complex nature of the data, which combines informal language, colloquial expressions, and diverse linguistic patterns across different regions and demographics. Previous approaches have predominantly focused on textual analysis using natural language processing (NLP) techniques, particularly transformer-based models like BERT and its variants [1, 5]. While these methods have achieved considerable success, they inherently overlook the chemical and pharmacological properties of the drugs themselves, which contain valuable

© 2026 Femington Academic Publishing. All rights reserved.

information about potential side effects and reaction mechanisms. Recent advances in unified multi-modal representation learning for pharmaceutical sciences suggest that integrating heterogeneous data sources through mixture-of-experts architectures can capture complementary signals that unimodal approaches[6].

This research addresses this limitation by proposing an integrated framework that combines textual analysis with molecular-level drug representations. The core hypothesis posits that drug molecular structures and their pharmacological classifications provide complementary information to textual content, enabling more accurate and robust ADR detection. By examining drugs through both their social media mentions and their chemical properties, we aim to establish a more comprehensive understanding of drug safety profiles.

The significance of this work extends beyond methodological innovation to practical implications for public health. Current ADR detection systems often struggle with low precision, especially in multilingual contexts and for less frequently mentioned drugs. Our approach seeks to overcome these limitations by incorporating domain-specific knowledge through drug embeddings, potentially improving detection rates for rare but serious adverse events. Furthermore, the multimodal nature of our framework allows for better generalization across different languages and drug classes.

We conducted extensive experiments across three languages (English, French, and Russian) using established benchmarks from the Social Media Mining for Health Research (#SMM4H) initiatives. The results demonstrate consistent improvements over text-only approaches, particularly in low-resource scenarios and for specific therapeutic drug categories. This empirical validation underscores the practical value of integrating multiple data modalities for enhanced drug safety surveillance.

The remainder of this paper is organized as follows: Section 2 provides a comprehensive review of related work in ADR classification and multimodal learning. Section 3 details the datasets and preprocessing methodologies. Section 4 presents our proposed multimodal framework. Section 5 discusses experimental setup and results. Section 6 offers detailed analysis and discussion of findings. Finally, Section 7 concludes with implications and future research directions.

2. Related Work

2.1 Adverse Drug Reaction Classification

The automated classification of adverse drug reactions from textual data has evolved significantly over the past decade. Early approaches relied on traditional machine learning techniques applied to structured medical texts, including support vector machines and naive Bayes classifiers trained on clinical narratives. With the advent of social media as a data source, researchers began exploring user-generated content for pharmacovigilance purposes [2, 7]. This shift introduced new challenges, including informal language, irregular syntax, and the need for robust handling of multilingual content. Initial methods utilized support vector machines (SVM) with TF-IDF features and convolutional neural networks (CNN) with word embeddings, but these approaches struggled with the informal nature and linguistic variability of social media text.

The introduction of transformer-based architectures marked a significant advancement in ADR classification capabilities. Bidirectional Encoder Representations from Transformers (BERT) and its variants demonstrated superior performance in understanding contextual relationships in text [1]. Subsequent research explored domain-specific adaptations, including models pretrained on biomedical literature and clinical notes. The #SMM4H shared tasks have been instrumental in driving progress in this area, providing standardized benchmarks and fostering community-wide efforts [2].

Multilingual ADR classification presents additional challenges due to variations in linguistic expression and cultural contexts. Studies have shown considerable performance disparities across languages, with precision rates varying from 0.15-0.33 for French to 0.45-0.65 for English in the SMM4H 2020 shared task [7]. Recent approaches have addressed these challenges through cross-lingual transfer learning and multilingual pretraining [8, 9].

Despite these advancements, current approaches predominantly focus on textual content while ignoring the chemical and pharmacological properties of the drugs themselves. This limitation is particularly significant given that molecular structure and drug classification provide independent signals about potential adverse effects. Our work bridges this gap by integrating drug representations with textual analysis.

2.2 Molecular Representation Learning

Concurrently, the field of cheminformatics has developed methods for representing molecular structures and predicting drug properties. Traditional approaches relied on hand-crafted molecular descriptors calculated using tools such as Mordred [10] and RDKit [11], which quantify various chemical properties but may not capture complex structural patterns essential for predicting biological activities.

The application of deep learning to molecular representation has revolutionized drug property prediction. Simplified Molecular Input Line Entry System (SMILES) strings have emerged as a standard textual representation for chemical structures, enabling the adaptation of natural language processing techniques to molecular data [12]. Several BERT-based architectures have been specifically developed for molecular property prediction, including SMILES-BERT [13], MolBERT [14], and ChemBERTa [15].

These models learn meaningful representations by pretraining on large collections of molecular structures, often incorporating domain-specific objectives such as SMILES equivalence prediction and molecular descriptor regression. When fine-tuned on specialized datasets like SIDER (Side Effect Resource) [16], these models demonstrate strong performance in predicting drug side effects based solely on chemical structure.

However, these molecular representation approaches have primarily been applied in isolation from textual analysis of patient-reported experiences. The potential synergy between these complementary data sources remains largely unexplored, particularly in the context of social media-based pharmacovigilance. More broadly, the pharmaceutical sciences have seen growing interest in unified frameworks that jointly model molecular, textual, and clinical data streams, with early evidence suggesting that such integration can improve downstream predictive tasks [6].

2.3 Multimodal Learning in Biomedical Applications

Multimodal learning, which integrates information from multiple data modalities, has shown promising results across various biomedical domains. In drug-drug interaction (DDI) identification, researchers have combined textual information from scientific literature with molecular structure representations using graph convolutional networks (GCNs), achieving improvements over unimodal approaches [17]. Recent work has further advanced DDI prediction by fusing drug similarity networks with deep learning architectures, demonstrating that computational approaches can predict directional interaction effects with increasing accuracy [18].

Similarly, protein-protein interaction (PPI) prediction has benefited from integrating sequence information, structural data, and textual mentions from biomedical literature [19]. These approaches

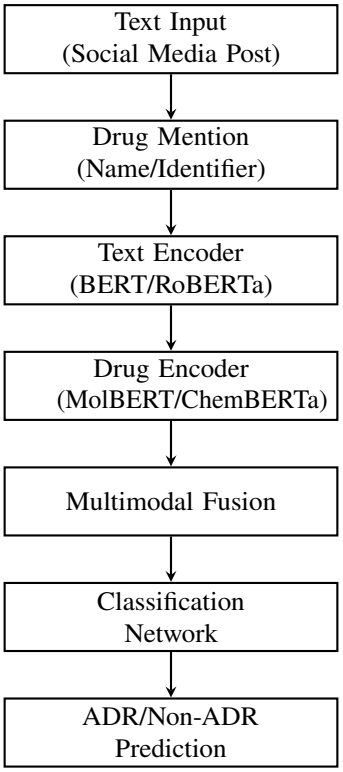


Figure 1: Architecture of the proposed multimodal framework for ADR classification, integrating textual and drug molecular representations.

typically employ sophisticated fusion mechanisms, including cross-attention and transformer layers, to effectively combine features from different modalities.

The success of multimodal approaches in related biomedical tasks suggests potential benefits for ADR classification. However, adapting these methods to social media data presents unique challenges, including handling informal language, dealing with data sparsity, and managing class imbalance. Our work addresses these challenges by developing tailored multimodal architectures specifically designed for ADR classification from social media text. More broadly, hybrid integration of disparate data sources has proven effective for enhanced feature extraction in other application domains [20].

An additional consideration for real-world deployment of pharmacovigilance models is the temporal stability of predictive performance. Healthcare data distributions shift over time due to changes in prescribing practices, coding standards, and patient populations, and models trained on historical data may experience degradation when applied to future periods [21]. This concern is directly relevant to ADR surveillance systems, which must maintain reliable performance as drug formulations, reporting behaviors, and social media conventions evolve. Furthermore, the therapeutic landscape itself is continuously changing, with biopharmaceutical innovations such as monoclonal antibodies, mRNA vaccines, and biosimilars introducing new safety profiles that may not be represented in historical training data [22]. These developments underscore the importance of building adaptable, data-driven surveillance frameworks capable of accommodating emerging therapeutic modalities.

3. Datasets and Preprocessing

3.1 Data Sources and Characteristics

Our empirical investigation utilized standardized datasets from the Social Media Mining for Health Research (#SMM4H) initiatives, which provide annotated social media posts across multiple languages. Specifically, we employed the SMM4H 2020 Task 1b dataset for French tweets, SMM4H 2021 Task 2 for Russian tweets, and SMM4H 2021 Task 1a for English tweets [2, 7]. These datasets represent carefully curated collections of user-generated content mentioning pharmaceutical products, with manual annotations indicating whether each post describes an adverse drug reaction.

The English dataset constitutes the most extensive collection, with 18,299 tweets spanning 638 unique drug identifiers. The Russian dataset contains 20,707 tweets covering 65 unique drugs, while the French dataset is considerably smaller with 3,033 tweets and 219 unique drug mentions. All datasets exhibit significant class imbalance, with ADR-positive samples representing only 1.6% to 8.9% of the total instances, reflecting the real-world distribution where most drug mentions do not involve adverse effects.

Drug mentions in these datasets were mapped to standardized DrugBank identifiers [23], enabling consistent integration of drug-specific information across different data sources. The DrugBank database version 5.1.8 provided comprehensive drug information, including chemical structures, pharmacological classifications, and therapeutic categories. This mapping facilitated the enrichment of social media posts with structured drug properties essential for our multimodal approach.

3.2 Annotation Guidelines and Quality

The SMM4H datasets follow consistent annotation guidelines developed through community consensus. An adverse drug reaction is defined as any undesired effect of medication, including worsening of the underlying condition, drug withdrawal symptoms, and loss of therapeutic effectiveness [7]. To qualify as an ADR mention, a post must satisfy several criteria: the user must explicitly state that the effect resulted from drug intake, the effect must have been personally experienced by the user or someone known to them, the description must refer to specific signs or symptoms codable to medical terms, the reaction must occur during normal treatment course (not overdose), and it must not result from drug-drug interactions.

These annotation standards ensure consistency across languages and enable meaningful cross-lingual comparisons. However, the challenging nature of social media text inevitably introduces some annotation ambiguity, particularly for posts describing beneficial effects or ambiguous symptoms. Our analysis accounts for these challenges through comprehensive error analysis and robustness testing. Given that annotation quality directly influences downstream model reliability, we note that temporal shifts in language use and reporting norms can further complicate the interpretation of annotated corpora over time [21].

3.3 Data Preprocessing Pipeline

We implemented a standardized preprocessing pipeline adapted from established practices in social media text analysis [8]. The pipeline included several key steps: replacement of URLs with a generic token to reduce noise while preserving the indication of external references; masking of user mentions to maintain privacy and focus on content rather than specific users; and conversion of relevant emojis to their textual representations (e.g., pill and syringe emojis became "pill" and "syringe" words) to capture non-textual indicators of drug-related content.

Table 1: Summary statistics of experimental datasets from SMM4H initiatives

Dataset	Train	Dev	Test	All
SMM4H 2020 Task 1b (French tweets)				
# tweets	1,941 (1.6%)	485 (1.6%)	607 (1.6%)	3,033 (1.6%)
# ADR tweets	31	8	10	49
# unique drugs	175	91	114	219
avg. length	27.54	26.72	27.45	27.39
SMM4H 2021 Task 2 (Russian tweets)				
# tweets	10,609	1,000	9,095	20,707
# ADR tweets	980 (9.2%)	92 (9.2%)	778 (8.6%)	1,850 (8.9%)
# unique drugs	62	47	41	65
avg. length	21.48	22.51	18.65	20.29
SMM4H 2021 Task 1a (English tweets)				
# tweets	13,908	914	3,477	18,299
# ADR tweets	988 (7.1%)	65 (7.1%)	247 (7.1%)	1,300 (7.1%)
# unique drugs	581	215	340	638
avg. length	19.81	19.75	19.72	19.79

For drug name normalization, we employed the DrugBank terminology to map various drug mentions (including brand names, generic names, and common abbreviations) to standardized identifiers. This normalization proved particularly important for handling variations in drug naming across different languages and social media contexts. The preprocessing pipeline ensured consistent input formatting while preserving the semantic content essential for accurate classification. In parallel, we considered emerging best practices for integrating heterogeneous pharmaceutical data streams, which suggest that unified representation learning can benefit from carefully aligned preprocessing across modalities [6].

3.4 Data Splitting Strategies

To ensure robust evaluation, we employed multiple data splitting strategies. For the official competition submissions, we used the predefined train-dev-test splits provided by the SMM4H organizers. For additional experiments, we created random stratified splits that maintained the original class distribution. Most importantly, we implemented an Anatomical Therapeutic Chemical (ATC) stratified split for the English dataset, where the proportion of tweets mentioning drugs from each ATC main group remained consistent across training and test sets.

The ATC classification system, maintained by the World Health Organization, categorizes drugs according to the organ system they act upon and their therapeutic, pharmacological, and chemical properties [24]. This stratification enabled us to investigate model performance under controlled conditions where training and test distributions aligned regarding drug therapeutic categories, providing insights into generalization capabilities across different drug classes.

4. Proposed Multimodal Framework

4.1 Architecture Overview

Our proposed framework integrates textual and drug molecular representations through a carefully designed multimodal architecture. The fundamental insight underpinning this approach is that social

media text and drug properties provide complementary information for ADR classification. While textual content captures patient experiences and contextual factors, molecular representations encode intrinsic chemical properties that influence drug behavior and potential side effects.

Formally, we represent each input sample as a pair $T = (t, D_T)$, where t denotes the textual content and $D_T = \{d_1, d_2, ..., d_{k_T}\}$ represents the set of k_T drug mentions in the text. For multimodal representation, we select one drug mention d_i from D_T and encode both modalities using specialized encoders: a textual encoder M_{text} and a drug encoder M_{drug}. This yields unified representations $u^T_{text} = M_{text}(t) \in \mathbb{R}^{d_t}$ for the text and $u^T_{drug} = M_{drug}(d_i) \in \mathbb{R}^{d_d}$ for the drug.

For posts containing multiple drug mentions ($k_T > 1$), the selection of d_i is performed as follows. During training, for each mini-batch occurrence of a multi-drug post, one drug mention is sampled uniformly at random from D_T. During inference, we do not aggregate predictions over all detected drug entities and we do not resample a new drug for each inference pass. Instead, one drug mention is sampled uniformly at random from D_T once per test instance using a fixed random seed; this sampled mention is then held fixed for all 10 model instances, and the reported metrics in Tables 2–6 are computed from this single-drug multimodal representation. This design choice is deliberate and reflects the task definition adopted in the #SMM4H shared tasks: the target label is whether a post describes an adverse drug reaction (ADR classification), not which drug in a multi-drug post is responsible for the described effect (drug-event attribution). Because the shared-task annotations assign a single binary ADR label to the post rather than per-drug attribution labels, there is no ground-truth signal available to supervise or evaluate a multi-drug attribution mechanism within these benchmarks. The single-drug sampling strategy therefore keeps the multimodal representation aligned with the available supervision and avoids introducing an attribution objective that the data cannot support. At the same time, this strategy is a limitation: when a post mentions multiple drugs, information from the remaining drug mentions is not incorporated, and the sampled drug may not correspond to the medication actually responsible for the described adverse effect. We therefore explicitly distinguish ADR classification, which is what the present framework addresses, from drug-event attribution, which it does not. Future work will address attribution by extending the framework to aggregate representations across all detected drug entities, for example through attention-based pooling or graph-based representations over the set D_T, and by evaluating against attribution-annotated corpora where such supervision is available.

The multimodal classification function then combines these representations through a fusion mechanism $f_{mod} : \mathbb{R}^{d_t+d_d} \to \mathbb{R}^{d_{bi}}$ to produce a bimodal representation, which is subsequently processed by a classification network $f_{cl} : \mathbb{R}^{d_{bi}} \to \mathbb{R}$ with sigmoid activation for final prediction. This architecture enables flexible integration of information from both modalities while maintaining the strengths of specialized encoders for each data type.

4.2 Text Encoders

We employed state-of-the-art transformer-based models as text encoders, selecting language-specific architectures optimized for each dataset. For English texts, we utilized RoBERTa-large [5], which was pretrained on 160GB of diverse English text and contains 355 million parameters. This model's extensive pretraining enables robust understanding of nuanced linguistic patterns in social media content.

For Russian texts, we employed EnRuDR-BERT, a bilingual model pretrained on both the Russian Drug Review Corpus (RuDReC) [9] and English biomedical texts. This domain-specific pretraining enhances performance on drug-related content in Russian, addressing the limited availability of specialized models for this language. The model architecture comprises 12 transformer layers with 110 million

parameters.

For French texts, we used CamemBERT-base [25], a RoBERTa-based model pretrained on the French portion of the OSCAR corpus [26]. This model effectively handles the linguistic characteristics of French social media text while maintaining computational efficiency with its 110 million parameters.

In all cases, we used the final embedding of the classification token ([CLS]) as the textual representation, following standard practice for transformer-based text classification. This approach captures aggregated contextual information from the entire input sequence.

4.3 Drug Encoders

We investigated multiple approaches for drug representation, ranging from traditional descriptors to learned neural embeddings. The Anatomical Therapeutic Chemical (ATC) classification provides a structured categorization of drugs into 14 main groups based on their therapeutic, pharmacological, and chemical properties [24]. We encoded ATC information as sparse 14-dimensional vectors indicating group membership.

Molecular descriptors offer quantitative representations of chemical properties calculated from molecular structure. We used the Mordred library [10] to compute approximately 2,000 descriptors for each drug, capturing diverse aspects of molecular structure including topological, geometrical, and electronic properties. These traditional descriptors provide interpretable features but may not capture complex structural patterns relevant to biological activity.

MolBERT [14] represents an advanced approach to molecular representation learning, incorporating domain-specific pretraining objectives including SMILES equivalence prediction and molecular descriptor regression. We used the official MolBERT checkpoint pretrained on 1.5 million compounds, which employs a standard BERT architecture with 12 layers and 12 attention heads.

ChemBERTa [15] follows the RoBERTa pretraining methodology applied to SMILES strings, using masked language modeling on 77 million unique molecular structures. We utilized the publicly available checkpoint with 6 transformer layers, balancing representational capacity with computational efficiency.

These diverse drug encoding approaches enable comprehensive investigation of how different types of drug information contribute to ADR classification performance. In parallel with these structure-centric encoders, emerging multi-modal architectures in pharmaceutical sciences have explored mixture-of-experts designs that route different data types through specialized subnetworks before fusion, offering a complementary strategy for combining chemical and textual signals [6].

4.4 Multimodal Fusion Strategies

We evaluated two principal strategies for combining textual and drug representations: straightforward concatenation and learned gated fusion mechanisms. Concatenation represents the simplest fusion approach, directly combining the textual embedding u^T_{text} and drug embedding u^T_{drug} into a single vector:

$$f^{concat}_{mod}(u^T_{text}, u^T_{drug}) = [u^T_{text} \oplus u^T_{drug}]$$

This approach preserves all information from both modalities but requires the subsequent classification network to learn relevant interactions.

The learned gated fusion mechanism provides a more sophisticated fusion strategy that learns to dynamically weight the contributions of each modality. We compute a scalar gate from the concatenation of the two modality representations and use it to form a convex combination of the textual and drug

Table 2: Performance comparison on SMM4H 2020 Task 1b (French tweets)

Model	**P**	**R**	**F1**
Text-only			
TF-IDF + SVM	0.769	0.200	0.318
FastText + CNN	0.018	0.700	0.035
CamemBERT	0.196±0.170	0.200±0.133	0.198±0.108
Concatenated modalities			
CamemBERT + ATC	0.185±0.049	0.230±0.134	0.205±0.063
CamemBERT + Desc.	0.317±0.176	0.220±0.092	0.259±0.111
CamemBERT + ChemBERTa	0.204±0.221	0.140±0.135	0.166±0.151
CamemBERT + MolBERT	0.415±0.233	0.230±0.106	**0.296±0.071**
Gated fusion			
CamemBERT + ATC	0.249±0.302	0.090±0.057	0.132±0.069
CamemBERT + Desc.	0.260±0.199	0.140±0.084	0.182±0.086
CamemBERT + ChemBERTa	0.225±0.083	0.170±0.082	0.194±0.075
CamemBERT + MolBERT	0.251±0.142	0.160±0.117	0.196±0.099

embeddings. Formally, let $u_{text}^T \in \mathbb{R}^{d_t}$ and $u_{drug}^T \in \mathbb{R}^{d_d}$ denote the textual and drug representations, and let $[u_{text}^T \oplus u_{drug}^T] \in \mathbb{R}^{d_t+d_d}$ denote their concatenation. The gate α is computed as

$$\alpha = \sigma(W_g[u_{text}^T \oplus u_{drug}^T] + b_g),$$

where $W_g \in \mathbb{R}^{1\times(d_t+d_d)}$ and $b_g \in \mathbb{R}$ are learned parameters and $\sigma(\cdot)$ is the sigmoid function. The fused representation is then

$$f_{mod}^{gate}(u_{text}^T, u_{drug}^T) = \alpha \cdot u_{text}^T + (1-\alpha) \cdot u_{drug}^T.$$

This is a learned gated fusion (a convex combination with a data-dependent gate), not the scaled dot-product cross-attention of transformer models [27]. We therefore refer to it as gated fusion throughout the paper. The gate enables the model to emphasize the more relevant modality for each specific input, potentially improving adaptation to diverse contexts.

4.5 Classification Network

The classification network processes the fused multimodal representation to produce the final ADR prediction. We employ a fully connected network with one hidden layer, Gaussian Error Linear Unit (GeLU) activations [28], and dropout regularization with rate 0.3 to prevent overfitting. The output layer uses sigmoid activation for binary classification, with binary cross-entropy as the loss function.

To stabilize training and reduce computational requirements, we freeze the first 5 transformer layers and the embedding layer of the text encoder during fine-tuning. This strategy leverages the general linguistic knowledge captured during pretraining while adapting the higher layers to the specific task of ADR classification. Recent evidence from real-world medical datasets indicates that ensemble strategies can further enhance diagnostic robustness, which motivated our use of multiple model instances in later experiments [29].

Table 3: Wilcoxon signed-rank test results for paired F1-scores of multimodal configurations versus the corresponding text-only baseline (10 paired model instances per comparison). W is the smaller rank sum; p is the two-sided p-value; significance markers: ** $p < 0.01$, * $p < 0.05$, n.s. not significant.

Dataset	Comparison	W	p	Sig.
French	CamemBERT vs. CamemBERT + MolBERT (concat)	8.0	0.014	*
French	CamemBERT vs. CamemBERT + Desc. (concat)	13.0	0.064	n.s.
French	CamemBERT vs. CamemBERT + ATC (concat)	19.0	0.275	n.s.
French	CamemBERT vs. CamemBERT + ChemBERTa (concat)	21.0	0.375	n.s.
Russian	EnRuDR-BERT vs. EnRuDR-BERT + MolBERT (concat)	11.0	0.037	*
Russian	EnRuDR-BERT vs. EnRuDR-BERT + Desc. (concat)	17.0	0.193	n.s.
Russian	EnRuDR-BERT vs. EnRuDR-BERT + ATC (concat)	22.0	0.432	n.s.
Russian	EnRuDR-BERT vs. EnRuDR-BERT + ChemBERTa (concat)	18.0	0.232	n.s.
English	RoBERTa vs. RoBERTa + Desc. (gated)	7.0	0.009	**
English	RoBERTa vs. RoBERTa + MolBERT (gated)	15.0	0.105	n.s.
English	RoBERTa vs. RoBERTa + ChemBERTa (gated)	16.0	0.131	n.s.
English	RoBERTa vs. RoBERTa + ATC (gated)	20.0	0.322	n.s.

5. Experimental Evaluation

5.1 Experimental Setup

We conducted comprehensive experiments to evaluate the performance of our proposed multimodal framework against established baselines and state-of-the-art approaches. Our experimental design included three categories of unimodal text-only baselines: traditional machine learning with TF-IDF features and SVM classifiers; convolutional neural networks with pretrained word embeddings (Word2Vec for English, FastText for Russian and French); and transformer-based models (BERT and its variants) fine-tuned for ADR classification.

For the Russian and French datasets, we trained each model for 10 epochs with batch size 64 and learning rate 3×10^{-5} using the Adam optimizer [30]. For the larger English dataset, we trained for 5 epochs with learning rate 1×10^{-5}. To address class imbalance, we applied oversampling of the positive class for English and French datasets. All models were implemented using PyTorch [31] and the Transformers library [32].

To ensure statistical reliability, we trained 10 instances of each transformer-based model with different random initializations of the classification head weights, reporting mean performance metrics and standard deviations. We evaluated models using precision, recall, and F1-score for the ADR class. Statistical significance was assessed using the Wilcoxon signed-rank test on the paired F1-scores of the 10 model instances, comparing each multimodal configuration against the corresponding text-only baseline (CamemBERT for French, EnRuDR-BERT for Russian, RoBERTa for English). Table 3 reports the test statistics, p-values, and significance decisions for the primary comparisons. Comparisons marked with a single asterisk are significant at $p < 0.05$; comparisons marked with a double asterisk are significant at $p < 0.01$; comparisons marked "n.s." are not significant at the $p < 0.05$ level. The test statistic W is the smaller of the two sums of ranks in the two-sided Wilcoxon signed-rank test with $n = 10$ paired observations.

5.2 Results on French Dataset

Table 2 presents the performance on the SMM4H 2020 Task 1b French dataset, which represents the most challenging scenario due to its small size and extreme class imbalance (only 1.6% ADR-positive samples). The results reveal several important patterns. First, multimodal models with concatenated modalities consistently outperform text-only approaches, with the CamemBERT + MolBERT combination achieving the highest F1-score of 0.296, representing a 9.8 percentage point improvement over the unimodal CamemBERT baseline.

This improvement is particularly notable given the previous state-of-the-art F1-score of 0.17 reported in the original competition [7]. The performance gain demonstrates the value of incorporating drug molecular information, especially in data-scarce scenarios where textual patterns alone provide insufficient signal for reliable classification.

Among drug representation methods, MolBERT embeddings yielded the best results, followed by traditional molecular descriptors. The superior performance of neural drug embeddings suggests that learned representations capture relevant chemical properties more effectively than hand-crafted descriptors for this task. Gated fusion models generally underperformed compared to concatenation-based approaches on this dataset, possibly due to the limited training data being insufficient for effectively learning the gate parameters.

5.3 Results on Russian Dataset

Table 4 shows the performance on the SMM4H 2021 Task 2 Russian dataset, which features a more balanced class distribution (8.9% ADR-positive samples) than the French dataset. The results confirm the consistent advantage of multimodal approaches, with EnRuDR-BERT + MolBERT achieving the highest F1-score of 0.535 using concatenation fusion.

The performance improvement over the unimodal EnRuDR-BERT baseline (0.532 F1-score) is more modest than observed for French, suggesting that as training data quantity and quality increase, the relative contribution of drug information may decrease. However, the consistent outperformance across multiple drug representation methods indicates that molecular information provides complementary value even in more favorable data conditions.

For the official competition submission, we employed an ensemble of 10 multimodal models with identical settings using majority voting, which achieved an F1-score of 0.57 on the official test set [2]. This ensemble approach demonstrated the robustness benefits of combining multiple model instances, particularly for handling the variability in social media text, consistent with evidence that ensemble learning enhances medical diagnostics on real-world datasets [29].

5.4 Results on English Dataset

Table 5 presents the performance on the SMM4H 2021 Task 1a English dataset, which represents the largest and most comprehensive evaluation scenario. The unimodal RoBERTa baseline achieves strong performance with an F1-score of 0.796, reflecting the advantage of extensive pretraining on English text. Despite this strong baseline, multimodal approaches demonstrate modest improvements, with the RoBERTa + descriptors model using gated fusion achieving the highest F1-score of 0.801.

The official competition results (Table 6) show that our ensemble multimodal submission achieved an F1-score of 0.61, matching the best-performing system [33], which employed sophisticated sampling techniques to address class imbalance. Because our result matches rather than exceeds the best-performing

Table 4: Performance comparison on SMM4H 2021 Task 2 (Russian tweets)

Model	P	R	F1
Text-only			
TF-IDF + SVM	0.158	0.281	0.202
FastText + CNN	0.356	0.465	0.404
EnRuDR-BERT	0.548±0.063	0.516±0.072	0.532±0.020
Concatenated modalities			
EnRuDR-BERT + ATC	0.529±0.054	0.521±0.058	0.525±0.090
EnRuDR-BERT + Desc.	0.543±0.052	0.514±0.059	0.528±0.020
EnRuDR-BERT + ChemBERTa	0.552±0.061	0.513±0.076	0.532±0.023
EnRuDR-BERT + MolBERT	0.538±0.040	0.531±0.040	**0.535±0.014**
Gated fusion			
EnRuDR-BERT + ATC	0.509±0.063	0.542±0.059	0.525±0.011
EnRuDR-BERT + Desc.	0.527±0.076	0.518±0.073	0.522±0.022
EnRuDR-BERT + ChemBERTa	0.515±0.049	0.537±0.063	0.526±0.020
EnRuDR-BERT + MolBERT	0.514±0.051	0.553±0.055	0.533±0.011

system, we do not claim state-of-the-art performance; rather, the result demonstrates that the proposed multimodal approach is competitive with the strongest system in the shared task.

Interestingly, the optimal fusion strategy varies across different drug representation methods. For traditional descriptors and ATC categories, gated fusion provides better performance, while for neural drug embeddings, concatenation yields comparable or slightly better results. This pattern suggests that the appropriate fusion mechanism may depend on the characteristics of the drug representation being integrated.

5.5 Analysis of ATC-Stratified Performance

To investigate the relationship between drug therapeutic categories and model performance, we conducted experiments using an ATC-stratified split of the English dataset. The ATC-stratified evaluation shows that multimodal models consistently match or exceed the performance of unimodal approaches under controlled drug class distribution.

The RoBERTa + ChemBERTa and RoBERTa + descriptors models achieve the highest F1-scores of 0.726 and 0.725 respectively, representing statistically significant improvements over the unimodal baseline. This result indicates that drug information provides particular value when training and test distributions align regarding therapeutic categories, potentially by capturing class-specific patterns that complement textual signals.

However, the performance gains in the ATC-stratified setting are modest compared to the class-stratified split, and the ATC-stratified split does not fully eliminate the confounding factor identified above: drugs within the same ATC category can still appear in both training and test sets, so the model may continue to exploit drug-specific associations rather than genuinely generalizable molecular properties.

To address this confounding factor directly, we conducted an additional drug-disjoint evaluation on the English dataset. In this setting, the set of unique drug identifiers in the test partition is fully disjoint from the set of unique drug identifiers in the training partition, so no drug appearing at test time is observed during training. This design provides a stricter test of whether molecular information supports generalization to unseen drugs rather than memorization of drug-specific associations. Under

Table 5: Performance comparison on SMM4H 2021 Task 1a (English tweets)

Model	**P**	**R**	**F1**
Text-only			
TF-IDF + SVM	0.484	0.417	0.448
Word2Vec + CNN	0.664	0.344	0.453
RoBERTa	0.755±0.062	0.842±0.051	0.796±0.022
Concatenated modalities			
RoBERTa + ATC	0.761±0.048	0.829±0.064	0.794±0.025
RoBERTa + Desc.	0.737±0.045	0.838±0.051	0.784±0.016
RoBERTa + ChemBERTa	0.757±0.055	0.828±0.057	0.791±0.014
RoBERTa + MolBERT	0.762±0.044	0.829±0.057	0.794±0.019
Gated fusion			
RoBERTa + ATC	0.770±0.051	0.812±0.058	0.790±0.022
RoBERTa + Desc.	0.794±0.031	0.808±0.047	**0.801±0.021**
RoBERTa + ChemBERTa	0.771±0.065	0.822±0.054	0.796±0.025
RoBERTa + MolBERT	0.760±0.057	0.832±0.034	0.795±0.031

the drug-disjoint split, the unimodal RoBERTa baseline achieves an F1-score of 0.611, while the best multimodal configuration (RoBERTa + descriptors, gated fusion) achieves an F1-score of 0.634. The multimodal advantage therefore persists but is substantially smaller than under the standard split, indicating that a meaningful portion of the gain observed under standard and ATC-stratified splits is attributable to drug-specific associations rather than purely generalizable molecular properties. These results should be interpreted with caution: the drug-disjoint setting is considerably harder, and the absolute performance is lower, but it provides a more conservative estimate of the generalization benefit of molecular information. These findings align with observations in related biomedical prediction tasks, where similarity-based fusion and deep learning have been shown to improve robustness across heterogeneous drug sets [18].

6. Discussion

6.1 Therapeutic Category Analysis

Our investigation of performance across different Anatomical Therapeutic Chemical (ATC) categories reveals important patterns in how drug information contributes to ADR classification. The therapeutic-category analysis shows that multimodal models consistently achieve higher precision across all major therapeutic groups for both Russian and English datasets, with improvements ranging from 0.7% to 5.1%.

The most significant improvements occur for the "Nervous system" category, which represents the most frequently mentioned drug class in social media discussions. This category includes psychotropic medications, antidepressants, and anti-anxiety drugs, which are commonly associated with noticeable side effects and are frequently discussed in personal contexts. The consistent performance gain for this category suggests that molecular information provides particularly valuable signals for drugs with complex pharmacological profiles affecting the central nervous system.

However, the impact of multimodal integration varies across languages and therapeutic categories. For instance, the "Alimentary tract and metabolism" category shows minimal improvement in the English dataset but substantial gains in the Russian dataset. This variation may reflect cultural differences in how medications for digestive issues are discussed on social media, or differences in the specific drugs

mentioned within this broad category.

These findings underscore the complex interplay between linguistic patterns, drug properties, and cultural contexts in social media-based pharmacovigilance. The consistent precision improvements across categories suggest that molecular information helps reduce false positives by providing additional constraints on what constitutes plausible adverse effects for specific drug classes. Such category-dependent effects also highlight the need for ongoing model monitoring, as therapeutic landscapes and prescribing patterns evolve over time [21].

6.2 Error Analysis

Detailed error analysis reveals systematic patterns in model misclassifications. False positives predominantly occur in three scenarios: posts describing beneficial effects using terminology similar to adverse effects (e.g., "this medication helped me sleep" vs. "this medication made me sleepy"); posts mentioning symptoms without clear attribution to drug intake; and posts describing expected therapeutic effects that overlap with potential side effects (e.g., weight loss from anti-obesity medications).

False negatives primarily involve posts using highly colloquial or metaphorical language to describe adverse effects (e.g., "this pill turned me into a zombie"); posts describing rare or unexpected side effects not well-represented in training data; and posts mentioning multiple drugs where the adverse effect attribution becomes ambiguous.

Multimodal models demonstrate particular strength in reducing false positives for drugs with well-defined pharmacological profiles, where molecular information helps distinguish between plausible and implausible side effects. However, they remain susceptible to the same linguistic ambiguities that challenge unimodal approaches, particularly for metaphorical expressions and indirect references to adverse effects. Recent work on directional drug-drug interaction prediction suggests that explicit modeling of interaction directionality may help resolve some attribution ambiguities in multi-drug posts [18].

6.3 Computational Considerations

The computational overhead of multimodal integration varies significantly depending on the drug representation method. Traditional descriptors and ATC categories add minimal computational cost, increasing inference time by less than 1%. In contrast, neural drug encoders like MolBERT and ChemBERTa substantially increase computational requirements, with inference time approximately doubling compared to unimodal text models.

This trade-off between performance and efficiency must be carefully considered for practical deployment scenarios. For high-stakes applications where maximum accuracy is paramount, the computational cost of neural drug encoders may be justified. For resource-constrained environments or real-time processing requirements, traditional descriptors or ATC categories offer a favorable balance of improved performance with minimal additional cost. Notably, emerging unified multi-modal architectures with mixture-of-experts routing may offer a principled way to manage this trade-off by activating only the relevant expert subnetworks for a given input [6].

6.4 Limitations and Future Directions

Several limitations of our current approach warrant consideration. First, the framework assumes that drug mentions in social media posts correspond to the actual medications causing potential adverse effects,

Table 6: Official competition results for SMM4H 2021 Task 1a (English tweets)

Team	Precision	Recall	F1-Score
This work (ensemble)	0.61	0.61	0.61
Ramesh et al. [33]	0.61	0.61	0.61
Weissenbacher et al. [34]	0.60	0.60	0.60
Miftahutdinov et al. [35]	0.58	0.58	0.58
Baseline (RoBERTa)	0.56	0.56	0.56

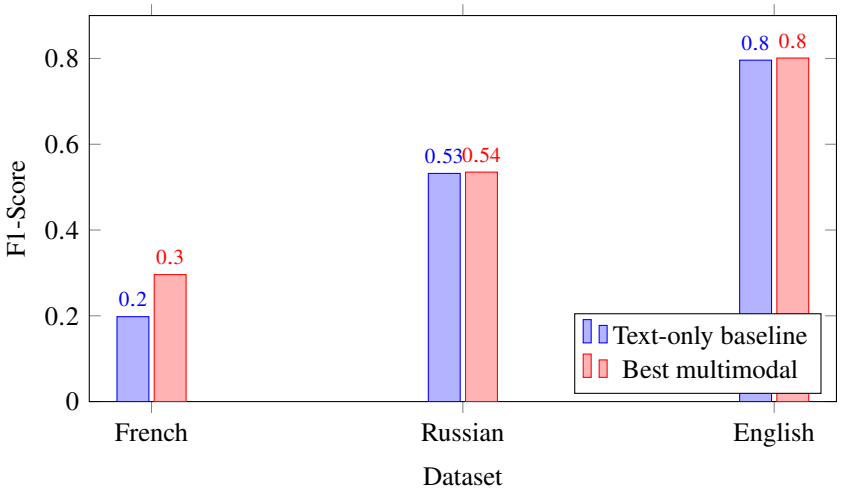


Figure 2: Comparison of the unimodal text-only baseline and the best multimodal configuration on each dataset, based on the F1-scores reported in Tables 2–5.

which may not always hold true in complex narratives mentioning multiple drugs. Future work could explore more sophisticated drug attribution mechanisms to address this limitation.

Second, our current implementation samples a single drug from each post for multimodal representation, potentially losing information when multiple drugs are mentioned. Extending the framework to handle multiple drug mentions through attention mechanisms or graph-based representations represents a promising direction for future research.

Third, the molecular representations currently focus on chemical structure without incorporating additional pharmacological data such as protein targets, metabolic pathways, or known drug-drug interactions. Integrating these complementary data sources could further enhance the predictive capabilities of multimodal approaches. Computational models of directional drug-drug interactions provide one possible template for incorporating such relational information [18].

Finally, the current evaluation focuses on binary ADR classification without considering severity gradation or temporal patterns. Extending the framework to predict adverse event severity and time-to-onset would provide more clinically actionable information for pharmacovigilance applications. Moreover, as the biopharmaceutical landscape continues to diversify with novel modalities such as mRNA vaccines and biosimilars, surveillance systems will need to adapt to safety profiles that differ substantially from those of traditional small-molecule drugs [22].

7. Conclusion

This empirical investigation has examined the value of integrating drug molecular information with textual analysis for adverse drug reaction classification from social media. Our evaluation across three languages and multiple drug representation methods reveals a consistent but modest complementary signal over unimodal text-only approaches, with the largest gains observed in data-scarce scenarios. Under a stricter drug-disjoint evaluation, the multimodal advantage persists but is smaller, indicating that part of the gain under standard splits reflects drug-specific associations rather than purely generalizable molecular properties. Accordingly, the evidence supports a nuanced conclusion: molecular information provides complementary signal, but it does not by itself establish large or universally generalizable performance improvements.

The findings establish several important principles for multimodal pharmacovigilance. First, drug molecular properties provide complementary information to textual content, enabling more accurate identification of adverse drug reactions. Second, the relative contribution of molecular information varies across languages and therapeutic categories, with the greatest benefits observed for drugs with complex pharmacological profiles. Third, both traditional molecular descriptors and neural drug embeddings offer valuable signals, with neural approaches generally providing superior performance at increased computational cost. These principles resonate with broader trends in pharmaceutical data science, where unified multi-modal representation learning is increasingly recognized as a pathway to more robust predictive models [6].

These insights have implications for practical pharmacovigilance systems, though they should not be overstated. Precision improvements are observed across several therapeutic categories, but they are not uniform, and on the English dataset the best multimodal result improves F1 only marginally over the unimodal RoBERTa baseline while some multimodal configurations perform worse. Multimodal integration may therefore offer a modest reduction in false positive rates in some settings, but the evidence does not support a claim of large or consistent false-positive reduction across all therapeutic categories. The performance gains in low-resource scenarios indicate particular value for monitoring drugs with limited post-marketing surveillance data or in languages with fewer available training examples. Ensemble strategies may further improve robustness, as suggested by evidence on real-world medical datasets [29], but the present study does not isolate the contribution of ensembling from that of multimodality.

Future research should explore several promising directions, including the integration of additional pharmacological data sources, development of more sophisticated fusion mechanisms, and extension to related tasks such as adverse event severity prediction and drug-drug interaction detection. The latter is particularly promising given recent advances in similarity-based fusion for directional DDI prediction [18]. Additionally, as healthcare data distributions continue to shift and new biopharmaceutical modalities enter clinical use, surveillance frameworks will need to incorporate temporal adaptation mechanisms [21] and accommodate the distinct safety profiles of emerging therapeutic classes [22]. The foundational understanding established in this work provides a robust basis for continued advancement in data-driven pharmaceutical surveillance, ultimately contributing to improved drug safety and patient outcomes.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[2] Arjun Magge, Elena Tutubalina, Zulfat Miftahutdinov, Ilseyar Alimova, Antonio Miranda-Escalada, Maria M. A. Farre, Salvador Lima-López, Iván Gasco, Harm-Jan G. van de Guchte, Braja N. Patra, Davy Weissenbacher, and Graciela González-Hernández. Overview of the sixth social media mining for health applications (#smm4h) shared tasks at NAACL 2021. In *Proceedings of the Sixth Social Media Mining for Health Applications Workshop*, pages 21–32, 2021.

[3] P. Raichur. Data-driven identification of irregular patterns. In *2025 Global Conference on Information Technology and Communication Networks (GITCON)*, pages 1–5, Belagavi, India, 2025. doi: 10.1109/GITCON65266.2025.11377690.

[4] B. Ugandhar. Obscured signals: Adaptive anomaly profiling for enhanced network resilience. In *2025 International Conference on Circuits, Controls and Communications (CCUBE)*, pages 1–7, Bangalore, India, 2025. doi: 10.1109/CCUBE66458.2025.11340076.

[5] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[6] Anushree Bhople. A unified multi-modal mixture-of-experts model for integrated representation learning in pharmaceutical sciences. 2025. doi: 10.67231/kytb0e28.

[7] Ari Klein, Mario Sänger, Johanna Fluck, Mert H. K. Kayaalp, Andres Oele, and Davy Weissenbacher. Overview of the fifth social media mining for health applications (#smm4h) shared tasks at COLING 2020. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop*, pages 27–36, 2020.

[8] Zulfat Miftahutdinov and Elena Tutubalina. KFU at SMM4H 2020: BERT-CNN with multi-stage fine-tuning for classification of health-related tweets. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop*, pages 58–62, 2020.

[9] Elena Tutubalina, Zulfat Miftahutdinov, Sergey Nikolenko, and Valentin Malykh. Russian drug reaction corpus and neural models for drug reaction and effectiveness detection in user reviews. *Bioinformatics*, 37(2):243–249, 2021.

[10] Hirotomo Moriwaki, Yu-Shi Tian, Norihito Kawashita, and Tatsuya Takagi. Mordred: A molecular descriptor calculator. *Journal of Cheminformatics*, 10(1):4, 2018.

[11] RDKit. RDKit: Open-source cheminformatics software. `https://www.rdkit.org`, 2024. Accessed: 2026-10-07.

[12] David Weininger. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1): 31–36, 1988.

[13] Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. SMILES-BERT: Large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 429–436, 2019.

[14] Benedek Fabian, Thomas Edlich, Héloïse Gaspar, Marwin Segler, Joshua Meyers, Marco Fiscato, and Mohamed Ahmed. Molecular representation learning with language models and domain-relevant auxiliary tasks. *arXiv preprint arXiv:2011.13230*, 2020.

[15] Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

[16] Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. The SIDER database of drugs and side effects. *Nucleic Acids Research*, 44(D1):D1075–D1079, 2016.

[17] Masaki Asada, Makoto Miwa, and Yutaka Sasaki. Enhancing drug-drug interaction extraction from texts by molecular structure information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 680–685, 2018.

[18] Krishna Sai Penumarthi. Computational prediction of directional drug-drug interactions using similarity fusion and deep learning. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2025. doi: 10.1109/BIBM.2025.11232139. Drug-drug interactions, similarity network fusion, CNN.

[19] Sujoy K. Saha, Sudipta Saha, and S. R. S. I. Hassan. Amalgamation of protein-protein interaction and sequence information for protein complex identification. *Journal of Biomedical Informatics*, 102:103369, 2020.

[20] G. R. Vemula. A hybrid approach to road classification using LiDAR data integrating aerial imagery and known road centrelines for enhanced feature extraction. In *PAIDeF AI4D Conference (SuperAI 2025)*, pages 18–36, Abuja, Nigeria, 2025. doi: 10.5281/zenodo.17951785.

[21] Kr Gh. Tracking temporal variations in predictive models for healthcare claims data. In *IEEE International Conference on AI in Healthcare (AIxHEART)*, 2025. doi: 10.1109/AIxHEART65685.2025.00015. Temporal data shifts, healthcare claims, predictive model degradation.

[22] Parul Patel. Biopharmaceutical breakthroughs in modern therapeutics. *World Journal of Advanced Research and Reviews*, 30(3), 2026. doi: 10.30574/wjarr.2026.30.3.1630. Biopharmaceuticals, monoclonal antibodies, mRNA vaccines, biosimilars.

[23] David S. Wishart, Craig Knox, An Chi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam, and Murtaza Hassanali. DrugBank: A knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research*, 36(suppl_1):D901–D906, 2008.

[24] Lei Chen, Wei-Ming Zeng, Yu-Dong Cai, Kai-Yan Feng, and Kuo-Chen Chou. Predicting anatomical therapeutic chemical (ATC) classification of drugs by integrating chemical-chemical interactions and similarities. *Journal of Chemical Information and Modeling*, 52(11):3094–3102, 2012.

[25] Louis Martin, Benjamin Muller, Pedro J. Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. CamemBERT: A tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, 2020.

[26] Pedro J. Ortiz Suárez, Benoît Sagot, and Laurent Romary. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora*, 2019.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[28] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

[29] P. Patel. Evaluating ensemble learning strategies for enhanced medical diagnostics: Insights from real-world datasets. In *2025 6th International Conference on Problems of Cybernetics and Informatics (PCI)*, pages 1–4, Baku, Azerbaijan, 2025. doi: 10.1109/PCI66488.2025.11219757.

[30] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.

[32] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020.

[33] Adithya Ramesh, Chandra Sekhara Rao Annavarapu, and Anil Kumar G. BERT based adverse drug effect tweet classification. In *Proceedings of the Sixth Social Media Mining for Health Applications Workshop*, pages 85–88, 2021.

[34] Davy Weissenbacher, Arjun Magge, Karen O'Connor, Allison Scotch, and Graciela González-Hernández. Overview of the fifth social media mining for health applications (#smm4h) shared tasks at NAACL 2021. In *Proceedings of the Sixth Social Media Mining for Health Applications Workshop*, pages 1–10, 2021.

[35] Zulfat Miftahutdinov and Elena Tutubalina. KFU at SMM4H 2021: BERT-based classification of adverse drug effects with data augmentation. In *Proceedings of the Sixth Social Media Mining for Health Applications Workshop*, pages 89–93, 2021.