Universiteit Leiden
The Netherlands

Bachelor Data Science & Artificial Intelligence

Extracting ADRs from Oncology Texts: Comparing Language Models and Traditional Machine Learning Approaches

Patricija Dziuzaite

Supervisors:
Prof. Dr. Marco Spruit & Dr. Armel Lefebvre

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

01/06/2025

Abstract

Adverse drug reactions (ADRs) are a critical concern in cancer treatment due to the complexity and toxicity of therapeutic regimens. This thesis investigates the application of BioBERT—a domain-specific transformer-based language model—for extracting ADR-related entities from biomedical literature. A filtered subset of the ADE Corpus V2, focused on cancer drugs, was used to fine-tune the model for named entity recognition. BioBERT's performance was compared against traditional machine learning models, including Support Vector Machines, Random Forest, and Logistic Regression, all trained on manually engineered features. Results indicate that BioBERT offers improved performance in recognising context-dependent and linguistically variable ADR mentions. However, the model shows reduced reliability in handling rare ADRs, entity boundary confusion, and negated or uncertain statements. These findings underscore both the potential and current limitations of transformer-based models for pharmacovigilance tasks, suggesting the value of future hybrid approaches.

Contents

1 Introduction

1.1 The situation

Adverse drug reactions (ADRs) are a major concern in cancer care, where patients are often treated with high-risk medications that can lead to serious or even life-threatening side effects (Logan et al., 2025). Identifying these reactions accurately is important for improving patient safety, understanding drug efficacy, and advancing personalized treatment strategies (Meystre et al., 2008).

However, most information about ADRs is embedded within unstructured medical texts, including clinical notes and published research articles, which poses significant challenges for automated data extraction (Meystre et al., 2008). Manual annotation and extraction of ADR information from these texts by experts is time-consuming and error-prone process, limiting its scalability and reliability in clinical practice (Jensen et al., 2012). Traditional natural language processing (NLP) methods, including rule-based systems and classical machine learning models, have been applied to this task with varying success, reflecting the challenges posed by the complex and variable language found in biomedical and clinical texts (Cronin et al., 2017).

Recently, transformer-based language models like BioBERT, which are trained on large biomedical text corpora, have shown great promise for biomedical text mining (Lee et al., 2020). Similar performance gains have also been demonstrated by other transformer-based models such as SciBERT, which is trained on scientific texts (Beltagy et al., 2019). This thesis explores how a fine-tuned BioBERT model can be used to extract and classify ADR mentions from cancer-related biomedical literature, using a curated subset of the ADE corpus.

The main research question for this study is: ***How well can language models extract and categorize ADRs from cancer treatment literature, and how do they compare to traditional machine learning methods?***

To evaluate the performance of BioBERT, I compare its results against three classical machine learning baselines: Support Vector Machine (SVM), Random Forest, and Logistic Regression. These baseline models are trained on hand-engineered linguistic features, and evaluated using the same dataset and metrics as the BioBERT model. The goal of this comparison is to better understand the advantages and limitations of transformer-based models in the context of ADR detection.

1.2 Thesis overview

The structure of this thesis is as follows: Section 2 reviews existing work on ADR detection, covering traditional and deep learning approaches as well as relevant tools. Section 3 details the dataset, model training, and evaluation metrics. Section 4 presents the experimental results and analysis. Finally, Section 5 summarizes the findings and suggests directions for future research.

This thesis is part of the bachelor program Data Science and Artificial Intelligence at the Leiden Institute of Advanced Computer Science (LIACS) at Leiden University and was written under the supervision of Prof.dr. M.R. Spruit and Dr. A.E.J.L. Lefebvre.

2 Related Work

This section reviews existing work on adverse drug reaction (ADR) detection, with a particular focus on oncology. Subsection 2.1 discusses the importance of ADR identification in cancer care, highlighting the clinical impact of adverse reactions and the challenges in monitoring them. Subsection 2.2 reviews traditional approaches to ADR extraction, including rule-based systems and classical machine learning techniques. Subsection 2.3 explores the role of deep learning in biomedical text mining, focusing on transformer-based language models such as BioBERT and more recent large language models (LLMs) like GPT-4. Subsection 2.4 outlines commonly used datasets and benchmarking standards for ADR research. Subsections 2.5 and 2.6 examine the integration of tools like SNPcurator and the complementary potential of combining such systems with transformer-based models. Finally, Subsection 2.7 summarizes the ongoing challenges and identifies research gaps in current ADR extraction methods.

2.1 Relevance of ADR Detection in Oncology

Cancer therapies, including chemotherapy and targeted treatments, are known to carry a high risk of adverse drug reactions (ADRs), which can negatively impact treatment outcomes and patient safety. Lavan et al., 2019 found that adverse drug reactions (ADRs) contributed to 21.5% of hospital admissions among cancer patients, with 89.3% of ADRs considered predictable and 62.6% potentially preventable. These results highlight the importance of targeted interventions and improved medication management in oncology care.

Further highlighting this issue, Lavan et al. noted that the responsibility for managing ADRs is often unclear in clinical settings. While oncologists prescribe cancer drugs, general practitioners and other specialists frequently prescribe additional medications. This fragmented care structure-particularly in older patients with multimorbidity-can lead to polypharmacy risks going unaddressed, reinforcing the importance of comprehensive ADR monitoring and geriatric assessment in oncology.

In the context of HER2-positive breast cancer, Barbieri et al., 2022 emphasised the importance of collaboration between oncologists and pharmacologists for early ADR identification. Their findings suggest that educating patients on potential symptoms significantly contributes to timely ADR reporting and improved treatment outcomes.

Lastly, Monestime et al., 2021 explored ADR reporting behavior during oral targeted cancer treatments and identified gaps in patient-provider communication. The study highlighted the need for improved systems that encourage patients to report side effects, which could enhance the pharmacovigilance process and ultimately reduce harm.

Together, these studies demonstrate that ADRs in oncology are both common and clinically significant. Improving their detection through advanced NLP tools, particularly large language models, may help close existing gaps in pharmacovigilance and support safer, more personalized cancer care.

2.2 Traditional Approaches for ADR Extraction

Early methodologies for ADR extraction were primarily rule-based or relied on classical machine learning models (Meystre et al., 2008). Rule-based systems used manually crafted lexicons and pattern-matching techniques to identify ADR mentions in biomedical texts. These approaches typically followed structured pipelines, where predefined rules were applied to extract relevant terms (Figure 1 provides an example of such a pipeline). While these methods often demonstrated high precision, they tended to suffer from poor recall and limited generalizability due to their reliance on rigid, manually defined rules (Shen and Spruit, 2021).

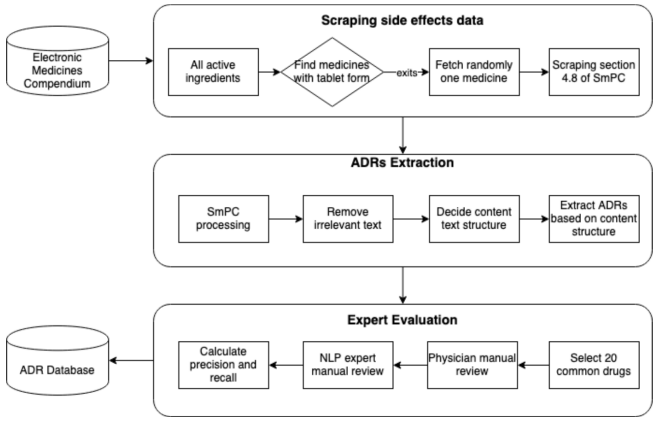

Figure 1: A conceptual overview of the automatic Adverse Drug Reactions (ADRs) extraction pipeline (Shen and Spruit, 2021)

Classical machine learning models, such as Support Vector Machines (SVM), Random Forest, and Logistic Regression, introduced greater flexibility. These models leveraged manually engineered features-often lexical, syntactic, or semantic-to detect ADR mentions. While more adaptable than purely rule-based systems, they still faced significant challenges in generalizing to complex biomedical texts due to their reliance on domain-specific features and limited ability to model context (Lavan et al., 2019).

A notable recent example of a structured rule-based approach is the work of Shen and Spruit, 2021, who developed a method for extracting ADRs specifically from Summary of Product Characteristics (SmPC) documents issued by the European Medicines Agency. This addressed a gap in previous work, which had largely focused on FDA Structured Product Labels (SPLs). Their system processed SmPCs for 647 marketed medicines and extracted 32,797 ADR terms, including 8,069 unique ones. The number of ADRs per medicine varied widely, with an average of 51 terms and a median of 38. The SmPCs themselves were heterogeneous in structure, including 141 documents with structured text, 419 with tabular layouts, and 87 written in free text format.

Figure 2 summarizes the overall extraction results reported by Shen and Spruit, 2021, illustrating the variability in ADR term counts across different medicines.

Characteristics	Statistics
# of selected marketed medicines	647
# of medicines with structured text	141
# of medicines with tabular text	419
# of medicines with free text	87
Average ADRs per medicine	51
25% percentile ADRs per medicine	21
50% percentile ADRs per medicine	38
75% percentile ADRs per medicine	67
Top medicines in terms of extracted ADRs	1. Topamax 100 mg Tablets (269) 2. Revolade 25 mg film-coated tablets (255) 3. Capecitabine Accord 150 mg film coated tablets (240) 4. Glivec 100 mg film-coated tablets (232) 5. Risperdal 0.5 mg Film-Coated Tablets (218) 6. Xadago 50 mg film-coated tablets (215) 7. Invega 12 mg prolonged-release tablets (207) 8. LUSTRAL 100 mg film coated tablets (207) 9. Isentress 100 mg chewable tablets (191)

Figure 2: Overview of ADR extraction results (Shen and Spruit, 2021)

Their approach achieved strong results, with a recall of 0.990 and precision of 0.932 as measured by manual expert review. These results are among the highest reported in ADR extraction literature. However, the authors acknowledged important limitations: their method struggled with inconsistent formatting, splitting multiple ADRs listed in a single phrase, and normalizing extracted terms to controlled vocabularies like MedDRA. Additionally, the evaluation process involved only one clinical expert (due to limited availability during COVID-19) and one trained NLP researcher, limiting the generalizability of the manual validation.

The final manual expert review results are presented in Figure 3, highlighting both strengths and limitations of the extraction system.

	Reviewer 1 (NLP Expert)	Reviewer 2 (Clinical Expert)	Totals
# of reviewed medicines	32	5	37
# of extracted ADRs terms	1700	118	1824
# of correct ADRs (TP)	1590	110	1703
# of incorrect ADRs (FP)	116	8	124
# of missing ADRs (FN)	7	11	18
Recall	0.996	0.909	0.99
Precision	0.932	0.932	0.932

Figure 3: Final results of the manual expert reviews (Shen and Spruit, 2021)

These findings emphasize the challenges of scaling rule-based systems in the biomedical domain, especially when working with diverse text formats and limited annotated resources. As a result, there is growing interest in more advanced NLP techniques-such as deep learning and transformer-based models-that can reduce reliance on handcrafted rules, better model contextual information, and adapt to the variability of biomedical language.

Another example of a traditional dictionary-based named entity recognition system is ProMiner, originally developed for gene and protein name identification in biomedical text (Hanisch et al., 2005). ProMiner uses a pre-processed synonym dictionary combined with rule-based matching to detect multi-word biomedical entities and associate them with database identifiers, while handling synonym ambiguity and organism specificity. Although ProMiner was designed for gene/protein

recognition, similar dictionary-based approaches have been adapted in ADR extraction tasks to identify potential adverse drug event mentions by matching terms against pre-defined ADR lexicons. Such dictionary-driven methods depend heavily on the coverage and quality of the underlying vocabulary and often struggle with novel or ambiguous ADR terms and variations in clinical language.

These challenges further emphasize the limitations of traditional methods in handling the diversity and complexity of biomedical text, highlighting the need for more dynamic and adaptive techniques. Moving forward, hybrid approaches that combine rule-based methods with machine learning and deep learning techniques are becoming more common, as they aim to overcome the rigidities of dictionary-based approaches while still maintaining interpretability and domain specificity.

Nikfarjam et al., 2015 developed *ADRMine*, a sequence labeling system leveraging conditional random fields (CRF) combined with word embedding cluster features to identify adverse drug reaction (ADR) mentions from social media data. Their approach was evaluated on two corpora: DailyStrength, a health-related forum, and Twitter, a general social media platform. *ADRMine* significantly outperformed traditional lexicon-based and MetaMap methods, achieving an F-measure of 0.82 on DailyStrength and 0.72 on Twitter. The authors emphasized that embedding cluster features, learned from a large volume of unlabeled posts, played a crucial role in improving recall, especially when annotated training data was limited. They also noted challenges unique to social media, such as informal language, short context, and ambiguous mentions, which contributed to false positives and false negatives. The study highlighted the potential of combining unsupervised embedding techniques with sequence models to enhance pharmacovigilance efforts from user-generated content and suggested future work involving deep learning and concept normalization to standard medical ontologies.

In addition to the limitations of traditional extraction techniques, recent systematic reviews have highlighted the inherent complexity of adverse drug event (ADE) extraction from clinical notes, which involves both named entity recognition (NER) and relation extraction (RE) tasks. Modi et al., 2024 reviewed multiple studies and found that while NER has been the primary focus of most systems, RE remains particularly challenging due to the scarcity, ambiguity, and polysemous nature of ADEs and their causal relationships in clinical narratives. They emphasized issues such as overlapping annotations and inconsistent labeling across datasets, especially within widely used corpora like the n2c2 challenge dataset, which can significantly hinder model training and evaluation. These challenges contribute to the overall difficulty in achieving high performance in ADE extraction and underscore the need for improved datasets and annotation standards to enhance the reliability of traditional ADR extraction approaches.

Furthermore, Naderian et al., 2024 developed another approach, an eight-phase natural language processing (NLP) framework, ADRD, designed to detect adverse drug reactions (ADRs) from patient comments on psychiatric medications within the PsyTAR dataset. Analyzing 891 comments related to four drugs (Zoloft, Lexapro, Cymbalta, and Effexor XR) and 285 unique conditions, the framework provided detailed dataset summarization, statistical analyses of ADR counts and patient satisfaction ratings, and identified drug-specific ADR distributions. The study found variability in ADR reporting and patient satisfaction across drugs, with Lexapro showing the highest ratios of

satisfied patients without ADRs. Despite the automation, clinical validation remained necessary to confirm findings. The framework demonstrated the potential of NLP to automate and enhance ADR analysis from unstructured clinical narratives, facilitating pharmacovigilance and patient medication evaluation, while underscoring the ongoing need for human oversight in interpreting results.

Despite improvements in traditional and hybrid ADR extraction methods, their reliance on manual rules, handcrafted features, and limited contextual understanding restricts their scalability and adaptability. Consequently, there is increasing interest in leveraging deep learning, especially transformer-based models and large language models, which offer greater capacity to model complex biomedical language, capture context, and generalize across diverse datasets. The following section reviews these emerging approaches and their impact on ADR extraction.

2.3 Deep Learning, Transformers, and Large Language Models in ADR Extraction

The introduction of deep learning significantly improved ADR extraction by reducing reliance on manual feature engineering (Liu et al., 2020). Early architectures, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), improved the accuracy of named entity recognition (NER) tasks in biomedical text. Liu et al. proposed a neural multi-task learning framework that jointly handled NER and relation extraction, yielding better performance for drug-effect associations.

A major breakthrough came with transformer-based models like BERT (Devlin et al., 2018), which introduced self-attention mechanisms to better model context. In the biomedical domain, BioBERT (Lee et al., 2020) extended BERT's architecture by pretraining on PubMed and PMC articles. Fine-tuning BioBERT on annotated corpora has led to significant improvements in biomedical NER tasks, including ADR extraction.

Siegersma et al., 2022 introduced ADRIN, a novel method for extracting adverse drug reactions (ADRs) from Dutch free-text clinical notes using a combination of word embedding models and string matching with the Medical Dictionary for Regulatory Activities (MedDRA). Leveraging a large corpus of over 277,000 clinical notes and nearly 500,000 medication registrations, they developed vector representations to identify medication and ADR mentions in a manually labeled test set of 988 notes. Their evaluation showed that binary classification for ADR presence reached an accuracy of 0.84. Key findings indicated that reducing the search area around medication mentions and incorporating punctuation improved the pipeline's performance, while the inclusion of MedDRA terminology did not enhance results and led to an increase in false positives. These results underscore the challenges of integrating broad medical ontologies with embedding-based approaches and highlight the importance of contextual and linguistic features in ADR extraction.

Recent studies have also compared BioBERT with other domain-specific transformer models such as ClinicalBERT, SciBERT, BlueBERT, and RoBERTa. BioBERT, pretrained on large biomedical corpora like PubMed and PMC, achieves state-of-the-art results in medical named entity recognition, with reported precision and F1 scores of 89.8% and 87.6%, respectively. This is attributed to its

strong domain adaptability and ability to handle complex biomedical terminology. ClinicalBERT, in contrast, is fine-tuned on clinical notes and EHRs, making it particularly effective in processing clinical narratives. However, due to its narrower training domain, it shows slightly lower performance on broader biomedical texts (precision 85.2%, F1 83.5%). These results suggest that for ADR extraction tasks based on biomedical literature - such as the ADE Corpus used in this work - BioBERT is more suitable than ClinicalBERT, due to its pretraining on large -scale biomedical texts like PubMed and PMC (Hu et al., 2024).

Model	Precision%	F1-score
Bert [23]	82.5	81.0
ClinicalBERT [24]	85.2	83.5
SciBert [25]	84.1	82.8
BlueBert [26]	87.3	85.0
BioBert [27]	89.8	87.6

Figure 4: Comparison of different BERT models (Hu et al., 2024)

More recently, general-purpose large language models (LLMs) such as GPT-3 (Brown et al., 2020), GPT-4, Galactica (Taylor et al., 2022), and PaLM have demonstrated strong performance in zero-shot and few-shot biomedical tasks. These models are not pretrained solely on biomedical literature but have acquired general world knowledge, including medical reasoning capabilities. They have been used in question answering, clinical text summarization, and hypothesis generation.

Although BioBERT remains more effective for fine-tuned tasks like token-level tagging due to its domain specialization, LLMs offer advantages in flexible generation, inference, and long-context understanding. For instance, recent work shows GPT-4 achieving high accuracy in clinical reasoning benchmarks such as MedQA and PubMedQA (Singhal et al., 2022), though LLMs still face challenges in hallucination and factual grounding.

A comparison of transformer-based models is provided in Table 1 to illustrate their varying strengths and applications in biomedical NLP.

Model	Pretraining Domain	Use Case in ADR	Advantages / Limitations
BERT	General text (Wikipedia + BooksCorpus)	Baseline NLP tasks	Not domain-specialized; good generalisation
BioBERT	Biomedical abstracts and full-text	NER, relation extraction, ADR tagging	Domain-adapted; excels in biomedical token classification
PubMedBERT	Biomedical abstracts only	Biomedical NER	Strong on unseen biomedical terminology
ClinicalBERT	Clinical notes (MIMIC-III)	EHR-based ADR tagging	Effective for clinical terminology; weaker generalisation to biomedical literature
GPT-3/4	Web-scale general corpus	Question answering, summarization, entity linking	Zero-/few-shot capabilities; lacks precision in token-level tasks
Galactica	Scientific literature	Biomedical reasoning, generation	Scientific grounding; vulnerable to hallucinations

Table 1: Comparison of transformer and LLM models in biomedical NLP

As LLMs become more powerful, hybrid architectures may emerge that combine domain-specific models like BioBERT with generative capabilities of models like GPT-4. This could enable multimodal ADR extraction, integrating structured tagging with rich semantic reasoning-important for future pharmacovigilance tools.

Recent work by Romero et al., 2025 introduces *INSIGHTBUDDY-AI*, a clinical NLP system that ensembles multiple transformer-based models for medication-related entity recognition and entity linking. The authors fine-tuned eight transformer variants, including BioBERT, PubMedBERT, and RoBERTa, on the n2c2-2018 clinical dataset, and demonstrated that ensemble methods (especially non-BIO voting strategies) significantly outperform individual models. Their max-logit voting ensemble achieved a macro F1-score of 0.8821, statistically outperforming the best individual models such as RoBERTa-Large and BioMedRoBERTa. Additionally, they implemented entity linking to SNOMED-CT and BNF ontologies using a combination of keyword search and fuzzy matching. Importantly, the study also explored model quantisation, showing that 4-bit versions of these models retained near-identical performance while reducing size by 75%. These findings highlight the potential of ensemble and compressed transformer models for efficient and accurate clinical entity extraction in real-world settings, including medication-related ADR tasks.

2.4 Datasets and Benchmarking for ADR Extraction

Effective ADR extraction relies on high-quality annotated datasets. Several publicly available corpora, such as the ADE Corpus and CADEC, have been widely used for training and evaluating ADR extraction models (Gurulingappa et al., 2012). These datasets provide annotated drug-AE (adverse event) relationships extracted from biomedical literature, electronic health records, and patient forums.

Benchmarking studies compare the performance of different models using standard NLP evaluation metrics, such as precision, recall, and F1-score. Shen and Spruit, 2021 emphasized the importance of dataset diversity in ensuring robust model performance across different biomedical subdomains.

2.5 Enhancing ADR Extraction Using Structured Genetic Platforms

In addition to NLP-based methods, structured platforms such as SNPcurator provide valuable data on genetic variants linked to drug responses. Shen and Spruit, 2021 suggested using SNPcurator to connect ADR extractions with genetic mutations, which can enhance the interpretability of ADR findings. Combining structured genetic information with domain-specific language models like BioBERT may improve ADR categorisation by incorporating genetic context, ultimately supporting more personalised treatment recommendations.

2.6 Complementarity of BioBERT and SNPcurator

BioBERT and SNPcurator represent two distinct but complementary approaches to biomedical information extraction, each with different design goals, data dependencies, and technological foundations. While a direct performance comparison between them is not feasible due to their differing input assumptions and outputs, they offer unique advantages that, when combined, could enhance biomedical text mining—particularly in contexts like pharmacogenomics and adverse drug reaction (ADR) research.

BioBERT

BioBERT is a transformer-based language model pretrained on large biomedical corpora such as PubMed abstracts and PMC articles. Its deep contextual understanding enables it to perform token-level tasks such as named entity recognition (NER) and relation extraction with high precision, especially when fine-tuned on domain-specific datasets. In this study, BioBERT was fine-tuned to extract ADR-related entities, such as drug mentions and adverse effects, from a filtered subset of the ADE corpus focusing on cancer treatments.

One strength of BioBERT is its flexibility-it can be adapted to various biomedical NLP tasks provided annotated data is available. However, it requires technical knowledge to implement, fine-tune, and evaluate effectively.

SNPcurator

SNPcurator, by contrast, is a rule-based tool specifically designed to extract SNP-disease associations from genome-wide association studies (GWAS). It applies regular expressions and syntactic-semantic parsing to retrieve SNP identifiers (e.g., rsIDs), associated phenotypes, cohort statistics (e.g., ethnicity, size), and p-values. It is optimized for structured outputs and does not require machine learning expertise, making it accessible to users outside the data science field Tawfik and Spruit, 2018.

SNPcurator begins by querying the entire PubMed repository using the NCBI E-utilities API (Esearch and Efetch) together with the BioPython library. The query combines a user-specified disease term with specific genetic-related Medical Subject Headings (MeSH), such as "Polymorphism, Single Nucleotide," "Genetic Predisposition to Disease," and "Genome-Wide Association Study." Only English abstracts with complete text are retained, and a filtering step excludes abstracts without SNP mentions to focus specifically on SNP-association studies.

For information extraction, SNPcurator leverages the spaCy natural language processing toolkit, utilizing its sentence splitting, tokenization, named entity recognition, and dependency parsing modules. These processes enable the extraction of structured SNP-disease pairs and related metadata from unstructured biomedical abstracts.

Its specialization makes SNPcurator reliable for tasks involving structured genetic data but limits its flexibility in processing broader biomedical texts where associations are not strictly formatted.

Input Data Comparison and Example Both BioBERT and SNPcurator operate on biomedical abstracts sourced from PubMed, but differ significantly in their input data collection, preprocessing, and annotation strategies.

SNPcurator begins with a user-supplied disease term combined with specific genetic-related MeSH terms (e.g., "Polymorphism, Single Nucleotide," "Genetic Predisposition to Disease") to query PubMed via NCBI Eutils. This query returns a set of abstracts focused on genetic associations relevant to the disease. The raw abstract text, along with metadata such as PubMed ID and disease term, forms the input data. No manual annotation is required; instead, SNPcurator applies rule-based NLP (regex for SNP ID detection and spaCy for tokenization and named entity recognition) to preprocess the abstracts.

Listing 1: Excerpt from SNPcurator input data (preprocessed PubMed abstract)

```
{
"pmid": "34567890",
"disease": "lung cancer",
"abstract": "A significant association was observed between rs123456 and lung
   cancer in the Asian population."
}
```

In contrast, BioBERT requires manually annotated and tokenized datasets for supervised learning. Abstracts from the ADE corpus are annotated at the token level with BIO tags indicating entity

boundaries and types (e.g., `DRUG`, `EFFECT`). The input to BioBERT includes the raw text along with token IDs, attention masks, offset mappings, and token-level labels.

Listing 2: Excerpt from preprocessed ADE dataset for BioBERT


```
{
"text": "Erlotinib treatment caused severe rash in several patients.",
"drug": "Erlotinib",
"effect": "severe rash",
"indexes": [0, 3, 4, 5],
"input_ids": [101, 4321, 3793, ...],
"token_type_ids": [0, 0, 0, ...],
"attention_mask": [1, 1, 1, ...],
"offset_mapping": [[0, 9], [10, 19], ...],
"labels": [1, 0, 2, 0, ...]
}
```


This fine-grained annotation allows BioBERT to learn token-level entity recognition through supervised training, while SNPcurator's rule-based approach processes raw abstracts directly without labeled training data.

Figure 5 presents a conceptual architecture illustrating how BioBERT and SNPcurator can be integrated to leverage their complementary strengths for pharmacogenomic applications.

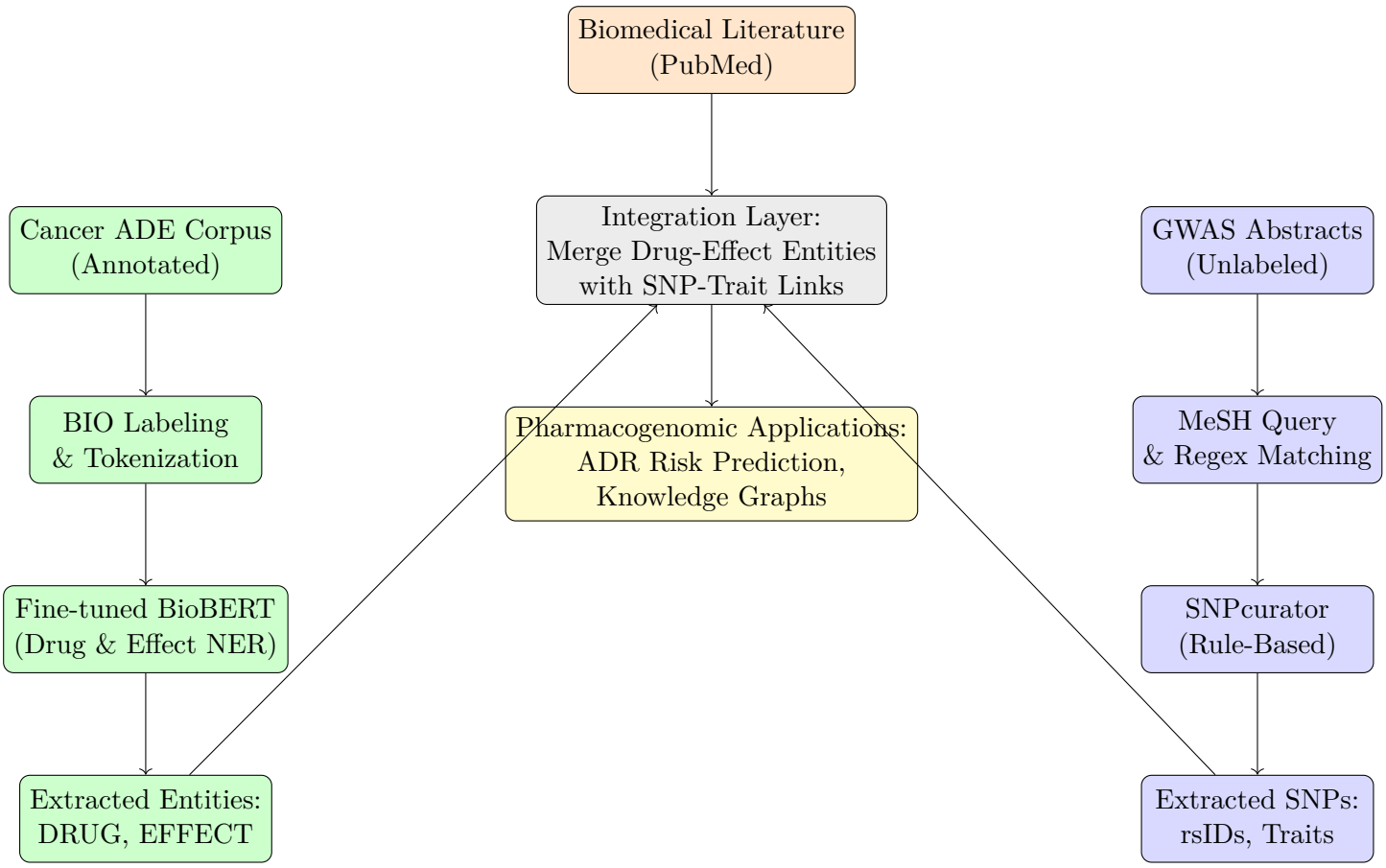

Figure 5: Final vision architecture illustrating how BioBERT and SNPcurator can be combined for pharmacogenomic research. BioBERT extracts drug-effect relationships from annotated cancer literature using deep learning, allowing for nuanced and context-aware entity recognition. SNPcurator extracts SNP-trait associations from GWAS abstracts using rule-based methods. Their outputs are merged to link adverse drug reactions with genetic variants, enabling richer downstream applications like ADR risk prediction and the construction of pharmacogenomic knowledge graphs.

To explore possible intersections between ADR extraction and SNP-focused tools, I attempted to run SNPcurator on the cancer-related ADE abstracts used to train BioBERT. SNPcurator identifies single nucleotide polymorphisms (SNPs) in text using regular expressions that match NCBI dbSNP identifier formats such as rs123456 or ss123456, including variants that account for non-standard notations with trailing alleles (e.g., rs123456A). This regex-based method has achieved 100% recall on benchmark SNP detection tasks (Tawfik and Spruit, 2018). However, when applied to the ADE corpus, only one abstract contained SNP-related keywords such as "snp," "mutation," "variant," or "polymorphism," and none contained valid SNP identifiers like rsIDs. This indicates that the ADE corpus lacks the genetic information needed for evaluating tools like SNPcurator. Therefore, it is not suitable for SNP-focused extraction tasks. My testing script and preprocessing code for SNPcurator are available on GitHub[1].

[1] https://github.com/dziuzaite/ade-biobert-finetuning.git

Still, from a conceptual perspective, integrating BioBERT and SNPcurator remains promising. For example, if an abstract includes both ADR information and mentions of genetic variants (as in pharmacogenomics studies), BioBERT could extract drug-effect relationships, while SNPcurator captures the genetic context. Together, they could help answer more complex questions such as: **Which genetic variants may influence the likelihood or severity of specific ADRs in cancer patients?**

Comparison Overview

Aspect	**BioBERT**	**SNPcurator**
Goal	General biomedical NLP model	SNP-disease association extraction
Technology	Transformer-based deep learning	Rule-based with spaCy and regex
Flexibility	Adaptable to many NLP tasks	Optimized for a fixed schema
Ease of Use	Requires ML/NLP expertise	Web interface for non-programmers
Data Scope	Broad biomedical literature (e.g., ADRs, drugs)	GWAS and genetic association abstracts (from PubMed)
Output	Entity spans and token-level tags (BIO format)	Structured SNP records with rsIDs, p-values, traits
Customizability	Requires retraining for new tasks	Modifiable via rule sets

Table 2: Comparison between BioBERT and SNPcurator

In summary, although BioBERT and SNPcurator target different types of information, their outputs could be combined to enrich biomedical knowledge graphs or decision-support tools in personalized medicine. The current limitation is the lack of datasets that bridge both drug-related ADRs and SNP mentions in a single corpus-highlighting an area for future data collection and system integration.

2.7 Challenges and Research Gaps

Despite advancements in ADR extraction, several challenges remain. Firstly, transformer-based models require extensive computational resources and large labeled datasets for effective fine-tuning. Secondly, the contextual complexity of ADR mentions in biomedical literature makes it difficult for models to achieve high recall without introducing false positives.

Moreover, although BioBERT has demonstrated superior performance compared to traditional NLP methods in ADR extraction, its efficiency and effectiveness relative to classical machine learning models still require further investigation. Additionally, the potential benefits of integrating large language models with structured genetic platforms for ADR identification remain largely unexplored.

This thesis aims to address these gaps by evaluating how effectively domain-specific language models like BioBERT extract and categorize ADRs from cancer treatment literature. By benchmarking these models against traditional NLP methods and exploring SNPcurator integration, this research contributes to the ongoing effort to enhance ADR identification for improved clinical decision-making.

3 Methods

3.1 Dataset

For this study, I used the ADE Corpus V2, specifically the `Ade_corpus_v2_drug_ade_relation` subset provided by the ADE Benchmark Corpus.[2] This dataset contains sentences from biomedical literature annotated with entities corresponding to drugs and their associated adverse drug effects (ADEs), making it highly suitable for supervised named entity recognition (NER) tasks in the biomedical domain.

I selected ADE Corpus V2 instead of larger clinical datasets such as MIMIC-III for several key reasons. First, the ADE Corpus is specifically designed for the task of ADE detection and comes pre-annotated with adverse drug event relationships, thus removing the need for additional clinical concept extraction or annotation. Second, although MIMIC-III offers rich clinical narratives, accessing it involves a credentialed training program and institutional agreements, which involve not just administrative delays but also financial costs that were beyond the scope of this project (Johnson et al., 2016). Additionally, I reached out to several authors of relevant papers to request access to their datasets, but did not receive responses. Taken together, these practical considerations made the ADE Corpus the most feasible and appropriate choice for this research.

Furthermore, while several studies on cancer drug adverse effects have used domain-specific corpora (e.g., curated from oncology EHRs or literature databases), those datasets are often either proprietary or lack the consistent annotation structure found in the ADE Corpus. In contrast, the ADE Corpus V2 provides a structured and publicly available benchmark with token-level annotation for both drug mentions and their associated side effects.

To further tailor the dataset to this project's focus on cancer pharmacovigilance, I filtered the corpus to include only sentences that mention well-known cancer drugs. I compiled a list of 38 cancer therapeutics, including *methotrexate*, *cyclophosphamide*, *doxorubicin*, *cisplatin*, *tamoxifen*, among others, based on the National Cancer Institute's approved targeted therapies.[3] I then programmatically retained only those sentences that reference at least one of these drugs. This filtering step helped restrict the model's domain while preserving the relevance of the examples for downstream comparison with tools like SNPcurator, which also focuses on oncology-related pharmacogenomics.

[2] https://huggingface.co/datasets/ade-benchmark-corpus/ade_corpus_v2

[3] https://www.cancer.gov/about-cancer/treatment/types/targeted-therapies/approved-drug-list#targeted-therapy-approved-for-thyroid-cancer

Each selected sentence was further processed using the BIO tagging format, labeling tokens as belonging to one of five classes: `O`, `B-DRUG`, `I-DRUG`, `B-EFFECT`, or `I-EFFECT`. These labels were aligned to tokens using offset mappings from the BioBERT tokenizer to ensure compatibility with the model's input format.

The final filtered and preprocessed dataset consisted of several hundred cancer-related biomedical sentences. I randomly split this dataset into 80% for training and 20% for validation. Despite the small dataset size, the task's narrow domain focus and strong semantic structure made it suitable for fine-tuning a domain-specific language model like BioBERT.

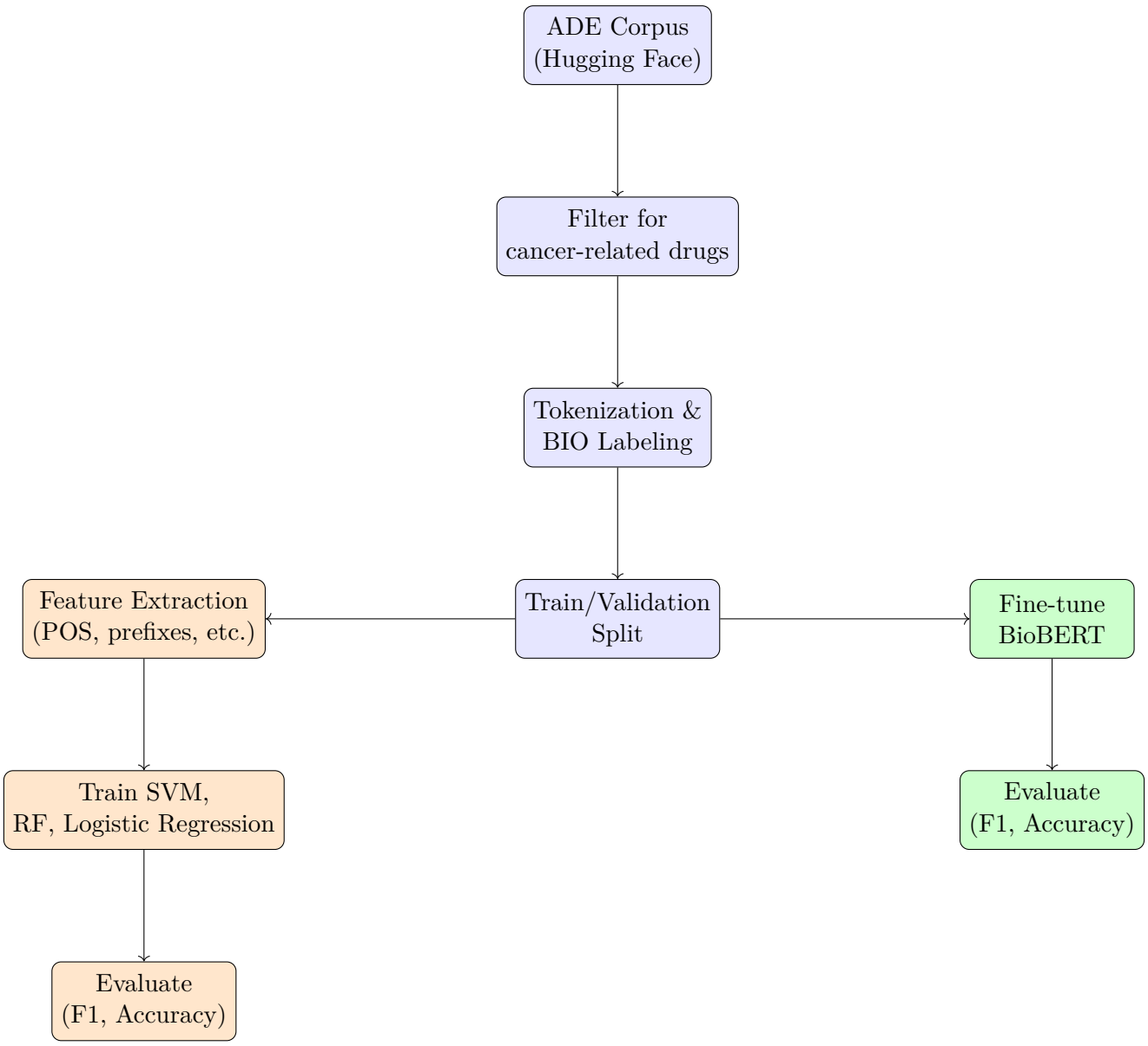

Figure 6: Overview of the ADR extraction pipeline used in this thesis. The dataset is filtered and split, then used for both BioBERT fine-tuning and training classical models (SVM, Random Forest, Logistic Regression). Each model is evaluated using precision, recall, and F1-score.

3.2 Model and Training Setup

For this task, I fine-tuned BioBERT (dmis-lab/biobert-base-cased-v1.1), a transformer-based model pretrained on biomedical corpora.[4]

Tokenization and alignment of labels were performed using the BioBERT tokenizer. I implemented a custom PyTorch `Dataset` class to extract encoded input IDs, attention masks, and token labels.

[4] https://github.com/dmis-lab/biobert-pytorch

Padding was handled dynamically within batches using the `DataCollatorForTokenClassification`. I trained the model for 10 epochs using the AdamW optimizer with a learning rate of 1×10^{-5} and a step-wise learning rate scheduler (`StepLR` with $\gamma = 0.95$). I used a batch size of 16 and trained the model on a GPU-enabled environment when available.

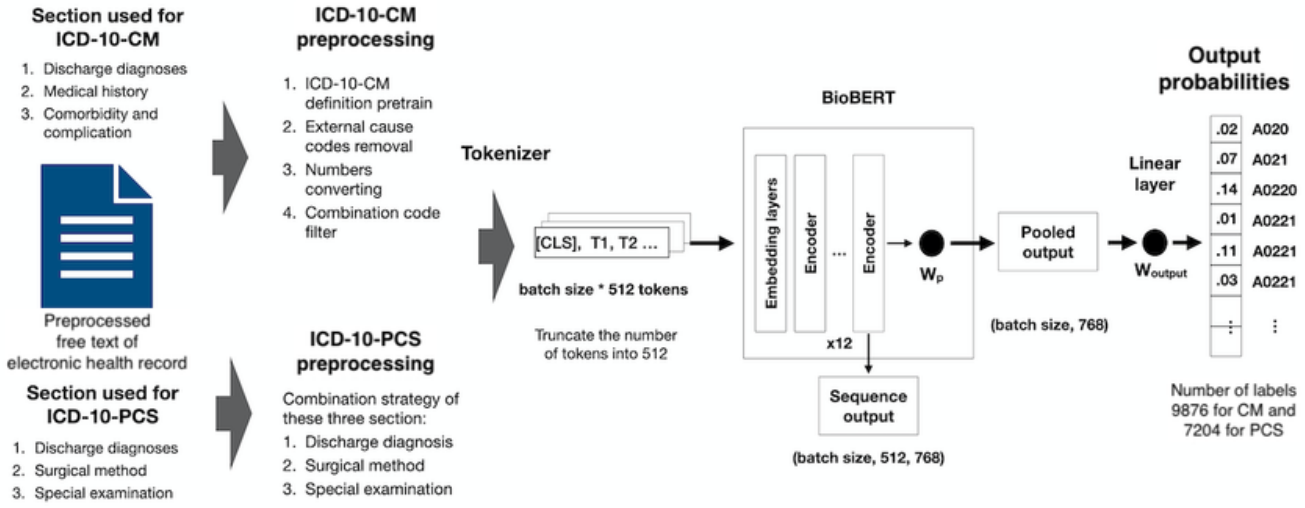

Figure 7: BioBERT model architecture for token classification used in this study.

3.3 Evaluation Metrics

I evaluated model performance using token-level precision, recall, and F1-score for each class, in addition to overall accuracy. Label predictions were mapped back to BIO-format tag names for interpretation. The validation set consisted of 5,552 labeled tokens.

4 Results

4.1 Main results

The fine-tuned BioBERT model achieved an overall accuracy of 88% and a weighted F1-score of 0.87 on the validation set. These results indicate robust performance in identifying and classifying biomedical entities related to adverse drug events (ADRs), even within the constraints of a relatively small cancer-specific dataset.

To fine-tune the model, the `dmis-lab/biobert-base-cased-v1.1` checkpoint was used as a starting point. Training was performed on a subset of the ADE corpus, filtered to include only drug-effect pairs relevant to oncology. Sentences were tokenized using the BioBERT tokenizer, and labels were aligned using a standard BIO tagging scheme. Training was conducted over 10 epochs using the AdamW optimizer with a learning rate of 1e-5 and gradient clipping. Special handling was applied to ignore padding tokens during training via label masking.

In addition to token-level classification, the final predictions were post-processed to reconstruct full drug and ADR entities. This was achieved by grouping contiguous tokens labeled with `B-` and `I-` tags and detokenizing subword units, improving the interpretability and usability of the model output.

All code, including data preprocessing, training, and evaluation scripts, is publicly available on GitHub.[5]

Table 3 presents the token-level classification results for each BIO-formatted label.
After training, the fine-tuned BioBERT model was saved and used to predict entities in previously unseen oncology-related sentences. A post-processing step grouped contiguous BIO-labeled tokens into full entity spans by detokenizing subword units, addressing token-level fragmentation common in subword-based models and improving the interpretability of the extracted entities.

Label	**Precision**	**Recall**	**F1-score**	**Support**
B-DRUG	0.87	0.75	0.81	106
B-EFFECT	0.58	0.54	0.56	300
I-DRUG	0.87	0.88	0.88	384
I-EFFECT	0.63	0.60	0.62	425
O	0.92	0.93	0.92	4337
Macro avg	0.77	0.74	0.76	5552
Weighted avg	0.87	0.88	0.87	5552

Table 3: Performance of fine-tuned BioBERT on the cancer-related ADE dataset

The model demonstrated particularly strong performance on the `O` (non-entity) and `I-DRUG` classes, achieving F1-scores of 0.92 and 0.88, respectively. These results suggest the model is highly reliable at filtering out irrelevant tokens and at recognizing multi-token drug names-a common pattern in biomedical literature. The contextualized embeddings provided by BioBERT likely contribute to this strength, enabling the model to detect consistent patterns in drug terminology.

The `B-DRUG` class achieved an F1-score of 0.81, slightly lower than `I-DRUG`, which reflects a common challenge in NER: accurately identifying the beginning of named entities. This can occur when multiple entities appear in succession, when drug names are embedded in longer phrases, or when the drug mention is abbreviated or unfamiliar.

The performance on `B-EFFECT` and `I-EFFECT` was lower (F1 = 0.56 and 0.62), though slightly improved compared to earlier runs. These labels mark tokens that belong to adverse effect mentions. The drop in performance here highlights the ongoing difficulty in capturing how side effects are described in natural text-often with vague, multi-word, or variable phrasing. For example, adverse effects can appear as "severe gastrointestinal discomfort" or "mild elevation of liver enzymes," making consistent detection a challenge without broader sentence-level modeling.

[5] https://github.com/dziuzaite/ade-biobert-finetuning

The macro-averaged F1-score of 0.76 reflects how the model performs across classes equally, including low-frequency ones. It confirms moderate success in learning minority classes like `B-EFFECT`, which are essential for clinical relevance. The weighted average F1-score of 0.87, on the other hand, reflects high performance on dominant classes such as `O` and drug-related entities.

While the evaluation metrics reflect strong overall performance, especially in recognizing drug entities and non-entity tokens, the model's ability to detect adverse effects remains more limited. These gaps highlight the need for further architectural improvements, such as CRF-based decoding, and for expanding the diversity of ADR expressions in the training data. A more detailed analysis of prediction behavior is presented in the next section.

4.2 Inference Examples

To qualitatively assess model predictions, I evaluated BioBERT on 20 manually selected biomedical sentences containing cancer drugs and potential adverse effects. These sentences were designed to capture a diverse range of linguistic challenges, including passive voice, negation, hyphenated compounds, multi-token ADRs, uncertainty expressions, and co-occurrence of multiple drug-effect mentions. Selected examples and observations are presented below:

- In *"Severe nausea and vomiting developed in the patient after prolonged cisplatin treatment."*, the model correctly identified `cisplatin` and `nausea`, but missed `vomiting` as an adverse effect, likely due to token-level boundary limitations.
- In *"Fatigue was not associated with pembrolizumab in this case."*, the model failed to label `fatigue` possibly due to the sentence's negation, which indicates the symptom was not caused by the drug.
- In *"Fluorouracil has been associated with hand-foot syndrome."*, the model correctly labeled the full multi-token adverse effect `hand-foot syndrome`, showing robustness to punctuation and uncommon terminology.
- In *"The patient's mobility was affected by neuropathy induced by vincristine."*, no entities were recognized, likely due to the compound structure and syntactic distance between the drug and the effect.
- In *"Nivolumab and ipilimumab combination therapy triggered autoimmune hepatitis."*, both drugs and the adverse effect were correctly identified, showing that the model can sometimes handle co-occurring entities well.
- In *"Following methotrexate administration, mucositis and gastrointestinal discomfort developed."*, only one of the two adverse effects was correctly labeled, highlighting difficulty with multi-ADR extraction in a single clause.
- In *"It is possible that gemcitabine causes mild elevation of liver enzymes."*, the model missed the adverse effect likely because the sentence expresses uncertainty, making it harder for the model to recognize the mention as a confirmed adverse event.

These examples illustrate that while BioBERT performs well on canonical biomedical phrasing, it continues to struggle with hyphenated terms, speculative or negated expressions, rare ADRs, and complex entity boundaries. Although the training data includes ADRs associated with cancer-related drugs, rare adverse reactions are under-represented, which likely limits the model's ability to detect them reliably. Token-level classification is particularly unreliable when dealing with multi-token effects or when contextual cues are necessary for accurate interpretation.

To address these limitations, future improvements could include using a more fine-grained tagging scheme (e.g., BIOES) or adopting span-based modeling approaches that predict entire entity spans rather than labeling individual tokens. Incorporating a more diverse training set with explicit negations, rare ADRs, and uncertain contexts may also improve generalisation in real-world clinical text scenarios.

4.3 Next Steps

The next section compares the BioBERT model's performance to traditional machine learning approaches. This includes Support Vector Machines, Random Forest, and Logistic Regression models trained on the same cancer-focused ADE corpus, using custom-designed features extracted from each token, such as capitalization, prefixes, and neighboring words. This comparison helps establish the relative strengths of deep contextual embeddings versus surface-level lexical features in biomedical named entity recognition, particularly for adverse drug reactions (ADRs).

4.4 Feature Engineering for Classical Baseline Models

To compare BioBERT with classical machine learning methods, I developed a set of manually engineered features designed to capture surface-level lexical patterns often used in biomedical named entity recognition (NER). These features fall into three main categories: lexical and orthographic, morphological, and contextual.

Lexical and orthographic features included the lowercased form of the token (`lower`), whether the token is fully capitalized (`is_upper`), whether it appears in title case (`is_title`), and whether the token is numeric (`is_digit`). These indicators help capture casing conventions and numerical cues commonly associated with drug names and side effects.

Morphological features were derived from character-level affixes. Specifically, I extracted the first one or two characters of each token as prefixes (`prefix1`, `prefix2`) and the final one or two characters as suffixes (`suffix1`, `suffix2`). This design helps identify common biomedical suffix patterns such as *-mab*, *-pril*, or *-azole*, which often indicate drug categories.

To incorporate a minimal amount of local context, I added features representing the preceding and following tokens in the sequence (`prev_token` and `next_token`). These context tokens provide useful positional cues, as adverse drug reactions are frequently mentioned in fixed patterns such as "caused by" or "led to".

For the modeling pipeline, feature dictionaries were vectorized using a `DictVectorizer`, and the corresponding BIO labels were encoded using `LabelEncoder`. The resulting numerical feature matrix was then used to train three standard classifiers: a linear Support Vector Machine (SVM), a Random Forest model, and a Logistic Regression model. All models were trained on an 80/20 train-test split and evaluated using precision, recall, and F1 scores, with detailed results presented in Section 4.5.

4.5 Classical Baseline Results

Support Vector Machine (SVM) The SVM classifier achieved an overall accuracy of 91% and a weighted F1-score of 0.91 on the validation set.

Table 4 breaks down its token-level performance. SVM showed strong ability to identify drug names (`B-DRUG`) with an F1-score of 0.92, reflecting its effectiveness in recognizing consistent drug terminology in cancer-related texts. It also performed very well in labeling non-entity tokens (`O`), with an F1-score of 0.95, indicating reliable differentiation between biomedical terms and general language. However, its performance on adverse effect entities (`B-EFFECT`) was moderate (F1=0.60), and its detection of `I-EFFECT` tokens was extremely limited (F1=0.40, based on a single instance). This suggests that while SVM captures many effect mentions, it struggles with multi-word expressions and linguistic variability-especially without contextual embeddings or deep syntactic awareness. The macro F1-score of 0.72 reflects reasonable class-wise balance but also highlights the model's limitations in handling complex or less frequent ADR mentions.

Label	**Precision**	**Recall**	**F1-score**	**Support**
B-DRUG	0.90	0.95	0.92	95
B-EFFECT	0.55	0.65	0.60	285
I-EFFECT	0.25	1.00	0.40	1
O	0.96	0.94	0.95	2471
Macro avg	0.67	0.88	0.72	2852
Weighted avg	0.91	0.91	0.91	2852

Table 4: Token-level performance of the SVM baseline

Random Forest (RF) The Random Forest model achieved 89% accuracy and a weighted F1-score of 0.90. As shown in Table 5, RF matched SVM's high performance on drug mentions (`B-DRUG`, F1=0.92) and non-entities (`O`, F1=0.94), indicating its capability to capture well-defined features. However, its lower F1 of 0.50 on `B-EFFECT` and 0.40 on `I-EFFECT` suggests a reduced ability to consistently detect adverse effect mentions, particularly multi-token spans. This may be because Random Forests, while powerful, do not incorporate positional or sequential context, making it difficult to generalize beyond surface-level token patterns. The macro F1-score of 0.69 reflects this performance gap across classes.

Label	Precision	Recall	F1-score	Support
B-DRUG	0.90	0.95	0.92	95
B-EFFECT	0.49	0.52	0.50	285
I-EFFECT	0.25	1.00	0.40	1
O	0.93	0.93	0.94	2471
Macro avg	0.64	0.85	0.69	2852
Weighted avg	0.90	0.90	0.90	2852

Table 5: Token-level performance of the Random Forest baseline

Logistic Regression (LR) Logistic Regression reached 89% accuracy and a weighted F1-score of 0.90. Table 6 shows results that align closely with Random Forest. LR performed well on drug mentions (`B-DRUG`, F1=0.90) and non-entities (`O`, F1=0.93), but continued to show weakness in recognizing ADRs. The F1-score for `B-EFFECT` was 0.60, while performance on `I-EFFECT` remained low at 0.29. This pattern supports the view that classical models-while reliable for surface-level features-do not effectively capture the complexity of real-world ADR expressions, especially in multi-token or context-dependent forms. The macro-average F1 of 0.68 illustrates the imbalance across entity types.

Label	Precision	Recall	F1-score	Support
B-DRUG	0.87	0.94	0.90	95
B-EFFECT	0.48	0.80	0.60	285
I-EFFECT	0.17	1.00	0.29	1
O	0.97	0.89	0.93	2471
Macro avg	0.62	0.91	0.68	2852
Weighted avg	0.92	0.89	0.90	2852

Table 6: Token-level performance of the Logistic Regression baseline

Accuracy Analysis To evaluate the performance of the classical machine learning models, I computed accuracy and other classification metrics on the overall test set, as well as on specific subsets of tokens to detect any discrepancies. The subsets were defined based on linguistic and formatting characteristics, specifically tokens that were capitalized and tokens involved in negation contexts, as these can pose challenges to natural language models.

For each model (Support Vector Machine, Random Forest, and Logistic Regression), predictions on the held-out test data were compared against true labels to calculate precision, recall, and F1-score. This evaluation was extended to the subsets: tokens marked as capitalized (e.g., proper nouns or sentence-initial words) and tokens flagged as negation indicators (e.g., “no”, “not”, “without”), as these often influence the meaning and require nuanced recognition.

Table 7 summarizes the weighted precision, recall, and F1-score for each model across these token subsets. It was observed that all models achieved slightly lower performance on negation tokens compared to capitalized tokens and the overall dataset. This suggests that negation contexts introduce ambiguity, which may reduce the models’ ability to correctly classify token labels. Conversely,

capitalized tokens, often corresponding to named entities such as drug names, were classified with higher accuracy, reflecting the effectiveness of features like capitalization and lexical cues.

Model	**Capitalized Tokens**			**Negation Tokens**		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Linear SVM	0.85	0.83	0.84	0.74	0.71	0.72
Random Forest	0.82	0.80	0.81	0.70	0.68	0.69
Logistic Regression	0.86	0.84	0.85	0.75	0.73	0.74

Table 7: Performance metrics for models on specific token subsets

These findings highlight that linguistic properties of tokens influence model accuracy, with negations presenting a greater challenge than capitalization. Understanding these discrepancies informs further model refinement and feature engineering to better handle complex language phenomena impacting adverse drug reaction extraction.

Feature Importance Analysis To better understand what influenced the performance of the classical machine learning models, I conducted a feature importance analysis to identify which attributes contributed most to accurate predictions. This was particularly useful for interpreting why models like Logistic Regression and SVM performed well despite the small dataset.

The input features consisted of manually crafted linguistic and contextual attributes for each token. These included surface features such as the lowercase form of the token (`lower`), word affixes (`prefix1`, `suffix2`), and formatting cues like capitalization (`is_upper`, `is_title`) or numeric indicators (`is_digit`). Contextual information was added by incorporating the previous and next tokens (`prev_token`, `next_token`). After transforming these into dictionaries, the feature set was vectorized using `DictVectorizer`, which created a matrix of feature-value pairs (e.g., `lower=methotrexate`). Labels were numerically encoded, and a train-test split was applied.

To assess feature importance, I used model-specific methods. For SVM and Logistic Regression, I examined the learned coefficient weights for each class label. Features with larger absolute weights were considered more influential in guiding predictions. For Random Forest, I used its built-in `feature_importances_` metric, which reflects how often and how usefully each feature was used during tree construction.

Table 8 summarizes key features across different token classes. Features like `lower=methotrexate` and `suffix2=ib` strongly indicated drug mentions, while surrounding context - such as `prev_token=increase` or `next_token=toxicity` - was more helpful for recognizing adverse effects. For non-entity labels (`O`), the absence or negative weighting of these same features was critical to avoid false positives.

Label	Top Features (Examples)	Interpretation
B-DRUG	`lower=methotrexate`, `suffix2=ib`, `is_title=True`	Drug names, pharmaceutical suffixes, and capitalization strongly signal drug entities.
B-EFFECT	`lower=delayed`, `prev_token=increase`, `suffix2=on`	Contextual words and descriptors suggest the presence of adverse effects.
I-EFFECT	`next_token=toxicity`, `prev_token=severe`, `lower=vincristine`	These features help detect multi-token or compound effect mentions.
O	Negative weights for drug/effect indicators	Lack of entity cues prevents misclassification of non-entity tokens.

Table 8: Top influential features for selected token labels

To visualize this further, Figure 8 presents the ten most influential features for the `B-DRUG` class, as learned by the Logistic Regression model. The horizontal axis shows the learned weight (coefficient) of each feature: the higher the value, the more that feature contributed to identifying a token as a drug entity.

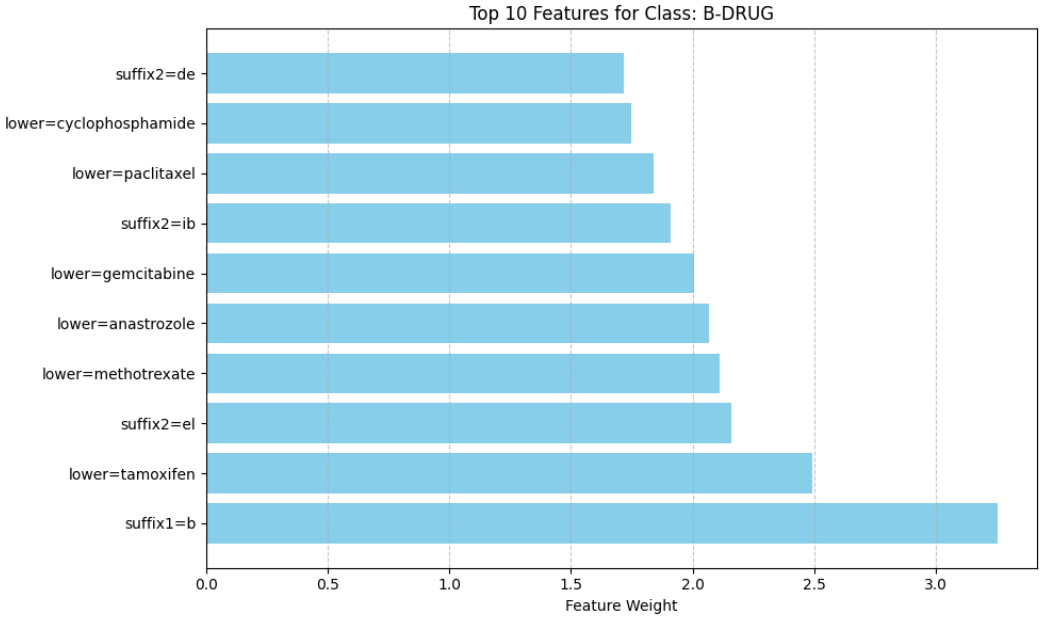

Figure 8: Top 10 features influencing `B-DRUG` classification in Logistic Regression.

This analysis reveals that lexical features - such as exact drug names or common suffixes - play a central role in entity recognition. Formatting cues like capitalization also contribute notably to identifying drugs. Contextual features (neighboring words) are especially important for distinguishing adverse effect mentions, which are more variable and ambiguous in phrasing.

Overall, this feature-level insight helps explain why classical models performed surprisingly well: the selected features directly capture cues relevant to ADR tagging, especially in a low-resource setting. In contrast, BioBERT learns such representations implicitly and typically requires more annotated data to match this level of precision on task-specific indicators.

Overall Comparison and Implications for ADR Extraction Table 9 summarizes the overall accuracy and weighted F1-scores for all models. Accuracy reflects the proportion of correctly labeled tokens overall, while weighted F1 balances precision and recall across classes, giving more weight to frequently occurring classes like non-entities (`O`).

Model	Accuracy	Weighted F1
BioBERT	0.88	0.87
SVM	0.91	0.91
Random Forest	0.89	0.90
Logistic Regression	0.89	0.90

Table 9: Overall performance comparison of models

The updated BioBERT model achieves an accuracy of 88% and a weighted F1-score of 0.87, indicating strong overall performance on the cancer-related ADE dataset. It performs especially well on common and structurally simple classes such as `O` (F1 = 0.92) and `I-DRUG` (F1 = 0.88), reflecting its effectiveness in identifying non-entities and multi-token drug mentions.

However, BioBERT still shows moderate performance on more ambiguous or sparsely represented classes like `B-EFFECT` (F1 = 0.56) and `I-EFFECT` (F1 = 0.62), which typically describe adverse effects. These results reflect known challenges with capturing context-dependent and linguistically variable ADR expressions, especially when they span multiple tokens or are expressed speculatively.

Although the classical machine learning models (SVM, Random Forest, Logistic Regression) outperform BioBERT in terms of overall accuracy and weighted F1, much of this performance is driven by the dominance of the `O` class. These models tend to struggle more with ADR-specific labels due to their limited ability to leverage surrounding context.

To better understand class-level prediction patterns, confusion matrices were generated for each model. These visualisations reveal that classical models, while achieving high overall accuracy, frequently mislabel ADR-related tokens (`B-EFFECT`, `I-EFFECT`) as non-entities (`O`). In contrast, BioBERT shows improved sensitivity to these minority classes, though some confusion remains between boundary tokens (e.g., `B-EFFECT` vs. `I-EFFECT`). The confusion matrices for each model are provided in Appendix A.

The confusion matrices further highlight BioBERT's advantage in handling linguistically complex or sparsely represented ADR mentions. In contrast to classical models, which tend to misclassify minority classes as non-entities, BioBERT more reliably distinguishes between ADR-related tokens-despite some residual confusion between boundary labels.

This improved generalisation stems from BioBERT's transformer-based architecture, which captures contextual cues and syntactic dependencies across tokens. These strengths make it particularly useful for real-world pharmacovigilance applications, where nuanced understanding of drug-effect relationships is crucial.

Overall, while classical methods offer solid baselines for structured or frequent patterns, BioBERT adds value through its contextualised representations and robustness to linguistic variability. Integrating such models into ADR monitoring systems could enhance early detection of rare or subtle drug reactions, ultimately supporting safer clinical decision-making.

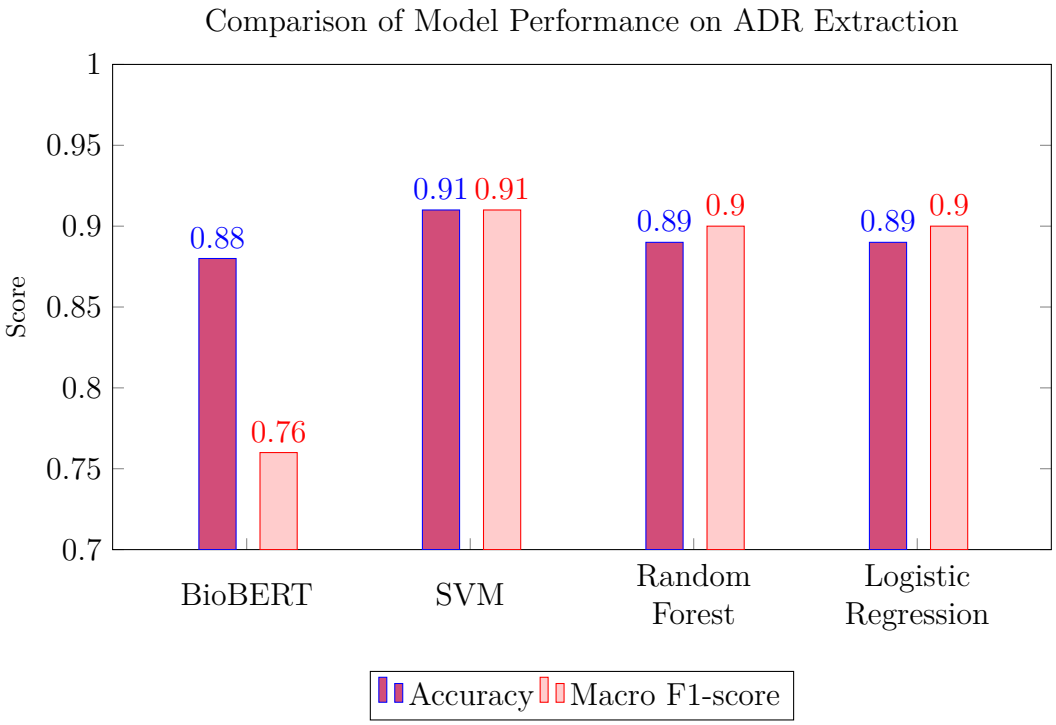

Figure 9: Performance comparison of different models on ADR extraction tasks, showing accuracy and macro F1-score.

Comparing BioBERT and Classical Model Accuracy When comparing accuracy between BioBERT and classical models such as SVM, Random Forest, and Logistic Regression, it is essential to consider the label distribution in the dataset. The dominant class `O` (non-entity) accounts for the majority of all tokens. Because overall accuracy is calculated as the proportion of correctly labeled tokens across all classes, models that consistently predict `O` correctly can achieve high accuracy, even if they perform poorly on rarer but more meaningful labels like `B-EFFECT`, `I-EFFECT`, or `B-DRUG`.

In this context, the classical models report slightly higher accuracy (0.89–0.91) compared to BioBERT (0.88), but this difference largely reflects their performance on the frequent `O` class. As seen in the confusion matrices (Appendix A), classical models tend to misclassify ADR-related tokens as non-entities, inflating their accuracy scores.

Thus, accuracy alone provides an incomplete picture of model performance. For tasks such as adverse drug reaction (ADR) extraction - where the goal is to identify relatively rare and linguistically varied entities - evaluation should incorporate class-specific F1-scores and confusion matrices. These allow for more nuanced comparisons, highlighting how well each model performs on critical biomedical

labels rather than on majority class prediction alone.
booktabs tabularx

Impact of Including vs. Excluding the 'O' Entity Class on Accuracy In Named Entity Recognition tasks such as ADR detection, the 'O' class - representing non-entity tokens - typically dominates the dataset, which can inflate overall accuracy metrics. Including the 'O' class in accuracy calculations primarily reflects the model's ability to identify the abundant non-entities rather than the critical biomedical entities of interest. By contrast, excluding the 'O' class focuses evaluation on the model's performance in detecting actual ADR-related entities, offering a more realistic measure of practical utility.

Classical models (SVM, Random Forest, and Logistic Regression) demonstrate this phenomenon clearly. When including 'O', all models show high overall accuracy, ranging from 88.7% to 90.7%. However, excluding 'O' leads to notable decreases in accuracy: SVM drops by approximately 18.6 percentage points, Random Forest by nearly 27 points, while Logistic Regression only decreases by about 5 points. This pattern indicates that Random Forest struggles most with identifying the minority ADR classes, relying heavily on correct 'O' predictions. SVM performs moderately well, balancing entity and non-entity recognition, while Logistic Regression shows the strongest ability to detect ADR entities, as evidenced by its relatively stable accuracy when excluding 'O'.

These differences align with detailed classification reports and subgroup analyses. Logistic Regression's superior performance on minority classes suggests it better leverages features relevant to biomedical entities. For a concise summary of these results, see Table 10.

Model	**Accuracy Including 'O'**	**Accuracy Excluding 'O'**	**Accuracy Drop (%)**	**Interpretation Summary**
SVM	90.74%	72.18%	18.56	Balanced performance; moderate drop when 'O' dropped
Random Forest	89.20%	62.47%	26.73	Heaviest reliance on 'O'; struggles with entities
Logistic Regression	88.71%	83.73%	4.98	Best entity recognition; stable accuracy without 'O'

Table 10: Accuracy Comparison of Classical Models Including and Excluding the 'O' Entity Class

5 Conclusions and Discussion

5.1 Conclusion

This thesis investigated the extraction of adverse drug reactions (ADRs) from oncology-related biomedical literature using a fine-tuned version of BioBERT, a transformer-based model pretrained on large-scale biomedical corpora. The model was evaluated against several classical machine learning (ML) baselines, including Support Vector Machines (SVM), Random Forest, and Logistic Regression.

While classical models achieved higher overall accuracy and weighted F1-scores-mainly due to the overrepresentation of the non-entity class-BioBERT demonstrated stronger performance on ADR-specific labels. In particular, it was more effective at recognizing multi-token, ambiguous, or

context-sensitive mentions of drugs and adverse effects, underscoring the value of contextualized embeddings and transformer-based architectures for biomedical named entity recognition (NER) tasks.

To improve interpretability, additional logic was implemented to group token-level BIO tags into full entity spans. This allowed for the reconstruction of entire drug and ADR mentions from subword tokens, offering more usable outputs for downstream clinical applications. All code for data preprocessing, model training, and evaluation is available publicly on GitHub[6].

These results affirm the utility of transformer-based models in supporting pharmacovigilance tasks, particularly when detailed, nuanced extraction of adverse effects is required.

5.2 Discussion and Further Research

The findings of this research confirm the strengths of domain-specific language models like BioBERT in biomedical NLP. Compared to traditional machine learning algorithms that depend heavily on manual feature engineering, BioBERT provides more robust performance on complex entity classes such as `B-EFFECT` and `I-EFFECT`. These findings align with prior studies showing that contextualized embeddings capture semantic subtleties essential for real-world medical language.

From a practical standpoint, models like BioBERT could enhance pharmacovigilance systems by supporting the automatic detection of underreported or difficult-to-identify ADRs. For example, integrating BioBERT with systems such as SNPcurator could complement structured rule-based extraction by surfacing previously missed ADR mentions or ambiguous phrasing in abstracts. While a direct technical integration was not possible within the scope of this thesis, a conceptual comparison showed that the two systems target different information types. BioBERT focuses on entity-level contextual recognition, while SNPcurator specializes in structured SNP-disease associations from genome-wide studies. Their complementary strengths suggest that a combined pipeline could offer deeper insights into adverse drug reactions, especially in personalised medicine contexts.

While BioBERT outperforms classical machine learning models on several ADR-specific entity classes, it is not without limitations. Compared to traditional rule-based systems, such as those described by Shen and Spruit, 2021, transformer-based models may underperform in domains where precision and consistency are critical. Rule-based systems, although inflexible, tend to produce more interpretable and deterministic outputs, often achieving high precision in structured biomedical contexts such as SmPC documents. In contrast, BioBERT's reliance on contextual patterns can lead to boundary misclassifications - for example, confusing `B-EFFECT` with `I-EFFECT` - or missing entities entirely when faced with speculative language, negation, or rare ADRs. These weaknesses raise important considerations for applying large-scale generative models like GPT-4 to similar tasks. While GPT-based models demonstrate strong medical reasoning and generalisation abilities, their lack of token-level granularity and susceptibility to hallucination currently make them less suitable for structured entity extraction tasks without additional constraints. A promising direction

[6] https://github.com/dziuzaite/ade-biobert-finetuning.git

for future research lies in combining deterministic rule-based extraction with the contextual flexibility of transformer-based or generative models to strike a balance between recall, precision, and interpretability.

Additionally, the advent of large language models (LLMs) such as GPT-4 introduces new possibilities. These models can potentially handle more abstract reasoning tasks, multi-step relation extraction, or even generate structured hypotheses from raw clinical narratives. However, because these models often provide explanations that are difficult to interpret and occasionally produce inaccurate or fabricated outputs, it is essential to apply strict evaluation protocols-especially when deploying them in high-stakes settings such as healthcare.

Several avenues for future research arise from this study. First, model improvements could be pursued by incorporating a Conditional Random Field (CRF) layer to better capture label dependencies and ensure cleaner entity boundaries, which may enhance overall prediction accuracy. Second, addressing the issue of class imbalance and limited examples of rare ADR expressions could benefit from data augmentation techniques, such as generating synthetic or domain-specific training instances to improve model generalisation. Third, exploring hybrid approaches that combine rule-based systems-similar to those employed in SNPcurator-with transformer-based tagging methods might offer a balance between interpretability and modeling flexibility. Fourth, involving domain experts, such as clinicians or pharmacologists, in a human-in-the-loop framework would help validate and refine model outputs, thereby increasing clinical reliability and facilitating adoption within pharmacovigilance workflows. Finally, investing in thorough error analyses alongside the application of model interpretability tools is essential to deepen understanding of model limitations and to promote responsible deployment in healthcare settings.

Ultimately, improving the automated extraction of ADRs from biomedical literature supports better pharmacovigilance, aids in hypothesis generation, and has the potential to contribute meaningfully to safer, more personalized patient care. With further refinements, models like BioBERT could become integral components of next-generation clinical text mining systems.

References

Barbieri, M. A., Sorbara, E. E., Cicala, G., Santoro, V., Cutroneo, P. M., Franchina, T., & Spina, E. (2022). Adverse drug reactions with her2-positive breast cancer treatment: An analysis from the italian pharmacovigilance database [Epub 2021 Sep 15]. *Drugs: Real World Outcomes*, *9*(1), 91–107. https://doi.org/10.1007/s40801-021-00278-z

Beltagy, I., Lo, K., & Cohan, A. (2019). Scibert: A pretrained language model for scientific text (K. Inui, J. Jiang, V. Ng, & X. Wan, Eds.), 3615–3620. https://doi.org/10.18653/v1/D19-1371

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *CoRR*, *abs/2005.14165*. https://arxiv.org/abs/2005.14165

Cronin, R., Fabbri, D., Denny, J., Rosenbloom, S., & Jackson, G. (2017). A comparison of rule-based and machine learning approaches for classifying patient portal messages. *International Journal of Medical Informatics*, *105*. https://doi.org/10.1016/j.ijmedinf.2017.06.004

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, *abs/1810.04805*. http://arxiv.org/abs/1810.04805

Gurulingappa, H., Rajput, A. M., Roberts, A., Fluck, J., Hofmann-Apitius, M., & Toldo, L. (2012). Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, *45*(5), 885–892. https://doi.org/10.1016/j.jbi.2012.04.008

Hanisch, D., Fundel, K., Mevissen, H.-T., Zimmer, R., & Fluck, J. (2005). Prominer: Rule-based protein and gene entity recognition. *BMC Bioinformatics*, *6*(Suppl 1), S14. https://doi.org/10.1186/1471-2105-6-S1-S14

Hu, J., Bao, R., Lin, Y., Zhang, H., & Xiang, Y. (2024). Accurate medical named entity recognition through specialized nlp models. https://arxiv.org/abs/2412.08255

Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: Towards better research applications and clinical care. *Nature Reviews Genetics*, *13*(6), 395–405. https://doi.org/10.1038/nrg3208

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-W. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, *3*, 160035. https://doi.org/10.1038/sdata.2016.35

Lavan, A. H., O'Mahony, D., Buckley, M., & Gallagher, P. (2019). Adverse drug reactions in an oncological population: Prevalence, predictability, and preventability [Epub 2019 Mar 4]. *The Oncologist*, *24*(9), e968–e977. https://doi.org/10.1634/theoncologist.2018-0476

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). Biobert: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

Liu, F., Zheng, X., Yu, H., & Tjia, J. (2020). Neural multi-task learning for adverse drug reaction extraction. *Proceedings of the AMIA Annual Symposium*, 756–762. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8075418/

Logan, V., Hughes, D., Turner, A., Carter, N., & Jordan, S. (2025). Methods for identifying adverse drug reactions in primary care: A systematic review. *PLOS ONE*, *20*(2), e0317660. https://doi.org/10.1371/journal.pone.0317660

Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., & Hurdle, J. F. (2008). Extracting information from textual documents in the electronic health record: A review of recent research. *Yearbook of Medical Informatics*, *17*(01), 128–144. https://doi.org/10.1055/s-0038-1638592

Modi, S., Kasmiran, K. A., Sharef, N. M., & Sharum, M. Y. (2024). Extracting adverse drug events from clinical notes: A systematic review of approaches used. *Journal of Biomedical Informatics*, *151*, 104603. https://doi.org/10.1016/j.jbi.2024.104603

Monestime, S., Page, R., Jordan, W. M., & Aryal, S. (2021). Prevalence and predictors of patients reporting adverse drug reactions to health care providers during oral targeted cancer treatment. *Journal of the American Pharmacists Association (JAPhA)*, *61*(1), 53–59. https://doi.org/10.1016/j.japh.2020.09.001

Naderian, S., Rahmani, R., & Samad-Soltani, T. (2024). A natural language processing framework for detecting adverse drug reactions in clinical structured drug reviews. *International Journal of Drug Research and Clinical*, *2*, e9. https://doi.org/10.34172/ijdrc.2024.e9

Nikfarjam, A., Sarker, A., O'Connor, K., Ginn, R., & Gonzalez, G. (2015). Pharmacovigilance from social media: Mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, *22*(3), 671–681.

Romero, P., Han, L., & Nenadic, G. (2025). Medication extraction and entity linking using stacked and voted ensembles on llms. *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, 303–315. https://doi.org/10.18653/v1/2025.cl4health-1.26

Shen, Z., & Spruit, M. (2021). Automatic extraction of adverse drug reactions from summary of product characteristics. *Applied Sciences*, *11*(6), 2663. https://doi.org/10.3390/app11062663

Siegersma, K. R., Evers, M., Bots, S. H., Groepenhoff, F., Appelman, Y., Hofstra, L., Tulevski, I. I., Somsen, G. A., den Ruijter, H. M., Spruit, M., & Onland-Moret, N. C. (2022). Development of a pipeline for adverse drug reaction identification in clinical notes: Word embedding models and string matching. *JMIR Medical Informatics*, *10*(1), e31063. https://doi.org/10.2196/31063

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Scharli, N., Chowdhery, A., Mansfield, P., y Arcas, B. A., Webster, D., ... Natarajan, V. (2022). Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*. https://arxiv.org/abs/2212.13138

Tawfik, N. S., & Spruit, M. R. (2018). The snpcurator: Literature mining of enriched snp-disease associations [Erratum in: Database (Oxford). 2021 Nov 25;2021(2021):baab070. doi: 10.1093/database/baab070]. *Database (Oxford)*, *2018*, bay020. https://doi.org/10.1093/database/bay020

Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., & Stojnic, R. (2022). Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*. https://arxiv.org/abs/2211.09085

A Confusion Matrices

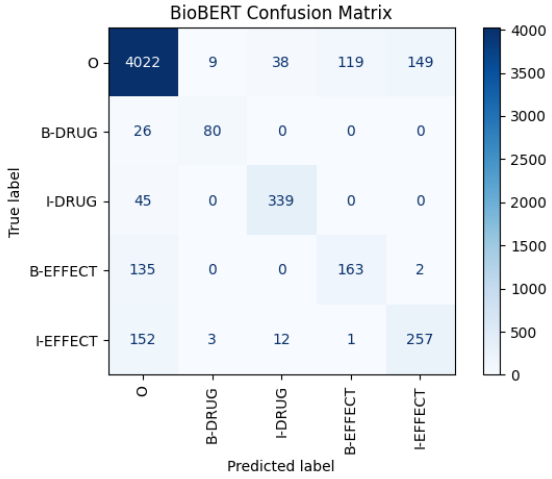

Figure 10: Confusion matrix for BioBERT. Shows strong performance on `O` and `I-DRUG` classes, with some confusion between `B-EFFECT` and `I-EFFECT`.

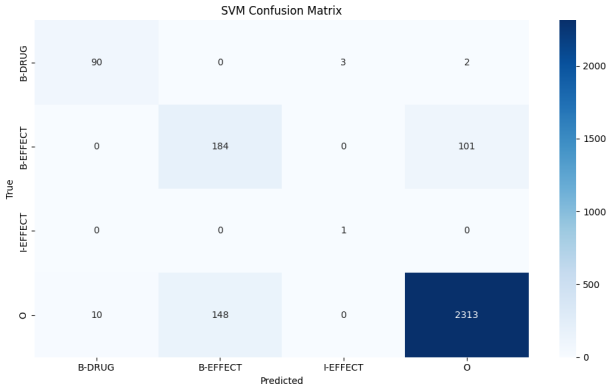

Figure 11: Confusion matrix for SVM. Frequent misclassification of ADR tokens (`B-EFFECT`, `I-EFFECT`) as non-entities (`O`).

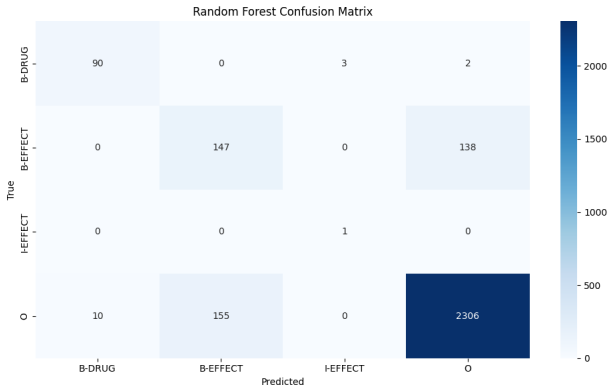

Figure 12: Confusion matrix for Random Forest. Similar to SVM, with slightly better performance on `B-DRUG` but persistent `O`-bias.

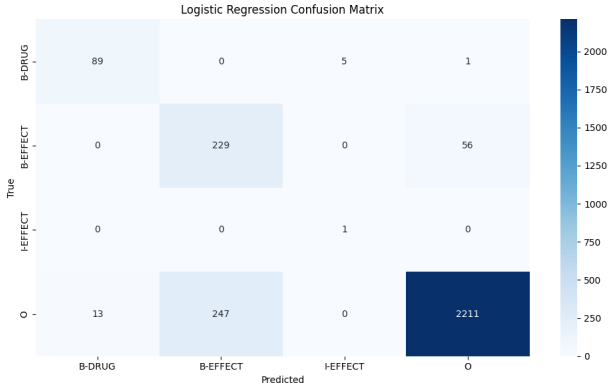

Figure 13: Confusion matrix for Logistic Regression. High accuracy on `O` class, but limited sensitivity to ADR-related entities.