Journal of Biomedical Informatics 152 (2024) 104621

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Methodological Review

Artificial intelligence-powered pharmacovigilance: A review of machine and deep learning in clinical text-based adverse drug event detection for benchmark datasets

Yiming Li [a], Wei Tao [b], Zehan Li [a], Zenan Sun [a], Fang Li [c], Susan Fenton [a], Hua Xu [d], Cui Tao [c,*]

[a] McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA
[b] Department of Biostatistics & Data Science, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA
[c] Department of Artificial Intelligence and Informatics, Mayo Clinic, Jacksonville, FL 32224, USA
[d] Section of Biomedical Informatics and Data Science, School of Medicine, Yale University, New Haven, CT 06510, USA

ARTICLE INFO

Keywords:
Pharmacovigilance
Machine learning/Deep learning
Adverse drug event (ADE) extraction
named-entity recognition (NER)
Relation extraction (RE)
Natural language processing (NLP)

ABSTRACT

Objective: The primary objective of this review is to investigate the effectiveness of machine learning and deep learning methodologies in the context of extracting adverse drug events (ADEs) from clinical benchmark datasets. We conduct an in-depth analysis, aiming to compare the merits and drawbacks of both machine learning and deep learning techniques, particularly within the framework of named-entity recognition (NER) and relation classification (RC) tasks related to ADE extraction. Additionally, our focus extends to the examination of specific features and their impact on the overall performance of these methodologies. In a broader perspective, our research extends to ADE extraction from various sources, including biomedical literature, social media data, and drug labels, removing the limitation to exclusively machine learning or deep learning methods.

Methods: We conducted an extensive literature review on PubMed using the query "(((machine learning [Medical Subject Headings (MeSH) Terms]) OR (deep learning [MeSH Terms])) AND (adverse drug event [MeSH Terms])) AND (extraction)", and supplemented this with a snowballing approach to review 275 references sourced from retrieved articles.

Results: In our analysis, we included twelve articles for review. For the NER task, deep learning models outperformed machine learning models. In the RC task, gradient Boosting, multilayer perceptron and random forest models excelled. The Bidirectional Encoder Representations from Transformers (BERT) model consistently achieved the best performance in the end-to-end task. Future efforts in the end-to-end task should prioritize improving NER accuracy, especially for 'ADE' and 'Reason'.

Conclusion: These findings hold significant implications for advancing the field of ADE extraction and pharmacovigilance, ultimately contributing to improved drug safety monitoring and healthcare outcomes.

Abbreviations: ADE, Adverse drug event; AMIA, American Medical Informatics Association; Att-BiLSTM, Attention-based Bidirectional long-short term memory; BERT, Bidirectional Encoder Representations from Transformers; BiLSTM, Bidirectional long-short term memory; CapNet, Capsule network; CNN, Convolutional neural network; CRF, Conditional Random Field; DL, Deep learning; EHR, Electronic health record; FDA, Food and Drug Administration; GB, Gradient boosting; GloVe, Global Vectors; i2b2, Informatics for Integrating Biology & the Bedside; IE, Information extraction; KB, Knowledge base; KNN, K-nearest neighbors; LLM, Large language model; LR, Logistic regression; MADE, Medication and Adverse Drug Event; MEDLINE, Medical Literature Analysis and Retrieval System Online; MeSH, Medical Subject Headings; MIMIC-III, Medical Information Mart for Intensive Care III; ML, Machine learning; MLP, Multilayer perceptron; n2c2, National NLP Clinical Challenge; NER, Named-entity recognition; NLP, Natural language processing; PoS, Part of speech; PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses; RC, Relation classification; RCNN, Recurrent convolutional neural network; RE, Relation extraction; RF, Random forest; RNN, Recurrent neural network; SemMedDB, Semantic MEDLINE Database; SNOMED CT, Systematized Nomenclature of Medicine Clinical Terms; SSEL, Semi-Supervised Ensemble Learning; SVM, Support vector machine; TAC, Text Analysis Conference; TransE, Translating Embeddings for Modeling Multi-relational Data; UMLS, Unified Medical Language System; US, United States.

* Corresponding author at: 4500 San Pablo Rd, Jacksonville, FL 32224. Tel.: +1 904 953 0255.
E-mail address: tao.cui@mayo.edu (C. Tao).

https://doi.org/10.1016/j.jbi.2024.104621
Received 21 November 2023; Received in revised form 19 February 2024; Accepted 3 March 2024
Available online 5 March 2024

1. Introduction

Adverse Drug Event (ADE) refers to harm caused by medical interventions associated with drugs, involving overdoses, allergic reactions, adverse drug reactions, and medical errors [1,2]. It is estimated that ADEs constitute 30 % of all hospital adverse events [3,4]. Medical errors, including ADEs, were the third leading cause of death in the United States (US) in 2013 [5,6]. Experiencing an ADE can result in longer hospitalizations, serious morbidity and mortality, and additional healthcare costs [7–11].

Majority of ADEs are preventable, and lessons learned from previous ADEs provide guidance for the prevention of future ADEs [8]. Hence, pharmacovigilance aims to characterize, detect, monitor, and prevent ADEs [12,13].

Electronic health records (EHRs), which contain multiple documents that record detailed patient treatments and responses, are nearly ubiquitous in the United States in clinical settings. Documents within EHR, such as progress notes and discharge summaries can be utilized as resources for ADE detection and research [8,14]. However, reviewing and manually extracting ADE-related information from clinical notes in EHRs is hard, costly, and more specifically, not realistic [8]. Hence, it is imperative to make use of natural language processing (NLP) techniques for the automatic collection of ADEs [5].

Typically, there are two tasks associated with ADE extraction: named entity recognition (NER) and relation classification (RC). ADE mentions and drug-related attributes should be identified in the NER task [15]. The RC task aims to assign the relation between the drug entity and other entities [8].

There were several challenges covering the task of ADE collections. The 2018 National NLP Clinical Challenge (n2c2) Shared Task on Adverse Drug Events and Medication Extraction and the 2018 Medication and Adverse Drug Event (MADE) 1.0 Challenge are chosen for methodological review because they offer standardized benchmark datasets for extracting medication and adverse drug event information from clinical text, a crucial task in healthcare informatics. Researchers prefer these challenges due to their direct relevance to clinical NLP, real-world applications, and the opportunity to advance NLP techniques, compare methods, and publish results in reputable venues. They provide a structured and transparent platform for evaluating and improving NLP methods in the context of clinical data extraction. Therefore, the objective of the article is to conduct a review about the use of machine learning (ML)/ deep learning (DL) models in ADE extraction from clinical text. In this review, only the studies using the two benchmark datasets in second track of the 2018 n2c2 Shared Task on ADEs and Medication Extraction, and 2018 MADE 1.0 Challenge were considered for the convenience of comparison. In this article, our primary emphasis is on examining the variations in methods, including models and inherent features, for ADE extraction in the NER and RC tasks using the benchmark data sets of the two challenges. We aim to investigate the impact of ensembles and knowledge-based features on the NER task. Additionally, the influence of embeddings on both NER and RC tasks will be explored.

2. Method

There are numerous articles available studying the ADE extraction from text. However, the focus of the research is on extracting ADEs from clinical notes through either ML or DL method. Those articles discussing ADE extraction from biomedical literature, social media posts, or drug labels as well as articles applying either rule-based or knowledge-based methods to extract ADEs will not be reviewed in detail [16].

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flowchart indicating the article selection process is illustrated in Fig. A.1.

2.1. Article search process

PubMed was the database used to search for articles. The query was "(((machine learning [Medical Subject Headings (MeSH) Terms]) OR (deep learning [MeSH Terms])) AND (adverse drug event [MeSH Terms])) AND (extraction)". There were 59 results returned from the query.

In addition to the database search, two articles were included by snowballing from the 275 references from the articles from PubMed. Two thousand one hundred seventy-five articles were identified from publishers, 176 from the American Medical Informatics Association (AMIA) (https://academic.oup.com/jamia), and 1999 from Elsevier (https://www.elsevier.com/) by searching with the keyword "machine learning for adverse drug event extraction".

2.2. Inclusion/Exclusion criteria

The criteria used for including articles in the literature review:

- Articles published in or after 2015, and
- A clinical text dataset as the study corpus, and
- The use of ML methods, and/or
- The use of DL methods, and
- The methods general to all types of ADEs

The criteria used for excluding articles in the literature review:

- Articles published prior to 2015, and
- The study corpus not in English, and
- Not ML/DL methods, and
- Not applied to clinical text, and
- Limited to specific kinds of ADEs

As Fig. A.1 illustrates, for the 59 articles retrieved from PubMed, 35 articles were not related to the study topic. For the remaining 24 articles, two articles were not accessible. There were no articles published before 2015. Two articles were excluded because the corpus used was not in English, and three articles because they did not apply ML or DL methods. Two articles were screened out because the corpus was either social media text or biomedical literature. And there were eight articles limited to specific kinds of ADEs, which were also removed.

For the 275 articles identified from snowballing, 273 were not related to the study topic and were therefore not reviewed.

For the 176 articles retrieved from AMIA, 12 records were duplicates. Ninety-eight articles were not associated with the study topic and 52 articles were published before 2015. For the remaining 14 articles, one article was not included because the corpus employed was not in English. Four articles were excluded because ML/DL methods were not used. Additionally, one article was excluded because the corpus was not clinical text. And six articles were screened out because the studies reported were limited to specific kinds of ADEs.

For the 1999 articles found in Elsevier, 43 were duplicate records, and 1786 articles were not correlated to the study topic. Additionally, four articles were inaccessible and 153 articles were published prior to 2015. For the remaining articles, two articles were screened out because the corpus used was not in English. Three articles were removed because they did not use ML or DL methods, two articles because they did not apply to clinical text, and five articles because they were limited to specific kinds of ADEs.

2.3. Corpora used

Five corpora were involved in the reviewed articles: 2018 n2c2 dataset, 2018 MADE 1.0 Challenge dataset, the cardio corpus, Informatics for Integrating Biology & the Bedside (i2b2) 2009 Medication Challenge dataset, and Text Analysis Conference (TAC) 2017 dataset.

2018 n2c2 dataset consisted of "ADE" mentions, drug-related attributes ("Drug", "Strength", "Route", "Form", "Dosage", "Reason", "Frequency"), and eight drug-related relations between drug and other non-drug entity types from 505 patient discharge summaries extracted from the Medical Information Mart for Intensive Care III (MIMIC-III) database [4]. The 2018 MADE 1.0 Challenge dataset contained nine entity types ("Medication", "Indication", "Frequency", "Severity", "Dosage", "Duration", "Route", "ADE", and "SSLIF" (any sign, symptom, and disease that is not an ADE or indication)), and seven relation types among these entities from 1089 EHRs of patients with cancer [17]. The cardio corpus consisted of the same entity types and relation types as the 2018 MADE 1.0 Challenge dataset from 485 EHRs of patients with cardiovascular diseases [17]. The i2b2 2009 Medication Challenge dataset focused on the identification of medications and their associated dosages, routes, frequencies, and durations from 1243 clinical notes [17]. The TAC 2017 dataset contained adverse drug reaction mention, five attributes ("Animal", "DrugClass", "Factor", "Negation" and "Severity"), and three relation types ("Effect", "Hypothetical", and "Negated") from 200 drug labels [18]. To allow direct comparison, this review focuses on the 2018 n2c2 dataset and 2018 MADE 1.0 Challenge dataset.

2.4. Included articles & datasets

There were twelve articles included in the review (Table A.1). Among them, seven articles were related to the second track of the 2018 n2c2 Shared Task on ADEs and Medication Extraction (Dataset details in Table A.2), five articles discussed the 2018 MADE 1.0 Challenge (Dataset details in Table A.3).

3. Results

3.1. NER task

3.1.1. Overview of model performance

As we delve into the specifics of ADE extraction methodologies, the diversity of approaches in the 2018 n2c2 and MADE 1.0 Challenge becomes evident. In the 2018 n2c2 NER task (Table 1), teams employed various models. Conditional random field (CRF) was the sole ML-based model used for entity extraction. Several DL-based models included Bidirectional long-short term memory (BiLSTM)-CRF, convolutional neural network (CNN)-BiLSTM-CRF, LSTM-BiLSTM-CRF, CNN-RNN combination, and Bidirectional Encoder Representations from Transformers (BERT)-related models. Ensembles were explored by multiple teams [10,13,15], with Wei et al. [13] comparing different ensemble methods: committee strategy, stacking strategy, and category-level best. These methods aimed to enhance performance by combining predictions from multiple models.

The models used in the NER task for the 2018 MADE 1.0 Challenge are summarized in Table 2. Similar to the 2018 n2c2 NER task, CRF was the only ML-based model considered. BiLSTM-CRF and LSTM-CRF were the DL-based models investigated. No teams studied the use of ensembles for the NER task.

Comparing the NER task of ADE extraction in the 2018 n2c2 and the 2018 MADE 1.0 Challenge, CRF was the only ML-based model. An LSTM-based model was used by the majority of the teams as the DL-based model. Local features, context features, and orthographic features were the most popular features utilized. Embeddings mean representing words/characters/positions as numeric vectors without human intervention [19,20]. Character embedding and word embedding were adopted by Chapman et al. [19] and Yang et al. [20]. Yang et al. [14] also introduced knowledge-based features to their deep learning based-models, which performed better than the simply knowledge-based models in the NER task.

Fig. 1 illustrates the performance comparison between different NER methods and models from different research teams for 2018 n2c2 ADE extraction. The knowledge-based method was much inferior to the ML-based model and the DL-based models with an F1-score of only 0.85. However, there was not much difference in performance between CRF and other DL-based models. The CRF model by Dai et al. [10] produced the same F1-score as the C (convolutional neural network) -BiLSTM-CRF and the L (long short-term memory)-BiLSTM-CRF model they used. The F1 scores of the C-BiLSTM-CRF and L-BiLSTM-CRF model proposed by Dai et al. [10] were even lower than the CRF model submitted by Ju et al. [15] and Yang et al. [14]. Ju et al. [15] developed a CRF model of which the F1-score could reach up to 0.93, and was comparable to other DL-based models. The F1 scores of LSTM-CRF based models by Wei et al. [13] and Yang et al. [14] were stable at 0.93. The SciBERT, BioBERT, and BlueBERT achieved the best F1-score of 0.94.

Fig. A.2 and Fig. A.3 show the performance of each method for each entity type in the 2018 n2c2 and 2018 MADE 1.0 Challenge, respectively. Every entity achieved an F1-score exceeding 0.9 in the 2018 n2c2, and an F1- score greater than 0.8 in the 2018 MADE 1.0 Challenge except "Duration", "Reason", "ADE", "Severity", and "Indication".

Table 1
Summary of included articles for NER model in 2018 n2c2 Shared Task on ADEs and Medication Extraction.

Articles	NER model		Features
Chen et al. [8]	DL-based Knowledge-based[b]	BiLSTM-CRF[a] NA	NA
Dai et al. [10][c]	ML-based DL-based	CRF CNN-BiLSTM-CRF LSTM-BiLSTM-CRF	Part of speech (PoS), context words, chunk, morphological features, orthographic features, common medical abbreviations, ADE, drug and disease dictionary features, word cluster features
Ju et al. [15][c]	ML-based	CRF	Orthographic features, lexical features, syntactic features, dictionary features, cluster features
	DL-based	BiLSTM-CRF	Orthographic features, context features,
Wei et al. [13][c]	ML-based	CRF	Word shape, n-gram, prefix-suffix, orthographic, discretized word embeddings, etc.
	DL-based	BiLSTM-CRF	Word embedding and character embedding
		BI-LSTM-CRF (Joint learning)	Recognized drug entities, etc.
Yang et al. [14]	DL-based	LSTM-CRF CNN and RNN	Knowledge-based features, word embeddings
El-allaly et al. [18]	DL-based	BERT, BioBERT (a language representation model on biomedical text mining tasks), SciBERT, RoBERTa (a robustly optimized BERT pretraining approach), BlueBERT (a language model trained on Biomedical and Clinical texts)	NA

[a] For Frequency and Duration entity.
[b] For all other entities except Frequency and Duration
[c] Ensemble applied.

Table 2
Summary of included articles for NER model in 2018 MADE 1.0 Challenge.

Articles	NER model		Features
Chapman et al. [19]	ML-based	CRF	Local features (token, stem, PoS tag, patterns of capitalization, digits, and punctuation, prefix and suffix characters (n = 2, 3), embedding clusters from unigram and bigrams, drug lexicon match, drug lexicon match to the left or right of the current word
Dandala et al. [21]	DL-based	BiLSTM-CRF	NA
F. Li et al. [5]			Words, whether capitalization for the initial character, PoS tag, character representation
Yang et al. [20]	ML-based	CRF	N-gram, prefixes, suffixes, word shape, sentence-level feature, brown clustering, discrete word embedding
	DL-based	LSTM-CRF	Character embedding, word embedding

Fig. 2 illustrates the performance comparison between different NER methods and models proposed by different research teams in the 2018 MADE 1.0 Challenge. There is a large performance difference between CRF and DL-based models. Chapman et al. [19] developed two CRF models with 0.796 and 0.809 in F-1 score. The CRF model by Yang et al. [20] achieved an F1-score of only 0.725. All DL-based models have an F1-score greater than 0.8134 (one RNN model by Yang et al. [20]).

3.2. RC task

3.2.1. Overview of model performance

In this comprehensive review of ADE extraction methodologies, we delve into an analysis of models and techniques utilized in the 2018 n2c2 and MADE 1.0 Challenge. Table 3 and Table 4 provide a detailed breakdown of models and features deployed by different research teams in the RC task, offering valuable insights into the diverse landscape of approaches. From Support Vector Machines (SVM) and Random Forests (RF) to advanced deep learning architectures like CNN and Bidirectional Long Short-Term Memory networks (BiLSTM), the spectrum of methodologies reflects the dynamic nature of ADE research. This analysis sets the stage for an in-depth exploration of the performance and implications of these strategies in the subsequent sections.

Table 3 summarizes the models and their related features employed by different research groups in the 2018 n2c2 RC task. SVM, as an ML-based model, was studied by Wei et al. [13] and Yang et al. [14] in the RC task. Additionally, Yang et al. [14] also developed both RF and gradient boosting (GB) models for the RC task. For the DL-based model, CNN was proposed by Chen et al. [8] and Mahendran et al. [4]. Wei et al. [13] integrated CNN and RNN for the RC task. BiLSTM-based models were investigated by Chen et al. [8] and Wei et al. [13]. BERT-related models were explored by Mahendran et al. [4] and El-allaly et al. [18]. For the features employed by the teams, entity features, context features, and position features were generally employed. Wei et al. [13] also introduced the embeddings into their CNN-RNN model.

Table 4 summarizes the models and corresponding features utilized by different research teams for the RC task in the 2018 MADE 1.0 Challenge. Similar to the 2018 n2c2, SVM and RF were also employed as the ML-based models for the RC task. Dandala et al. [21] and F. Li et al. [5] investigated the application of Att-BiLSTM to the RC task. F. Li and Yu [17] explored the use of multilayer perceptron (MLP) and Capsule network (CapNet). Entity features, position features, and context features were the features commonly used in the models. F. Li and Yu [17] also introduced embedding features into the model.

Fig. 3 illustrates the overall RC performance for the 2018 n2c2. The two CNN models proposed by Mahendran and McInnes [4] did not achieve a satisfactory performance, with an F1-score of only 0.78. There

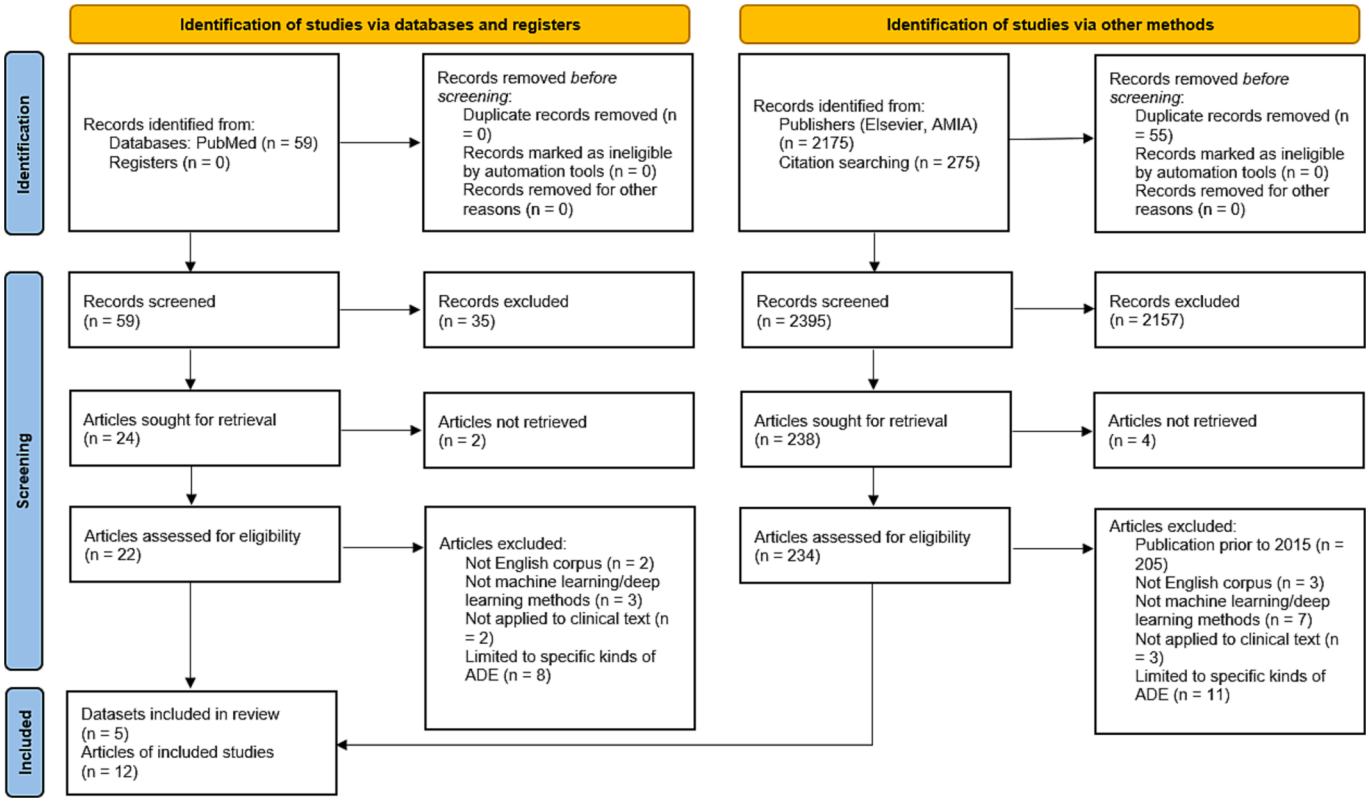
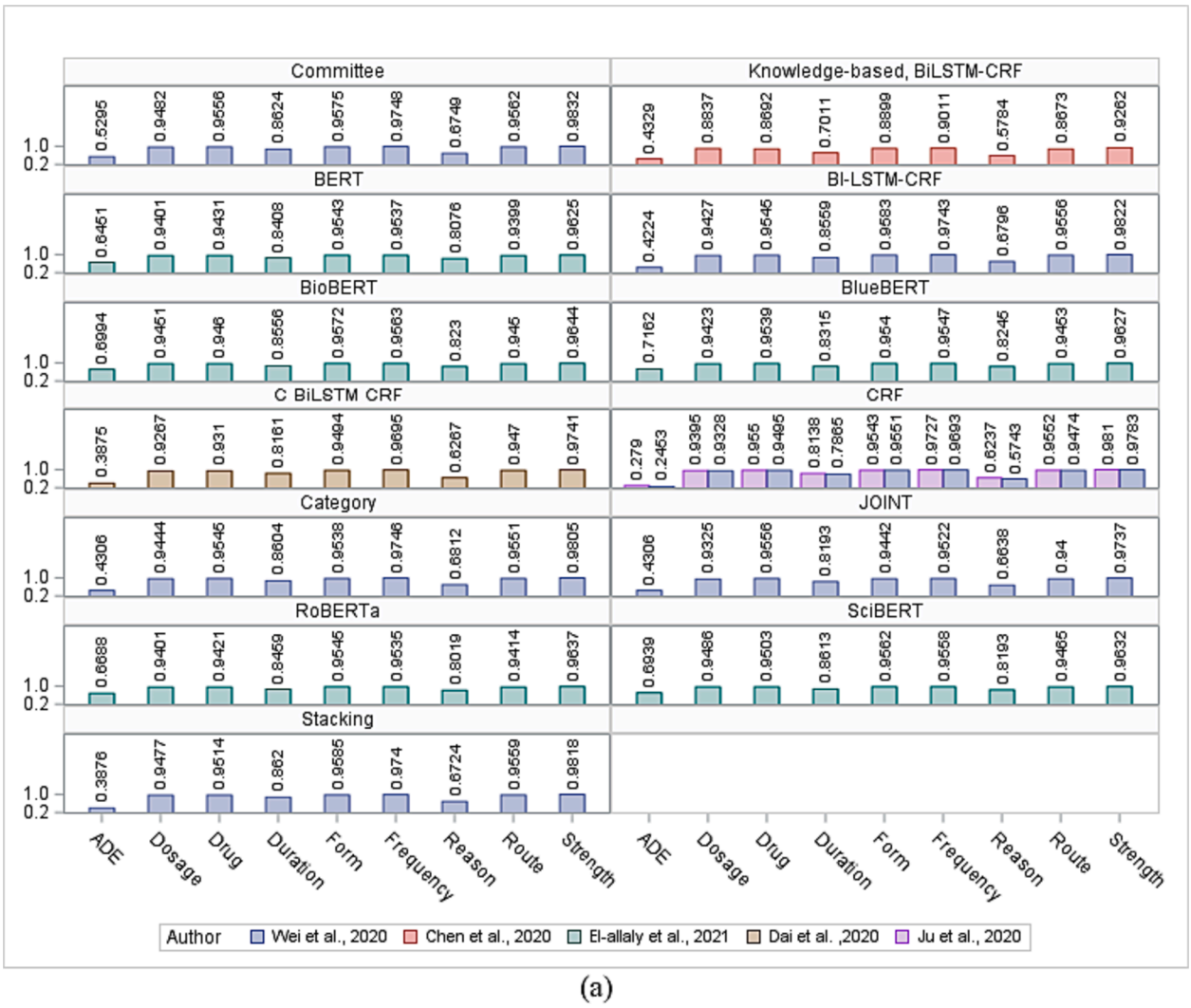


Fig. 1. Overall NER Performance for 2018 n2c2 Shared Task on ADEs and Medication Extraction.



(a)

Fig. 2. Overall NER Performance in 2018 MADE 1.0 Challenge.

was an increase in the F1-score to 0.93 with the integration of CNN and RNN developed by Wei et al. [13]. The performances of ML-based models-SVM, RF, and GB proposed by Yang et al. [14], and SVM developed by Wei et al. [13] were comparable or even outperformed the DL-based models derived from Bi-LSTM and BERT.

Fig. 4 illustrates the overall RC performance in the 2018 MADE 1.0 Challenge. RF proposed by Chapman et al. [19] was equivalent or even better than most of the DL-based models explored. However, the performance of the SVM and RFs model developed by Yang et al. [20] was much inferior to other models.

Fig. A.4 and Fig. A.5 show the F1-score for each relation type drawn from the models employed in the 2018 n2c2 and 2018 MADE 1.0 Challenge, respectively. Similar to the NER task, the extractions of "Duration-Drug", "Reason-Drug", and "ADE-Drug" were difficult for most of the teams. However, Fig. A.4 shows that BERT and BERT-related models (BERT, BioBERT etc.) improved the performance in these relation extractions. For the training set in the 2018 n2c2 shared task on ADE and medication extraction challenge, approximately 30 % of "ADE" and "Reason" were not in the same sentences as the related drugs [8,13]. Some models were trained on a single sentence without the context information in the NER task [13,14,21]. Similarly, the models trained on a single sentence in the RC task were not able to classify relations across several sentences [13]. This issue can be handled by utilizing an attention-based neural network architecture that can exploit more context information [10]. Second, the co-location information was insufficient to detect the "Drug-ADE" and "Drug-Reason" relations correctly when several drugs were located in the same sentence [4].

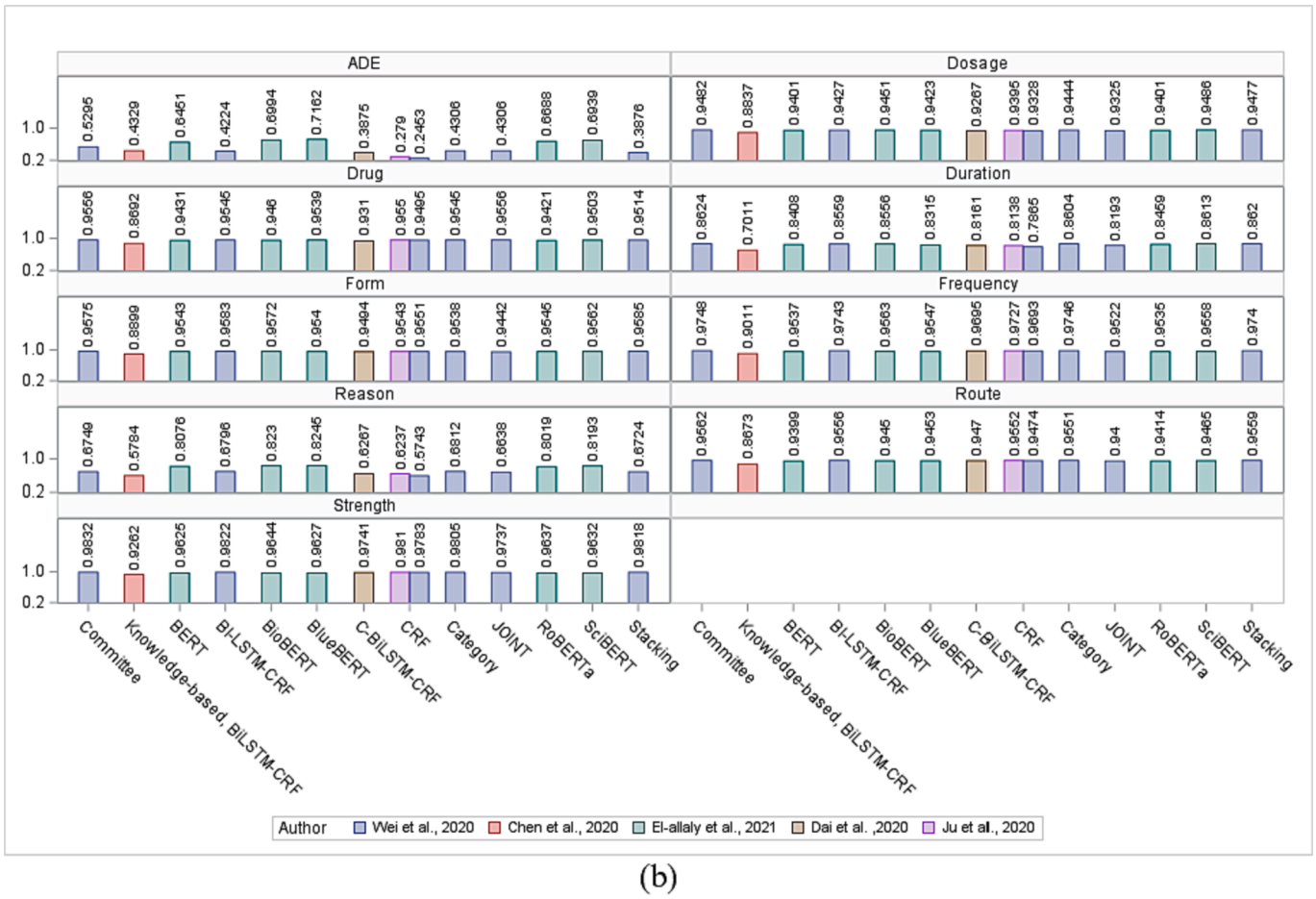
3.3. End-to-End system

Table A.1 shows the end-to-end systems proposed by each team and their corresponding performance. The BERT model developed by Mahendran and McInnes [4] achieved the highest F-1 score of 0.94.

4. Discussion

Through studying the methods published by the different research teams and analyzing their submitted results, the models with state-of-the-art performance, as well as the current state of the literature reviewed, significant gaps or flaws in existing knowledge, future areas for study, and associations between our research and existing knowledge were identified.

We provided the summary of advantages, limitations, and prospective avenues for further research for the methodologies under consideration in Table A.4. The methodologies considered for NER tasks, including CRF, BiLSTM-CRF, CNN-BiLSTM-CRF, LSTM-BiLSTM-CRF, CNN, RNN, LSTM-CRF, BERT, BioBERT, SciBERT, RoBERTa, and BlueBERT, each offer distinct strengths and weaknesses. CRF, while efficient



(b)

Fig. 2. (*continued*).

and capable of capturing complex patterns, requires manual feature engineering [22–24]. Deep learning models like BiLSTM-CRF and its variants excel in learning representations but demand substantial annotated data and computational resources [25,26]. Transformer-based models such as BERT family variants have state-of-the-art performance due to their ability to capture long-range dependencies, yet they are computationally intensive and require careful fine-tuning [27,28]. In terms of RC tasks, SVM, RF, and GB provide interpretability but may struggle with complex patterns [29]. Deep learning models like BiLSTM and CNN can capture intricate relationships but require large amounts of data and computational power [29,30]. Transformer-based models for RC, including BERT variants and CapNet, offer high performance but require optimization for efficiency and domain adaptation [31]. Future research should focus on optimizing these methods for performance, efficiency, and generalization to enhance their utility in NER and RC tasks.

4.1. Named-entity recognition task

For the NER task, it was found that SciBERT and BlueBERT performed best in the NER task for the 2018 n2c2. LSTM-based models outperformed others in the NER task for the 2018 MADE 1.0 Challenge, achieving F1 scores comparable to those of BERT-based models. The performances of the NER task for the included studies in the 2018 MADE 1.0 Challenge were generally lower than that in the 2018 n2c2 shared task. DL-based models typically performed better than ML-based models. The ensemble was considered by Dai et al. [10], Ju et al. [15], and Wei et al. [13]. External resources were adopted and attempted by a few research groups but did not enhance performance significantly. NER performance was high for most entities in both challenges except for “ADE”, “Duration”, “Reason”, and “Indication”. The insufficient number of instances for training, imbalanced dataset, misclassified entity types, and mismatched boundaries (only part of the entities is correctly extracted) accounted for most of the errors. Possible improvements can be made by using a larger sample size and additional feature engineering.

4.1.1. The effect of ensembles

The results reported by Wei et al. [13] showed that the ensembles did make a small difference to the final F1-score compared to the models without ensembles. The ensemble was used mainly to deal with nested entities (entities embedded in other entities) [10,13,15]. For example, Dai et al. [10] used ensembles to deal with the overlapping annotations through training independent models on subsets of nonoverlapping entities and then incorporating the results by voting. In their study, the ensemble structure exploited more context information to boost the performance in drug-related entity recognition [10]. Both ensemble systems (CRF, C-BiLSTM-CRF, and L-BiLSTM-CRF models vs C- and L-BiLSTM-CRF models) by Dai et al. [10] achieved the best F-1 score in all entities except “ADEs”, “Durations”, and “Reasons”, in addition to better precision in “Route”, “Reason”, “Form”, and “ADE” entity and high recall in “Drug”, “Dosage”, and “Strength” entity. Ju et al. [15] concluded that the application of an ensemble to DL-based models reduced the number of errors, such as the category error (entities that have correct lenient spans but incorrect categories) and span error (entities have both incorrect categories and spans), which indicated the effect of ensembles. However, the precisions of “Reason” and “ADE” were greatly improved at the cost of a reduction in recalls with the

Table 3
Summary of included articles for RC model in 2018 n2c2 Shared Task on ADEs and Medication Extraction.

Articles	RC model		Features
Chen et al. [8]	DL-based	Attention-based Bidirectional long-short term memory (Att-BiLSTM)	NA
		BiLSTM	
		CNN	
Mahendran and McInnes [4]	DL-based	CNN BERT, BioBERT, Clinical BERT	NA
Wei et al. [13]	ML-based	SVM	Entity feature, context feature, position feature
	DL-based	CNN-RNN	Word embeddings, PoS embeddings, and position embeddings
		BI-LSTM-CRF (Joint learning)	Predicted drugs from the committee method etc.
Yang et al. [14]	ML-based	SVM RF GB	1) Local context information of entities including cased words inside each entity, unigrams of each entity, words inside the entity, 2) the distance between two entities at the token level, 3) unigrams, bigrams, and trigrams before and after each entity, 4) semantic information
El-allaly et al. [18]	DL-based	BERT, BioBERT, SciBERT, RoBERTa, BlueBERT	NA

Table 4
Summary of included articles for RC model in 2018 MADE 1.0 challenge.

Articles	RC model		Features
Chapman et al. [19]	ML-based	RF	Candidate entities, other entities between candidates, surface features
Dandala et al. [21]	DL-based	Att-BiLSTM	Tokens, entity types, positional indicators around the source and target concepts
F. Li et al. [5]			Words, PoS tag, position embeddings
Yang et al. [20]	ML-based	SVM RFs	1. Local context information 2. Distance between two clinical concepts in character and word level 3. Unigram, bigram, and trigram before and after each entity 4. Semantic feature
F. Li and Yu [17]	DL-based	MLP	Sequence-level feature (e.g., word embeddings), instance-level feature (e.g., entity type)
		CapNet	NA

ensembles. The reason lies in the strict threshold set for the two entities where more than half of the votes (optimal count) were required [10]. For instance, all involved models predicted labels and assigned them to tokens [10]. The algorithm tested whether the assigned entity label received more votes than the optimal count [10]. The label with the highest votes will be assigned if more than one label exceeds the count threshold [10]. Second, as the DL-based model is capable of learning abstract and high-level features from data, the ensemble of the DL-based model and CRF showed limited advantages [15]. The factors mentioned above explained why the ensemble could improve the F-1 score for the NER task, but not by much.

4.1.2. The effect of the knowledge-based feature

Early ADE studies commonly employed knowledge-based models, prioritizing precision via lookup and matching, albeit at the expense of recall. As ML and DL were developed, simply applying knowledge-based models to deal with ADE extraction is inferior. This is demonstrated by the NER model proposed by Chen et al. [8], which utilized knowledge-based models to recognize the majority of the entities, resulting in a low F1-score of 0.85. The knowledge base (KB) is currently frequently used as a feature for ML or DL-based models rather than as a whole model to identify entities.

According to Yang et al. [14], the knowledge-based feature added into the recurrent convolutional neural network (RCNN) model gave rise to a 0.01 improvement in F-1 score (Fig. 1) compared to the RCNN model alone. With the integration of the knowledge-based feature, both LSTM-CRFs and RCNN reached a high precision but low recall, which demonstrated that medical knowledge facilitates the precision of ADE and medication detection [14]. A possible reason for the low recall is that the KB employed does not have as wide a coverage as the Unified Medical Language System (UMLS) for lab tests, treatment, and disease [14]. One reason for the limited improvement in F1-score after the knowledge-based feature was added was that F1-score for "ADE" and "Reason" were extremely low (less than 0.7), which reduced the overall F-1 score. Unlike other entities, "ADE" and "Reason" are more context-dependent. These cannot be improved as easily as other entities by just adding the knowledge-based feature. Another reason lies in that the DL-based models have strengths in learning abstract and high-level features from data automatically, which reduces the dependence on external knowledge resources [15].

4.1.3. Limitations and improvement

A common issue was that entities like "ADE" and "Reason" were very difficult to extract. Wei et al. [13] found most of the samples in the clinical text were negative and would not be annotated with "ADE" or "Reason", which accounted for the low performance in extraction. It is estimated that only 2 %-8% of medical entities in the training set were classified as "ADEs" or "Reasons", which made it challenging for models to learn a general pattern to identify them [13,14,20,21]. Wei et al. [13] and Dandala et al. [21] demonstrated the feasibility of implementing the joint method to handle rare entities.

Additionally, some entities were labeled both "ADE" and "Reason". Some methods proposed can only predict one label for one mention and therefore missed some entities [10,13,15]. Drug entities commonly overlap with "ADE" and "reason" [10]. For example, [[narcotic]$_{\mathbf{Drug}}$ induced respiratory distress]$_{\mathbf{ADE/reason}}$ [10]. To deal with the overlapping entity spans (the boundary of the entities, as some entities may consist of multiple tokens), Dai et al. [10] proposed a cascading structure in which classifiers were trained sequentially in the cascade, and the output of one stage in the cascade exerted an influence on the training samples given to the next. Yang et al. [14] proposed individual models for each of the mentions- "Drug", "ADE", and "Reason" to overcome the overlapping issues and only keep the specific annotations when training its model. The introduction of an external KB such as Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) improved the performance of "ADE" entity extraction [10,21]. The joint model with external resources can enhance the performance of extracting "Indication" and "ADE" by 2 % and 4 % individually compared with the joint model without external resources [21]. Additionally, BiLSTM-CRF modeled generic entity types which were labeled by substituting SSLIF for "ADE" or "indication" and then applied the predictions from the RC task to deduce the correct entity type [21].

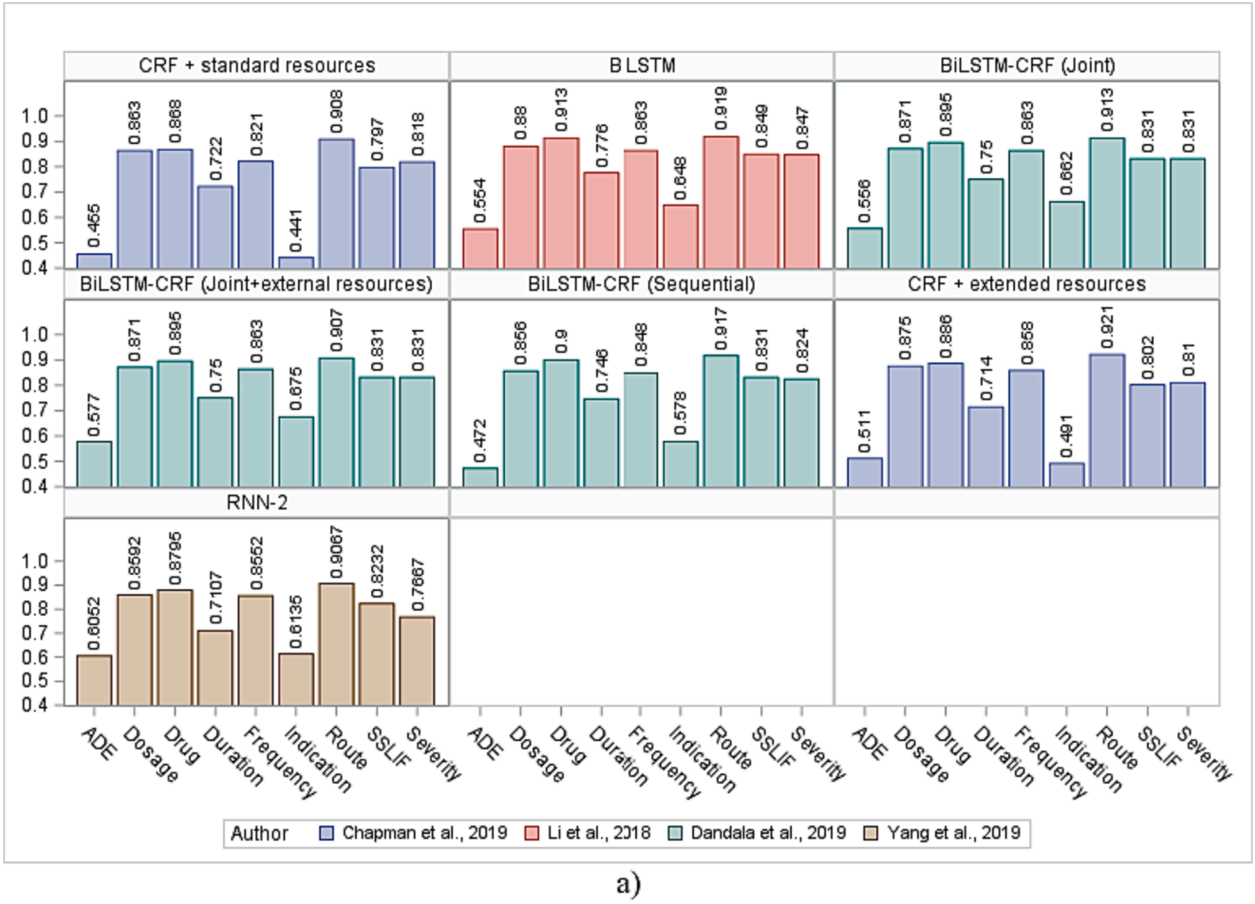


Fig. 3. Overall RC Performance for 2018 n2c2 Shared Task on ADEs and Medication Extraction.

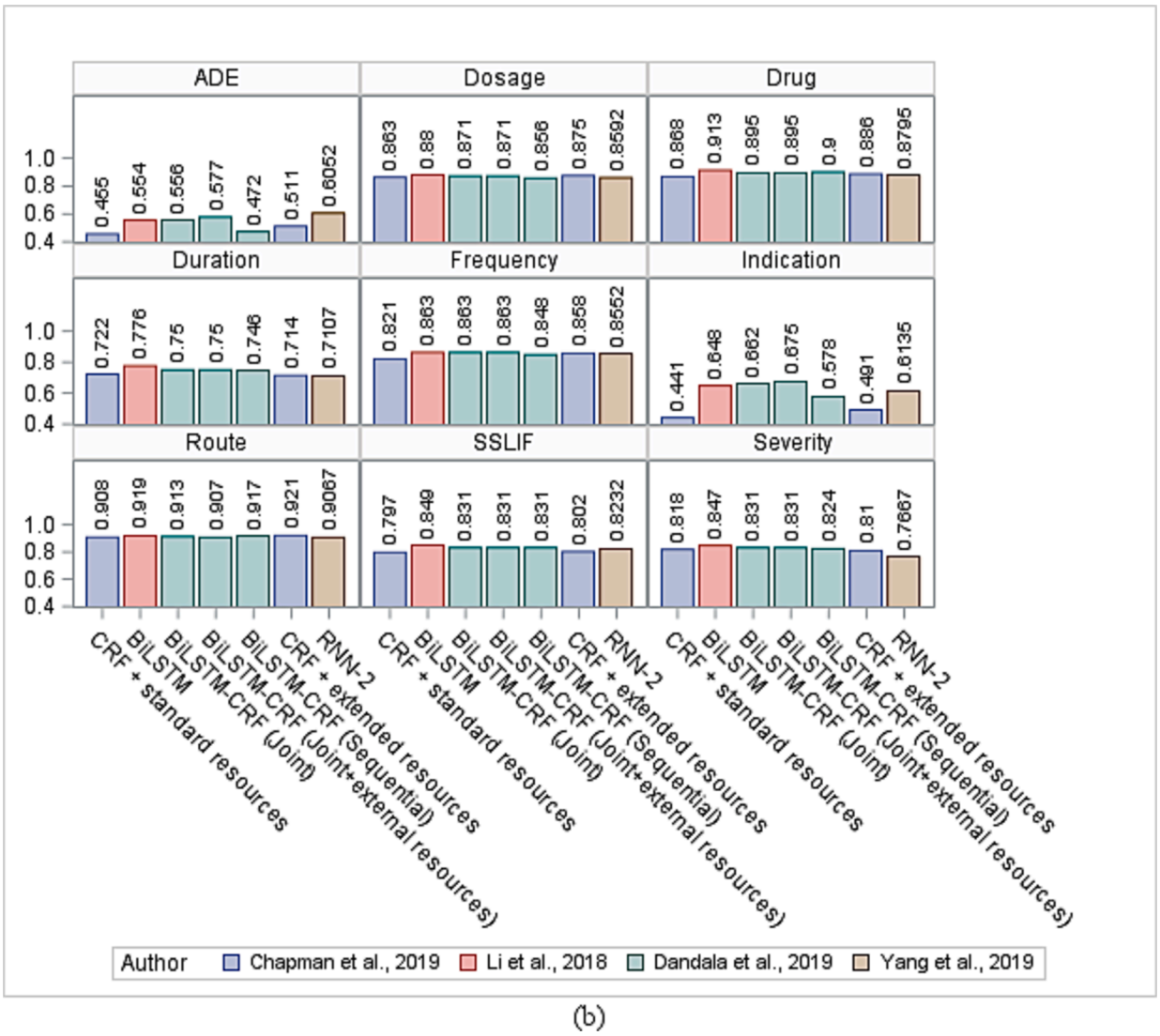
4.2. RC task

For the RC task, GB outperformed other models with an F1-score of 0.96 for the 2018 n2c2 shared task. MLP and RF performed best in terms of F1 in the 2018 MADE 1.0 Challenge. Similar to the NER task, the performance for each model in the 2018 MADE 1.0 Challenge was not as high as that in the 2018 n2c2. ML models in a few included studies produced a comparable performance to DL models. From the study proposed by Wei et al. [13], post-processing had an effect on the performance of the RC task, which improved the F1-score by approximately 2 % for each model studied. Similar to the performance on each entity type in the NER task, the performances of the relation types "ADE-Drug" and "Reason-Drug" were lower than other relation types for most of the included studies. Some ADE/reason-drug pairs were not located in the same sentence. In some sentences, there was more than one relation pair and this interfered with retrieving the relations. These factors combined made the relation task challenging. However, it is encouraging to find the BERT family produced state-of-the-art performance in ADE/reason-drug relations, which was comparable to other relation types. The BERT model is a new beginning for RC tasks. Future studies should be focused on models with postprocessing and larger context windows.

4.3. End-to-end task

For the end-to-end task, the performance is based on the NER task and the RC task. The BERT model achieved the best performance in the included studies. As the efficacy of the RC task is based on the performance of the NER task, it is vital to boost the accuracy to extract each entity. For future improvement, the emphasis of the end-to-end task should be placed on each entity extraction in the NER task, especially "ADE" and "Reason".

As it can be seen from Tables 1 — 4, the inclusion of embeddings as the features was preferred among many research teams in both challenges. Word embedding, character embedding, and position embedding were three common types of embeddings used in ADE extraction. Wei et al. [13] trained the word embedding on word2vec$_{MIMIC}$ as did most of the top teams. In contrast, Dai et al. [10] looked into the effect of different word embeddings and concluded that Global Vectors (GloVe) as word embedding had a strong performance. Moreover, they found that the optimal choice of embeddings will have a greater impact on the performance of the sequence labeling task than fine-tuned hyperparameters [10,32]. Dai et al. [10] found that the embeddings trained from the biomedical domain such as the MIMIC-III corpus did not perform better than GloVe, which was trained on the general domain. A possible explanation is that the domain-specific embeddings can increase the efficacy in identifying domain-specific mentions like



(b)

Fig. 3. (*continued*).

"Medication", but the narrative entities like "Reason" and "ADE" favor general domain embeddings [10,33]. F. Li et al. [5] found that the model with pretrained embeddings performed better than the fine-tuned embeddings due to the heterogeneity of the training data and the overfitting issue.

4.4. The current state of research

The literature under review includes numerous high-quality studies that exhibit superior performance in ADE extraction, while also providing valuable scientific insights in both NER and RC tasks. The models proposed in the included studies can serve as baselines for future studies or state-of-the-art methods for applications. Additionally, there have been comprehensive and quantitative studies investigating the effect of embeddings, post-processing, the inclusion of external resources, joint learning, and ensembles on final results. The model with GloVe word embedding had a strong performance. External resources and ensembles did not greatly improve performance. However, the introduction of postprocessing made a great difference.

Several flaws or gaps in existing knowledge were identified: First, although multi-task learning improved the performance in ADE extraction, it is largely dependent on the methods and data [5]. Second, although attempts were made to include external resources to improve the NER performance, how the external resources integrated into the DL framework had not been fully studied [13]. Third, the current sentence detection algorithm may not be perfect, and the division of documents into short sentences is unhelpful [19]. The sentence division is related to the context available to one sentence, which will greatly affect the performance in RC task [19]. Additionally, the longitudinality in the dataset was not considered [17]. Last but not least, limited annotated

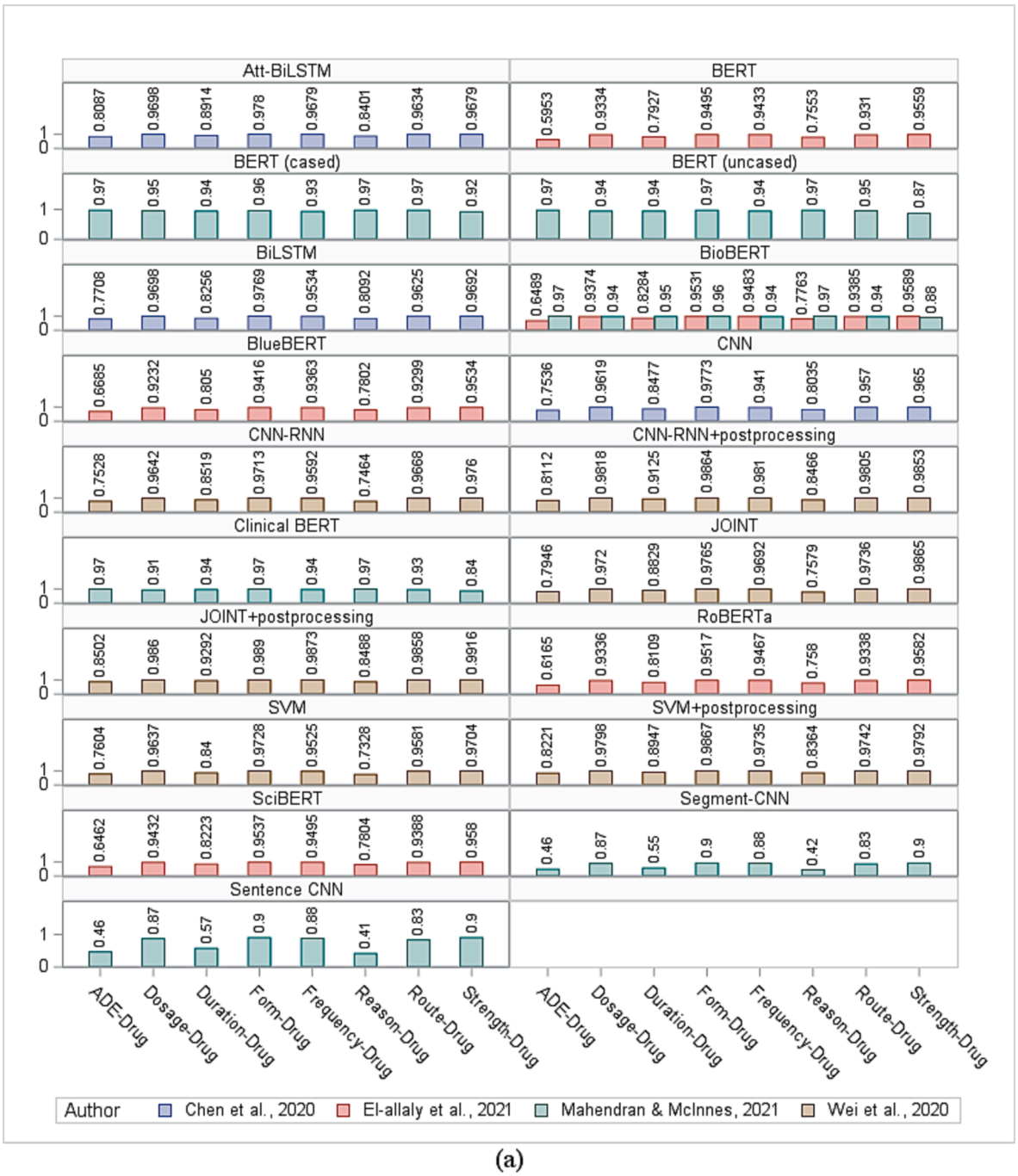


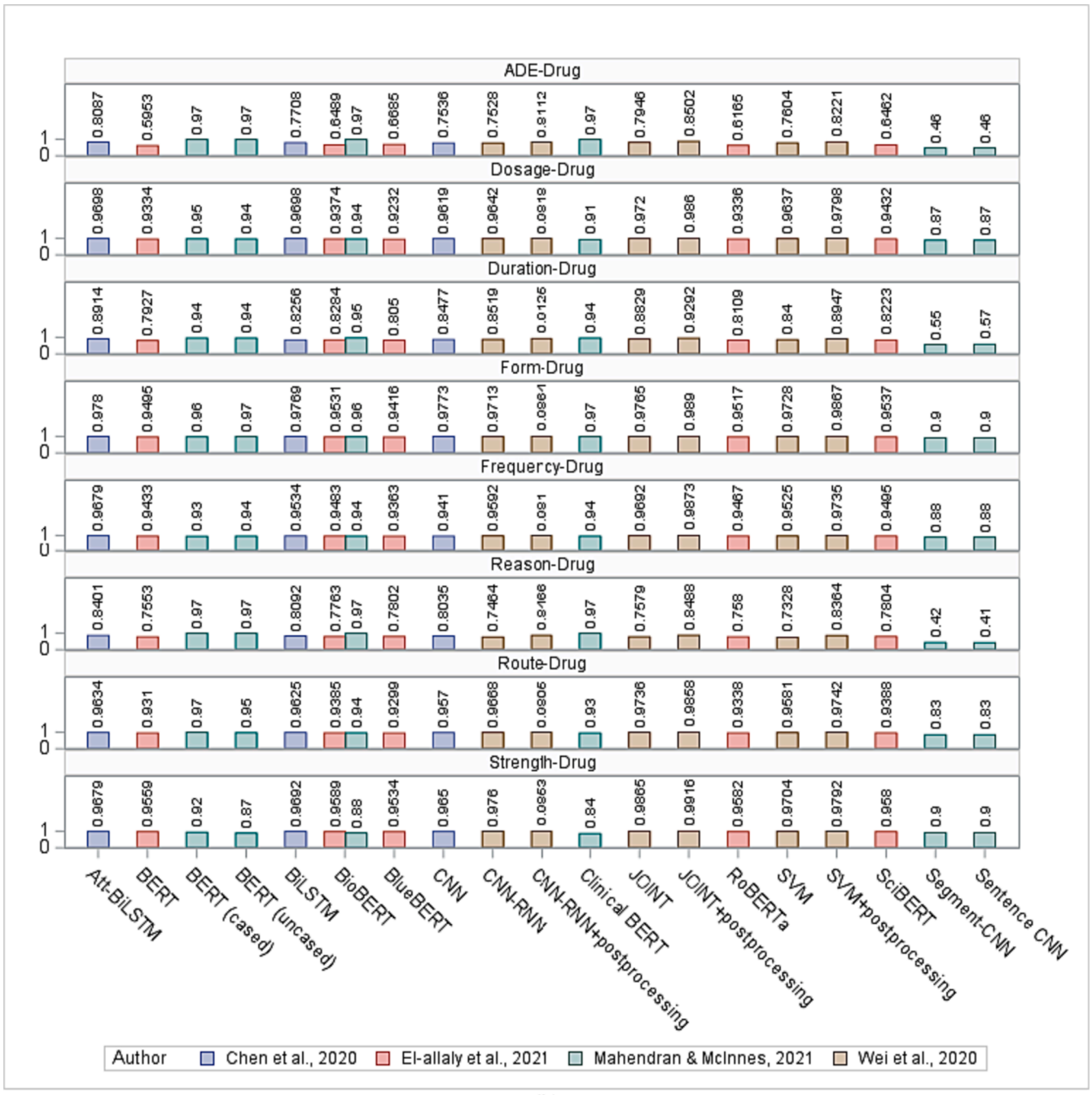
Fig. 4. Overall RC Performance in 2018 MADE 1.0 Challenge.

corpora and disease were included in the study [17].

Furthermore, while the literature predominantly focuses on machine learning and deep learning approaches, there is emerging interest in rule/ontology/knowledge-based methods [34,35] and semi-supervised learning methods [36,37]. Rule-based methods offer interpretability and explicit modeling of domain knowledge, but they can be limited by the complexity of rules and the need for manual intervention [35]. Semi-supervised learning, on the other hand, leverages both labeled and unlabeled data, which can be beneficial when labeled data is scarce or costly to obtain [37].

4.5. Future study

ADE is a critical and multifaceted aspect of biomedical research



(b)

Fig. 4. (*continued*).

[38,39]. Large language models (LLMs) have emerged as transformative tools in the domain of ADE extraction, promising unparalleled advancements in performance [40,41]. For example, Leas et al. found that ChatGPT accurately replicated human annotations for adverse event reports, achieving high agreement for both general (97.7 %; Kappa = 0.95) and serious adverse events (98.0 %; Kappa = 0.95) [42]. Li et al. applied and fine-tuned various LLMs, including GPT-2, GPT-3 variants, GPT-4, and Llama 2 to evaluate the effectiveness of LLMs in extracting AEs related to the Influenza vaccine [40]. In the reviewed articles, the groundbreaking success of the BERT model, exemplified by its state-of-the-art F1-score of 0.94 in end-to-end tasks, positions it as a pivotal focus for future ADE extraction studies. However, only one drug-entity pair for the entire sentence was considered for the BERT model in the reviewed study, which did not represent all the actual cases [4]. Multiple drug-

entity pairs may exist in a sentence, which challenged the RC task in ADE extraction. Therefore, practical methods to represent multiple drug-entity pairs within a sentence should be constructed. In the work proposed by Mahendran and McInnes [4], an independent binary model for each class in the dataset was developed. In the future, a model for multiclass classification should be implemented and the effect of ensembles incorporated into multiple individual models should be studied. It has become a trend to employ character embeddings, word embeddings, and position embeddings as features. The coverage of selected embeddings directly affected the performance of ADE extraction [10]. In the study proposed by Dai et al. [10], the model with GloVe embedding greatly outperformed that with word2vec$_{MIMIC}$ and BioWordVec [10]. It remains to be studied which pre-trained embeddings will work best with BERT. A small dataset was used and it was challenging to extract ADE and reasons with limited training samples [13,14,17,20,21]. In the future study, a larger dataset with diverse medical domains will be included. As shown in Fig. 1 and Fig. 2, the integration of external resources into BiLSTM-CRF, LSTM-CRFs, and CRF did not enhance the performance in NER. It still needs to be explored whether the introduction of external resources works for BERT.

5. Conclusion

In conclusion, this study highlights the current state of ADE extraction research, emphasizing the potential of BERT models, effective embeddings, and the need for advanced strategies in entity extraction. The performance variations in NER and RC tasks suggest the importance of tailoring methods to specific challenges. Future research should focus on optimizing the use of BERT models, tackling complex scenarios, leveraging pre-trained embeddings, and incorporating larger, more diverse datasets to further advance ADE extraction, ultimately enhancing patient safety and healthcare decision-making.

CRediT authorship contribution statement

Yiming Li: Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Wei Tao:** Visualization. **Zehan Li:** Writing – review & editing, Visualization. **Zenan Sun:** Data curation. **Fang Li:** Writing – review & editing. **Susan Fenton:** Writing – review & editing. **Hua Xu:** Writing – review & editing. **Cui Tao:** Writing – review & editing, Validation, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This article was partially supported by the National Institute of Allergy And Infectious Diseases of the National Institutes of Health, United States under Award Numbers R01AI130460 and U24AI171008. I would like to express my sincere appreciation to Irmgard Willcockson for her invaluable contributions as the editor of this publication.

Appendix A

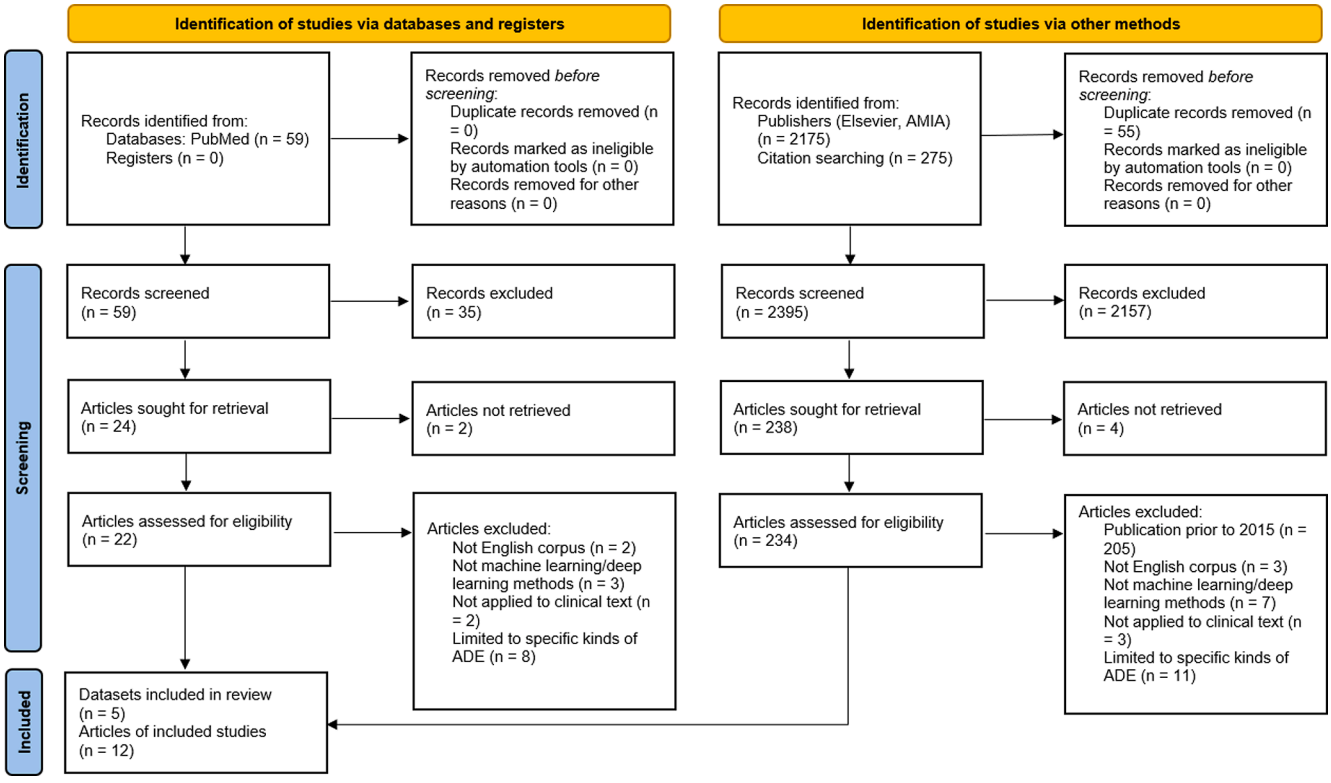
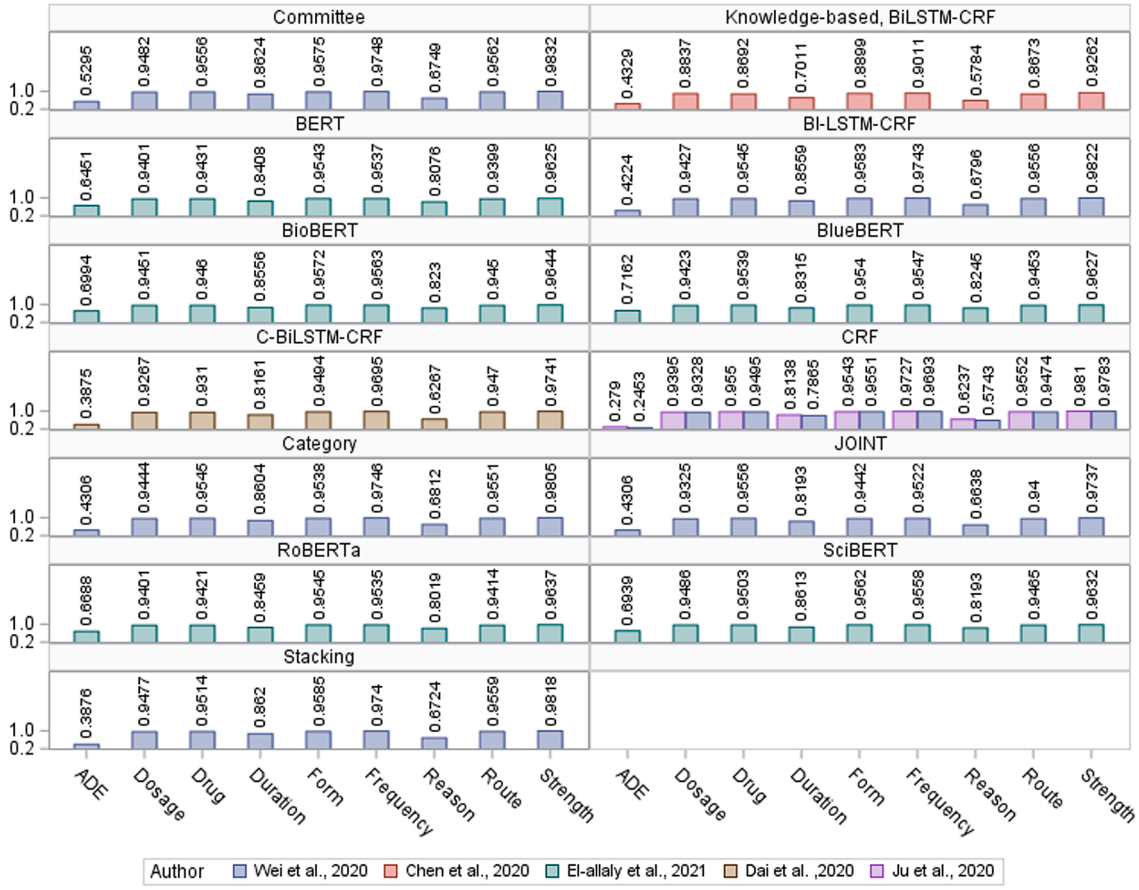
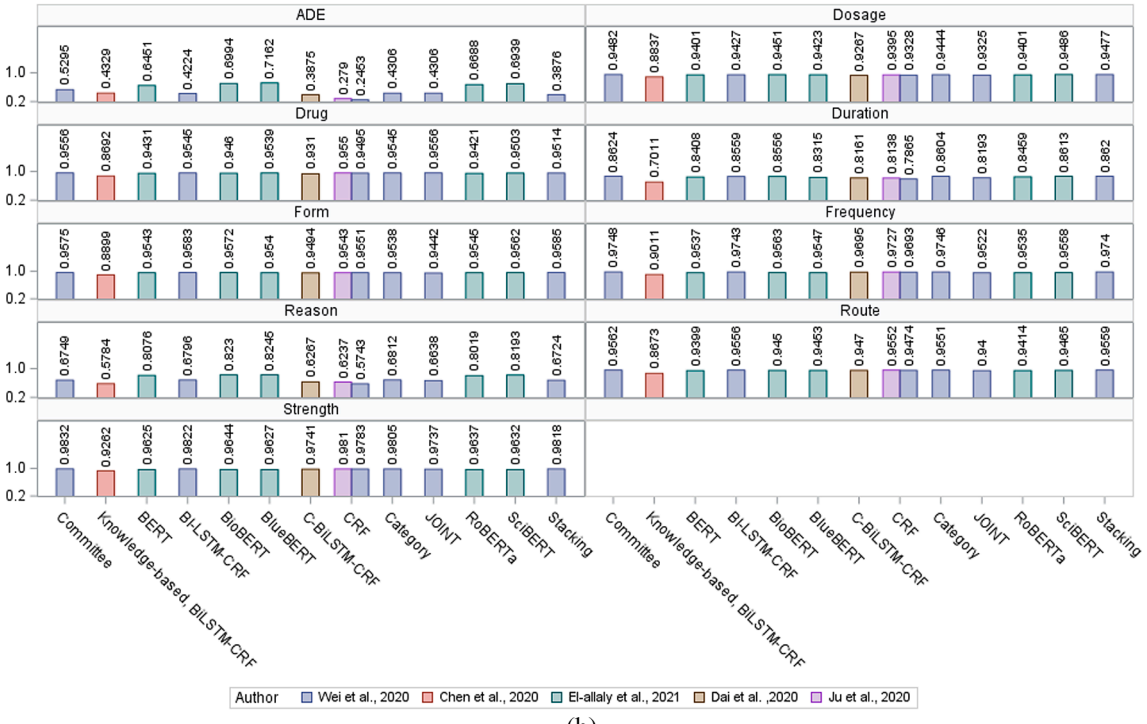


Fig. A1. PRISMA 2020 flow diagram for new systematic reviews which included searches of databases, registers and other sources. Adapted from " The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. " by Page et al. [43].

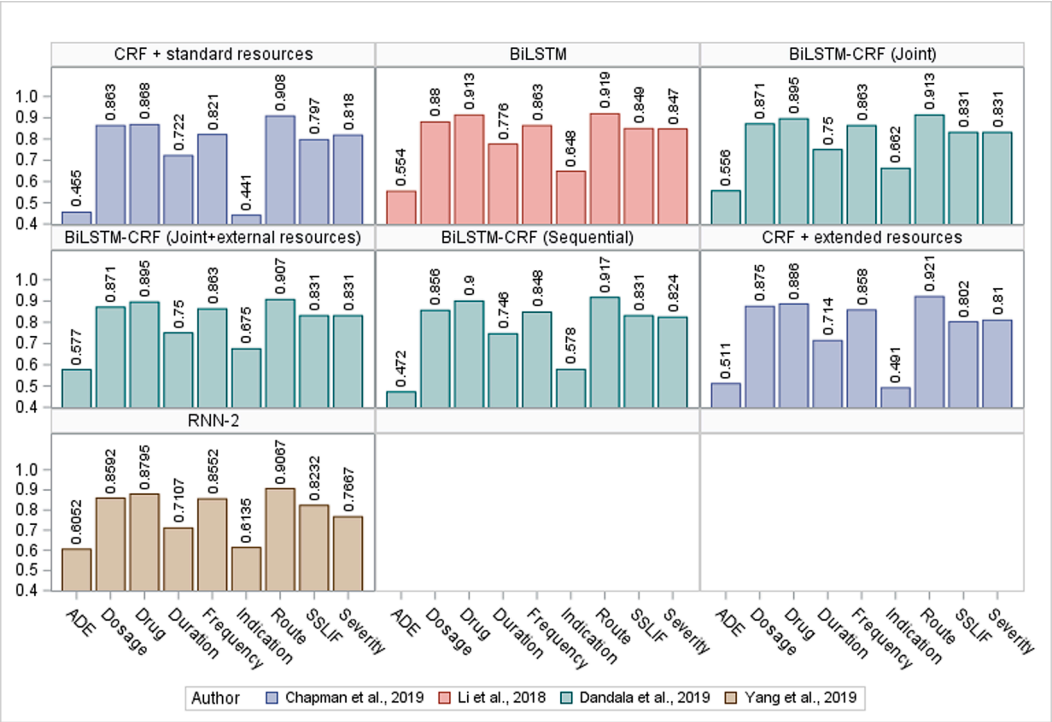


(a)

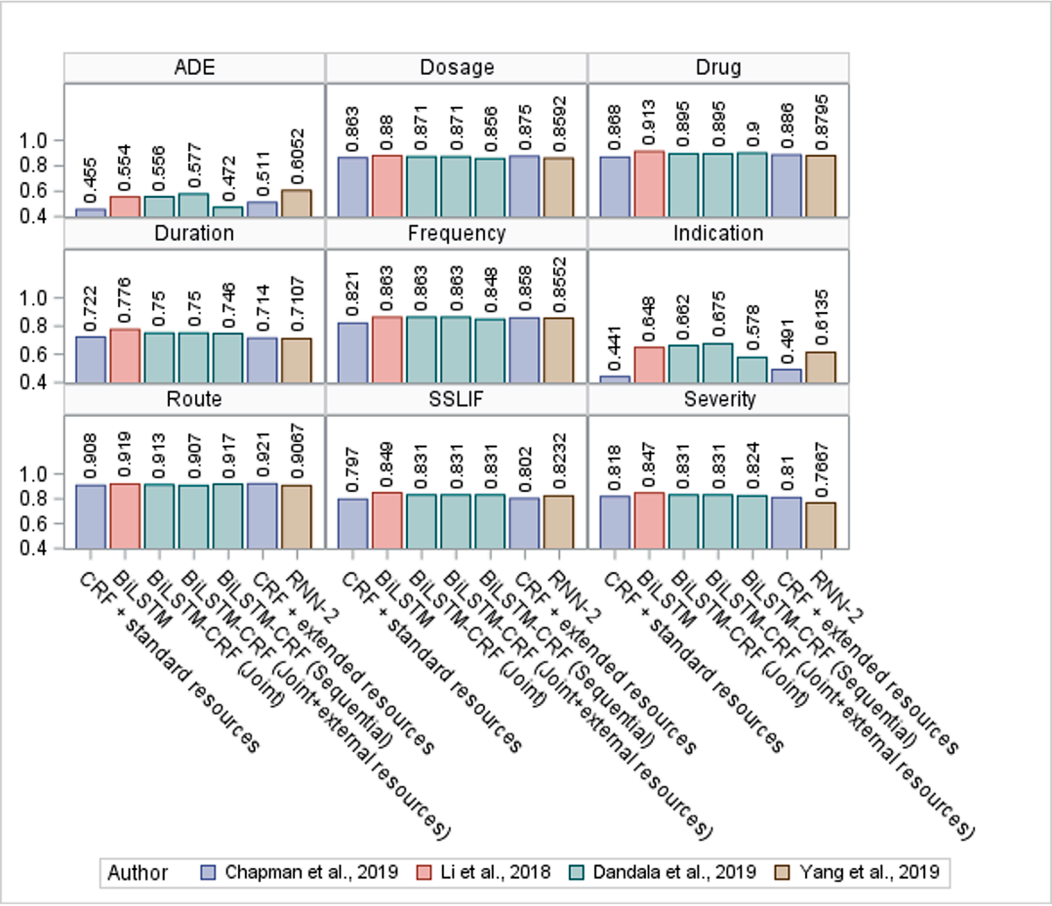


(b)

Fig. A2. a) F1 score for each entity type in 2018 n2c2 Shared Task on ADEs and Medication Extraction (method based) b) F1 score for each entity type in 2018 n2c2 Shared Task on ADEs and Medication Extraction (entity type based)

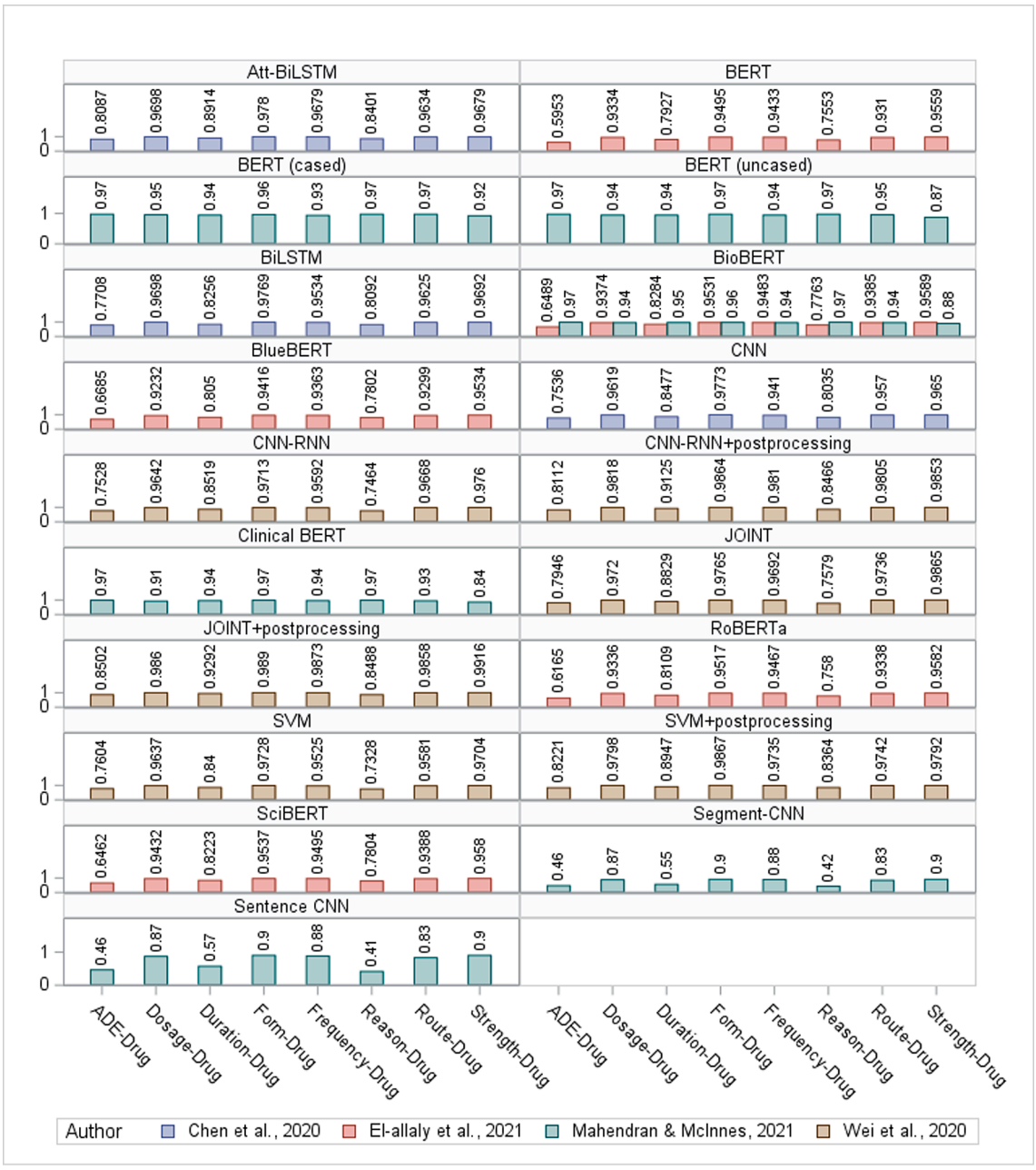


(a)



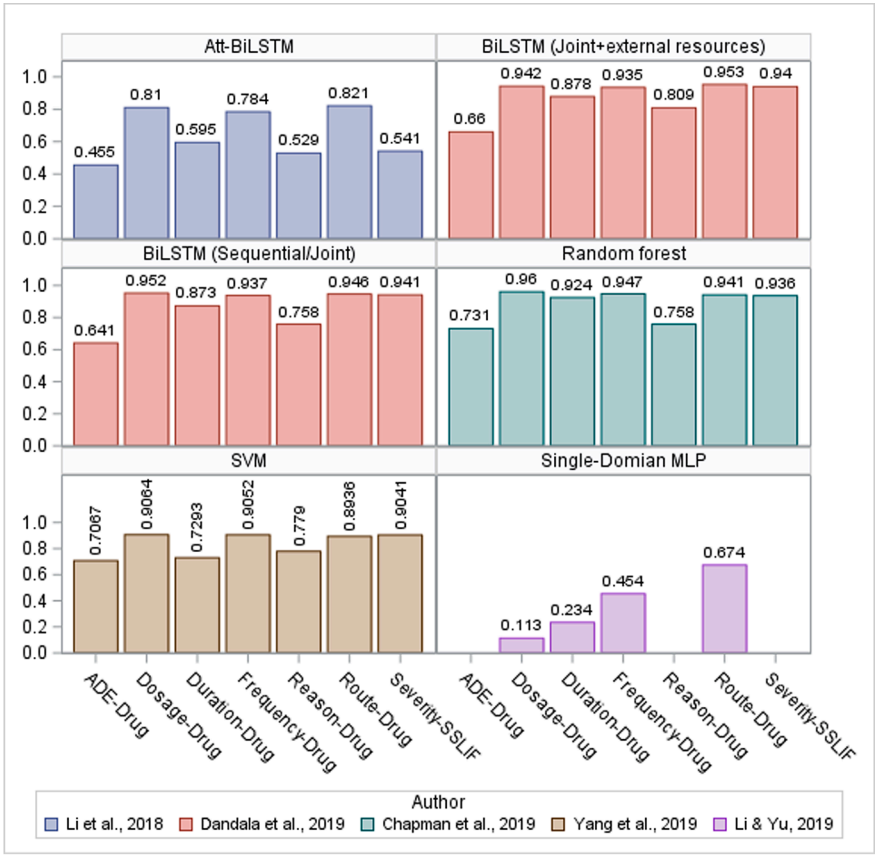
(b)

Fig. A3. a) F1 score for each entity type in 2018 MADE 1.0 Challenge (method based) b) F1 score for each entity type in 2018 MADE 1.0 Challenge (entity type based)

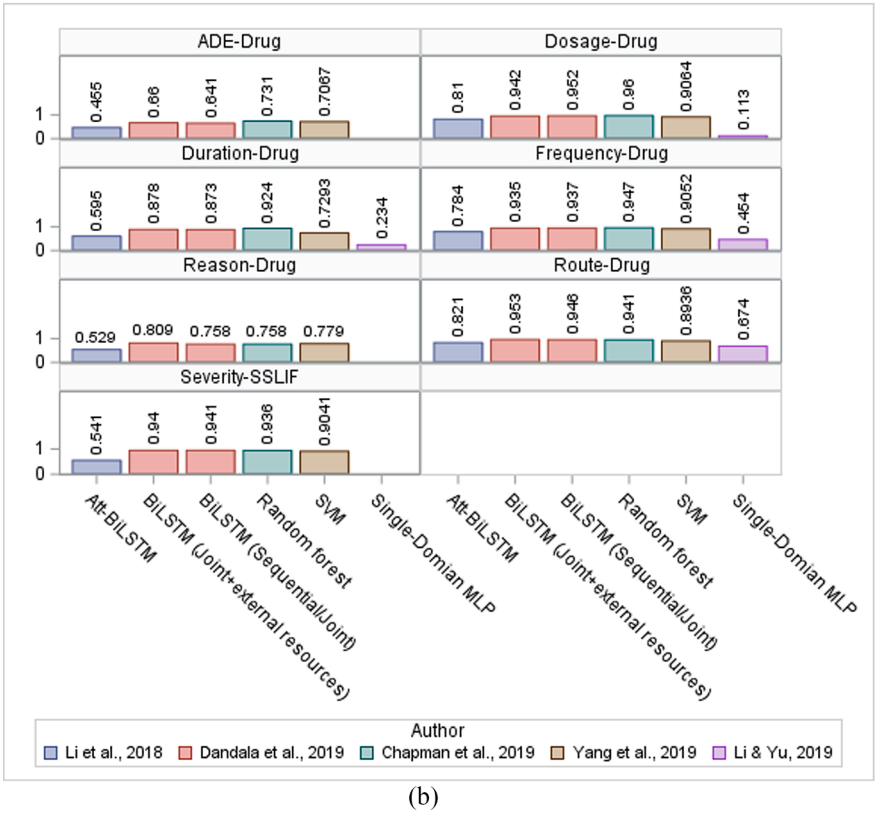


(a)

Fig. A4. a) F1 Score for Each Relation Type in 2018 n2c2 Shared Task on ADEs and Medication Extraction (method based) b) F1 Score for Each Relation Type in 2018 n2c2 Shared Task on ADEs and Medication Extraction (entity type based)

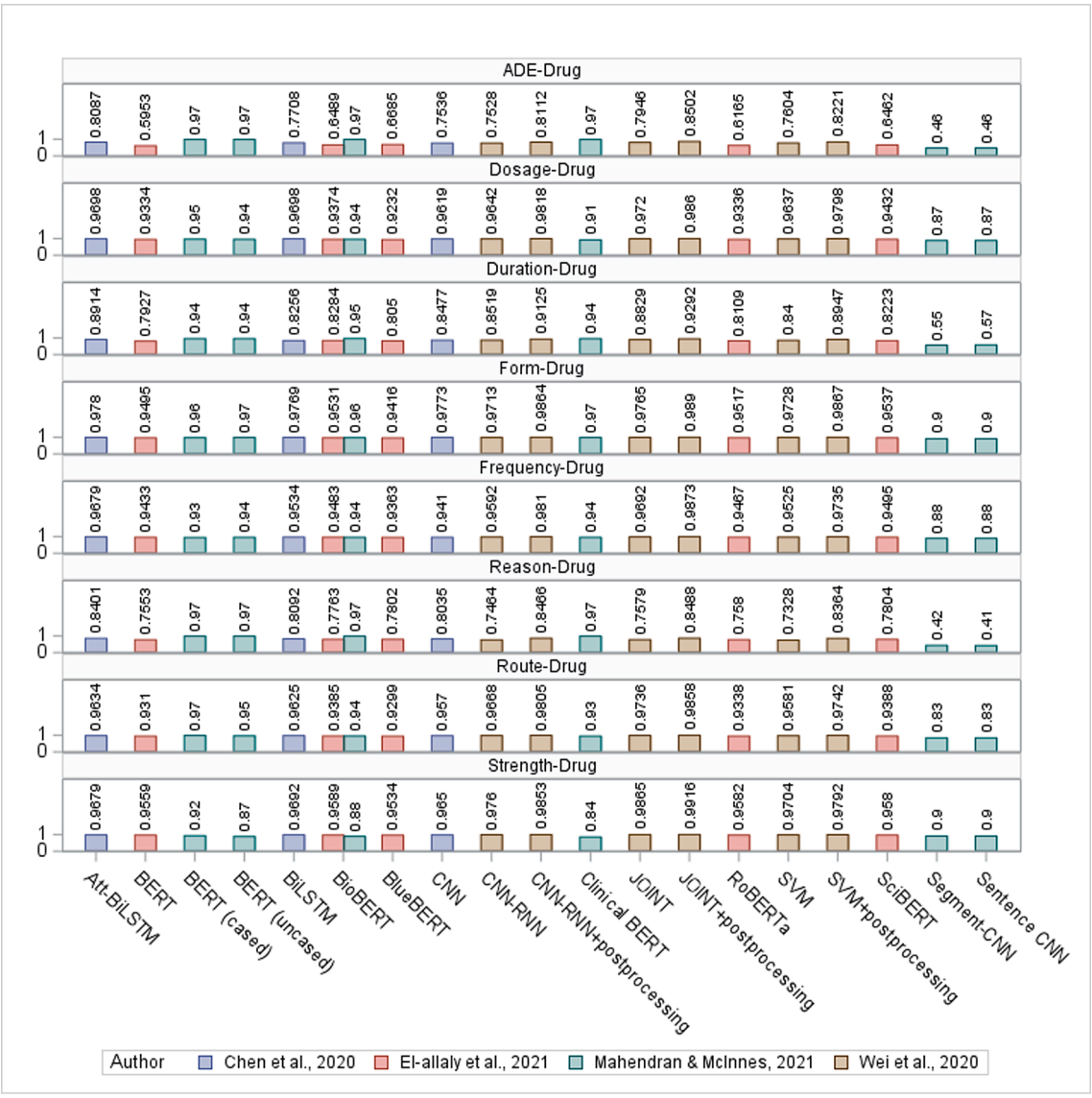


(a)



(b)

Fig. A5. a)F1 Score for Each Relation Type in 2018 MADE 1.0 Challenge (method based) b) F1 Score for Each Relation Type in 2018 MADE 1.0 Challenge (entity type based)



(b)

Fig. A4. (*continued*).

Table A1
Summary of included articles, related datasets, and their best end-to-end performances.

Dataset	Authors	Models used (NER model-RC model)	Precision	Recall	F1
2018 n2c2	Chen et al. [8]	Knowledge-based- Attention-based Bidirectional long-short term memory (Att-BiLSTM)	0.8382	0.7539	0.7938
	Dai et al. [10]	All Bidirectional long-short term memory (BiLSTM)- conditional random field (CRF) models	0.939	0.900	0.919
	Ju et al. [15]	BiLSTM-CRF (NER only)	0.9599	0.8979	0.9278
	Mahendran and McInnes [4]	Bidirectional Encoder Representations from Transformers (BERT)	0.93	0.96	0.94
	Wei et al. [13]	Convolutional neural network (CNN)- recurrent neural network (RNN)	NA	NA	0.8905
	Yang et al. [14]	LSTM-CRFs + gradient boosting (GB)-4	0.9187	0.8593	0.8880
2018 MADE 1.0 Challenge	Chapman et al. [19]	CRF- random forest (RF)	0.721	0.534	0.612
	Dandala et al. [21]	BiLSTM-CRF-BiLSTM (Joint + external resources)	0.696	0.632	0.662
	F. Li et al. [5]	BiLSTM-CRF- BiLSTM-Attention	NA	NA	0.667
	Yang et al. [20]	RNN-2-support vector machine (SVM)	0.5758	0.6542	0.6125
2018 MADE 1.0 Challenge Cardio corpus	F. Li and Yu [17]	Capsule network (CapNet)	NA	NA	0.827[a]

(*continued on next page*)

Table A1 (*continued*)

Dataset	Authors	Models used (NER model-RC model)	Precision	Recall	F1
i2b2 2009 Medication Challenge					
TAC 2017 2018 n2c2	El-allaly et al. [18]	SciBERT (pre-trained BERT-based language model for performing scientific tasks)	0.9193[b]	0.9057[b]	0.9125[b]

Note:
[a]Evaluated on 2018 MADE 1.0 challenge.
[b]Evaluated on 2018 n2c2.

Table A2
Statistics of 2018 n2c2 dataset Adapted from " MTTLADE: A multi-task transfer learning-based method for adverse drug events extraction." by El-allaly et al. [18].

Task	Type	Training	Test	Description
NER	*Drug*	*16,225*	*10,575*	*Specifies the name of the drug*
	ADE	*959*	*625*	*Indicates a side effect of a drug*
	Strength	*6691*	*4230*	*Specifies the strength of the drug*
	Dosage	*4221*	*2681*	*Specifies a drug dosage*
	Duration	*592*	*378*	*Indicates the duration of the drug*
	Frequency	*6281*	*4012*	*Specifies the frequency of the prescribed drug*
	Form	*6651*	*4359*	*Indicates a drug form*
	Route	*5476*	*3513*	*Specifies the route of administration for a drug*
	Reason	*3855*	*2545*	*Includes signs, symptoms, and disease for drug*
RC	*ADE-Drug*	*1107*	*733*	*Links an ADE with a drug name*
	Strength-Drug	*6702*	*4244*	*Connects a drug strength with its name*
	Dosage-Drug	*4225*	*2695*	*Relationship between drug dosage and its name*
	Duration-Drug	*643*	*426*	*Links a drug duration with its name*
	Frequency-Drug	*6310*	*4034*	*Connects a drug frequency and its name*
	Form-Drug	*6654*	*4374*	*Relationship between a drug form and its name*
	Route-Drug	*5538*	*3546*	*Links the route for a drug with its name*
	Reason-Drug	*5169*	*3410*	*Connects a reason for a drug with its name*

Table A3
Statistics of 2018 MADE 1.0 challenge dataset Adapted from Dandala et al. [21].

Task	*Type*	*Training*	*Test*	*Description*
NER	SSLIF	34,056	5328	Indicates all signs and symptoms
	Drug	13,507	2395	Specifies the name of the drug
	ADE	1509	431	Indicates a side effect of a drug
	Dosage	4893	801	Specifies a drug dosage
	Duration	765	133	Indicates the duration of the drug
	Frequency	4147	659	Specifies the frequency of the prescribed drug
	Severity	3374	534	Indicates Severity of disease or symptom
	Indication	3168	636	Specifies Affliction that is being treated with a drug
	Route	2278	389	Specifies the route of administration for a drug
RC	ADE-Drug	2082	530	Links an ADE with a drug name
	Dosage-Drug	5177	866	Relationship between drug dosage and its name
	Duration-Drug	906	147	Links a drug duration with its name
	Frequency-Drug	4419	730	Connects a drug frequency and its name
	Severity-SSLIF	3476	559	Relationship between severity and SSLIF
	Route-Drug	2551	455	Links the route for a drug with its name
	Reason-Drug	4554	876	Connects a reason for a drug with its name

Table A4
Summary of the advantages, limitations, and prospective avenues for further research for methodologies under consideration.

Tasks	Method	Advantages	Limitations	Prospective Avenues
NER	CRF	• particularly suitable for sequence labeling tasks like NER; • efficient and can capture complex patterns in data	• requires handcrafted features • struggles with capturing long-range dependencies	• incorporates more sophisticated features or embeddings into CRF models to improve performance
	BiLSTM-CRF, CNN-BiLSTM-CRF, LSTM-BiLSTM-CRF, CNN, RNN, LSTM-CRF	• high performance in NER tasks, especially with large amounts of data • automatically learn representations from data	• require a large amount of annotated data for training • computationally expensive • struggle with capturing rare or unseen entities	• explore techniques for training deep learning models with limited annotated data • develop methods for improving their ability to generalize to unseen entities

(*continued on next page*)

Table A4 (*continued*)

Tasks	Method	Advantages	Limitations	Prospective Avenues
	BERT, BioBERT, SciBERT, RoBERTa, BlueBERT	• state-of-the-art performance in NER tasks • can capture long-range dependencies • have contextual understanding	• require substantial computational resources and memory for training and inference. • Fine-tuning for specific tasks can also be challenging	• optimize the architecture and training process of transformer-based models for NER tasks • explore methods for domain adaptation
RC	SVM, RF, GB	• can handle high-dimensional data well	• struggle with capturing complex patterns in text data • may require careful tuning of hyperparameters	• improve the ability to capture subtle relationships between entities in text • explore ensemble methods for combining predictions
	BiLSTM, CNN, Att-BiLSTM, CNN-RNN, MLP	• can capture complex patterns in text data • promising results in RC tasks	• require large amounts of annotated data • computationally expensive	• develop more efficient deep learning architectures for RC tasks • explore methods for incorporating external knowledge
	BERT, BioBERT, Clinical BERT, SciBERT, RoBERTa, BlueBERT, CapNet	• impressive performance in RC tasks, especially with large amounts of data. • can capture complex relationships between entities	• require substantial computational resources • struggle with rare or unseen relationships • Fine-tuning for specific tasks can also be challenging	• develop more efficient training strategies for transformer-based models for RC tasks • explore methods for incorporating domain knowledge

References

[1] Institute of Medicine. To Err Is Human: Building a Safer Health System. Washington, DC: The National Academies Press; 2000. 10.17226/9728.

[2] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, et al., Graph neural networks: A review of methods and applications, AI Open 1 (2020) 57–81, https://doi.org/10.1016/j.aiopen.2021.01.001.

[3] S. Henry, K. Buchan, M. Filannino, A. Stubbs, O. Uzuner, 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records, J. Am. Med. Inform. Assoc. 27 (2020) 3–12, https://doi.org/10.1093/jamia/ocz166.

[4] D. Mahendran, B.T. McInnes, Extracting Adverse Drug Events from Clinical Notes, AMIA Jt Summits Transl Sci Proc AMIA Jt Summits Transl Sci 2021 (2021) 420–429.

[5] F. Li, W. Liu, H. Yu, Extraction of Information Related to Adverse Drug Events from Electronic Health Record Notes: Design of an End-to-End Model Based on Deep Learning, JMIR Med. Inform. 6 (2018) e12159.

[6] M.A. Makary, M. Daniel, Medical error—the third leading cause of death in the US, BMJ (2016:i2139.), https://doi.org/10.1136/bmj.i2139.

[7] D.W. Bates, N. Spell, D.J. Cullen, E. Burdick, N. Laird, L.A. Petersen, et al., The Costs of Adverse Drug Events in Hospitalized Patients, J. Am. Med. Assoc. 277 (1997) 307–311, https://doi.org/10.1001/jama.1997.03540280045032.

[8] L. Chen, Y. Gu, X. Ji, Z. Sun, H. Li, Y. Gao, et al., Extracting medications and associated adverse drug events using a natural language processing system combining knowledge base and deep learning, J. Am. Med. Inform. Assoc. 27 (2020) 56–64, https://doi.org/10.1093/jamia/ocz141.

[9] D.C. Classen, Adverse Drug Events in Hospitalized PatientsExcess Length of Stay, Extra Costs, and Attributable Mortality, JAMA 277 (1997) 301, https://doi.org/10.1001/jama.1997.03540280039031.

[10] H.-J. Dai, C.-H. Su, C.-S. Wu, Adverse drug event and medication extraction in electronic health records via a cascading architecture with different sequence labeling models and word embeddings, J. Am. Med. Inform. Assoc. 27 (2020) 47–55, https://doi.org/10.1093/jamia/ocz120.

[11] J. Fanikos, J.L. Cina, S. Baroletti, K. Fiumara, L. Matta, S.Z. Goldhaber, Adverse Drug Events in Hospitalized Cardiac Patients, Am. J. Cardiol. 100 (2007) 1465–1469, https://doi.org/10.1016/j.amjcard.2007.06.041.

[12] Y. Luo, W.K. Thompson, T.M. Herr, Z. Zeng, M.A. Berendsen, S.R. Jonnalagadda, et al., Natural Language Processing for EHR-Based Pharmacovigilance: A Structured Review, Drug Saf. 40 (2017) 1075–1089, https://doi.org/10.1007/s40264-017-0558-6.

[13] Q. Wei, Z. Ji, Z. Li, J. Du, J. Wang, J. Xu, et al., A study of deep learning approaches for medication and adverse drug event extraction from clinical text, J. Am. Med. Inform. Assoc. 27 (2020) 13–21, https://doi.org/10.1093/jamia/ocz063.

[14] X. Yang, J. Bian, R. Fang, R.I. Bjarnadottir, W.R. Hogan, Y. Wu, Identifying relations of medications with adverse drug events using recurrent convolutional neural networks and gradient boosting, J. Am. Med. Inform. Assoc. 27 (2020) 65–72, https://doi.org/10.1093/jamia/ocz144.

[15] M. Ju, N.T.H. Nguyen, M. Miwa, S. Ananiadou, An ensemble of neural models for nested adverse drug events and medication extraction with subwords, J. Am. Med. Inform. Assoc. 27 (2020) 22–30, https://doi.org/10.1093/jamia/ocz075.

[16] Fenton S. Final Paper and Poster Assignment Instruction for BMI 7303 Critical Review of Biomedical Informatics Literature Seminar 2021.

[17] F. Li, H. Yu, An investigation of single-domain and multidomain medication and adverse drug event relation extraction from electronic health record notes using advanced deep learning models, J. Am. Med. Inform. Assoc. 26 (2019) 646–654, https://doi.org/10.1093/jamia/ocz018.

[18] El-allaly E, Sarrouti M, En-Nahnahi N, Ouatik El Alaoui S. MTTLADE: A multi-task transfer learning-based method for adverse drug events extraction. Inf Process Manag 2021;58:102473. 10.1016/j.ipm.2020.102473.

[19] A.B. Chapman, K.S. Peterson, P.R. Alba, S.L. DuVall, O.V. Patterson, Detecting Adverse Drug Events with Rapidly Trained Classification Models, Drug Saf. 42 (2019) 147–156, https://doi.org/10.1007/s40264-018-0763-y.

[20] X. Yang, J. Bian, Y. Gong, W.R. Hogan, Y. Wu, MADEx: A System for Detecting Medications, Adverse Drug Events, and Their Relations from Clinical Notes, Drug Saf. 42 (2019) 123–133, https://doi.org/10.1007/s40264-018-0761-0.

[21] B. Dandala, V. Joopudi, M. Devarakonda, Adverse Drug Events Detection in Clinical Notes by Jointly Modeling Entities and Relations Using Neural Networks, Drug Saf. 42 (2019) 135–146, https://doi.org/10.1007/s40264-018-0764-x.

[22] AyaData. What Is Named Entity Recognition in NLP? Aya Data 2023. https://www.ayadata.ai/blog-posts/what-is-named-entity-recognition-in-nlp/ (accessed February 17, 2024).

[23] H. Xu, B. Hu, Legal Text Recognition Using LSTM-CRF Deep Learning Model, Comput. Intell. Neurosci. 2022 (2022) 9933929, https://doi.org/10.1155/2022/9933929.

[24] Y. Li, X. Peng, J. Li, S. Peng, D. Pei, C. Tao, et al., Development of a Natural Language Processing Tool to Extract Acupuncture Point Location Terms, in: 2023 IEEE 11th Int. Conf. Healthc. Inform. ICHI, 2023, pp. 344–351, https://doi.org/10.1109/ICHI57859.2023.00053.

[25] Panchendrarajan R, Amaresan A. Bidirectional LSTM-CRF for Named Entity Recognition. In: Politzer-Ahles S, Hsu Y-Y, Huang C-R, Yao Y, editors. Proc. 32nd Pac. Asia Conf. Lang. Inf. Comput., Hong Kong: Association for Computational Linguistics; 2018.

[26] Jie Z, Lu W. Dependency-Guided LSTM-CRF for Named Entity Recognition 2019.

[27] M.H. Syed, S.-T. Chung, MenuNER: Domain-Adapted BERT Based NER Approach for a Domain with Limited Dataset and Its Application to Food Menu Domain, Appl. Sci. 11 (2021) 6007, https://doi.org/10.3390/app11136007.

[28] walidamamou. Mastering Named Entity Recognition with BERT. UBIAI 2023. https://ubiai.tools/mastering-named-entity-recognition-with-bert/ (accessed February 17, 2024).

[29] Y. Zhao, X. Yuan, Y. Yuan, S. Deng, J. Quan, Relation extraction: advancements through deep learning and entity-related features, Soc. Netw. Anal. Min. 13 (2023) 92, https://doi.org/10.1007/s13278-023-01095-8.

[30] T. Chen, M. Wu, H. Li, A general approach for improving deep learning-based medical relation extraction using a pre-trained model and fine-tuning, Database (2019;2019:baz116.), https://doi.org/10.1093/database/baz116.

[31] P. Su, K. Vijay-Shanker, Investigation of improving the pre-training and fine-tuning of BERT model for biomedical relation extraction, BMC Bioinf. 23 (2022) 120, https://doi.org/10.1186/s12859-022-04642-w.

[32] Reimers N, Gurevych I. Optimal Hyperparameters for Deep LSTM-Networks for Sequence Labeling Tasks 2017. 10.48550/ARXIV.1707.06799.

[33] Y. Wang, S. Liu, N. Afzal, M. Rastegar-Mojarad, L. Wang, F. Shen, et al., A comparison of word embeddings for the biomedical natural language processing, J. Biomed. Inform. 87 (2018) 12–20, https://doi.org/10.1016/j.jbi.2018.09.008.

[34] V. G, H. R, J. Hareesh. Relation Extraction in Clinical Text using NLP Based Regular Expressions. 2019 2nd Int. Conf. Intell. Comput. Instrum. Control Technol. ICICICT, vol. 1, 2019, p. 1278–82. 10.1109/ICICICT46008.2019.8993274.

[35] G. Personeni, E. Bresso, M.-D. Devignes, M. Dumontier, M. Smaïl-Tabbone, A. Coulet, Discovering associations between adverse drug events using pattern structures and ontologies, J Biomed Semant 8 (2017) 29, https://doi.org/10.1186/s13326-017-0137-x.

[36] N. Liu, C.-B. Chen, S. Kumara, Semi-Supervised Learning Algorithm for Identifying High-Priority Drug-Drug Interactions Through Adverse Event Reports, IEEE J.

Biomed. Health Inform. 24 (2020) 57–68, https://doi.org/10.1109/JBHI.2019.2932740.
[37] S. Gupta, S. Pawar, N. Ramrakhiyani, G.K. Palshikar, V. Varma, Semi-Supervised Recurrent Neural Network for Adverse Drug Reaction mention extraction, BMC Bioinf. 19 (2018) 212, https://doi.org/10.1186/s12859-018-2192-4.
[38] Y. Li, J. Li, Y. Dang, Y. Chen, C. Tao, Temporal and Spatial Analysis of COVID-19 Vaccines Using Reports from Vaccine Adverse Event Reporting System, JMIR Prepr (2023), https://doi.org/10.2196/preprints.51007.
[39] Y. Li, S.K. Lundin, J. Li, W. Tao, Y. Dang, Y. Chen, et al., Unpacking adverse events and associations post COVID-19 vaccination: a deep dive into vaccine adverse event reporting system data, Expert Rev. Vaccines 23 (2024) 53–59, https://doi.org/10.1080/14760584.2023.2292203.
[40] Y. Li, J. Li, J. He, C. Tao, AE-GPT: Using Large Language Models to Extract Adverse Events from Surveillance Reports-A Use Case with Influenza Vaccine Adverse Events, 2023. ArXivorg. https://doi.org/10.48550/arXiv.2309.16150.
[41] Y. Hu, I. Ameer, X. Zuo, X. Peng, Y. Zhou, Z. Li, Y. Li, J. Li, X. Jiang, H. Xu, Zero-shot Clinical Entity Recognition using ChatGPT, 2023. ArXivorg. https://doi.org/10.48550/arXiv.2303.16416.
[42] E.C. Leas, J.W. Ayers, N. Desai, M. Dredze, M. Hogarth, D.M. Smith, Using Artificial Intelligence (ChatGPT) to Support Biomedical Text Analysis: A Case Study of Adverse Event Detection, JMIR Prepr (2023).
[43] M.J. Page, D. Moher, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, Prisma,, et al., explanation and elaboration: updated guidance and exemplars for reporting systematic reviews, BMJ 2021 (2020) n160, https://doi.org/10.1136/bmj.n160.