Advances in Artificial Intelligence and Machine Learning;Research (Ahead of Print) Received 03-12-2025; Accepted 05-03-2026; Published online 12-03-2026

Biomedical Named Entity Identification using Machine Learning

Mehroz Sadiq mehroz.buic@bahria.edu.pk
Department of Computer Science, Bahria University
E-8, Islamabad, Pakistan

Fatima Khalique fkhalique.buic@bahria.edu.pk
Department of Computer Science, Bahria University
E-8, Islamabad, Pakistan

Saba Mahmood smahmood.buic@bahria.edu.pk
Department of Computer Science, Bahria University
E-8, Islamabad, Pakistan

Riad Alherby ralharbi@uj.edu.sa
Department of Information Systems and Technology
College of Computer Science and Engineering, University of Jeddah
Jeddah, Saudi Arabia

Sachi Arafat Sarafat@kau.edu.sa
King AbdulAziz University
Saudi Arabia

Ali Daud Adaud@ra.ac.ae
Faculty of Resilience, Rabdan Academy
Abu Dhabi, Abu Dhabi, United Arab Emirates

Corresponding Author: Ali Daud and Saba Mahmood

Abstract

In recent times, teething dispute in recognizing the drug, chemical name entities and automatic extracting of relevant information from biological literature causes difficulties for the experts. There is an essential need of data mining techniques to develop a system which can help in automatic extraction of information so that the problem to manually find the information could be minimized. To handle this assortment, this paper focuses on the proposed methodology of recognizing the biological entities, in which five chemical entities (Protein, DNA, RNA, Cell type, Cell line) are recognized accurately. The presented Conditional Random Fields (CRFs) in the core of solution, Biomedical Name Entity Recognizer, are trained on orthographic and contextual features to segment and label the sequence data. The system is also capable of interpreting chemical formulas. The system is successful in annotating the chemical entities containing 3000 abstracts as training data, 3500 abstracts as development data sets, and 14000 records containing 7000 subset records as test data. The obtained results are encouraging, with 92.2% of precision, 93.2% of recall, and 92.48%

Citation: Mehroz Sadiq, et al. Biomedical Named Entity Identification using Machine Learning. **Advances in Artificial Intelligence and Machine Learning**. 2026. (Ahead of Print). https://dx.doi.org/10.54364/AAIML.2026.62288

of F-score measures for Chemical Entity Mention in Patent (CEMP) and 92% of precision, 95.21% of recall, 93.4% of F-score for Chemical Passage Detection (CPD).

Keywords: A biomedical Name entity recognizer, Conditional random fields, Chemical entities, Regular expressions, Machine learning.

1. INTRODUCTION

The past decade shows that the number of biological documents has been increasing at a high rate with the rise towards inclusion of iot,sensors, and other such technologies in healthcare[1]. The rapidly increasing volume of such biological documents causes a serious problem in manually finding and extracting relevant information from these documents [2]. Indeed, the need for automatic classification of unlabeled data becomes crucial. New-fangled statistics can be easier to access by automatic indexing of individual publications of biomedical corpus, containing the names of chemical entities mentioned in them. Knowledge about the location of the chemical entities in published reports is helpful in building relationships with other chemical compounds or concepts [3, 4]. Annotation of biological corpus is now a vital issue in all areas of Natural Language Processing (NLP) and Information Extraction (IE).

Research in the biomedical sciences and bioinformatics domain is mainly centered on conventional entities such as proteins, genes, DNA, RNA, chemicals, and drugs. The demand for facilitating a more proficient way to retrieve documents and sentences describing these entities is at its peak [5]. The recognition of chemical entities and drugs must deal with a substantial degree of variable names encountered between and within different chemical subdiscipline [6]. There exists a number of different representations and nomenclatures for chemical and drug-recognizing names [7]. Some of the well-known standards are provided by SMILES, InChi [8], and IUPAC. SMILES and InChi permit a direct structure search [9]. The different approaches for name entity recognition are divided into three categories: dictionary-based, morphology-based (grammar-based), and context-based [10]. In the dictionary-based approach, the matching between the dictionary terms is detected in the text usually in the form of two ways: partial matches and complete matches. Usually, well-known chemical databases are required to build dictionaries. The problem arises in dictionary-based approach with its drawback that it is approximately impossible to identify and include all the identifiers such as the IUPAC-names [11]and SMILES which follow precise grammar. Grammar-based approaches gain the advantage over dictionary-based approaches in extracting the names of chemical entities through the capture of systematic terms utilizing this set of predefined rules [12]. The drawback of both the approaches, dictionary-based, and grammar-based, is that these approaches may undergo from tokenization problem. The name entity recognition of chemical compounds is a context aware approach. This approach uses machine learning techniques and algorithms and natural language processing (NLP) for the extraction of chemical entities terms. The problem arises in machine learning approaches when there is the need for sufficiently large, annotated corpus is required for training set [13].

Although, several tools, which are efficient in use for the purpose of biomedical text mining and name entity recognition [14], suffer with limitations. Initially, they offer annotations for a limited set of biomedical entity types (e.g., five types. Additionally, they frequently utilize several single types Named Entity Recognition (NER) models to identify different entity types. This approach demands substantial Graphics Processing Unit (GPU) memory for parallel processing but is notably slow during inference when run sequentially. Furthermore, many systems use Named Entity Normalization (NEN) models that rely on rule-based methods with dictionaries [15]. These models often struggle with the diverse morphological forms of biomedical entities. For example, a straightforward dictionary-based NEN model cannot convert 'oxichlorochine' to its standard form 'hydroxychloroquine' unless the dictionary specifically includes 'oxichlorochine' [16]. In the biomedical field, Named Entity Recognition (NER) is typically followed by Relation Detection (RD), also referred to as relation extraction or entity association. This process involves linking various biomedical entities to uncover significant interactions that warrant further investigation. Given the vast array of named entity classes in biomedicine, the potential combinations of these entities can quickly become overwhelming. Conducting biological experiments to identify the most critical relationships would be both expensive and time intensive [17]. However, by utilizing computational methods to analyze millions of biomedical research papers, it is possible to discover numerous associations and construct networks. For example, detecting protein interactions can lead to the development of protein-protein interaction networks [18]. Similarly, mapping gene-disease relationships can connect molecular data with phenotypic information. These relational networks not only help in pinpointing previously unknown connections for further exploration but also provide a comprehensive overview of the interactions among various biological entities, including diseases, genes, food, drugs, side effects, pathways, and toxins, thereby paving the way for new research opportunities [19].

Hence, the fact reveals that the biomedical literature becomes voluminous due to the emerging integration of IoT, enhanced technologies, and involvement of the digital systems in all departments including healthcare [1]. There are major challenges which occur during the extraction of meaningful information from these literatures. Although, it is time-consuming and inefficient approach if the data is extracted through manual procedures. Although there exist a number of tools for Biomedical Name Entity Recognition (BNER), the tools are still facing the critical limitations, as the dictionary-based and rule-based approaches are lacking in recognition of original or non-standard chemical entities. This drawback occurs especially for entities that follow the formats and structures of the SMILIES, InChI, or IUPAC nomenclature. However, deep learning models showcase their performance on widespread annotated datasets, which are not always accessible to cover all biomedical entity types or languages. Moreover, there exist real-time applicability gap in several models and are not able to generalize on multiple languages in biomedical subdomains [20].

However, there are loop holes in detecting the names of chemical and drug entities which require the dire need of efficient and adaptable biomedical NER systems that must be capable of recognizing a assorted range of entity types such as protein, DNA, RNA, cell line, and cell types. Moreover, these systems should be computationally efficient and generalizable across several types of dataset and literature formats.

This paper depicts its importance by presenting a robust Conditional Random Fields (CRFs)-based approach for the recognition of biomedical entities that integrates number of methods like contextual features, regular expressions, and dictionary matching techniques. The proposed system is capable of demonstrating high performance in recognition of key biomedical entities and also, shows the relationships from unstructured literature. This work has its significant importance for progressive biomedical text mining, improved knowledge discovery, and usage in critical applications such as drug detection, disease modelling, and helps in making decision at clinical level approaches. Also, this proposed technique contributes to the emerging demand for practical and intelligent systems for extraction of biomedical information by highlighting both scalability and accuracy.

The key contributions of this research article include:

- **Biomedical Document Retrieval:** Emphasized the growing need for efficient document and sentence retrieval mechanisms related to biomedical entities.
- **Challenges in Chemical and Drug Recognition:** Discussed the complexities of recognizing chemical entities and drugs due to varied nomenclatures and representations [21], including SMILES, InChi, and IUPAC standards.
- **Categorization of NER Approaches:** Classified Named Entity Recognition (NER) techniques into dictionary-based, morphology-based (grammar-based), and context-aware methods, analyzing their respective strengths and weaknesses.
- **Advancements in Context-Aware Methods:** Reviewed modern context-aware NER techniques using machine learning and natural language processing (NLP), emphasizing their reliance on large, annotated datasets for effective training.

This work is primarily empirical, it focuses on a practically effective CRF-based framework that integrates contextual and orthographic features with explicit chemical formula handling, evaluated at scale on biomedical datasets.

2. LITERATURE REVIEW

Indeed, it is tedious to extract the chemical entities from the texts[21]. The foremost reasons behind this difficulty are the large number of terms and synonyms within the chemical domain, failure to act upon the rules while creating systematic terms by authors, the use of specific characters such as hyphens, commas, exclamatory marks, and quotation marks within the chemical terms, the use of stop words such as it, and, or, while, when, etc. in biomedical text, and the uncertainty within and across the chemical databases. These difficulties are tackled by the use of approaches which are previously mentioned [20]. Kocaman et al. [22] works for the name entity recognition by developing and implementing a Bi-LSTM-CNN-Char deep learning model on Apache Spark, designed to enhance Named Entity Recognition (NER) tasks. This new model achieves cutting-edge performance on seven widely used biomedical datasets, surpassing previous benchmarks without relying on extensive contextual embeddings like BERT. The model demonstrates significant improvements across various datasets, including achieving an accuracy of 93.72% on BC4CHEMD, which represents a 4.1% improvement; 80.91% on Species800, marking a 4.6% gain; and 81.29% on JNLPBA, reflecting a 5.2% enhancement. Another work demonstrated the use of deep neural network and word embeddings for urdu language[23].

The model is available without any cost through open source spark NLP library in contribution of its massive performance. The design of model is made to be compatible with any of the Spark cluster, which allows scalable training and interference. In addition, model is also capable of supporting GPU acceleration, which is beneficial in increasing of processing speed and efficiency.

Moreover, it also assists by the libraries of various programming languages, including Python, R, Java, and Scala. The architecture of the model is flexible and capable to support additional human languages without code changes, helping to behave like a versatile tool for different linguistic and computational applications.

The proposed work by Asghari et al. [24] has the advantage for the word entity recognition by utilizing the Begin-Inside- Outside technique. This model is capable to use the several algorithms during training to group the various words altogether. The understanding of the given entity is highly depended on upon the context that is provided. The intricate nature of the word structure, usually, cause the doldrums for traditional neural networks. The sentence is considered as the collection of well-defined words, is ideally investigated using Recurrent Neural Network (RNN) [25] architectures, which are compatible for handling sequential data. RNNs are usually used in linguistic models but they suffer from significant limitations including trouble in apprehending long-range dependencies within a sequence known as the waning gradient problem. To handle this assortment, the Long Short-Term Memory (LSTM) [26] model was presented. LSTMs improve the RNN by permitting sequences to be preserved and maintained over longer periods, thus providing a mechanism to retain and utilize information throughout the sequence. In Name Entity Recognition (NER), the given approach by author plays the crucial role for tasks which are in charge of both backward and forward processing to highlight the relationships. As the entities are made up of groups of relational words, identification of these entities is necessary in understanding the context within the both directions of the sequences. Therefore, authors employ a Bidirectional LSTM (BiLSTM) [27] approach, which helps in leveraging the information from past and future contexts to improve entity detection.

Hakala et al. [28], have presented their system to the Turku NLP group, who have detailed their approach to manage the PharmaCoNER task, which emphases on Spanish biomedical name entity recognition. Their method to detect he name entities employs a amalgamation of a CRF-based baseline model and multilingual BERT, a state-of-the-art language model. They have achieved prominent performance, with an F-score of 88% on the development dataset and 87% on the test dataset using BERT by integrating these techniques. These results highlight the efficiency of smearing a sophisticated multilingual model to the task, even though the algorithm was not specifically personalized for either the Spanish language or the biomedical domain. This approach determines how leveraging has been made advance, and general-purpose models can yield powerful results in dedicated applications.

Weber, Leon et al. [29], have introduced HunFlair, a NER tagger designed to meet these criteria. HunFlair is integrated with the popular NLP [30] framework Flair and can recognize five types of biomedical entities. It matches or surpasses state-of-the-art performance across various evaluation datasets and is trained in a cross-corpus manner to minimize bias related to specific corpora. The tool utilizes a character-level language model that has been pretrained on approximately 24 million biomedical abstracts and three million full-text documents. HunFlair outperforms other commercially available biomedical NER tools, showing an average improvement of 7.26 percentage points over the next best tool in cross-corpus evaluations, and delivers comparable results to leading research prototypes in in-corpus tests. NER models and systems are providing their aid in healthcare departments in several ways [31]. Additionally, HunFlair can be easily installed with a single command and used with just four lines of code. It also comes with standardized versions of 23 biomedical NER corpora. These type of systems are helpful in health care departments to increase the capablity of work flow. Although, several researchers have contributed in chemical entities recognitions, there exists some key limitations which are required to focus on.

The following section outlines the challenges and limitations identified in previous research efforts:

- **Vanishing Gradient Problem in RNNs:** Traditional RNNs struggle to capture long-range dependencies, making them less effective for biomedical entity recognition. While LSTMs address this issue, challenges in maintaining long-term dependencies still exist [24].
- **Lack of Domain-Specific Adaptation in Multilingual Models:** Although multilingual BERT demonstrates strong performance in biomedical NER, it is not specifically designed for the biomedical or Spanish language domain, limiting its effectiveness for specialized tasks [29].
- **Sequential Processing Slows Inference:** Existing NER tools process entities sequentially, resulting in slower inference times, which can hinder real-time applications [32].
- **Cross-Corpus Bias in NER Models:** Despite efforts to improve generalizability, cross-corpus evaluations highlight biases that impact the performance of biomedical NER models across different datasets [28].
- **Limited Support for Underrepresented Languages:** While certain models claim multilingual capabilities, the lack of adaptation for underrepresented languages restricts their applicability in diverse linguistic contexts [29].

TABLE 1 summarizes the contributions of different authors.

Table 1: Comparison of Biomedical NER Approaches

Ref.	**Approach**	**Dataset**	**Evaluation Measures**
Hakala et al. [28]	CRF-based model, multilingual BERT	PharmaCoNER	F1-score: 88% (development), 87% (test)
Kocaman et al. [32]	Bi-LSTM-CNN	BC4CHEMD, Species800	80.91% on species, 81.29% on JNLPBA
Asghari et al. [24]	B-I-O technique, RNN, LSTM	JNLPBA, BIONLP13PC, ADEs	0.92 precision, 0.89 recall, 0.89 F1-score
Leon et al. [13]	HunFlair	24M biomedical abstracts, 3M full-text documents	F1-score: 59.69 (chemical), 72.19 (gene), 85.05 (species)

3. PROPOSED METHODOLOGY

To deal with the assortment and variety in chemical entities, this research reveals the invocation of our system to handle the Biomedical Name Entity Recognition (BNER) tasks. To handle this assortment, this paper focuses on the proposed methodology of recognizing the biological entities, in which five chemical entities (Protein, DNA, RNA, Cell type, Cell line) are recognized accurately. The system is also capable of interpreting chemical formulas. The system is successful in annotating the chemical entities containing 3000 abstracts as training data, 3500 abstracts as development data sets, and 14000 records containing 7000 subset records as test data.

The proposed system pipeline involves the integration of pre-processing module to work on data to handle tokenization, SBD, stop words removal, and POS tagging. Pre-processed data is passed to the proposed algorithm Conditional Random Fields to recognize the chemical entities. The entities are classified into six classes, i.e. gene, protein, cell line, cell type, DNA, and RNA.

System architecture as shown in FIGURE 1, demonstrates the several stages of proposed work: section of the participating system.

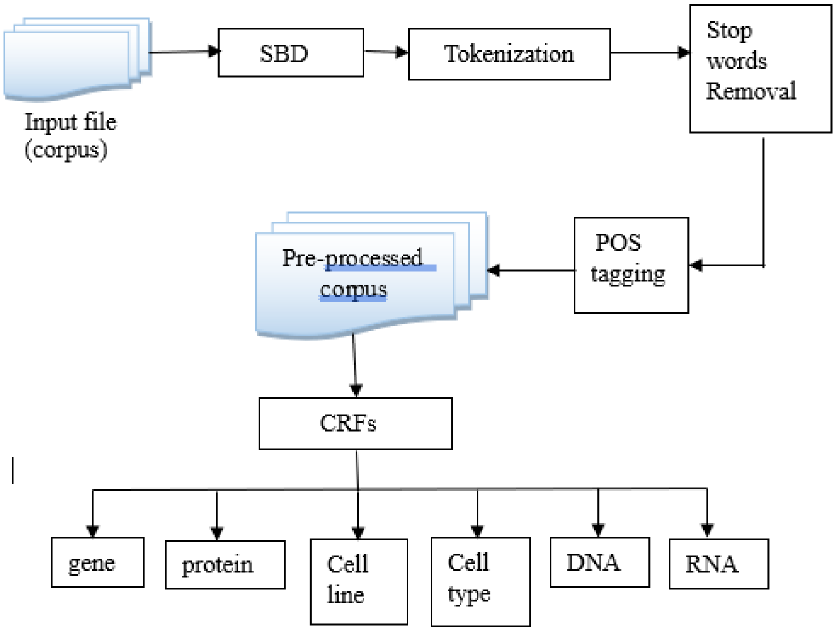

Figure 1: Process Flowchart

3.1 Dataset

The system achieves its precision by annotating chemical entities using a global data set, which comprises 3000 abstracts for training, 3,500 abstracts for development, and 14,000 records available for testing. The test data set is split into subset of 7,000 records, ensuring a vigorous evaluation of the system's performance in chemical name entity recognition [33] [34].

3.2 Pre-processing

There are several sub-tasks involved in preprocessing to increase the precision and recall rate of the system and are applied on the training and development data. The pre-processing stage includes the following sub-tasks:

3.2.1 Sentence boundary detection

SBD seems to be easy task as the sentence boundary is detected on basis of punctuation marks (PMs): period, question mark, exclamation mark, and suspension points [35]. In some cases, there is the need of special care to notice as these PMs may also play other roles in sentences other than the detection of boundaries. For example, periods in date and in serial numbers. This problem can be handled by inspection that there is no blank character after the period.

3.2.2 Lexicalization (tokenization)

This stage of pre-processing involves the division of the sentence into tokens (words, numbers, punctuation etc.). In proposed methodology, the devised tokenization algorithm is based on the regular expressions which chops the stream into pieces. In tokenization algorithm, we tokenize the word stream on the basis of white spaces and other kinds of symbols depicted below.

{ ’ ’, ’\r’, ’\n’, ’.’, ’:’, ’;’, ’,’, ’=’, ’(’, ’)’, ’[’, ’]’, ’{’, ’}’, ’!’, ’@’, ’#’, ’$’, ’%’, ’^’, ’&’, ’*’, ’|’, ’?’, ’<’, ’>’, ’/’, ’\’, □, ’"’, ’`’, ’~’ }

3.2.3 Stop words removal

Some extremely common words that implement the role of very little importance in a document in searching matches and necessarily to be excluded from documents. These words are commonly known as “stop words” or “stop word list”. The essential feature to remove these words from the documents is to reduce the number of terms so the performance of the system would not affect and to achieve time efficiency. Stop word list includes determiners (nouns), coordinating conjunctions (connecting words), and prepositions Stop words list consisting of 36 words are depicted in TABLE 2. The algorithm proposed to confiscate the stop words is dictionary based in which the matches hit for any of the word represent in TABLE 2, the algorithm eliminates it from the document.

Table 2: Stop Words List of 36 Words

a	an	and	is	am	it	of	at	be
In	he	on	as	to	so	by	its	has
was	but	for	yet	are	not	too	this	that
were	will	would	before	under	towards	another	there	from

3.2.4 Part-of-speech (POS) Tagging

This phase of pre-processing involves the tagging of each token with part of speech such as noun, pronoun, adjective, verb etc. POS tagger is actually, a piece of software and the tagger used to tag the biological data in our proposed methodology is the MaxentTagger by Standford Log-Linear Part-Of-Speech Tagger. MaxentTagger is efficient enough time to produce the results of huge file in few seconds and it is licensed under the GNU General Public License [36]. MaxentTagger is made by a constructionist where the location of trained tagger's parameter files is provided as arguments as: MaxentTagger tagger = new MaxentTagger("models/left3words-wsj-0-18.tagger"); [37].

3.3 Name Entity Recognition (NER)

Name entity recognition is the crucial NLP step for the information management aims and considers tagging the tokenized biomedical terms (DNA, RNA, protein, cell line, cell type etc.). Statistic machine learning system which uses linear-chain conditional random fields (CRFs) with the number of contextual and orthographic features for biomedical terms analysis, is used to achieve the required goal. The user-friendly interface of Biomedical Name Entity Recognition (BNER) is easy to use where the biomedical text commands can be typed manually or loaded from a file and multiple entities can be tagged automatically in real time. The interesting feature of the proposed methodology is that each biomedical entity is highlighted with a specified unique color simultaneously (green = DNA, yellow = Protein, pink = Cell type etc.) and annotated document can be saved in multiple file formats.

In addition to standard biomedical entities, the proposed system is capable of recognizing chemical formulae present in unstructured biomedical text. This functionality is achieved by incorporating domain-specific regular expression patterns as orthographic features within the Conditional Random Fields (CRF) framework. These patterns are designed to capture common structural characteristics of chemical formulae, including element symbols, numeric subscripts, and alphanumeric combinations. The CRF model utilizes these formula-specific features in conjunction with contextual information from neighboring tokens to accurately distinguish chemical formulae from general text during sequence labeling.

3.3.1 Algorithm and implementation

Conditional random fields (CRFs), belong to statistical modeling methods used in machine learning and pattern matching. These are probabilistic graphical models correspond to conditional finite state machines to label and segment the sequence data. These decomposable probabilistic graphical models by which CRF approach is derived also known as Markov random fields (MRKs). CRF's has its application in the areas of part-of-speech tagging, shallow parsing, and name entity recognition etc. [38]. During name entity recognition, CRF framework assigns the tokens with labels from the defined states.

Suppose $m = \{m_1, m_2, m_3, \ldots, m_n\}$ defines the sequence of words of length n. Let X be the set of labels DNA, RNA, others, etc., corresponds to the states in finite state machine. So, $X = \{x_1, x_2, x_3, \ldots, x_n\}$ be the labels from X assigned to the tokens from input m.. The conditional probability of a given input sequence 'm' is defined by the first order linear-chain CRF as given by equation 1.

$$P(X|m) = \frac{1}{G_o} \exp\left(\sum_{a=1}^{n} \sum_{b=1}^{k} \lambda_b t_b(x_{a-1}, x_a, m, a) \right) \tag{1}$$

Where G_0 refers to the normalization factor over the sequence of labels S, and defines one of the k binary functions that refer to the feature at position a from the sequence m of length n, and the weight of that feature is w_a. For example, the text "... the ATPase ..." has the feature Word = ATPase and feature value 1 with the transitions, where x_{a-1} represents the label state "other", and x_a represents the label state "protein". The weight of the feature w_a has a positive value for features that are correlated to the target label, a negative value for anti-correlated features, and is set to zero for uninformative features.

To set the weights of the feature value up to its maximum, the conditional log of the labeled sequence for the training data set $Z = \{(m,x)_1, (m,x)_2, (m,x)3 \dots, (m,x)_n\}$ is refered in equation 2.

$$LL(Z) = \sum_{a=1}^{t} \log\left(P(X_{(a)}|m_{(a)})\right) - \sum_{b=1}^{k} \frac{\lambda b^2}{2\sigma^2} \tag{2}$$

The feature sets of biomedical name entity recognition, i-e: orthographic and contextual features, mostly based on the neighboring tokens and the regular expressions. Furthermore, the feature set includes regular-expression-based indicators specifically designed for chemical formula structures. These indicators enable the model to effectively capture formula-level patterns and enhance the recognition of chemical expressions that do not follow conventional word-based representations.

3.3.2 Model complexity and trainable parameters

The complexity of the model is considered to be the critical aspect of the systems used for the recognition of biomedical names. Additionally, the number of trainable parameters gains the same importance in BNER systems. These are essential during practical deployment in clinical or embedded environments where the particular resources for computation are in limit. Moreover, we worked to develop this system by using the Conditional Random Fields (CRFs)-based model, comprising of both orthogonal and contextual features with rich input space. Our CRF model does not depend on entity embeddings or pre-trained language models and therefore has a suggestively smaller number of trainable parameters as compare to the deep learning models (e.g., BERT, BiLSTM-CRF).

The proposed system use CRF-based model which consists of approximately 25,000-40,000 trainable parameters, smaller than transformer based models such as BioBERT (110M parameters) and also smaller than BiLSTM models, which characteristically range from 500K to 10M parameters. This lower complexity permits faster training, slight memory necessities, and smooth deployment on control devices or legacy clinics and hospitals IT organizations.

4. RESULTS AND DISCUSSIONS

The data set provided by the challenge organizers is organized into three distinct sets: the training set, the development set, and the test set. The training and test sets come up with their crucial capabilities in the procedure of developing and evaluating the proposed system. In addition, the training dataset is comprised of 3,000 abstracts, which is helpful to train our system by allocating it to learn and train on underlying patterns and relations of the data entities. This considers to be the foundation on which the proposed model builds its capability to perform effectively on new, real-world unseen data.

Following the BioCreative V evaluation protocol, the performance of the proposed system is reported using overall precision, recall, and F1-score metrics rather than entity-wise class-specific measures.

Furthermore, the development set consisting of 3,500 abstracts, behaves as a middle stage in between of training and testing of the proposed methodology. The development set has been utilizing in fine tuning the model, make it possible to provide adjustments to enhance the performance of the system. It also, generalizes the system through training data. Similarly, the identification of potential problems or areas where the model suffers from overfitting problem can be performed by testing the proposed system on the development set. These steps allow us to correct these instead of moving towards the next step.

Moreover, the test dataset is comprised of 14,000 records, consider to be the largest one. This dataset is further divided into subsets of 7,000 abstracts each and used in final evaluation of the system's performance. No doubt, it is difficult to handle the working on test set, as it represents the measure of how our proposed system behave during the unseen data. The ability of the system depends on the process to handle the large amount of data in the test set, while considering the accuracy and robustness as important as before. This inclusive approach guarantees that our proposed model is not only well designed, but also capable of solving the real-time world and give responses effectively.

The proposed system achieves the accuracy with 92.2% precision, 93.2% recall, and 92.48% F1 score for CEMP (Chemical Entity mention in Patent) and 92% precision, 95.21% recall, and 93.% F-score for the CPD (Chemical Passage Detection).

4.1 System Performance Evaluation and Optimization

The effectiveness and efficiency of the proposed system is measured systematically by using training, development and test datasets. These datasets ensure about the desirable goals. This proposed approach helps us to access the system's functionality across several stages and allow us to fix it as per the requirement to optimize the performance. Specifically, we examine the numerous amalgamations of Conditional Random Fields (CRF), regular expressions, and dictionary matching techniques. These combinations are tested with extreme care to control and examine the most active configuration for extracting pertinent information and attaining high accuracy in the tasks at hand.

4.2 Performance Analysis for Chemical Entity Mention in Patents (CEMP)

The system is exposed to multiple runs on the test datasets, with the final results being summarized in TABLE 1, to ensure a thorough evaluation. In addition, the tasks involved in Chemical Entity Mention in Patents (CEMP), the performance of the system reaches to conclude average over five runs, yielding an average 92.2% precision, 93.2% recall, and 92.48% F1 score as shown in TABLE 3. These metrics designate the ability of the system to identify and retrieve relevant chemical entities accurately, from the data while creating balance in precision and recall.

4.3 Performance Analysis for Chemical-Protein-Disease (CPD)

In the same way, when the Chemical-Protein-Disease (CPD) task is concerned, the system is capable of achieving an average precision of 92% precision, 95.21% recall, and 93.% F-score across the five runs as shown in TABLE 3. These results are helpful in highlighting the effectiveness of the system in detecting, recognizing, and extracting relationships between chemicals, proteins, and diseases, which is considered to be crucial for progressing research in the biomedical field.

Table 3: Experimental Results for CEMP and CPD for Test Sets

CEMP	Run	Precision	Recall	F-score
	1	91.24%	98.69%	94%
	2	92.74%	90.75%	91.74%
	3	94.35%	90.82%	92.08%
	4	90%	92.54%	91.27%
	5	93.21%	93.43%	93.32%
CPD	**Run**	**Precision**	**Recall**	**F-score**
	1	93.65%	95.78%	94.21%
	2	91.89%	93.27%	92.58%
	3	92.01%	94.34%	93.17%
	4	92%	95.45%	93.22%
	5	90.45%	97.21%	93.83%

The precision metric imitates the correctness and accuracy of the system in perceiving the relevant entities, whereas recall measures its aptitude to recognize all conceivable relevant entities inside the dataset. The F-score, a harmonic means of precision and recall, delivers a well-adjusted measure of the overall performance of the system. The constancy of the proposed system's results across several runs establishes the robustness and consistency of the proposed system in managing complex biomedical tasks.

4.4 Evaluating and Refining the System for Optimal Accuracy

Overall, the iterative testing process enables us to fine-tune the components of the system, foremost to a well-proportioned and operative solution for the CEMP and CPD tasks. These results underscore the potential of the proposed system to be applied in real-world scenarios, where accurate and efficient extraction of chemical entities and their relationships is critical for advancing biomedical research and applications performance of the proposed system is obtained on training, development and test data to achieve the desired goal. This permits us to test the working of the system with the several combinations of CRF/regular expressions/dictionary matching.

The improvements in the results are demonstrated through a comparative analysis with previous studies, TABLE 4 presents a detailed comparison of the proposed approach with existing works. According to the comparison, it is clearly visible that the proposed method using CRFs to detect the chemical and drug name entities has achieved the highest accuracy. The system is successful in diagnosing the five chemical entities to classify in to five different classes i.e. gene, protein, DNA, RNA, cell type and cell line.

Table 4: Comparison of the Proposed Approach

Reference	Evaluation Measures
Hakala et al., 2019	F-score of 87% on the test dataset
Koca-man et al., 2021	80.91% on species, 81.29% on JNLPBA
Asghari et al., 2022	0.92 precision, 0.89 recall, 0.89 F1 score
Leon et al., 2022	F1 score: 59.69 for chemical, 72.19 for gene, 85.05 for species
Kumar et al., 2023	Reduction in MASE of 94% for wheat, 72% for groundnut, and 71% for barley
Proposed Methodology	92.2% precision, 93.2% recall, and 92.48% F-score for the CEMP and 92% precision, 95.21% recall, and 93.4% F-score for the CPD.

5. CONCLUSION

This article presents our contribution in recognizing the chemical and drug names by the CRF based solution working in the core of proposed methodology. The proposed robust system is helpful in extracting the chemical entities, passage drug entities, and their relationships by utilizing the amalgamation of regular expressions, dictionary matching techniques, contextual methods, and Conditional Random Fields (CRFs). Pre-processing modules are also integrated to remove noise and for time efficient results. Exploratory data analysis reveals the importance of pre-processing techniques to be included on the given data set to enhance the efficiency of system. The procedures involves are stop words removal, tokenization, sentence boundary detection and POS tagging. The running time performance of the proposed system is marvelously beneficial; as it performs its functionality on huge data comprising of 14000 records with subsets of 7000 records in a short span of time, almost in less than 2 hours.

The system is successful in systematically evaluating the data using training, development, and test data sets, and it also ensures its ability to predict the unseen data effectively. The results demonstrate the system's high precision, recall, and F1 scores, with 92.2% precision, 93.2% recall, and 92.48% F1 score for the Chemical Entity Mention in Patents (CEMP) task, and 92% precision, 95.21% recall, and 93% F1 score for the Chemical-Protein-Disease (CPD) task. Performance metrics culminate in the productiveness of the proposed approach by identifying and retrieving admissible chemical and drug entity names with a high accuracy between precision and recall.

The proposed project has proven to be accurate and robust, has the ability to handle a large amount of data set with high maintenance by using iterative testing and optimization. Moreover, the analysis, which leads to a comparison with previously discussed methods, validates the advances of the system in the recognition of chemical and drug entity names. This reveals the importance of the system for real-world biomedical applications. Due to the effectiveness of the proposed system, this contribution can contribute further in extracting information from biomedical research, facilitate knowledge discovery, and increase scientific advancements in the field.

6. FUTURE WORK

Expanding the training dataset with more diverse biomedical corpora and extending the system to recognize additional entity types, such as drugs and pathways, would increase its applicability. Additionally, integrating semantic contextualization through knowledge graphs could improve the extraction of meaningful insights, while developing real-time processing capabilities and a user-friendly interface would make the system more practical for researchers. To expand its global usability, incorporating multilingual support and testing the system in longitudinal studies could further validate its robustness and applicability in real-world scenarios.

7. DECLARATIONS

7.1 Ethics approval and consent to participate:

Not applicable

7.2 Consent for publication:

All authors agree to publication.

7.3 Availability of data and material:

The simulation files/data used to support the findings of this study are available from the corresponding author upon request.

7.4 Competing interests:

It is declared that the authors have no competing interests.

7.5 Funding:

No funding involved.

7.6 Authors contributions

Mehroz developed the main idea and conducted experiments. Fatima and Saba worked on methodology and wrote the manuscript. Riad and Tariq validated the experimental results. Ali carried out editing and proofread of the manuscript for final submission.

7.7 Acknowledgement

Not applicable

References

[1] Abbas T, Khan AH, Kanwal K, Daud A, Irfan M, et al. Iomt-Based Healthcare Systems: A Review. Computer Systems Science and Engineering. 2024;48:871-895.

[2] Xu D, Chen W, Peng W, Zhang C, Xu T, et al. Large Language Models for Generative Information Extraction: A Survey. Front Comput Sci. Springer Nature. 2024;18:186357.

[3] Keloth VK, Hu Y, Xie Q, Peng X, Wang Y, et al. Advancing Entity Recognition in Biomedicine via Instruction Tuning of Large Language Models. Bioinformatics. 2024;40:btae163.

[4] Wang Y, Tong H, Zhu Z, Hou F, Li Y. Enhancing Biomedical Named Entity Recognition With Parallel Boundary Detection and Category Classification. BMC Bioinform. Springer Nature. 2025;26:63.

[5] Chiruzzo L, Jiménez-Zafra SM, Rangel F. Overview of Iberlef 2024: Natural Language Processing Challenges for Spanish and Other Iberian Languages. InIberLEF@ SEPLN. 2024;3756.

[6] Maojo V, Fritts M, Martin-Sanchez F, De la Iglesia D, Cachau RE, et al. Nanoinformatics: Developing New Computing Applications for Nanomedicine. Computing. Springer Nature. 2012;94:521-539.

[7] Liu S, Wang A, Xiu X, Zhong M, Wu S. Evaluating Medical Entity Recogni-Tion in Health Care: Entity Model Quantitative Study. JMIR Med Inform. 2024;12:e59782.

[8] Kim S, Chen J, Cheng T, Gindulyte A, He J, et al. Pubchem 2025 Update. Nucleic Acids Res. 2025;53:D1516-D1525.

[9] Blanke G, Brammer J, Baljozovic D, Khan NU, Lange F, et al. Making the InCHI FAIR and Sustainable While Moving to Inorganics. Faraday Discuss. 2025;256:503-519.

[10] Peiris TD, Asanka PD. Sinhala Document Clustering Using Named Entity Recog-Nition Technique. In2024 4th International Conference on Advanced Research in Computing. ICARC. IEEE. 2024:179-183.

[11] Jehangir B, Radhakrishnan S, Agarwal R. A Survey on Named Entity Recognition – Datasets Tools and Methodologies. Nat Lang Proc J. 2023;3:100017.

[12] Dash A, Darshana S, Yadav DK, Gupta V. A Clinical Named Entity Recogni-Tion Model Using Pretrained Word Embedding and Deep Neural Networks. Decis Anal J. 2024;10:100426.

[13] Lee JU, Klie JC, Gurevych I. Annotation Curricula to Implicitly Train Non-Expert Annotators. Comput Linguist. 2022;48:343-373.

[14] Khan W, Daud A, Shahzad K, Amjad T, Banjar A, et al. Named Entity Recognition Using Conditional Random Fields. Appl Sci. 2022;12:6391.

[15] Neuberger J, Ackermann L, Jablonski S. Beyond Rule-Based Named Entity Recognition and Relation Extraction for Process Model Generation From Natural Language Text. InInternational Conference on Cooperative Information Systems. Cham:Springer Nature. 2023:179-197.

[16] Rajamanickam D. Improving Legal Entity Recognition Using a Hybrid Transformer Model and Semantic Filtering Approach. 2024. arXiv preprint:https://arxiv.org/pdf/2410.08521.

[17] Kpanou R, Dallaire P, Rousseau E, Corbeil J. Learning Self-Supervised Molecular Representations for Drug–Drug Interaction Prediction. BMC Bioinform. 2024;25:47.

[18] Waqar M, Majeed N, Dawood H, Daud A, Aljohani NR. An Adaptive Doctor-Recommender System. Behav Inform Technol. 2019;38:959-973.

[19] Jin M, Choi SM, Kim GW. Comcare: A Collaborative Ensemble Frame-Work for Context-Aware Medical Named Entity Recognition and Relation Extraction. Electronics. 2025;14:328.

[20] https://theses.hal.science/tel-04877187/file/142011_BOROVIKOVA_2024_archivage-1.pdf.

[21] Alharbey R, Kim JI, Daud A, Song M, Alshdadi AA, et al. Index-Ing Important Drugs From Medical Literature. Scientometrics. 2022;127:2661-2681.

[22] Kocaman V, Talby D. Biomedical Named Entity Recognition at Scale. In 2021 International Conference on Pattern Recognition. Cham: Springer International Publishing. 2021:635-646.

[23] Khan W, Daud A, Alotaibi F, Aljohani N, Arafat S. Deep Recurrent Neural Networks With Word Embeddings for Urdu Named Entity Recognition. ETRI J. 2020;42:90-100.

[24] Asghari M, Sierra-Sosa D, Elmaghraby AS. Biner: A Low-Cost Biomedical Named Entity Recognition. Inf Sci. 2022;602:184-200.

[25] Aljuaydi F, Zidan M, Elshewey AM. A Deep Learning CNN-GRU-RNN Model for Sustainable Development Prediction in Al-Kharj City. Eng Technol Appl Sci Res. 2025;15:20321-20327.

[26] Hasan MW. Design of an Iot Model for Forecasting Energy Consumption of Residential Buildings Based on Improved Long Short-Term Memory (LSTM). Meas Energy. 2025;5:100033.

[27] Kumar VK, Ramesh KV, Rakesh V. Optimizing Lstm and Bi-Lstm Models for Crop Yield Prediction and Comparison of Their Performance With Traditional Machine Learning Techniques. Appl Intell. Springer Nature. 2023;53:28291-28309.

[28] Hakala K, Pyysalo S. Biomedical Named Entity Recognition With Multilingual Bert. InProceedings of the 5th workshop on BioNLP open shared tasks. ACL Anthology. 2019:56-61.

[29] Weber L, Sänger M, Münchmeyer J, Habibi M, Leser U, et al. Hunflair: An Easy-To-Use Tool for State-Of-The-Art Biomedical Named Entity Recognition. Bioinformatics. 2021;37:2792-2794.

[30] Zhang Y, Liu J, Zhong X, Wu L. Seclmner: A Framework for Enhanced Named Entity Recognition in Multi-Source Cybersecurity Data Using Large Language Models. Expert Syst Appl. 2025;271:126651.

[31] Masood I, Wang Y, Daud A, Aljohani NR, Dawood H. Towards Smart Healthcare: Patient Data Privacy and Security in Sensor-Cloud Infrastructure. Wirel Commun Mob Comput. 2018;2018:2143897.

[32] Kocaman V, Talby D. Biomedical Named Entity Recognition at Scale. InInternational Conference on Pattern Recognition. Cham: Springer Nature. 2021:635-646.

[33] https://www.ncbi.nlm.nih.gov/research/bionlp/biocreative#bioc-5.

[34] http://www.biocreative.org/tasks/biocreative-v/.

[35] Sheik R, Ganta SR, Nirmala SJ. Legal Sentence Boundary Detection Using Hybrid Deep Learning and Statistical Models. Artif Intell Law. 2024;33:519-549.

[36] https://sergey-tihon.github.io/Stanford.NLP.NET/other/POSTagger.html

[37] http://nlp.stanford.edu/nlp/javadoc/javanlp/edu/stanford/nlp/tagger/maxent/MaxentTagger.html.

[38] Goyal N, Singh N. Named Entity Recognition and Relationship Extraction for Biomedical Text: A Comprehensive Survey Recent Advancements and Future Research Directions. Neurocomputing. 2024;618:129171.