

Comparison of CRF and BiLSTM Models for Disease Named Entity Recognition on the NCBI Disease Dataset

Agus Siswanto¹, Muhammad Wardani², Yulia Arvita³, Jasmir⁴

Department of Computer Engineering, Faculty of Computer Science, Universitas Sriwijaya

**agussiswanto@unama.ac.id*

ABSTRACT

Named Entity Recognition (NER) is a fundamental task in Natural Language Processing (NLP) aimed at identifying and classifying named entities in text. In the biomedical domain, disease entity recognition is crucial for supporting automated medical information extraction. This study compares two NER model approaches, Conditional Random Field (CRF) and Bidirectional Long Short-Term Memory (BiLSTM), for disease entity recognition using the NCBI Disease dataset. The CRF model utilizes handcrafted linguistic features including prefixes, suffixes, POS tags, and word shapes, while BiLSTM employs GloVe 200-dimensional embeddings and automatically learns features from sequential context. BiLSTM experiments were conducted using 5 different seeds for robust statistical validation, reporting mean F1-score, standard deviation, and 95% confidence interval. Comprehensive analysis includes confusion matrix, error analysis, learning curve, and performance comparison per entity length. Results show that CRF achieves an F1-score of 78.49% (P=82.37%, R=74.95%), while BiLSTM achieves 76.54% $\pm$ 1.32% (95% CI $\pm$ 1.15%). CRF outperforms BiLSTM by 1.95 percentage points under baseline conditions, although BiLSTM demonstrates notably fewer false positives (FP=124 vs FP=153) and a standard deviation of 1.32% across 5 experimental seeds.

Keywords: Named Entity Recognition, CRF, BiLSTM, NCBI Disease, Biomedical, NLP.

1. INTRODUCTION

The rapid advancement of artificial intelligence and Natural Language Processing (NLP) has opened significant opportunities for automated biomedical text analysis [1]. One of the core tasks in NLP is Named Entity Recognition (NER), which involves identifying and classifying named entities such as disease names, genes, proteins, and drugs from unstructured text [2]. In the biomedical domain, the ability to accurately recognize disease entities is crucial for supporting downstream applications such as relation extraction, medical knowledge base construction, and clinical decision support systems [3].

The NCBI Disease dataset is one of the most widely used standard benchmarks in biomedical NER research [4]. It consists of PubMed article abstracts manually annotated with disease entities using the BIO (Beginning-Inside-Outside) tagging scheme. A key characteristic of this dataset is the significant class imbalance between disease entity tokens and non-entity tokens.

Two of the most widely studied approaches in NER are the statistical Conditional Random Field (CRF) model and the deep learning-based Bidirectional Long Short-Term Memory (BiLSTM) model. CRF is a discriminative probabilistic model proven effective for sequential labeling due to its ability to model dependencies between adjacent labels [5][6]. BiLSTM, on the other hand, can automatically learn feature representations from raw data and exploit contextual information from both directions within a sequence [7].

Several prior studies have compared these two models across various domains and datasets [8][9][10]. However, systematic comparisons that include multi-seed validation with comprehensive statistical reporting on the NCBI Disease dataset remain limited. Most studies report results from a single execution without considering inter-run variance, making reliability difficult to assess [11].

This study aims to : compare the performance of CRF and BiLSTM models on the disease NER task using the NCBI Disease dataset, conduct robust statistical validation through multi-seed experiments (5 seeds) on BiLSTM with reporting of mean, standard deviation, and 95% confidence interval, and conduct an in-depth analysis of the strengths and limitations of each model across multiple evaluation dimensions including confusion matrix, error analysis, learning curve, and performance analysis by entity length [12][13].

Unlike previous studies that primarily report single-run evaluation results, this study provides a more comprehensive comparison between CRF and BiLSTM on the NCBI Disease dataset through multi-seed validation, confidence interval reporting, confusion matrix analysis, entity-level error analysis, and learning curve evaluation. These analyses provide a more reliable assessment of model performance and stability in biomedical disease named entity recognition.

2. LITERATURE REVIEW

2.1 Named Entity Recognition (NER)

Named Entity Recognition (NER) is a sequence labeling task that aims to identify text spans referring to specific entities and classify them into predefined categories [2]. In the biomedical context, NER is a critical component in information extraction pipelines because medical texts contain dense and complex specialized terminology [13]. NER approaches have evolved from rule-based methods, statistical machine learning, to deep learning [14].

2.2 Conditional Random Field (CRF)

Conditional Random Field (CRF) was introduced as a discriminative probabilistic model for sequential data labeling [6]. CRF directly models the conditional distribution $P(y|x)$, enabling it to consider arbitrary features from the entire input sequence without strict independence assumptions. In NER, linear-chain CRF is widely used for its ability to capture dependencies between adjacent labels [15]. CRF model quality relies heavily on manually engineered features [16]. Key advantages of CRF include computational efficiency, no GPU requirement, and deterministic outputs that are easy to reproduce.

2.3 Bidirectional LSTM (BiLSTM)

Long Short-Term Memory (LSTM) is a variant of Recurrent Neural Networks designed to address the vanishing gradient problem in long sequences [7]. Bidirectional LSTM (BiLSTM) extends LSTM by processing sequences from both directions simultaneously, so each token representation captures context from both sides [17]. The combination of BiLSTM with pre-trained word embeddings such as GloVe has proven to deliver competitive performance on NER tasks [18]. Unlike CRF, BiLSTM is non-deterministic due to random weight initialization and dropout, making multi-seed evaluation necessary for statistically valid reporting [11].

2.4 NCBI Disease Dataset

The NCBI Disease dataset was developed by Doğan et al. [4] and consists of PubMed abstracts annotated with disease entities. The dataset is divided into training, development/validation, and test splits with BIO tagging: B-Disease (entity beginning), I-Disease (entity continuation), and O (non-entity). It serves as a standard evaluation benchmark for biomedical NER models due to its high annotation quality [19].

2.5 GloVe Word Embeddings

Global Vectors for Word Representation (GloVe) is a word representation learning method that combines the advantages of global co-occurrence matrix-based methods [20]. GloVe 200-dimensional representations trained on large corpora have proven effective as embedding initializations for various NLP tasks including biomedical NER [21]. Using pre-trained embeddings allows models to leverage semantic knowledge from large amounts of unlabeled data [22].

2.6 Related Work

Several studies have compared CRF and BiLSTM models for NER. Ramachandran and Arutchelvan [8] demonstrated the benefits of machine learning-based approaches on biomedical texts. Ramos-Vargas et al. [9] conducted a comprehensive evaluation of pre-trained models for biomedical NER and found that contextual representations yield significant improvements. Liu et al. [10] investigated knowledge distillation for biomedical NER with high efficiency. Moscato et al. [23] proposed a generative approach for biomedical entity linking that complements the NER pipeline. Studies that specifically compare pure CRF versus BiLSTM on the NCBI Disease dataset with standardized multi-seed validation remain scarce [24][25][26].

3. RESEARCH METHODOLOGY

3.1 Dataset and Preprocessing

This study uses the NCBI Disease dataset in CSV format with three splits: train.csv, validation.csv, and test.csv. Each row contains one sentence with a tokens column

(string array) and ner_tags column (integer label array). Integer labels are converted to IOB2 format: 0=O, 1=B-Disease, 2=I-Disease.

TABLE 1.
Disease Dataset Statistics

Split	Sentence	Total Token	Entity Token
Training	5.393	134.796	11.173
Validation	919	23.803	1.862
Test	938	24.430	2.027
Total	7.250	183.029	15.062

The dataset follows the official train-validation-test split provided by the NCBI Disease corpus.

3.2 CRF Model

The CRF model was implemented using the sklearn-crfsuite library with the L-BFGS algorithm (c1=0.1, c2=0.1, max_iterations=100, all_possible_transitions = True). Extracted features include 20 feature types per token: word representations (lowercase, isupper, istitle, isdigit), compressed word shape, POS tag with its first 2 characters, boolean flags (is_alpha, has_digit, has_hyphen), prefix/suffix of 1–4 characters, contextual features of neighboring tokens ± 1 , and BOS/EOS markers [6][16]. The CRF model is deterministic and thus requires only a single training run.

3.3 BiLSTM Model

The BiLSTM model was implemented using PyTorch. The architecture consists of an embedding layer of 200 dimensions initialized with GloVe 6B.200d and fine-tuned during training, two BiLSTM layers with hidden size 128 per direction (total output 256), dropout rate 0.2, and a fully connected layer Linear (256→3) [7][18]. Training uses the Adam optimizer (lr=0.001, weight_decay=0.0), CrossEntropyLoss with padding masking, gradient clipping (max_norm=5.0), and early stopping (patience=5) based on validation F1-score with a maximum of 20 epochs. Complete hyperparameters are presented in Table 2.

TABLE 2.
BiLSTM Model Hyperparameters

Hyperparameter	Value
Embedding Dimension	200 (GloVe 6B.200d, fine-tuned)
Hidden Dimension	$128 \times 2 = 256$
Num Layers	2
Dropout	0.2
Learning Rate	0.001 (Adam)
Weight Decay	$1e-5$
Batch Size	8
Max Epochs	15
Early Stopping Patience	5
Gradient Clipping	max_norm = 5.0

3.4 Evaluation Protocol

Evaluation uses entity-level Precision, Recall, and F1-score via the seqeval library following CoNLL standards [2][19]. The CRF model is evaluated once (deterministic). The BiLSTM model is evaluated 5 times with different seeds; results are reported as mean $\pm$ std and 95% confidence interval ($CI = 1.96 \times \text{std} / \sqrt{5}$). Additional analysis includes : token-level confusion matrix, entity error analysis (TP/FP/FN), performance analysis by entity length (single/multi-2/multi-3+), and learning curve on data proportions from 20% to 100% [27].

4. RESULTS AND DISCUSSION

4.1 Exploratory Data Analysis (EDA)

Prior to modeling, exploratory analysis was conducted to understand the dataset characteristics. The NCBI Disease dataset consists of 7,250 sentences total (train= 5,393, validation= 919, test= 938) with 134,796 tokens in the training set. The class imbalance ratio is 11.06:1 (123,623 non-entity tokens vs 11,173 entity tokens), which is more moderate than initially expected. Figure 1 shows the sentence length distribution across each split, which are relatively similar. Figure 2 displays the BIO label distribution in pie charts confirming the dominance of O tokens. Figure 3 shows the disease entity length distribution on the training set entity lengths range from 1 to 22 tokens with a mean of 2.19 tokens per entity, indicating many multi-token entities that present a distinctive challenge for NER models.

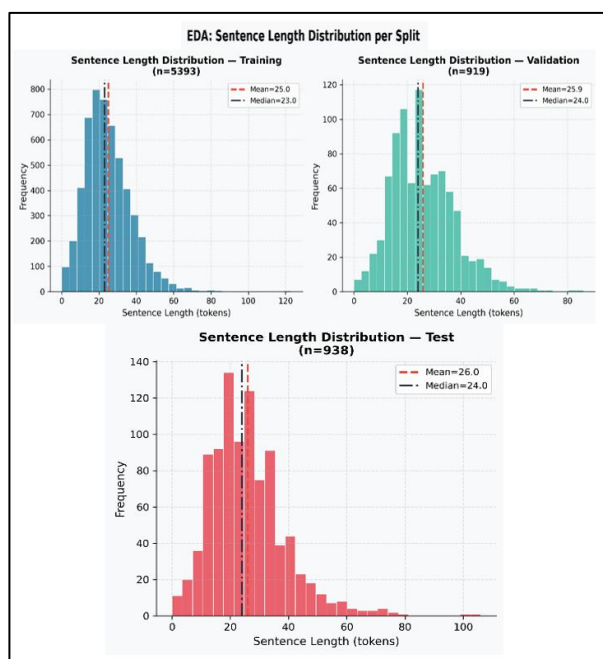


FIGURE 1. Sentence Length Distribution per Split (Training, Validation, Test)

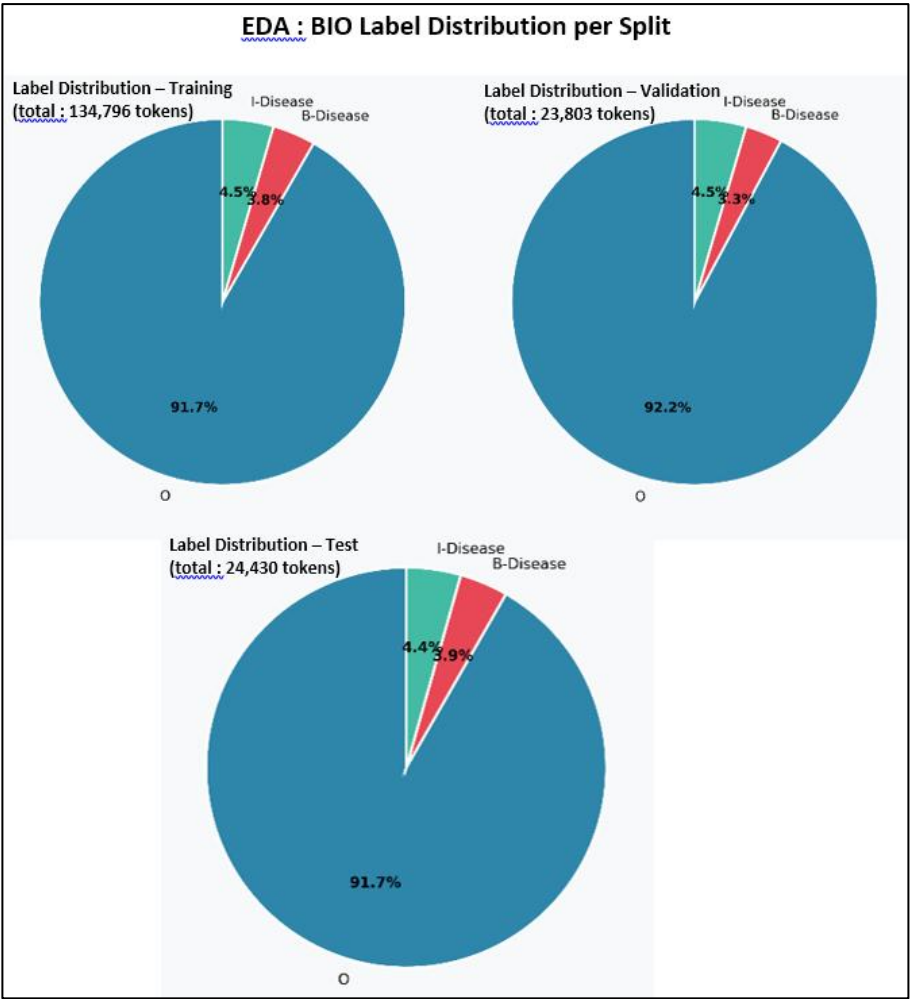


FIGURE 2. BIO Label Distribution per Split (Pie Chart)

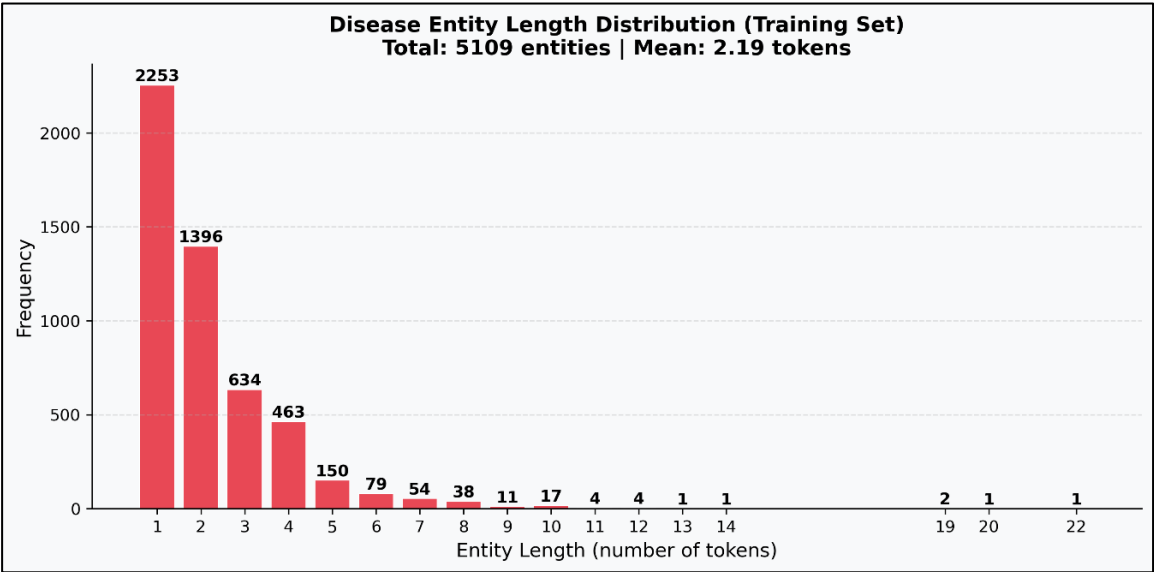


FIGURE 3. Disease Entity Length Distribution on Training Set

4.2 CRF Model Results

The CRF model was trained once on the full training set (5,393 sentences). Table 3 summarizes the CRF evaluation results on the test set. Being deterministic, no repetition of experiments is required. CRF training time is remarkably short at 9.44 seconds without GPU, making it an attractive choice in resource-constrained scenarios. The constructed vocabulary covers 9,712 words with a GloVe coverage of 78.2% (7,594 out of 9,712 words).

TABLE 3.
CRF Model Evaluation Results on NCBI Disease Test Set

Model	Precision (%)	Recall (%)	F1-Score (%)
CRF	82.37	74.95	78.49

4.3 BiLSTM Model Results (5 Seeds)

The BiLSTM model was run 5 times with different seeds using GloVe 200d with 78.2% coverage on a vocabulary of 9,712 words. Figure 4 shows the training loss and validation F1-score per epoch for all five seeds, illustrating the convergence behavior of each run. Seed 1024 converged fastest at 9 epochs, while seeds 123 and 789 stopped at 14 epochs. Seed 42 stopped at 19 epochs, and seed 456 reached the full 20 epochs without triggering early stopping, indicating slower convergence. The best seed is 42 with F1=78.24%, while the weakest is seed 789 with F1=74.77%. Figures 5 and 6 visualize the F1-score distribution across seeds. Table 4 presents the complete per-seed results along with summary statistics.

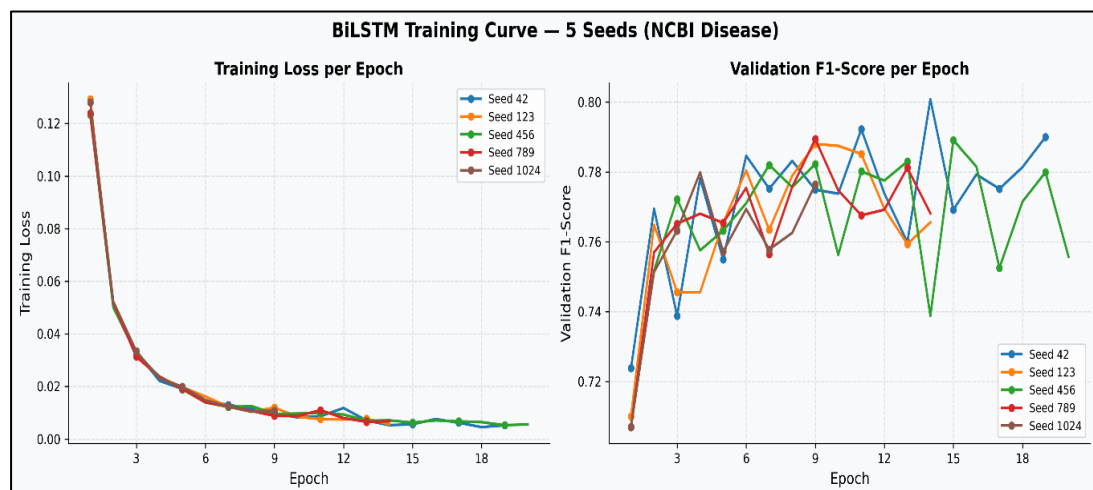


FIGURE 4. Training Loss and Validation F1-Score per Epoch (5 Seeds)

Agus Siswanto, Muhammad Wardani, Yulia Arvita, Jasmir
Comparison of CRF and BiLSTM Models for Disease Named Entity Recognition on
the NCBI Disease Dataset

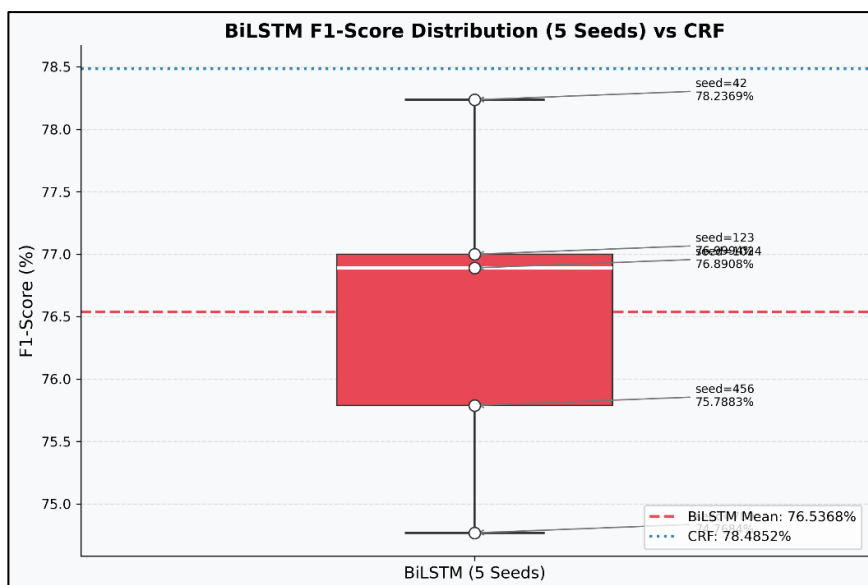


FIGURE 5. Boxplot of BiLSTM F1-Score Distribution (5 Seeds) vs CRF

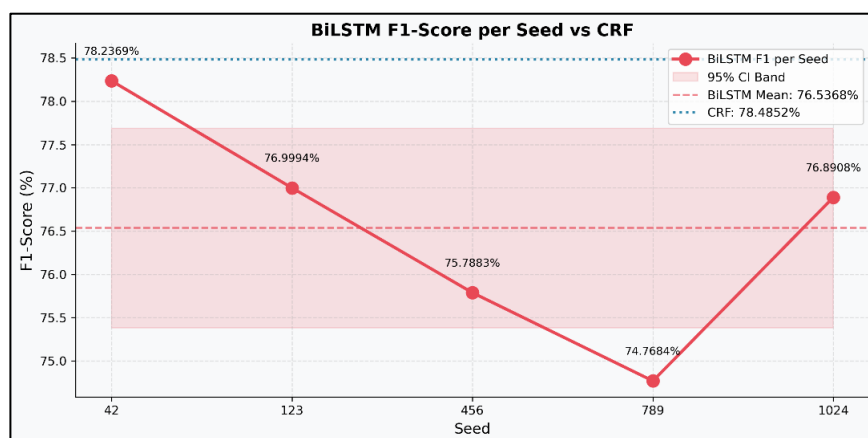


FIGURE 6. Line Plot of BiLSTM F1-Score per Seed with 95% CI Band

TABLE 4.
BiLSTM Model Evaluation Results per Seed on NCBI Disease Test Set

Seed	Precision (%)	Recall (%)	F1-Score (%)	Time (s)	Epochs
42	82.4623	74.4235	78.2369	91.4	19
123	81.2573	73.1656	76.9994	69.3	14
456	81.8735	70.5451	75.7883	101.7	20
789	77.8661	71.9078	74.7684	73.1	14
1024	77.0526	76.7296	76.8908	50.9	9
Mean $\pm$ Std	80.1024 $\pm$ 2.4668	73.3543 $\pm$ 2.3747	76.5368 $\pm$ 1.3150	77.3 $\pm$ 19.8	-
95% CI	$\pm$ 2.1617	$\pm$ 2.0825	$\pm$ 1.1527	-	-

4.4 Comprehensive Comparison: CRF vs BiLSTM

Figure 7 visually presents the F1-score comparison between both models, while Figure 8 shows all three metrics (Precision, Recall, F1) in a grouped bar chart. Figure 9 visualizes the performance profile of both models in a radar chart to highlight

strengths across each metric dimension. Figure 10 summarizes all values in a heatmap format for at-a-glance comparison. Table 5 presents a comprehensive comparison across multiple aspects.

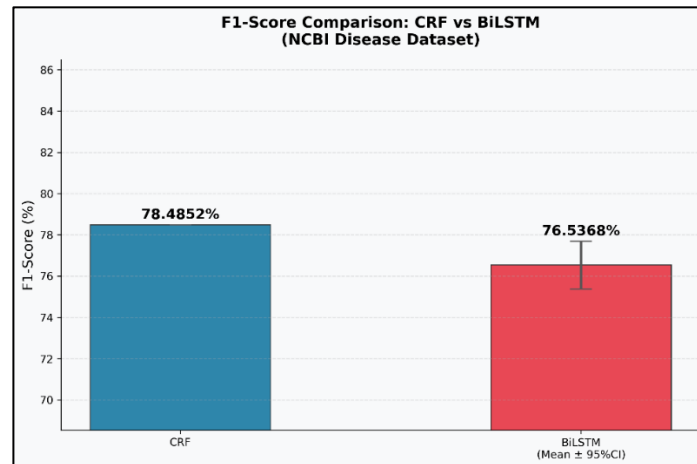


FIGURE 7. F1-Score Comparison: CRF vs BiLSTM (Mean $\pm$ 95%CI)

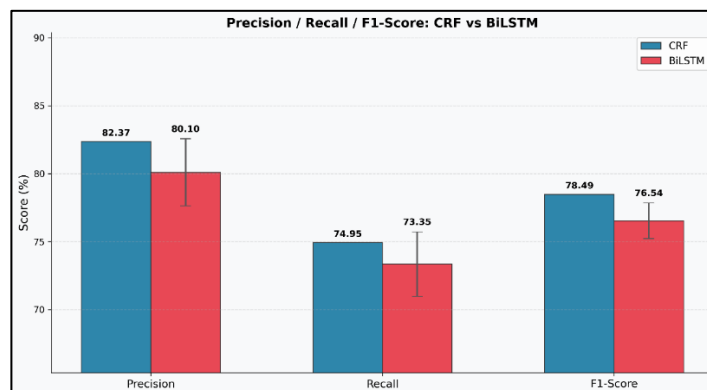


FIGURE 8. Precision, Recall, and F1-Score Comparison: CRF vs BiLSTM

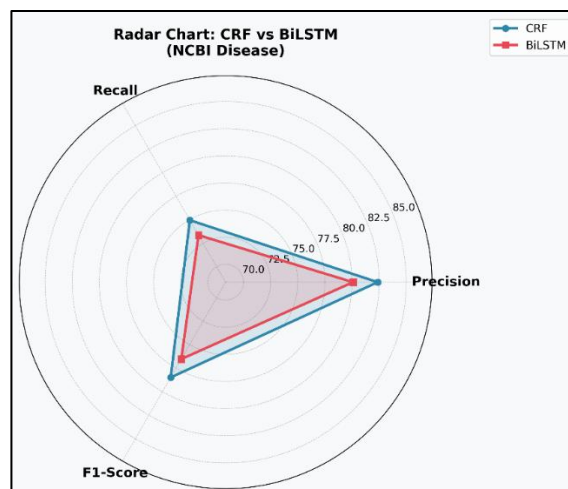


FIGURE 9. Radar Chart of P/R/F1 Comparison: CRF vs BiLSTM

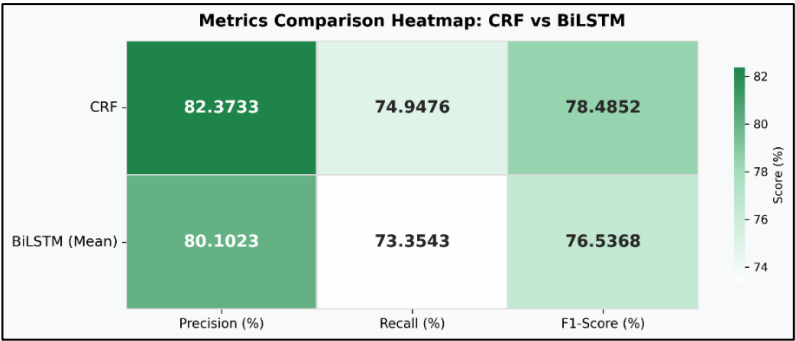


FIGURE 10. Metrics Comparison Heatmap: CRF vs BiLSTM

TABLE 5.
Comprehensive Comparison of CRF vs BiLSTM on NCBI Disease

Aspect	CRF	BiLSTM
Precision (%)	82.3733	80.1024 ± 2.4668
Recall (%)	74.9476	73.3543 ± 2.3747
F1-Score (%)	78.4852	76.5368 ± 1.3150
95% CI F1	N/A (deterministic)	±1.1527%
Δ F1 (CRF – BiLSTM)	+1.9484 pp	Winner: CRF
Number of Experiments	1 run	5 seeds
Feature Engineering	Manual (20 features/token)	Automatic (GloVe 200d)
Training Time	9.44 seconds	77.3 seconds (mean)
GPU Requirement	No	Recommended
Model Nature	Deterministic	Stochastic
Interpretability	High	Low (black box)
Stability (Std F1)	N/A	1.3150%

4.5 Confusion Matrix Analysis

Figure 11 shows the token-level confusion matrix for both models. The main diagonal represents correct predictions, while off-diagonal elements indicate classification errors. CRF tends to generate more false positives (predicting O tokens as entities), while BiLSTM generates fewer false alarms at the cost of missing more entities. This difference in error patterns reflects the distinct learning strategies of both models.

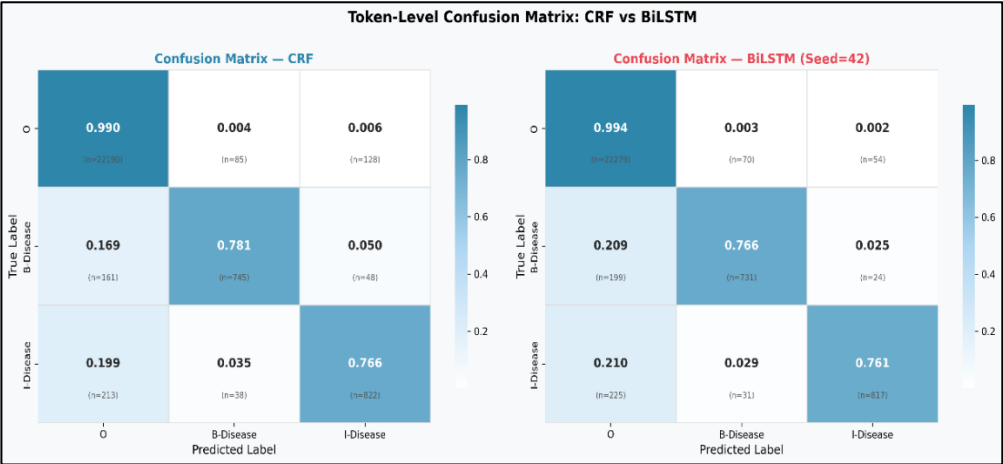


FIGURE 11. Token-Level Confusion Matrix: CRF vs BiLSTM

Figure 12 presents the per-seed metrics heatmap for BiLSTM, demonstrating performance consistency across different random initializations. The low variance across seeds confirms that BiLSTM training is stable and the results are reliable for deployment.

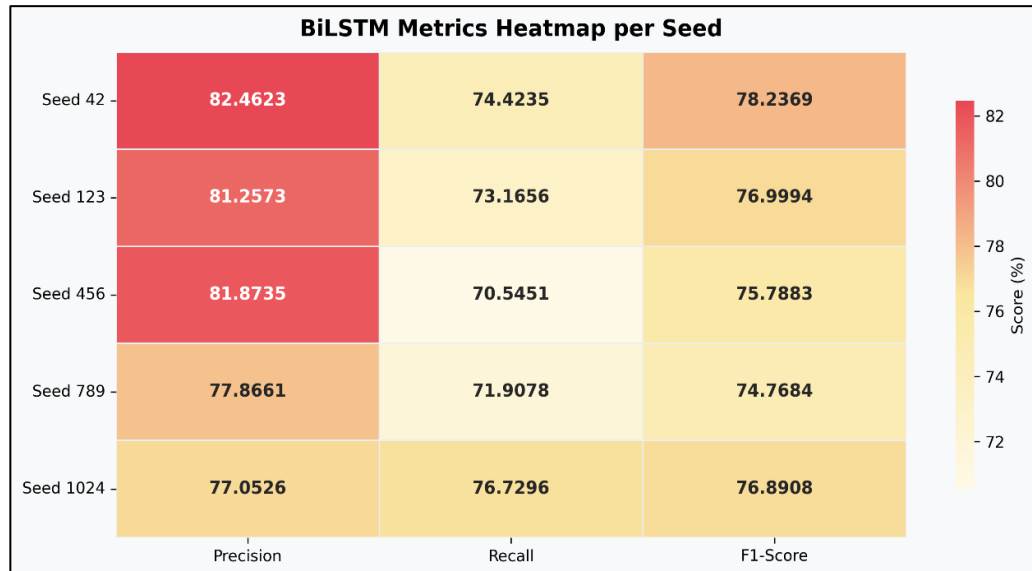


FIGURE 12. BiLSTM Precision-Recall-F1 Heatmap per Seed

4.6 Entity Error Analysis

Figure 13 presents entity-level error analysis. CRF yields TP=715, FP=153, FN=239, while BiLSTM yields TP=708, FP=124, FN=246. A key finding is that BiLSTM produces fewer false positives (FP=124 vs FP=153, a difference of 29 predictions), meaning BiLSTM is more conservative and precise in its predictions. However, BiLSTM misses slightly more entities (FN=246 vs FN=239), indicating marginally lower recall. CRF detects more true entities (TP=715 vs TP=708) but at the cost of more false alarms. This trade-off explains why CRF achieves higher overall F1, though BiLSTM's lower FP rate is advantageous in precision-critical applications.

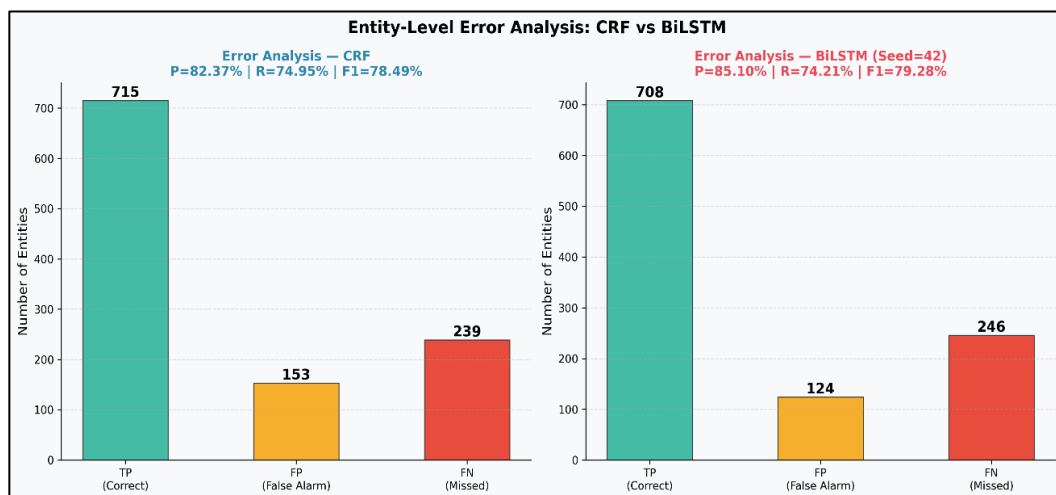


FIGURE 13. Entity-Level Error Analysis (TP/FP/FN): CRF vs BiLSTM

Figure 14 presents performance by entity length category: single (1 token), multi-2 (2 tokens), and multi-3+ (3 or more tokens). This analysis is important as NER models generally struggle more with multi-token entities that require accurate boundary detection.

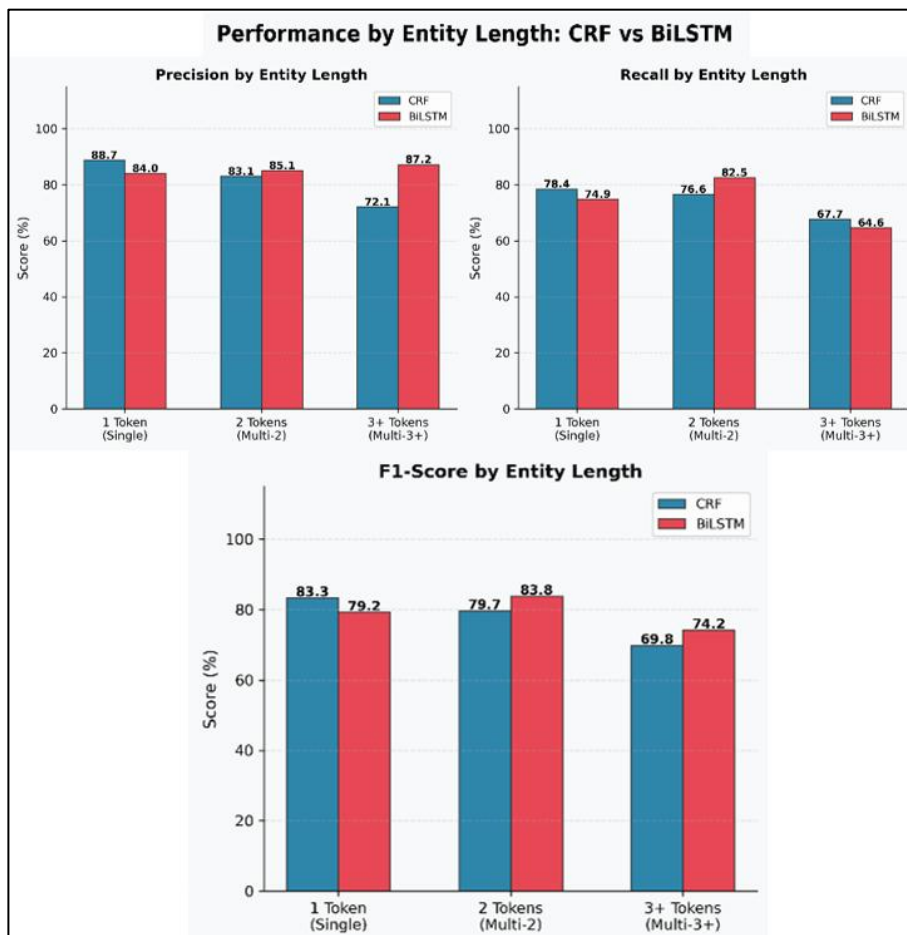


FIGURE 14. Performance by Entity Length: CRF vs BiLSTM

4.7 Learning Curve

Figure 15 shows the learning curve of CRF and BiLSTM at various training data proportions (20%–100%). At 20% data (1,078 sentences), CRF achieves F1=69.69% vs BiLSTM=66.74%, a gap of 2.95 pp. As data increases, both models improve consistently: at 40% (CRF=74.30% vs BiLSTM=72.21%), 60% (CRF=76.57% vs BiLSTM=74.74%), 80% (CRF=77.05% vs BiLSTM=75.87%), and 100% (CRF=78.49% vs BiLSTM=77.62%). Notably, the performance gap narrows consistently from 2.95 pp at 20% data to 1.18 pp at 80% data and further to 0.87 pp at 100%. This trend strongly suggests that BiLSTM has the potential to match or surpass CRF with larger training data. Figure 16 shows training time CRF takes only 9.44 seconds vs BiLSTM mean of 77.3 seconds, approximately 8× faster.

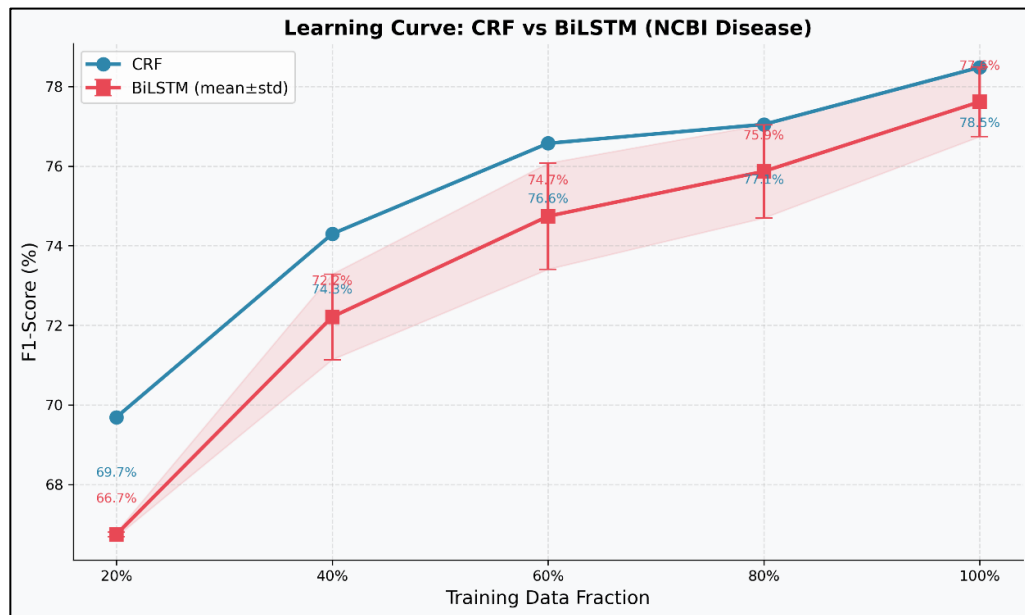


FIGURE 15. Learning Curve: CRF vs BiLSTM (Training Data Proportion 20%–100%)

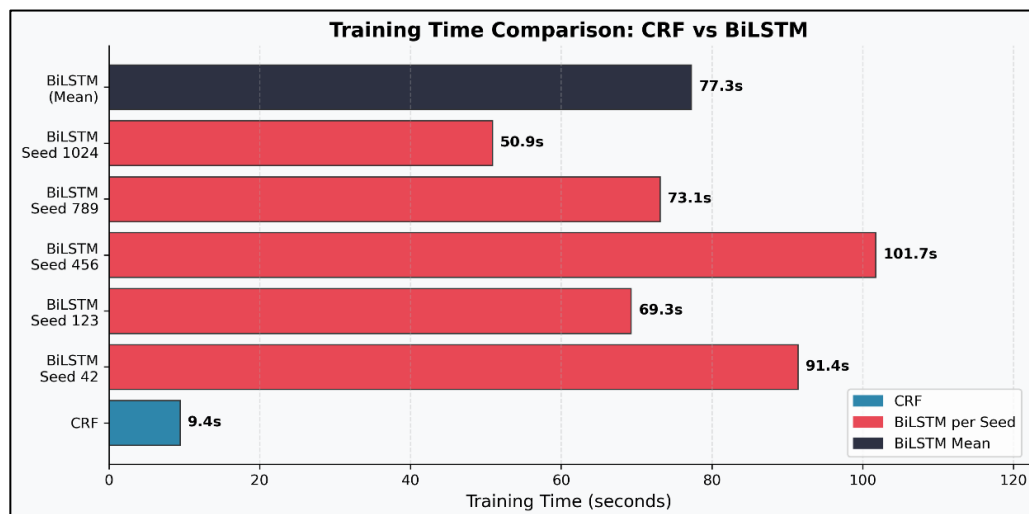


FIGURE 16. Training Time Comparison: CRF vs BiLSTM per Seed

4.8 Discussion

The superior performance of CRF may be attributed to the relatively limited size of the NCBI Disease dataset and the effectiveness of handcrafted linguistic features such as prefixes, suffixes, POS tags, and word shape features. In contrast, BiLSTM relies heavily on learning feature representations from data, which may require larger training corpora to fully exploit its representation learning capability.

Furthermore, the GloVe embedding coverage of 78.2% indicates that a portion of biomedical terminology is not represented in the pretrained embedding space, potentially limiting BiLSTM performance on domain-specific disease entities.

4.9 Limitations

This study has several limitations. First, experiments were conducted using only a single benchmark dataset, namely the NCBI Disease dataset. Second, only GloVe embeddings were utilized for the BiLSTM model. Third, more advanced contextual language models such as BioBERT and PubMedBERT were not included in the comparison. Therefore, the findings should be interpreted within the scope of the evaluated models and dataset.

5. CONCLUSION

This study compared CRF and BiLSTM models for disease Named Entity Recognition on the NCBI Disease dataset (7,250 sentences, 134,796 training tokens, IR=11.06:1) using a standardized evaluation protocol. The CRF model achieved an F1-score of 78.49% (P=82.37%, R=74.95%) in just 9.44 seconds without GPU. The BiLSTM model achieved a mean F1-score of 76.54% $\pm$ 1.32% (95% CI $\pm$ 1.15%) across 5 seeds, with the best seed being 42 (F1=78.24%) and the weakest being seed 789 (F1=74.77%). CRF outperforms BiLSTM by 1.95 percentage points under baseline conditions.

In-depth analysis reveals several important findings. First, CRF achieves higher overall F1 (78.49% vs 76.54%) driven by higher Precision (82.37% vs 80.10%) and more true positive detections (TP=715 vs TP=708). Second, BiLSTM demonstrates a lower false positive rate (FP=124 vs FP=153), making it more conservative and less prone to spurious predictions — a desirable property in precision-critical biomedical applications. Third, the learning curve shows the performance gap narrowing consistently from 2.95 pp at 20% data to 0.87 pp at 100% data, strongly indicating The observed learning curve trend suggests that BiLSTM may continue to benefit from larger training datasets and has the potential to reduce the performance gap with CRF. Fourth, CRF is approximately 8 $\times$ faster (9.44s vs 77.3s mean) without requiring GPU hardware. For future research, it is recommended to explore :

1. Hybrid BiLSTM-CRF models that combine the representation learning capability of BiLSTM with the label transition modeling of CRF.
2. Contextual pre-trained language models such as BioBERT or PubMedBERT which have shown substantial gains on biomedical NER benchmarks.
3. Data augmentation techniques to address class imbalance, which are expected to particularly benefit BiLSTM given the observed trend of gap narrowing in the learning curve.

ACKNOWLEDGEMENTS

The authors would like to express their sincere gratitude to all parties who have contributed to the completion of this research. We especially thank the institution and colleagues for their support, guidance, and valuable resources throughout the research process. We also appreciate the constructive feedback from reviewers, which has significantly improved the quality of this work. Finally, we acknowledge all individuals who have directly or indirectly contributed to this study.

REFERENCES

- [1] L. B. Elvas, A. Almeida, and J. C. Ferreira, "Natural language processing in medical text processing: A scoping literature review," *Int. J. Med. Inf.*, vol. 204, p. 106049, Dec. 2025, doi: 10.1016/j.ijmedinf.2025.106049.
- [2] Z. Nasar, S. W. Jaffry, and M. K. Malik, "Named Entity Recognition and Relation Extraction: State-of-the-Art," *ACM Comput. Surv.*, vol. 54, no. 1, pp. 1-39, Jan. 2022, doi: 10.1145/3445965.
- [3] N. Goyal and N. Singh, "Named entity recognition and relationship extraction for biomedical text: A comprehensive survey, recent advancements, and future research directions," *Neurocomputing*, vol. 618, p. 129171, Feb. 2025, doi: 10.1016/j.neucom.2024.129171.
- [4] R. Islamaj Dogan et al., "Overview of the BioCreative VI Precision Medicine Track: mining protein interactions and mutations for precision medicine," *Database*, vol. 2019, Jan. 2019, doi: 10.1093/database/bay147.
- [5] M. Altalhan, A. Algarni, and M. Turki-Hadj Alouane, "Imbalanced Data Problem in Machine Learning: A Review," *IEEE Access*, vol. 13, pp. 13686-13699, 2025, doi: 10.1109/access.2025.3531662.
- [6] S. Singh and U. S. Tiwary, "ACRF: Aggregated Conditional Random Field for Out of Vocab (OOV) Token Representation for Hindi NER," *IEEE Access*, vol. 12, pp. 22707-22717, 2024, doi: 10.1109/access.2024.3362645.
- [7] S. Arslan, "Application of BiLSTM-CRF model with different embeddings for product name extraction in unstructured Turkish text," *Neural Comput. Appl.*, vol. 36, no. 15, pp. 8371-8382, May 2024, doi: 10.1007/s00521-024-09532-1.
- [8] R. Ramachandran and K. Arutchelvan, "Named entity recognition on bio-medical literature documents using hybrid based approach," *J. Ambient Intell. Humaniz. Comput.*, Mar. 2021, doi: 10.1007/s12652-021-03078-z.
- [9] R. E. Ramos-Vargas, I. Roman-Godinez, and S. Torres-Ramos, "Comparing general and specialized word embeddings for biomedical named entity recognition," *PeerJ Comput. Sci.*, vol. 7, p. e384, Feb. 2021, doi: 10.7717/peerj-cs.384.
- [10] J. Liu et al., "A hybrid deep-learning approach for complex biochemical named entity recognition," *Knowl.-Based Syst.*, vol. 221, p. 106958, Jun. 2021, doi: 10.1016/j.knosys.2021.106958.
- [11] M. Wozniak, M. Wiczorek, and J. Silka, "BiLSTM deep neural network model for imbalanced medical data of IoT systems," *Future Gener. Comput. Syst.*, vol. 141, pp. 489-499, Apr. 2023, doi: 10.1016/j.future.2022.12.004.
- [12] K. Pakhale, "Comprehensive Overview of Named Entity Recognition: Models, Domain-Specific Applications and Challenges," Sep. 25, 2023, arXiv: arXiv:2309.14084. doi: 10.48550/arXiv.2309.14084.
- [13] Y. Park, G. Son, and M. Rho, "Biomedical Flat and Nested Named Entity Recognition: Methods, Challenges, and Advances," *Appl. Sci.*, vol. 14, no. 20, p. 9302, Oct. 2024, doi: 10.3390/app14209302.
- [14] J. Sun, Y. Liu, J. Cui, and H. He, "Deep learning-based methods for natural hazard named entity recognition," *Sci. Rep.*, vol. 12, no. 1, p. 4598, Mar. 2022, doi: 10.1038/s41598-022-08667-2.

- [15] Y. An, X. Xia, X. Chen, F.-X. Wu, and J. Wang, "Chinese clinical named entity recognition via multi-head self-attention based BiLSTM-CRF," *Artif. Intell. Med.*, vol. 127, p. 102282, May 2022, doi: 10.1016/j.artmed.2022.102282.
- [16] N. Alshammari and S. Alanazi, "The impact of using different annotation schemes on named entity recognition," *Egypt. Inform. J.*, vol. 22, no. 3, pp. 295-302, Sep. 2021, doi: 10.1016/j.eij.2020.10.004.
- [17] H. Pooja and M. P. P. Jagadeesh, "A Deep Learning Based Approach for Biomedical Named Entity Recognition Using Multitasking Transfer Learning with BiLSTM, BERT and CRF," *SN Comput. Sci.*, vol. 5, no. 5, Apr. 2024, doi: 10.1007/s42979-024-02835-z.
- [18] R. P. Kusumawardani and K. N. Kusumawati, "Named entity recognition in the medical domain for Indonesian language health consultation services using bidirectional-lstm-crf algorithm," *Procedia Comput. Sci.*, vol. 245, pp. 1146-1156, 2024, doi: 10.1016/j.procs.2024.10.344.
- [19] S. M. Archana and J. Prakash, "Biomedical named entity recognition through improved balanced undersampling for addressing class imbalance and preserving contextual information," *Int. J. Inf. Technol.*, vol. 16, no. 8, pp. 4995-5003, Dec. 2024, doi: 10.1007/s41870-024-02137-w.
- [20] E. H. Houssein, R. E. Mohamed, and A. A. Ali, "Machine Learning Techniques for Biomedical Natural Language Processing: A Comprehensive Review," *IEEE Access*, vol. 9, pp. 140628-140653, 2021, doi: 10.1109/access.2021.3119621.
- [21] A. Dash, S. Darshana, D. K. Yadav, and V. Gupta, "A clinical named entity recognition model using pretrained word embedding and deep neural networks," *Decis. Anal. J.*, vol. 10, p. 100426, Mar. 2024, doi: 10.1016/j.dajour.2024.100426.
- [22] D. Vithanage, P. Yu, L. Wang, and C. Deng, "Contextual Word Embedding for Biomedical Knowledge Extraction: a Rapid Review and Case Study," *J. Healthc. Inform. Res.*, vol. 8, no. 1, pp. 158-179, Mar. 2024, doi: 10.1007/s41666-023-00157-y.
- [23] V. Moscato, M. Postiglione, G. Sperli, and A. Vignali, "ALDANER: Active Learning based Data Augmentation for Named Entity Recognition," *Knowl.-Based Syst.*, vol. 305, p. 112682, Dec. 2024, doi: 10.1016/j.knosys.2024.112682.
- [24] B. Jehangir, S. Radhakrishnan, and R. Agarwal, "A survey on Named Entity Recognition - datasets, tools, and methodologies," *Nat. Lang. Process. J.*, vol. 3, p. 100017, Jun. 2023, doi: 10.1016/j.nlp.2023.100017.
- [25] K. Bose and K. Sarkar, "Named Entity Recognition in Bengali and Hindi Using MuRIL and Conditional Random Fields," *SN Comput. Sci.*, vol. 5, no. 7, Sep. 2024, doi: 10.1007/s42979-024-03211-7.
- [26] H. Chen, L. Dan, Y. Lu, M. Chen, and J. Zhang, "An improved data augmentation approach and its application in medical named entity recognition," *BMC Med. Inform. Decis. Mak.*, vol. 24, no. 1, Aug. 2024, doi: 10.1186/s12911-024-02624-x.
- [27] S. M. Archana, J. Prakash, P. K. Singh, and W. Ahmed, "An Effective Biomedical Named Entity Recognition by Handling Imbalanced Data Sets Using Deep Learning and Rule-Based Methods," *SN Comput. Sci.*, vol. 4, no. 5, Aug. 2023, doi: 10.1007/s42979-023-02068-6.