**Indonesian Journal of Electrical Engineering and Computer Science**
Vol. 35, No. 3, September 2024, pp. 1722~1731
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v35.i3.pp1722-1731 

# Clinical named entity extraction for extracting information from medical data

**Dhanasekaran Kuttaiyapillai[1], Anand Madasamy[1], Shobanadevi Ayyavu[1], Md Shohel Sayeed[2]**
[1]Department of Data Science and Business Systems, School of Computing, CET, SRM Institute of Science and Technology, Kattankulathur, India
[2]Faculty of Information Science and Technology (FIST), Multimedia University, Bukit Beruang, Malaysia

**Article Info**

***Article history:***

Received Nov 30, 2023
Revised Apr 25, 2024
Accepted May 7, 2024

***Keywords:***

Clinical data analysis
Deep learning
Information extraction
Medical data analytics
Named entity extraction

**ABSTRACT**

Clinical named entity extraction (NER) based on deep learning gained much attention among researchers and data analysts. This paper proposes a NER approach to extract valuable Parkinson's disease-related information. To develop an effective NER method and to handle problems in disease data analytics, a unique NER technique applies a "recognize-map-extract (RME)" mechanism and aims to deal with complex relationships present in the data. Due to the fast-growing medical data, there is a challenge in the development of suitable deep-learning methods for NER. Furthermore, the traditional machine learning approaches rely on the time-consuming process of creating corpora and cannot extract information for specific needs and locations in certain situations. This paper presents a clinical NER approach based on a convolutional neural network (CNN) for better use of specific features around medical entities and analyzes the performance of the proposed approach through fine-tuning NER with effective pre-training on the BC5CDR dataset. The proposed method uses annotation of entities for various medical concepts. The second stage develops a clinically NER method. This proposed method shows interesting results on the performance measures achieving a precision of 92.57%, recall of 92.22%, and F1-measure of 91.6%.



***Corresponding Author:***

Md Shohel Sayeed
Faculty of Information Science and Technology (FIST), Multimedia University
Jalan Ayer Keroh Lama, Bukit Beruang, 75450 Melaka, Malaysia
Email: shohel.sayeed@mmu.edu.my

## 1. INTRODUCTION

Named entity recognition provides a way to find insights from unstructured medical data. Due to the challenges in handling new disease cases, medical information extraction demands urgent solutions to address disease-related problems. Traditional methods rely on the time-consuming process of creating a large dictionary, rule-set, or clean-structured corpora. The new deep learning need not depend upon hand-crafted features. Although conditional random fields (CRF) capture correlations between nearest class labels, the existing methods do not make use of complete contextual information in the hidden layers [1].

Medical people nowadays have the option to utilize artificial intelligence technology for making informed decisions and to provide evidence-based medicine. Moreover, intelligent systems are trying to derive valuable insights for disease monitoring. To deal with the ambiguity, the named entity recognition approach proposed a fully self-attentive mechanism [2] and integrated the encoding vector with the contextual information. While extracting features, instead of CRF, it has used multivariate convolutional

decoding on hidden vectors to associate the current word at a specified position with neighbouring words of the sentence.

Named entity extraction (NER) requires a method to solve polysemy problems on financial information. In this direction, Xin and Xiaoyan [3] proposed an ERNIE-Doc-Bi-LSTM-CRF. This pre-trained model shows the accuracy of 86.72%, the recall 83.39%, and the F1-score 85.02%. Figure 1 shows example named entities.


```
hibition of thrombosis by 55 % and 35 % , respectively , while acetylsalicylic acid ( ASA , 50 mg /
kg , i . p . ) , a positive control , showed only 30 % inhibition .', {'entities': [(115, 125, 'B_Di
sease')]}], ['In the vitro human platelet aggregations induced by the agonists used in tests , TET a
nd FAN showed the inhibitions dose dependently .', {'entities': [(19, 27, 'B_Disease'), (28, 40, 'I_
Disease')]}], ['Angioedema due to ACE inhibitors : common and inadequately diagnosed .', {'entitie
s': [(0, 10, 'B_Disease')]}], ['The estimated incidence of angioedema during angiotensin - convertin
g enzyme ( ACE ) inhibitor treatment is between 1 and 7 per thousand patients .', {'entities': [(27,
37, 'B_Disease')]}], ['Cocaine - induced mood disorder : prevalence rates and psychiatric symptoms i
n an outpatient cocaine - dependent sample .', {'entities': [(18, 22, 'B_Disease'), (23, 31, 'I_Dise
ase'), (55, 66, 'B_Disease')]}], ['This paper attempts to examine and compare prevalence rates and s
ymptom patterns of DSM substance - induced and other mood disorders .', {'entities': [(118, 122, 'B_
Disease'), (123, 132, 'I_Disease')]}], ['243 cocaine - dependent outpatients with cocaine - induced
mood disorder ( CIMD ) , other mood disorders , or no mood disorder were compared on measures of psy
chiatric symptoms .', {'entities': [(59, 63, 'B_Disease'), (64, 72, 'I_Disease'), (75, 79, 'B_Diseas
e'), (90, 94, 'B_Disease'), (95, 104, 'I_Disease'), (113, 117, 'B_Disease'), (118, 126, 'I_Diseas
e'), (156, 167, 'B_Disease')]}], ['The prevalence rate for CIMD was 12 % at baseline .', {'entitie
s': [(24, 28, 'B_Disease')]}], ['Introduction of the DSM - IV diagnosis of CIMD did not substantiall
y affect rates of the other depressive disorders .', {'entities': [(42, 46, 'B_Disease'), (95, 105,
'B_Disease'), (106, 115, 'I_Disease')]}], ['Patients with CIMD had symptom severity levels between t
hose of patients with and without a mood disorder .', {'entities': [(14, 18, 'B_Disease'), (92, 96,
```


Figure 1. Example named entities

In this work, the proposed method for clinical NER based on convolutional neural network (CNN) aims to perform well in the NER task. It generates a clinical information vector attached to the entities for effective mapping to the class labels of the disease concept. This study investigated the effects of related medical entities on medical concepts for disease handling. While earlier studies have explored the impact of medical entities on other diseases, they have not explicitly addressed its influence on Parkinson's disease.

The major contributions are:

- Eliminating unwanted tokens and utilization of annotated named entities.
- Building clinical NER models that perform recognize-map-extract (RME) tasks around entities, reducing irrelevant extraction through effective mapping to Parkinson's disease-related topics.
- Developing an effective clinical information extraction approach based on CNN.
- Better performance during testing and validation.
- Significant improvement in precision, recall, and F1-measure.

The rest of the section covers related work, methodology, results and discussion and then the conclusion. The present work applies a rich feature list which includes named-entity-features and localized contextual features.

## 2. RELATED WORKS

Named entity recognition has recently been used to classify entities from e-health records for medical concepts. There is a huge interest in clinical NER for medical concepts. The work on clinical NER in the paper [4] has aimed at improving the performance of supervised classification using medical terminologies. In this existing work, a French corpus APcNER consisting of 5 types of entities has been built and it has used 147 annotated documents. The hybrid system using the terminologies has performed better than biGRU-CRF on i2b2 small corpus and APcNER, where the F-measure was 87.8% and 86.4% respectively.

Several NER systems identify named entities from biomedical documents. However, the manually annotated feature representation requires a lot of time and effort. In contrast, word embedding is considered a better choice because word embedding gives a more general representation with an automatic feature generation for learning by neural network and multiple kernel learning [5]. The integrated method containing dictionary look-up and machine learning has shown benefits in terms of F1-score on the CRAFT corpus which has 67 full documents. In the earlier work, confidence scores of evidence on events are not reliable as far as informative documents-handling is considered for supervised learning. An event extraction employed in the work on active learning uses NER and the details of event participants [6]. The event extraction has focused on filtering false positives from unlabeled documents. This related work has used committee-based learning to support multiple event types. The word similarity measure in this work finds unknown words in

the test data. This existing method used named entities to rank false negatives and measured the expression probabilities of events.

Linking named entities to an ontology concept is a very important step in information extraction. Karadeniz and Özgür [7] proposed an approach to normalize the bacteria biomedical entities through biotope ontology. The dictionary is used to normalize the drug reaction entities. This work has shown a precision of 65.9% using the BioNLP shared task bacteria Biotape test data.

Research scientists who work in chemical text data mining research aim to develop deep learning models to extract molecules and contextual features which will help in determining the properties of molecules and their activities. In this direction, the paper [8] has presented a review of chemical NER. A chemical entity recognition based on machine learning presented in the paper [9] used two conditional random field models. In that research work, the F-measure of 87.48% has been achieved in the NER task. Further, the research for effective extraction of named entities and associations between medical entities would allow automatic medical information extraction.

Nguyen *et al.* [10] outlined the creation of the COPIOUS corpus. This corpus includes annotations for five entity categories essential for biodiversity research and 668 documents with a total of 28,801 entity annotations, making it suitable for training and assessing text mining tools. The experiments has indicated that this corpus is valuable for extracting information related to biodiversity texts, specifically in the areas of NER and occurrence extraction. The NER investigated in the study [11] used Spark NLP without contextual embeddings and without using transformer models. As mentioned by the authors of the paper, there is a demand for easy-to-use NER models or tools for biomedical applications.

Another study on transfer learning [12] aimed to develop predictive disease models. However, this work used limited data to identify named entities present in the texts. The potential of NLP has to be explored in the field of disease detection. The integration of electronic health records (EHRs) and clinical reports obtained in real-time can improve the effectiveness of fighting against diseases like COVID-19.

The method discussed in the research article [13] aimed at extracting named entities for pituitary disorders using Chinese electronic medical records which contain information related to diagnosis and treatment. The authors have constructed a domain dictionary to perform feature matching through the CRF-based model. To make data analysis and data management easier, the research article [14] reviewed practices to identify named entities and relationships between proteins and drugs or genes and diseases. Information extraction can help doctors in the decision-making by generating necessary information about diseases, and treatments and helps to avoid medical errors. The research article [15] discussed several techniques used in the NER and relation extraction. From the literature review, it was understood that the process of merging several entity types to perform proper matching to the entities helps to reduce ambiguity.

The manual analysis of e-health records to extract valuable information is a tedious task. Silva *et al.* [16] have introduced an approach to extract valuable information from different resources. They have obtained 78.24% accuracy on the annotated Brazilian Portuguese dataset for entity and relation extraction. The authors have planned to use ontology concepts in future work.

An active learning-based approach introduced in the article [17] used an intrinsic strategic sampling to generate the training data for clinical text understanding from radiological reports. The major contributions of the article include consistent labelling and human annotation. The authors of the article have used German-MedBERT and R-BERT models for information extraction. The methodology presented in the paper [18] describes the Danish NER and relation extraction dataset aiming to utilize the e-health records.

A case study presented in the article [19] discussed a method to apply artificial intelligence in the NER for developing an ophthalmic disease registry. The NER in this existing work has shown 81.57% for precision, 80.99 recall and 81.28 F-score. The article [20] proposed a keyword-matching method to identify tumour-related entities and information. This existing method has applied regular expressions and rules to acquire sites of tumours.

Jiang *et al.* [21] implemented machine learning for developing a hybrid named entity recognition and evaluated their method using a training corpus that consisted of 349 annotated notes. The test dataset used in this work had 477 annotated notes. To utilize important clinical notes from the Japanese case report corpus, Shibata *et al.* [22] evaluated the information extraction approach using NLP and machine learning. In that study, they have used manually annotated 113 types of entities. After preprocessing, there were 2,194 sentences. The results of their research method reported a micro-averaged F1-score of 0.91 on the NER task.

Zhang *et al.* [23] applied two machine learning models, namely, the CRF, and the long short-term memory (LSTM)-CRF, to identify entities from Chinese EHRs. This existing work has used a dictionary-based approach as the baseline method. Regarding performance, the CRF-based model has improved precision and the LSTM-CRF-based model has improved the F1-score.

Clinical information is required at various stages for patient handling. Machine learning-based methods presented in the paper [24] applied text mining to extract information from discharge summaries.

This existing work used large-dimensional bags of features. The F1-score obtained on the concept extraction was 85.2. Bio-medical text mining plays a very important role in extracting entities from bio-medical articles. In this perspective, Cho and Lee [25] proposed the NER method using Bi-directional long short-term memory (Bi-LSTM) and CRF. This recent work has used the BioCreative II gene corpus and the disease corpus. The author has reported that they have achieved an F-score of 81.44%.

## 3. METHOD

The proposed clinical information extraction method works based on CNN. The first phase of the method involves pairs of sentence encoding for the given corpus. The knowledge represented in the form of contextual information vectors contains names of entities for the respective concepts (for example, disease names, medicines, treatment, and symptoms) towards Parkinson's disease-related data analytics. The set of false positives is fully analyzed to generate tokens. The features are evaluated to define the medical concepts properly.

Furthermore, the proposed method performs word recognition, addition, normalization, mapping, and adaptive combining. As part of convolutional learning on features, the fine-tuning looks at the neighbouring entity which leads to better information in the combining stage. The softmax function gives the output of named entity classification for different categories.

### 3.1. Dataset

The BC5CDR dataset is collected from the Kaggle repository which includes training data, testing data, and validation data. Text annotations are increasing the expressiveness of texts present in the document. The following labels are used in the dataset: B-begin entity, I-Inside entity, and O-outside entity. Table 1 provides the details of the type of data along with size.

Table 1. Dataset details

| Data type | Size |
|---|---|
| Train data | 919 KB |
| Test data | 960 KB |
| Train_Dev data | 1.83 MB |

### 3.2. Detecting clinically named entities using CNN-RME

The main functional components of the proposed method are input clinical dataset, tokenization, text embedding, CNN-RME, POS Tagger, dependency parser, named entity recognition, and visualization. Tokenization splits the input texts into words. As part of encoding, the words were embedded in the form of vectors. The input to this task is the maximum length of an entity-embedded sentence and an entity-embedded sentence. Once completed with embedding, the algorithm proceeds with the "RME" cycle, convolutional learning, max-pooling, and Softmax function. During the extraction of entities and related information, we used 1-D convolutional filters which are of varied widths. The width of each filter denotes the n-gram length used for the filter. Figure 2 illustrates the clinical NER-RME architecture, where 'r' denotes the recurring nature of the process until the stopping condition is reached.

### 3.3. The RME mechanism

The "RME" mechanism involves multiple steps such as recognizing entities in the text, mapping those entities to target labels or entity types, and then extracting structured information based on recognized entities. In the recognition step, named entities, for example, persons, disease names, symptoms, dates, hospitals, and locations, are located and identified from the dataset. This step uses tokenization, part-of-speech tagging, and NER. Clinical NER models extract words or phrases that are entities and assign them specific labels or types.

In the mapping step, named entities were mapped to predefined categories. These categories represent the specific medical information extraction task. For example, disease, symptom, treatment, or medicine. The mapping step helps to extract useful information.

The extraction step produces structured information from the mapped entities. For each entity type, attributes and their associated values were extracted for the specific information task. For example, the entity type "patient" includes attributes such as "name", "age", "place", "city", and "country". The extracted information can be presented in the form of a graph or table. Figure 3 shows an example of named entities recognized. Figure 4 shows an example of mapped entity types and Figure 5 shows an example of extracted information.

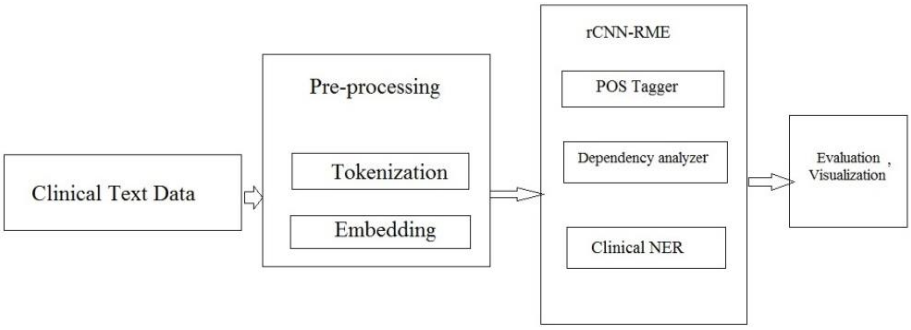


Figure 2. The architecture of the clinical NER method

Entities recognized:
Parkinson: (DISEASE)
Psychiatric disorder: (SYMPTOM)
DSM-IV: (DIAGNOSIS)

Figure 3. Example named entities recognized

Mapped entity types:
"Parkinson" is mapped to the entity type "Disease"
"Psychiatric disorder" is mapped to the entity type "Symptom"
DSM-IV is mapped to the entity type "Diagnosis"

Figure 4. Example mapped entity types

Structured information extracted:
Disease Name: "Parkinson"
Disorder: "Psychiatric disorder"
Diagnosis: "DSM-IV"

Figure 5. Example extracted information

### 3.4. Loss function

Let N be the number of tokens, C be the number of entity classes, $y_{ij}$ be a binary indicator to check whether the correct label for the token 'i' is class j, and $\widehat{y_{ij}}$ be the predicted probability that the token 'i' belongs to class j. Now, the loss function is defined as (1).

$$L(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}.\log(\widehat{y_{ij}}) \quad (1)$$

The mini-batches formed from the dataset improve the efficiency during training. The regularization techniques applied in the present work add terms to adjust the loss function to avoid overfitting. The dropout method reduces the co-adaptation between different units during training.

## 4. RESULTS AND DISCUSSION

The Spacy 2.0 for the experimental analysis used necessary features for advanced natural language processing. The experiment is conducted in the Jupyter Notebook with an Anaconda environment. The present method based on CNN utilized functions such as POS tagger, dependency analyzer, and NER. The NER part of the Spacy module was adapted for this present clinical NER approach. The training of a clinical NER model is carried out for detecting and classifying entities of interest including disease entities, medications, procedures, and symptoms. Dropout used with the CNN prevents overfitting by randomly

setting a fraction of input neurons during training. It helps to prevent co-adaptation of units for improving the generalization performance. During training, when the dropout was 0.5 it had 50% drop-out units. The results are presented in Figures 6 to 9.

The following measures are used to assess the performance of the proposed NER technique. Precision calculates the accuracy of positive entity predictions whereas recall considers the number of relevant entities captured in the process. The F1-measure is used to maintain the balance between precision and recall because missing important clinical entities leads to low recall. Further, incorrect identification of non-existent entities leads to low precision.

- TP = Number of correctly predicted positive entities.
- FP = Number of incorrectly predicted entities as positive.
- FN = Number of incorrectly predicted negative entities.
- TN = Number of entities correctly predicted as negative.
- Precision = TP/(TP+FP)
- Recall = TP/(TP+FN)
- F1-score = 2*precision*recall/(precision+recall)

The load_data_spacy() returns training data and unique labels. The cross-entropy loss helps to improve the effectiveness of mapping each token to one of the entity classes.

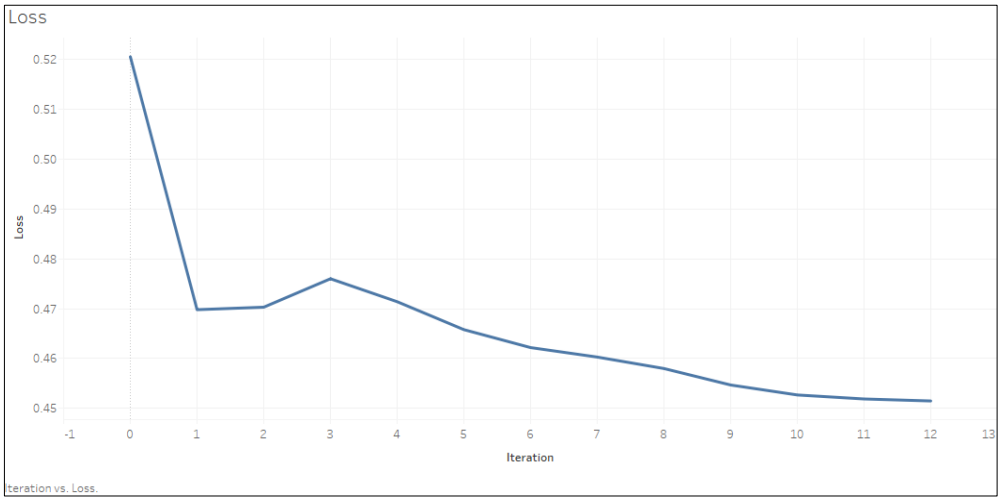


Figure 6. Loss

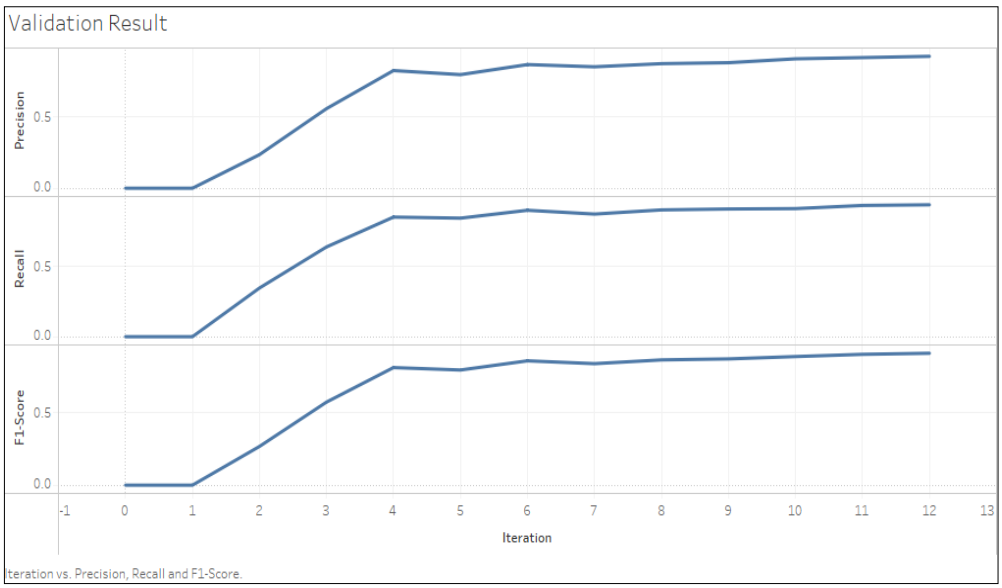


Figure 7. Validation result

The precision and recall of the NER model depend on the correctness of tagged named entities. The degree of correctness is measured based on the phrase boundaries. Let us consider the phrase "acute Parkinson's disease". If the tag misses "acute", the NER system needs to decide whether the entity is a true positive or not.

In a normal case, it is possible to consider this phrase as a "true positive". But, for a severe disease that requires more effective treatment, it may be required to consider the whole term. Therefore, the phrase will be treated as a false positive in this case. The train_spacy() returns an F1-score for validation and an F1-score for testing.

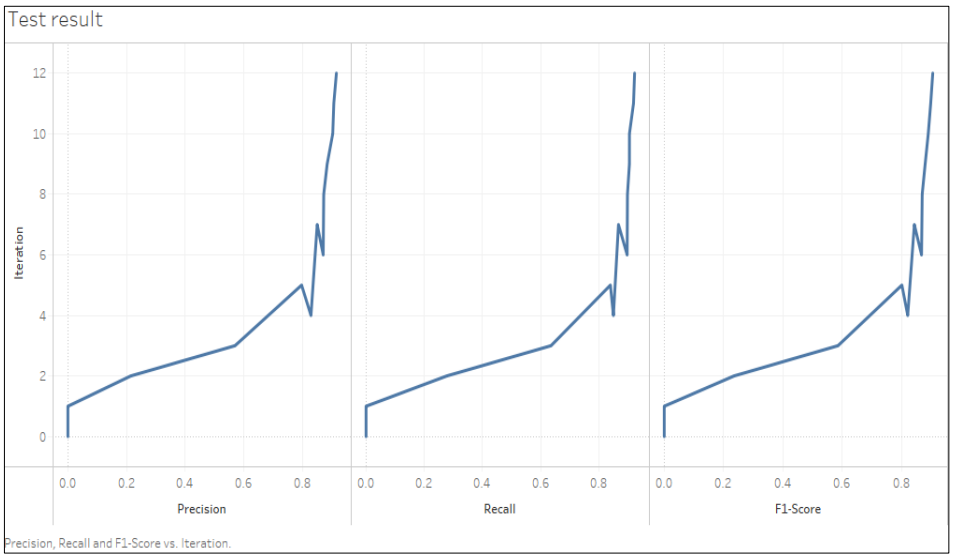


Figure 8. Test result

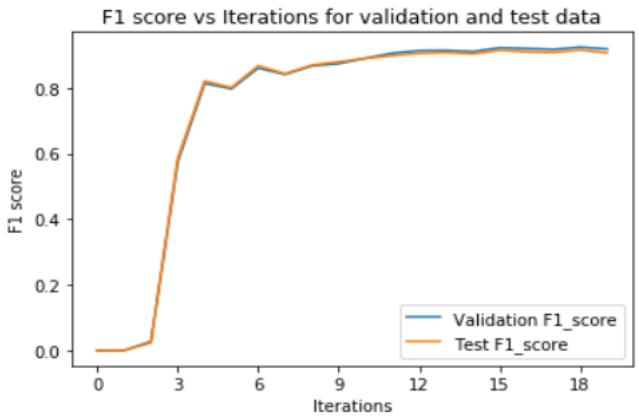


Figure 9. Iteration-wise F1-score

Figure 10 shows the test result after RME tasks. This paper compares the present method with other methods like HMM, CRF, ERNIE-Doc-BiLSTM-CRF NER, and biGRU-CRF. Table 2 presents the comparative analysis of the proposed method and other methods. The task of the proposed clinical NER method is to determine the entity type and its boundary. This paper uses a unique NER method which checks the consistency of how the entity-related information accurately corresponds to the true target labels. The quantitative analysis of our method shows that this CNN-RME method can obtain the optimal performance on the performance measures, namely, precision, recall, and F1-score. CNN depends on the network structure that contains rich feature maps. Each convolutional operation performed on these feature maps leads to more feature acquisition that handles complex relationships present in the medical data. This provides more information about the model. The proposed method may benefit from CNN-RME

without adversely impacting NER performance. However, further in-depth studies may be needed to confirm its performance, especially regarding NER in the complex domain.

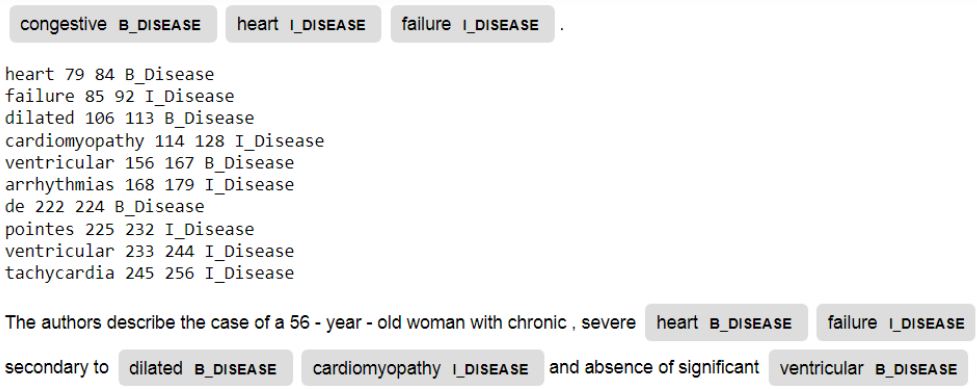


Figure 10. Test result after RME tasks

Table 2. Comparison of results

| Method/Measures | Precision | Recall | F1-measure |
|---|---|---|---|
| Hybrid approach | 72.63-79.10 | 70.14-77.70 | 73.79 |
| MCD | 92.4 | 90.7 | 91.5 |
| CRF | 92.03 | 89.7 | 89.4 |
| ERNIE-Doc-BiLSTM-CRF NER | 86.72 | 83.39 | 85.02 |
| biGRU-CRF | - | - | 87.8 |
| Proposed method | 92.57 | 92.22 | 92.1 |

## 5. CONCLUSION

The automatic mapping of text to knowledge bases or databases is a difficult task. This paper presented research gaps identified through the review of related methods that have been discussed for disease data analytics. Moreover, a clinically NER method presented in this paper was developed with a combination of CNN-based deep learning and RME analysis for analyzing medical data for information extraction. This proposed method has created a better data science approach and effective mapping of named entities to medical concepts. Furthermore, this research work has studied existing works related to NER in the information extraction process. The future study will try to increase the effectiveness of clinical data analytics through other deep learning methods.

## REFERENCES

[1] R. Ramachandran and K. Arutchelvan, "Named entity recognition on bio-medical literature documents using hybrid based approach," *Journal of Ambient Intelligence and Humanized Computing*, Mar. 2021, doi: 10.1007/s12652-021-03078-z.
[2] T. Yang, Y. He, and N. Yang, "Named entity recognition of medical text based on the deep neural network," *Journal of Healthcare Engineering*, vol. 2022, pp. 1–10, Mar. 2022, doi: 10.1155/2022/3990563.
[3] L. Xin and H. Xiaoyan, "Recognition of unknown entities in specific financial field based on ERNIE-Doc-BiLSTM-CRF," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–8, May 2022, doi: 10.1155/2022/3139898.
[4] I. Lerner, N. Paris, and X. Tannier, "Terminologies augmented recurrent neural network model for clinical named entity recognition," *Journal of Biomedical Informatics*, vol. 102, p. 103356, Feb. 2020, doi: 10.1016/j.jbi.2019.103356.
[5] I. Lauriola, F. Aiolli, A. Lavelli, and F. Rinaldi, "Learning adaptive representations for entity recognition in the biomedical domain," *Journal of Biomedical Semantics*, vol. 12, no. 1, p. 10, Dec. 2021, doi: 10.1186/s13326-021-00238-0.
[6] X. Han, J. Kim, and C. K. Kwoh, "Active learning for ontological event extraction incorporating named entity recognition and unknown word handling," *Journal of Biomedical Semantics*, vol. 7, no. 1, p. 22, Dec. 2016, doi: 10.1186/s13326-016-0059-z.
[7] İ. Karadeniz and A. Özgür, "Linking entities through an ontology using word embeddings and syntactic re-ranking," *BMC Bioinformatics*, vol. 20, no. 1, p. 156, Dec. 2019, doi: 10.1186/s12859-019-2678-8.
[8] S. Eltyeb and N. Salim, "Chemical named entities recognition: a review on approaches and applications," *Journal of Cheminformatics*, vol. 6, no. 1, p. 17, Dec. 2014, doi: 10.1186/1758-2946-6-17.
[9] D. Campos, S. Matos, and J. L. Oliveira, "A document processing pipeline for annotating chemical entities in scientific documents," *Journal of Cheminformatics*, vol. 7, no. S1, p. S7, Dec. 2015, doi: 10.1186/1758-2946-7-S1-S7.
[10] N. Nguyen, R. Gabud, and S. Ananiadou, "COPIOUS: a gold standard corpus of named entities towards extracting species occurrence from biodiversity literature," *Biodiversity Data Journal*, vol. 7, pp. 1–23, Jan. 2019, doi: 10.3897/BDJ.7.e29626.
[11] V. Kocaman and D. Talby, "Accurate clinical and biomedical named entity recognition at scale," *Software Impacts*, vol. 13, p. 100373, Aug. 2022, doi: 10.1016/j.simpa.2022.100373.
[12] S. Raza and B. Schwartz, "Entity and relation extraction from clinical case reports of COVID-19: a natural language processing approach," *BMC Medical Informatics and Decision Making*, vol. 23, no. 1, p. 20, Jan. 2023, doi: 10.1186/s12911-023-02117-3.

[13] A. Fang *et al.*, "Extracting clinical named entity for pituitary adenomas from Chinese electronic medical records," *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, p. 72, Dec. 2022, doi: 10.1186/s12911-022-01810-z.

[14] N. Perera, M. Dehmer, and F. Emmert-Streib, "Named entity recognition and relation detection for biomedical information extraction," *Frontiers in Cell and Developmental Biology*, vol. 8, Aug. 2020, doi: 10.3389/fcell.2020.00673.

[15] M. Y. Landolsi, L. Hlaoua, and L. Ben Romdhane, "Information extraction from electronic medical documents: state of the art and future research directions," *Knowledge and Information Systems*, vol. 65, no. 2, pp. 463–516, Feb. 2023, doi: 10.1007/s10115-022-01779-1.

[16] D. P. da Silva, W. R. Fröhlich, B. H. de Mello, R. Vieira, and S. J. Rigo, "Exploring named entity recognition and relation extraction for ontology and medical records integration," *Informatics in Medicine Unlocked*, vol. 43, p. 101381, 2023, doi: 10.1016/j.imu.2023.101381.

[17] M. Jantscher, F. Gunzer, R. Kern, E. Hassler, S. Tschauner, and G. Reishofer, "Information extraction from German radiological reports for general clinical text and language understanding," *Scientific Reports*, vol. 13, no. 1, p. 2353, Feb. 2023, doi: 10.1038/s41598-023-29323-3.

[18] M. Laursen, J. Pedersen, R. Hansen, T. R. Savarimuthu, and P. Vinholt, "Danish clinical named entity recognition and relation extraction," *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pp. 655–666, 2023, [Online]. Available: https://aclanthology.org/2023.nodalida-1.65.

[19] C. Z. Macri *et al.*, "A case study in applying artificial intelligence-based named entity recognition to develop an automated ophthalmic disease registry," *Graefe's Archive for Clinical and Experimental Ophthalmology*, vol. 261, no. 11, pp. 3335–3344, Nov. 2023, doi: 10.1007/s00417-023-06190-2.

[20] Z. Liang, J. Chen, Z. Xu, Y. Chen, and T. Hao, "A pattern-based method for medical entity recognition from Chinese diagnostic imaging text," *Frontiers in Artificial Intelligence*, vol. 2, May 2019, doi: 10.3389/frai.2019.00001.

[21] M. Jiang *et al.*, "A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 601–606, Sep. 2011, doi: 10.1136/amiajnl-2011-000163.

[22] D. Shibata, E. Shinohara, K. Shimamoto, and Y. Kawazoe, "Towards structuring clinical texts: joint entity and relation extraction from japanese case report corpus," in *Studies in Health Technology and Informatics*, vol. 310, 2024, pp. 559–563.

[23] Y. Zhang, X. Wang, Z. Hou, and J. Li, "Clinical named entity recognition from Chinese electronic health records via machine learning methods," *JMIR Medical Informatics*, vol. 6, no. 4, p. e50, Dec. 2018, doi: 10.2196/medinform.9965.

[24] B. de Bruijn, C. Cherry, S. Kiritchenko, J. Martin, and X. Zhu, "Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 557–562, Sep. 2011, doi: 10.1136/amiajnl-2011-000150.

[25] H. Cho and H. Lee, "Biomedical named entity recognition using deep neural networks with contextual information," *BMC Bioinformatics*, vol. 20, no. 1, p. 735, Dec. 2019, doi: 10.1186/s12859-019-3321-4.


## BIOGRAPHIES OF AUTHORS

**Dr. Dhanasekaran Kuttaiyapillai** is currently working as an Assistant Professor in the Department of Data Science and Business Systems, School of Computing, SRM Institute of Science and Technology, Chengalpattu, Chennai, Tamilnadu, India. He has several publications to his credit published in reputed journals, and he is regularly contributing articles of national and international importance that are highly appreciated. He is a member of ACM, ISTE, and IAENG. He is a reviewer of various international journals and was honoured with the "Excellent Young Researcher Award" in 2022. His research interests include machine learning, data analytics, and natural language processing. During his service, he has worked as a research coordinator, NBA coordinator, NAAC Coordinator, and NIRF Coordinator. As a project guide, he has received the "Best Project of the Year" award from KSCST. He has received grants from funding Agencies, namely ISRO, CSIR, DRDO, Government of India, for organizing national-level seminars, national-level conferences, and national-level workshops, respectively. He has published a Book chapter in IGI Global, USA, presented papers at various conferences and participated in seminars, workshops, and training programs. He has also delivered several talks and chaired several sessions at conferences. He can be contacted at email: dhanasek1@srmist.edu.in.

**Dr. Anand Madasamy** received a B.Tech. degree in Information Technology from Anna University, India, in 2006, an M.E. degree in Computer Science and Engineering from Anna University, India in 2010 and a Ph.D. in Computer Science and Engineering from Anna University, India in 2020. He is currently working as an Assistant Professor at, the Department of Data Science and Business Systems at the School of Computing, SRM Institute of Science and Technology, Chennai. His research interests include Cross-layer design, Energy-aware routing and MAC protocol design, link adaptation, network optimization for wireless networks, and machine learning. He is an active member of IEEE. He can be contacted at email: anandmenscall@gmail.com.

**Dr. Shobanadevi Ayyavu** is working as an Assistant Professor in the Department of Data Science and Business Systems, School of Computing at SRM Institute of Science and Technology. She obtained her BE degree in Computer Science and Engineering from Anna University, Chennai, India in the years 2007 and 2008 respectively. After completing her post-graduation M.Tech. (CSE) in the year 2016 from Jawaharlal Nehru University Ananthapur (JNTUA), she completed her PhD in Computer Science and Engineering in the year 2021, SRM Institute of Science and Technology, Kattankulathur, Chennai. She has more than 10 years of teaching and industry experience. She has published 15 papers in highly reputed international SCI/Scopus Indexed Journals and presented several papers at National and International conferences. She has also worked as a Research Scientist at the National Institute of Wind Energy, Velachery, Chennai, India. Her main area of research interest includes data analytics, data mining, machine learning, deep learning, forecasting models Python, and other data science-related frameworks. She can be contacted at email: shobanaa3@srmist.edu.in.

**Dr. Md Shohel Sayeed** has been a member of Multimedia University since 2001 and now he serves as a Professor of the Faculty of Information Science and Technology. Dr. Shohel's core research interest is in the area of Biometrics, information security, image and signal processing, pattern recognition and classification. Till date, he has published over 90 research papers in international peer-reviewed journals and international conference proceedings as a result of his research work. His research works have been accepted by journals such as IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), the International Journal of Pattern Recognition and Artificial Intelligence (IJPRI), and Discrete Dynamics in Nature and Society (DDNS). These papers have been cited in various international journals and conferences several times. Several of his findings have been presented at several well-recognized IEEE conferences such as ICSP2006, ICIAS2007, ITSIM2008, CSECS2009, and ITSIM2010. He has been appointed technical paper reviewer for the Journal of Pattern Recognition Letters, IEEE Transaction on Neural Networks, IEEE Transactions on Automation Science and Engineering, Journal of Computer Methods and Programs in Biomedicine and International Journal of Computer Theory and Engineering. He has also been invited to review technical papers for several international conferences. In recognition of his professional contribution, he has obtained recognition as a senior member of the IEEE Computer Society. He can be contacted at email: shohel.sayeed@mmu.edu.my.