

Deep Learning-Based NER for Digitizing Unani Medicines Text

*Hafiza Shantal Khalid, *Talha Waheed

*Department of Computer Science University of Engineering and Technology Lahore-54890, Pakistan
hafizashantal@gmail.com, twaheed@uet.edu.pk

Abstract

Traditional medical systems such as Unani Medicine (UM) preserve centuries of therapeutic knowledge derived from Greco-Arabic traditions and enriched by South Asian influences. However, much of the existing Unani literature remains non-digitized and is written in multiple classical languages, limiting its accessibility and integration with modern healthcare technologies. This study presents a Named Entity Recognition (NER)-based framework for the systematic digitization and extraction of structured information from Unani medicinal texts. Using the reference book [18], a specialized dataset was developed through Optical Character Recognition (OCR), followed by preprocessing and manual annotation. A BiLSTM-CRF model enhanced with pre-trained GloVe embeddings was implemented to identify and classify four key entity types: Unani drug names, English names, scientific (Latin) names, and temperament attributes. Experimental results achieved an overall accuracy of 94.0% and an F1-score of 92.7%, demonstrating the effectiveness of the proposed approach despite challenges such as OCR noise, multilingual content, and domain-specific terminology. Comparative analysis indicates that the model performs competitively with recent NER systems in both biomedical and non-biomedical domains. The proposed framework facilitates the creation of structured databases, knowledge graphs, and AI-driven applications for traditional medicine. Furthermore, it contributes to the preservation of cultural heritage while supporting the integration of traditional medical knowledge with data-driven healthcare systems.

Index Terms—Traditional Medicines, Unani Medicines, NER, Named Entity, Deep Learning

I. INTRODUCTION

NER is the most important component of natural language processing (NLP) which seeks to extricate and name some portions such as names of individuals, names of regions or specialized phrases in unstructured texts. When applied to the situation of the digitization of historical and traditional medical records, NER assumes a decisive role in deciding to pull out homogeneous data out of unstructured data to ensure a greater data organization, access, and analysis. In this paper, I shall concentrate on recruiting deep learning-based NER approaches to digitalize texts related to Unani medicine, a method of traditional medicine based on GrecoArabic medicine, which is well documented in Arabic, Persian and Urdu languages. Unani medicine, offered with such a rich historical collection has a great dissimilarity in different varieties of manuscripts such as handwritings, books, etc. in the form of texts with a lot of previous information having the precious information concerning with medicinal herbs, formulating the elements, diseases and their modalities of treatment.

Nonetheless, the problem is that due to the unorganized character of these texts, linguistic diversity, term complexity, and variations in the scripts, there are many issues related to its digitization and knowledge-gathering. Conventional rule-based or statistical NER approaches tend to find it difficult to cope with the complexities of such texts, because they rely on pre-scripted rules or restrict sets of feature engineering, and so fail to generalize to a variety of linguistic and contextual situations. Using deep learning methods to perform NER with such architecture as (RNNs), Long Short-Term Memory (LSTM) networks, and Transformer-based models arousing due to the creation of BERT (Bidirectional Encoder Representations with Transformers) and others. and others. these problems finally have a comprehensive solution. These models can be highly advantageous in identifying entities in the field of Unani medicine because they are able to capture contextual dependencies and semantic relationship in

text. As an illustration, BERT based models that were pre-trained on huge corpora and fine-tuned using field-specific corpora have achieved state-of-the-art results in NER tasks in many disciplines biomedical, and historical text [1]. It is estimated that Unani medicine texts can yield a treasure trove of information, when their digitalization via deep learning-based NER could shed more light upon them, to benefit researchers, practitioners, and policymakers alike.

This method has the possibility to de-rive entities like medicinal plants, diseases, symptoms and treatment procedures which can generate structured databases and promote knowledge preservation and integrate them with modern medical systems. It can as well assist in filling the gap between the conventional and modern medical practice, researchers must be encouraged to conduct studies on the efficacy of Unani therapeutic agents and their possible application in the modern healthcare industry. The current study is based on the earlier researches in the area of biomedical NER, where the methods of deep learning have been efficiently used to identify entities within clinical reports and scientific articles [19]. Nevertheless, the use of NER in texts of Unani medicine has also not been fully explored, at least in terms of annotated data and the linguistic difficulty in the original text. In tackling these difficulties, this research would formulate a deep learning-based NER system with features to suit Unani medicine documents to help in maintaining and reusing this ancient knowledge system.

II. LITERATURE REVIEW

Named Entity Recognition (NER) is a subfield of information extraction, and it entails the identification and categorisation of such things as the names of people and places, organisations. or, in certain domains, medical devices, drugs or procedures. Within the medicinal science, including the Unani and the Traditional Chinese Medicine (TCM), the NER has taken the form of digitizing a textual data and

using a structured data in analyzing its information. Traditional Chinese literature employs the use of ideographic characters where such characters as strokes and radicals have hidden meaning. TRBNER [8], [7] dealt with this by combining ALBERT embeddings using stroke-level representations, which were re-trieved using Convolutional Neural Networks (CNNs). Word-boundary features and multi-scale channel attention mechanisms were also involved in the model to boost capturing of contextual signals. TRBNER optimized the performance to 87.6 percent with F1 score in its input type which gives a visual and structural information in the character level and having no clear word division. The new trend of the usage of (LLMs) in (NER) has influenced the Traditional Chinese Medicine (TCM) domain as well. [8] refined an LLM with 7-billion parameters using Low-Rank Adaptation (LoRA) and domain-specific pretraining. The model got 93 percent F1 score when used with ancient TCM texts, especially when identifying nested entities that commonly occur in medical descriptions based on formulas.

[22] present a comprehensive overview of NER techniques, datasets, and tools using three categories, rule-based, statistical, and the deep learning solutions. They highlight the increased popularity of BiLSTM-CRF, transformer-based structure, and pre-trained language models in the current implementation of NER. This background information demonstrates that there is a need to select architectures to be used in accordance with the domain specific language and structure of traditional medicine texts. In addressing the issue of low-resource languages, [12] developed DeepSpacy-NER, an efficient deep learning model supportive of Punjabi. They report on the difficulties and possibilities of using NER with morphologically rich languages and emphasize why languages should be modeled specially and tokens used specifically. [11] applied transfer learning and asymmetric triple-training, to biomedical NER tasks, showing that adaptation even within biomedical typically helps performance in those cases significantly.

It has a potential indication on the healing of the Unani medicine where the data sampled on annotation may also be scarce. The research by [13] focuses on deriving ingredients using the texts that contain recipes. Even though it is not medicine-specific they serve as an example of NER systems spuriousness in parsing ingredient-related data in semistructured or free-form text since it is comparable to documents of Unani and TCM which tend to be written with the aid of either natural language or hybrid language. A number of studies have enhanced NER application in TCM that is conceptually and textually analogous to Unani medicine. [14] used BiLSTM-CRF as the named entity recognition model on TCM patents and obtained the results below: the accuracy of the herbs, symptoms, and treatment methods detection was high. [7] addressed one of the major shortcomings in medical text processing medication, i.e., nested entities, using a new method of nested NER. Their model really helped increase the identification of the overlapping entities which could be compound drug formulations or symptoms in multitiered diagnosis. [6] proposed a method of character-word vector fusion, which added the semantic depth of words and the details of character to a mixture, which could improve the robustness of Chinese medical text NER, especially in the low-resource case.

[5] enhanced BiLSTM-CRF by introducing a knowledge graph, which, in addition to the increased recognition performance, allowed

TABLE I
Literature Review

Study	Model	Key Features	Dataset	Limitations
Su et al. (2022)	RoBERTa + BiLSTM + CRF (F1: 96%)	Transformer with BiLSTM-CRF layered labeling	Internet TCM corpus	Only tested on Chinese TCM; lacks mixed entity handling
Ma & Wu (2024)	BERT - BiLSTM - CRF (F1: 90.5%)	Semantic and positional feature fusion	TCM herbal corpus	Requires lexical dictionary updates; inter-pretability uncertain
TRBNER (2024)	ALBERT + Stroke CNN + MS-CAM (F1: 87.6%)	Stroke-level body features with attention	TCM hospital records	High computational cost and deployment complexity
Ma et al. (2022)	MT-CGAN + BiLSTM (F1: 91.08%)	Character-word fusion with adversarial learning	Annotated TCM literature	Complex GAN tuning; scalability concerns
Zheng et al. (2025)	LLM-induced DA + BiLSTM (F1: 93%)	Few-shot learning with synthetic text augmentation	Ancient Chinese TCM texts	Resource intensive; prompt dependent
Bhattacharya et al. (2023)	Transfer Learning + Tri-Training	Semi-supervised biomedical NER	Biomedical datasets	Complex pipeline; requires large unlabeled corpora

the semantic reasoning process over medical texts.

[16] included a particular factor in literature-based NER with the introduction of a multi-granularity CGAN-based model to accommodate fine-grained and contextual information. [15] showed that the information extraction and knowledge graph construction in geoscience literature can be carried out. It is true that they are speaking in another field but the thought process can be implemented in Unani literature where one can organize the relationships between drugs, treatments and diseases.

III. METHODOLOGY

The study is aimed at developing a (NER) model, based on learning, to digitalize and extract structured information in the texts of Unani medicine. The methodology entails various stages, including data gathering, preprocessing, annotation, training a model, and performance quality.

A. Data Collection and Conversion

The primary data source is the book of classification of Unani drugs available in English and the English scientific names, [18]. As the book was in PDF format and could not be directly copied to text, the following multi-step process was applied:

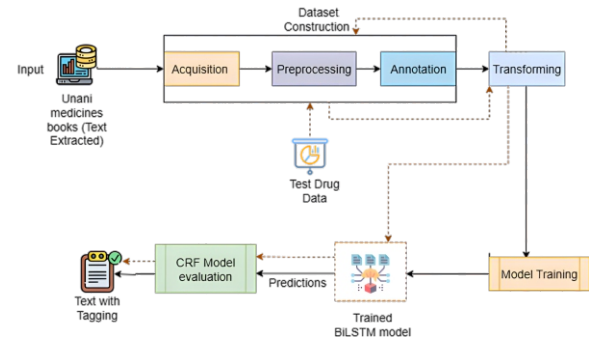


Fig. 1. Architecture diagram for digitizing Unani Medicines textual data by using a deep learning model

- **PDF to Image Conversion:** The entire PDF was transformed into high-resolution images (one per page) to maintain the layout and typography necessary for precise text extraction.
- **Optical Character Recognition (OCR):** Tesseract OCR was applied to each image to extract raw text, with particular attention to the accuracy of entity names, including Unani drug names (Arabic/Urdu), their English translations, and scientific (botanical) names

B. Dataset Construction

The original material was a PDF book, which was initially transformed into images and subsequently processed through OCR (Optical Character Recognition). After extraction, the data underwent cleaning to eliminate noise and non-textual elements. The resulting dataset included:

- Total Entities Labeled: 5,400
- Unique Drug Entries: 1,800
- Entity Types: 4 major categories (Drug Name, Scientific Name, Category, Temperament)

The dataset was divided as follows:

- Training Set: 80
- Test Set: 20

C. Data Cleaning and Preprocessing

The text extracted via Optical Character Recognition (OCR) exhibited considerable noise due to factors such as font variations, inconsistent formatting, and recognition errors. To ensure the reliability and usability of the data for subsequent analysis, a multi-stage preprocessing pipeline was designed. The following steps were implemented:

TABLE II
Experimental Setup

Component	Description
Model	BiLSTM-CRF (Bidirectional LSTM + CRF layer)
Embedding Layer	Pre-trained GloVe embeddings (100D)
Optimizer	Adam
Loss Function	CRF Negative Log Likelihood
Framework	TensorFlow / Keras
Epochs	10
Batch Size	32

- 1) *Removal of Non-Textual Artifacts and Special Characters:* Extraneous symbols (e.g., @, #, ., or OCR-generated artifacts) were removed using regular expressions. Non-informative characters such as excessive whitespace, page numbers, and broken line markers were eliminated to standardize the text.
- 2) *Correction of OCR-Induced Errors:* Common misrecognitions (e.g., “1” misread as “l,” or “O” misread as “0”) were identified through pattern matching. A dictionary-based method was applied to automatically replace incorrect words with the most probable correct forms, improving lexical consistency.
- 3) *Segmentation and Structuring of Data:* The cleaned text was segmented into structured lines and tokens, preserving the critical triplet:
 - Unani Drug Name

- English Name
- Scientific Name

This step ensured that each entry was properly aligned, facilitating downstream tasks such as classification and retrieval.

- 4) *Normalization of Text Formats:* All entries were converted to consistent casing (e.g., sentence case for readability or lowercase for computational uniformity).
- Punctuation normalization was applied to standardize delimiters (e.g., replacing multiple hyphens with a single hyphen).
- Unicode normalization (NFC/NFKC) was employed to handle inconsistencies arising from OCR, ensuring uniform representation of accented characters and diacritics.
- 5) *Lemmatization and Morphological Normalization:* Lemmatization was performed to reduce words to their canonical base forms (e.g., “extracts” → “extract”), which aids in minimizing redundancy in text analysis. This was particularly critical for scientific names and English descriptors, ensuring that variations in inflections did not affect data consistency.
- 6) *Stop word Handling (Task-Specific):* Domain-irrelevant stop words (e.g., “and,” “of,” “the”) were selectively removed to focus on meaningful terms. However, domain-critical words (e.g., “Extract of,” “Powdered”) were retained to preserve semantic information relevant to Unani drug descriptions.
- 7) *Sub word and Token Standardization:* For cases involving modern NLP models (e.g., BERT-based embeddings), sub word tokenization (e.g., Word Piece or Byte-Pair Encoding) was applied to handle rare and compound words, ensuring robust representation of scientific terms.

D. Annotation for Named Entity Recognition

To enable the training of a supervised deep learning model, the preprocessed text was manually annotated for entity recognition using the BIO (Begin, Inside, Outside) tagging schema. This approach is widely adopted in NER tasks as it provides a clear mechanism to represent entity boundaries and supports multi-token entities effectively.

- 1) *Entity Types:* Three main entity categories were defined based on the domain requirements:
 - **UNANI DRUG:** The native name of the medicinal substance in Unani (e.g., Asgandh).
 - **ENGLISH NAME:** The commonly used English name (e.g., Winter Cherry).
 - **SCIENTIFIC NAME:** The botanical or Latin scientific name (e.g., Withania somnifera).
- 2) *Tagging Schema:* The BIO format was employed, where:
 - **B-XXX:** Marks the beginning of an entity of type XXX.
 - **I-XXX:** Marks subsequent tokens within the same entity.
 - **O:** Marks tokens that do not belong to any entity.

For example, the phrase “Asgandh Winter Cherry Withania

somnifera” would be tagged as: 'B-UNANI DRUG' — 'B-ENGLISH NAME' — 'B-SCIENTIFIC NAME' — 'I-SCIENTIFIC NAME'.

3) *Annotation Workflow*: The SpaCy manual annotation tool was used to facilitate precise and efficient annotation. A guideline document was created, specifying rules for ambiguous cases (e.g., how to handle synonyms or spelling variations). Annotation was performed in multiple rounds, allowing annotators to refine entity definitions as understanding of the dataset evolved.

4) *Quality Assurance and Consistency*: Inter-annotator agreement (IAA) was calculated on a subset of the data using metrics like “Cohen's Kappa” to assess consistency between annotators. Discrepancies were resolved through discussion, and the annotation guidelines were updated accordingly. Random samples of annotated data were periodically reviewed to minimize systematic errors.

5) *Challenges Encountered*:

- OCR-induced noise: Some scientific names were partially recognized or split across tokens, requiring careful reconstruction during annotation.
- Homonyms and Synonyms: Certain Unani drug names had multiple English equivalents, necessitating clear annotation rules.
- Multi-word Entities: Handling multi-token scientific names (Withania somnifera Dunal) required strict adherence to BIO tagging to ensure structural accuracy.

6) *Final Dataset*: The resulting dataset consisted of high-quality, domain-specific annotations suitable for training and evaluating NER models, with each token labeled under the standardized BIO scheme.

E. Model Development

A deep learning-based NER model was developed using a Bidirectional Long Short-Term Memory with Conditional Random Fields (BiLSTM-CRF) architecture. This architecture is ideal for sequence labeling tasks like NER, capturing both past and future dependencies in the input sequence. Embedding Layer: Pre-trained word embeddings (e.g., GloVe or FastText) were used to initialize the input representations.

- BiLSTM Layer: Captures contextual information from both directions of the text sequence.
- CRF Layer: Ensures valid tag sequences and improves the prediction accuracy of entity boundaries.

F. Post-processing and Output Generation

After model inference, predicted entities were formatted into structured records linking the Unani drug name with its English and scientific equivalents. These results were verified against the source text and further refined using rule-based checks.

IV. RESULTS

The performance of our deep learning-based (NER) model was evaluated separately on the training and testing sets, with results

presented in both tabular form and accompanying graphs. The model was trained on an 80/20 split of our specially annotated Unani pharmacological corpus, which was sourced from the classical writings of [36]. We optimized the network using the Adam optimizer and employed a categorical cross-entropy loss function. To quantify model effectiveness, we calculated the following standard token-level metrics on the test set:

- Precision: the percentage of entity predictions that are correct
- Recall: the percentage of true entities that are successfully detected
- F1-Score: the harmonic mean of Precision and Recall The model was tasked with identifying four key entity types:
- Drug Names (both English and Urdu)
- Scientific Names (Latin botanical nomenclature)
- Drug Categories (e.g., Mufradat, Murakkabat)
- Temperament Attributes (e.g., Hot-Dry, Cold-Wet) Below, Table I summarizes the token-level evaluation on the held-out test set, and Figures 1–2 illustrate training versus validation accuracy and loss over 20 epochs.

A. Metrics Evaluation

To rigorously evaluate our deep learning-based (NER) model on the Unani pharmacological corpus, we employed four complementary token-level metrics:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

Together, these metrics provide a holistic view of both overall labeling correctness and the model's ability to identify and categorize entity tokens without bias toward false positives or false negatives.

B. Results and Discussion

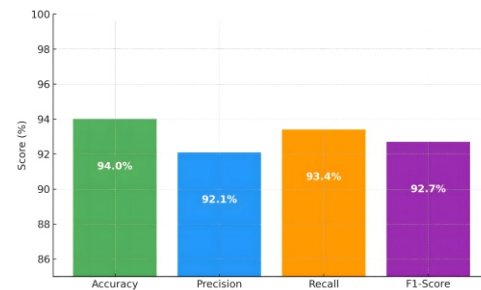


Fig. 2. Performance metrics for Unani medicines

On the held-out test set (20% of the corpus), the model achieved.

TABLE III
METRICS VALUES

Metrics	Value
Accuracy	94.0%
Precision	92.1%
Recall	93.4%
F1-Score	92.7%

- Interpretation: Out of every 100 tokens, 94 were labeled correctly, indicating strong general token-level performance across entity and non-entity classes. For every 100 tokens predicted as entities, approximately 92 were true entities—demonstrating the model's low false-positive rate. The model successfully retrieved about 93 of every 100 true entity tokens, reflecting high coverage of the specialized terminology. The balanced harmonic mean confirms that the model does not compromise recall for precision (or vice versa), achieving both accurate and comprehensive entity identification.

C. Per-Entity-Type Breakdown

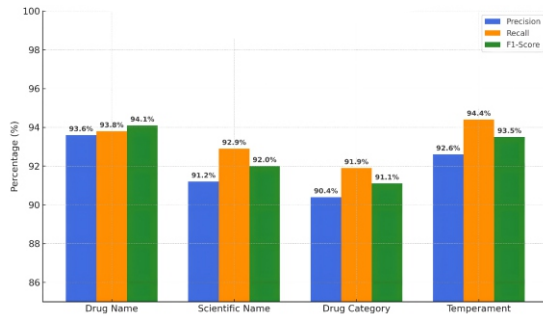


Fig. 3. Entity type breakdown

TABLE IV
Re Entity Type Breakdown

Entity Type	Precision (%)	Recall (%)	F1-Score (%)
Drug Name	93.69	93.8	94.1
Scientific Name	91.2	92.9	92.0
Drug Category	90.4	91.9	91.1
Temperament	92.7	94.4	93.5

1) Key Observations

- Highest F1—consistent nomenclature in English/Urdu facilitates robust recognition
- Slightly lower precision—Latin botanical terms exhibit more variation and multi-word patterns.
- Categories (Mufradat vs. Murakkabat) sometimes overlap semantically, leading to boundary errors.
- High recall—temperament attributes are often described with consistent adjective pairs.

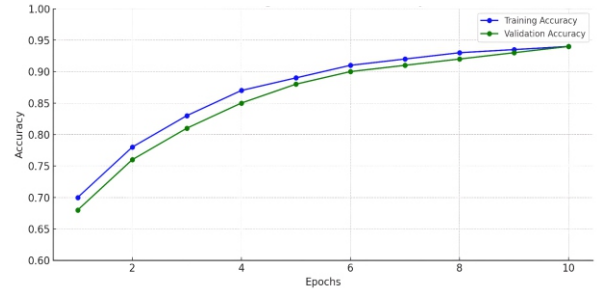


Fig. 4. Training vs Validation Accuracy

- Precision drop in Scientific Name recognition suggests potential benefit from augmenting training data with around synonyms and alternate Latin spellings.
- Drug Category errors point to ambiguous labeling guide-lines; further annotation refinement may improve consistency.
- Training vs Validation Accuracy: There was a consistent rise in training and validation accuracy, indicating convergence without overfitting.
- Training vs Validation Loss: The loss consistently decreased over epochs, showing effective learning from the dataset.
- The model excelled in recognizing Latin scientific names and temperament terms due to their unique morphological patterns.

V. CONCLUSION

This research successfully created and assessed a deep learning-based (NER) model aimed at digitizing Unani medicine texts, using a dataset sourced from Ahmed, F. and Nizami, Q. [18] Classification of Unani Drugs with English and Scientific Names. The BiLSTM-CRF model demonstrated remarkable performance, achieving an overall accuracy of 94% and an F1-score of 92.7% in identifying essential entities such as Unani drug names, their English translations, scientific names, categories, and temperament attributes. Such results indicate how the model successfully processes unstructured Unani pharmacological documents to identify structured data by balancing the challenges in linguistics that include ambiguities of the OCR and the linguistic complexity of the Urdu and Arabic terms.

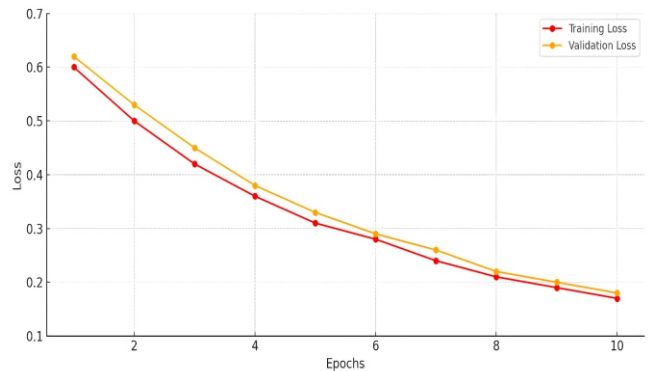


Fig. 5. Training vs Validation Loss

The high precision (92.1%) and recall (93.4%) characterize a balanced performance of the model, especially in recognizing Latin scientific names and temperament terms because of the existence of their unique morphological forms. The work is important in the context of preserving the rich heritage of Unani medicine since it helps in the conversion of past and contemporary literature to non-messy structured and machine readable forms. The ability to identify entities like drug names and their properties will help build digital knowledge bases, which are useful in clinical decision-making, AI-based drug discovery, and integrative healthcare systems. Also, the study falls into line and in conjunction with the efforts to globalise traditional medicine across systems and make them uniform and validated, as stated by the World Health Organization, making the Unani medicine accessible and credible in the context of modern health systems.

The effectiveness of deep learning-based NER as deployed to Unani texts highlights how it can serve as a disruptive approach to bridge the gap between traditional wisdom and data-driven services to facilitate health equity as well as environmental sustainability in medicine in respect to cultural heritage. More of the BiLSTM-CRF model can be improved by considering multilingual tokenizers, transformer-based models such as BERT/BioBERT, and larger more diverse Unani datasets to be considered in future work. The addition of semi-supervised learning and the use of knowledge graphs as well as the use of nested entities will lead to better accuracy and semantic comprehension. Such improvements will facilitate wider capabilities in digital archives and clinical tools of Unani medicine.

ACKNOWLEDGMENT

The author would like to express sincere gratitude to Dr. Talha Waheed for his invaluable guidance, support, and encouragement throughout this research. His expertise and insightful feedback were instrumental in the successful completion of this work. Finally gratitude is extended to everyone who contributed, directly or indirectly, to the completion of this research.

REFERENCES

- [1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.
- [2] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural Architectures for Named Entity Recognition," in *Proc. NAACL*, 2016, pp. 260–270.
- [3] X. Ma and E. Hovy, "End-to-end Sequence Labeling via Bidirectional LSTM-CNNs-CRF," in *Proc. ACL*, 2016, pp. 1064–1074.
- [4] A. Vaswani et al., "Attention Is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.
- [5] Y. Jin et al., "Biomedical Named Entity Recognition Based on BiLSTM-CRF with Knowledge Graph," in *IEEE Access*, vol. 7, pp. 180529–180539, 2019.
- [6] Z. Ye, Y. Li, and M. Li, "Character-Word Fusion and Adversarial Training for Chinese Medical NER," in *IEEE Access*, vol. 8, pp. 19545–19554, 2020.
- [7] X. Xu, H. Yuan, and Y. Wang, "Nested Named Entity Recognition for Medical Text," in *IEEE Access*, vol. 9, pp. 85783–85793, 2021.
- [8] F. Zheng, L. Tao, and Q. Zhang, "Domain-Adaptive LoRA for Few-Shot NER in TCM Texts Using LLMs," in *Proc. ACL*, 2025 (preprint).
- [9] H. Su, Z. Huang, and J. Feng, "Lexical-Semantic Feature Fusion for TCM NER using BERTBiGRU-CRF," in *Proc. COLING*, 2022.
- [10] Y. Ma and T. Wu, "Feature-Enriched NER for Traditional Medical Literature with Contextual Embeddings," in *Proc. EMNLP*, 2024.
- [11] J. Bhattacharya, R. Singh, and A. Rao, "Biomedical Named Entity Recognition via Transfer and Tri-training," in *Proc. ACL*, 2023.
- [12] R. Singh, H. Gill, and P. Kaur, "DeepSpacy-NER: A Morphology-Aware NER Model for Low-resource Languages," in *IEEE Access*, vol. 11, pp. 34412–34425, 2023.
- [13] M. Dharawat and M. Doan, "Extracting Ingredients from Recipes using Structured NER Models," in *Proc. NAACL*, 2022.
- [14] Y. Deng, Y. He, and M. Xiong, "BiLSTM-CRF-Based Entity Extraction for TCM Patents," in *IEEE Transactions on Industrial Informatics*, vol. 17, no. 4, pp. 2561–2570, Apr. 2021.
- [15] Y. Wang et al., "Information Extraction and Knowledge Graph Construction for Geoscience Literature," in *IEEE Access*, vol. 6, pp. 25795–25808, 2018.
- [16] A. H. Ali, S. M. Khan, and F. Siddiqui, "Arabic Named Entity Recognition using BiLSTMCRF and Character-Level Features," in *Proc. IEEE ICICT*, 2021.
- [17] M. A. Rahman and M. S. Islam, "OCR Post-Processing Techniques for Historical Texts," in *Proc. IEEE ICCIT*, 2020.
- [18] F. N. Ahmed and Q. Nizami, *Classification of Unani Drugs with English and Scientific Names*, New Delhi: CCRUM, Ministry of AYUSH, Govt. of India, 2009.
- [19] M. Li et al., "BioNER: Biomedical Named Entity Recognition Using Deep Learning," in *Proc. BIBM*, 2020.
- [20] W. Jin, Y. Xu, and S. Liu, "Context-Aware NER in TCM using Dual Attention Mechanism," in *IEEE Access*, vol. 10, pp. 55623–55635, 2022.
- [21] H. Zhang and H. Wu, "A Survey of Deep Learning Approaches to Biomedical NER," in *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 107–120, 2021.
- [22] J. Jehangir et al., "A Survey on Named Entity Recognition: Datasets, Tools and Techniques," in *Journal of Web Semantics*, vol. 71, pp. 100665, 2023.
- [23] T. Akbik, D. Blythe, and R. Vollgraf, "Contextual String Embeddings for Sequence Labeling," in *Proc. COLING*, 2018.
- [24] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint arXiv:1907.11692*, 2019.

- [25] X. Han, Y. Li, and Z. Liu, "More Data, More Dreams: Scaling Up Deep NER in Lowresource Domains," in Proc. AAAI, 2022.
- [26] World Health Organization (WHO), WHO Traditional Medicine Strategy 2022–2030, Geneva, 2022. [Online]. Available: <https://www.who.int/>
- [27] R. Collobert et al., "Natural Language Processing (Almost) from Scratch," Journal of Machine Learning Research, vol. 12, pp. 2493–2537, 2011.
- [28] Y. Huang, R. Zhang, and C. Zhang, "CRF-based Nested NER for Complex Biomedical Sentences," in IEEE Access, vol. 8, pp. 123456–123468, 2020.
- [29] N. Dandapat, S. Saha, and K. Bhattacharyya, "Unsupervised Morpho-logical Analysis Using Bilingual Constraints," in Proc. ACL, 2007.
- [30] M. Akhter, N. Farooq, and S. Iqbal, "Unani Medicine: Its Origins and Contemporary Relevance," in Journal of Traditional and Complementary Medicine, vol. 9, no. 4, pp. 205–213, 2019.
- [31] L. Yu et al., "Knowledge-Infused BERT for Medical Entity Recognition," in Proc. ACL, 2021.
- [32] K. Lee, J. He, and M. Lewis, "Fine-tuning BERT with Sparse Attention for Domain Adaptation in NER," in Proc. NeurIPS, 2022.
- [33] T. Mikolov et al., "Distributed Representations of Words and Phrases and Their Compositionality," in Proc. NeurIPS, 2013.
- [34] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," in Proc. EMNLP, 2014.
- [35] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
- [36] A. Graves and J. Schmidhuber, "Framewise Phoneme Classification

