

A Sequential Approach for Conditional Random Fields Based Entity Recognition and Routing in Mixture of Experts State Space Models

A PROJECT REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE AWARD OF THE DEGREE
OF
MASTER OF TECHNOLOGY by RESEARCH
IN
COMPUTER SCIENCE AND ENGINEERING

Submitted by

CHANDRESH PANDEY

2K24/RCO/06

Under the supervision of

Prof. Rahul Katarya



**COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY**

(Formerly Delhi College of Engineering) Bawana
Road, Delhi 110042

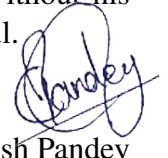
MAY, 2026

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY**

(Formerly Delhi College of Engineering) Bawana
Road, Delhi-110042

ACKNOWLEDGEMENT

I wish to express my sincerest gratitude to Prof. Rahul Katarya for his continuous guidance and mentorship that he provided me during the project. He showed me the path to achieve my targets by explaining all the tasks to be done and explained to me the importance of this project as well as its industrial relevance. He was always ready to help me and clear my doubts regarding any hurdles in this project. Without his constant support and motivation, this project would not have been successful.



Place: Delhi

Chandresh Pandey

Date: MAY, 2026

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY**

(Formerly Delhi College of Engineering) Bawana
Road, Delhi-110042

CANDIDATE'S DECLARATION

I, Chandresh Pandey, Roll No – 2K24/RCO/06, student of M.Tech by Research (Computer Science and Engineering), hereby declare that the project dissertation titled “A Sequential Approach for Conditional Random Fields Based Entity Recognition and Routing in Mixture of Experts State Space Models.”, which is submitted by me to the Department of Computer Science and Engineering, Delhi Technological University, Delhi, in partial fulfillment of the requirement for the award of the degree of Master of Technology by Research, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma, Associate ship, Fellowship, or other similar title or recognition.



Candidate's Signature

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor

Signature of External Examiner

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY**

(Formerly Delhi College of Engineering) Bawana
Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Dissertation titled “A Sequential Approach for Conditional Random Fields Based Entity Recognition and Routing in Mixture of Experts State Space Models.” which is submitted by Chandresh Pandey, Roll No – 2K24/RCO/06, Computer Science and Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology by Research, is a record of the project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

Prof. Rahul Katarya

Date: MAY, 2026

SUPERVISOR

A Sequential Approach for Conditional Random Fields Based Entity Recognition and Routing in Mixture of Experts State Space Models

by Chandresh Pandey

Submission date: 27-May-2026 06:15PM (UTC+0530)

Submission ID: 2970557188

File name: thesis_revision-5-1.pdf (4.43M)

Word count: 10215

Character count: 65904

A Sequential Approach for Conditional Random Fields Based Entity Recognition and Routing in Mixture of Experts State Space Models

Chandresh Pandey

ABSTRACT

Sequential learning systems have become increasingly important for modeling temporal and structured data in domains such as language processing, dynamical systems, and biomedical text mining. Recent architectures including State Space Models (SSMs) and Mixture-of-Experts (MoE) have improved scalability and long-range sequence modeling, while structured prediction methods such as Conditional Random Fields (CRFs) continue to provide effective sequence-level inference. The behavior of modular sequential architectures, by contrast, under heterogeneous conditions remains insufficiently understood.

This thesis presents a study of modular sequential learning systems through empirical analysis and application-oriented evaluation. The first part investigates Mixture-of-Experts based spectral State Space Models in controlled regime-switching environments containing chaotic, oscillatory, and noisy temporal patterns. Experimental results demonstrate routing collapse, instability, and inverse scaling behavior, where increasing the number of experts does not consistently improve performance. The second part presents a sequential Biomedical Named Entity Recognition framework combining BioBERT, Bidirectional LSTM, multi-head self-attention, and a Conditional Random Field decoder. Evaluated on the BioCreative V Chemical–Disease Relation dataset, the proposed model achieves an entity-level micro-averaged F1-score above 86% with improved sequence-level consistency.

The findings of this thesis show that modularity alone does not guarantee improved sequential learning performance. Instead, effective architectures depend on the alignment between model structure and the underlying characteristics of the data. The study highlights the importance of structured inference, stable routing mechanisms, and alignment-aware design for future sequential learning systems.

CHAPTER 1

INTRODUCTION

Sequential learning has become one of the central problems in modern machine learning due to its importance in modeling temporal, contextual, and structured data. Applications such as language modeling, biomedical text mining, forecasting, and dynamical system analysis all require models capable of learning dependencies across long sequences while preserving relevant contextual information. Recent advances in deep learning have improved the ability of neural architectures to process sequential data, particularly through transformer-based models and structured sequential frameworks [1]. As sequence modeling tasks become increasingly heterogeneous, and computationally demanding, the limitations of conventional architectures have become more apparent [2].

Recent developments make great use of SSMs as an efficient tool for long-range sequence modelling. SSMs offer efficient sequential computation without changing the representation of time, unlike the traditional recurrent architecture, which is highly ineffective. In parallel, MoE architectures have attracted interest for their modular computation, achieved through an expert specialization, and a routing component [3]. These are based on the idea that, for sequential data, multiple specialized functions should be able to model some of the data's variability. These architectures have been empirically proven competitive in various domains; however, their performance in case of dynamically changing conditions is still not well understood. However, while having more experts may lead to better results, it can also generate routing instabilities, expert collapses, and inconsistent specializations at training time, among others [4].

Structured prediction methods provide a different perspective on sequential modeling. Instead of introducing modularity within latent dynamics, approaches based on CRFs impose structure directly on output dependencies. Such methods remain highly effective in sequence labeling tasks where contextual consistency is important, particularly in BioNER. Biomedical text contains complex terminology, long entity spans, abbreviations, and domain-specific ambiguity, making accurate entity recognition a challenging problem. While transformer encoders provide strong contextual representations, independent token-level predictions often fail to maintain sequence-level consistency. Sequential architectures combining contextual encoders with sequential and structured decoding components therefore continue to achieve competitive performance in biomedical sequence labeling tasks [2].

In order to explore modular sequential architectures, the following two comple-

mentary lines of research have been explored in this thesis. The first part explores the application of MoE-based spectral SSMs in controlled environments with various chaotic, oscillatory and noisy temporal regimes [2]. Analysis methods include the study of routing behavior, stability and the interplay between modular decomposition and system’s dynamics. The second part introduces a sequential BioNER framework, which combines BioBERT, BiLSTM, self-attention, and structured decoding methods for biomedical entity recognition with CRF. The study of the two systems will allow the research to analyze the effects of modularity on sequential learning and its negative and positive influences on increasing an Architecture’s complexity. This work aims at grasping the impact of structural assumptions of sequential architectures on the learning behavior over different heterogeneous domains [5], [6], [1]. The results indicate that in addition to model capacity and/or model architecture, successful sequential modeling is also dependent on matching model structure with the nature of the underlying data.

1.1 Motivation

As the domains of applications for sequential and temporal data such as models of dynamical systems, biomedical text mining, and natural language processing grow in complexity, it is also becoming increasingly important to have efficient and reliable sequential learning architectures. Patterns evolve over time and in many modern applications it is required to model heterogeneous data distributions with patterns that behave differently in various situations with stable models representing these data distributions in the long-term. While recent progress in deep learning has increased the power of the models that can be represented and solved as sequences, current deep learning architectures are not always well suited for cases of dynamically changing temporal structures, noisy regimes, or structured prediction constraints.

New perspectives in SSMs and MoE architectures have emerged recently, which bring new opportunities to scalable sequential computing and modular learning. These techniques try to address the above problem through breaking down a problem into smaller, specialized tasks (components). Although modules are good across the board in some tasks, they are often found to be unstable in other routing tasks, exhibiting representation collapse, expert specialization inconsistencies and performance degradation when coping with diverse conditions. Structured prediction tasks, like BioNER, also demand for architectures that are able to maintain a sequence-level consistency, to deal with contextual ambiguity, and to capture the complex entity boundaries. Structured prediction tasks (like BioNER) simultaneously demand architectures capable of maintaining sequence-level consistency, having the ability to handle contextual ambiguity and to capture the complex entity boundaries. This underlines the importance of understanding when decomposition in modules would benefit learning behaviors and when modulating more will be in these ways counterproductive.

The objective of this research is to investigate the efficacy of modular sequential architectures both in temporal modelling as well as in structured prediction. This study explores MoE-based spectral SSMs in controlled regime switching environments, and proposes a sequential BioNER framework that combines the BioNER-based contextual

encoding, sequential modeling, attention mechanisms, and CRF-based decoding. The ultimate goal of the study is to gain deeper insights into these types of stability and routing behavior, and to help build more reliable and alignment-aware architectures for future sequential learning applications.

1.2 Objectives

The principal aims of this study are:

- To study the behavior of modular sequential learning architectures under heterogeneous conditions, investigate the effectiveness of MoE-based spectral SSMs for regime-switching temporal modeling, and analyze routing behavior, expert specialization, and inverse scaling in modular sequential systems.
- To develop a sequential BioNER framework integrating BERT, BiLSTM, self-attention, and CRF-based structured decoding, evaluate model performance using standard metrics such as precision, recall, F1-score, and NMSE, and perform comparative and error-based analysis for understanding the strengths and limitations of sequential architectures.
- To explore alignment-aware design principles for improving stability and structured inference in future sequential learning systems.

1.3 Challenges

There are several challenges associated with designing efficient and reliable sequential learning architectures for heterogeneous temporal and structured data:

- Sequential data often contains dynamically changing temporal patterns, making stable long-range modeling difficult, and modeling heterogeneous regimes such as chaotic, oscillatory, and noisy temporal patterns within a unified framework is computationally challenging.
- MoE-based architectures frequently suffer from routing instability, expert collapse, and inconsistent specialization during training, and increasing the number of experts does not always improve performance and may lead to inverse scaling behavior.
- Biomedical sequence labeling tasks contain contextual ambiguity, long entity spans, abbreviations, and imbalanced label distributions, and independent token-level predictions may fail to maintain sequence-level consistency in BioNER tasks.

This research aims to address these challenges through the study of modular spectral SSMs and the development of a sequential BioNER framework combining contextual encoding, sequential modeling, attention mechanisms, and CRF-based structured decoding.

1.4 Thesis Organization

- **Chapter 1: Introduction** Introduces the fundamental concepts of sequential learning, modular architectures, and structured prediction. The chapter presents the motivation, objectives, research challenges, and overall scope of the thesis within the context of heterogeneous temporal modeling and biomedical sequence labeling.
- **Chapter 2: Literature Review** — Reviews existing research on sequential learning architectures including SSMs, MoE frameworks, transformer-based models, and CRF-based structured prediction methods. The chapter also discusses recent developments in BioNER and highlights limitations related to routing instability, expert collapse, and sequence-level consistency.
- **Chapter 3: Proposed Work** — Describes the datasets, preprocessing pipeline, modular spectral SSM framework, MoE routing mechanism, proposed GMS architecture, and the BioNER framework integrating BioBERT, BiLSTM, self-attention, and CRF-based structured decoding.
- **Chapter 4: Results and Discussion** — Presents the experimental setup, evaluation metrics, routing analysis, phase-space reconstruction studies, comparative performance evaluation, BioNER benchmarking, ablation analysis, and detailed discussion of the observed findings across both sequential learning tasks.
- **Chapter 5: Conclusion and Future Scope** — Summarizes the major findings of the study, discusses the limitations of modular sequential architectures, and outlines future research directions related to stability-aware routing, geometry-aware evaluation, structured inference, and efficient sequential learning systems.

38 CHAPTER 2

LITERATURE REVIEW

2.1 Literature Survey

Sequential learning and structured prediction have recently achieved intensive focus, because of their use in natural language processing, temporal prediction and biomedical text mining. Previous sequential models mainly employed hand-crafted feature engineering approaches along with recurrent neural networks. These techniques could model context at short-range, but were often hard to scale and had poor performance in modelling long-range context, or in coping with weak gradients. New transformer architectures brought a significant advancement to the study of contextual learning, but their high quadratic complexity posed problems for long-sequence processing and for the deployment while keeping a limited resource footprint.

The recent progress in SSM has opened the possibility of alternative methods to efficiently model long range sequence data. Other architectures like Mamba and spectral SSM related architectures achieved effective sequential learning in the chosen models with less computation as compared to the classical transformer based models [3]. At the same time, MoE architectures introduced modular computation through routing mechanisms and expert specialization, enabling larger model capacity without proportional increases in computation. Several recent studies combined MoE with SSM-based architectures to improve scalability and sequential representation learning [7], [8]. Researchers have also reported issues such as routing instability, expert collapse, inconsistent specialization, and degradation under heterogeneous conditions. Studies on sparse MoE routing further showed that increasing the number of experts does not always guarantee improved performance and may lead to representation collapse during training [9].

In parallel, structured prediction methods based on CRFs continued to remain effective for sequence labeling tasks, particularly in BioNER. Earlier approaches relied heavily on handcrafted linguistic features and statistical sequence models, but modern systems increasingly combine contextual encoders with sequential and structured decoding components [10]. Sequential architectures integrating transformer encoders, BiLSTM layers, attention mechanisms, and CRF-based decoding achieved strong performance in biomedical entity recognition by improving contextual understanding and maintaining sequence-level consistency. Recent studies demonstrated that CRF-based

structured decoding still plays an important role in handling complex biomedical terminology, long entity spans, and contextual ambiguity [11].

From the reviewed literature, it can be observed that both modular sequential architectures and structured prediction frameworks offer important advantages for complex sequence modeling tasks. Issues related to routing behavior, stability, heterogeneous temporal dynamics, and structured consistency remain open research challenges. These observations form the foundation of the present study, which investigates modular spectral SSMs and sequential CRF-based BioNER architectures within a unified sequential learning framework. Table 2.1 summarises representative recent work on modular sequential and MoE-based SSM architectures, while Table 2.2 lists representative BioNER systems evaluated on the BCSCDR corpus together with their reported F_1 values and limitations.

Table 2.1: Summary of existing research on modular sequential architectures and MoE-based SSMs.

Reference & Year	Objective	Architecture / Dataset	Key Findings	Limitations
Gu et al., 2021	Efficient long-range sequence modeling	S4 / Long Range Arena	Strong long-sequence performance	High implementation complexity
Dao & Gu [3], 2023	Linear-time selective sequence modeling	Mamba / Language modeling benchmarks	Efficient scalable sequential learning	Limited analysis of heterogeneous dynamics
Fedus et al. [8], 2021	Sparse expert scaling for transformers	Switch Transformer	Improved scalability through sparse routing	Routing instability and expert imbalance
Lepikhin et al. [7], 2020	Large-scale MoE training	GShard / Multilingual translation	Increased model capacity with sparse experts	Communication overhead and routing collapse

Table 2.2: Recent BioNER systems evaluated on the BC5CDR corpus.

Reference & Year	Aim	Model / Dataset	Results (F1 %)	Limitations
Keloth et al., 2024	Instruction-tuned generation	BioNER-LLaMA (LLaMA-7B) / NCBI, BC5CDR, BC2GM	Chemical: 92.6, Disease: 86.9	High cost, LLMs
Gou & Jie, 2023	Lightweight BERT	LWNER (BERT + CNN + BiLSTM + CRF) / BC5CDR	Chemical: 91.29, Disease: 81.23	Low disease F1
Zhang & Chen [11], 2022	Multi-task + syntax	BioBERT + POS / Syn / Dep Attn / Multi-datasets	Chemical: 94.26, Disease: 87.76	Complexity
Sun et al., 2021	MRC reformulation	BioBERT / BlueBERT / Clinical-BERT-MRC / BC5CDR	Chemical: 94.19, Disease: 87.83	Templates, training
Lee et al. [17], 2020	Domain pre-training	BioBERT / BC5CDR	Chemical: 93.47, Disease: 87.15	No CRF, labels

2.2 Dataset

The datasets used in this research correspond to two different sequential learning problems: heterogeneous temporal sequence modeling and biomedical named entity recognition. For the analysis of modular sequential architectures, synthetic regime-switching temporal datasets were generated to simulate heterogeneous dynamical environments. The generated sequences contain chaotic, oscillatory, and noise-dominated temporal regimes in order to evaluate the stability and routing behavior of MoE-based spectral SSMs under changing sequential conditions. Chaotic sequences were generated using the Mackey–Glass system, while oscillatory and autoregressive noise regimes were introduced to create controlled temporal transitions. The combined dataset allows evaluation of long-range sequential modeling, expert specialization, and routing consistency under heterogeneous dynamics [12].

The generated temporal sequences were divided into training and testing sets while maintaining consistent regime distributions across both splits. All sequences were normalized before training to improve numerical stability and reduce scale-related bias during optimization. The dataset design was selected specifically to analyze whether modular decomposition through expert routing improves sequential modeling performance in dynamically changing environments. Since the dataset contains multiple temporal regimes with significantly different characteristics, it provides an appropriate benchmark for studying routing collapse, inverse scaling, and stability issues in modular SSM architecture [15].

For the BioNER task, the BioCreative V Chemical–Disease Relation (BC5CDR) dataset was used [13]. The dataset consists of annotated PubMed abstracts containing chemical and disease entities labeled using standard biomedical annotation guidelines. The corpus contains manually curated biomedical terminology with complex entity boundaries, abbreviations, and contextual ambiguity, making it suitable for evaluating structured sequence labeling systems. The dataset is divided into predefined training,

validation, and testing splits, ensuring consistent evaluation across different BioNER models. In this work, the textual sequences are processed using a sequential framework combining BioBERT, BiLSTM, self-attention, and CRF-based structured decoding for biomedical entity recognition. Table 2.3 additionally lists alternative biomedical NER corpora considered for context, and representative samples of the temporal regimes and annotated biomedical entities used in the study are illustrated in Fig. 2.1.

Table 2.3: Publicly available biomedical NER corpora considered in this study.

Dataset (Year)	Description	Entities (Count & Type)
BC5CDR [13], 2016	1,500 PubMed articles annotated with chemicals, diseases, and interactions	2 (Chemical, Disease)
NCBI-disease [14], 2014	793 PubMed abstracts with disease mentions and concepts	1 (Disease)
BC2GM [15], 2007	20,000 sentences annotated for gene and protein names	1 (Gene/Protein)
GENIA [16], 2003	2,000 MEDLINE abstracts annotated with 36 biological entities	36 (Protein, DNA, RNA, Cell, etc.)

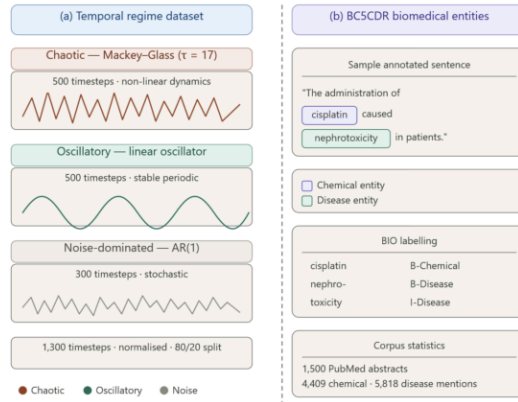


Figure 2.1: Sample temporal regimes and annotated biomedical entity examples used in the study.

2.3 Problem Statement

Modeling heterogeneous sequential data remains a challenging problem in modern machine learning. Sequential datasets often contain dynamically changing temporal

patterns, long-range dependencies, noisy transitions, and contextual ambiguity, making accurate representation learning difficult. Traditional sequential architectures frequently struggle to maintain stability and consistent performance when the underlying data distribution changes across time. Although recent architectures based on SSMs and MoE have improved scalability and long-sequence processing, several limitations still remain unresolved. Studies have reported routing instability, representation collapse, poor expert specialization, and inconsistent performance scaling in modular sequential systems.

Another important challenge exists in structured sequence labeling tasks such as BioNER. Biomedical text contains complex terminology, long entity spans, abbreviations, and highly imbalanced label distributions, making entity recognition difficult even with powerful contextual encoders. Independent token-level prediction approaches often fail to preserve sequence-level consistency, leading to boundary errors and ambiguous entity labeling. While transformer-based architectures provide strong contextual understanding, additional sequential and structured decoding mechanisms are often required for reliable biomedical entity recognition.

The primary problem addressed in this thesis is the development and analysis of efficient modular sequential learning frameworks for heterogeneous temporal modeling and structured prediction tasks. Instead of relying solely on large end-to-end architectures, the proposed approach investigates MoE-based spectral SSMs for temporal sequence modeling and a sequential BioNER framework combining BioBERT, BiLSTM, self-attention, and CRF-based structured decoding. The objective is to improve sequential learning stability, structured consistency, and overall performance while analyzing the limitations of modular decomposition under heterogeneous conditions.

CHAPTER 3

PROPOSED WORK

12

3.1 Research Design

The present study follows an empirical research approach to investigate modular sequential learning architectures under heterogeneous sequential environments. The work focuses on two complementary problems: temporal sequence modeling using MoE-based spectral SSMs and structured sequence labeling using a sequential BioNER framework. The proposed work is designed to analyze both the effectiveness and limitations of modular sequential architectures through controlled experimentation, comparative evaluation, and structured performance analysis.

The overall research workflow is divided into multiple stages beginning with dataset preparation and preprocessing, followed by model design, training, evaluation, and comparative analysis. For temporal sequence modeling, heterogeneous regime-switching datasets containing chaotic, oscillatory, and noise-dominated patterns are generated in order to evaluate routing behavior, expert specialization, and stability within modular spectral SSM architectures. For the BioNER task, biomedical textual sequences from the BCSCDR dataset are processed using contextual encoding, sequential modeling, attention mechanisms, and CRF-based structured decoding. The complete research workflow, including dataset preparation, preprocessing, model design, training, and evaluation, is illustrated in Fig. 3.1.

The overall research process is structured into the following stages:

1. **Data Preparation** — Generating heterogeneous temporal sequences for modular SSM evaluation and preparing biomedical textual datasets for BioNER experiments.
2. **Data Preprocessing** — Normalizing sequential data, tokenizing biomedical text, and preparing structured input representations for training.
3. **Model Design** — Developing MoE-based spectral SSM architectures for temporal modeling and designing a sequential BioNER framework integrating BioBERT, BiLSTM, self-attention, and CRF-based decoding.
4. **Model Training** — Training the proposed architectures under controlled experimental settings using multiple evaluation configurations and optimization

strategies.

5. **Evaluation and Analysis** — Evaluating model performance using metrics such as NMSE, precision, recall, and F1-score, followed by routing analysis, ablation studies, and comparative performance assessment.

This structured work enables systematic experimentation, reproducibility, and comprehensive evaluation of modular sequential learning architectures under heterogeneous conditions.

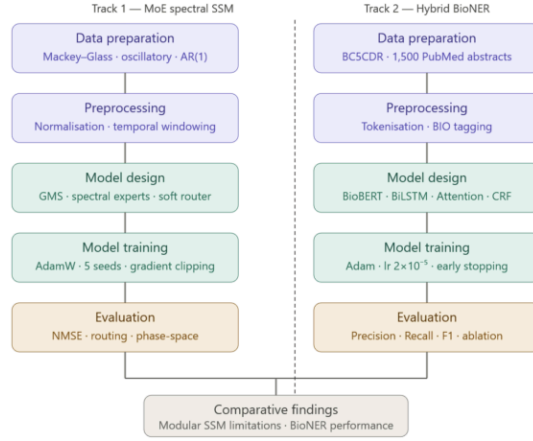


Figure 3.1: Flowchart illustrating the complete proposed work and experimental workflow.

3.2 Dataset Description

The datasets used in this research correspond to two different sequential learning tasks: heterogeneous temporal sequence modeling and biomedical named entity recognition.

For temporal modeling experiments, synthetic regime-switching sequential datasets were generated to evaluate the behavior of modular spectral SSM architectures under dynamically changing conditions. The generated sequences consist of chaotic, oscillatory, and noise-dominated temporal regimes combined within a unified sequential environment. Chaotic behavior was generated using the Mackey-Glass system, while oscillatory and autoregressive noise sequences were incorporated to simulate heterogeneous temporal transitions. The dataset was specifically designed to analyze routing stability, expert specialization, and performance scaling in MoE-based sequential architectures.

The temporal dataset was divided into training and testing sequences while maintaining balanced representation of all temporal regimes across both splits. All sequential data was normalized before training in order to improve numerical stability and optimization consistency. The diversity of temporal patterns within the dataset makes it suitable for studying heterogeneous sequence modeling and modular routing behavior under controlled experimental conditions.

For the BioNER task, the BC5CDR dataset was used, which is a publicly available biomedical corpus developed for the BioCreative V challenge. The dataset consists of 1,500 annotated PubMed abstracts containing chemical and disease entity mentions together with manually curated biomedical annotations. The corpus includes complex biomedical terminology, long entity spans, abbreviations, and contextual ambiguity, making it highly suitable for evaluating structured sequence labeling systems. The dataset is divided into predefined training, validation, and testing splits to ensure consistent experimental evaluation.

- Chaotic temporal sequences represent complex non-linear dynamical behavior generated using the Mackey–Glass system.
- Oscillatory sequences correspond to stable periodic temporal patterns.
- Noise-dominated sequences simulate stochastic temporal behavior under autoregressive conditions.
- BC5CDR biomedical sequences contain annotated chemical and disease entities extracted from PubMed abstracts.

The selected datasets provide both heterogeneous temporal environments and real-world biomedical sequence labeling challenges, enabling comprehensive evaluation of modular sequential learning architectures.

24

3.3 Data Preprocessing

Data preprocessing plays an important role in preparing sequential and biomedical data for efficient model training and stable learning behavior. Since the proposed research involves both temporal sequence modeling and biomedical named entity recognition, separate preprocessing strategies were applied for each task in order to improve consistency, reduce noise, and ensure compatibility with the proposed architectures.

For the temporal modeling experiments, all generated sequential data was normalized before training to reduce scale variation across different temporal regimes. Chaotic, oscillatory, and noise-dominated sequences were converted into fixed-length temporal windows suitable for sequential learning. The generated sequences were then divided into training and testing sets while maintaining balanced representation of all temporal regimes. Temporal normalization improved numerical stability during optimization and reduced instability caused by heterogeneous sequence magnitudes.

For the BioNER task, preprocessing involved tokenization of biomedical text sequences using the BioBERT tokenizer. Annotated biomedical entities were converted

into BIO tagging format to preserve sequence-level label consistency during training. Long biomedical abstracts were segmented into manageable token sequences while maintaining contextual continuity. Padding and attention masks were applied to ensure uniform sequence lengths for batch-based training. In addition, special biomedical abbreviations and domain-specific terminology were preserved during preprocessing to avoid loss of contextual information.

- Sequence normalization and temporal window generation for heterogeneous sequential data.
- Tokenization and BIO tag conversion for biomedical textual sequences.
- Preparation of fixed-length input representations and attention masks.
- Separation of training, validation, and testing datasets while maintaining balanced distributions.
- Formatting sequential data for modular SSM architectures and sequential CRF-based decoding frameworks.

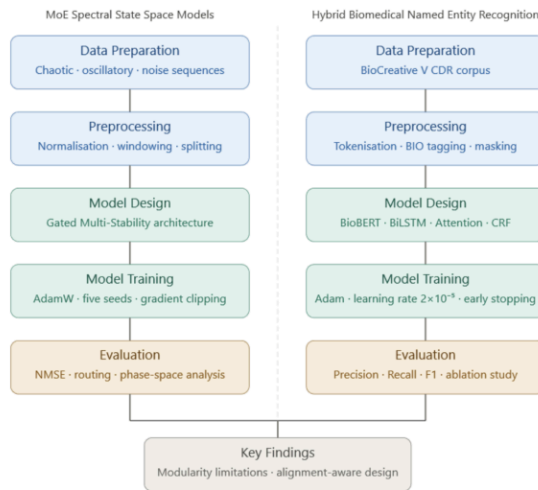


Figure 3.2: Preprocessing pipeline for temporal sequence modeling and biomedical named entity recognition.

After preprocessing, the generated sequential representations and biomedical token sequences were used as inputs for modular spectral SSM training and sequential BioNER model evaluation. The complete preprocessing pipeline for both temporal and

biomedical inputs is illustrated in Fig. 3.2, and ensures consistent data representation, stable optimization, and efficient learning across both sequential modeling tasks.

3.4 Model Architectures and Working Principles

3.4.1 Spectral State Space Models (SSMs)

SSMs have recently emerged as an efficient alternative for long-range sequential modeling. Unlike conventional recurrent architectures, SSMs model temporal dependencies through structured state transition dynamics, allowing stable sequence processing with lower computational complexity. These models are particularly effective for long sequential inputs because they maintain continuous latent state representations while avoiding the quadratic complexity associated with transformer attention mechanisms. Recent sequential architectures such as Mamba and related spectral SSM variants achieved competitive performance across language modeling and temporal forecasting tasks by combining efficient recurrence with stable long-range dependency modeling.

In the proposed work, spectral SSMs are used as the foundational sequential operators for heterogeneous temporal modeling. The hidden state evolves according to a linear state transition process in which the current state depends on the previous latent representation¹ and the current input sequence [3]. The continuous-time state-space formulation can be represented as:

$$\dot{h}(t) = Ah(t) + Bx(t), \quad y(t) = Ch(t) + Dx(t) \quad (3.1)$$

where $h(t)$ represents the latent hidden state, $x(t)$ denotes the input sequence, $y(t)$ represents the output sequence, and A, B, C, D correspond to learnable state transition matrices.

For practical sequential learning, the continuous formulation is discretized into a recurrent update process:

$$h_t = \bar{A}h_{t-1} + \bar{B}x_t, \quad y_t = \bar{C}h_t + \bar{D}x_t \quad (3.2)$$

⁹ where h_t denotes the hidden state at time step t , x_t represents the input at time step t , and y_t is the predicted output sequence. The overall spectral SSM architecture, with its latent state transitions and sequential recurrence, is shown in Fig. 3.3.

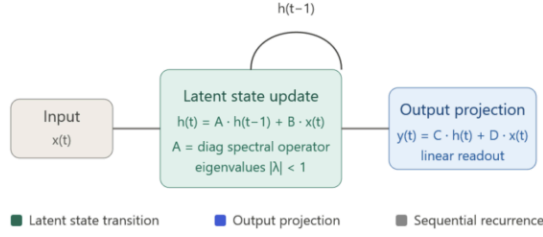


Figure 3.3: Spectral State Space Model architecture illustrating latent state transitions and sequential recurrence.

3.4.2 Mixture-of-Experts Routing Mechanism

MoE architectures introduce modular computation by dividing a model into multiple specialized expert networks and a routing mechanism that determines which expert should process a given input. Instead of activating the entire model for every sequence element, the router selectively assigns inputs to a subset of experts, improving computational efficiency and increasing model capacity. Recent studies combining MoE with SSM architectures demonstrated that modular routing can improve scalability and long-sequence processing while maintaining efficient inference.

In the proposed work, the routing mechanism is used to dynamically combine multiple spectral SSM experts for heterogeneous temporal modeling. Each expert is designed to capture different temporal behaviors such as chaotic dynamics, oscillatory patterns, or stochastic noise. The router observes the current sequential state and produces routing probabilities that determine the contribution of each expert during sequence processing. This enables the architecture to adaptively allocate computational focus according to the characteristics of the input sequence.

The routing operation can be represented as:

$$y = \sum_{k=1}^K g_k(x) f_k(x) \quad (3.3)$$

where $f_k(x)$ represents the output of the k^{th} expert, $g_k(x)$ denotes the routing weight assigned to that expert, and K represents the total number of experts.

The routing probabilities are generated using the softmax function:

$$g_k(x) = \frac{e^{h_k(x)}}{\sum_{j=1}^K e^{h_j(x)}} \quad (3.4)$$

where $h_k(x)$ represents the routing score produced for expert k and $g_k(x)$ corresponds to the normalized routing probability.

The routing mechanism allows the model to distribute sequential processing across multiple experts while preserving efficient computation. Previous studies have reported challenges such as routing collapse and inconsistent expert specialization,

where a single expert dominates the routing distribution during training. The proposed research investigates these behaviors under heterogeneous temporal environments in order to analyze the stability and effectiveness of modular sequential learning systems. The router-expert interaction and sequential expert allocation used in the proposed framework are depicted in Fig. 3.4.

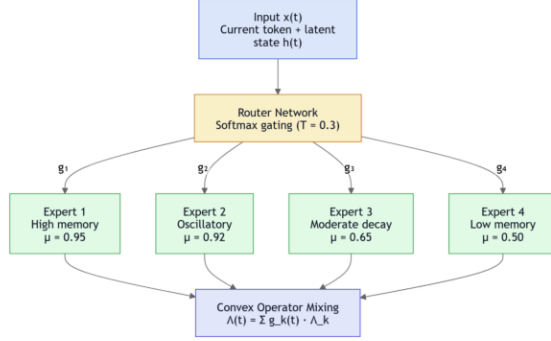


Figure 3.4: Mixture-of-Experts routing architecture showing router-expert interactions and sequential expert allocation.

3.4.3 Gated Multi-Stability (GMS) Architecture

The proposed Gated Multi-Stability (GMS) architecture combines spectral SSMs with an MoE-based routing mechanism in order to model heterogeneous temporal dynamics more effectively. Instead of relying on a single sequential operator, the architecture utilizes multiple spectral experts where each expert is expected to capture different temporal behaviors such as chaotic, oscillatory, or stochastic patterns [8]. The routing mechanism dynamically assigns importance weights to the experts according to the characteristics of the input sequence, enabling adaptive sequential computation across varying temporal regimes [9].

In the proposed framework, each expert consists of an independent spectral SSM operator with stable state transition dynamics. The router observes the current sequential representation and generates routing probabilities that determine the contribution of each expert during temporal processing. The weighted outputs from all experts are then combined to produce the final sequential representation. This modular design increases representational flexibility while preserving efficient sequential computation. Previous studies on sparse MoE systems also report challenges such as routing instability and expert collapse during training.

The effective state transition operator of the GMS architecture is represented as:

$$\overline{A}_{\text{mix}}(t) = \sum_{k=1}^K g_k(t) \overline{A}_k \quad (3.5)$$

where $\bar{A}_k$ represents the spectral state transition operator of the k^{th} expert, $g_k(t)$ denotes the routing weight assigned at time step t , and K represents the total number of experts.

The final sequential representation is obtained through adaptive expert mixing, allowing the model to process heterogeneous temporal patterns within a unified framework. The overall GMS architecture with its spectral experts, routing mechanism, and adaptive sequential mixing is shown in Fig. 3.5. The proposed GMS architecture is specifically designed to analyze routing behavior, expert specialization, stability, and inverse scaling under regime-switching sequential environments.

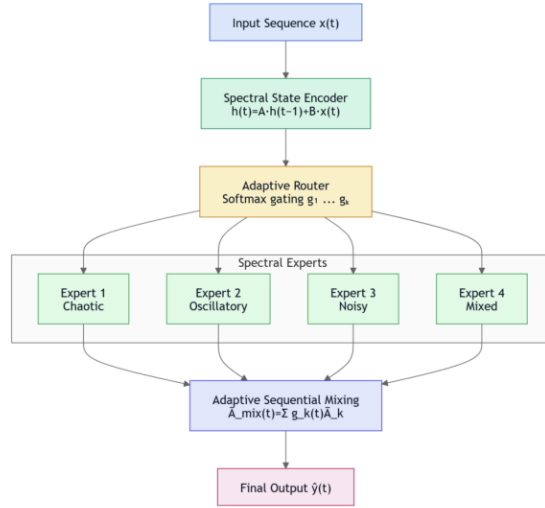


Figure 3.5: Proposed Gated Multi-Stability architecture showing spectral experts, routing mechanism, and adaptive sequential mixing.

3.4.4 BioBERT Architecture

BioBERT is a domain-specific transformer architecture developed for biomedical text mining and natural language processing tasks. Unlike traditional language models trained on general-domain corpora, BioBERT is pre-trained on large-scale biomedical literature including PubMed abstracts and PMC full-text articles. This enables the model to better understand biomedical terminology, contextual relationships, abbreviations, and domain-specific semantic patterns commonly present in biomedical text. BioBERT has been shown effective across several biomedical NLP tasks including BioNER, relation extraction, and question answering.

In the proposed work, BioBERT is used as the contextual embedding layer for biomedical sequence representation [17]. The input biomedical text is first tokenized

and converted into contextual embeddings using the transformer encoder. Unlike static word embeddings, BioBERT generates context-aware representations where the meaning of a token depends on surrounding biomedical context. This improves the model's ability to recognize complex entity boundaries and ambiguous biomedical terminology.

The contextual representation generated by BioBERT can be expressed as:

$$H = \text{BioBERT}(X) \quad (3.6)$$

where X represents the input biomedical token sequence and H denotes the contextual embedding representation generated by BioBERT.

The generated contextual embeddings are then forwarded to sequential and structured decoding components including BiLSTM, self-attention, and CRF layers for biomedical entity recognition [18]. The complete sequential BioNER framework, integrating BioBERT, BiLSTM, self-attention, and CRF-based decoding, is shown in Fig. 3.6. The use of BioBERT enables the proposed framework to capture semantic and contextual relationships more effectively than conventional embedding methods, improving sequence-level understanding in biomedical text processing tasks.

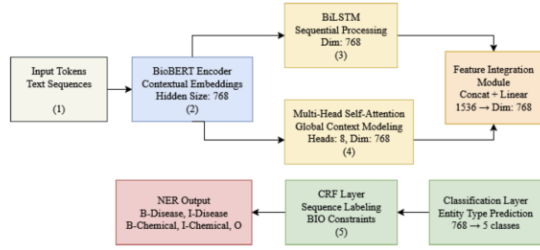


Figure 3.6: Proposed sequential BioNER framework combining BioBERT, BiLSTM, self-attention, and CRF-based structured decoding.

3.4.5 Sequential and Structured Decoding Components

The contextual embeddings generated by BioBERT are further processed using sequential and structured decoding layers in order to improve biomedical entity recognition performance. Although transformer encoders provide strong contextual understanding, independent token-level predictions may fail to preserve sequential consistency across biomedical entities. To address this limitation, the proposed framework combines BiLSTM, multi-head self-attention, and CRF-based structured decoding for improved sequence labeling performance.

25

Bidirectional Long Short-Term Memory (BiLSTM)

BiLSTM is used to capture sequential dependencies from both forward and backward directions within biomedical text sequences. Unlike conventional recurrent networks, BiLSTM preserves contextual information from past and future tokens simultaneously, allowing better understanding of long biomedical entity spans and contextual relationships. The hidden representations generated from both directions are combined to form a unified sequential representation [5].

The hidden state computation in BiLSTM is represented as:

$$\begin{aligned} \vec{h}_t &= \text{LSTM}(x_t, \vec{h}_{t-1}), & \overleftarrow{h}_t &= \text{LSTM}(x_t, \overleftarrow{h}_{t+1}), & h_t &= [\vec{h}_t; \overleftarrow{h}_t] \end{aligned} \quad (3.7)$$

where x_t denotes the input token embedding at time step t , $\vec{h}_t$ and $\overleftarrow{h}_t$ correspond to the forward and backward LSTM hidden states, and h_t is the combined contextual representation obtained by concatenating both directional hidden states.

BiLSTM improves contextual continuity and enables the model to better capture long-range dependencies in biomedical text sequences.

Multi-Head Self-Attention

The self-attention mechanism is incorporated to improve contextual representation by assigning different importance weights to tokens within a sequence [19]. Multi-head attention allows the model to learn multiple contextual relationships simultaneously, improving recognition of complex biomedical entities and ambiguous terminology.

The attention mechanism is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \quad (3.8)$$

where Q represents the query matrix, K denotes the key matrix, V corresponds to the value matrix, and d_k represents the dimensionality scaling factor.

The attention layer enables the model to focus on contextually important biomedical tokens during sequence processing.

Conditional Random Field (CRF)

The CRF layer is used as the final structured decoding component for biomedical entity recognition. Unlike independent token-level classification, CRF considers dependencies between adjacent output labels and ensures valid sequence transitions during prediction [20]. This improves sequence-level consistency and reduces invalid BIOES tag combinations.

The conditional probability of the output label sequence is represented as:

$$P(y | x) = \frac{e^{s(x,y)}}{\sum_{y' \in Y_x} e^{s(x,y')}} \quad (3.9)$$

where $s(x, y)$ represents the sequence score for input sequence x and label sequence y , and Y_x denotes the set of all possible label sequences.

The CRF decoder improves boundary detection and structured consistency for biomedical entity recognition tasks involving long entity spans and contextual ambiguity.

Sequential BioNER Framework

The proposed Sequential BioNER framework combines contextual embedding, sequential modeling, attention mechanisms, and structured decoding in a unified architecture for biomedical entity recognition. The framework integrates BioBERT, BiLSTM, multi-head self-attention, and CRF layers in order to capture semantic context, sequential dependencies, and sequence-level label consistency simultaneously. This sequential design enables the model to effectively recognize complex biomedical entities containing long spans, abbreviations, and contextual ambiguity.

In the proposed architecture, biomedical text sequences are first processed using BioBERT to generate contextual embeddings. These embeddings are then forwarded to the BiLSTM layer for bidirectional sequential modeling. The multi-head self-attention mechanism further refines the contextual representation by assigning importance weights to relevant biomedical tokens. Finally, the CRF decoder predicts the optimal label sequence while enforcing valid BIO tag transitions during structured decoding.

The overall sequential representation can be expressed as:

$$Y = \text{CRF}(\text{Attention}(\text{BiLSTM}(\text{BioBERT}(X)))) \quad (3.10)$$

where X represents the input biomedical token sequence and Y denotes the predicted entity label sequence.

The integration of contextual encoding and structured decoding improves both token-level understanding and sequence-level consistency. The proposed sequential framework is specifically designed to reduce boundary prediction errors and improve recognition performance for complex biomedical entities. In addition, the architecture enables effective learning from domain-specific biomedical corpora while maintaining stable sequential representations during training. Similar sequential transformer-CRF architectures have achieved competitive results in biomedical sequence labeling tasks by combining contextual semantics with structured inference.

3.5 Training and Evaluation Strategy

The proposed architectures are trained and evaluated separately for heterogeneous temporal modeling and biomedical named entity recognition tasks. For the modular spectral SSM experiments, the generated sequential datasets are divided into training and testing sequences while maintaining balanced representation of chaotic, oscillatory, and noise-dominated temporal regimes. Each experiment is conducted under controlled settings using different numbers of experts in order to analyze routing behavior, expert specialization, stability, and inverse scaling. The models are optimized using gradient-based learning with normalized sequential inputs to ensure stable convergence during training.

For the BioNER task, the BC5CDR dataset is divided into predefined training, validation, and testing splits. Biomedical text sequences are processed using the proposed sequential framework consisting of BioBERT, BiLSTM, multi-head self-attention, and CRF-based structured decoding. The contextual embeddings generated by BioBERT are fine-tuned during training while the sequential and structured decoding layers learn biomedical entity dependencies and valid BIO transitions. Recent studies on sequential transformer-CRF frameworks have shown that combining contextual embeddings with structured decoding improves biomedical entity recognition performance and sequence consistency.

To evaluate the effectiveness of the proposed models, multiple evaluation metrics are employed for both tasks:

- Normalized Mean Squared Error (NMSE) is used to evaluate temporal prediction quality for spectral SSM experiments, and routing distribution analysis is performed to examine expert specialization and routing collapse in modular architectures.
- Precision measures the proportion of correctly predicted entities among all predicted entities, recall evaluates the ability of the model to identify all relevant biomedical entities, and F1-score provides a balanced measure between precision and recall.
- Ablation studies are conducted to analyze the contribution of BioBERT, BiLSTM, self-attention, and CRF components within the sequential framework.

The evaluation strategy enables systematic comparison of modular sequential architectures across heterogeneous temporal modeling and biomedical sequence labeling tasks.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Experimental Setup

All experiments in this research were conducted using GPU-accelerated environments for efficient training and evaluation of sequential learning models. The proposed modular spectral SSM architectures and sequential BioNER framework were implemented using PyTorch and related deep learning libraries. For the temporal modeling experiments, synthetic heterogeneous sequential datasets were generated under controlled settings containing chaotic, oscillatory, and noise-dominated temporal regimes. The chaotic sequences were generated using the Mackey–Glass system with delay $\tau = 17$, while oscillatory and autoregressive noise regimes were introduced to simulate heterogeneous temporal transitions.

The generated sequential datasets were divided into training and testing splits while maintaining balanced representation across all temporal regimes. All sequences were normalized to zero mean and unit variance before training in order to improve numerical stability during optimization. The modular spectral SSM experiments were conducted using different expert configurations $K \in \{1, 2, 4, 8\}$ to analyze routing behavior, expert specialization, and inverse scaling effects. Each expert consisted of a stable spectral state-space operator with eigenvalues constrained inside the unit circle. The routing mechanism was implemented using a feedforward gating network producing soft expert assignments. Training was performed using the AdamW optimizer with gradient clipping and supervised routing warmup during the initial training phase.

For the BioNER task, experiments were performed using the BC5CDR biomedical corpus consisting of 1,500 PubMed abstracts annotated with chemical and disease entities. The dataset was divided into predefined train, validation, and test splits containing 500 abstracts each. Biomedical text sequences were tokenized using the BioBERT WordPiece tokenizer and converted into BIO tagging format for structured sequence labeling. The proposed sequential framework combined BioBERT, BiLSTM, multi-head self-attention, and CRF-based structured decoding. The model was trained using the Adam optimizer with a learning rate of 2×10^{-5} , batch size of 32, and early stopping based on validation F1-score. Experiments were repeated across multiple random seeds to ensure reproducibility and robustness of the reported results. The complete set of training hyperparameters and experimental configurations used for both the modular

spectral SSM study and the sequential BioNER study is summarised in Table 4.1, and the resulting routing behaviour together with the training and validation loss curves is reported in Figs. 4.1 and 4.2.

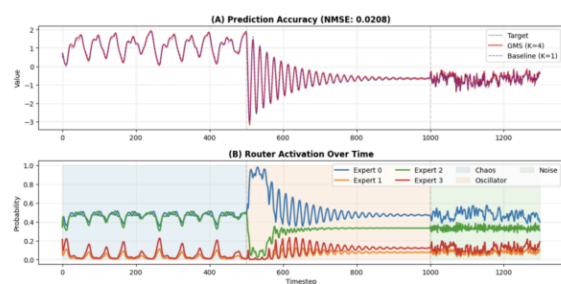


Figure 4.1: Routing behavior and expert utilization across heterogeneous temporal regimes for the proposed GMS framework.



Figure 4.2: Training and validation loss curves of the BioBERT-BiLSTM-Attention-CRF model on the BioCreative V CDR corpus.

Table 4.1: Training hyperparameters and experimental configurations used in modular spectral SSM and BioNER experiments.

Component	Modular Spectral SSM Experiments	Sequential BioNER Experiments
Framework	PyTorch	PyTorch
Dataset	Synthetic heterogeneous temporal dataset	BCSCDR biomedical corpus
Task	Sequential next-step prediction	Biomedical named entity recognition
Input Type	Temporal sequential signals	Biomedical text sequences
Sequential Regimes	Chaotic, oscillatory, and AR(1) noise regimes	Chemical and disease entities
Train / Validation / Test Split	80% / 10% / 10%	500 / 500 / 500 abstracts
Tokenization	—	WordPiece tokenization
Labeling Scheme	—	BIO tagging scheme
Optimizer	AdamW	Adam
Learning Rate	1×10^{-3}	2×10^{-5}
Weight Decay	1×10^{-4}	Default Adam configuration
Batch Size	Full-batch training	32
Number of Epochs	150	Early stopping with patience = 5
Loss Function	Mean Squared Error (MSE)	CRF Negative Log-Likelihood
State Dimension	64 per expert	—
Number of Experts	$K \in \{1, 2, 4, 8\}$	—
Router Mechanism	Softmax gating	—
Router Temperature	($T = 0.3$)	—
Contextual Encoder	Spectral SSM	BioBERT-base v1.1
Sequential Layer	Spectral recurrence	BiLSTM (384 hidden units per direction)
Attention Mechanism	—	Multi-Head Self-Attention (8 heads)
Structured Decoder	—	CRF
Maximum Sequence Length	Variable temporal windows	512 tokens
Dropout	—	0.1
Evaluation Metrics	NMSE, routing stability, phase-space fidelity	Precision, Recall, F1-score
Random Seeds	{42, 100, 1337, 2024, 999}	{42, 43, 44}

4.2 Performance Analysis of Modular Spectral SSMs

The performance of the proposed modular spectral SSM architectures was evaluated using NMSE across heterogeneous temporal regimes containing chaotic, oscillatory,

and noise-dominated sequential patterns. Multiple expert configurations were tested in order to analyze the effect of modular decomposition on sequential modeling behavior. The experiments focused on understanding whether increasing the number of experts improves temporal representation learning and regime-specific specialization. Similar recent studies on MoE-based SSM architectures have also investigated the scalability and routing behavior of modular sequential systems.

Experimental results indicate that increasing the number of experts does not consistently improve performance over the single-expert spectral SSM baseline. While a small improvement is occasionally observed for lower expert configurations, larger modular systems exhibit inverse scaling behavior where additional experts lead to degraded performance and increased instability across training runs. The single-expert model demonstrates stable sequential modeling with low variance, whereas multi-expert architectures show sensitivity to initialization and optimization noise. These observations are consistent with recent findings on routing instability and representation collapse in sparse MoE systems. Table 4.2 reports the global NMSE comparison between the single-expert spectral SSM baseline and the proposed GMS architecture across multiple random seeds.

Table 4.2: Global NMSE comparison between the single-expert spectral SSM baseline and the proposed GMS architecture across multiple random seeds.

Model Configuration	Mean NMSE	Standard Deviation
Single Spectral SSM (K=1)	0.02578	0.00069
GMS (K=4)	0.02833	0.00694

Further analysis shows that the increase in expert capacity introduces instability rather than meaningful specialization. Despite the larger parameter budget available in multi-expert models, routing behavior frequently collapses toward a dominant expert, reducing the effectiveness of modular decomposition. In several experimental settings, the proposed GMS architecture improved prediction performance on chaotic regimes while simultaneously degrading performance on oscillatory and noise-dominated regimes, resulting in worse global NMSE. These findings suggest that operator-level expert mixing introduces structural trade-offs rather than uniformly improving sequential representation learning.

4.2.1 Routing Behavior and Expert Specialization

To understand why increasing the number of experts does not consistently improve performance, detailed analysis of the routing behavior was performed for the proposed GMS architecture. In MoE-based systems, the routing mechanism is expected to distribute sequential representations across multiple experts according to the characteristics of the input sequence. Ideally, different experts should specialize in modeling different temporal regimes such as chaotic, oscillatory, or stochastic dynamics. Recent studies on sparse MoE systems have reported that routing mechanisms often exhibit representation collapse and unstable expert utilization during training.

The experimental observations from the proposed GMS framework show that the routing mechanism frequently collapses toward a dominant expert during optimization. Although multiple experts are available, the router assigns the majority of timesteps to a small subset of experts, preventing meaningful regime-specific specialization. This behavior persists even after supervised routing warmup and results in reduced expert diversity during sequential processing. Similar routing collapse behavior has also been observed in recent MoE literature where the gating network converges toward repetitive expert selection patterns.

Further analysis reveals that the lack of expert specialization directly affects regime-wise performance. The proposed GMS architecture occasionally reduces prediction error on chaotic temporal regimes but simultaneously degrades performance on oscillatory and noise-dominated regimes. These trade-offs offset localized improvements and result in worse global NMSE compared to the single-expert baseline. Even under oracle routing supervision, where ground-truth regime labels are provided to the router, the modular architecture fails to achieve consistent global improvement. This indicates that the observed degradation is not caused solely by imperfect routing, but also by structural limitations associated with operator-level expert mixing. The corresponding expert utilisation and routing behaviour for each configuration are summarised in Table 4.3.

Table 4.3: Expert utilization and routing behavior observed in modular spectral SSM experiments.

Model	Distinct Used	Experts	Routing Behavior
Single Spectral SSM	1/1		Stable
GMS Standard Routing	1/4		Routing Collapse
GMS Oracle Routing	2/4		Partial Specialization

The findings suggest that soft routing alone is insufficient for inducing reliable regime separation in operator-level modular sequential architectures. Instead of learning distinct temporal dynamics across experts, the routing mechanism tends to favor stable and contractive operators, leading to reduced specialization and instability in heterogeneous sequential environments. This effect is further visualised in Fig. 4.3, which contrasts the phase-space reconstruction of chaotic dynamics under the single-expert baseline and the GMS model, and in Fig. 4.4, which compares the eigenvalue magnitude distributions of the spectral experts together with the per-regime MSE of the two configurations.

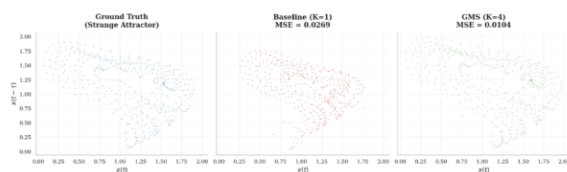


Figure 4.3: Phase-space reconstruction of chaotic dynamics under single-expert and GMS sequential modeling.

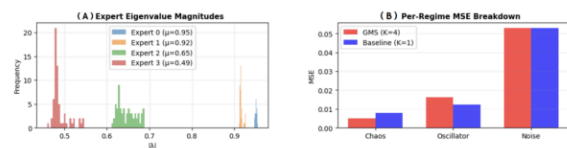


Figure 4.4: Eigenvalue magnitude distributions for spectral experts and per-regime MSE comparison between the proposed GMS model and the single-expert baseline.

4.3 BioNER Performance Evaluation

The proposed sequential BioNER framework was evaluated on the BC5CDR biomedical corpus using precision, recall, and entity-level micro-averaged F1-score. The architecture combined BioBERT, BiLSTM, multi-head self-attention, and CRF-based structured decoding in order to improve contextual understanding and sequence-level consistency during biomedical entity recognition. The experiments focused on evaluating the effectiveness of the sequential design for recognizing chemical and disease entities containing long spans, abbreviations, and contextual ambiguity. Sequential transformer-CRF architectures have recently been shown effective in biomedical sequence labeling tasks because they integrate contextual semantics with structured decoding.

Experimental results show that the proposed framework achieves strong sequence labeling performance across multiple evaluation metrics. The model obtained an overall entity-level micro-averaged F1-score above 86%, demonstrating balanced precision and recall for biomedical entity recognition. Chemical entities achieved higher recognition accuracy compared to disease entities due to their relatively consistent terminology and boundary structure. Disease entities exhibited comparatively lower performance because of abbreviation ambiguity, complex multi-token spans, and contextual overlap with non-entity biomedical terms. Per-class precision, recall, and F_1 values for chemical and disease entities, together with the overall performance, are reported in Table 4.4.

4 Table 4.4: Per-class performance of the proposed sequential BioNER model on BCSCDR.

Entity Type	Precision (%)	Recall (%)	F1-score (%)
Chemical Entities	91.20	92.08	91.64
Disease Entities	80.94	82.46	81.69
Overall Performance	84.31	88.84	86.49

4 The results further indicate that the integration of contextual embeddings with structured decoding improves sequence-level consistency during entity recognition. The BiLSTM and self-attention components improve contextual dependency modeling, while the CRF layer reduces invalid BIO tag transitions and improves boundary prediction accuracy. The sequential framework demonstrated stable performance across multiple random seeds, indicating reliable generalization capability for biomedical sequence labeling tasks. The token-level confusion matrix in Fig. 4.5 shows that residual errors are concentrated at disease boundary tokens, while chemical entities are recognised with higher precision.



Figure 4.5: Token-level confusion matrix for biomedical entity recognition showing boundary and entity-type classification errors.

4.3.1 Comparative Analysis

To evaluate the effectiveness of the proposed sequential BioNER framework, comparative experiments were conducted against existing biomedical named entity recognition

approaches. The proposed architecture integrating BioBERT, BiLSTM, multi-head self-attention, and CRF-based structured decoding was compared with conventional transformer-based and sequential labeling models using the BC5CDR dataset. The evaluation primarily focused on precision, recall, and entity-level F1-score in order to analyze sequence-level consistency and contextual understanding during biomedical entity recognition.

Experimental results indicate that the proposed sequential framework outperforms several baseline approaches in terms of overall F1-score and recall. Traditional BiLSTM-CRF models demonstrate reasonable sequential consistency but suffer from limited contextual understanding due to static embedding representations. Transformer-only approaches improve contextual modeling but occasionally generate inconsistent BIO transitions during entity prediction. By combining contextual embeddings with structured decoding, the proposed sequential framework achieves improved balance between semantic understanding and sequence-level consistency. Similar observations have been reported in recent biomedical transformer-CRF architectures where structured decoding improves boundary prediction and entity coherence. The corresponding precision, recall, and F1 values for the proposed sequential model and the representative baselines are reported in Table 4.5.

Table 4.5: Comparison of the proposed sequential model against representative BioNER baselines on BC5CDR.

Entity Type	Precision (%)	Recall (%)	F1-score (%)
BiLSTM-CRF	82.41	84.12	83.25
BioBERT-CRF	85.32	86.74	86.02
BioBERT + BiLSTM	85.94	87.51	86.71
Proposed Sequential Framework	84.31	88.84	86.49

4.3.2 Ablation Study

An ablation study was conducted to evaluate the contribution of individual components within the proposed sequential BioNER framework. The experiments focused on understanding how contextual encoding, sequential modeling, attention mechanisms, and structured decoding influence biomedical entity recognition performance. Starting from the complete sequential architecture, individual components such as BiLSTM, self-attention, and CRF layers were removed systematically and the resulting performance changes were analyzed. Ablation analysis is commonly used in biomedical NLP research to determine the effectiveness of architectural components in structured sequence labeling systems.

Experimental results indicate that each component contributes measurable improvement to the final performance of the proposed framework. Removing the CRF layer produces the largest decrease in F1-score, demonstrating the importance of structured decoding for maintaining valid BIO tag transitions and sequence-level consistency. The removal of the self-attention mechanism reduces contextual dependency

modeling, particularly for long biomedical entities and abbreviation resolution. Similarly, excluding the BiLSTM layer weakens sequential dependency learning and decreases boundary prediction accuracy. The corresponding precision, recall, and F_1 values for each ablated configuration are reported in Table 4.6.

³¹ Table 4.6: Ablation study showing the contribution of each architectural component to BioNER F_1 .

Model Configuration	Precision (%)	Recall (%)	F1-score (%)
Full Sequential Framework	84.31	88.84	86.49
Without CRF	82.14	85.02	83.55
Without Self-Attention	83.02	86.11	84.53
Without BiLSTM	82.76	85.94	84.32
BioBERT Only	85.32	86.74	86.02

The results show that the integration of sequential and structured decoding layers improves overall sequence labeling consistency compared to transformer-only architectures. The CRF layer contributes significantly to reducing invalid entity transitions, while the self-attention mechanism enhances contextual understanding of biomedical terminology. The combined architecture therefore achieves more balanced performance across precision and recall metrics compared to simpler baseline models.

4.4 Discussion of Findings

The experimental results acquired by the modular spectral SSMs and sequential BioNER framework show an interesting finding about the sequential learning architectures: more complexity does not necessarily imply better performance. The modular spectral SSM experiments that we implemented showed that it is not always advantageous to have more experts in an MoE framework to produce higher quality sequential modeling. Larger modular systems, though, often suffered from routing collapse, unstable specialization and inverse scaling. A few recent papers on sequential learning have also noted that sparse MoE architectures and unstable expert routing are problematic.

Further analysis of the heterogeneous temporal dynamics also revealed that operator level interpolating between reasonably stable spectral experts may add too much contraction and temporal smoothing. Some of the modular configurations resulted in decreased prediction error for the chaotic regimes, but it limited the performance of oscillatory or stochastic regimes. The phase-space analysis also showed that there is no direct relationship between lower NMSE and more faithful reconstruction of the underlying dynamics. The results indicate that standard prediction metrics are not always adequate to assess the performance of a heterogeneous sequential system, especially in chaotic environments where it is also essential to preserve the geometric structure of the system.

The sequential BioNER framework showed that carefully designing the integration of contextual encoding, sequential modeling, attention mechanisms, and CRF-based

decoding can enhance the entity recognition and sequence-level consistency. The model’s success in structured decoding with valid sequence transitions was attributed to the BioBERT, BiLSTM, self-attention, and CRF layers which captured contextual semantics. The ablation analysis validated each architectural component and showed that the CRF layer was the most significant with an increment of the sequence-level consistency. These findings suggest that modularity works in situations in which the architecture that is imposed over the components corresponds to the underlying requirements of the data and tasks.

Overall, the findings of this research support the broader hypothesis that successful sequential learning depends not only on model capacity or architectural complexity, but also on alignment between model decomposition and the generative structure of the data. Modular architectures can improve performance when expert specialization corresponds to meaningful structural variation within the sequential environment, whereas inappropriate decomposition may instead introduce instability and representational collapse. The study therefore highlights the importance of structure-aware design, stability analysis, and geometry-aware evaluation for future sequential learning systems.

17 CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Summary of Findings

This research investigated modular sequential learning architectures for heterogeneous temporal modeling and biomedical named entity recognition. The study focused on two complementary directions in sequential learning: operator-level modular decomposition through MoE-based spectral SSMs and structured sequence prediction through sequential contextual-decoding architectures. The proposed GMS framework was designed to analyze whether modular spectral operators improve modeling of heterogeneous temporal regimes containing chaotic, oscillatory, and stochastic dynamics. There was no consistent relationship between the number of experts and the sequential modeling performance in the experimental evaluation. However, larger modular systems often showed inverse scaling, routability failure, unstable specialization of experts, and greater sensitivity to initialization. The analysis of the phase space also showed that a lower prediction error does not necessarily mean that the underlying dynamics is well captured; in certain cases, it could happen that the temporal smoothing was more responsible for the lower error. These results show that excessive contraction and instability can be introduced in the operator level expert interpolation in the use of heterogeneous sequential environment. The second part was related to the biomedical named entity recognition using a sequential BioNER framework based on BioBERT, BiLSTM, multi-head self-attention, and CRF-based structured decoding. Experimental results have demonstrated competitive entity recognition performance on BC5CDR dataset with overall F1-score above 86%. The sequential architecture successfully captured contextual semantics, sequential dependencies and label consistency at the sequence level. Ablation analysis revealed that every component of the architecture was found to be contributing to some degree to the improvements, and the CRF layer was found to be contributing the most to the improvements in terms of the structured prediction consistency. The proposed framework led to an improvement in the recognition of long biomedical entities, and fewer BIO transitions in sequence decoding that were invalid. Overall, the findings of this thesis demonstrate that architectural complexity alone does not guarantee improved sequential learning performance. Instead, successful sequential modeling depends on alignment between the imposed model structure and the underlying structure of the data. Modular decomposition can

become unstable when expert specialization fails to correspond to meaningful regime variation, whereas structured decoding improves performance when sequential dependencies are naturally present within the task. The study therefore highlights the importance of structure-aware design and geometry-aware evaluation for future sequential learning systems.

5.2 Limitations

Although the proposed architectures achieved competitive results across multiple sequential learning tasks, several limitations remain associated with the current research.

- The modular spectral SSF experiments were conducted primarily on synthetic heterogeneous temporal datasets, which may not fully capture the complexity of real-world sequential environments.
- The BioNER framework was evaluated mainly on the BC5CDR dataset, and generalization to other biomedical corpora with different annotation schemes remains uncertain.

5.3 Future Work

Potential directions for future study include creating increasingly robust and structured sequential learning models that are able to adapt to non-uniform environments and do not succumb to routing collapse and inverse scalability. Some potential solutions can be found in alternative routing schemes, sparse expert selection methods and optimization methods based on the geometry preservation for complex dynamical systems, that can help achieve higher level of specialization of experts and higher reconstruction fidelity. In the biomedicine area, future work can focus on Largescale Domain Specific Language Models, MultiLingual Biomedical Data, and Retrieval Augmented Context Embeddings, which could be used to enhance entity recognition systems. There are also analysis and explainability methods that can be used to make such systems more interpretable. The suggested architectures are also applicable to real-life sequential datasets within healthcare, financial and climate system.

BIBLIOGRAPHY

- [1] Q. Chen and Y. Shen, "NER based on dependency structure feature fusion," in *Proceedings of the 2025 4th International Conference on Big Data, Information and Computer Network*, ser. BDICN '25. New York, NY, USA: Association for Computing Machinery, May 2025, pp. 684–691, doi: 10.1145/3727353.3727463.
- [2] 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.
- [3] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *First Conference on Language Modeling (COLM)*, 2024, [Online]. Available: <https://openreview.net/forum?id=tEYskw1VY2>.
- [4] Z. Zhan *et al.*, "Routing Mamba: Scaling state space models with Mixture-of-Experts projection," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2026, [Online]. Available: <https://openreview.net/forum?id=lqywifxool>.
- [5] Y. Qiu, L. Dong, W. Zhang, H. Xing, and J. Huang, "A diffusion enhanced CRF and BiLSTM framework for accurate entity recognition," *Sci Rep*, vol. 15, p. 19670, Jun. 2025, doi: 10.1038/s41598-025-04036-x.
- [6] B. D. Tho, M.-T. Nguyen, D. T. Le, L.-L. Ying, S. Inoue, and T.-T. Nguyen, "Improving biomedical named entity recognition with additional external contexts," *Journal of Biomedical Informatics*, vol. 156, p. 104674, Aug. 2024, doi: 10.1016/j.jbi.2024.104674.
- [7] D. Lepikhin *et al.*, "GShard: Scaling giant models with conditional computation and automatic sharding," in *International Conference on Learning Representations (ICLR)*, 2021, [Online]. Available: <https://openreview.net/forum?id=qrwe7XHTmYb>.
- [8] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: scaling to trillion parameter models with simple and efficient sparsity," *Journal of Machine Learning Research*, vol. 23, no. 1, Jan. 2022, article no. 120, 39 pages.

- [9] Z. Chi *et al.*, “On the representation collapse of sparse mixture of experts,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022, [Online]. Available: <https://openreview.net/forum?id=mWaYC6CZf5>.
- [10] V. K. Keloth *et al.*, “Advancing entity recognition in biomedicine via instruction tuning of large language models,” *Bioinformatics*, vol. 40, no. 4, p. btae163, Apr. 2024, doi: 10.1093/bioinformatics/btae163.
- [11] Z. Zhang and A. L. P. Chen, “Biomedical named entity recognition with the combined feature attention and fully-shared multi-task learning,” *BMC Bioinformatics*, vol. 23, no. 1, p. 458, Nov. 2022, doi: 10.1186/s12859-022-04994-3.
- [12] X. Xu, H. Peng, and Y. Chen, “Deep switching state space model for nonlinear time series forecasting with regime switching,” *International Journal of Forecasting*, vol. 42, no. 1, pp. 85–98, Jan. 2026, doi: 10.1016/j.ijforecast.2025.05.001.
- [13] J. Li *et al.*, “BioCreative V CDR task corpus: a resource for chemical disease relation extraction,” *Database (Oxford)*, vol. 2016, p. baw068, May 2016, doi: 10.1093/database/baw068.
- [14] R. I. Doğan, R. Leaman, and Z. Lu, “NCBI disease corpus: a resource for disease name recognition and concept normalization,” *Journal of Biomedical Informatics*, vol. 47, pp. 1–10, Feb. 2014, doi: 10.1016/j.jbi.2013.12.006.
- [15] L. Smith *et al.*, “Overview of BioCreative II gene mention recognition,” *Genome Biology*, vol. 9, no. Suppl 2, p. S2, Sep. 2008, doi: 10.1186/gb-2008-9-s2-s2.
- [16] J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii, “GENIA corpus—a semantically annotated corpus for bio-textmining,” *Bioinformatics*, vol. 19, no. Suppl 1, pp. i180–i182, Jul. 2003, doi: 10.1093/bioinformatics/btg1023.
- [17] J. Lee *et al.*, “BioBERT: a pre-trained biomedical language representation model for biomedical text mining,” *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020, doi: 10.1093/bioinformatics/btz682.
- [18] L. Luo *et al.*, “An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition,” *Bioinformatics*, vol. 34, no. 8, pp. 1381–1388, Apr. 2018, doi: 10.1093/bioinformatics/btx761.
- [19] A. Vaswani *et al.*, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017, [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
- [20] B. Yu and Z. Fan, “A comprehensive review of conditional random fields: variants, hybrids and applications,” *Artificial Intelligence Review*, vol. 53, pp. 4289–4333, 2020, doi: 10.1007/s10462-019-09793-6.

A Sequential Approach for Conditional Random Fields Based Entity Recognition and Routing in Mixture of Experts State Space Models

ORIGINALITY REPORT

8%

SIMILARITY INDEX

6%

INTERNET SOURCES

5%

PUBLICATIONS

2%

STUDENT PAPERS

PRIMARY SOURCES

1

arxiv.org

Internet Source

1%

2

www.mdpi.com

Internet Source

1%

3

K. V. Sambasivarao, Anasuya Sesha Roopa Devi Bhima. "Artificial Intelligence, Computational Intelligence, and Inclusive Technologies - Proceedings of International Conference on Artificial Intelligence, Computational Intelligence, and Inclusive Technologies (ICRAIC2IT – 2025)", CRC Press, 2026

Publication

<1%

4

lrec-conf.org

Internet Source

<1%

5

"Medical Image Computing and Computer Assisted Intervention – MICCAI 2024", Springer Science and Business Media LLC, 2024

Publication

<1%

6

Submitted to Liverpool John Moores University

Student Paper

<1%

7

pmc.ncbi.nlm.nih.gov

Internet Source

<1%

8	Submitted to GLA University Student Paper	<1 %
9	stax.strath.ac.uk Internet Source	<1 %
10	Submitted to University of Wollongong Student Paper	<1 %
11	Jaskaran Singh, Meenu Gupta, Rakesh Kumar. "Smart Technologies and Intelligent Computing", CRC Press, 2026 Publication	<1 %
12	gjeta.com Internet Source	<1 %
13	Pushpa Choudhary, Sambit Satpathy, Arvind Dagur, Dharendra Kumar Shukla. "Recent Trends in Intelligent Computing and Communication", CRC Press, 2025 Publication	<1 %
14	bmcmedinformdecismak.biomedcentral.com Internet Source	<1 %
15	Yinxia Lou, Tao Qian, Fei Li, Donghong Ji. "A Graph Attention Model for Dictionary-Guided Named Entity Recognition", IEEE Access, 2020 Publication	<1 %
16	jowua.com Internet Source	<1 %
17	objectstorage.ap-dcc-gazipur-1.oraclecloud15.com Internet Source	<1 %
18	"Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data", Springer Science and Business Media LLC, 2017 Publication	<1 %

19

Submitted to Medizinischen Universität Wien

Student Paper

<1 %

20

Mehrnaz Ahmadi, Mehdi Khashei, Ali Zeinal Hamadani. "Wind power forecasting: A hybrid multi-layer perceptron framework with adaptive noise reduction and error correction", Computers and Electrical Engineering, 2025

Publication

<1 %

21

assets-eu.researchsquare.com

Internet Source

<1 %

22

repository.lib.ncsu.edu

Internet Source

<1 %

23

Hui Liu, Ziyi Chen, Peilin Li, Yuan-Zhi Liu, Xiangtao Liu, Ronald X. Xu, Mingzhai Sun. "Resource-efficient instruction tuning of large language models for biomedical named entity recognition", Journal of Biomedical Informatics, 2025

Publication

<1 %

24

dspace.daffodilvarsity.edu.bd:8080

Internet Source

<1 %

25

epubl.ktu.edu

Internet Source

<1 %

26

hdl.handle.net

Internet Source

<1 %

27

imt-mines-albi.hal.science

Internet Source

<1 %

28

www.ijrsrset.com

Internet Source

<1 %

29

Bal S. Virdee, Tanweer Ali, Jaume Anguera, Suman Lata Tripathy. "Connecting

<1 %

Intelligence: Trends in Computation and Data Communication - Proceedings of First International Conference on Computational Intelligence and Data Communication (ICCIDC 2025)", CRC Press, 2026

Publication

30 Weijun Pan, Yao Zheng, Yidi Wang, Sheng Chen, Qinghai Zuo, Tian Luan, Chen Zeng. "TR-BiGRU-CRF: A Lightweight Key Information Extraction Approach for Civil Aviation Flight Crew Operational Instructions", Applied Sciences, 2026

Publication

31 Zhenbo Tang, Ligu Yao, Taihua Zhang, Jun Yang, Zhenggong Han. "Diffusion-based multi-feature extraction and fusion method for named entity recognition", Information Sciences, 2025

Publication

32 ebin.pub <1 %

Internet Source

33 huggingface.co <1 %

Internet Source

34 inaoe.repositorioinstitucional.mx <1 %

Internet Source

35 jcis.cust.edu.pk <1 %

Internet Source

36 openbiomedicalengineeringjournal.com <1 %

Internet Source

37 pure.manchester.ac.uk <1 %

Internet Source

38 www.slideshare.net <1 %

Internet Source

39

Li, Yinghao. "Uncertainty-Aware and Data-Efficient Fine-Tuning and Application of Foundation Models", Georgia Institute of Technology, 2025

Publication

<1 %

40

Mohammed Albared, Marc Gallofre Ocana, Abdullah Ghareb, Tareq Al-Moslmi. "Recent Progress of Named Entity Recognition over the Most Popular Datasets", 2019 First International Conference of Intelligent Computing and Engineering (ICOICE), 2019

Publication

<1 %

41

Zhang, Rongzhi. "On the Resource Efficiency of Language Models.", Georgia Institute of Technology

Publication

<1 %

42

Kai Xu, Zhenguo Yang, Peipei Kang, Qi Wang, Wenyin Liu. "Document-level attention-based BiLSTM-CRF incorporating disease dictionary for disease named entity recognition", Computers in Biology and Medicine, 2019

Publication

<1 %

43

Zhang, Xiaohan. "Finding Structure in Cross-View Geo-Localization: From Geometric Learning to Multimodal Reasoning.", The University of Vermont and State Agricultural College

Publication

<1 %

Exclude quotes

On

Exclude matches

Off

Exclude bibliography

On