

Reproducing Time-Aware Transformer-based Network for Clinical Notes Series Prediction

Benjamin Miller
Siebel School of Computing and Data Science
University of Illinois Urbana-Champaign
bm69@illinois.edu

March 30, 2025

1 Introduction

1.1 Problem Statement

Clinical notes in Electronic Health Records (EHRs) contain rich patient information and expert insights critical for decision-making [1]. However, effectively utilizing clinical notes for predictive tasks presents several significant challenges:

First, clinical notes naturally follow a *multi-level sequential structure* - they form a sequence of documents over time, with each document containing sequences of words [2]. Second, clinical notes have *complex interrelations* between notes and their constituent parts (chunks) [3]. Third, the *timing and order* of clinical notes can be crucial indicators of patient progression, with irregularly spaced notes potentially containing vital temporal information that most current models ignore.

The paper proposes to address these challenges by developing a novel hierarchical architecture that can effectively capture both the multi-level sequential structure and temporal information in clinical notes for improved patient-level predictions such as mortality, readmission, and infection risks [4].

1.2 Citation to Original Paper

Zhang, D., Thadajarassiri, J., Sen, C., & Rundensteiner, E. (2020). Time-Aware Transformer-based Network for Clinical Notes Series Prediction [4]. In *Proceedings of Machine Learning Research*, 126:1-22. Machine Learning for Healthcare.

2 Methodology

2.1 Specific Approach

The authors proposed a model called FTL-Trans (Flexible Time-aware LSTM Transformer) [4], a hierarchical architecture with four main components:

1. **Chunk Content Embedding:** Transformer-encoder using ClinicalBERT weights encodes fixed-length note chunks [3].
2. **Position-Enhanced Chunk Embedding Layer:** Merges content embeddings with sequential information using novel global position embeddings (for note position) and local position embeddings (for chunk position within a

note), following principles established in transformer architectures [5].

3. **Time-Aware Layer:** Implements a novel Flexible Time-aware LSTM (FT-LSTM) cell that incorporates temporal information by learning a flexible time decay function to capture patterns of temporal importance, extending prior work on time-aware LSTMs [6].
4. **Classification Layer:** Generates patient-level predictions using the learned representations.

Evaluation Metrics: The authors evaluated their model using standard classification metrics: Area Under the Receiver Operating Characteristic curve (AUROC), Accuracy, and Area Under Precision-Recall curve (AUPR). They compared against baselines including BERT [7], ClinicalBERT [3], and variations of the proposed architecture.

Clinical Tasks: The model was evaluated on five prediction tasks using MIMIC-III data [1]: in-hospital mortality, 30-day readmission, and three infection predictions (Escherichia Coli, Enterococcus Sp., and Klebsiella pneumoniae).

2.2 Novelty and Relevance

The key innovations of this work are:

1. **Multi-level Sequential Structure:** Unlike flat models like ClinicalBERT [3] that lose the hierarchical structure of clinical notes, FTL-Trans preserves the relationships between notes and chunks through a hierarchical design and multi-level position embeddings [4].
2. **Flexible Time-Awareness:** The proposed FT-LSTM contains a trainable time decay function that combines convex, linear, and concave sub-functions, allowing it to learn task-specific patterns of temporal importance rather than using fixed decay functions [6].
3. **Jointly Modeling Content and Context:** The model simultaneously leverages linguistic content and contextual information (position and timing), capturing the full complexity of clinical note series [4].

These innovations address significant limitations in existing models. Transformer-based models like ClinicalBERT [3] typically disregard interrelations among notes and chunks, while standard time-aware models [6] make restrictive assumptions about temporal influence decay. The authors hypothesized

that a model capturing both multi-level structure and flexible temporal importance would outperform existing approaches, which their empirical results supported (improvements of up to 5% in AUROC and 6% in accuracy).

2.3 Planned Ablations and Extensions

Given the complexity of the model and time constraints, I will focus on two key extensions that offer the best balance of feasibility and research value:

1. **Enhanced Time Decay Function:** The original FT-LSTM uses a flexible time decay function with trainable parameters to model temporal importance [4]. I will experiment with alternative formulations of this decay function, including:

- Different initialization strategies based on clinical domain knowledge
- Alternative mathematical forms (e.g., exponential, logarithmic, and sigmoid variants)
- Task-specific parameterizations for different prediction tasks

This extension directly targets the core temporal modeling capability without requiring architectural changes.

2. **Additional Clinical Tasks:** I will extend the evaluation to other clinically relevant prediction tasks using the same MIMIC-III dataset [1], such as:

- Length-of-stay prediction
- Decompenation prediction (predicting acute deterioration)
- Medication recommendation

This will help assess the model’s generalizability across different clinical prediction scenarios while leveraging the same data preprocessing pipeline.

For future work beyond this project’s scope, more ambitious extensions could include attention-based temporal modeling [5], specialized pre-training strategies, and note type importance modeling. The current focus will allow for thorough validation of the original hypothesis (that capturing both multi-level structure and temporal information improves clinical prediction), particularly regarding the flexible time decay function’s contribution to performance.

3 Data Access and Implementation Details

3.1 Data Access

The paper uses MIMIC-III [1], a critical care database with de-identified data from 40,000+ patients. Access requires:

1. Completion of CITI training courses
2. Signing of a data use agreement

3. Registration on the PhysioNet website and credentialing through the class event

I have completed the pre-requisite Conflicts of Interest course and Data or Specimens Only Research Course, and have joined the class’s PhysioNet event for data access credentialing. I am currently awaiting final access approval. The authors extract clinical notes from the NoteEvents table and create five cohorts for different prediction tasks (mortality, readmission, and three infection types). I will follow their detailed cohort construction process as described in Section 3 of the paper.

The source code for the FTL-Trans model is publicly available at <https://github.com/zdy93/FTL-Trans>, which will facilitate exact reproduction of the original results.

3.2 Computational Feasibility

The computational requirements for reproducing this work are substantial but manageable within the Google Colab environment:

1. **Hardware & Memory:** Google Colab Pro provides GPUs comparable to the authors’ V100. I’ll use their approach of 64 chunks per patient to manage memory constraints.
2. **Training Time:** Training requires 1.56 hours/epoch for 3 epochs, feasible within Colab’s time limits.
3. **Pre-trained Models:** The model utilizes pre-trained ClinicalBERT [3] weights, which are publicly available and will reduce the computational burden of training from scratch.

For ablation studies, I may need to optimize hyperparameters or model components to fit within computational constraints, potentially using model parallel training across multiple Colab sessions if necessary.

3.3 Implementation Approach

I plan to use the authors’ existing codebase (<https://github.com/zdy93/FTL-Trans>) as the foundation for reproduction [4]. This approach offers several advantages:

1. Ensures fidelity to the original implementation details
2. Speeds up reproduction, allowing more time for validation and extension
3. Provides a verified baseline against which to compare ablations

I will inspect the code for transparency and develop extension modules that maintain compatibility with the original preprocessing and evaluation pipelines.

For the implementation of the transformer-based components, I will leverage the Hugging Face Transformers library [8], which provides efficient and well-tested implementations of transformer architectures.

References

- [1] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.
- [2] Cansu Sen, Thomas Hartvigsen, Xiangnan Kong, and Elke Rundensteiner. Patient-level classification on clinical note sequences guided by attributed hierarchical attention. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 930–939. IEEE, 2019.
- [3] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clin-icalbert: Modeling clinical notes and predicting hospital readmission. In *arXiv preprint arXiv:1904.05342*, 2019.
- [4] Dongyu Zhang, Jidapa Thadajarassiri, Cansu Sen, and Elke Rundensteiner. Time-aware transformer-based network for clinical notes series prediction. In *Proceedings of Machine Learning Research*, volume 126, pages 1–22. Machine Learning for Healthcare, 2020.
- [5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
- [6] Inci M Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K Jain, and Jiayu Zhou. Patient subtyping via time-aware lstm networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 65–74, 2017.
- [7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 4171–4186, 2019.
- [8] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface’s transformers: State-of-the-art natural language processing. In *arXiv preprint arXiv:1910.03771*, 2019.