xLSTM for Irregular Multivariate Clinical Time-Series Forecasting

Laura Legat
Institute for Machine Learning
Johannes Kepler University Linz
laura.legat@jku.at

Erich Kobler
Institute for Machine Learning
LIT AI Lab
Department of Virtual Morphology
Clinical Research Institute Medical AI
Johannes Kepler University Linz
erich.kobler@jku.at

Abstract

Intensive care units (ICUs) provide lifesaving treatments to patients with severe medical conditions, producing large amounts of clinical time-series data that reflect patient health trajectories. Forecasting future trajectory changes helps clinicians anticipate adverse events. While prior work addresses the challenges of missing values and irregularities in clinical time-series, designing effective forecasting architectures for such data remains an open research area. At the same time, limitations of Transformer-based models are prompting a renewed interest in recurrent architectures for processing time-series. Among them, the recently proposed xLSTM demonstrates strong forecasting capabilities across several domains, yet its potential for clinical use-cases remains largely unexplored. In this work, we address this gap by extending xLSTM to forecast irregular multivariate clinical time-series with missing values. To this end, we replace the temporal and cross-channel modeling components of an established forecasting architecture with xLSTM blocks. Our models achieve competitive predictive performance compared to several baselines on a subset of MIMIC-III, highlighting xLSTM's potential as a powerful backbone for clinical time-series forecasting.

1 Introduction

Continuous monitoring in the ICU is employed to improve clinical outcomes and provide insight into individual health journeys [20], generating large volumes of patient data organized as electronic health records (EHR) [6]. A substantial portion of these records constitute clinical time-series in the form of sequential, time-indexed, multivariate observations [13]. We refer to them as multivariate clinical time-series (MCTS), which capture a patient's physiological trajectory over time. Anticipating future trajectory changes early-on is crucial, as it helps to avoid adverse events and prolonged hospital stays [19]. This can be framed as a time-series forecasting task [14] and remains a challenging endeavor, since MCTS often contain a multitude of channels [22], all sampled at different frequencies, with different levels of missingness, outliers, and artifacts [22, 29].

Transformer-based architectures [21] remain a popular choice for this task [29, 17], yet they face limitations, such as large training data requirements as well as the quadratic complexity of the attention mechanism with respect to sequence length and number of channels [12, 1]. This results in a renewed interest in linear-complexity recurrent models for time-series forecasting [12]. Beck et al. [2] introduce such an architecture with the Extended Long Short-Term Memory (xLSTM), a successor to the LSTM, which shows strong forecasting results across several non-medical domains, such as weather, solar, or electricity [12, 1]. However, the potential of xLSTM for MCTS forecasting remains unexplored.

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

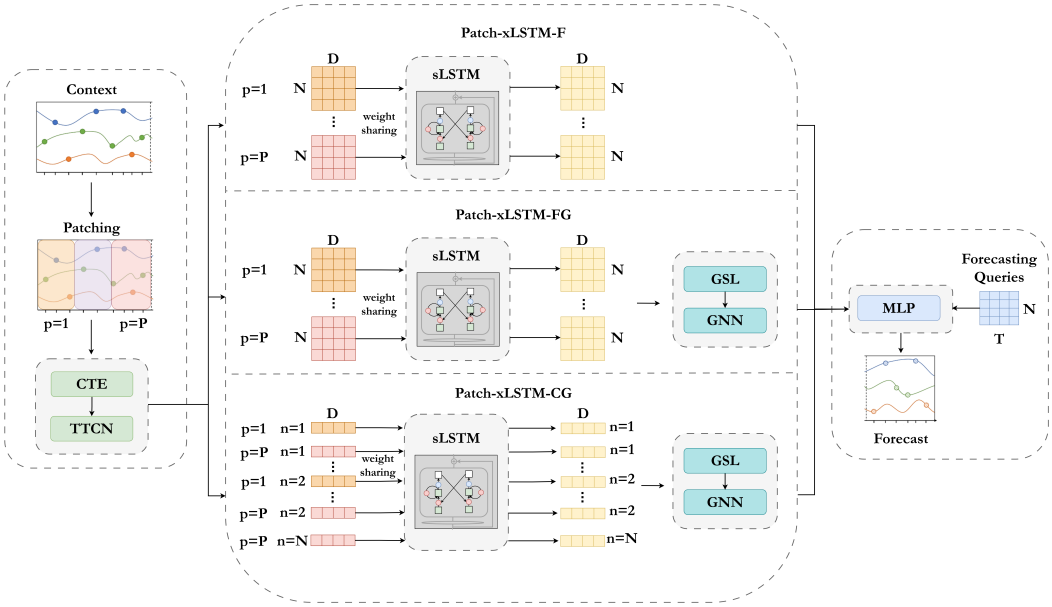

Figure 1: Overview of our proposed approaches. On the left, each patient's time-series are first split into fixed-length patches and encoded into patch embeddings. These are processed either jointly (full-patch), or for each channel separately (per-channel). Given the final representations and future forecasting queries, a MLP produces the irregular MCTS forecast.

Motivated by this limitation, we propose three xLSTM-based model variants for forecasting irregular MCTS with missing values. Our approach integrates xLSTM into the established forecasting architecture T-PatchGNN [29] by replacing components responsible for temporal and cross-channel modeling with xLSTM blocks. We assess whether xLSTM can function as an effective backbone for irregular MCTS forecasting by evaluating our models on a subset of the MIMIC-III dataset, and comparing their performance with various baselines. The findings show that our xLSTM-based variants achieve competitive performance, indicating their suitability for forecasting irregular MCTS.

2 Method

In our work, we investigate xLSTM-based architectures for forecasting irregularly-sampled MCTS with missing values. For this, we integrate xLSTM blocks in the established forecasting model T-PatchGNN [29], which we choose due to its open-source implementation and strong forecasting performance.

Problem Let $\mathbf{X} = \{\{(t_i^n, x_i^n, m_i^n)\}_{i=1}^{T_n}\}_{n=1}^{N}$ denote irregular MCTS data with missing values, where N is the number of channels and the n-th channel contains T_n observations. An observation at time t for channel n consists of a chronological time delta $t_i^n \in \mathbb{R}^+$ as the minutes elapsed since the measurement start, a value $x_i^n \in \mathbb{R}$, and a binary missingness mask $m_i^n \in \{0, 1\}$. Specifically, $m_i^n = 1$ if the measurement exists at t, otherwise $m_i^n = 0$. Irregular forecasting then aims to predict future values at specific requested timestamps along a continuous time axis, which are given through forecasting queries [29]. We denote such queries by $\mathbf{Q} = \{\{q_j^n\}_{j=1}^{Q_n}\}_{n=1}^{N}$, where q_j^n is the j-th forecasting request for channel n. Then, given historical MCTS data $\mathbf{X}$ and queries $\mathbf{Q}$, we want to forecast $\hat{\mathbf{X}} = \{\{\hat{x}_j^n\}_{j=1}^{Q_n}\}_{n=1}^{N}$ according to $\mathcal{F}_\theta(\mathbf{X}, \mathbf{Q}) \to \hat{\mathbf{X}}$, where $\mathcal{F}(\cdot)_\theta$ denotes a forecasting function parameterized by θ whose parameters are learned from data.

Models For forecasting models to be applicable in clinical settings, they should naturally handle irregularly-sampled data and missing values. To achieve this, we propose three forecaster variants $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3$ parameterized by $\theta_1, \theta_2, \theta_3$, and illustrated in Figure 1. The raw MCTS are first partitioned into P non-overlapping consecutive patches, each covering fixed time intervals, resulting in a unified

temporal resolution despite varying observation counts in the patches. Following prior work [29], the continuous time-embedding (CTE) and the transformable time-aware convolution network (TTCN) encode the patches into a sequence of patch embeddings $\mathbf{Z} = \{\mathbf{z}_p\}_{p=1}^{P}$, $\mathbf{z}_p \in \mathbb{R}^{N \times D}$ that are concatenated with the missingness masks, where N is the number of channels and D the embedding dimension. We distinguish two input processing settings. In the full-patch setting, each sLSTM time-step comprises a multivariate patch embedding, allowing the capture of cross-channel dependencies. In the per-channel setting, the time-step is a univariate patch embedding, meaning all channels are processed independently in parallel. In both cases, weights are shared across time-steps and patients. Originally, T-PatchGNN [29] employs a Transformer for temporal processing, followed by graph structure learning (GSL) and a graph neural network (GNN) for cross-channel modeling. Our first model $\mathcal{F}_1$, abbreviated as Patch-xLSTM-F, replaces both with a 2-layer sLSTM, a variant of xLSTM, for jointly modeling channel and temporal dynamics. It processes the full-patch input representation, producing patch representations of dimensions $N \times D$. These are then passed to a multilayer perceptron (MLP) that generates forecasts at given query timestamps $\mathbf{Q}$. The second full-patch variant is $\mathcal{F}_2$, or Patch-xLSTM-FG, which, in contrast to $\mathcal{F}_1$, separates temporal and cross-channel modeling between a 2-layer sLSTM and the GNN before producing forecasts. Finally, $\mathcal{F}_3$ is Patch-xLSTM-CG, where the C denotes the per-channel setting. The temporal dependencies between the patches are processed for each channel independently by a 2-layer sLSTM, before employing GSL and GNN for cross-channel processing.

Data For reproducibility, we evaluate our models on a publicly-available subset of the MIMIC-III dataset, which we further refer to as MIMIC-III-TP.[1] MIMIC-III is a large single-center ICU dataset containing EHR data of patients admitted to Beth Israel Deaconness Medical Center in Boston [9]. After pre-processing, we obtain irregularly-sampled MCTS from 23.457 unique patients, each covering the first 48 hours after ICU admission across 96 different channels. The channels exhibit varying but strong levels of missingness, as well as irregularity, and include laboratory measurements, medication administrations, fluid inputs like insulin, and fluid outputs, such as urine. Each observation is concatenated to a corresponding binary missingness mask indicating whether or not a value is observed at some timestamp.

3 Experiments and results

We evaluate all models on the task of forecasting irregular MCTS over a 24-hour prediction window given the previous 24-hour context, together with a set of forecasting queries. The data is split admission-wise into training, validation and test sets according to a 60-20-20% ratio. Channel values and timestamps are minimum-maximum normalized based on training set statistics, before each admission is partitioned into non-overlapping context and prediction windows. Finally, the sequences are zero-padded to equal length, constituting the input to the model. For this, we follow the protocol established by Zhang et al. [29].

Baselines We compare our proposed variants to 18 baselines covering both regular and irregular forecasting architectures: DLinear [27], TimesNet [23], PatchTST [15], Crossformer [31], CrossGNN [8], Graph WaveNet [24], MTGNN [25], FourierGNN [26], StemGNN [4], SeFT [7], Latent-ODE [17], Neural Flows [3], mTAN [17], CRU [18], GRU-D [5], RAINDROP [30], Warpformer [28], and the original T-PatchGNN [29].

Training and evaluation All models use the Adam optimizer [10] with a learning rate of 0.001, no weight decay, and betas (0.9, 0.999). We set the batch size to 32 and apply early stopping based on validation performance with a patience of 10 epochs. The hidden dimension is fixed to 64 for all models, and hyperparameters unrelated to sLSTM are set as in the original work [29]. The models are optimized using the masked mean squared error (MSE) loss between the predicted and ground truth values at the query timestamps. Model performance is evaluated using the masked MSE and masked mean absolute error (MAE), following prior protocols [29, 11]. We repeat all experiments five times with different random seeds and report the mean and standard deviations of the metrics in Table 1. Separate 100-iteration random hyperparameter searches are conducted over the sLSTM-related parameters, selecting the final parameters based on the lowest validation masked MSE.

[1] https://physionet.org/content/mimic-iii-ext-tpatchgnn/1.0.0/

Table 1: Performance comparison on MIMIC-III-TP for forecasting 24 hours from a 24-hour context window. Lower is better. **Bold** shows the best performance, underlined shows the second-best. † indicates that this result is reported by Zhang et al. [29].

Algorithm	**MSE** $\times 10^{-2}$	**MAE** $\times 10^{-2}$
TimesNet† [23]	5.88 ± 0.08	13.62 ± 0.07
DLinear† [27]	4.90 ± 0.00	16.29 ± 0.05
PatchTST† [15]	3.78 ± 0.03	12.43 ± 0.10
CrossGNN† [8]	2.95 ± 0.16	10.82 ± 0.21
Graph WaveNet† [24]	2.93 ± 0.09	10.50 ± 0.15
MTGNN† [25]	2.71 ± 0.23	9.55 ± 0.65
Crossformer† [31]	2.65 ± 0.19	9.56 ± 0.29
FourierGNN† [26]	2.55 ± 0.03	10.22 ± 0.08
RAINDROP† [30]	1.99 ± 0.03	8.27 ± 0.07
Latent-ODE† [16]	1.89 ± 0.11	8.11 ± 0.52
Neural Flows† [3]	1.87 ± 0.05	8.03 ± 0.06
SeFT† [7]	1.87 ± 0.01	7.84 ± 0.08
mTAN† [17]	1.85 ± 0.06	7.73 ± 0.13
CRU† [18]	1.81 ± 0.05	8.06 ± 0.07
GRU-D† [5]	1.76 ± 0.03	7.53 ± 0.09
Warpformer† [28]	1.73 ± 0.04	7.58 ± 0.13
StemGNN† [4]	1.73 ± 0.02	7.71 ± 0.11
T-PatchGNN [29]	1.71 ± 0.03	7.33 ± 0.10
Patch-xLSTM-FG (ours)	1.76 ± 0.05	7.63 ± 0.2
Patch-xLSTM-F (ours)	1.74 ± 0.06	7.51 ± 0.03
Patch-xLSTM-CG (ours)	**1.68 ± 0.02**	**7.29 ± 0.09**

Results The results of our experiments are summarized in Table 1. Patch-xLSTM-CG achieves the lowest MSE and MAE among the evaluated models, performing on par with other strong baselines, while the original T-PatchGNN [29] ranks second. These findings suggest that sLSTM's powerful state tracking capabilities, as well as capturing temporal dynamics separately for each channel before cross-channel aggregation can be beneficial for learning MCTS. Accordingly, joint modeling temporal and cross-channel dependencies with a recurrent component instead of separating these responsibilities can limit the model's ability to learn channel-specific dynamics, as indicated by Patch-xLSTM-F's results. Patch-xLSTM-FG, performs the worst out of our proposed models, suggesting that mixing channel and temporal information early on leads to the loss of channel-specific information which cannot be recovered by subsequent components. In summary, our results indicate that both per-channel temporal modeling, as well as separate modules for temporal and cross-channel modeling, are beneficial for effective MCTS learning.

4 Discussion and Conclusion

To investigate the potential of xLSTM for forecasting irregular MCTS with missing values, we integrate it as a temporal modeling backbone into an established forecasting architecture [29]. We compare the three resulting model variants to a range of regular and irregular forecasting approaches on MIMIC-III-TP. Our results indicate that employing separate architectural components for temporal and cross-channel modeling, as well as capturing temporal dependencies per-channel, are beneficial for MCTS learning. Looking the beyond predictive performance, sLSTM offers linear complexity and a sequential inductive bias [2], both favorable attributes for clinical time-series forecasting. Overall, our proposed models achieve competitive forecasting performance to strong baselines, positioning xLSTM as a promising temporal backbone for irregular MCTS forecasting.

Acknowledgments and Disclosure of Funding

The authors acknowledge support from the DFG within the SPP2298 under project number 543939932 and from the Austrian Science Fund (FWF) project number 10.55776/COE12.

References

[1] Musleh Alharthi and Ausif Mahmood. xlstmtime: Long-term time series forecasting with xlstm. *AI*, 5(3): 1482–1495, 2024.

[2] Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xlstm: Extended long short-term memory. In *Advances in Neural Information Processing Systems*, volume 37, pages 107547–107603, Red Hook, NY, USA, 2024. Curran Associates Inc.

[3] Marin Biloš, Johanna Sommer, Syama Sundar Rangapuram, Tim Januschowski, and Stephan Günnemann. Neural flows: Efficient alternative to neural odes. In *Advances in Neural Information Processing Systems*, volume 34, pages 21325–21337, Red Hook, NY, USA, 2021. Curran Associates Inc.

[4] Defu Cao, Yujing Wang, Juanyong Duan, Ce Zhang, Xia Zhu, Conguri Huang, Yunhai Tong, Bixiong Xu, Jing Bai, Jie Tong, and Qi Zhang. Spectral temporal graph neural network for multivariate time-series forecasting. In *Advances in Neural Information Processing Systems*, volume 33, Red Hook, NY, USA, 2020. Curran Associates Inc.

[5] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Scientific Reports*, 8(1):6085, 2018.

[6] Chris Giordano, Meghan Brennan, Basma Mohamed, Parisa Rashidi, François Modave, and Patrick Tighe. Accessing Artificial Intelligence for Clinical Decision-Making. *Frontiers in Digital Health*, 3:645232, 2021.

[7] Max Horn, Michael Moor, Christian Bock, Bastian Rieck, and Karsten Borgwardt. Set functions for time series. In *Proceedings of the International Conference on Machine Learning*, pages 4353–4363. JMLR.org, 2020.

[8] Qihe Huang, Lei Shen, Ruixin Zhang, Shouhong Ding, Binwu Wang, Zhengyang Zhou, and Yang Wang. Crossgnn: Confronting noisy multivariate time series via cross interaction refinement. In *Advances in Neural Information Processing Systems*, volume 36, pages 46885–46902, Red Hook, NY, USA, 2023. Curran Associates Inc.

[9] Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, 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(1):160035, 2016.

[10] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[11] Christian Klötergens, Tim Dernedde, Lars Schmidt-Thieme, and Vijaya Krishna Yalavarthi. Mixing it up: Exploring mixer networks for irregular multivariate time series forecasting, 2026.

[12] Maurice Kraus, Felix Divo, Devendra Singh Dhami, and Kristian Kersting. xlstm-mixer: Multivariate time series forecasting by mixing via scalar memories, 2025.

[13] Jeong Min Lee and Milos Hauskrecht. Modeling multivariate clinical event time-series with recurrent temporal mechanisms. *Artificial Intelligence in Medicine*, 112:102021, 2021.

[14] Bryan Lim and Stefan Zohren. Time Series Forecasting With Deep Learning: A Survey. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2194): 20200209, 2021.

[15] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

[16] Yulia Rubanova, Ricky T. Q. Chen, and David Duvenaud. Latent odes for irregularly-sampled time series. In *Advances in Neural Information Processing Systems*, volume 32, Red Hook, NY, USA, 2019. Curran Associates Inc.

[17] Satya Narayan Shukla and Benjamin Marlin. Multi-time attention networks for irregularly sampled time series. In *International Conference on Learning Representations*, 2021.

[18] Sunghyun Sim, Dohee Kim, and Hyerim Bae. Correlation recurrent units: A novel neural architecture for improving the predictive performance of time-series data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):14266–14283, 2023.

[19] Mahanazuddin Syed, Shorabuddin Syed, Kevin Sexton, Hafsa Bareen Syeda, Maryam Garza, Meredith Zozus, Farhanuddin Syed, Salma Begum, Abdullah Usama Syed, Joseph Sanford, and Fred Prior. Application of machine learning in intensive care unit (icu) settings using mimic dataset: Systematic review. *Informatics*, 8(1), 2021.

[20] Davy Van De Sande, Michel E. Van Genderen, Joost Huiskens, Diederik Gommers, and Jasper Van Bommel. Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit. *Intensive Care Medicine*, 47(7):750–760, 2021.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[22] Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Yong Liu, Chen Wang, Mingsheng Long, and Jianmin Wang. Deep time series models: A comprehensive survey and benchmark, 2025.

[23] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023.

[24] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph wavenet for deep spatial-temporal graph modeling. In *Proceedings of the International Joint Conference on Artificial Intelligence*, page 1907–1913. AAAI Press, 2019.

[25] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang. Connecting the dots: Multivariate time series forecasting with graph neural networks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, page 753–763, New York, NY, USA, 2020. Association for Computing Machinery.

[26] Kun Yi, Qi Zhang, Wei Fan, Hui He, Liang Hu, Pengyang Wang, Ning An, Longbing Cao, and Zhendong Niu. Fouriergnn: Rethinking multivariate time series forecasting from a pure graph perspective. In *Advances in Neural Information Processing Systems*, volume 36, Red Hook, NY, USA, 2023. Curran Associates Inc.

[27] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11121–11128, 2023.

[28] Jiawen Zhang, Shun Zheng, Wei Cao, Jiang Bian, and Jia Li. Warpformer: A multi-scale modeling approach for irregular clinical time series. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, page 3273–3285. Association for Computing Machinery, 2023.

[29] Weijia Zhang, Chenlong Yin, Hao Liu, Xiaofang Zhou, and Hui Xiong. Irregular multivariate time series forecasting: A transformable patching graph neural networks approach. In *International Conference on Machine Learning*, 2024.

[30] Xiang Zhang, Marko Zeman, Theodoros Tsiligkaridis, and Marinka Zitnik. Graph-guided network for irregularly sampled multivariate time series. In *International Conference on Learning Representations*, 2022.

[31] Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *International Conference on Learning Representations*, 2023.