Interpretable Vital Sign Forecasting with Model Agnostic Attention Maps

Yuwei Liu[1], Chen Dan[1], Anubhav Bhatti[1], Bingjie Shen[1], Divij Gupta[1], Suraj Parmar[1] and San Lee[1]

[1]*SpassMed Inc., Toronto, Ontario, Canada*

Abstract

Sepsis is a leading cause of mortality in intensive care units (ICUs), representing a substantial medical challenge. The complexity of analyzing diverse vital signs to predict sepsis further aggravates this issue. While deep learning techniques have been advanced for early sepsis prediction, their *'black-box'* nature obscures the internal logic, impairing interpretability in critical settings like ICUs. This paper introduces a framework that combines a deep learning model with an attention mechanism that highlights the critical time steps in the forecasting process, thus improving model interpretability and supporting clinical decision-making. We show that the attention mechanism could be adapted to various black box time series forecasting models such as N-HiTS and N-BEATS. Our method preserves the accuracy of conventional deep learning models while enhancing interpretability through attention-weight-generated heatmaps. We evaluated our model on the eICU-CRD dataset, focusing on forecasting vital signs for sepsis patients. We assessed its performance using mean squared error (MSE) and dynamic time warping (DTW) metrics. We explored the attention maps of N-HiTS and N-BEATS, examining the differences in their performance and identifying crucial factors influencing vital sign forecasting.

Keywords

Time Series Forecasting, Deep Learning, Interpretable Machine Learning, Attention Map, Vital Signs, Sepsis, Explainable AI

1. Introduction

Sepsis is a life-threatening condition that occurs when the immune system of the body responds incorrectly to an infection and causes rapid organ dysfunction and failure. A meta-analysis conducted on articles published in PubMed and the Cochrane Database revealed that the average 30-day mortality rate for sepsis was 24.4%, and the average 90-day mortality rate was 32.2% between 2009 and 2019 [1]. While sepsis has been acknowledged for a long time, its clinical definition did not emerge until the late 20^{th} century [2]. In 1991, a consensus conference posited that sepsis arises from the individual's inflammatory response to infection, marked by systemic inflammatory response syndrome (SIRS), emphasizing the human response to invading organisms. This syndrome is characterized by variations in temperature, heart rate (HR), respiratory rate (RR), blood pressure (BP), and white blood cell (WBC) count [3]. In 2016, the definition of sepsis was revised to multiple organ dysfunction syndrome (MODS) [4]. Systolic

The 2nd World Conference on eXplainable Artificial Intelligence, 17-19 July, 2024, Malta, Valletta

✉ yuwei.liu@spassmed.ca (Y. Liu); chen.dan@spassmed.ca (C. Dan); anubhav.bhatti@spassmed.ca (A. Bhatti); bingjie.shen@spassmed.ca (B. Shen); divij.gupta@spassmed.ca (D. Gupta); suraj.parmar@spassmed.ca (S. Parmar); sanlee@spassmed.ca (S. Lee)

blood pressure (SBP) and RR abnormalities indicate organ dysfunction [5]. Thus, creating precise models for forecasting vital signs becomes essential in predicting sepsis [6]. Accurate vital sign predictions can promptly aid clinicians in identifying and intervening in sepsis cases, potentially saving lives and improving the intensive care unit (ICU) patient outcomes.

The growth in explainable artificial intelligence (XAI) research is mainly attributed to the rapid growth in the popularity of deep learning with widespread healthcare applications. However, most models developed using these technologies are considered *'black-boxes'* by experts due to their intricate, non-linear structures that are challenging for non-experts to understand [7]. The proposed research contributes to the following two aspects: **(1)** Adding an attention mechanism to show the relationship between input time steps and forecasted results; **(2)** Providing analysis and interpretation of the findings derived from the attention map.

1.1. Literature Review

In recent years, the significance of model explainability has been widely recognized, leading to the integration of an increasing number of explainable methods into data-driven models [8]. Prior research has demonstrated the development of deep learning neural networks incorporating attention mechanisms, resulting in interpretable models with strong performance within the medical field. Kaji et al. demonstrated that integrating an attention mechanism into the LSTM network, trained with Electronic Health Record (EHR) data, not only improves the daily sepsis onset prediction's Area Under the Receiver Operating Characteristic Curve (AUROC) score to 0.876 but also highlights critical time points for prediction [9]. An attention-based gated recurrent unit (GRU) was developed by Shickel et al. Self-attention was applied to focus on significant time steps when predicting in-hospital mortality [10]. Choi et al. proposed reverse time attention (RETAIN), processing EHR data in reverse order, achieving an Area Under the ROC Curve (AUC) of 0.87 in heart failure prediction. It adds interpretability using a two-level neural attention model [11].

While previous XAI research integrating deep learning models with interpretable modules has excelled in time series classification, attention mechanisms in interpretable time series forecasting remain underexplored. Our approach aims to explore attention mechanism interpretability in time series forecasting.

2. Method

In this section, we begin by detailing the information of the eICU Collaborative Research Database (eICU-CRD) [12], followed by an outline of the composition of our input data. Subsequently, we dive into the specifics of the attention mechanism and the frameworks of our forecasting models.

2.1. Dataset Description and Data Preprocessing

The eICU-CRD data is a publicly accessible repository containing data from over 200,000 ICU admissions across 208 hospitals in the United States between 2014 and 2015 [12]. This comprehensive dataset comprises diverse patient information, including demographics, diagnoses, medications, and laboratory results. Our research focuses on the ’diagnosis’ and ’vitalAperiodic’ tables, from which we extract dynamic physiological data such as temperature, HR, and BP at 5-minute intervals. The core of our study revolves around forecasting two crucial dynamic variables: HR and mean blood pressure (MBP), derived from SBP and diastolic blood pressure (DBP) measurements. Following the works of [13, 14], we create one or more groups within a 9-hour time window for each patient to predict vital signs for the subsequent 3 hours based on the preceding 6 hours of data. Data preprocessing involves imputing missing values, filtering outliers, and scaling using domain-specific knowledge. Clinically reasonable boundaries for each critical vital sign were set using this specialized knowledge: HR ranged from 0 to 300 bpm, MBP from 0 to 190 mmHg, and RR from 0 to 100 bpm.

2.2. Experiment Setup

The dataset is divided into training, validation, and test sets in an 80:10:10 ratio. Within these intervals, the initial 6 hours consist of 72 time steps, while the subsequent 3 hours encompass 36 time points. The forecasting model integrates either HR alone or HR combined with RR as covariates to forecast MBP or conversely. Training of the model occurs over the first 72 time steps, followed by predictions for the remaining 36 time steps. Ultimately, model performance is assessed through Mean Squared Error (MSE) and Dynamic Time Warping (DTW) evaluations.

2.3. Deep Learning Forecasting Model

Based on the forecasting performance of the N-HiTS and N-BEATS model [15, 16, 17], as well as the idea proposed by Pantiskas et al. [18] we aim to address their inherent lack of interpretability and understand why the model has different performances. To achieve this, we implemented an attention mechanism that can be applied to the N-HiTS and N-BEATS architecture, which may also be applied to other black-box deep learning models. The N-HiTS and N-BEATS model consists of a series of stacks, each responsible for learning residual values from the preceding stack.

Within each stack are blocks comprising several fully connected layers, which generate backward (θ_l^b) and forward (θ_l^f) expansion coefficients according to Equation 1 , where $h_{l,4}$ represents the output of the fourth fully connected layer in the basic block, and $Linear$ denotes

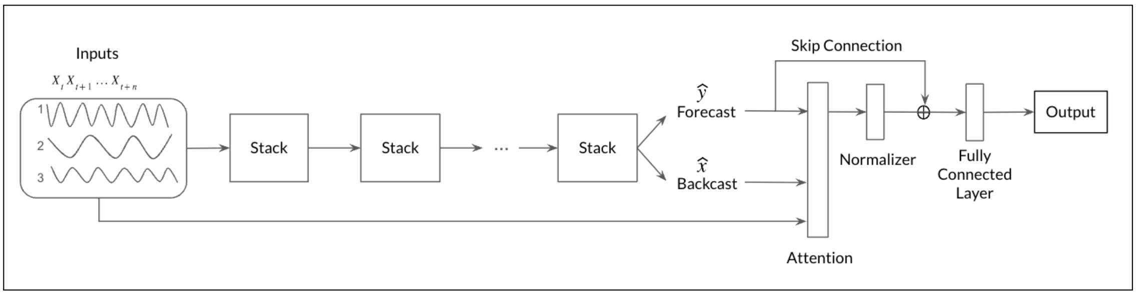


Figure 1: Structure of our implementation. Adding the attention layer at the top of stacks, getting the results from the output.

a linear projection layer [15]:

$$\theta_l^b = \text{Linear}_l^b(h_{l,4}), \quad \theta_l^f = \text{Linear}_l^f(h_{l,4}), \tag{1}$$

Additionally, each block includes backward (g_l^b) and forward (g_l^f) basis layers that produce backcast and forecast outputs as per Equation 2, where $\widehat{y}_l$ and $\widehat{x}_l$ denote forecast and backcast outputs, respectively:

$$\widehat{y}_l = \sum_{i=1}^{\dim(\theta_l^f)} \theta_{l,i}^f \mathrm{v}_{l,i}^f, \quad \widehat{x}_l = \sum_{i=1}^{\dim(\theta_l^b)} \theta_{l,i}^b \mathrm{v}_{l,i}^b. \tag{2}$$

Here, $v_{l,i}^f$ and $v_{l,i}^b$ represent forecast and backcast basis vectors. Notably, for N-HiTS, it has a max-pooling layer (Equation 3) before passing the values to the fully connected layer, which is applied to enable multi-rate signal sampling for the l^{th} basic block [15]:

$$y_{t-L:t,l}^{(p)} = \text{MaxPool}\left(y_{t-L:t,l}, k_l\right), \tag{3}$$

where k_l is the kernel size of the MaxPool layer.

Subsequently, inspired by Pantiskas et al. [18] idea, we introduced an attention mechanism to explore the relationship between learned information and original inputs after obtaining the residuals from the final stack. The forecasted result is utilized to construct the Query (Q), while the original input forms the basis for the Value (V) and Key (K) [18]. The resulting output is computed as follows:

$$O^{N*H} = D \cdot V = softmax(\frac{QK^T}{\sqrt{L}})V \tag{4}$$

$$K^{N*1*L} = I^{N*1*L} \cdot W_K^{N*L*L} + b_K^{N*1*L} \tag{5}$$

$$V^{N*1*L} = I^{N*1*L} \cdot W_V^{N*L*L} \tag{6}$$

and N is the number of input multi time seires, H is the forecasting horizon length, and L is the history input horizon. As shown in Figure 1, after the attention layer, a normalizer is applied, and skip connections are employed to mitigate the vanishing gradient issue. Finally, a fully connected layer is utilized to generate the forecasted results.

2.4. Interpretable Attention Map

To illustrate the attention map for a specific item, we computed [18]:

$$A^{H*L*N} = D^{H*L} \cdot abs(W_v^{N*L*L})^T \tag{7}$$

Here, A^{H*L*N} denotes the attention map, where A_i^{H*L} represents the i^{th} series in the multivariate time series. Each row j in A_i^{H*L} signifies the relationship between the j^{th} forecasted data point and the historical input of length L.

This computation enables the visualization of how the model attends to different historical inputs when forecasting specific data points across the multivariate time series.

Table 1
Performance of forecasting models on forecasting MBP and HR. Here, covariates (W C) for MBP are HR & RR, and covariates for HR are MBP & RR. *The MSE values are scaled by $1e^{-4}$ for better representation. †The DTW values are scaled by $1e^{-3}$ for better representation.

Models	Cov.	MBP (MSE*)	MBP (DTW†)	HR (MSE*)	HR (DTW†)
Persistence [17]	-	24.55	34.50	7.35	17.52
N-HiTS [17]	W C	18.46	18.70	7.37	13.12
N-HiTS [17]	W/o C	**18.02**	20.46	7.22	13.97
N-BEATS [17]	W C	19.79	19.37	8.73	14.36
N-BEATS [17]	W/o C	18.52	**17.63**	7.48	**10.71**
TFT [17]	W C	18.89	25.93	7.71	16.12
TFT [17]	W/o C	19.45	25.65	8.12	16.65
N-BEATS with Attention	W C	21.86	21.07	8.04	14.32
N-BEATS with Attention	W/o C	18.71	18.03	8.40	11.33
N-HiTS with Attention	W C	18.78	20.44	7.24	13.32
N-HiTS with Attention	W/o C	19.73	20.42	**6.97**	12.24

3. Results and Discussion

3.1. Forecasting Benchmarks

Here, table 1 shows the results using different deep learning time series forecasting models. We compared N-HiTS [15], N-BEATS [16], Temporal Fusion Transformer (TFT) [19], which are computed by Bhatti et al. [17] using MSE and DTW as the evaluation metrics.

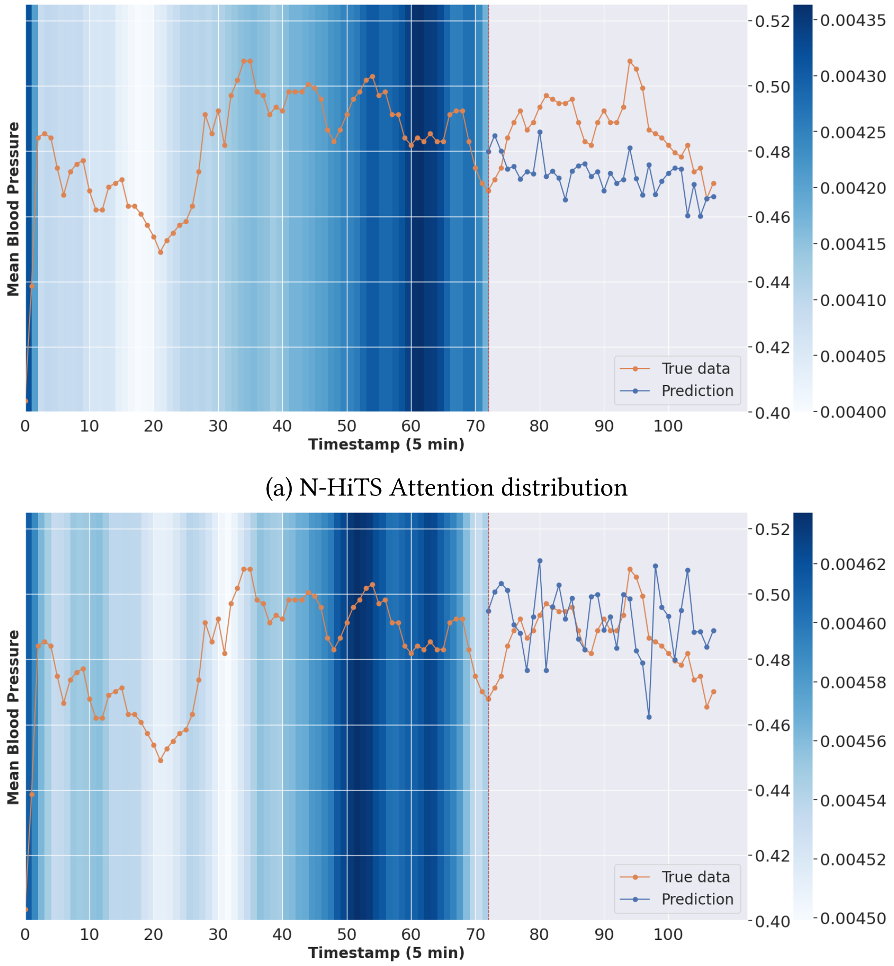
The results indicate that the N-HiTS model, both with and without an attention mechanism, consistently outperforms other models across MBP and HR predictions when considering MSE. Similarly, the N-BEATS model also performs well both with and without attention mechanisms.

Furthermore, the TFT model demonstrates competitive performance, especially when considering MSE. But in the previous paper by Bhatti et al. [17], the forecasting result of TFT is relatively smooth and doesn't show fluctuations.

In conclusion, the N-HiTS model, when augmented with an attention mechanism, emerges as a robust choice for forecasting MBP and HR, showcasing its efficacy in capturing complex temporal patterns. However, further exploration and experimentation are warranted to optimize model performance, particularly regarding temporal alignment and covariate incorporation.

3.2. Interpretability Analysis

In the heatmap provided (Fig 2a, Fig 2b), darker colors indicate higher attention weights at specific time points, which correspondingly have a greater influence on prediction outcomes. Conversely, lighter colors suggest a lesser impact. The "N-HiTS + Attention" in Fig 2a demonstrates that areas after the 20^{th} time point exhibit darker shades compared to earlier sections. Notably, significant changes or peaks at certain points (like the 35^{th}, 54^{th}, and 63^{rd} points) increasingly darken, highlighting their crucial role in shaping the prediction. This pattern suggests that N-HiTS places a stronger emphasis on data after the 20^{th} points, effectively capturing



(a) N-HiTS Attention distribution

(b) N-BEATS Attention distribution.

Figure 2: N-HiTS & N-BEATS with attention using covariates to forecast MBP after minmax filter.

both data fluctuations and overall trends. As a result, the predictions closely align with the actual data and accurately reflect downward trends.

On the other hand, the predictions from N-BEATS do not closely follow the downward trend of the actual data and display considerable fluctuation. This model's attention map reveals that N-BEATS in Fig 2b assigns larger weights to almost every rise and fall (such as at the 3^{rd}, 10^{th}, and 29^{th} points), but without considering if it's worth to focus on the trend, which contributes to less effective information capture. Moreover, it appears that N-BEATS prioritizes data from the initial 1-2 hours more than N-BEATS, contributing to less stable prediction outcomes.

Both models indicate that the initial 1-3 hours are crucial for prediction, suggesting that medical staff should focus on interventions during this period. Significant changes occurring up to three hours prior also substantially impact the predictions.

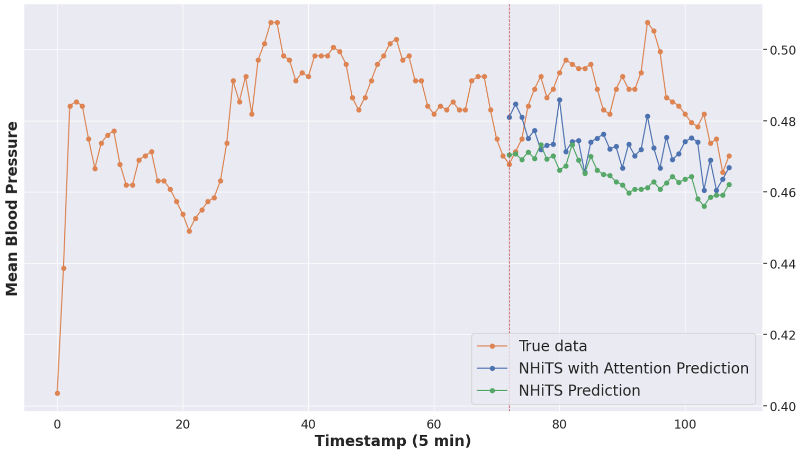


Figure 3: N-HiTS forecasting results with attention using covariates after minmax filter

4. Conclusion

In this paper, we presented an interpretable time series forecasting algorithm that combines black-box deep learning models (N-HiTS & NBEATS) with a general attention mechanism. This approach allows us to observe how the deep learning algorithm assigns importance to inputs while transparently generating each step of its output. Upon applying this advanced architecture to the eICU-CRD dataset, our findings demonstrate that the attention mechanism can enhance interpretability in deep learning time series forecasting models with minimal reduction or even no change in accuracy. By visualizing attention distributions, clinicians can identify which vital signs and historical data points are most influential in predicting sepsis. Furthermore, our model-agnostic attention mechanism is applicable to various deep learning forecasting models.

References

[1] M. Bauer, H. Gerlach, T. Vogelmann, F. Preissing, J. Stiefel, D. Adam, Mortality in sepsis and septic shock in europe, north america and australia between 2009 and 2019—results from a systematic review and meta-analysis, Critical Care 24 (2020) 1–9.

[2] J. E. Gotts, M. A. Matthay, Sepsis: pathophysiology and clinical management, Bmj 353 (2016).

[3] J.-L. Vincent, S. M. Opal, J. C. Marshall, K. J. Tracey, Sepsis definitions: time for change, The Lancet 381 (2013) 774–775.

[4] Z. Cheng, S. T. Abrams, J. Toh, S. S. Wang, Z. Wang, Q. Yu, W. Yu, C.-H. Toh, G. Wang, The critical roles and mechanisms of immune cell death in sepsis, Frontiers in immunology 11 (2020) 1918.

[5] M. Singer, C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G. R. Bernard, J.-D. Chiche, C. M. Coopersmith, et al., The third international consensus definitions for sepsis and septic shock (sepsis-3), Jama 315 (2016) 801–810.

[6] B. Behinaein, A. Bhatti, D. Rodenburg, P. Hungler, A. Etemad, A transformer architecture for stress detection from ecg, in: Proceedings of the 2021 ACM International Symposium on Wearable Computers, 2021, pp. 132–134.

[7] G. Vilone, L. Longo, Notions of explainability and evaluation approaches for explainable artificial intelligence, Information Fusion 76 (2021) 89–106.

[8] L. Longo, R. Goebel, F. Lecue, P. Kieseberg, A. Holzinger, Explainable artificial intelligence: Concepts, applications, research challenges and visions, in: International cross-domain conference for machine learning and knowledge extraction, Springer, 2020, pp. 1–16.

[9] D. A. Kaji, J. R. Zech, J. S. Kim, S. K. Cho, N. S. Dangayach, A. B. Costa, E. K. Oermann, An attention based deep learning model of clinical events in the intensive care unit, PloS one 14 (2019) e0211057.

[10] B. Shickel, T. J. Loftus, L. Adhikari, T. Ozrazgat-Baslanti, A. Bihorac, P. Rashidi, Deepsofa: a continuous acuity score for critically ill patients using clinically interpretable deep learning, Scientific reports 9 (2019) 1879.

[11] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, W. Stewart, Retain: An interpretable predictive model for healthcare using reverse time attention mechanism, Advances in neural information processing systems 29 (2016).

[12] T. J. Pollard, A. E. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, O. Badawi, The eicu collaborative research database, a freely available multi-center database for critical care research, Scientific data 5 (2018) 1–13.

[13] A. Bhatti, N. Thangavelu, M. Hassan, C. Kim, S. Lee, Y. Kim, J. Y. Kim, Interpreting forecasted vital signs using n-beats in sepsis patients, arXiv preprint arXiv:2306.14016 (2023).

[14] H. M. O'Halloran, K. Kwong, R. A. Veldhoen, D. M. Maslove, Characterizing the patients, hospitals, and data quality of the eicu collaborative research database, Critical Care Medicine 48 (2020) 1737–1743.

[15] C. Challu, K. G. Olivares, B. N. Oreshkin, F. G. Ramirez, M. M. Canseco, A. Dubrawski, Nhits: Neural hierarchical interpolation for time series forecasting, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, 2023, pp. 6989–6997.

[16] B. N. Oreshkin, D. Carpov, N. Chapados, Y. Bengio, N-beats: Neural basis expansion analysis for interpretable time series forecasting, 2020. arXiv:1905.10437.

[17] A. Bhatti, Y. Liu, C. Dan, B. Shen, S. Lee, Y. Kim, J. Y. Kim, Vital sign forecasting for sepsis patients in icus, arXiv preprint arXiv:2311.04770 (2023).

[18] L. Pantiskas, K. Verstoep, H. Bal, Interpretable multivariate time series forecasting with temporal attention convolutional neural networks, in: 2020 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, 2020, pp. 1687–1694.

[19] B. Lim, S. Ö. Arık, N. Loeff, T. Pfister, Temporal fusion transformers for interpretable multi-horizon time series forecasting, International Journal of Forecasting 37 (2021) 1748–1764.