ISSN: 2835-7329

Research Article

Dermatology Journal of Cosmetic and Laser Therapy

Physiological Vital Time Series Forecasting using Fractional Calculus and Deep Neural Network

Sama Nemati[1], Seyed Amin Seyed Jafari[1], Mostafa Fakhri[1], Kosar Seraji[2], Farzane Vosoughi-Motlagh[3] and Mojtaba Hajihasani[4]*

[1]*Institute of Medical Sciences and Technology, Shahid Beheshti University, Tehran, Iran*

[2]*Department of Computational Quran Mining at Quran Miracle Research Institute, Shahid Beheshti University, Tehran, Iran*

[3]*Management and Accounting Faculty, Department of Industrial and Information Management, Shahid Beheshti University, Tehran, Iran*

[4]*Department of Modern Engineering, Amol University of Special Modern Technology, Tehran, Iran*

***Corresponding Author**
Mojtaba Hajihasani, Department of Modern Engineering, Amol University of Special Modern Technology, Tehran, Iran.

Submitted: 2024, Aug 06; **Accepted:** 2024, Aug 28; **Published:** 2024, Sep 10

Citation: Nemati, S., Seyed Jafari, S. A., Fakhri, M., Seraji, K., Vosoughi-Motlagh, F., et al. (2024). Physiological Vital Time Series Forecasting using Fractional Calculus and Deep Neural Network. *Dearma J Cosmetic Laser Therapy, 3*(2), 01-11.

Abstract

Continuous physiological monitoring integrated with time series analysis and multi-step forecasting is vital when encountering postoperative cases either hospitalized in intensive care units (ICU) or given home health care will experience adverse cardiac events. The low-cost common vital signs, i.e., heart rate and arterial blood pressure are captured and predicted with adjustable horizons up to 30 minutes in advance to achieve punctual clinical decision-making to prevent the events of bradycardia, tachycardia, hypo-tension, and hypertension. Scaling properties of physiological stationary/non-stationary signals are necessarily determined and drastically affected by the selection and architecture design of time series forecasting models. In contrast to integer-order difference that achieves stationary memory-erased series, fractional order difference ensures the stationary of the data while preserving as much memory as possible. The deep learning architecture for multi-step forecasting is the combination of two direct and iterative methods which utilizes the convolutional neural networks with skip connections inspired by the concepts of U-Net convolutional networks and multi-layer bi-directional long short-term memories (Bi-LSTMs). Various scenarios of observe-target windows e.g. (20, 30, 60, or 120) - (7, 15, 20, or 30) minutes are trained using hyper-parameter tuning and evaluated by mean absolute percentage error (MAPE). The results of the proposed method indicate that crucial vital signs such as heart rate, systolic blood pressure and mean arterial blood pressure will be predictable in an adjustable observe-target window size from 20-7 to 120-30 minutes with narrow ranges of MAPE values between [2.78%, 4.17%], [4.69%, 6.47%] and [4.45%, 6.86%], respectively.

Keywords: Time Series Forecasting, Non-Stationary Process, Fractional Calculus, Machine Learning Modeling, Adverse Clinical Events

1. Introduction

Continuous physiological monitoring, crucial for preventing postoperative complications, has spurred biomedical research interest [1-3]. Despite advancements in prediction through classification algorithms, accurately forecasting vital signs remains a challenge, necessitating exploration into deep learning methods [4-6]. Such forecasts enable early intervention, enhancing patient care and informing the design of intelligent alarm systems [7]. Advanced deep learning models have shown promise in prognostic prediction for patients in intensive care units (ICUs) [3,8]. Bontempi et. al. and Masum et. al. described and compared five different forecast strategies including the Recursive strategy, Direct strategy, Direct-Recursive (DirRec) strategy, Multi-Input Multi-Output (MIMO) strategy, and Direct Multi-Output (DIRMO) strategy [7,9]. All strategies utilized the combination of long short-term memory (LSTM), Bidirectional-LSTMs (Bi-LSTM), and Convolutional Neural Networks (CNN). In 2019, Liu et. al. proposed a new approach called generative boosting that includes two parts of the predictive and generative models [10]. Generative boosting utilizes LSTM for both parts leading

to a scheme called generative LSTM (GLSTM). The first model consists try to generate synthetic data for the next few time steps, and the second models, try to make long-range predictions based on observed and generated data. Generative boosting mitigates the error propagation in the generative models and reduces the effective prediction horizon in the predictive models. They showed that GLSTM outperforms efficient benchmark models, in such a way that the mean absolute percentage errors (MAPE) of 7.41% and 6.17% were achieved to predict heart rate (HR) and systolic blood pressure (SBP) 20 minutes in advance, respectively [10]. In 2020, Youssef et. al. proposed a hybrid machine learning algorithm of KNN-LS-SVM instead of LSTM-based models for real-time early warning scores (EWS) estimation and vital signs time-series prediction [11]. They preserved at least one-hour statistical attributes of the different vital signs (i.e., minimum, mean) as input data to forecast statistical attributes one, two, and three hours in advance [11]. They achieved the MAPE of predicting a one-hour average heart rate are 4.1, 4.5, and 5% for the next one, two, and three hours, respectively, for cardiology patients.

In general, the development of intelligent monitoring systems faces two notable challenges. Firstly, the relative scarcity of adverse events, particularly during the early postoperative phase, poses a significant obstacle [12,13]. Secondly, ensuring stationarity in time series data is imperative for effective modeling strategies [14]. The rarity of adverse events, particularly in the early postoperative period, poses a significant challenge to the clinical precision of intelligent monitoring systems. Late detection of clinical instability often leads to delayed recognition and reduced successful clinical intervention, as evidenced by investigations into patterns of in-hospital deaths [15]. To address this challenge, ElMoaqet et al. (2016) developed a framework for multi-step ahead prediction models, introducing a performance metric tailored to compensate for and resolve issues in intelligent monitoring systems. This metric evaluates near-term predictions of critical levels of anomaly in physiological time series [1]. The second challenge involves ensuring stationarity in time series data, which is crucial for effective analysis. Stationary and non-stationary time series determine not only the form of auto-correlations and moments but also impact the selection of estimators and models [16]. A time series is considered stationary when its key properties, such as mean, variance, and auto-correlation structure, remain constant over time [14,17].

To better understand the underlying dependencies of time series data, employing decomposition models or performing a basic data cleaning process to classify signals into stationary or non-stationary categories are essential for subsequent analysis and forecasting [14]. However, applying an analysis or modeling technique developed for stationary conditions to a non-stationary signal can lead to ambiguity or significant performance reduction. Formal statistical tests for stationarity include unit root tests, with the augmented Dickey–Fuller test being one common approach [18]. Additionally, the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test is widely used to determine if a time series is non-stationary due to a unit root or stationary around a mean or linear trend [19].

The sources of data may lead to non-stationary time series, influenced by systematic components and dependencies on past values. Removing systematic trends and seasonal effects that are not of interest at the mean level of the series is essential. The primary method for achieving this is differentiation, which transforms non-stationary time series into stationary ones by eliminating trend effects [14]. In cases where the time series exhibits a varying trend, a first difference may not suffice to achieve complete stationarity. Higher orders of differentiation may be required. However, in practice, first or second differences often render the mean stationary, and further differentiation is rarely necessary [17,20]. While integer order difference transformations effectively render the series stationary, they come at the cost of removing all memory from the original series [21]. This poses a dilemma, particularly in time series forecasting, where preserving memory is essential for predictive modeling purposes.

This study introduces a novel preprocessing approach based on fractional-order differentiation, followed by a deep learning architecture comprising convolutional and multi-layer Bidirectional Long Short-Term Memory (Bi-LSTM) networks.

2. Materials and Methods

2.1 Dataset

Waveform Database Matched Subset of the Multi-parameter Intelligent Monitoring in Intensive Care (MIMIC)-III, a dataset containing 22,247 numeric records from 2001 to 2012 for 10,282 patients in ICU, has been collected at the Beth Deaconess Medical Center in Boston, Massachusetts [22]. Records that contain periodic measurements including HR (times/min), SBP (mm-Hg), and diastolic blood pressure (DBP, mm-Hg) have been selected. We use records in which all three desired signals were present simultaneously. Records with a frequency of 1/60 Hz i.e. one sample per minute and a length of more than 180 minutes have been selected. Therefore, 4141 out of 22247 records have been nominated according to the above criteria. The total length of all selected records is 212947 hours, in which the minimum, maximum and mean record length are 180, 34615, and 3085 minutes, respectively.

2.2 Proposed Method

Pre-processing

2.2.1 Windowing: The various time windows are defined for training of the deep learning model. The observe lengths vary from 20-, 30-, 60-, and 120-minutes time windows followed consecutively by the different predictive target length from 7, 15, 20, and 30 minutes ahead. Therefore, number of datasets have been created based on different combinations of observe-target length such as 20-7, 20-15, 30-7, 30-15, 30-20, 60-7, 60-15, 60-20, 60-30, 120-7, 120-15, 120-20, and 120-30. In each dataset the entire records break down into the pieces of one of the observe-target length consecutively without any overlap i.e. a piece of 20-7 window is include 27 consecutive samples of a record that is supposed to predict the last 7 minutes from the first 20 minutes observant input. Each dataset is dedicated to train one related DL model architecture.

2.2.2 Filtering and Scaling: To prepare windows for training

the model, at the first stage, they smoothed, and the noise was reduced by a convolution filter with order of 2. Then, the whole records were globally scaled by the Min-Max scaler.

2.2.3 Fractional Differencing: The notion of fractional differentiation applied to the time series has been developed by Hosking in 1981 [23]. In the following, the concept of fractional differentiation is described in detail. Assume a time series X_t is not stationary and let B the backshift operator ($B^k X_t = X_{t-k}$) which is traditionally denoted for integer difference as following equation:

$$\nabla^d X_t = (1-B)^d X_t \quad (1)$$

In a fractional differentiation, the exponent d can be a real number, with the following binomial series expansion:

$$(1-B)^d = \sum_{k=0}^{\infty} \binom{d}{k} (-B)^k = 1 - dB + \frac{d(d-1)}{2!} B^2 - \frac{d(d-1)(d-2)}{3!} B^3 + \cdots \quad (2)$$

Despite integer d, the weights, ω_k, in Eq. 3 will not be zero in real value d that means to preserve memory. Therefore, the current value in time series depends on all the past values that occurred. From Eq. 2 the weights can be generated by following iterative scheme:

$$\omega_k = -\omega_{k-1} \frac{d-k+1}{k} \quad (3)$$

Although, there is an explicit expression (Eq. 3) for fractional order difference but in practice due to data limitations, the fractionally differentiated values cannot be computed on an infinite series of weights. Therefore, two alternative implementations of fractional differentiation have been proposed, i) the standard "expanding window" method, and ii) an efficient method based on FFD. In FFD, the weights are kept based on their modulus ($|\omega_k|$) values more than a given threshold while the remains are dropped. This modification results in the advantage that the same vector of weights is used across the entire time series differentiating, thus avoiding the negative drift caused by an expanding window's added weights.

2.2.4 Model: We used a hybrid deep learning network architecture consists of CNN and Bi-LSTM layers, as illustrated in Fig 1. CNNs are well suited for learning and extracting salient features from an input feature, while LSTMs can capture temporal information from time series data. The network architecture comprises eight one-dimensional convolutional layers, forming a U-Net-like structure. As the original U-Net [24] model, it incorporates skip-connections to combine low-level feature maps with high-level feature maps. Our proposed CNN model is facilitated by the advantages of the skip connections. Each convolutional layer performs a convolution operation using a kernel size of 16 followed by a Rectified Linear Unit (ReLU) activation function, which introduces non-linearity into the model.

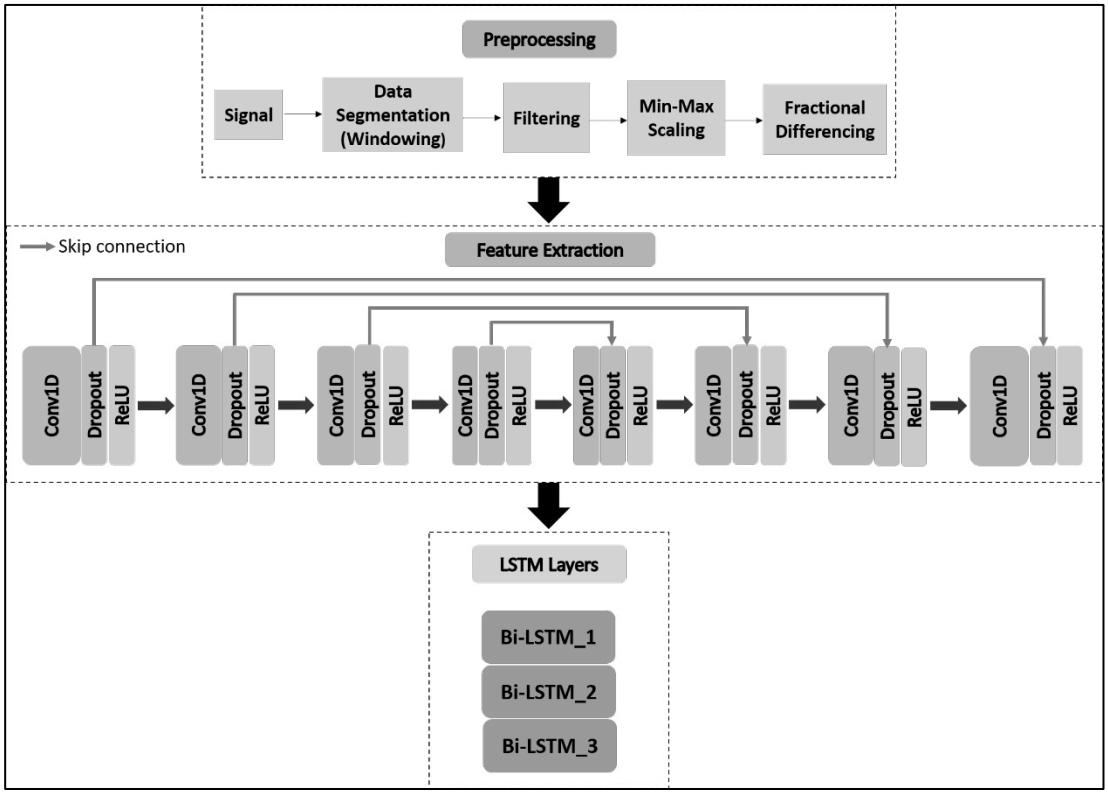

Figure 1: The Whole Framework of the Proposed Method including Preprocessing and Dl Model Structure

After eight convolutional layers, the resulting feature map, after flattening, is fed into a dense layer with as many units equal to the target length. The output of the dense layer is fed into the first Bi-LSTM layer. Three Bi-LSTM layers are employed in the model architecture. Multiple Bi-LSTM layers allow the hidden state at each layer of the network to operate at a different time scale, thus enabling the model to capture a wide range of temporal dynamics. A Bi-LSTM layer produces an output sequence represented as a vector, which is then used as input to a subsequent Bi-LSTM layer. The output shape of the last Bi-LSTM layer is subsequently fed into a time-distributed layer. A time-distributed layer is a wrapper that allows dense layers to process time-series inputs. The output of the time-distributed layer is the model's output, the shape of which is dependent on the target length. The model's output includes the observed sequence (input) and predicted values for subsequent time points. Although the output of our proposed model is a multi-step forecast, some adverse clinical events including Bradycardia, Tachycardia, and Hypotension are classified. There is no gold standard for these events, however, we use the following thresholds:

According to the national institutes of health, if the HR of an adult goes below 60 bits/min, then it is a bradycardia [25]. If a patients' Systolic BP goes above 140 mmHg for at least two consecutive samples then that event is known as a Hypertension. If a patients' Mean Arterial Blood Pressure (MAP) drops below 70 mmHg for at least two consecutive samples then that event is known as a Hypotension.

2.2.5 Training: We utilized Adam's optimization algorithm to update the weights. The loss function employed in the model output was the MAPE. We trained each network with an initial learning rate of 0.0001, which was dropped by a factor of 10 when validation loss did not improve after 5 epochs. The training was stopped using the Early Stopping callback, and a batch size of 16 was used to fit in the GPU memory.

The MAPE utilized for training the model by ADAM optimization. The other metrics including mean square error (MSE) and mean absolute error (MAE) have been calculated in addition to MAPE providing better insight about comparison of the results. The performance of the suggested structure was evaluated by these common regression criteria MAPE, MAE, and MSE as well as some criteria driven by the confusion matrix.

Furthermore, we report accuracy (ACC), positive predictive value (PPV), and Mathews Correlation Coefficient (MCC) in addition to four confusion matrix categories (True Positive, False Positive, True Negative, and False Negative) which are more interpretable for evaluating predicted signals. ACC simply means the number of values correctly predicted (it measures the fraction of correct predictions). PPV is used to indicate the probability that in case of a positive test that the patient has the specified disease. MCC is a more reliable statistical rate that produces a high score only if the prediction obtained good results in all of the four confusion matrix categories.

3. Results

In the present study, fractional differentiation is computed using FFD, which is, since the coefficients are driven by Eq. (3) tend to zero (Figure. 2), a given threshold determines how many coefficients will be preserved [21]. The first eight coefficients are kept for applying FFD, therefore the driven signal is lost in the first seven entities. There are two well-known unit root tests to determine whether a given time series is stationary, including the augmented Dickey–Fuller (ADF) test and the KPSS test [26]. The null hypothesis of the ADF test assumes non-stationary, whereas the null hypothesis of the KPSS test is stationary.

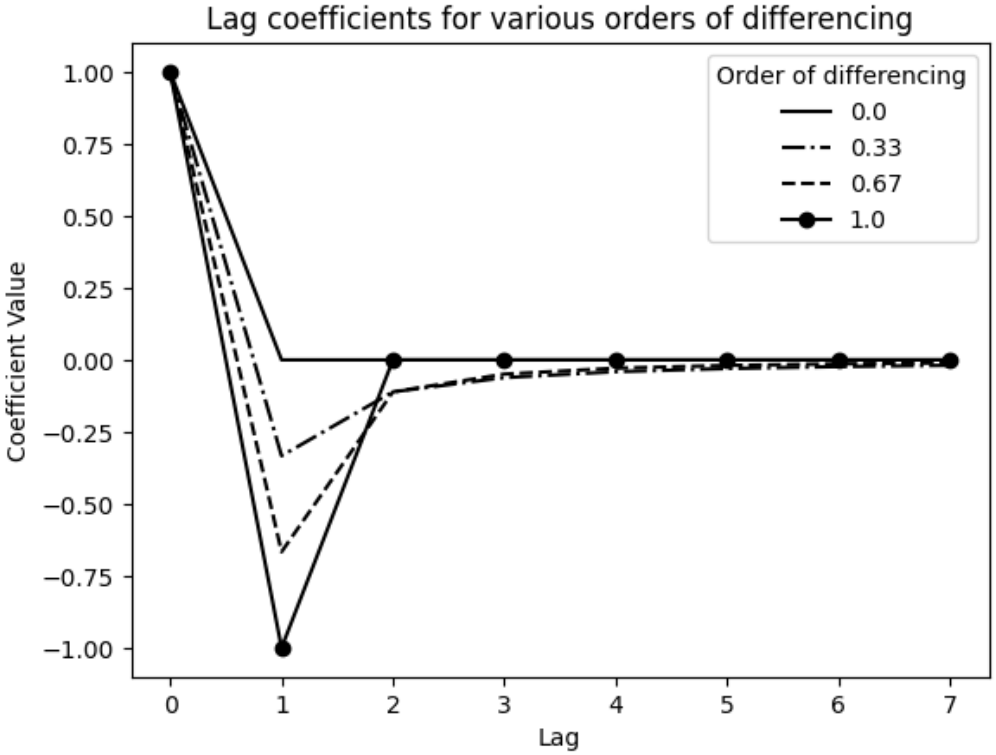

Figure 2: Lag Coefficients for various orders of differentiating between 0 and 1

To find the optimized order of differentiating, the combination of the ADF and KPSS test statistics and Pearson correlation coefficient were utilized. These curves are depicted in Figure. 3. The dashed horizontal line represents the threshold if the ADF test statistic passes through, then it ensures that the fractionally difference series is stationary. The same concept with dash-dotted horizontal line which is associated with KPSS test statistic. A sample HR time series was examined in Figure. 3 that represents the behavior of such signal from our dataset and can help to adjust the difference order for further simulation. In Figure. 3 for the sample HR series, the difference order is about 0.57 and 0.91 accordingly to ADF and KPSS criteria, respectively, which shows a substantial variation. On the other hand, finding different orders based on this approach depends on the length of the time series. Additionally, the inverse fractional-order difference operation is required in practice to demonstrate the predicted series in the original scales. The inverse operation contains an intrinsic deterioration, especially in the difference order close to 1. On another note, the Pearson correlation is on the left y-axis, showing the correlation between the original series and the driven series which is almost 0.6 and 0.25 respectively to a difference order of 0.57 and 0.91. When the first-order difference is applied to the series, the Pearson correlation drops drastically down. For all these essential reasons, we decrease and adjust the difference order to 0.3 in this study to utilize fractional difference features while tolerating a little non-stationary trait. However, we examined the different order of a range between 0.3 to 0.6 which consequence a subtle performance variation.

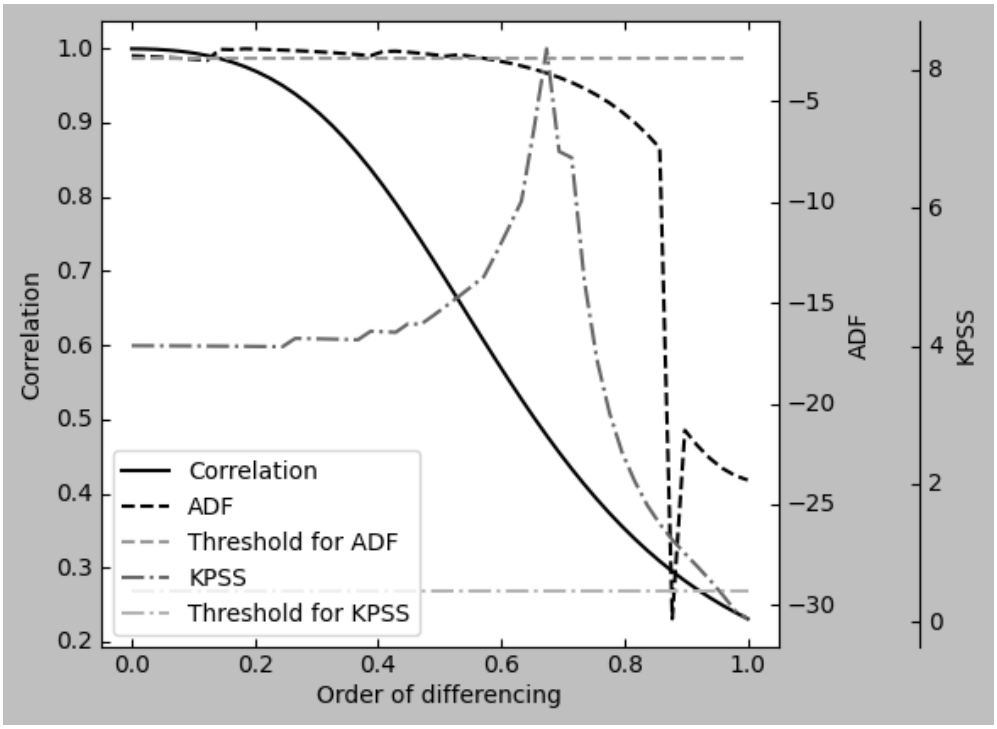

Figure 3: ADF and KPSS Test Statistics and Pearson Correlation Coefficients with the Original Time-Series for Fractional Orders of Difference in the Range Of [0,1], Applied to a Sample of HR Signal

Our study used the CNN with skip connections and Bi-LSTM recurrent neural network to forecast the future of HR, SBP, and mean arterial blood pressure (MBP) [5] from univariant time-series data. In U-Net, by exploiting the auto-encoder structure, it is possible to reach salient features in the bottleneck. Also, utilizing skip connections serve to push more details between two linked down-sample and up-sample layers [5, 24]. We operate hyper-parameter tuning to set the hyper-parameters such as the number of layers and filters. Unlike the original U-Net in which filters were designed in ascending order, we got the best result with filters in descending order which can be related to a number of salient features that are required to restore the rest of the signal. Each model was evaluated using MAPE, MAE, and MSE of different observe-target windows (20-7 min, 20-15 min, 30-7 min, 30-15 min, 30-20 min, etc.) according to tables1-4. In all tables, the best value of each criterion is bolded and the worst one has been underlined.

We experimentally demonstrate the model performance in terms of classification of clinical events using confusion matrix criteria on predicted target window.

Table. 1 indicates the result of evaluating predicted HR signals based on three classes Normal, Bradycardia, and Tachycardia. Among all results for HR, the MAPE, MAE, and MSE had the lowest values (2.78, 2.17, and 15.2, respectively) all for a window size of 120-7 while their maximum values are 4.17, 3.24, and 29.09 respectively all related to the window size of 120-30. Moreover, the true positive of normal class has a maximum value of 98.8% for a window size of 30-15 and a minimum value of 97.0% for a window size of 20-7. True positive of bradycardia has a maximum value of 92.03% related to 120-7 and the minimum value of 76.4% related to 20-15. True positive ratio of tachycardia also has a maximum value of 86.9% for 20-7 and the minimum value of 72.8% for 60-30. The false positive of bradycardia has been a range of [0.4%, 1.1%] for 30-

15 and 30-7 respectively. The false positive of tachycardia has been minimum in both window sizes of 30-15 and 60-15 with the same values of 0.8 while it has been maximum in 20-7 with the value of 2.3. Accuracy (ACC) has also been calculated in a range of [82.96%, 92.03%] related to 60-30 and 120-7. Positive predictive value (PPV) for bradycardia has a minimum value of 98.67% for 30-7 and a maximum value of 99.49% for 30-7. PPV for tachycardia has a minimum value of 97.14% for 30-20 and a maximum value of 99.03% for 60-15. The Matthew's correlation coefficient (MCC) for bradycardia has a minimum value of 0.81% for 60-20 while it has a maximum value of 0.93% for 120-7. MCC for tachycardia has a minimum value of 0.78% for 60-30 while it has a maximum value of 0.88 for all windows with 7 minutes target length.

Performance	**Metrics**	**Observation length (min)**	**20**		**30**			**60**				**120**			
		Target length (min)	**7**	**15**	**7**	**15**	**20**	**7**	**15**	**20**	**30**	**7**	**15**	**20**	**30**
Loss Functions	MAPE		3.19	3.75	3.03	3.68	4.06	2.83	3.43	3.84	4.17	**2.78**	3.23	3.56	3.94
	MAE		2.49	2.94	2.33	2.84	3.03	2.18	2.65	3.01	3.24	**2.17**	2.5	2.77	3.07
	MSE		20.85	26.74	18.39	25.7	28.08	15.21	20.85	27.87	29.09	**15.2**	19.29	21.82	27.41
Confusion Matrix	T-Normal	(%)	97.0	97.5	**97.8**	98.8	97.3	98.0	98.5	98.0	98.0	97.4	98.2	98.0	98.1
	FP (Brady)	(%)	0.8+0.	0.6+0.	1.0+0.1	0.4+0.	0.7+0.	0.6+0.	0.7+0.0	0.3+0.2	0.7+0.1	0.8+0.00	0.8+0.00	0.5+0.00	0.9+0.0
	FP (Tachy)	(%)	2.2+0.1	1.9+0.1	1.2+0.	**0.8+0.**	2+0.3	1.4+0.	**0.8+0.0**	1.7+0.3	1.3+0.1	1.8+0.2	1.0+0.1	1.5+0.2	1+0.1
	T-Brady	(%)	82.4	76.4	82.1	79.1	76.7	83.9	84.0	75.4	77.9	**92.3**	89.0	85.0	80.4
	T-Tachy	(%)	**86.9**	81.5	86.2	79.2	78.3	86.1	82.5	80.5	72.8	86.2	83.4	79.7	76
Confusion Matrix Metrics	ACC	(%)	88.8	85.16	88.73	85.7	84.2	89.33	88.33	84.8	82.96	**92.03**	90.23	87.63	84.86
	PPV (Brady)	(%)	99.03	99.22	98.67	**99.49**	99.09	99.28	99.17	99.34	98.98	99.14	99.1	99.41	98.89
	PPV (Tachy)	(%)	97.42	97.6	98.62	99.0	97.14	98.4	**99.03**	97.57	98.11	97.73	98.69	97.91	98.57
	MCC (Brady)		0.86	0.82	0.85	0.84	0.82	0.87	0.87	0.81	0.83	**0.93**	0.91	0.88	0.84
	MCC (Tachy)		**0.88**	0.84	**0.88**	0.84	0.82	**0.88**	0.86	0.83	0.78	**0.88**	0.86	0.83	0.81

Table 1: The Validation Loss Values as well as Classification Performance for Various Observation and Target Lengths in HR.

Table. 2 shows the performance of the proposed model for SBP assumed on two normal and hypo-tension classes. According to the table, the SBP model performed best for the MAPE and MAE both in 120-7observe-target windows with values of 4.69% and 5.93 while MSE performed best in 60-7 window size with a value of 103.4. Also, MAPE and MAE have had maximum values of 6.88% and 8.46 for a window size of 60-30 and MSE has had the maximum value of 164.31 for 120-30. True positive ratio besides false positive ratios have shown their best values of 98.38% and 1.62% respectively both for 60-30. They also have had their worst values of 96.75% and 3.25% both for 20-7 window sizes. True positive ratio and false negative ratio both have had the best performance in 120-7 window size and the worst one in 60-30 with values of 79.95% and 20.05% besides 61.30% and 38.7% respectively. Accuracy, MCC, and PPV have the maximum values of 88.56%, 97.51% and 0.78% all related to the 120-7 observe-target window size. On the other hand, ACC has a minimum value of 79.83% for 60-30, PPV has a minimum value of 95.81% for 20-7 and MCC has a minimum value of 0.64 for two 60-30 and 120-30 window sizes.

Performance	Metrics	Observation length (min)	20		30			60				120			
		Target length (min)	7	15	7	15	20	7	15	20	30	7	15	20	30
Loss Functions	MAPE		5.61	6.37	5.33	6.19	6.52	4.93	6.02	6.44	6.88	**4.69**	5.67	6.01	6.82
	MAE		7.11	7.97	6.73	7.73	8.12	6.26	7.50	7.86	8.46	**5.93**	7.09	7.47	8.36
	MSE		137.98	156.55	121.76	145.8	156.5	**103.4**	139.2	143.05	161.83	103.91	128.1	136.37	164.31
Confusion Matrix	TNR	(%)	96.75	97.38	97.38	98.08	97.97	97.97	97.70	97.78	**98.38**	97.18	98.2	98.2	97.85
	FPR	(%)	3.25	2.62	2.62	1.93	2.02	2.02	2.30	2.23	**1.62**	2.82	1.8	1.8	2.15
	TPR	(%)	74.45	67.47	76.08	66.90	66.40	75.88	70.25	67.12	61.30	**79.95**	70.53	67.8	62.3
	FNR	(%)	25.55	32.52	23.93	33.10	33.60	24.12	29.75	32.88	38.7	**20.05**	29.48	32.2	37.7
Confusion Matrix Metrics	ACC	(%)	85.6	82.42	86.72	82.48	82.18	86.92	83.97	82.45	79.83	**88.56**	84.36	83.0	80.07
	PPV	(%)	95.81	96.25	96.66	97.2	97.04	97.4	96.82	96.79	97.41	96.59	**97.51**	97.41	96.66
	MCC		0.73	0.68	0.75	0.68	0.67	0.75	0.7	0.68	0.64	**0.78**	0.71	0.69	0.64

Table 2: The Validation Loss Values as well as Classification Performance for Various Observation and Target Lengths in SBP.

Table. 3 is a description of the MBP model based on two different classes of normal and hypo-tension. MAPE, MAE, and MSE are in the ranges of [4.45, 6.47], [3.31, 4.93], and [35.58, 61.38] respectively which all minimums are related to 120-7 and the maximums of MAPE and MAE both occurred in 60-30 though the maximum of MSE is in 120-30. Also, the true negative ratio with the range of [93.3, 95.9] along with the false positive ratio in the range of [4.1, 6.7] and PPV with the range of [92.39, 94.61] have had their best and worst values in the same window sizes of 60-30 and 30-7 respectively. True positive ratios with values in the range of [72.1, 86.6] and false negative ratios with values in the range of [13.4, 27.9] both have performed well in 120-7 while they have had their lowest achievements in 60-30. ACC with values between 84.52 related to 30-15 and 90.58 related to 120-7 has carried the least and greatest results respectively. Finally, the results of MCC with the minimum amount of 0.7 associated with 120-7 and the maximum amount of 0.81 associated with 30-15, 30-20, and 60-30 are seen in the table below.

Performance	Metrics	Observation length (min)	30			60				120			
		Target length (min)	7	15	20	7	15	20	30	7	15	20	30
Loss Functions	MAPE		5.09	5.97	6.44	4.63	5.61	6.11	6.47	**4.45**	5.4	5.88	6.44
	MAE		3.8	4.53	4.85	3.47	4.21	4.59	4.93	**3.31**	4.01	4.43	4.88
	MSE		44.26	55.45	59.94	36.40	47.9	57.16	59.64	**35.58**	45.23	56.48	61.38
Confusion Matrix	TNR	(%)	93.3	94.67	94.58	95.17	95.2	95.03	**95.9**	94.58	95	94.88	95.33
	FPR	(%)	6.7	5.33	5.42	4.83	4.8	4.98	**4.1**	5.42	5	5.12	4.67
	TPR	(%)	81.4	74.38	74.72	83.93	77.68	76.33	72.1	**86.6**	82.04	79.33	75.15
	FNR	(%)	18.6	25.62	25.27	16.07	22.32	23.67	27.9	**13.4**	17.95	20.67	24.85
Confusion Matrix Metrics	ACC	(%)	87.35	84.52	84.65	89.55	86.43	85.67	84.0	**90.58**	88.52	87.1	85.23
	PPV	(%)	92.39	93.31	93.23	94.56	94.18	93.88	**94.61**	94.1	93.67	93.93	94.14
	MCC		0.75	0.7	0.7	0.79	0.74	0.72	0.7	**0.81**	0.77	0.75	0.71

Table 3: The Validation Loss Values as well as Classification Performance for Various Observation and Target Lengths in MBP.

Tables 4 and 5 present a comparative analysis of network performance based on varying orders of differencing and with the inclusion of no differencing input data, for both HR and MBP datasets. The observation-target length considered was 20-7, and the evaluation metrics used include MAPE, MAE, and MSE. The results indicate that the most favorable performances were achieved through differencing with an order of 0.3. However, it is noteworthy that, in Table 4, the MSE for HR was marginally lower when differencing was performed with an order of 0.4. Additionally, it is observed that as the order of differencing increases, the performance deteriorates. This increasing trend keep preserving in incremental difference orders, such as 0.5 and 0.6.

	0	**0.3**	**0.4**	**0.5**	**0.6**
MAPE	4.44	**3.19**	3.22	3.60	4.72
MAE	3.36	**2.49**	2.51	2.73	3.41
MSE	28.06	20.85	**18.89**	24.11	33.71

Table 4: Comparison of Fractional Differences with Orders 0.3, 0.4, 0.5, and 0.6, along with the Standard Difference of 0, for Observation-Target Lengths of 20-7 in HR.

	0	**0.3**	**0.4**	**0.5**	**0.6**
MAPE	7.54	**5.09**	5.77	7.15	10.26
MAE	6.04	**3.8**	4.27	5.4	8.23
MSE	79.64	**44.26**	**53.5**	70.56	19360

Table 5: Comparison of Fractional Differences with Orders 0.3, 0.4, 0.5, and 0.6, along with the Standard Difference of 0, for Observation-Target Lengths of 20-7 in MBP.

Tables 6 and 7 provide a comparison between the proposed method and ARIMA (Autoregressive Integrated Moving Average), as a baseline solution, for forecasting HR, SBP, and MBP. The evaluation criteria include MAPE, MAE, and MSE. The assessment is conducted based on two distinct observation-target lengths: 120-30 for Table 6 and 30-15 for Table 7. The results clearly demonstrate the superior performance of the proposed method across all three metrics when compared to ARIMA.

	ARIMA			**Proposed Network**		
	HR	**SBP**	**MBP**	**HR**	**SBP**	**MBP**
MAPE	7.30	10.91	10.22	3.94	6.82	6.44
MAE	5.91	13.19	7.83	3.07	8.36	4.88
MSE	77.96	324.69	110.81	27.41	164.31	61.38

Table 6: Comparison of the proposed method with ARIMA for Observation-Target Lengths of 120-30 in HR, SBP, and MBP.

	ARIMA			**Proposed Network**		
	HR	**SBP**	**MBP**	**HR**	**SBP**	**MBP**
MAPE	7.42	7.79	8.68	3.68	6.19	5.97
MAE	5.97	9.87	6.57	2.84	7.73	4.53
MSE	80.61	208.30	98.77	25.7	145.8	55.45

Table 7: Comparison of the proposed method with ARIMA for Observation-Target Lengths of 30-15 in HR, SBP, and MBP.

To assess the agreement between the deep learning model's predictions and the actual observed values for the three different vital signs of HR, MBP, and SBP, we employed Bland-Altman plots and presented the results in Figure 4. In the subplots (Figure. 4a, b, and c), the model's output for HR, MBP, and SBP are compared to actual measurements across various time windows, respectively, that shows both acceptable bias and standard deviation. In Figure 5 the deep learning model predictions are plotted against the actual values for HR, MBP, and SBP in 120-30 time window.

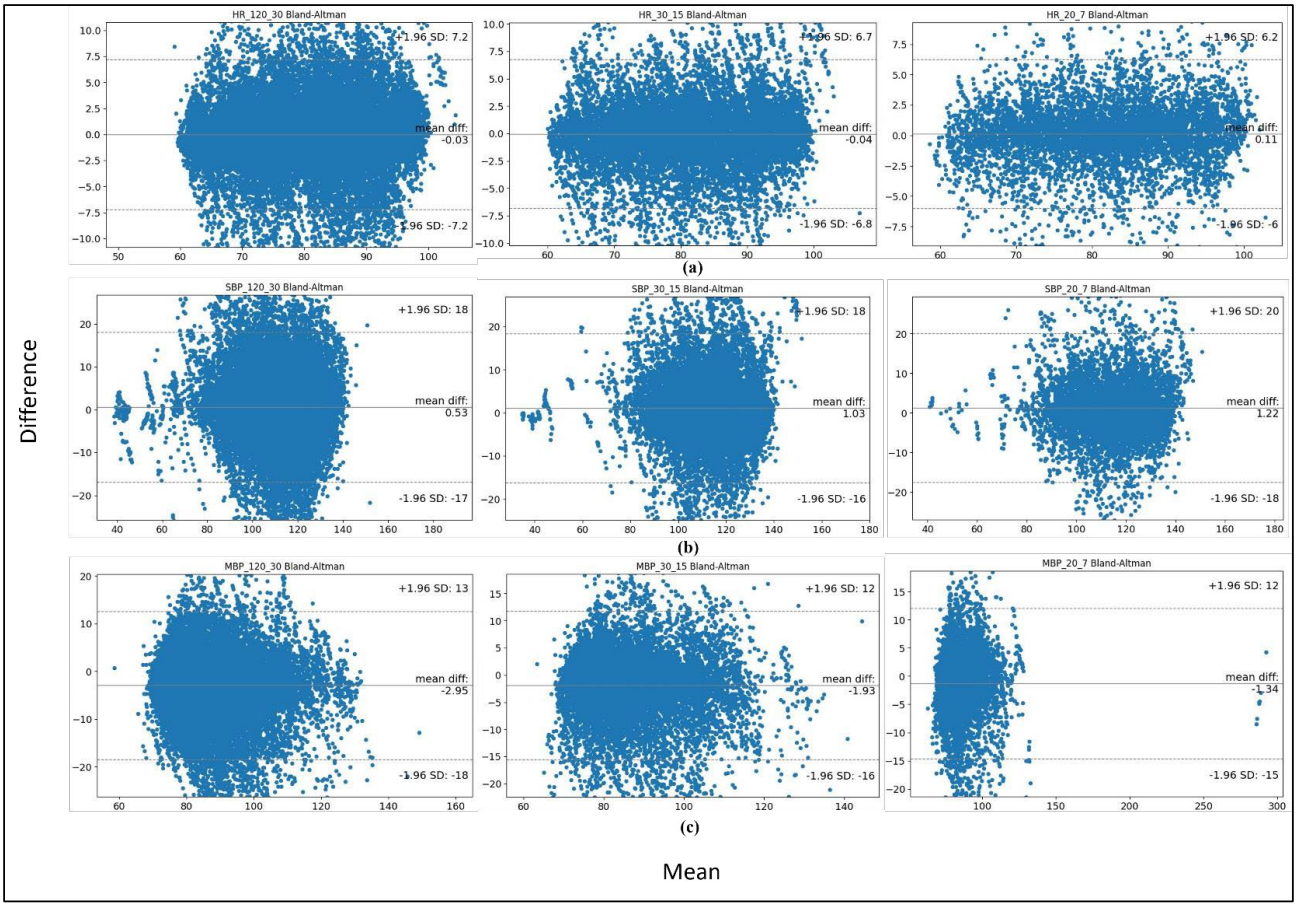

Figure 4: Bland-Altman Assessments for the Agreement between the Predicted Values and the Observed Values

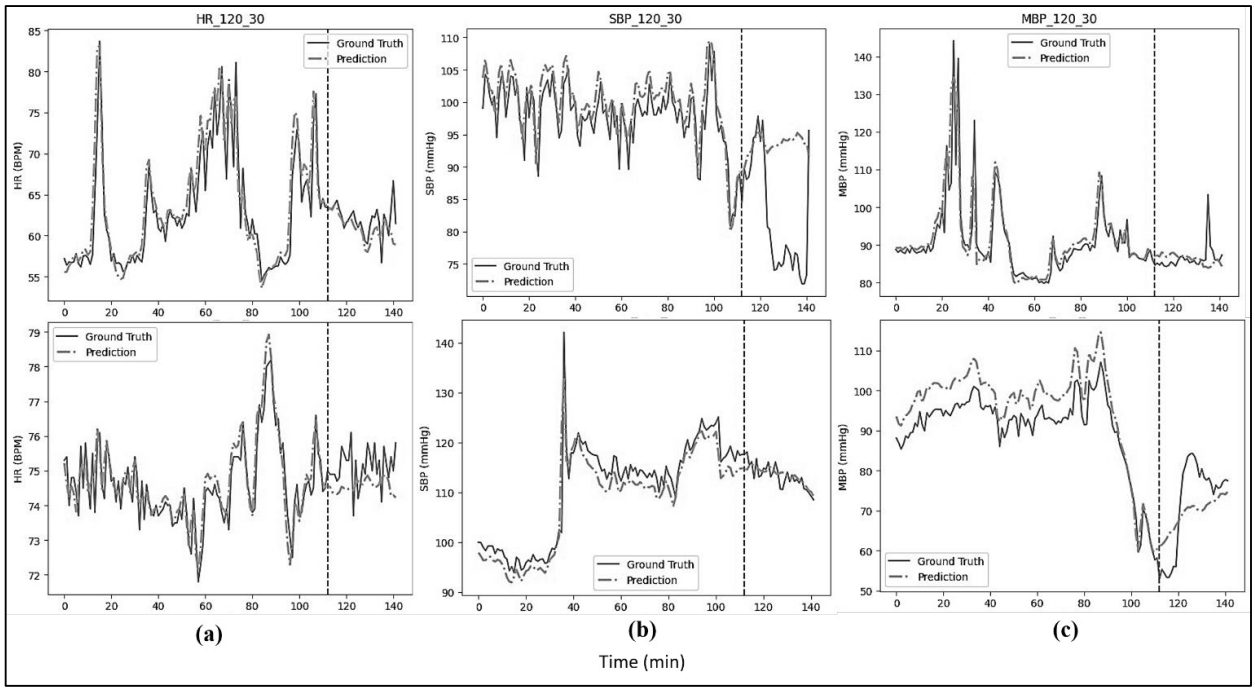

Figure 5: The Predicted Values vs the Observed Values

4. Discussion

Our study used the three HR, SBP, and MBP time-series data from the MIMIC-III database for prediction. Each time series has been made stationary in preprocessing step by using fractional derivatives and preserving most of the time series information. Noticeable improvements are visible in using fractional order in all criteria according to Table 4 and Table 5. Complementary results provided for higher order differences (0.4, 0.5, and 0.6) in Tables 6 and 7 in which the deterioration of the network's performance shown by calculating MAPE, MAE, and MSE. It

shows that preserving history of time series utilizing by fractional differencing has remarkable impact on the forecasting algorithm, however, the sophisticated part is to find out the optimum order.

In the following step to forecast time series in a multi-step manner, we exploited a deep neural network which is a combination of CNNs with skip connections and Bi-LSTM. The proposed architecture is more efficient than the normal CNNs [5, 7, and 24]. Evaluating the performance of the model based on the regression criteria reveals how much the predicted signal is similar to the actual signal. These criteria are beneficial to show the power of the model for the simulation of exact values of signal even with its fluctuation in the future which is important to make sense of events probably will occur in the specified future time for the clinicians. On the other hand, evaluating the performance of the model based on criteria calculated from the confusion matrix is essential due to classifying the predicted signal based on a threshold to alarm when an intended event is likely to happen. Although, this kind of evaluation cannot consider a main reference due to the inherent error of threshold besides the lack of global standard values. Therefore, we calculated ACC, PPV, and MCC in addition to four confusion matrix categories (True Positive, False Positive, True Negative, and False Negative). It seems that the suggested model performance with ACC, PPV, and MCC scores of more than 0.8 could be a reliable model to forecast vital signs such as HR, SBP, and MBP during surgery or in ICU [27]. However, when the target size increases to 30 min this performance decay to almost 0.7. Such prognostic performance for patients in ICUs could be considering as promising results for preventing adverse clinical events and enhancing patient care [3,5,7,8].

Furthermore, the study showed a relatively narrow difference between the maximum and the minimum values of regression criteria (MAPE, MAE, and MSE). Model performance based on regression criteria as described in the results section with details, have mostly the best values in a window size that has the longest observation length and the lowest target length i.e., 120-7. This makes sense with the theory that says the model performs better when it sees longer signals as it can find more patterns and also it is quite clear that less target window has been a more precise prediction [10]. But it is worth noting that the longer the observed length, the more cost we pay, and it is preferred to make a system that can forecast more future time based on less input length. In this regard, the proposed structure has the advantage of having a narrow range of maximum and minimum for each criterion which means it also performs well for lower cost strategies such as 60-30 observe-target window. As a comparison of the results, Liu et al. in 2019 has reported MAPE values of 7.41% and 6.17% for HR and SBP respectively for an observe-target window size of 20-7 when the results of the proposed method are 3.19% for HR and 5.61% for SBP in the same window size which indicate that our proposed method performed better [10].

As illustrated in Figure. 4, the small mean difference in all nine plots suggests minimal bias in the model's predictions, indicating reasonable accuracy. As the prediction windows get longer, the spread on the y-axis widens which means the model prediction may tolerate the acceptable range. More specificly, in Figure. 4a the deviation range is of approximately ±7 beats per minutes in HR signal forecasting for various time windows. It indicates that the model can estimate heart rate values within this error range. According to inherent noise measurment and accuracy of devices, such an error range in HR is promising. In Figure. 4b, the higher mean differences in SBP models, compared to HR models, imply a greater bias. Also, deviation increases to approximately ±20 mmHg. It showed that forecasting the blood pressure vital sign is complicated as it is the holistic complex signal derived from cardivascular system interacting with the whole body. In Figure. 4c, the mean difference in MBP models is greater than in HR models but comparable to SBP models. The deviation is narrower than the SBP models i.e. ±12 mmHg, indicating that MBP signal is well-behaved than SBP to model and forecast.

As for time series visual assemnet, Figure. 5 compares the forecasted HR, MBP, and SBP values with the actual measurements over time. The prediction line closely follows the ground truth line in Figure. 5a, indicating that the model captures the overall trend of HR. There are minor deviations, but overall, the model seems to have produced reasonabe results. In Figure. 5b, the predicted SBP shows more variation than in the HR plot, with noticeable differences at certain points. However, there is still alignment between predicted and actual values, suggesting that the model provides fair predictions. In Figure. 5c, SBP predictions are compared to actual measurements. Similar to MBP, there are deviations between the lines, especially during sharp rises or falls in blood pressure estimations.

5. Conclusion

In this paper, we proposed an effective way to directly forecast vital time series such as HR, SBP, and MBP for several minutes in advance. We make each time series stationary by fractional difference to preserve their main history. We showed that this algorithm has improved the MAPE reported by the last study. The forecasting process provides golden few minutes preventing adverse events and allows the physicians to be prepared and alerted and to intervene properly. For this purpose, we demonstrated the efficiency of the proposed method for use in the ICU to alarm adverse clinical events such as Bradycardia, Tachycardia, and Hypo-tension by evaluating criteria driven by a confusion matrix. Since the model structure is in a high interaction with the input type, it seems that our proposed model is not compatible with the multivariate input strategy and it is required some structural modification which would be considered in our future work.

References

1. ElMoaqet, H., Tilbury, D. M., & Ramachandran, S. K. (2016). Multi-step ahead predictions for critical levels in physiological time series. *IEEE transactions on cybernetics, 46*(7), 1704-1714.
2. Fossion, R., Alvarez-Millán, L. A., Miranda-Velazco, E., Garduño, F. G., Padilla, S. R. M., Zapata-Fonseca, L. I., ... & Estañol, B. (2019, April). On the role of continuous physiological monitoring and time-series analysis in medical prognosis. In *AIP Conference Proceedings* (Vol.

2090, No. 1). AIP Publishing.
3. Deng, Y., Liu, S., Wang, Z., Wang, Y., Jiang, Y., & Liu, B. (2022). Explainable time-series deep learning models for the prediction of mortality, prolonged length of stay and 30-day readmission in intensive care patients. *Frontiers in Medicine, 9,* 933037.
4. Sun, G., Matsui, T., Watai, Y., Kim, S., Kirimoto, T., Suzuki, S., & Hakozaki, Y. (2018). Vital-SCOPE: Design and Evaluation of a Smart Vital Sign Monitor for Simultaneous Measurement of Pulse Rate, Respiratory Rate, and Body Temperature for Patient Monitoring. *Journal of Sensors, 2018*(1), 4371872.
5. Masum, S., Liu, Y., & Chiverton, J. (2018). Multi-step time series forecasting of electric load using machine learning models. In *Artificial Intelligence and Soft Computing: 17th International Conference, ICAISC 2018, Zakopane, Poland, June 3-7, 2018, Proceedings, Part I 17* (pp. 148-159). Springer International Publishing.
6. Lim, B., & Zohren, S. (2021). Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A, 379*(2194), 20200209.
7. Masum, S., Chiverton, J. P., Liu, Y., & Vuksanovic, B. (2019). Investigation of machine learning techniques in forecasting of blood pressure time series data. In *Artificial Intelligence XXXVI: 39th SGAI International Conference on Artificial Intelligence, AI 2019, Cambridge, UK, December 17–19, 2019, Proceedings 39* (pp. 269-282). Springer International Publishing.
8. Narayan Shukla, S., & Marlin, B. M. (2020). Integrating Physiological Time Series and Clinical Notes with Deep Learning for Improved ICU Mortality Prediction. *arXiv e-prints, arXiv-2003.*
9. Bontempi, G., Ben Taieb, S., & Le Borgne, Y. A. (2013). Machine learning strategies for time series forecasting. *Business Intelligence: Second European Summer School, eBISS 2012, Brussels, Belgium, July 15-21, 2012, Tutorial Lectures 2*, 62-77.
10. Liu, S., Yao, J., & Motani, M. (2019, November). Early prediction of vital signs using generative boosting via LSTM networks. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 437-444). IEEE.
11. Youssef Ali Amer, A., Wouters, F., Vranken, J., de Korte-de Boer, D., Smit-Fun, V., Duflot, P., ... & Vanrumste, B. (2020). Vital signs prediction and early warning score calculation based on continuous monitoring of hospitalised patients using wearable technology. *Sensors, 20*(22), 6593.
12. Watkinson, P. J., Barber, V. S., Price, J. D., Hann, A., Tarassenko, L., & Young, J. D. (2006). A randomised controlled trial of the effect of continuous electronic physiological monitoring on the adverse event rate in high risk medical and surgical patients. *Anaesthesia, 61*(11), 1031-1039.
13. Watkinson, P. J., & Tarassenko, L. (2012). Current and emerging approaches to address failure-to-rescue. *The Journal of the American Society of Anesthesiologists, 116*(5), 1158-1159.
14. Hyndman, R. J. (2018). *Forecasting: principles and practice*. OTexts.
15. Lynn, L. A., & Curry, J. P. (2011). Patterns of unexpected in-hospital deaths: a root cause analysis. *Patient safety in surgery, 5,* 1-25.
16. Kristoufek, L. (2014). Measuring correlations between non-stationary series with DCCA coefficient. *Physica A: Statistical Mechanics and its Applications, 402*, 291-298.
17. Jebb, A. T., Tay, L., Wang, W., & Huang, Q. (2015). Time series analysis for psychological research: examining and forecasting change. *Frontiers in psychology, 6,* 727.
18. Said, S. E., & Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika, 71*(3), 599-607.
19. Kwiatkowski, D., Phillips, P. C., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?. *Journal of econometrics, 54*(1-3), 159-178.
20. Chan, K.-S. & Cryer, J. D. (2008). Time series analysis with applications *in R, Springer.*
21. Lopez de Prado, M. (2018). Advances in financial machine learning (chapter 1). *Advances in Financial Machine Learning, Wiley, 1st Edition* (2018).
22. Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., ... & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific data, 3*(1), 1-9.
23. Hosking, J. R. M. 1981. Fractional differencing. *Biometrika*, 68, 165-176.
24. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18* (pp. 234-241). Springer International Publishing.