IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 9 February 2023, accepted 10 March 2023, date of publication 15 March 2023, date of current version 27 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3257406

RESEARCH ARTICLE

Rapid Response System Based on Graph Attention Network for Predicting In-Hospital Clinical Deterioration

THANH-CONG DO[1], HYUNG-JEONG YANG[1], (Member, IEEE), GUEE-SANG LEE[1], (Member, IEEE), SOO-HYUNG KIM[1], (Member, IEEE), AND BO-GUN KHO[2]

[1]Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju 61186, South Korea
[2]Department of Internal Medicine, Chonnam National University Hospital, Gwangju 61469, South Korea

Corresponding author: Hyung-Jeong Yang (hjyang@jnu.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.2023-00208397). This study was supported by a grant (HCRI 23038) Chonnam National University Hwasun Hospital Institute for Biomedical Science.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Independent Institutional Review Board (IRB) of Chonnam National University Hospital (CNUH).

ABSTRACT In-hospital clinical deterioration is a major worldwide healthcare burden in the intensive care units (ICUs), as it requires rapid intervention. Rapid response systems (RRSs) are widely used in many hospitals for the early detection of clinical deterioration to prevent cardiac arrest. Recently, with the increasing use of deep learning (DL) and electronic health records (EHR), many DL models have been developed for the intensive care domain, such as prediction of cardiac arrest, sepsis, or transferring to ICU. However, most existing methods do not explicitly learn the structure of multivariate time-series data, and this leads to high false-alarm rates and low sensitivity. In this research, we propose a novel DL-based framework that interpolates high-dimensional sequential data. Our approach combines two graph neural networks with an attention mechanism to learn the complex dependencies among multivariate time series. The experiments were conducted on two datasets: a private clinical dataset collected from Chonnam National University Hospital (CNUH) and a public dataset from the University of Virginia (UV). The experimental results show the potential performance of our model compared to some other related research.

INDEX TERMS Attention mechanism, cardiac arrest, clinical deterioration, deep learning, graph neural network, rapid response system.

I. INTRODUCTION

There has recently been an increase in the number of patients and in the overcrowding of emergency departments, which are causing adverse treatment outcomes. In-hospital events that represent clinical deterioration, such as cardiac arrest, are considered to be major burdens for most intensive care units (ICUs) and affect patient mortality [1], [2], [3]. Previous studies indicate that 80% of cardiac arrests show abnormalities of vital signs 8–24 hours before the event [4], [5], [6]. However, early detection of cardiac arrests faces many difficulties because a single vital sign usually does not accurately predict patient prognosis [7]. In fact, the survival discharge rate of cardiac arrest patients is less than 20% [8], [9].

The rapid respond system (RRS) has been introduced in many hospitals as a clinical supporting tool to prevent in-hospital emergencies such as cardiac arrest by proactively intervening in patients who are clinically deteriorating [10], [11]. Some traditional RRSs use the Modified Early Warning Score (MEWS), which calculates a weighted score for each

The associate editor coordinating the review of this manuscript and approving it for publication was Marco Giannelli.

vital sign and then finds patients with events based on the sum of the scores [12]. However, this method has limitations of low sensitivity and high false-alarm rates [12], [13]. The National Early Warning Score was recently introduced and shows outstanding predictive performance over previous systems in terms of cardiac arrest and death [14].

In recent years, the increased access to electronic health records (EHR) has motivated the development of artificial intelligence models to predict clinical events in the ICU [15], [16], [17]. There have been several attempts to use artificial neural networks (ANN) to detect in-hospital events, and these methods have attracted considerable attention [18], [19], [20]. ANNs have shown better performance than traditional scoring systems, which predict cardiac arrest earlier and with better accuracy [20]. However, these deep learning (DL)-based models usually lack interpretability, as they are viewed as a black box that does not provide any insight about the features learned from the data [21]. There are also several approaches that utilize recurrent neural networks (RNNs) to capture sequence features in health records data [21], [22]. These approaches require regularly spaced data points. However, the EHR is usually sparse, noisy, and incomplete. In addition, most of these methods do not present multivariate correlations explicitly, which may lead to missing some inter-relationships in multivariate time-series data.

Recently, graph neural networks (GNNs) have shown success in modeling complex patterns in graph-structured data. Some recent studies have applied graph-based attention mechanisms to incorporate medical knowledge from EHR into DL models [23], [24]. In fact, these graph-based attention models are compatible with discrete data but not with sequential data [25]. In this paper, we propose an end-to-end DL-based RSS that tackles the limitations of previous methods for detecting the clinical deterioration of in-hospital cardiac arrest patients. Our proposed architecture is based on double graph attention networks (DGATs) and includes two GNNs that learn the graphs of the dependencies between features and the temporal relations between time points at the same time. In DGATs, two graph attention networks [26] are trained in parallel, namely the feature-based graph attention network that captures the correlations among multiple input features and the time-based graph attention network that discovers the dependencies between different time points.

The primary contributions of this paper are as follows:

- We propose a DL-based framework to perform clinical deterioration prediction tasks. The proposed DGAT takes advantage of the complex inter-relationships in multivariate time series by learning their correlations in both the time and feature domains. To the best of our knowledge, this is the first study that applies the graph attention models to the problem of clinical deterioration prediction.
- In the medical context, interpretability is considered to be one of the key components of clinical utility [27]. Our model presents good interpretability by utilizing an attention mechanism that emphasizes those parts of a multivariate time series that are most relevant to the output target.
- Our proposed system shows better performance on two clinical datasets than the baseline approaches. We also conduct experiments using a cross-validation strategy to present the generalization of the model.

The rest of the paper is organized as follows: Section II begins with some description about related works on RRS models for clinical deterioration. Section III presents a detailed architecture of our proposed system. Dataset information and the experiment results are reported in Section IV. Finally, Section V presents the conclusion and some future work for our study.

II. RELATED WORKS

In this section, we review some recent DL-based applications for predicting in-hospital clinical deterioration. RNN-based models have recently shown great potential for the medical and healthcare domain because of their capability to capture sequential patterns in time series [28]. The study in [22] developed prediction models for three events: sepsis, acute kidney injury (AKI), and death. Each model utilized bidirectional long short-term memory (Bi-LSTM) architecture to design binary classification models. The experimental results showed superior performances by the Bi-LSTM models relative to some other machine learning methods. A deep early warning system (DEWS) was proposed in [18] to predict in-hospital cardiac arrest between 0.5 and 24 hours before the event. The DEWS consisted of three RNNs with an LSTM unit to solve the long-term dependency problem. The model was tested on two clinical datasets, and it outperformed MEWS, logistic regression, and random forest on all metrics. Another study [20] used the architecture with three ANNs for the early detection of patients at risk for cardiac arrest. The first network was a multilayer perceptron (MLP) that used baseline variables (age, sex, initial vital signs, ...). The second network consisted of LSTM layers stacked with MLP layers that used 12 recent updates of vital signs within the previous 6 hours as input. The third model was a hybrid LSTM and MLP model, where the sequence data and static data were processed separately first, and then fused to generate the output. All of the above methods usually outperform traditional scoring systems. However, these approaches do not have good interpretability and may not be capable of capturing multivariate correlations explicitly. These traditional DL methods with RNN-based models only consider the sequential patterns in time-series. Some other information such as correlations among features and dependencies among time points are also necessary.

Some recent studies have applied an attention mechanism to their models, which improves the interpretability as it highlights those parts of the input data that contribute the most to the model's decision. The study in [21] proposed a deep interpretable early warning system based on an encoder

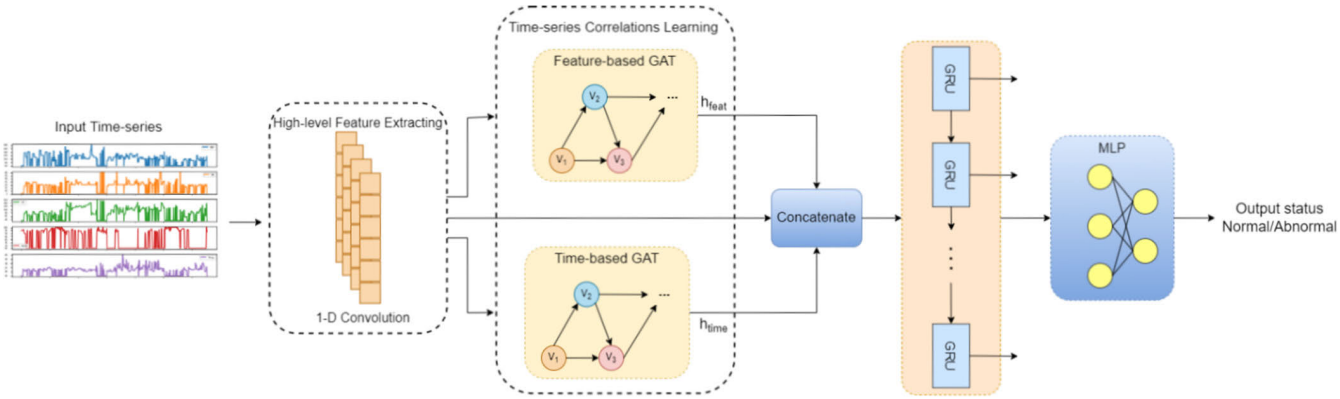

FIGURE 1. Overall architecture of the proposed method for clinical deterioration prediction. The input time series data are fed to a 1-D convolutional layer to extract the high-level features. These features then go through two graph attention networks which are trained parallelly. The output of the graph networks is concatenated with the features extracted from the convolution layer and go through a gated recurrent unit (GRU) layer followed by the MLP layers for final prediction.

with a Bi-LSTM architecture and an attention block. This model achieved state-of-the-art performance even when using a limited set of features. A DL-based framework that utilizes an MLP along with an attention mechanism was proposed in [29]. This approach shows competitive performance relative to MLP, linear regression, and stacked denoising autoencoder (SDAE) on the task of clinical prediction for heart failure patients. The two attention-based methods above successfully captured the trends of input vital signs that most contribute to the model behaviors, but they did not explicitly learn which features were related to one another, which could lead to missing some inter-relationships among multivariate time series.

III. PROPOSED METHOD

This section presents our proposed DL-based method for in-hospital clinical deterioration prediction. In this paper, we define the task of detecting clinical deterioration as a binary classification problem, as was done in recent studies [18], [20], [21]. Thus, we define our input time series as $x_t \in R^{n \times m}$, where n is the length of the input sequence time points before time t, and m is the number of clinical features. In this work, we generate input of a fixed length $n \in \{8, 12, 24\}$ by sliding a time window with a sliding step $k = 1$. The output of the model is a vector $y \in R^n$, where $y_t \in \{0, 1\}$ denotes the normal or abnormal status at time point t. The abnormal status is defined as the time at which clinicians recognize abnormal changes in patients' measurements that would likely lead to an event. In other words, the event would occur after a series of abnormal time points. Because this work is defined as an anomaly detection problem, our system can detect clinical events earlier, which helps medical staff to make the necessary interventions before the occurrence of the events.

A. METHOD OVERVIEW

The overall architecture of our proposed method is shown in Fig. 1. After data preprocessing step, the input time series go through a 1-D convolution layer to extract the high-level features. A recent study [30] indicates that the convolutional process can capture the local features of time series. The outputs of the 1-D convolution layer are fed to the DGAT, which includes the feature-based graph attention network and the time-based attention network. These networks are trained in parallel; we concatenate their outputs with the extracted features from a 1-D convolution layer, and then feed them into a gated recurrent unit (GRU) layer to capture the sequential information from input time series. Finally, the latent features from the GRU layer are fed to the MLP to obtain the final prediction result.

B. DATA PREPROCESSING

For the task of clinical deterioration prediction, we use two types of input features: vital signs, which are measured every hour, and laboratory tests, which are recorded discontinuously. In the medical context, most healthcare datasets face the problem of missing values. In this work, for the vital signs, we apply the carry-forward method for missing data imputation if data existed before the missing time points. If not, we fill the missing data with the mean value of the non-missing values for each patient. The mean is a measure of central tendency and can provide a good estimate of the typical value in the dataset. For the lab tests, because each patient has only a few records, we duplicate these values for every time point between the two records.

Class imbalance is one of the main problems in medical anomaly detection because most of the data are labeled with normal status [18], [31]. On such imbalanced datasets, DL models usually perform poorly for the minority class (positive class). Several time series augmentation techniques were mentioned in [32] that generate more samples of minority class to control the imbalance problem. In this work, we apply two data augmentation methods as in [32], which are cropping and label expansion. In addition to augmentation, we also apply data normalization to enhance the robustness of the model. In this work, we normalize the input time

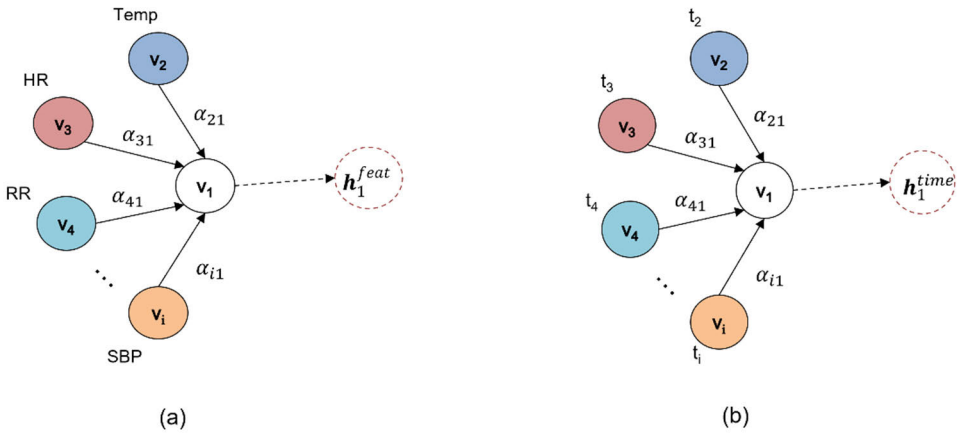

FIGURE 2. Description of the two graph networks: (a) Feature-based graph network; (b) Time-based graph network. In the feature-based graph network, each node vector v represents a certain feature meanwhile in time-based graph network, each node represents a feature vector of a certain time point in the input sequence. In both graph networks, the edges denote the correlations between two nodes.

series using standard normalization as in (1):

$$\bar{x} = \frac{x - \mu}{\sigma} \tag{1}$$

where x is the input time series, and μ and σ are the mean and standard deviation of the input data.

C. GRAPH ATTENTION NETWORK

In this research, we implement two complete graph networks, namely the feature-based graph attention network and the time-based graph attention network.

1) TIME-BASED GRAPH ATTENTION NETWORK

This network captures temporal relationships in time series. Each node represents one time point, and all time points in a sliding window create a complete graph. The time-based graph includes m nodes $\{v_1, v_2, \ldots, v_m\}$, where m is the number of time points in the current sliding window; v_i is the vector representation of each node with $v_i = \{v_{i,t} | t \in [0, n)\}$, where n is the number of multivariate features. The output of the graph is a matrix of size $m \times n$.

2) FEATURE-BASED GRAPH ATTENTION NETWORK

This network learns the correlations between multivariate time series. We consider the whole input time series as a complete graph where each node represents a certain feature, and each edge denotes the dependencies between a pair of nodes. Therefore, the model can explicitly learn the inter-relationships between the multivariate time series features. The output of the feature-based graph attention network is a matrix having the shape of $n \times m$. Finally, we concatenate the output of both graph attention networks with the output from the 1-D convolutional layer to keep the original sequential information.

For both the time-based graph network and the feature-based graph network, we utilize the graph attention mechanism [26] to fuse the node information. The input of the attention layer is a set of vectors representing every node in the graph network $V = \{\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_N\}$, where N is the number of nodes. We first compute the attention coefficients between each pair of nodes as in (2):

$$e_{ij} = LeakyReLU(w^T(W\vec{v}_i \oplus W\vec{v}_j)) \tag{2}$$

where j is one of the adjacent nodes of node i, w and W are weight parameters, $\oplus$ denotes the concatenation of two node representations, and $LeakyReLU$ is a nonlinear activation function [33]. Then, we normalize the coefficients across all choices of node j by using the SoftMax function:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{K} \exp(e_{ik})} \tag{3}$$

where K is the number of adjacent nodes for node i. We compute the final representation for each node by combining the attention weight with the corresponding feature vector:

$$h_i = \sigma(\sum_{j=1}^{K} \alpha_{ij} v_j) \tag{4}$$

Here, σ denotes the sigmoid activation function, and h_i is the output of node i, which has the same shape as input vector v_i. The two graph networks in DGAT are illustrated in Fig. 2. By utilizing two graph attention networks, our proposed DGAT can explicitly learn the correlations between different features and time points.

IV. EXPERIMENTS

In the experimental section, we evaluate our proposed DGAT for clinical deterioration prediction on two healthcare datasets.

A. DATASET

This research was approved by the Independent Institutional Review Board (IRB) of Chonnam National University Hospital (CNUH). The primary database used for the experiment

TABLE 1. CNUH data characteristics. 5 vital signs and 18 laboratory tests are input features for proposed architecture. The data is highly imbalance with only 1.5% of the patients with events.

Characteristic	Value
Number of patients	2,615
Age (years), mean (95% CI)	65.66 (65.14–66.17)
Gender (% male)	36.8%
Number of vital signs	5
Number of laboratory tests	18
Patient with events (%)	1.5%

TABLE 2. UV data characteristics. The data is imbalance with only 4.5% of the patients with events. 1000 non-event samples are selected.

Characteristic	Value
Number of patients	8,105
Age (years), mean (95% CI)	63.74 (63.42–64.07)
Race (% white/black/other)	80.9%/16.9%/2.2%
Number of vital signs	7
Number of laboratory tests	24
Number of ECG monitoring	15
Patient with events (%)	4.5%

contains licensed patient data provided by Chonnam National University Hwasun Hospital's doctors. In addition, we validated our model on a public dataset from the University of Virginia to ensure the generalization of the model.

The CNUH dataset contains information from 2,615 patients from two hospitals. The main types of features include demographic, vital signs, and laboratory tests. There are five vital signs: body temperature (BT), systolic blood pressure (SBP), heart rate (HR), respiration rate (RR), and oxygen saturation (SaO2), which are measured every hour. The laboratory tests contain information collected from patients' blood, including alanine transaminase (ALT), blood urea nitrogen (BUN), aspartate aminotransferase (AST), white blood cell count (WBC count), C-reactive protein (CPR), albumin, lactate, total protein, platelet, hemoglobin (Hgb), alkaline phosphatase, total calcium, total bilirubin, creatinine, glucose, sodium, chloride, potassium. Some characteristics of the dataset are shown in Table 1.

The dataset from the University of Virginia was collected within 63 years from 8,150 patients who were admitted to a tertiary care academic medical center. It also contains vital signs and lab test features, similar to the CNUH dataset. In addition, it contains 15 features related to cardiorespiratory dynamics that were measured from continuous ECG heart monitoring, which tracks the heart rate and electrical activities of patients. Table 2 represents some characteristics of the UV dataset. Both datasets contain the clinical status normal or abnormal, which is determined by medical staff every hour. The abnormal status is specified when medical doctors recognize abnormal changes in patients' measurements that are likely to lead to an event. The first time point with an abnormal label is considered to be the detection time, and the event is the last abnormal time point for each patient. The patients with events have both normal and abnormal time points, whereas the non-event patients have only normal time points.

TABLE 3. Number of samples in each fold for two datasets. 5-fold cross validation is applied for both datasets. In each fold, 20% of the patients are used for validation.

		Train		Test	
Dataset	Fold	Normal	Abnormal	Normal	Abnormal
CNUH	1	238,161	799	59,533	208
	2	238,164	797	59,530	210
	3	238,161	800	59,533	207
	4	238,147	814	59,547	193
	5	238,143	818	59,551	189
UV	1	325,572	18,340	81,349	4,629
	2	325,499	18,413	81,422	4,556
	3	325,508	18,404	81,413	4,565
	4	325,580	18,332	81,341	4,637
	5	325,525	18,387	81,396	4,582

The imbalance problem exists in both datasets. For the CNUH dataset, there are only 41 patients with events (1.5%). In the UV dataset, the number of event patients is 367 (4.5%), and there are 7,738 non-event patients (95.4%). Because of the large amount of non-event data, we randomly select 1000 normal samples from the UV dataset for our experiments to control the imbalance problem and mitigate the computational burden.

B. TRAINING AND IMPLEMENTATION SETTING

In this work, we apply a 1-D convolution with a kernel size of 7 to the input time series to extract the high-level features. The extracted features then are processed by two graph attention networks to learn the multivariate time series correlations. The output of the graph networks is concatenated with the high-level features from the 1-D CNN and then goes through a GRU layer with 128 hidden dimensions. The output of the GRU is fed to two 128-neuron fully connected (FC) layers, each followed by ReLU activation and a dropout rate of 0.2, before the last output layer to predict the binary output value of normal or abnormal. The experiments are carried out in PyTorch using the Adam optimizer and a learning rate of 0.0001. We also apply five-fold cross validation for training to present the generalization of the model. Table 3 shows the details for the number of time points over the five folds on both datasets. We use the original CNUH dataset with a total of 317,006 time points but only 1,042 abnormal time points. We exclude 6,738 non-event patients from the UV dataset, and the data used for this study contains 493,333 time points with 23,881 abnormal samples.

TABLE 4. Experimental results for CNUH dataset. The experiments are performed with different input sequence lengths of 8 hours, 12 hours, and 24 hours. The results shown in the table are the mean of 5 folds with a standard deviation.

Input Window Size	Model	Metrics (Mean of 5 folds ± Std)	
		AUROC	AUPRC
8 hours	LSTM	0.827 ± 0.07	0.653 ± 0.08
	Bi-LSTM	0.873 ± 0.09	0.700 ± 0.06
	Bi-LSMT + Attention	0.918 ± 0.06	0.822 ± 0.04
	DGAT	**0.936 ± 0.06**	**0.842 ± 0.05**
12 hours	LSTM	0.822 ± 0.07	0.697 ± 0.07
	Bi-LSTM	0.884 ± 0.04	0.782 ± 0.06
	Bi-LSMT + Attention	0.932 ± 0.06	0.845 ± 0.04
	DGAT	**0.955 ± 0.07**	**0.879 ± 0.06**
24 hours	LSTM	0.812 ± 0.08	0.672 ± 0.07
	Bi-LSTM	0.860 ±0.06	0.694 ± 0.05
	Bi-LSMT + Attention	0.875 ± 0.05	0.767 ± 0.03
	DGAT	**0.905 ± 0.07**	**0.822 ± 0.05**

TABLE 5. Experimental results for UV dataset. The experiments are performed with different input sequence lengths of 8 hours, 12 hours, and 24 hours. The results shown in the table are the mean of 5 folds with a standard deviation.

Input Window Size	Model	Metrics (Mean of 5 folds ± Std)	
		AUROC	AUPRC
8 hours	LSTM	0.799 ± 0.07	0.698 ± 0.05
	Bi-LSTM	0.845 ± 0.07	0.769 ± 0.05
	Bi-LSMT + Attention	0.880 ± 0.05	0.802 ± 0.06
	DGAT	**0.924 ± 0.04**	**0.863 ± 0.05**
12 hours	LSTM	0.853 ± 0.09	0.704 ± 0.03
	Bi-LSTM	0.897 ± 0.05	0.790 ± 0.07
	Bi-LSMT + Attention	0.923 ± 0.06	0.833 ± 0.06
	DGAT	**0.949 ± 0.06**	**0.889 ± 0.03**
24 hours	LSTM	0.846 ± 0.08	0.695 ± 0.05
	Bi-LSTM	0.898 ± 0.06	0.822 ± 0.07
	Bi-LSMT + Attention	0.900 ± 0.05	0.816 ± 0.08
	DGAT	**0.937 ± 0.06**	**0.865 ± 0.03**

To solve the imbalance problem, we utilize focal loss as the final objective function for the prediction task. It is an extension of the common cross-entropy loss with a scaling factor that can down-weight the effect of samples in the majority class and focus the model on hard samples. The focal loss can be expressed as follows:

$$FL\left(\rho_t\right) = -\alpha_t \left(1-\rho_t\right)^{\gamma} \log(\rho_t) \tag{5}$$

where ρ_t is a predicted probability, and $(1-\rho_t)^{\gamma}$ is the scaling factor. With $\gamma = 0$, the focal loss is equivalent to the cross-entropy loss. α_t is the weighting parameter that represents the inverse class frequency.

C. EXPERIMENTAL RESULTS

In this paper, we present an evaluation of our proposed DGAT and compare the performance with some recent research

TABLE 6. Experimental results for CNUH dataset. We evaluate the performance of each graph network by comparing the results when using only each graph network with using both of them. The experiments are performed with different input sequence lengths of 8 hours, 12 hours, and 24 hours. The results shown in the table are the mean of 5 folds with a standard deviation.

Input Window Size	Model	Metrics (Mean of 5 folds ± Std)	
		AUROC	AUPRC
8 hours	GAT (feature-based)	0.892 ± 0.05	0.821 ± 0.07
	GAT (time-based)	0.796 ± 0.07	0.711 ± 0.04
	DGAT	**0.936 ± 0.06**	**0.842 ± 0.05**
12 hours	GAT (feature-based)	0.933 ± 0.06	0.844 ± 0.08
	GAT (time-based)	0.865 ± 0.03	0.792 ± 0.05
	DGAT	**0.955 ± 0.07**	**0.879 ± 0.06**
24 hours	GAT (feature-based)	0.876 ± 0.05	0.813 ± 0.06
	GAT (time-based)	0.826 ± 0.09	0.804 ± 0.05
	DGAT	**0.905 ± 0.07**	**0.822 ± 0.05**

TABLE 7. Experimental results for UV dataset. We evaluate the performance of each graph network by comparing the results when using only each graph network with using both of them. The experiments are performed with different input sequence lengths of 8 hours, 12 hours, and 24 hours. The results shown in the table are the mean of 5 folds with a standard deviation.

Input Window Size	Model	Metrics (Mean of 5 folds ± Std)	
		AUROC	AUPRC
8 hours	GAT (feature-based)	0.913 ± 0.06	0.809 ± 0.04
	GAT (time-based)	0.825 ± 0.08	0.694 ± 0.06
	DGAT	**0.924 ± 0.04**	**0.863 ± 0.05**
12 hours	GAT (feature-based)	0.934 ± 0.05	0.871 ± 0.04
	GAT (time-based)	0.857 ± 0.07	0.804 ± 0.06
	DGAT	**0.949 ± 0.06**	**0.889 ± 0.03**
24 hours	GAT (feature-based)	0.795 ± 0.08	0.774 ± 0.06
	GAT (time-based)	0.914 ± 0.07	0.823 ± 0.05
	DGAT	**0.937 ± 0.06**	**0.865 ± 0.03**

about clinical deterioration prediction. The first one is an LSTM-based DEWS in [18] that consists of three RNN networks with LSTM cells. The second model is based on a Bi-LSTM architecture that is proposed in [22]. The last model that we mention in our experiment is a Bi-LSTM model with an attention mechanism (ABiLSTM) in [21]. As we describe in previous sections, we perform the experiment with input window sizes $n \in \{8, 12, 24\}$ and sliding step $k = 1$. The two-evaluation metrics that we use to estimate the performance of the models are the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC), which represent well the minority class in an imbalance dataset [34].

Table 4 shows the performance of our model on the CNUH dataset compared with other approaches having different lengths of input sequences. It is obvious that the proposed method outperforms the others on both metrics with an AUROC of 0.955 and an AUPRC of 0.871. The table also indicates that the input sequence length of 12 hours achieves the best results, whereas the input sequence of 24 hours generally shows a poor performance. The same trend is observed in Table 5, which shows the experimental results on the UV dataset. A window time of 12 hours also achieves the best outcome, with an AUROC of 0.978 and an AUPRC of 0.873. It can be inferred from the experiment results that the input window time length of 24 hours is too long for the model

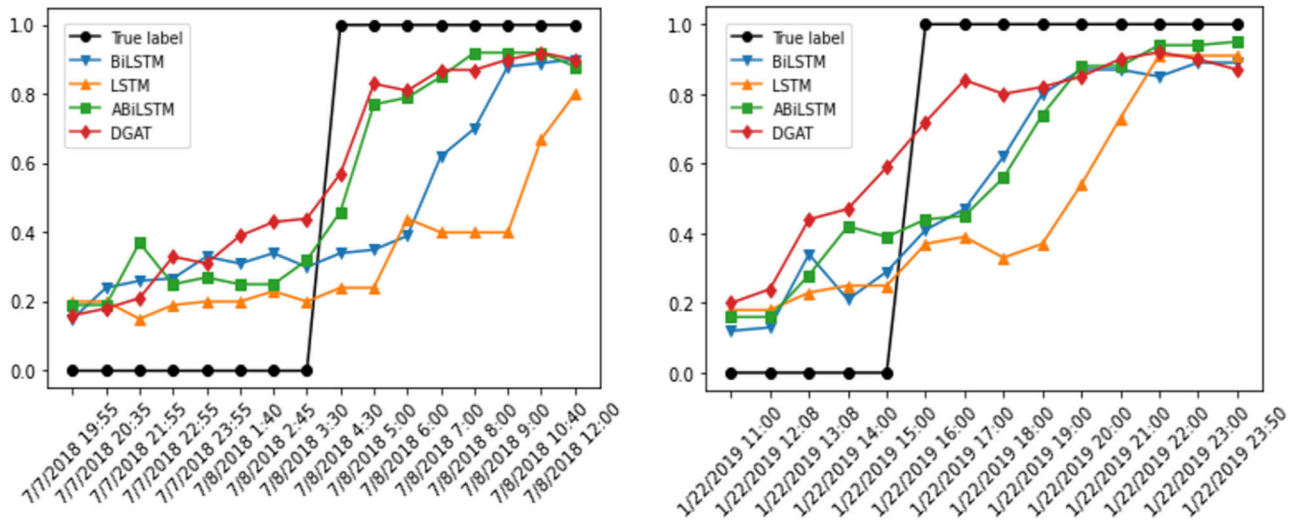

FIGURE 3. Two prediction result samples in CNUH dataset. The y-axis represents the probability of being abnormal, the x-axis represents the time points. The proposed method (green line) usually returns high probabilities for abnormal cases with lower false alarm rates.

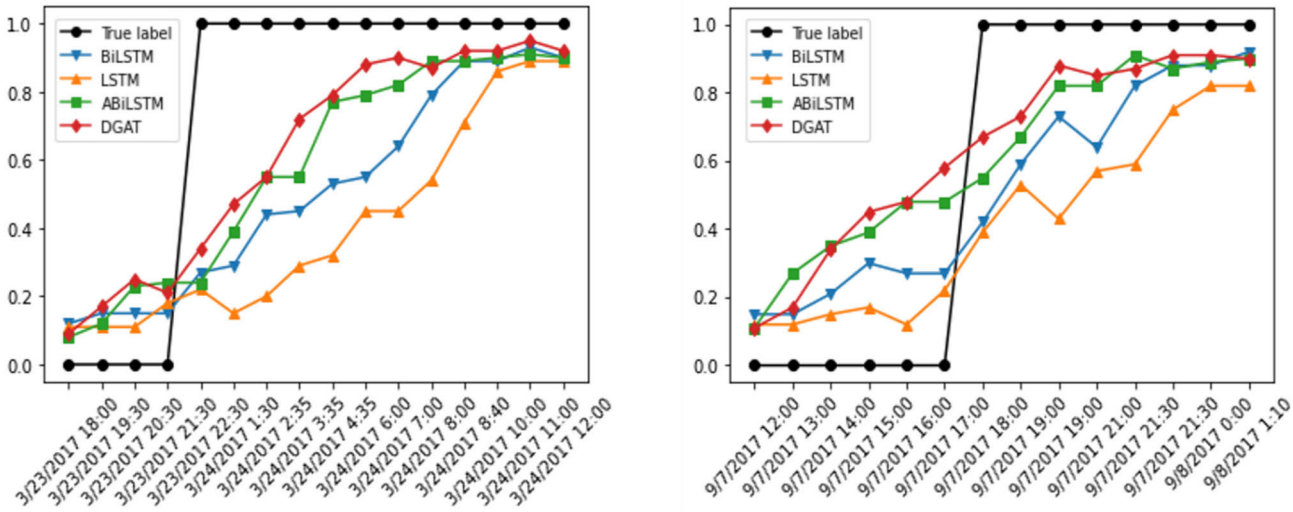

FIGURE 4. Two prediction result samples in UV dataset. The y-axis represents the probability of being abnormal, the x-axis represents the time points. The proposed method (green line) usually returns high probabilities for abnormal cases with lower false alarm rates.

to learn the time series correlations effectively, whereas the window size of 8 hours is too short to capture all the necessary information.

We also evaluate the contribution of the two graph networks in DGAT. Table 6 and 7 show the results of a comparison for only the time-based graph network or the feature-based graph network with the DGAT on both datasets. The results show that DGAT achieves better performance than using only one of the graph networks. It is obvious that using only the time-based GAT shows the worst performance with an AUROC of 0.865 and an AUPRC of 0.792. The feature-based GAT performs much better with the AUROC of 0.933 and the AUPRC of 0.844. However, combining both feature-based and time-based GATs achieve the highest results with the AUROC of 0.955 and the AUPRC of 0.879. Therefore, it can be concluded that exploring the multivariate time series feature correlations and the time domain dependencies are both necessary and significant for the prediction of clinical deterioration.

We also visualize some prediction results for samples on both datasets in Fig 3 and 4. With a classification threshold of 0.5, it is shown that the proposed DGAT performs better than the other models, and it generally returns high probabilities for abnormal time points. The figures also indicate the stability of our approach at low late alarm rates meanwhile other methods usually capture the positive status later.

It is obvious that the reason for the outstanding performance of DGAT is the contribution of two graph networks. The feature-based graph attention network explores the inter-relationships between features properly. Whenever

there is an abnormal change among features, the attention mechanism captures the abnormal correlations to speculate that an incident has occurred with the patient. Therefore, it can deal with many complex circumstances in multivariate time series. Although a GRU layer is applied to capture sequential patterns, the time-based graph attention network is also necessary for the final prediction. The main difference between the time-based graph network and the GRU is that it models the dependencies between a pair of time points directly even if they are not adjacent. Thus, some long-term correlations between time points can be learned explicitly.

V. CONCLUSION

In this paper, we propose a DL-based framework for clinical deterioration prediction using graph attention networks. Our proposed architecture includes a 1-D CNN layer to extract the high-level features, a feature-based graph attention network that explores the inter-relationships between time series features, a time-based graph attention network that can model the long-term dependencies among time points, and a GRU layer to capture the sequential patterns of time series. By combining information from different sources, our approach models the multivariate correlations explicitly and avoids missing some inter-relationship information. The attention mechanism in the graph networks also enhances the interpretability of our model by emphasizing the parts of multivariate time series that are most relevant to the final prediction. Our proposed framework was tested on two datasets: a public dataset from CNUH and a private dataset from UV. DGAT achieved the best performance on both datasets compared with the other state-of-the-art models for clinical deterioration prediction.

In conclusion, we successfully applied the proposed DGAT model to the task of clinical deterioration forecasting by defining it as an abnormal detection problem. In that way, our approach detects early the abnormal changes in patients' measurements to prevent the occurrence of events. Our proposed framework showed superior performance compared to other methods because of the contribution of two graph attention networks that explore the complex dependencies among multivariate time series. However, forecasting the time of the event is also one of the tasks that requires considerable attention from medical doctors. For future work, we plan to continue developing the current framework for application to an event detection problem to meet the requirements of the clinicians.

REFERENCES

[1] B. C. Sun, "Effect of emergency department crowding on outcomes of admitted patients," *Ann. Emerg. Med.*, vol. 61, pp. 605–611, Jun. 2013.

[2] E. C. Stecker, K. Reinier, E. Marijon, K. Narayanan, C. Teodorescu, A. Uy-Evanado, K. Gunson, J. Jui, and S. S. Chugh, "Public health burden of sudden cardiac death in the United States," *Circulat., Arrhythmia Electrophysiol.*, vol. 7, no. 2, pp. 212–217, Apr. 2014.

[3] R. Graham, M. A. McCoy, and A. M. Schultz, "Current status and future directions; board on health sciences policy; institute of medicine," in *Strategies to Improve Cardiac Arrest Survival: A Time to Act*. Washington, DC, USA: National Academic, 2015.

[4] T. J. Hodgetts, G. Kenward, I. G. Vlachonikolis, S. Payne, and N. Castle, "The identification of risk factors for cardiac arrest and formulation of activation criteria to alert a medical emergency team," *Resuscitation*, vol. 54, no. 2, pp. 125–131, Aug. 2002.

[5] V. M. Nadkarni, "First documented rhythm and clinical outcome from in-hospital cardiac arrest among children and adults," *J. Amer. Med. Assoc.*, vol. 295, no. 1, p. 50, Jan. 2006.

[6] E. J. Benjamin, "Heart disease and stroke statistics—2017 update: A report from the American Heart Association," *Circulation*, vol. 135, pp. e146–e603, Mar. 2017.

[7] W. Hong, A. Earnest, P. Sultana, Z. Koh, N. Shahidah, and M. E. H. Ong, "How accurate are vital signs in predicting clinical outcomes in critically ill emergency department patients," *Eur. J. Emergency Med.*, vol. 20, no. 1, pp. 27–32, Feb. 2013.

[8] R. M. Merchant, L. Yang, L. B. Becker, R. A. Berg, V. Nadkarni, G. Nichol, B. G. Carr, N. Mitra, S. M. Bradley, B. S. Abella, and P. W. Groeneveld, "Incidence of treated cardiac arrest in hospitalized patients in the United States," *Crit. Care Med.*, vol. 39, no. 11, pp. 2401–2406, Nov. 2011.

[9] J. P. Nolan, J. Soar, A. Cariou, T. Cronberg, V. R. M. Moulaert, C. D. Deakin, B. W. Bottiger, H. Friberg, K. Sunde, and C. Sandroni, "European resuscitation council and European society of intensive care medicine guidelines for post-resuscitation care 2015: Section 5 of the European resuscitation council guidelines for resuscitation 2015," *Resuscitation*, vol. 95, pp. 202–222, Dec. 2015.

[10] D. A. Jones, M. A. DeVita, and R. Bellomo, "Rapid-response teams," *New England J. Med.*, vol. 365, no. 2, pp. 139–146, 2011.

[11] S. L. Kronick, M. C. Kurz, S. Lin, D. P. Edelson, R. A. Berg, J. E. Billi, J. G. Cabanas, D. C. Cone, D. B. Diercks, J. Foster, R. A. Meeks, A. H. Travers, and M. Welsford, "Part 4: Systems of care and continuous quality improvement: 2015 American Heart Association guidelines update for cardiopulmonary resuscitation and emergency cardiovascular care," *Circulation*, vol. 132, pp. S397–S413, Nov. 2015.

[12] G. B. Smith, D. R. Prytherch, P. E. Schmidt, and P. I. Featherstone, "Review and performance evaluation of aggregate weighted 'track and trigger' systems," *Resuscitation*, vol. 77, no. 2, pp. 170–179, May 2008.

[13] G. B. Smith and D. R. Prytherch, "Widely used track and trigger scores: Are they ready for automation in practice?" *Resuscitation*, vol. 85, no. 10, p. e157, Oct. 2014.

[14] *National Early Warning Score (NEWS) 2 Standardizing the Assessment of Acute-Illness Severity in the NHS: Additional Implementation Guidance*, Rcplondon.Ac.Uk, Royal College of Physicians, London, U.K., Mar. 2020.

[15] J. Kim, Y. R. Park, J. H. Lee, J.-H. Lee, Y.-H. Kim, and J. W. Huh, "Development of a real-time risk prediction model for in-hospital cardiac arrest in critically ill patients using deep learning: Retrospective study," *JMIR Med. Informat.*, vol. 8, no. 3, Mar. 2020, Art. no. e16349, doi: 10.2196/16349.

[16] H. Thorsen-Meyer, A. B. Nielsen, A. P. Nielsen, B. S. Kaas-Hansen, P. Toft, J. Schierbeck, T. Strom, P. J. Chmura, M. Heimann, L. Dybdahl, L. Spangsege, P. Hulsen, K. Belling, S. Brunak, and A. Perner, "Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: A retrospective study of high-frequency data in electronic patient records," *Lancet Digit. Health.*, vol. 2, no. 4, pp. e179–e191, Apr. 2020, doi: 10.1016/s2589-7500(20)30018-2.

[17] S. L. Hyland, M. Faltys, M. Hüser, X. Lyu, T. Gumbsch, C. Esteban, C. Bock, M. Horn, M. Moor, B. Rieck, M. Zimmermann, D. Bodenham, K. Borgwardt, R. Gunnar, and T. M. Merz, "Early prediction of circulatory failure in the intensive care unit using machine learning," *Nature Med.*, vol. 26, no. 3, pp. 364–373, Mar. 2020, doi: 10.1038/s41591-020-0789-4.10.1038/s41591-020-0789-4.

[18] J. Kwon, Y. Lee, Y. Lee, S. Lee, and J. Park, "An algorithm based on deep learning for predicting in-hospital cardiac arrest," *J. Amer. Heart Assoc.*, vol. 7, no. 13, Jul. 2018, Art. no. e008678.

[19] A. Rajkomar, "Scalable and accurate deep learning with electronic health records," *NPJ Digit. Med.*, vol. 1, p. 18, May 2018.

[20] D.-H. Jang, J. Kim, Y. H. Jo, J. H. Lee, J. E. Hwang, S. M. Park, D. K. Lee, I. Park, D. Kim, and H. Chang, "Developing neural network models for early detection of cardiac arrest in emergency department," *Amer. J. Emergency Med.*, vol. 38, no. 1, pp. 43–49, Jan. 2020.

[21] F. E. Shamout, T. Zhu, P. Sharma, P. J. Watkinson, and D. A. Clifton, "Deep interpretable early warning system for the detection of clinical deterioration," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 2, pp. 437–446, Feb. 2020.

[22] M. Sung, S. Hahn, C. H. Han, J. M. Lee, J. Lee, J. Yoo, J. Heo, Y. S. Kim, and K. S. Chung, "Event prediction model considering time and input error using electronic medical records in the intensive care unit: Retrospective study," *JMIR Med. Informat.*, vol. 9, no. 11, Nov. 2021, Art. no. e26426, doi: 10.2196/26426.
[23] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun, "GRAM: Graph-based attention model for healthcare representation learning," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2017, pp. 787–795.
[24] C. Yin, R. Zhao, B. Qian, X. Lv, and P. Zhang, "Domain knowledge guided deep learning with electronic health records," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2019, pp. 738–747.
[25] L. J. Liu, V. Ortiz-Soriano, J. A. Neyra, and J. Chen, "KGDAL: Knowledge graph guided double attention LSTM for rolling mortality prediction for AKI-D patients," in *Proc. ACM BCB*, Aug. 2021, pp. 1–10, doi: 10.1145/3459930.3469513.
[26] P. Velićković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," 2017, *arXiv:1710.10903*.
[27] M. A. Ahmad, A. Teredesai, and C. Eckert, "Interpretable machine learning in healthcare," in *Proc. IEEE Int. Conf. Healthcare Informat. (ICHI)*, Jun. 2018, pp. 559–560.
[28] Z. C. Lipton, J. Berkowitz, and C. Elkan, "A critical review of recurrent neural networks for sequence learning," 2015, *arXiv:1506.00019*.
[29] P. Chen, W. Dong, J. Wang, X. Lu, U. Kaymak, and Z. Huang, "Interpretable clinical prediction via attention-based neural network," *BMC Med. Informat. Decis. Making*, vol. 20, no. S3, pp. 1–9, Jul. 2020.
[30] C. Dos Santos and M. Gatti, "Deep convolutional neural networks for sentiment analysis of short texts," in *Proc. COLING 25th Int. Conf. Comput. Linguistics, Tech. Papers*, 2014, pp. 69–78.
[31] J. Kong, W. Kowalczyk, S. Menzel, and T. Back, "Improving imbalanced classification by anomaly detection," *Parallel Problem Solving From Nature PPSN XVI* (Lecture Notes in Computer Science), vol. 12269. Cham, Switzerland: Springer, 2020, pp. 512–523.
[32] J. Gao, X. Song, Q. Wen, P. Wang, L. Sun, and H. Xu, "Robust time series anomaly detection via decomposition and convolutional neural networks," in *Proc. MileTS 6th KDD Workshop Mining Learn. From Time Ser.*, 2020, pp. 1–6.
[33] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," 2015, *arXiv:1505.00853*.

THANH-CONG DO received the B.S. degree from the Department of Information Technology, VNU University of Engineering and Technology, in 2015, the M.S. degree from the Department of Artificial Intelligence Convergence, Chonnam National University, South Korea, in 2021. His research interests include computer vision, time-series forecasting, and deep learning algorithms in medical healthcare applications.

HYUNG-JEONG YANG (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Chungbuk National University, Gwangju, South Korea. She is currently a Professor with the Department of Artificial Intelligence Convergence, Chonnam National University. Her main research interests include multimedia data mining, medical data analysis, social network service data mining, and video data understanding.

GUEE-SANG LEE (Member, IEEE) received the B.S. degree in electrical engineering and the M.S. degree in computer engineering from Seoul National University, Republic of Korea, in 1980 and 1982, respectively, and the Ph.D. degree in computer science from The Pennsylvania State University, in 1991. He is currently a Professor with the Department of Electronics and Computer Engineering, Chonnam National University, Republic of Korea. His research interests include image processing, computer vision, and video technology.

SOO-HYUNG KIM (Member, IEEE) received the B.S. degree in computer engineering from Seoul National University, in 1986, and the M.S. and Ph.D. degrees in computer science from Korea Advanced Institute of Science and Technology, in 1988 and 1993, respectively. Since 1997, he has been a Professor with the School of Electronics and Computer Engineering, Chonnam National University, South Korea. His research interests include pattern recognition, document image processing, medical image processing, and ubiquitous computing.

BO-GUN KHO received the B.S. degree from Yonsei University, Seoul, South Korea, and the M.S. degree from Chonnam National University, Gwangju, South Korea. He is currently an Assistant Professor with the Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, Chonnam National University Hospital. His main research interests include critical care and rapid response systems in hospital.