A Transformer-Based Prediction Method for Depth of Anesthesia During Target-Controlled Infusion of Propofol and Remifentanil

Yongkang He, Siyuan Peng, *Member, IEEE*, Mingjin Chen, Zhijing Yang, and Yuanhui Chen

***Abstract*— Accurately predicting anesthetic effects is essential for target-controlled infusion systems. The traditional (PK-PD) models for Bispectral index (BIS) prediction require manual selection of model parameters, which can be challenging in clinical settings. Recently proposed deep learning methods can only capture general trends and may not predict abrupt changes in BIS. To address these issues, we propose a transformer-based method for predicting the depth of anesthesia (DOA) using drug infusions of propofol and remifentanil. Our method employs long short-term memory (LSTM) and gate residual network (GRN) networks to improve the efficiency of feature fusion and applies an attention mechanism to discover the interactions between the drugs. We also use label distribution smoothing and reweighting losses to address data imbalance. Experimental results show that our proposed method outperforms traditional PK-PD models and previous deep learning methods, effectively predicting anesthetic depth under sudden and deep anesthesia conditions.**

***Index Terms*— Depth of anesthesia prediction, bispectral index, transformer, drug infusion history, data imbalance.**

I. Introduction

With the advancement of automated control technology, intravenous target-controlled infusion techniques are increasingly being utilized in anesthesia procedures [1], [2]. The pharmacokinetic-pharmacodynamic (PK-PD) model [3], [4] is currently widely adopted in infusion pumps to calculate the effector compartment concentration of anesthetic drugs. However, the traditional PK-PD model has a significant limitation. In clinical practice, it typically requires the selection of multiple parameters due to individual organism differences [5]. This is because the drug effect between the drug dose and a specific organism is unclear. Even if the same dose of an anesthetic drug is administered at the same time, physiological responses vary from person to person [6]. To date, no reasonable or effective research has been conducted on precisely administering drugs according to a specific individual, while combining with the existing PK-PD model.

An accurate drug efficacy prediction model is essential for intravenous target-controlled infusion systems [7], [8]. In recent years, deep learning methods have been investigated for addressing this problem [9], [10]. Compared to traditional PK-PD prediction models, deep learning methods have the advantage of complex nonlinear dynamic computation, resulting in good prediction performance under different situations such as complex environmental information, unclear knowledge background, and unclear inference rules. Lee et al. proposed a method in [11] that combines the PK-PD model framework with the long short-term memory (LSTM) network to extract features from drug injection history information, and then incorporates human physiological characteristics such as age, gender, height, and weight to predict the Bispectral Index (BIS). Although this method shows significant improvement in anesthesia depth prediction compared to previous PK-PD-based methods, it performs poorly on samples with large fluctuations in BIS. As a result, the deep learning-based prediction method proposed in [11] is less efficient at predicting the depth of anesthesia (DOA) during unexpected situations.

In addition, some researchers have successfully utilized the Electroencephalogram (EEG) signal [12] to calculate the BIS value. For example, Li et al. used the Butterworth filter to extract several features such as column entropy, sample entropy, wavelet entropy, and band power from EEG, and then input these features to the sparse denoising autoencoder and long short term memory (SDAE-LSTM) network to predict the DOA [13]. Combining the signal processing and deep learning technique, this method has high prediction accuracy for the DOA. However, the EEG-based prediction method is less practical than the PK-PD-based prediction method, since it requires the huge amount of the EEG signal data and is

Manuscript received 12 October 2022; revised 11 March 2023 and 20 July 2023; accepted 10 August 2023. Date of publication 15 August 2023; date of current version 25 August 2023. This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2021A1515011341 and Grant 2023A1515011185 and in part by the Guangdong Provincial Key Laboratory of Intellectual Property and Big Data under Grant 2018B030322016. *(Corresponding author: Siyuan Peng.)*

Yongkang He, Siyuan Peng, and Mingjin Chen are with the School of Information Engineering, Guangdong University of Technology, Guangzhou 510006, China (e-mail: 1181309500@qq.com; peng0074@gdut.edu.cn; 2112103033@mail2.gdut.edu.cn).

Zhijing Yang is with the School of Information Engineering, Guangdong University of Technology, Guangzhou 510006, China, and also with the Guangdong Provincial Key Laboratory of Intellectual Property & Big Data, Guangzhou 510006, China (e-mail: yzhj@gdut.edu.cn).

Yuanhui Chen is with Zhejiang Hospital of Integrated Traditional Chinese and Western Medicine, Baoshan 310005, China (e-mail: chenyuanhui_831@hotmail.com).

Digital Object Identifier 10.1109/TNSRE.2023.3305363

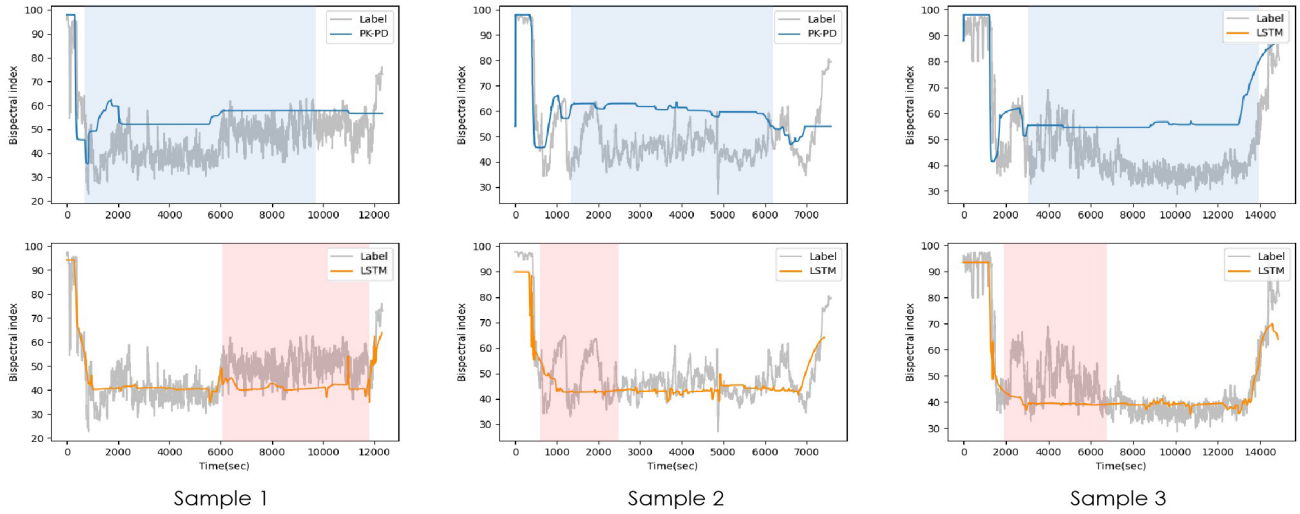


Fig. 1. Prediction results of PK-PD and LSTM for different samples. Top: the blue line is the predicted values of the PK-PD model, which shows a great deviation from the ground true value, especially in the light blue area. Bottom: the orange line is the predicted values of the LSTM model, which usually has relatively poor results (see the light pink area) under the conditions of abrupt change.

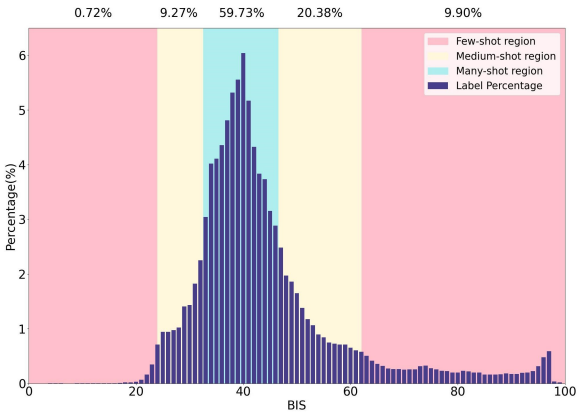


Fig. 2. Label distribution of the dataset, which is divided into three regions including the many-shot region (59.73%), the medium-shot region (29.65%), and the few-shot region (10.62%).

easily interfered by the electromagnetic. Furthermore, from the viewpoint of real-world applications, the proposed method in [13] is hard to directly establish a simulation environment for EEG signals in the field of anesthesia control.

Fig. 1 shows the prediction results of the PK-PD model [3] and the LSTM-based deep learning method [11] for different samples. The figure reveals two main drawbacks of previous approaches to predicting the depth of anesthesia (DOA). Firstly, when sudden changes in bispectral index (BIS) occur during the maintenance period, the prediction results of previous methods remain relatively stable and do not reflect the actual changes, as can be seen in the light blue and pink areas in Fig. 1. Secondly, the BIS data collected for anesthesia clinical records are often unbalanced, with most values falling in the $30-50$ range, as illustrated in Fig. 2. Previous works have neglected the few-shot region, leading to overfitting in many-shot regions and inaccurate predictions in other regions.

This paper proposes a new deep learning method, based on transformer architecture, to accurately predict the DOA using the drug infusion of propofol and remifentanil. The proposed method uses the fusion of human parameters, drug injection history, and derived multimodal features to enhance prediction accuracy. The PK-PD model is embedded at the beginning of the network to provide pseudo-historical information, which is corrected in the training phase using the LSTM network and bottleneck layer. The gate residual network (GRN) module is then applied to fuse multidimensional features and patient context information, suppress irrelevant variables, and aggregate physiological characteristics into each time step. An improved attention mechanism is used to learn long-term dependencies among mixed features for exploiting drug-drug interactions. To overcome data imbalance, the proposed method uses label distribution smoothing and reweighting losses to prevent overfitting in many-shot regions and exhibit good prediction ability in other regions.

In summary, the main contribution of this work are as follows:

1) A new transformer-based deep learning framework is proposed to predict the DOA by using the drug infusion history of propofol and remifentanil simultaneously, which can overcome the limitations of previous DOA prediction methods;
2) A feature fusion layer is developed in the proposed method to combine dynamic and static information from different modalities to achieve the fusion of temporal and textual information, enabling the entire network to fully consider the response of patients with different ages, genders, heights and weights for the same drug;
3) The label distribution smoothing and reweighting losses are used to solve the issue of data imbalance in different intervals in the filed of DOA.

The remainder of this paper is organized as follows. The related works are introduced in Section II. The proposed method is presented in Section III. The experimental results are shown in Section IV. Finally, the conclusion is given in Section V.

II. Related Works

A. Prediction of Anesthetic Efficacy

Anesthetic prediction methods based on PK-PD models have been widely used in clinical drug effect prediction [5], [14]. These methods model the transfer and metabolism of drugs in each component of the human body by solving a system of differential equations. However, PK-PD models with fixed parameters often have poor performance due to inter-patient variability. Although an optimization approach in [15] was used to identify the parameters for different patients, it still required measuring BIS values during the procedure for optimization. Recently, deep learning methods have been proposed for drug effect prediction based on time series prediction [10], [11]. For instance, in [11], an LSTM model was used to extract long-term and short-term memory of drug injection records, and combined with patient characteristics for BIS value prediction. However, the proposed method did not pay enough attention to sparse data samples (e.g. deep anesthesia states with BIS below 40), which often leads to poor performance on imbalanced data.

B. Multivariate Time Series Forecasting

Distinguishing from the univariate time series forecasting, the distinction between variables needs to be considered under the task of multivariate input, since different variables may be heterogeneous or even belong to different modalities. Previous studies have often considered in the location of the feature fusion layer. For example, Anastasopoulos et al. experimentally verified that multimodal data fusion usually have better performance in the middle or deep layers of the network [16]. Pérez-Rúa et al. discovered the optimal architecture from a large number of possible combinations of positions of the network by means of a sequential model-based exploration method approach [17]. In recent years, heterogeneous feature fusion based on the attention mechanisms [18] technique has gradually applied in the practical tasks [19], [20]. Bryan Lim et al. proposed to use a gating mechanism and an interpretable attention mechanism to achieve multilevel time series prediction [19]. Arevalo et al. developed a gated multimodal units, which found the best combination from different combinations of data and allowed to apply this fusion strategy anywhere in the model [20].

C. Data Imbalance

For the anesthesia clinical records, the collected data are often unbalanced (as shown in the Fig.2, most of the BIS values are in the range of 30-50). To solve this issue, the proposed method in [21] adjusted the distribution among the data by resampling the samples. However, such methods are often difficult to grasp the sampling ratio, leading to oversampling easily. In order to further address this problem of data imbalance, some improved methods have been successfully proposed in [22] and [23], which adjusted the learning weights of the loss function for a small number of samples. According to [24], one can observe that the methods that directly adjust the weights of the loss function requires a high degree of differentiation between categories. Hence, in [25], a novel method has been developed to improve the performance of reweighting loss by smoothing the label distribution of the samples.

III. Methodology

A. Problem Definition

In this paper, our goal is to design a transformer-based network for accurate drug effect prediction by using the drug injection history of propofol and remifentanil and body covariates together. The overall framework of our proposed method are shown in Fig. 3, which consists of three components: A) **drug effect encoder**, it is used for extracting the temporal features from the drug injection history, B) **feature fusion layer**, it is used for fusing different dynamic and static information, and C) **temporal fusion decoder**, it is used for learning the mixed features of different long-term dependencies. A re-weighted root mean square loss function is adopted in the training phase to overcome the drawback of data imbalance. First, the two anesthetic drugs that are often used in combination in anesthetic surgery are described in detail in the next subsection.

B. Anesthesia Clinical Record and Feature Extraction

1) Prorofol Infusion History: Propofol is a widely used anesthetic drug in general anesthesia procedures [26], which provides rapid and stable hypnosis function and has additional or synergistic hypnotic effects with other drugs used in anesthesia (such as barbiturates, benzodiazepines, opioids and ketamine) [27]. Thanks to its large absorption and rapid elimination by the body, propofol has become the best anesthetic target-controlled infusion (TCI) drug. In the automated target-controlled infusion systems, the injection rate of propofol is often used as one of the most important characteristics for calculating the BIS prediction values [14]. In our work, the injection history of propofol in the range of 1800s before t is adopted as a model feature to predict the BIS value at moment $t+1$.

2) Remifentanil Infusion History: Remifentanil is commonly used as a supplement for the general anesthesia and is extensively metabolized extrahepatically by blood and tissue non-specific esterases, resulting in an extremely rapid clearance efficiency (3 L/min) [3]. However, when the synergistic effect of remifentanil and propofol is given in [15], similarly, the injection history of remifentanil at 1800s before moment t into the proposed model is utilized to predict the BIS value at moment $t+1$.

3) Drug Effect Site Concentration: The concentration of the drug effect represents the ideal concentration of the drug at the site of action in the body. To some extent, it reflects the effect of the anesthetic drug on the DOA. Note that the effector compartment is not real and therefore cannot be measured directly. Based on the traditional three-compartment model [5], it can be calculated by a pharmacokinetic model. Combining the effector compartment concentration and the pharmacokinetic model, the simple pseudo-BIS values are able to initially calculate. Then the neural network method is applied to correct the pseudo-BIS values calculated by the PK-PD model to

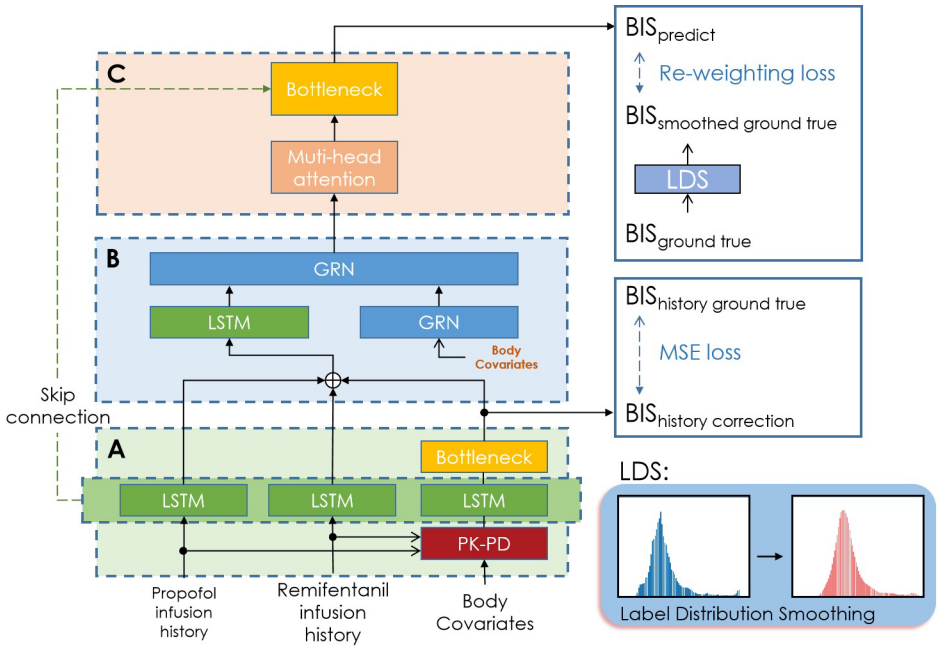


Fig. 3. An overview of our model. Our framework consists of three components, A: Pharmacodynamic encoder, it is used to extract drug history information. Right side of A: correction of pseudo-bis values calculated by PK-PD model; B: Feature fusion layer, it combines the dynamic temporal information with the static human physiological features; C: Temporal fusion decoder, it is adopted to learn the long-term dependencies of temporal hybrid features using a multi-headed attention mechanism.

briefly provide the historical information for the proposed model, which is illustrated in part A of Fig. 3.

C. Model Architecture

Our proposed method combines a recurrent neural network (RNN) with an attention mechanism and transformer architecture. RNNs are effective at capturing features from time series data and preserving memory, but using a single RNN can be difficult when dealing with multiple, heterogeneous inputs. Simply combining different types of data in the network can lead to the loss of important characteristics [17]. To address this issue, our method extracts features from different inputs using separate long short-term memory (LSTM) modules and combines them using a feature fusion layer that controls the mixing of multiple types of information. Furthermore, we incorporate static covariates at each time step to explore the relationship between dynamic and static information. To compensate for the uneven distribution of data caused by the small number of samples, we use the label distribution smoothing method to smooth data with highly unbalanced label categories. At the end of the model, we use a reweighting loss function to assign higher loss weights to categories with sparse numbers (such as deep anesthesia states with BIS values between 20 and 30). This encourages the network to pay more attention to deep anesthesia states with critical sample sizes, despite their small representation in the data.

Our model consists of the following three parts:

1) Drug Effect Encoder: To capture pharmacological feedback under different anesthesia stages, the proposed network first applies the PK-PD model to calculate the pseudo historical information of BIS, and then uses LSTM and bottleneck

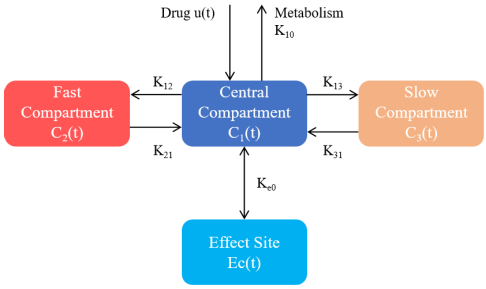


Fig. 4. Structure of the PK model based on three compartments and an effect site compartment.

(as shown on the right side of part A in Fig. 3) to correct the pseudo historical information. The PK-PD model is a classical model widely used in anesthesiology to calculate the effects of anesthetic drugs. The PK-PD model uses a three-compartment model that utilizes the drug infusion rate as input to simulate the transfer and metabolism of the drug between various regions of the body, and reflects the drug effect with an ideal effector compartment. Assuming that the drug effect with an ideal effector compartment has a negligible volume, the clinical effect of the drug is quantified as the effector concentration E_c. Fig. 4 shows the general structure of the PK model used in anesthesia, where the central chamber represents the plasma and tissue, and the fast and slow chambers represent the peripheral chambers, including less perfused organs [28].

According to this structure, the Schnider and Minto model [5] parameters is used to calculate the effect concentration E_c by solving the system of differential equations. Then the response surface model proposed by Short et al. [14] is

adopted to calculate the BIS values:

$$BIS = BIS_0 + (BIS_{\min} - BIS_0)\frac{\left(\frac{Ec_r}{Ec_{50r}} + \frac{Ec_p}{Ec_{50p}}\right)^{\gamma}}{1 + \left(\frac{Ec_r}{Ec_{50r}} + \frac{Ec_p}{Ec_{50p}}\right)^{\gamma}} \tag{1}$$

where Ec_r and Ec_p are the effect concentrations of propofol and remifentanil respectively. γ stands for the nonlinear shape of the sigmoid curve. Ec_{50r} and Ec_{50p} denotes the effector site concentration corresponding to 50% of the maximum clinical effect, which can calculated from Eqn.(2) to Eqn.(5):

$$V_1\frac{dC_1(t)}{dt} = V_2C_2(t)k_{21} + V_3C_3(t)k_{31} - V_1C_1(t)\left(k_{10} + k_{12} + k_{13}\right) + u(t) \tag{2}$$

$$V_2\frac{dC_2(t)}{dt} = V_1C_1(t)k_{12} + V_2C_2(t)k_{21} \tag{3}$$

$$V_3\frac{dC_3(t)}{dt} = V_1C_1(t)k_{13} + V_3C_3(t)k_{31} \tag{4}$$

$$\frac{dEc(t)}{dt} = C_1(t)ke_0 - C_e(t)ke_0 \tag{5}$$

where V_1, V_2, V_3 are the volumes of central compartment, fast compartment and slow compartment, respectively, and k_{ij} is the rate of drug transfer between the chambers. All of them are calculated from human physiological characteristics (age, sex, height, weight), as shown in Table II.

However, a PK model with a limited number of compartments and covariates may be insufficient to accurately account for propofol kinetics. Therefore, we used a separate LSTM modules and Bottleneck (a three fully connected layer) to increase PK-PD model's parameter quantity by extract features from pseudo-history BIS values and injection history of propofol and remifentanil. When encountering a population that is very different from our training set, we can fine-tune its parameters by using a small number of samples to retrain the model.

Because different drugs have different elimination times, For example, propofol takes longer to be absorbed (elimination of 66% in about 25 minutes) while remifentanil has a faster absorption and elimination. Therefore, we used three independent LSTM modules with different parameters to address this issue.

2) Feature Fusion Layer: It should be noted that the effects of drugs can vary among different populations, even when administered at the same dose and time [6]. Generally, the variability between patients makes it impossible to overlook their physiological characteristics. However, the relationship between a patient's physiological information and the corresponding drug effects is not always direct, and this can negatively affect the results. As a result, using a simple fully connected layer may degrade the performance of the model. In our proposed method, we use the gate residual network (GRN) from [19] to control the input variables. GRN uses a gating mechanism to eliminate irrelevant noisy variables and extract important parts from variables in multivariate regression, where the specific contributions of variables to the output are often unknown. The human physiological information (age, gender, height, and weight) is integrated into the network based on GRN to combine temporal and static information features.

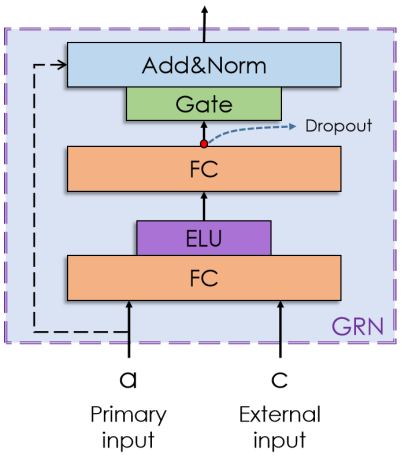


Fig. 5. Structure of GRN. The gated residual network blocks enable the efficient information flow with the skip connections and the gating layer.

Fig. 5 demonstrates the structure of GRN, in which the parameters $\boldsymbol{a}$ and $\boldsymbol{c}$ denote the primary input and external input of GRN respectively. Specific to our method, $\boldsymbol{a}$ is generated by drug effect encoder, which is temporal feature about drug infusion history and BIS pseudo history. And $\boldsymbol{c}$ is patient physiological information. After first fully connection (FC) layer, Exponential Linear Unit (ELU) activation function be used to accelerates the learn speed by reducing the effect of bias offset and making the normal gradient closer to the unit natural gradient [29]. The gated linear unit (GLU) is applied at the output layer to select the input information, which can effectively suppress the noisy information that is not relevant to the output result [30]. Finally, the primary input $\boldsymbol{a}$ is connected to the output layer after passing through the Layernorm Layer [31], which serves to normalize a single sample and accelerate the convergence of the entire network.

3) Temporal Fusion Decoder: In our work, the interpretable multi-headed attention mechanism is used to learn long-term and short-term dependencies between different time steps from the multidimensional features with a mixture of temporal and static information. Different from the classical multi-headed attention, the interpretable multi-headed attention modifies the calculation of the attention weights in multiple heads, for enhancing the characterization of specific features by:

$$\mathcal{F}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \tilde{\boldsymbol{H}}\boldsymbol{W}_{\boldsymbol{H}} \tag{6}$$

where $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ denote the keys, queries and values respectively, all of them are the vectors and originated from the input features. $\boldsymbol{W}_{\boldsymbol{H}}$ denotes the weights for $\tilde{\boldsymbol{H}}$, which is used for linear mapping. In addition, $\tilde{\boldsymbol{H}}$ is obtained by summing each head:

$$\begin{aligned}\tilde{\boldsymbol{H}} &= \tilde{A}(\boldsymbol{Q}, \boldsymbol{K})\boldsymbol{V}\boldsymbol{W}_V \\ &= \left\{\frac{1}{m_H}\sum_{h=1}^{m_H} A\left(\boldsymbol{Q}\boldsymbol{W}_Q^{(h)}, \boldsymbol{K}\boldsymbol{W}_K^{(h)}\right)\right\}\boldsymbol{V}\boldsymbol{W}_V \\ &= \frac{1}{m_H}\sum_{h=1}^{m_H} \text{Attention}\left(\boldsymbol{Q}\boldsymbol{W}_Q^{(h)}, \boldsymbol{K}\boldsymbol{W}_K^{(h)}, \boldsymbol{V}\boldsymbol{W}_V\right)\end{aligned} \tag{7}$$

where $\boldsymbol{W}_Q^{(h)}$ and $\boldsymbol{W}_K^{(h)}$ are the head-specific weights for $\boldsymbol{Q}$ and $\boldsymbol{K}$, and $\boldsymbol{W}_V$ are the value weights shared across all heads.

Specifically, for the temporal fusion features $Z(t) = [z(t-k), \cdots, z(t)]^T$ obtained by the feature fusion layer, the multi-headed attention mechanism is applied:

$$\boldsymbol{A}(t) = \mathcal{F}(Z(t), Z(t), Z(t)) \tag{8}$$

to yield $B(t) = [\beta(t-k), \cdots, \beta(t)]^T$. Finally, the last time step $\beta(t)$ of $B(t)$ is utilized as the output, and then the skip connection is adopted to combine the hidden states $h_\tau^{(1)}, h_\tau^{(2)}, h_\tau^{(3)}$ of the last layer of the three LSTM modules in the encoder and $\beta(t)$ into the bottleneck. The output is the final BIS prediction values. The skip connections are used to facilitate feature fusion as well as to prevent performance degradation caused by over-deepening the network.

4) Label Distribution Smoothing: In classification tasks, it is popular to increase the loss weights of samples from a few categories. This can lead the used network to focus on those categories with less data for solving the issue of data imbalance [24]. However, weighting of the loss function usually requires a high correlation between the label distribution of the samples and the error distribution. Furthermore, a dataset with a continuous label space usually has the following properties: the error distribution is smooth and no longer correlates well with the label density distribution. Therefore, to address the imbalance of the label distribution in our task, the label distribution smoothing proposed in [25] is utilized, which adopts a Gaussian kernel function convolved with the empirical density of the labels to extract a kernel-smoothed new label distribution, given by:

$$\tilde{p}\left(y'\right) \triangleq \int_{\mathcal{Y}} \mathrm{k}\left(y, y'\right) p(y) dy \tag{9}$$

where $p(y)$ is the number of appearances of label y in the training data, $\tilde{p}\left(y'\right)$ is the effective density of label y', and $k(y, y')$ is Gaussian kernel, which characterizes the similarity between target value y' and any y in the target space. The new distribution has an excellent negative Pearson correlation with the error. After obtaining a more efficient label density, the usual methods are able to apply for solving the label imbalance in our task. The details are described in the following section.

5) Loss Function: We first define some notations. We parameterize the encoder (part A in Fig. 3) as θ_{en} (excluding the fixed parameters in PK-PD), the feature fusion layer (part B) is denoted as θ_{fl}, and the decoder (part C) is denoted as θ_{de}. Suppose we are predicting BIS at moment τ, the pseudo-history BIS of PK-PD predictions corrected by neural networks is denotes as $\hat{Y}_i = [\hat{y}_i(\tau-1), \cdots, \hat{y}_i(\tau-l)]$, the ground true history of BIS is denotes as $Y_i = [y_i(\tau-1), \cdots, y_i(\tau-l)], i \in N$, N is training batch size. The final predicted BIS and ground true is denotes as $\hat{y}_i$ and y_i.

To provide the network with a preliminary BIS history, $L_h\left(\hat{Y}, Y; \theta_{en}\right)$ is adopted in the encoder part such that the encoder extracts a BIS value that is as close as possible to the historical true value trend:

$$L_h\left(\hat{Y}, Y; \theta_{en}\right) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\hat{y}_i(\tau-t) - y_i(\tau-t)\right)^2 \tag{10}$$

Based on the length of the drug in fusion history, we set $T = 180$ in the experiments, which corresponds to the length of the input sequence.

In the outset, we used the standard mean squared error (MSE) loss function as the objective function for gradient descent to train the model:

$$L_{MSE}\left(\hat{y}, y; \theta_{en}, \theta_{fl}, \theta_{de}\right) = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{y}_i - y_i\right)^2 \tag{11}$$

However, we found that using only the MSE loss function does not encourage the model to learn the mutation condition of BIS, particularly during the maintenance period of anesthesia (i.e., from 10 minutes after anesthetic injection to the end of drug infusion), where the model tends to learn the mean of the sample. Therefore the model can be further improved by introducing the reductive bias, a weighted combination of multiple loss functions, with each particular function that is used to focus on a different side. Therefore, we use a weighted MSE loss as follows:

$$L_w\left(\hat{y}, y; \theta_{en}, \theta_{fl}, \theta_{de}\right) = \frac{1}{N} \sum_{i=1}^{N} w_i \left(\hat{y}_i - y_i\right)^2, w_i \in W \tag{12}$$

$$W = \frac{1}{\tilde{p}(y')} = \frac{1}{\int_y k\left(y, y'\right) p(y) dy} \tag{13}$$

where the weight matrix $W = [w_1, w_2, \cdots, w_{100}]$ is the inverse of the new label distribution $\tilde{p}(y')$ after kernel smoothing. The smaller the number of samples, the larger the weights w_i. Correspondingly, the loss of the sample is larger. Overall, the optimal model can be obtained by:

$$\arg\min(\lambda_h L_h\left(\hat{Y}, Y; \theta_{en}\right) + \lambda_w L_w\left(\hat{y}, y; \theta_{en}, \theta_{fl}, \theta_{de}\right)) \tag{14}$$

where λ_h and λ_w are the loss weights. In the experiments, we set them to 5 and 10 respectively.

IV. Experiments and Result

A. Data Preparation

In our experiments, the used data is the VitalDB database, which includes the drug injection records and static covariates of patient physiological characteristics (age, gender, height, weight). The detailed description is illustrated in Section B of Methodology. Since the VitalDB database contains the real surgery records collected in real time, there is a lot of noise, interference, and incorrect records in the data, which greatly affect the learning of the model. Additional data processing is required to convert it into a suitable form for computer computation. Therefore, the database is cleaned to minimize the interference of the noisy signals.

Considered that a large amount of noise and error records is in the database, we first performed data cleaning. Since there are many missing values in some samples, the following

TABLE I
PATIENT CHARACTERISTICS, MEAN±STANDARD DEVIATION (MIN-MAX)

	Training Data Set	Validation Data Set	Testing Data Set
N	180	76	76
Age(yr)	56.1 ± 14.0 (17-82)	56.3 ± 15.0 (17-79)	56.2 ± 15.1 (17-79)
Sex(male/female)	113/67	47/29	40/36
weight(kg)	61.5 ± 10.2 (37.9-98.1)	60.7 ± 10.3 (37.9-98.1)	60.0 ± 9.8 (37.9-81.6)
Height(cm)	163.2 ± 8.2 (138.8-186.6)	162.3 ± 7.9 (138.8-182.0)	161.2 ± 7.5 (138.8-182.0)
Median BIS	41.1 ± 5.4 (25.9-59.5)	43.1 ± 6.1 (23.1-57.2)	42.5 ± 5.8 (30.6-55.9)
Propofol total dose(g)	1.19 ± 0.63 (0.28-3.41)	1.27 ± 0.71 (0.32-3.31)	1.32 ± 0.71 (0.30-4.24)
Propofol Median Ce($\mu g/ml$)	3.02 ± 0.47 (1.91-4.30)	3.06 ± 0.49 (2.00-4.00)	3.05 ± 0.50 (1.60-4.00)
Remifientanil total dose(g)	1.46 ± 1.01 (0.29-6.29)	1.43 ± 0.84 (0.25-3.70)	1.46 ± 0.91 (0.34-5.16)
Remifientanil Median Ce($\mu g/ml$)	3.73 ± 1.08 (1.50-6.01)	3.67 ± 0.95 (2.00-6.97)	3.70 ± 0.87 (2.00-6.00)

operations are performed: 1) interpolate the data outliers and nulls with linear interpolation; 2) discard samples with more than 30s data loss; 3) discard samples with only half-field surgery records. After that, the data are subjected to additional processing. Considering the retention time of the drug on the human body, propofol clears 66% of the time after 3 hours of injection for about 25 minutes [27], we set the time range of the drug input history within 1800s to extract the features, which is consistent with the hypothesis in [32]. It is worth noting that during the initial periods, the input, in the form of a zero sequence concatenate a medication record, is used for our proposed model to predict the BIS value. That means, when $t \in [1, 1800]s$, the input is a zero sequence of length $(1800 - t)$ seconds concatenate a t seconds of medication record.

The injection histories of propofol and remifentanil in the original database are the total drug injections recorded every 1s. To save the computing resources, the data are processed into the drug doses injected in every 10s (i.e., 180 data points). In addition, the drug dose is divided by the length of time to obtain the drug injection rate, which often is used in the PK-PD model to calculate BIS characteristics. The drug injection history and other static covariates are normalized to facilitate faster network convergence.

To suppress the negative influence of the noise contained in the database, the true BIS values of the training set is smoothed and the locally weighted scatter plot smoothing (LOWESS) with a smoothing parameter of 0.03 is used for the original BIS values to reduce the computational error during the training phase. To ensure the authenticity of the experiments, the validation set and test set are not smoothed.

B. Data Characteristics

We use the VitalDB database as our capital training set, which is a open access data source, freely downloadable from the website, https://vitaldb.net. In our experiments, 680 samples that contain the real surgical records with TIVA general anesthesia injection are randomly selected from the VitalDB database as the original database. After data cleaning, 348 samples with serious missing data (e.g., only half of the surgical records) are excluded. In this case, only 332 samples are used as the remaining samples. Among them, 180 samples are randomly selected as the training set, 76 samples are randomly selected as the validation set, and the other samples are used as the test set. The characteristics of these three sets of samples are shown in the Table I.

C. Experimental Settings

The evaluation metrics including median performance error (MDPE), median absolute performance (MDAPE) and mean square error (RMSE) are used to evaluate the performance of our proposed model. The predictive performance is evaluated for each period according to the definition of three periods in anesthesia surgery (induction period: from the start of the anesthetic drug propofol injection until 10 minutes later, maintenance period: from the end of the induction period until the cessation of prorofol injection, and recovery period: from the cessation of propofol injection until the end of the surgical record). In addition, a paired t-test is used to compare the performance of our model with other compared methods, and the experimental results are expressed as mean ± SD (range). Statistical analysis is performed with SPSS 21 (IBM, USA), and $P < 0.05$ is the considered significant for the paired t-test.

The model is optimized by using the Adam optimizer. Batchsize is set to 1024 while training, the initial learning rate is 0.03, and the learning rate decays to 0.1 times after every 10 epochs. Unless otherwise mentioned, the parameters of PK-DK model used in our proposed method are shown in Table II.

The pytorch 1.4.0 framework and python 3.9.1 are adopted in our implementations. All experiments are run on a single 24GB NVIDIA TITAN RTX GPU, which takes about 30 minutes to train 50 epochs with batch size 1024 on the entire dataset. Our code is made available at https://github.com/heeeyk/Transformer-DOA-Prediction.

D. Experimental Results

In the experiments, our proposed model has made a comparison with the LSTM method [11] and the PK-PD method [14], the experimental results are shown in the Table III.

TABLE II
PK AND PD PARAMETERS OF PROPOFOL AND REMIFENTANIL

Params	Propofol	Remifentanil
model	Schnider	Minto
V_1	4.27	5.1 - 0.0201*(age - 40) + 0.072*(lbm - 55)
V_2	18.9 - 0.391*(age - 53)	9.82 - 0.0811*(age - 40)
V_3	238	5.42
C_1	1.89 + (wgt - 77)*0.0456 - (lbm - 59)*0.0681 + (hgt - 18)*0.0264	2.6 - 0.0162*(age - 40) + 0.0191*(lbm - 55)
C_2	1.29 - 0.024*(age - 53)	2.05 - 0.0301*(age - 40)
C_3	0.836	0.076 - 0.00113*(age - 40)
ke_0	0.46	0.595 - 0.007*(age - 40)
$E_0 - E_{max}(BIS)$	98-0	98-0
$Ec_{50}(\mu g/mL)$	4.47	19.3
γ	1.43	1.43

• $Age = age(y); wgt = weight(kg); hgt = height(cm); lbm = lean\ body\ mass$
$lbm_{male} = 1.1 * wgt - 128 * (wgt/hgt)^2, lbm_{female} = 1.07 * wgt - 140 * (wgt/hgt)^2$

TABLE III
COMPARISON OF ERRORS BETWEEN OURS PROPOSED MODEL AND THE BASELINE MODEL DURING THREE ANESTHESIA PERIODS

	MDPE(%)			MDAPE(%)			RMSE		
Anesthesia Period	PK-PD	LSTM	Ours	PK-PD	LSTM	Ours	PK-PD	LSTM	Ours
All	21.75 ± 12.65	3.64 ± 14.96	**-2.08** ± 14.91	24.23 ± 10.16	15.97 ± 7.91	**15.51** ± 6.87	15.64 ± 5.19	10.20 ± 2.45	**9.52** ± 2.35
Induction	-16.09 ± 23.12	-6.35 ± 20.50	**4.64** ± 17.82	27.18 ± 13.89	22.75 ± 9.39	**18.52** ± 8.32	17.62 ± 4.56	14.57 ± 3.79	**12.91** ± 4.14
Maintenance	22.89 ± 12.74	-2.99 ± 15.24	**-2.62** ± 15.32	24.34 ± 10.70	**15.08** ± 8.34	15.26 ± 7.51	14.78 ± 5.84	8.72 ± 2.78	**8.50** ± 2.62
Recovery	16.54 ± 18.15	-13.92 ± 25.44	**-4.00** ± 22.23	22.29 ± 12.46	24.97 ± 16.03	**19.18** ± 13.64	16.93 ± 7.86	15.18 ± 6.43	**12.38** ± 5.55

From this table, one can observe that the proposed model outperforms the baseline method and the PK-PD method in all periods in terms of evaluation metrics except for MDAPE in the maintenance period, in which our proposed model has a slight performance degradation. The main reason is that the baseline method tends to predict the smooth BIS curves, which coincides with the overall trend in the maintenance period with a very large sample size, and thus has a better performance on the whole dataset. The detailed analysis is described in the next section. The performance comparison on different test samples is shown in Fig. 6. Obviously, our method has greatly improvements on the predictive capability for the mutation conditions and outperforms the baseline method, which are performed significantly in the light pink areas. In addition, the concordance correlation coefficient (CCC) is used to measure the correlation between BIS and ground true BIS. The experimental results for all methods are shown in Table IV. The CCC (95% Confidence Interval) is 0.677 [0.691 to 0.665] in our model, which is significantly larger than that in the LSTM method (0.590 [0.582 to 0.609]) and in the PK-PD method (0.556 [0.543 to 0.571]).

In the real-world applications, the DOA prediction is often set to predict the BIS value every 1 second. Therefore, we set the batch size to 1 to simulate the inference process of the DOA prediction, and find that our model only takes 0.014 seconds to predict the BIS value every single time. This indicates that our model can be used for real-time monitoring.

TABLE IV
CCC (CONCORDANCE CORRELATION COEFFICIENT) COMPARISONS

	PK-PD	LSTM	Ours
CCC	$0.556^{0.571}_{0.543}$	$0.595^{0.609}_{0.582}$	$0.677^{0.691}_{0.665}$

• C_b^a denotes concordance correlation coefficient with 95% confidence interval upper bound a and lower bound b.

E. Test Error Analysis

For a single case, one can observe the extreme unreasonableness of prediction curve for the baseline method, such as lacking variation and fluctuating in a rough range around 40 for the BIS values. This coincides with the label distribution of the dataset, as shown in the upper part of Fig. 8. By analyzing the label distribution of the dataset, it is verified that the baseline method is disappointing in terms of the overall

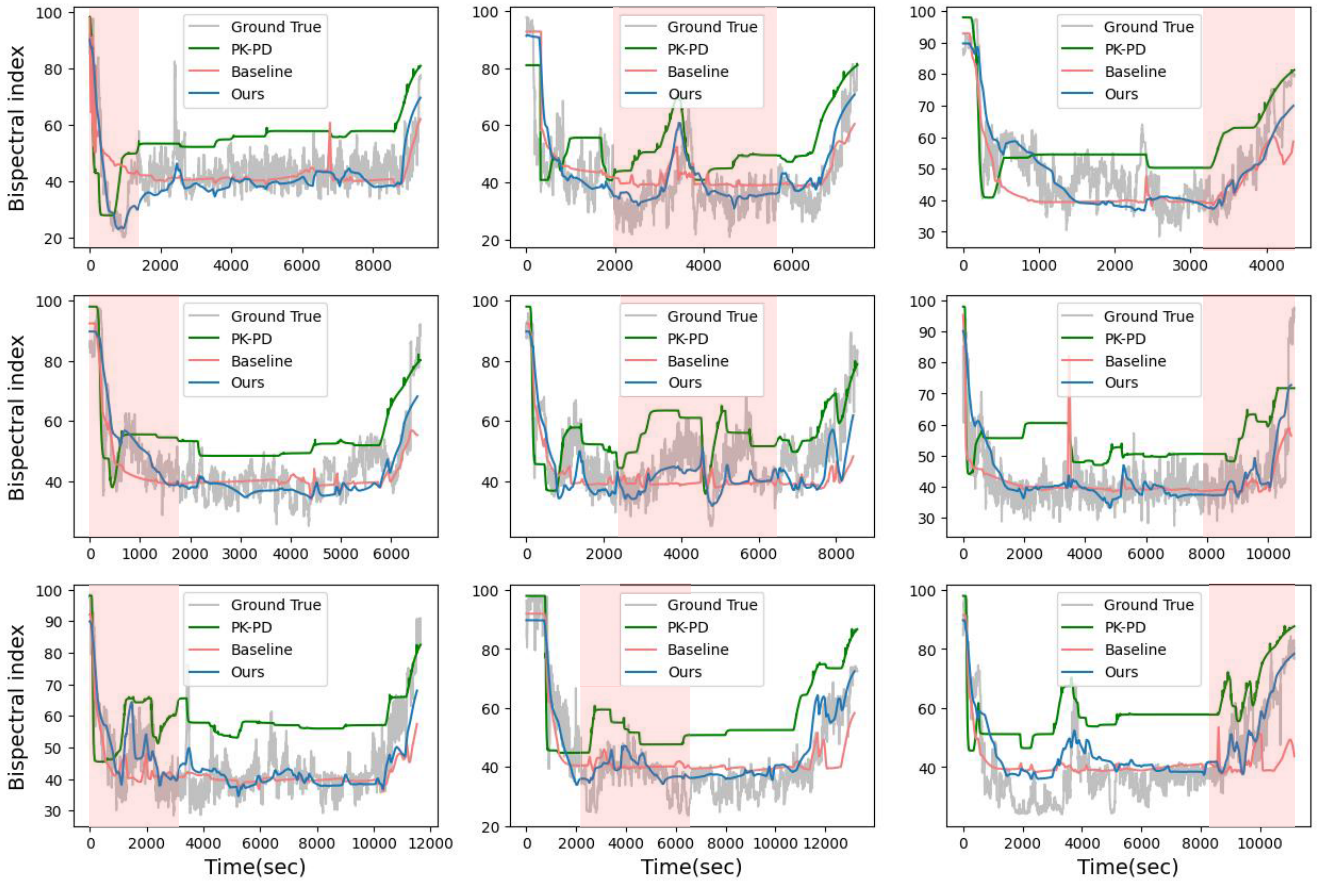
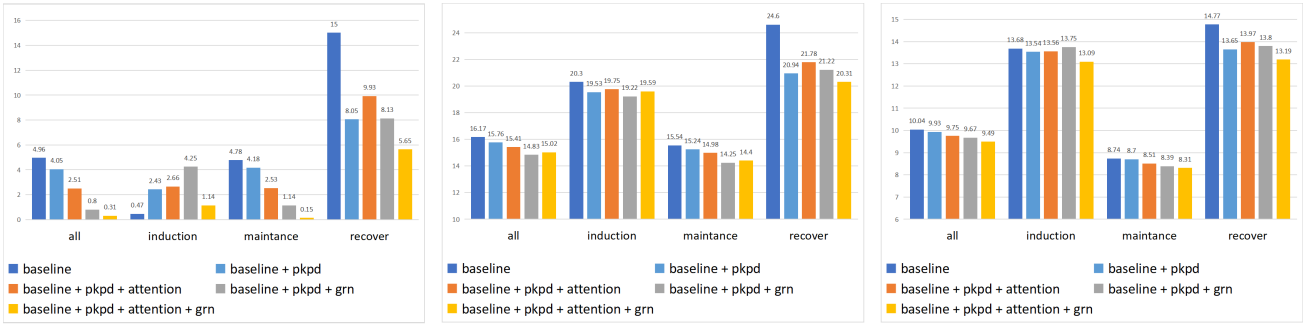


Fig. 6. Performance comparison between our proposed method and other compared methods (i.e., the baseline LSTM method [11] and the PK-PD [14] method) on different test samples. Each subfigure corresponds to an independent sample.



(a) (b) (c)

Fig. 7. Prediction results of ablation study on different components for our proposed model. (a): MDPE; (b): MDAPE; (c): RMSE.

prediction performance, because it has a overfitting problem in the many-shot region with large data volume and neglects the prediction ability for other regions, especially the unbalanced data distribution happens to be the medical data.

Therefore, the testing errors in the sample labels is adopted to verified the fact that our method outperforms the baseline method. Specifically, the test error is calculated by the following formula:

$$error^{(j)} = \frac{1}{n^{(j)}} \mid \sum_{i=1}^{l} \left([\hat{Y}_i]^{(j)} - [Y_i]^{(j)} \right) \mid, \quad j \in [0, 100] \tag{15}$$

where j denotes the range of the BIS values, $n^{(j)}$ is the number of points in the dataset, [.] stands for the rounding operation, $\hat{Y}_i$ and Y_i denote the predicted output of the model and the ground true for the i^{th} sample, respectively. As shown in Fig. 8, one can see that, compared with the baseline method, our proposed method has great improvements on the prediction performance in other regions, and at the same time, without reducing the predictive power in the many-shot region. Especially in the few-shot interval between 10 to 20 and between 60 to 90, the test error decreases by **15.23%** and **48.99%**, respectively. These experimental results are coincide with the rare deep anesthesia state and the shallow anesthesia

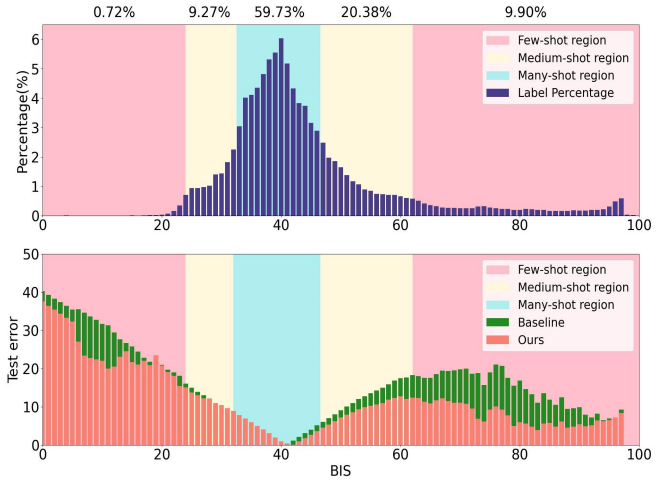


Fig. 8. Top: the sample label distribution of the dataset is divided into three regions, in which the largest amount of data in the many-shot region contains 59.73% and the smallest amount of data in the few-shot region contains 10.62%. Bottom: comparison of the baseline method with our proposed method on the test error.

state between wakefulness and anesthesia during the induction period. Moreover, the enhancement of the predictive ability over these two regions indicates that our proposed method solves the overfitting problem in the many-shot region.

F. The Region Where BIS Mutations Occur

For all experimental data, the main mutations of BIS occurs in the medium-shot region, in which the number of the samples is less than 30%. In order to improve the prediction performance, a weighted MSE loss function is adopted in the proposed model to solve the issue of data imbalance, such that much attention is paid to the medium-shot region. We have conducted a statistical analysis during the relatively stable anesthesia maintenance phase, and the percentage of the mutations of BIS in different regions is shown in Table V From this table, one can see that, if the BIS value, denoted as B_t at time t, satisfies one of the following conditions:

$$\begin{cases} |B_t - min(B_T)| > m \\ |B_t - max(B_T)| > m, \end{cases} \quad T \in (t-30, t+30)s \qquad (16)$$

where m denotes the magnitude of BIS changes, and a larger m indicates a greater mutation magnitude. In Table V, we have illustrated the results with $m = 5, 7$ and 10 respectively. In particular, when $m = 10$, approximately 33% of the BIS mutations occurred in the Many-shot region ($BIS \in [31, 48]$). However, the rest of the BIS mutations occurred in the Medium-shot and Few-shot regions. This means, the loss function assigns the weight to these data points in the Medium-shot and Few-shot regions more than twice the weight in the Many-shot region. Therefore, our loss function enables the network to pay more attention to the mutation condition of BIS. This indicates, the number of samples and the mutation condition of BIS has necessary relationship in our proposed model.

TABLE V
PERCENTAGE OF THE MUTATIONS OF BIS IN DIFFERENT REGIONS

Region	$m = 5$	$m = 7$	$m = 10$	The weights of loss function
Many-shot $BIS \in [31,\ 48)$	50.94%	34.82%	33.33%	$w \in [1,\ 2)$
Medium-shot $BIS \in [48,\ 54)$	21.23%	27.93%	33.33%	$w \in [2,\ 4)$
Medium-shot $BIS \in [54,\ 64)$	16.86%	26.89%	19.04%	$w \in [4,\ 8)$
Few-shot $BIS \geq 64$	7.46%	10.34%	14.28%	$w \geq 8$

G. Ablation Experiments

To illustrate the contribution of each individual module in our proposed model, the ablation experiments are conducted. Starting from our full network model and gradually removing some of the network components, each method is trained under the same conditions until the network converged, and the experimental results in terms of MDPE, MDAPE and RMSE are shown in Fig. 7 respectively. From this figure, one can observe that, each network component has a contribution to improve the performance metrics (e.g., MDAPE and RMSE) for the prediction of DOA. Our proposed model outperforms the baseline method in the entire anesthesia period. However, the prediction performance may degrade during the induction period. Based on the experimental results, this problem may be caused by the introduction of the PK-PD model, because the PK-PD model in the induction period predicts the BIS values with large deviations (e.g., there is a certain time lag). This will lead to the performance degradation. In addition, if the GRN component is discarded and only the attention layer is considered, the effect will not be greatly improved compared with the combination of the baseline method plus PK-PD. The main reason is that the lack of the GRN component prevents the effective inclusion of static covariates in the temporal information, which often leads to a bad prediction performance.

In general, incorporating the PK-PD model into the network can improve the prediction performance of the model because it contains a large number of hyperparameters calculated from human experiments itself. Furthermore, it provides some statistical data to the deep learning model for enhancing the robustness of the model. The PK-PD model also does regression calculations on the past BIS indices, which brings much time series information to the model. But the disadvantage is that the PK-PD model has bias in the induction period, which easily leads to partial performance degradation of the model. To analyze the contribution of the PK-PD model in our method, the combination of the baseline (i.e., LSTM) model and the PK-PD model and the baseline model are conducted in the ablation experiment. From the experimental result in Fig. 7(a), one can observe that the performance of the combination of them can improve in most periods, except for the induction period. In particular, the MDPE in the induction period increases from 0.47 (only LSTM) to 2.43 (LSTM+PK-PD), which degrades the performance of

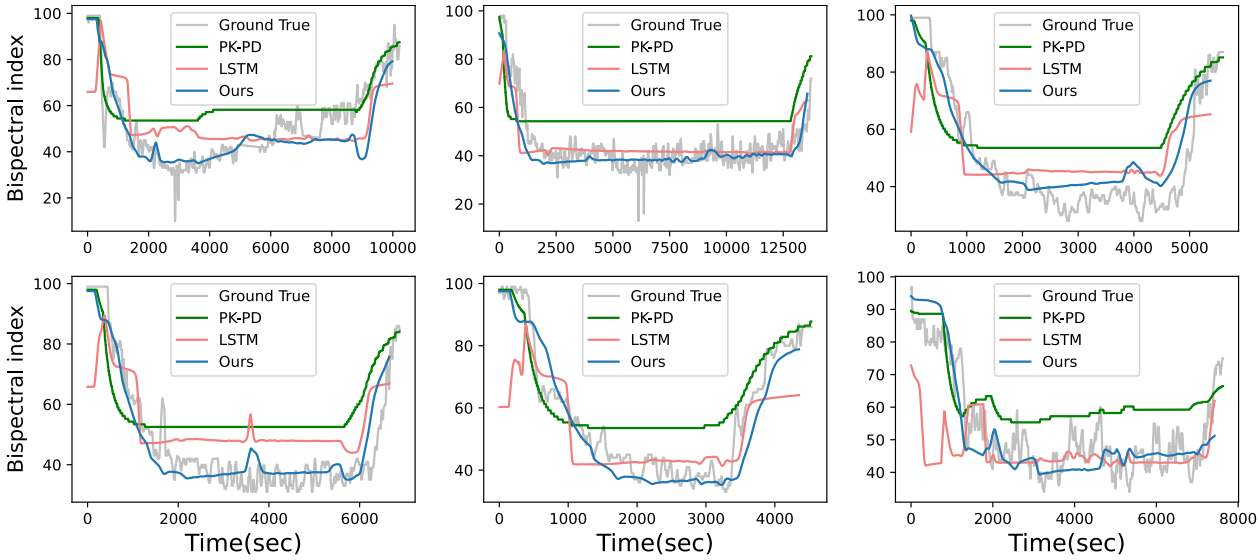


Fig. 9. Performance comparison between our proposed method and the baseline methods (LSTM [11] and PK-PD [14]) on our in-house dataset.

TABLE VI
COMPARISON OF ERRORS BETWEEN OURS PROPOSED MODEL AND OTHER METHODS IN OUR IN-HOUSE DATASET

	MDPE(%)			MDAPE(%)			RMSE		
Anesthesia Period	PK-PD	LSTM	Ours	PK-PD	LSTM	Ours	PK-PD	LSTM	Ours
All	24.54±16.62	3.79±20.55	**2.83**±19.46	28.1±13.24	20.83±11.18	**20.08**±9.40	17.94±6.27	13.23±3.88	**12.08**±3.73
Induction	-19.18±15.63	-8.35±20.77	**-0.46**±13.39	23.81±9.94	20.35±12.10	**12.91**±6.95	16.52±4.80	16.38±6.21	**12.86**±5.47
Maintenance	26.99±16.69	4.71±21.89	**2.55**±21.06	28.92±14.19	**20.76**±12.28	20.87±10.19	17.45±7.09	11.78±4.42	**10.95**±4.25
Recovery	20.27±21.10	-2.85±28.93	**2.37**±27.45	24.39±16.55	27.22±15.10	**24.32**±13.56	19.09±9.79	15.98±6.51	**15.08**±6.50

the model. However, when the other modules (i.e., Attention and GRN) are incorporated into our model, the MDPE in the induction period drops to 1.14, which illustrates our model can reduce the negative influence efficiently for the disadvantage of the PK-PD model.

The GRN module, on the other hand, can favourably reflect the physiological changes of drug doses in different populations by adding the static information to the time series (different ages have different drug elimination rates), and the unique gating mechanism of GRN can eliminate the signal noise and the static variables that have almost no contribution to the output. The attention mechanism can learn the long-term relationship between different time steps, but it is difficult to perform effectively when the dynamic-static information is missing in GRN. The confounding effect of temporal information and static covariance is considered to maximize the effect of the attention mechanism.

H. Model Generalization

To verify the generalization of our model, the experiments that training a model on the VitalDB source and testing it on our in-house dataset are conducted. Our in-house dataset contains 44 cases collected from a hospital in China. The main difference between the the VitalDB dataset and our dataset is shown in Table VII. Since the two datasets are collected from different race and data acquisition equipment, it may lead to the large domain gap. In particular, our dataset is mainly sampled from younger people, hence the average age of samples in our dataset is the larger that that of samples in the VitalDB dataset. In the experiments, 20 cases are randomly selected as the training set to fine-tune the baseline and our model. The PK-PD cannot be fine-tune, since its parameters is fixed. The experimental results are shown in Fig. 9 and Table VI, and obviously our method still has the

TABLE VII
DATASETS DIFFERENCE

	Our dataset	VitalDB(Training set)
N	44	180
Age(yr)	39.9 ± 13.4 (19-69)	56.1 ± 14.0 (17-82)
Sex(male/female)	22/22	113/67
weight(kg)	62.6 ± 10.5 (43-105)	61.5 ± 10.2 (37.9-98.1)
Height(cm)	166 ± 8.1 (147-183)	163.2 ± 8.2 (138.8-186.6)
Median BIS	43.5 ± 10.1 (23.0-68.2)	41.1 ± 5.4 (25.9-59.9)

best performance. It's worth noting that, due to the limitation of our collection equipment, we can only record the BIS values every 5 seconds. In this situation, the BIS and narcotic record may cause inaccurate label information after interpolation.

V. Conclusion

Accurate drug efficacy prediction is helpful for anesthesiologists to make suitable decisions in the clinical procedures. In this paper, a transformer-based prediction method is proposed for predicting the depth of anesthesia. Particularly, the proposed method adopts a LSTM based deep learning architecture and an enhanced attention mechanism to efficiently predict the sudden change of anesthesia depth under the effect of drugs. In addition, a weighted loss function is used in the network to solve the problem of data imbalance, improving the generalization in comparison to previous approaches. Experimental results show that our proposed model has better prediction performance than previous methods, especially in the few-shot region such as deep anesthesia stage and situation.

References

[1] B. J. Anderson and O. Bagshaw, "Practicalities of total intravenous anesthesia and target-controlled infusion in children," *Anesthesiology*, vol. 131, no. 1, pp. 164–185, Jul. 2019.

[2] M. White and G. N. C. Kenny, "Intravenous propofol anaesthesia using a computerised infusion system," *Anaesthesia*, vol. 45, no. 3, pp. 204–209, Mar. 1990.

[3] T. D. Egan, "Remifentanil pharmacokinetics and pharmacodynamics: A preliminary appraisal," *Clin. Pharmacokinetics*, vol. 29, no. 2, pp. 80–94, Aug. 1995.

[4] P. F. White, "Propofol-pharmacokinetics and pharmacodynamics," *Semn Anesthesia*, vol. 7, no. 1, pp. 4–20, 1988.

[5] C. Minto and T. Schnider, "Contributions of PK/PD modeling to intravenous anesthesia," *Clin. Pharmacol. Therapeutics*, vol. 84, no. 1, pp. 27–38, Jul. 2008.

[6] J. Schüttler and H. Ihmsen, "Population pharmacokinetics of propofol: A multicenter study," *J. Amer. Soc. Anesthesiol.*, vol. 92, no. 3, pp. 727–738, 2000.

[7] L. Merigo, F. Padula, N. Latronico, M. Paltenghi, and A. Visioli, "Optimized PID control of propofol and remifentanil coadministration for general anesthesia," *Commun. Nonlinear Sci. Numer. Simul.*, vol. 72, pp. 194–212, Jun. 2019.

[8] K. van Heusden et al., "Optimizing robust PID control of propofol anesthesia for children: Design and clinical evaluation," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 10, pp. 2918–2923, Oct. 2019.

[9] C. W. Connor, "Artificial intelligence and machine learning in anesthesiology," *Anesthesiology*, vol. 131, no. 6, pp. 1346–1359, Dec. 2019.

[10] O. Caelena, O. Caillouxb, D. Ghoundiwalb, A. A. Mirandaa, L. Barvaisb, and G. Bontempia, "Real-time prediction of an anesthetic monitor index using machine learning," *Artif. Intell. Med.*, vol. 7, pp. 1–15, 2011.

[11] H.-C. Lee, H.-G. Ryu, E.-J. Chung, and C.-W. Jung, "Prediction of bispectral index during target-controlled infusion of propofol and remifentanil: A deep learning approach," *Anesthesiology*, vol. 128, no. 3, pp. 492–501, 2018.

[12] J. Wang, J. Cao, D. Hu, T. Jiang, and F. Gao, "Eye blink artifact detection with novel optimized multi-dimensional electroencephalogram features," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 1494–1503, 2021.

[13] R. Li, Q. Wu, J. Liu, Q. Wu, C. Li, and Q. Zhao, "Monitoring depth of anesthesia based on hybrid features and recurrent neural network," *Frontiers Neurosci.*, vol. 14, p. 26, Feb. 2020.

[14] T. G. Short et al., "Refining target-controlled infusion: An assessment of pharmacodynamic target-controlled infusion of propofol and remifentanil using a response surface model of their combined effects on bispectral index," *Anesthesia Analgesia*, vol. 122, no. 1, pp. 90–97, 2016.

[15] J. M. Gonzalez-Cava, J. A. Reboso, J. L. Calvo-Rolle, and J. A. Mendez-Perez, "Adaptive drug interaction model to predict depth of anesthesia in the operating room," *Biomed. Signal Process. Control*, vol. 59, May 2020, Art. no. 101931.

[16] A. Anastasopoulos, S. Kumar, and H. Liao, "Neural language modeling with visual features," 2019, *arXiv:1903.02930*.

[17] J.-M. Pérez-Rúa, V. Vielzeuf, S. Pateux, M. Baccouche, and F. Jurie, "MFAS: Multimodal fusion architecture search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6959–6968.

[18] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[19] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *Int. J. Forecasting*, vol. 37, no. 4, pp. 1748–1764, Oct. 2021.

[20] J. Arevalo, T. Solorio, M. Montes-y-Gómez, and F. A. González, "Gated multimodal units for information fusion," 2017, *arXiv:1702.01992*.

[21] S. García and F. Herrera, "Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy," *Evol. Comput.*, vol. 17, no. 3, pp. 275–306, Sep. 2009.

[22] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[23] Q. Dong, S. Gong, and X. Zhu, "Imbalanced deep learning by minority class incremental rectification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 6, pp. 1367–1381, Jun. 2019.

[24] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu, "Large-scale long-tailed recognition in an open world," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2532–2541.

[25] Y. Yang, K. Zha, Y. Chen, H. Wang, and D. Katabi, "Delving into deep imbalanced regression," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 11842–11851.

[26] H. M. Bryson, B. R. Fulton, and D. Faulds, "Propofol: An update of its use in anaesthesia and conscious sedation," *Drugs*, vol. 50, no. 3, pp. 513–559, Sep. 1995.

[27] M. M. R. F. Struys et al., "Comparison of plasma compartment versus two methods for effect compartment–controlled target-controlled infusion for propofol," *J. Amer. Soc. Anesthesiol.*, vol. 92, no. 2, p. 399, Feb. 2000.

[28] H. J. B. van Oud-Alblas, M. J. E. Brill, M. Y. M. Peeters, D. Tibboel, M. Danhof, and C. A. J. Knibbe, "Population pharmacokinetic-pharmacodynamic model of propofol in adolescents undergoing scoliosis surgery with intraoperative wake-up test: A study using bispectral index and composite auditory evoked potentials as pharmacodynamic endpoints," *BMC Anesthesiol.*, vol. 19, no. 1, pp. 1–12, Dec. 2019.

[29] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," 2015, *arXiv:1511.07289*.

[30] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 933–941.

[31] J. Lei Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[32] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.