BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Anesthesia depth prediction from drug infusion history using hybrid AI

Liang Wang[1], Yiqi Weng[1] and Wenli Yu[1*]

Abstract

Background Accurately predicting the depth of anesthesia is essential for ensuring patient safety and optimizing surgical outcomes. Traditional regression-based approaches often struggle to model the complex and dynamic nature of patient responses to anesthetic agents. Machine learning techniques offer a promising alternative by capturing intricate relationships within physiological data. This study proposes a hybrid model integrating Long Short-Term Memory (LSTM) networks, Transformer architectures, and Kolmogorov-Arnold Networks (KAN) to improve the predictive accuracy of anesthesia depth.

Methods The proposed model combines multiple deep learning techniques to address different aspects of anesthesia prediction. The LSTM component captures the sequential nature of drug administration and physiological responses. The Transformer architecture utilizes attention mechanisms to enhance contextual understanding of patient data. The KAN models nonlinear relationships between drug infusion histories and anesthesia depth. The model was trained and evaluated on patient data from a publicly available anesthesia monitoring database. Performance was assessed using Mean Squared Error (MSE) and compared against other models.

Results The hybrid model demonstrated superior predictive performance compared to conventional regression approaches. Tested on the VitalDB database, the proposed framework achieved a MSE of 0.0062, which is lower than other methods. The inclusion of attention mechanisms and nonlinear modeling contributed to improved accuracy and robustness. The results indicate that the combined approach effectively captures the temporal and nonlinear characteristics of anesthesia depth, offering a more reliable predictive tool for clinical use.

Conclusions This study presents a novel deep learning framework for anesthesia depth prediction, integrating sequential, attention-based, and nonlinear modeling techniques. The results suggest that this hybrid approach enhances prediction reliability and provides anesthesiologists with a more comprehensive analysis of factors influencing anesthesia depth. Future research will focus on refining model robustness, exploring real-time applications, and addressing potential biases in predictive analytics to further improve clinical decision-making.

Keywords Depth of anesthesia, Deep learning, Drug infusion history, LSTM, Transformers, Kolmogorov-Arnold Networks

*Correspondence:
Wenli Yu
yzxwenliyu@163.com
[1]Department of Anesthesiology, Tianjin First Center Hospital, No. 24 Fukang Road, Nankai District, Tianjin 300192, China

Introduction

The depth of anesthesia (DoA) is a crucial factor in anesthesiology, significantly impacting patient outcomes during surgical interventions. Monitoring DoA is necessary to make sure that analgesia, amnesia and muscle relaxation are achieved by patients which are the main targets of anesthesia. Anesthetizing agents should be delivered in right dosages in order to avoid excessive or insufficient anesthesia leading to some adverse effects like intraoperative awareness or hemodynamic instability [1, 2]. The complexity of anesthesia makes it essential to understand better how different levels affect physiological responses because different anesthetics can have an inhibitory effect on organ function whereby various clinical signs and symptoms may appear [1, 3]. The consciousness of patients must be managed very carefully during surgical interventions. The purpose of anesthesia is to provide a temporary suspension of consciousness so as not to feel pain nor remember anything during the procedure. However, clinicians often find it difficult to define accurately what is the right amount of anesthetic needed for total unconsciousness since the use of muscle relaxants hides vital signs such as breathing and movement [4]. For this reason, proper monitoring systems must be implemented for accurate anesthesia depth assessment and appropriate drug adjustments [5, 6].

DoA can be evaluated by a variety of methods. Traditional clinical assessments that are commonly used such as monitoring vital signs like heart rate and blood pressure or observing patient responses, tend to be subjective and unreliable [4]. Among more advanced techniques, heart rate variability analysis, isolated forearm technique (IFT) or lower esophageal contractility (LOC) assessments provide other insights into the patient's anesthetic state [7]. However these methods may not always perfectly correlate with the actual depth of anesthesia. Hence, there is still need for precise monitoring tools. One widely recognized and applied means for monitoring depth of anesthesia is the Bispectral Index (BIS). It is a non-invasive system that monitors brain activity measuring it within a numerical range of zero meaning no brain activity and hundred indicating full consciousness [8, 9]. It has been endorsed by the U.S. Food and Drug Administration (FDA) as a reliable indicator for anesthetic depth and several studies have proven its effectiveness in multiple surgical settings [10]. Continuous BIS monitoring allows anesthesiologists to titrate anesthetic agents thereby enhancing patients' safety and comfort during procedures [11].

Traditional methods of monitoring DoA, particularly through target-controlled infusion (TCI) models, primarily rely on pharmacokinetic-pharmacodynamic (PK-PD) models that predict drug concentration based on propofol dosage. However, these models often do not incorporate the BIS as a measurement index, leading to inconsistencies during various stages of anesthesia. This limitation highlights the need for more robust monitoring systems that can accurately reflect the patient's state throughout the anesthesia process [12].

In recent years, there has been a significant development in DoA monitoring models, particularly those based on electroencephalogram (EEG) signals. The BIS has emerged as a standard in this domain, providing a continuous assessment of the patient's brain activity during anesthesia. However, the intricacies involved in interpreting EEG signals, combined with the limited sensitivity of BIS to specific anesthetic agents, present significant challenges that may undermine the accuracy and reliability of DoA evaluations [12]. Furthermore, the high costs associated with advanced EEG monitoring technologies limit their accessibility, particularly in resource-constrained settings, underscoring the necessity for more affordable and effective solutions [12]. The application of machine learning (ML) techniques has been explored in modeling DoA and propofol dosage, yet many studies have overlooked the effects of adjunctive agents such as remifentanil. This oversight may lead to incomplete models that fail to capture the full pharmacodynamic interactions occurring during anesthesia [13]. Recent advancements in deep learning have shown promise in predicting DoA based on EEG signals and fluid administration history. For instance, studies by Zhou et al. [14] and Abel et al. [15] have developed deep learning models that aim to enhance the accuracy of DoA predictions. However, these models have encountered difficulties, particularly during the anesthesia induction and recovery stages, where physiological responses can vary significantly among patients.

The incorporation of deep learning techniques into DoA monitoring marks a substantial leap forward compared to conventional methods. By utilizing extensive datasets and sophisticated algorithms, these models hold the potential to deliver more accurate DoA predictions, effectively addressing the diverse responses patients exhibit to anesthetic agents. For example, Chen et al.'s work on a deep learning framework for anesthesia depth prediction emphasizes the challenges posed by individual physiological differences, which can lead to inconsistent pharmacodynamic responses during anesthesia [16]. Moreover, the exploration of EEG features through ML techniques has been shown to classify unconsciousness effectively, providing a more nuanced understanding of the relationship between drug administration and brain activity [15].

Despite these advancements, the field still faces challenges, particularly in ensuring the generalizability of ML models across diverse patient populations and clinical scenarios. The need for robust validation of these models

is crucial, as the consequences of inaccurate DoA predictions can be severe, including the risk of awareness during surgery or inadequate anesthesia leading to patient distress [17]. Furthermore, the interaction between different anesthetic agents, such as propofol and remifentanil, necessitates a comprehensive understanding of their combined effects on brain activity, which current models may not fully address [13, 17]. To overcome these limitations, in this paper, we propose a computational framework that integrates the Long Short-Term Memory (LSTM) networks, Transformer architectures, and Kolmogorov-Arnold Networks (KAN). This hybrid AI approach, combining diverse machine learning models and deep learning architectures, has demonstrated promising outcomes across diverse applications in healthcare and biomedicine [18–20].

The LSTM networks [21] are particularly well-suited for handling sequential data, as they are designed to remember information over extended periods. In the context of drug infusion history, where the timing and dosage of medications can significantly influence patient outcomes, LSTMs are highly effective at capturing temporal dependencies. By retaining critical information about previous drug administrations, they enable the model to make informed predictions about the current depth of anesthesia. This capability is crucial, as the effects of anesthesia drugs can be.

Transformers [22], renowned for their proficiency in capturing data relationships via attention mechanisms, augment LSTM by strengthening the model's ability to interpret the context of drug infusion histories. The attention mechanism allows the model to focus on the most relevant parts of the input sequence, such as key drug doses or timing intervals, which are vital for assessing the depth of anesthesia. By utilizing the Transformer's capability to manage long-range dependencies, we can enhance the model's ability to comprehend intricate interactions among multiple drugs administered over extended periods.

KAN [23] introduces a layer of flexibility in modeling nonlinear relationships inherent in biological systems. Anesthesia depth is influenced by various factors, including patient-specific variables and drug interactions, which may not follow linear patterns. KAN's ability to approximate complex functions allows for a more nuanced understanding of how different drug infusion histories translate into varying depths of anesthesia. This feature is especially beneficial in medical settings, where personalized patient reactions can differ substantially.

By combining LSTM, Transformer, and KAN, our proposed model creates a robust framework that optimizes predictive performance while providing deeper insights into the underlying dynamics of anesthesia depth. LSTM captures the sequential nature of drug administration, the Transformer enhances contextual understanding through attention mechanisms, and KAN addresses the nonlinearities present in patient responses. This combined method not only enhances prediction precision but also enables a more thorough examination of the factors affecting anesthesia depth. Ultimately, this model aims to enhance patient safety and outcomes by providing anesthesiologists with reliable predictions based on detailed drug infusion histories.

Related work

Traditional methods for assessing DoA, such as the pharmacokinetic-pharmacodynamic (PK-PD) model, primarily utilize propofol doses to predict the effect-site concentration of the drug by Shalbaf et al. (2014) [24]. However, Huang et al. (2023) [25] highlighted the limitations of relying solely on propofol kinetics, arguing that a more comprehensive approach is necessary for accurate DoA assessment.

In recent years, the incorporation of machine learning and deep learning into anesthesia prediction models has transformed the field, enabling more accurate and personalized patient care. Taylor et al. (2016) [26] developed a hybrid approach integrating clustering and regression methods, employing electromyography (EMG) signals in conjunction with propofol infusion rates to forecast bispectral index (BIS) signals. This innovative approach demonstrates the potential of ML to enhance the accuracy of DoA predictions. Similarly, Zhou and Srinivasan (2021) [27] employed reinforcement learning to design a closed-loop anesthesia control system that uses BIS as a control parameter, while also incorporating mean arterial pressure (MAP) into the model. This dual consideration allows for more precise regulation of propofol infusion rates, ensuring that both BIS and MAP values remain within desired ranges. The advancements in ML have also led to the development of adaptive models that aim to improve the accuracy of anesthesia predictions. Mizuguchi and Sawamura (2023) [28] proposed a fuzzy logic-based adaptive model enhanced by a genetic algorithm, which seeks to provide more reliable predictions of DoA. Recently, Peng (2024) [29] focused on identifying optimal EEG features to classify DoA stages using an adaptive neuro-fuzzy inference system. Despite the promise of this method, earlier studies in this area were often constrained by limited experimental settings and small sample sizes, which may not adequately reflect the complexities of drug effects on DoA or the diversity of patient physiology. This concern is further underscored by Lee et al. [30], who developed a deep learning model that leverages infusion histories of propofol and remifentanil, along with patient characteristics, to predict BIS signals. Their model, trained on a substantial cohort of 231 subjects, offers a more robust framework for DoA prediction

compared to previous methodologies. Additionally, the integration of AI into clinical decision-making processes has been shown to improve outcomes in anesthesia, as highlighted by Hashimoto et al. [31]. Their review identified six key themes of AI applications in anesthesiology, including depth of anesthesia monitoring and event prediction. The potential of machine learning to enhance predictive analytics in anesthesia is further supported by studies such as those conducted by Kang et al. [32], who developed a prediction model for hypotension following anesthesia induction. Their results indicate that machine learning can substantially enhance prediction accuracy when compared to conventional logistic regression approaches. This is particularly relevant in the context of intraoperative management, where timely and accurate predictions can mitigate risks associated with anesthesia. In 2021, Kim et al. (2021) [33] made significant progress by integrating convolutional neural networks (CNNs), LSTM networks, and attention mechanisms to develop an innovative framework for predicting DoA using EEG signals. This comprehensive methodology highlights the promise of deep learning in improving the precision and dependability of DoA evaluations in clinical settings. However, it is important to note that anesthesiologists traditionally manage DoA based on the pharmacological effects of anesthetic agents, which may not always align with the predictions generated by these advanced models.

Material and method

Dataset

Dataset description

The dataset utilized in the experiments is the VitalDB database [34], accessed on January 1, 2022. This publicly available dataset is curated and managed by the Department of Anesthesiology and Pain Medicine at Seoul National University Hospital, affiliated with Seoul Metropolitan Medical College, located in Seoul, South Korea. It encompasses comprehensive information from 6,388 surgical patients, including demographic details such as height, weight, sex, and age. Additionally, it features over 60 clinical indicators related to operating room equipment, including patient monitors, anesthesia machines, BIS monitors, target-controlled infusion pumps, cardiac output monitors, and local oximeters. The dataset also includes BIS values and signal quality indices collected by BIS VISTA at one-second intervals, along with cumulative infusion volumes, effect-site concentrations (Ce), and plasma concentrations (Cp) of propofol and remifentanil, measured at the same one-second intervals by target-controlled infusion pumps.

In this study, we selected BIS as the measurement to calculate DoA, as described in Fig. 1. In the VitalDB public dataset, BIS is the exclusive measure of anesthesia depth. In addition, when compared to other DoA-type metrics (such as the Narcotrend index and Patient State index), BIS offers several significant benefits. It is the most prevalent indicator of anesthesia depth and has received FDA approval for use as a monitoring device for anesthetic effects on the brain. Moreover, BIS

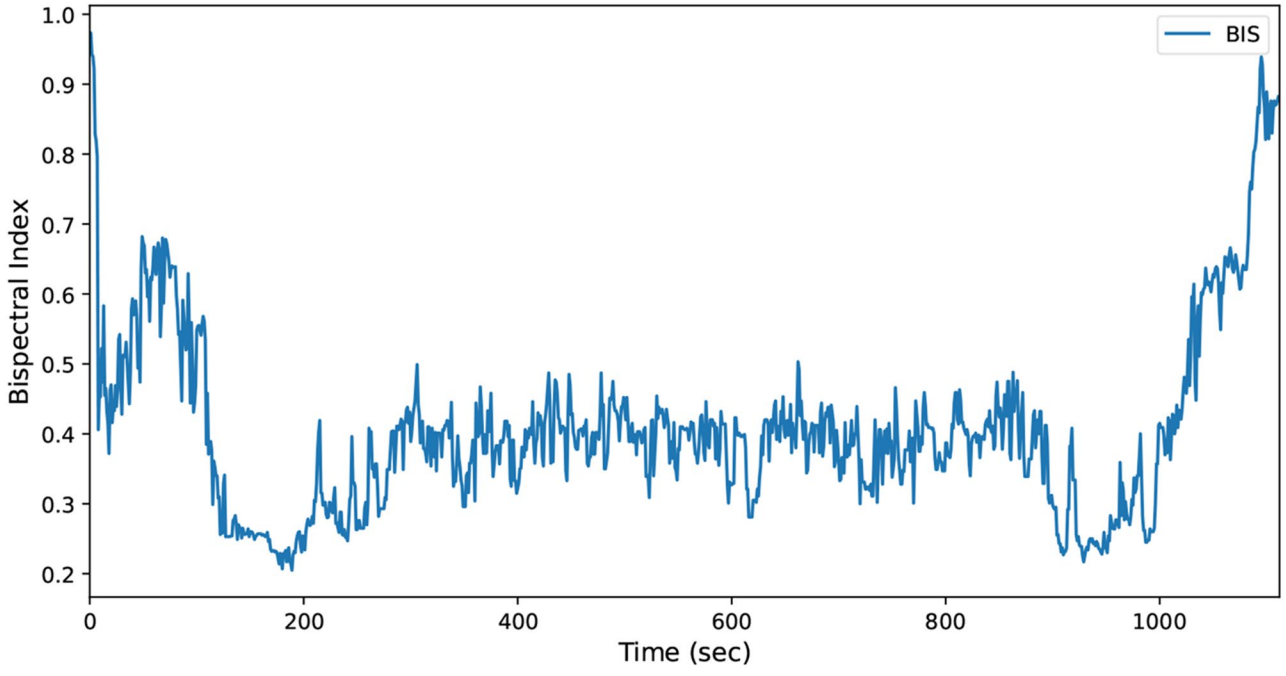

Fig. 1 Visualization of BIS for CaseID 1210

demonstrates a strong correlation with the sedative properties of various anesthetic agents, effectively reflecting their sedation levels, especially for frequently used drugs like propofol and sevoflurane.

Data processing

The data processing begins with initializing lists to store the doses of propofol and remifentanil, as well as patient demographic information (age, gender, height, weight), case identifiers (c), and the output BIS values (y). Data loading occurs for each case ID, extracting vital signs every 10 seconds. Missing values are filled using forward fill methods, and any initial missing data is replaced with zeros. Cases that do not involve drug infusion or where BIS values are absent are excluded from further analysis. The starting points for drug infusion are identified and previous data points are discarded. The volume of drug infusion is converted to a rate, ensuring that any negative values are set to zero, with additional conditions to invalidate rates exceeding 10 mL per 10 seconds.

Subsequent filtering steps exclude cases where the first BIS value is less than 80 and where the last BIS value drops below 70. For valid cases, the data is padded with zeros for the timepoints leading up to drug infusion. Patient demographics information are extracted and associated with the respective case IDs. Finally, input values for propofol and remifentanil doses are appended to the datasets with a time window of 180 timepoints, equivalent to 1800 seconds, alongside the normalized BIS values. Subsequently, we randomly selected 100 patients and divided them into training, validation, and testing datasets, comprising 91, 9, and 10 cases, respectively. The characteristics of these three datasets are presented in Table 1.

Methodology

Long short-term memory

The LSTM network [21] is an advanced variant of the Recurrent Neural Network (RNN), characterized by recurrent connections within its hidden layers. Its architecture incorporates a feedback mechanism across multiple layers, allowing it to effectively capture nonlinear temporal dependencies in time series data. LSTMs were specifically designed to address the vanishing and exploding gradient problems inherent in traditional RNNs. This is achieved through an external feedback loop that integrates the hidden state from the previous time step into the network's inputs, influencing subsequent predictions.

Table 1 Description of patien characteristics, mean ± standard deviation (min-max)

Characteristics	Training Dataset	Validation Dataset	Test Dataset
Number of Cases	91	9	10
Number of Samples	95189	11082	13040
Age (yr)	61.7 ± 11.2 (33–81)	56.0 ± 12.3 (33–72)	64.3 ± 6.3 (54–75)
Sex (male/female)	35/46	6/3	4/6
Weight (kg)	63.8 ± 11.2 (39–98)	61.4 ± 11.8 (42–76)	58.3 ± 9.3 (45–73)
Height (cm)	163.1 ± 8.3 (145–186)	159.7 ± 8.4 (148–177)	160.3 ± 8.2 (149–175)

At the core of the LSTM is the memory cell, a fundamental component of its internal feedback loop. This cell serves as an independent storage unit, preserving temporal information over extended periods and mitigating gradient-related issues faced by standard RNNs.

A typical LSTM unit consists of a memory cell and three key control gates: the input gate, output gate, and forget gate. Let x_t and h_t represent the input and hidden state at time t, respectively. The gates are defined as f_t (forget gate), i_t (input gate), and o_t (output gate), while $\tilde{H}_t$ represents the candidate information to be stored. The input gate regulates how much new information is retained, and the transformations governing these gates, as well as the cell state and hidden state, are described by the following equations:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f), \tag{1}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i), \tag{2}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o), \tag{3}$$

$$\tilde{H}_t = \tanh(W_H \cdot [h_{t-1}, x_t] + b_H), \tag{4}$$

$$H_t = f_t \cdot H_{t-1} + i_t \cdot \tilde{H}_t, \tag{5}$$

$$h_t = o_t \cdot \tanh(H_t). \tag{6}$$

Transformer

Unlike conventional architectures that depend on RNNs, the Transformer [22] utilizes a self-attention mechanism, removing the need for sequential data processing. This advancement enables the model to process input data in parallel, significantly boosting its efficiency and allowing it to capture global dependencies within the dataset. The Transformer is structured with multiple encoder and decoder components. The encoder, made up of several stacked layers, converts raw input data into a structured representation. The decoder then uses this encoded information to produce the desired output. A key feature of the encoder is its multi-head self-attention mechanism, which enables the model to identify and utilize dependencies across both short-term and long-term metrics. By simultaneously focusing on different parts of the input sequence, the model extracts and emphasizes

critical features, enhancing its ability to understand complex patterns.

The self-attention mechanism operates using three core matrices: Q (query), K (key) and V (value). These matrices interact to calculate the relationships between elements in the sequence. The dimensionality of the key vector is represented as d_K, and the relationships are mathematically defined as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_K}}\right)V. \quad (7)$$

The Transformer excels at contextual comprehension, equipping it with distinctive capabilities for tackling temporal data forecasting challenges.

Kolmogorov-Arnold network

KAN, as described by Liu et al. (2024) [23], consists of two distinct types of functions: the inner functions $\phi_{q,p}$ and the outer functions Φ_q. The inner functions, which operate on individual input variables x_p, serve as the initial layer of the network. They take single-variable inputs and convert them into intermediate representations, processing each feature independently. Conversely, the outer functions aggregate the outputs generated by the inner functions. This second layer synthesizes the intermediate values through a weighted summation of the inner function outputs, ultimately producing the final predictions.

In mathematical terms, a multivariate continuous function f is represented as follows:

$$f(x) = \sum_q \Phi_q \left(\sum_p \phi_{q,p}(x_p) \right), \quad (8)$$

where $\phi_{q,p}$ denotes the inner functions, while Φ_q represents the outer functions that utilize the processed information from the inner layer.

Proposed model

In this research, we present an advanced model for predicting DoA, leveraging a combination of LSTM, Transformer and KAN, as shown in Fig. 2. The model integrates three input sources: Propofol Dose (feature size 180), Remifentanil Dose (feature size 180), and AGWH (Age, Gender, Weight, Height) (feature size 4).

Initially, the Propofol Dose and Remifentanil Dose inputs are processed through their respective LSTM layers to effectively capture temporal dependencies and patterns over time. Each LSTM layer is configured as follows: Input layer (180, 1), LSTM layer (activation = Tanh, recurrent_activation = Sigmoid, return_sequences = true). Both LSTM layers are designed to enhance the model's robustness against noise and missing data, with a dropout layer (0.1) applied to prevent overfitting.

Simultaneously, the AGWH input undergoes a normalization layer to ensure consistent scaling and distribution of the features. This normalization layer is configured with epsilon is 1×10^{-6} and is set to scale = true, ensuring that the AGWH features are appropriately prepared for integration.

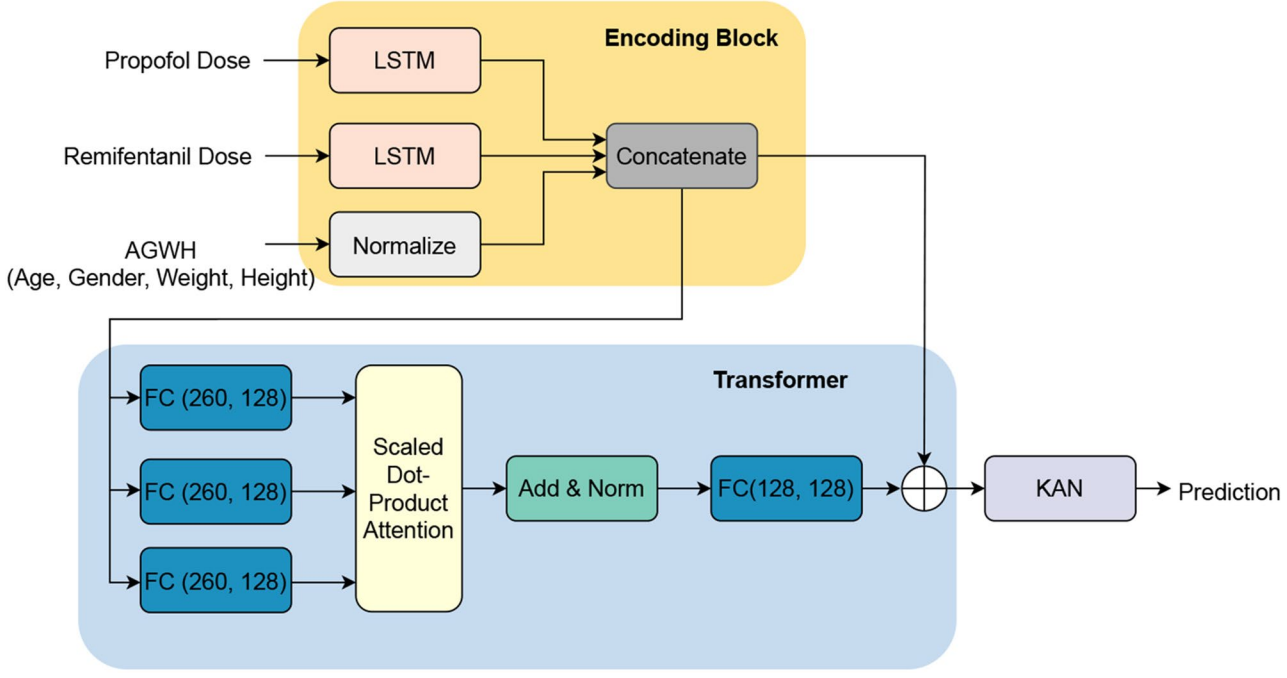

Fig. 2 Overview of the model architechture

The outputs from the three processing paths (the two LSTM networks and the normalized AGWH input) are then concatenated to form a combined feature vector of size 260. This concatenated vector is subsequently passed through a Transformer layer, which transforms the feature representation into a new vector of size 128. The self-attention mechanism in the Transformer enables the model to concentrate on the most significant elements of the feature sequence, improving its capacity to identify and interpret contextual relationships within the input data.

Finally, the output vector of size 128 is fed into KAN, which captures complex nonlinear relationships to predict the DoA.

Experiments

Setup

In the experiments, our proposed model is implemented using PyTorch 2.0.0, running on a 12 GB NVIDIA RTX 3060 GPU. The network is trained with the Adam optimizer, starting with an initial learning rate of 0.001, and undergoes 100 epochs of training. A consistent batch size of 256 is employed throughout the training, validation, and testing phases. The complete training process is completed in under 15 minutes, with a batch size of 256.

Evaluation metrics

To assess the effectiveness of the models, we employed various evaluation metrics including Mean Squared Error (MSE), Mean Directional Percentage Error (MDPE (%)), Mean Absolute Error (MAE), Root Mean Square Deviation (RMSE). Their mathematical formulas are described as:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{9}$$

$$\mathrm{MDPE}(\%) = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i - \hat{y}_i}{y_i} \times 100, \tag{10}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{11}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{12}$$

Table 2 Hyperparameters in machine learning algorithms

Model	Hyperparameters
Random Forest (RF)	Number of estimators: [100, 200, 300] Max features: [1, 10, 'log2', 'sqrt'] Criterion: squared error
Logistic Regression (LR)	Regularization: [L1, L2] C: [0.01, 0.1, 1, 10]
Naive Bayes (NB)	Model: [Gaussian, Multinomial] Laplace smoothing: [True, False]
AdaBoost (ADB)	Number of estimators: [100, 200, 300]
Gradient Boosting (GB)	Number of estimators: [100, 200, 300] Learning rate: [0.01, 0.1, 0.2] Max depth: [3, 5, 7]
XGBoost (XGB)	Number of estimators: [100, 200, 300] Learning rate: [0.01, 0.1, 0.2] Max depth: [3, 5, 7]

Baselines and benchmarking models

In this section, we conduct experiments to measure the performance of models, we initially compare our model's performance with several widely-used machine learning (ML) techniques considered as baseline models including Random Forest (RF) [35], Logistic Regression (LR) [36], Naive Bayes (NB) [37], and AdaBoost (ADB) [38], Gradient Boosting (GB) [39], and XGBoost (XGB) [40]. To ensure competitive comparison results, we conducted parameter tuning for these machine learning models. The relevant parameters for each model are described in Table 2.

In addition, to comprehensive performance evaluation, we compared our proposed model against several cutting-edge deep learning techniques considered as benchmarking models. We aim to highlight the effectiveness of our architecture and its potential benefits in predicting DoA through a systematic comparison with these models. These methods we examined include:

- **LSTM** [21]: LSTM networks are a specialized form of recurrent neural network engineered to model long-range dependencies in sequential data. LSTMs utilize memory cells and gating mechanisms to regulate information flow, enabling them to preserve context across lengthy sequences. This feature makes LSTMs highly effective for applications such as time series forecasting, speech recognition, and natural language processing. In our experiments, we additionally implemented an LSTM model enhanced with the Attention mechanism.
- **GRU** [41]: Gated Recurrent Unit (GRU) networks are a variant of LSTM that simplifies the architecture while retaining similar performance. GRU combines the forget and input gates into a single update gate, which streamlines the model and reduces computational complexity. This efficiency makes GRUs a popular choice for sequence modeling tasks, offering robust performance in various applications.
- **AdaRNN** [42]: AdaRNN is an adaptive recurrent neural network that learns a dynamic model through two key modules: Temporal Distribution

Characterization (TDC) and Temporal Distribution Matching (TDM). TDC segments the training data into diverse periods with significant distribution gaps, enhancing the understanding of temporal dynamics. Following this, TDM utilizes an RNN-based approach to minimize distribution divergence, resulting in improved time series forecasting capabilities.

- **Transformer** [22]: The Transformer architecture revolutionizes sequence modeling by employing self-attention mechanisms to process input data in parallel rather than sequentially. This design enables the model to capture complex dependencies across different parts of the input, making it highly effective for tasks in natural language processing, computer vision, and more. The Transformer's ability to scale and learn contextual representations has led to state-of-the-art performance in various applications.
- **FEDformer** [43]: FEDformer, or Frequency Enhanced Decomposed Transformer, combines seasonal-trend decomposition with the Transformer architecture to improve long-term time series forecasting. By breaking down time series into seasonal and trend components, FEDformer captures the overall global profile while utilizing Transformers to analyze detailed structures. This approach demonstrates enhanced efficiency and performance, achieving notable reductions in prediction error compared to conventional models.
- **Crossformer** [44]: Crossformer is a Transformer-based model specifically designed for multivariate time series forecasting. It emphasizes cross-dimension dependency by employing a Dimension-Segment-Wise (DSW) embedding to preserve both time and dimension information. The model utilizes a Two-Stage Attention (TSA) mechanism to capture dependencies across time and dimensions efficiently. Crossformer's architecture allows it to leverage hierarchical information, resulting in superior forecasting performance on various datasets.

Table 3 Performance results of the baseline models on the test dataset

Model	MAE	MSE	RMSE	MDPE (%)
Random Forest (RF)	0.0735	0.0090	0.0950	6.2435
Logistic Regression (LR)	0.0769	0.0100	0.0998	2.2953
Naive Bayes (NB)	0.0766	0.0099	0.0996	3.4667
AdaBoost (ADB)	0.0749	0.0094	0.0970	2.3463
Gradient Boosting (GB)	0.0714	0.0085	0.0923	3.1513
XGBoost (XGB)	0.0741	0.0091	0.0955	1.5293
Ours	**0.0620**	**0.0065**	**0.0808**	**0.6254**

Ablation study

In this section, we systematically evaluate the contribution of each component in our proposed hybrid model, which integrates LSTM, Transformer, and KAN. We perform a series of experiments where we individually remove each component from the model and assess the resulting performance metrics. First, we eliminate the LSTM layer and analyze how the absence of sequential data processing affects the overall accuracy and robustness of the model. Next, we exclude the Transformer component, which is responsible for capturing long-range dependencies, to understand its impact on the model's performance in handling complex relationships within the data. Finally, we remove the KAN component, which is designed to enhance the model's capacity for learning intricate patterns, allowing us to evaluate its significance in improving predictive capabilities. By comparing the performance of these ablated models against the full model, we aim to provide insights into the effectiveness of each component and their synergistic contributions to the overall performance.

Results and discussion

Performance of baseline models

Our results, as summarized in Table 3, demonstrate a marked improvement over traditional machine learning models in terms of various error metrics.

The MAE for our proposed model is 0.0620, significantly lower than that of the baseline models, such as RF at 0.0735 and GB at 0.0714. This reduction in MAE illustrates the enhanced capability of our architecture to accurately predict anesthesia depth, which is critical in clinical settings to ensure patient safety and optimal outcomes.

Furthermore, our model achieved a MSE of 0.0065 and an RMSE of 0.0808, outperforming all other models assessed in this study. The MSE and RMSE metrics are particularly relevant as they penalize larger errors more significantly, indicating that our model not only minimizes average error but also improves reliability in predicting depth fluctuations during anesthesia.

One of the most compelling aspects of our findings is the percentage of MDPE, which stands at 0.6254% for our model. This strong performance shows that our design keeps errors relatively low, which is important for tasks that need high accuracy. In contrast, the best-performing baseline model, XGB, recorded an MDPE of 1.5293%, underscoring the advantages of our approach.

The integration of LSTM and Transformer components within our architecture allows for the effective handling of sequential data and the capturing of long-range dependencies within the anesthesia monitoring signals. The KAN further enhances our model's ability to approximate complex non-linear relationships, contributing to

its superior performance. While traditional models such as RF, LR, and ADB achieved respectable results, they fell short of the accuracy and robustness demonstrated by our proposed method. The limitations of these models often stem from their reliance on handcrafted features and their inability to capture temporal dynamics inherent in the anesthesia data.

Overall, our findings advocate for the adoption of advanced hybrid models like the LSTM-Trans-KAN architecture in predicting anesthesia depth. Future research should explore the potential of this model in diverse clinical environments and consider its scalability across different patient demographics and anesthesia types. This exploration will not only validate the model's effectiveness but also contribute to the ongoing advancements in anesthesia monitoring technology, ultimately enhancing patient safety and care quality.

Performance of deep learning models

The results summarized in Table 4 highlight the superior performance of our proposed LSTM-Transformer-KAN model in predicting DoA compared to several established models. Notably, our model achieved a MAE of 0.0620, which is the lowest among all models evaluated. This significant reduction in MAE indicates a high level of accuracy in the model's predictions, suggesting that our architecture effectively captures the intricate patterns inherent in anesthesia depth data.

In terms of MSE, our model recorded 0.0065, again outperforming all competitors. This reduction in MSE demonstrates the robustness of our methodology and reinforces the model's capability to minimize prediction errors. The RMSE achieved by our model is 0.0808, which is the lowest among those tested. This metric is particularly important as it penalizes larger errors more harshly, further confirming the reliability of our approach in producing accurate depth predictions. Such precision is crucial in anesthesia management, where even minor discrepancies can have significant consequences.

Additionally, our model's MDPE of 0.6254% underscores its effectiveness in maintaining close alignment with actual depth values during predictions. This metric indicates that the model consistently provides estimates that are not only accurate but also reliable for practical applications in real-time monitoring. Comparing our results to other models such as Crossformer (MAE: 0.0653, MSE: 0.0071, RMSE: 0.0843, MDPE: 1.6298) and FEDformer (MAE: 0.0658, MSE: 0.0072, RMSE: 0.0849, MDPE: 2.1030), it is evident that our LSTM-Transformer-KAN model consistently outperforms these alternatives across all metrics. This highlights the enhanced capability of our architecture to leverage temporal dependencies and contextual information effectively.

To visualize the results, we visualize the outcomes of our proposed model and benchmarking models in predicting the BIS for case IDs 1210 and 1392 in Fig. 3. The results show that our model provides predictions that are closer to the actual BIS values compared to other models.

The demonstrated performance of our model not only emphasizes its potential for practical implementation in anesthesia monitoring systems but also suggests avenues for further research into hybrid modeling approaches that can enhance predictive accuracy in dynamic clinical environments.

Table 4 Performance results of benchmarking models on the test dataset

Model	MAE	MSE	RMSE	MDPE (%)
LSTM	0.0707	0.0084	0.0916	2.7481
GRU	0.0703	0.0082	0.0906	3.0585
AdaRNN	0.0674	0.0075	0.0865	1.4002
LSTM with Attention	0.0665	0.0073	0.0855	1.9850
Transformer	0.0700	0.0082	0.0908	4.1341
FEDformer	0.0658	0.0072	0.0849	2.1030
Crossformer	0.0653	0.0071	0.0843	1.6298
Ours	**0.0620**	**0.0065**	**0.0808**	**0.6254**

Impact of time window length on model performance

The results indicate that the 1800-second time window (180 timepoints) provides the best performance for the proposed model, achieving the lowest errors across all metrics (MAE = 0.0620, MSE = 0.0065, RMSE = 0.0808, MDPE = 0.6254). When using a shorter time window of 900 seconds, performance declines (MAE = 0.0652, RMSE = 0.0837), likely due to insufficient historical data to capture the full effects of drug infusion, leading to increased sensitivity to short-term fluctuations. Conversely, extending the time window to 3600 seconds slightly worsens performance (MAE = 0.0637, RMSE = 0.0824), potentially due to redundant information and increased noise, which may reduce model efficiency. The 1800-second window strikes an optimal balance, providing enough historical context to capture meaningful patterns while avoiding excess complexity. These findings suggest that selecting an appropriate time window is crucial for predictive accuracy, and future work could explore adaptive approaches to dynamically adjust time windows based on patient-specific responses.

Findings from the ablation study

In our discussion of the results presented in Fig. 4 for predicting the depth of anesthesia, we observe notable differences in performance across the various model configurations. The model excluding the LSTM component recorded a MAE of 0.0725 and an RMSE of 0.0929. This indicates that the lack of LSTM, which is adept at capturing temporal dependencies in sequential data, adversely

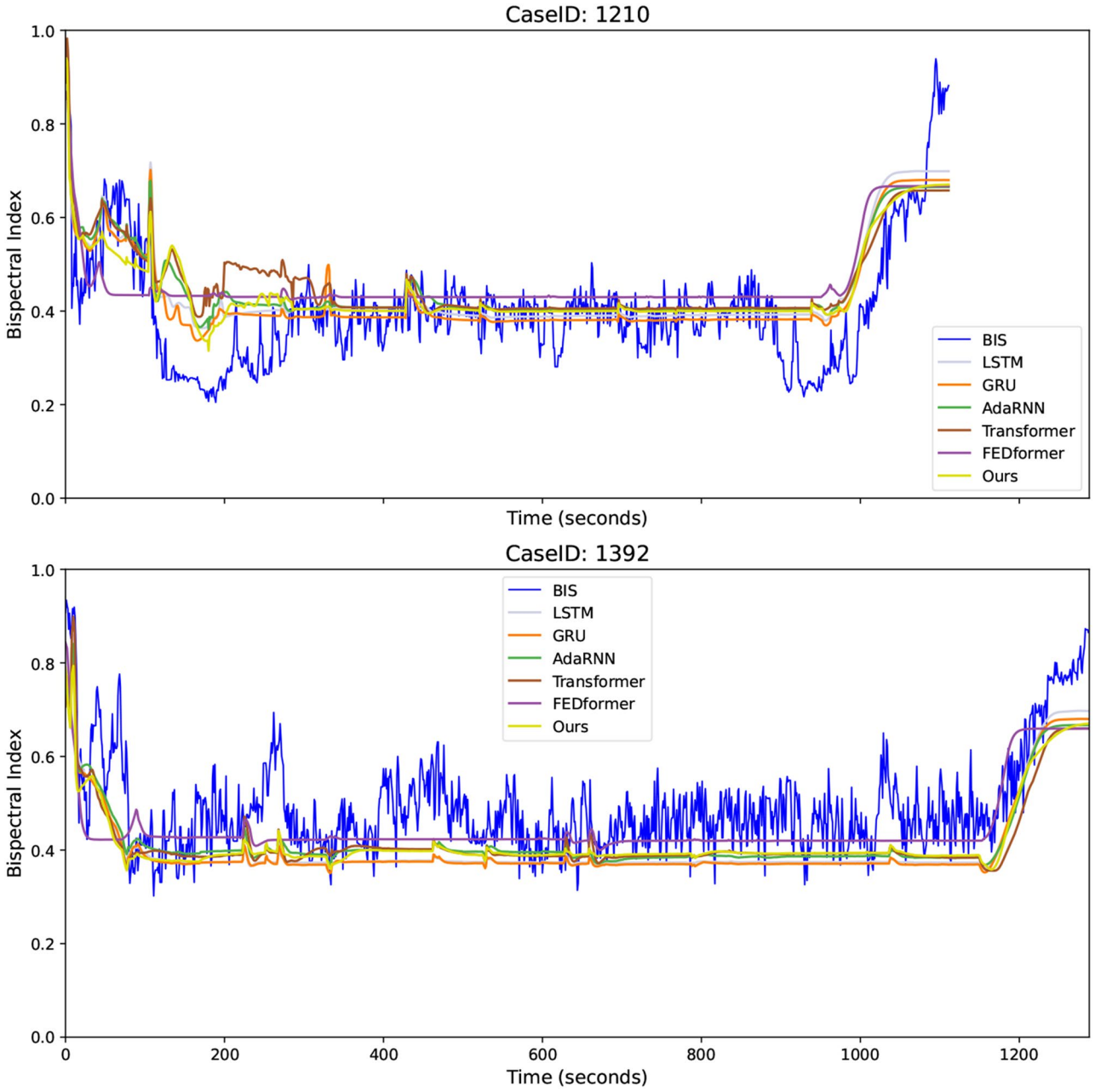

Fig. 3 Visualization of benchmarking models and our architecture for predicting the BIS for two cases with IDs of 1210 and 1392

affects the model's ability to accurately predict anesthesia depth.

When we removed the Transformer component, the model's performance further declined, resulting in a MAE of 0.0735 and an RMSE of 0.0950. This outcome emphasizes the importance of the Transformer's capability to manage long-range dependencies, which is crucial in understanding the complex interactions of physiological signals over time. In addition, the model without the KAN component performed slightly better, achieving a MAE of 0.0682 and an RMSE of 0.0872. This suggests that while KAN contributes to the model's learning capacity, its absence does not lead to a significant degradation in performance compared to the other configurations. In contrast, our proposed model, integrating LSTM, Transformer, and KAN, achieved the best results, with a MAE of 0.0620, MSE of 0.0065, and RMSE of 0.0808. The MDPE of 0.6254 further illustrates the model performance in accurately predicting the depth of anesthesia.

These findings highlight the complementary roles of each component in our hybrid model, reinforcing the notion that the combination of LSTM, Transformer, and KAN is essential for effectively addressing the complexities involved in anesthesia depth prediction. This

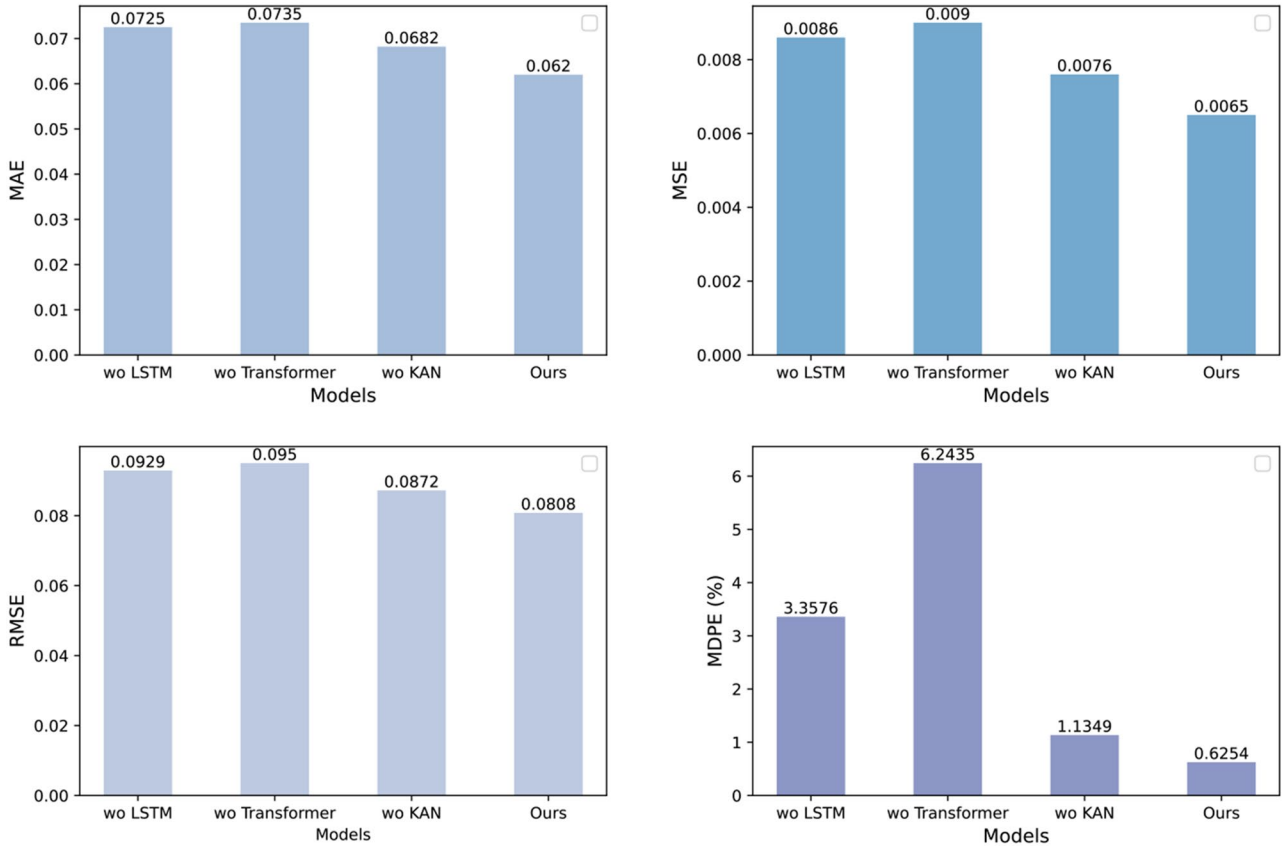

Fig. 4 Performance visualization on the test dataset in ablation study

comprehensive approach not only enhances predictive accuracy but also demonstrates the model's robustness in clinical settings.

Limitations and future work

While the proposed hybrid model leveraging LSTM, Transformer, and KAN demonstrates significant advancements in predicting anesthesia depth, it is not without limitations. One primary concern is the model's reliance on the quality and comprehensiveness of the input data. Although our dataset encompasses diverse patient scenarios, it may not fully represent the wide range of individual responses to anesthesia across different demographics and clinical conditions. This could introduce biases that affect the model's predictive accuracy in real-world applications.

Moreover, the complexity of the model, stemming from the integration of LSTM, Transformer, and KAN, raises practical challenges. The computational demands during training and inference may limit accessibility for healthcare institutions with constrained resources. To address this, future work could explore optimization strategies such as model pruning or quantization, which would enhance efficiency without significantly compromising performance.

Another limitation involves the interpretability of the model's predictions. As with many deep learning architectures, understanding the decision-making process remains a challenge. Enhancing model explainability through techniques like attention visualization could provide anesthesiologists with insights into how various factors contribute to predictions, thereby fostering trust and facilitating informed clinical decisions.

Looking ahead, several promising directions can be pursued to enhance the proposed framework. First, incorporating additional contextual variables–such as patient comorbidities and real-time physiological data–could significantly improve prediction accuracy and better adapt to the dynamic demands of anesthesia management. Second, adopting a multi-modal approach that integrates diverse data sources, including electronic health records and continuous monitoring systems, would offer a more comprehensive understanding of the factors influencing anesthesia depth. Third, future research should prioritize validating the model across a wide range of clinical environments and patient populations to ensure its robustness and generalizability. Finally, engaging with anesthesiology practitioners to gather insights and feedback could drive iterative refinements,

ultimately resulting in a model that more effectively addresses real-world clinical challenges.

Conclusion

This study introduces a comprehensive predictive model for assessing anesthesia depth based on drug infusion histories by integrating LSTM, Transformer, and KAN architectures. The results demonstrate that this hybrid approach not only significantly improves prediction accuracy but also offers valuable insights into the multifaceted factors that influence anesthesia depth. By providing anesthesiologists with reliable predictions, this model aims to enhance patient safety and outcomes, ultimately contributing to more informed decision-making in clinical settings. The advancement of such integrative frameworks underscores the potential of modern machine learning techniques in optimizing anesthesia management and improving overall patient care.

Author contributions

L.W. and Y.W. wrote the main manuscript text and W.Y. edited it. W.Y. designed the experiments and L.W. and Y.W performed the experiments including writing the code and preparing the figures. All authors reviewed the manuscript.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability

The VitalDB dataset used in this study is from Lee et al. [34] and can be accessed via https://vitaldb.net.

Code availability

The code used in this study is available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 26 February 2025 / Accepted: 24 March 2025
Published online: 08 April 2025

References

1. Xia L-Y, Zhang Q, Zhuo M, Deng Z-H, Huang K-N, Zhong M-L. Effects of different anesthetic depths monitored by Narcotrend on glandular secretion in patients undergoing laparoscopic total hysterectomy. 2022. Preprint at Research Square:rs-1508610/v1.
2. Lee TY, Kim MA, Eom DW, Jung JW, Chung CJ, Park SY. Comparison of remimazolam-remifentanil and propofol-remifentanil during laparoscopic cholecystectomy. Anesth Pain Med. 2023;18(3):252–59.
3. Xia L-Y, Zhang Q, Zhuo M, Deng Z-H, Huang K-N, Zhong M-L. Effects of different anesthetic depths monitored by processed EEG analysis on glandular secretion in patients undergoing laparoscopic total hysterectomy. Front Anesthesiol. 2023;2:1237970.
4. Ahmad T, Sheikh NA, Akhter N, Dar BA, Ahmad R. Intraoperative awareness and recall: a comparative study of dexmedetomidine and propofol in cardiac surgery. Cureus. 2017;9(8):1542.
5. Xuan H, Xu K. Warning and nursing experience of anesthesia depth monitoring for patients with general anesthesia delayed to leave anesthesia recovery room and delirium. Emerg Med Int. 2022;2022:1–5.
6. Eftekharian A, Amizadeh M, Mottaghi K, Safari F, Mahani MH, Ranjbar LA, Abdi A, Mokari N. Effect of depth of general anesthesia on the threshold of electrically evoked compound action potential in cochlear implantation. Europ Archiv Oto-Rhino-Laryngol. 2014;272(10):2697–701.
7. Liu Y, Lei P, Wang Y, Zhou J, Zhang J, Cao H. Boosting framework via clinical monitoring data to predict the depth of anesthesia. Technol Health Care. 2022;30:493–500.
8. Obata Y, Yamada T, Akiyama K, Sawa T. Time-trend analysis of the center frequency of the intrinsic mode function from the hilbert-huang transform of electroencephalography during general anesthesia: a retrospective observational study. BMC Anesthesiol. 2023;23:125.
9. Li Z, Cai J, Li J, Xu X, Zheng L. Comparative evaluation of the bispectral index (BIS) and BISpro during propofol anaesthesia. J Int Med Res. 2021;49(4).
10. Yu Y, Wang H, Wei L, Gao Y, Yan N, Chu J, Li H. Assessing the Index of Consciousness (IoC) as a monitoring tool for sedative effects of ciprofol in general anesthesia induction. 2024. Preprint at Research Square:rs-4622578/v1.
11. Smajic J, Praso M, Hodzic M, Hodzic S, Srabovic Okanovic A, Smajic N, Djonlagic Z. Assessment of depth of anesthesia: prst score versus bispectral index. Med Arch. 2011;65(4):216.
12. Nsugbe E, Connelly S, Mutanga I. Towards an affordable means of surgical depth of anesthesia monitoring: an EMG-ECG-EEG case study. BioMedInformatics. 2023;3(3):769–90.
13. Shalbaf A, Saffar M, Sleigh JW, Shalbaf R. Monitoring the depth of anesthesia using a new adaptive neurofuzzy system. IEEE J Biomed Health Inf. 2018;22(3):671–77.
14. Zhou P, Deng H, Zeng J, Ran H, Yu C. Unconscious classification of quantitative electroencephalogram features from propofol versus propofol combined with etomidate anesthesia using one-dimensional convolutional neural network. Front Med. 2024;11:1447951.
15. Abel JH, Badgeley MA, Meschede-Krasa B, Schamberg G, Garwood IC, Lecamwasam K, Chakravarty S, Zhou DW, Keating M, Purdon PL, Brown EN. Machine learning of EEG spectra classifies unconsciousness during GABAergic anesthesia. PLoS One. 2021;16(5):0246165.
16. Chen M, He Y, Yang Z. A deep learning framework for anesthesia depth prediction from drug infusion history. Sensors. 2023;23(21):8994.
17. Purdon PL, Sampson A, Pavone KJ, Brown EN. Clinical electroencephalography for anesthesiologists: part i: background and basic signatures. Anesthesiology. 2015;123(4):937–60.
18. Nguyen QH, Nguyen BP, Nguyen TB, Do TTT, Mbinta JF, Simpson CR. Stacking segment-based CNN with SVM for recognition of atrial fibrillation from single-lead ECG recordings. Biomed Signal Process Control. 2021;68:102672.
19. Nguyen L, Nguyen Vo T-H, Trinh QH, Nguyen BH, Nguyen-Hoang P-U, Le L, Nguyen BP. iANP-EC: identifying anticancer natural products using ensemble learning incorporated with evolutionary computation. J Chem Inf Model. 2022;62(21):5080–89.
20. Nguyen QH, Nguyen BP, Nguyen MT, Chua MCH, Do TTT, Nghiem N. Bone age assessment and sex determination using transfer learning. Expert Syst Appl. 2022;200:116926.
21. Hochreiter S, Schmidhuber J.Long short-term memory. Neural Comput. 1997;9(8):1735–80.
22. Vaswani A. Attention is all you need. Adv Neural Inf Process Syst. 2017;30:1–11.
23. Liu Z, Wang Y, Vaidya S, Ruehle F, Halverson J, Soljacic M, Hou TY, Tegmark M. KAN: Kolmogorov-Arnold networks. 2024. Preprint at arXiv:2404.19756IF: NA NA NA.
24. Shalbaf R, Behnam H, Jelveh Moghadam H. Monitoring depth of anesthesia using combination of EEG measure and hemodynamic variables. Cognit Neurodyn. 2014;9(1):41–51.
25. Huang H, Wang J, Zhu Y, Liu J, Zhang L, Shi W, Hu W, Ding Y, Zhou R, Jiang H. Development of a machine-learning model for prediction of extubation failure in patients with difficult airways after general anesthesia of head, neck, and maxillofacial surgeries. J Clin Med. 2023;12(3):1066.
26. Taylor RA, Pare JR, Venkatesh AK, Mowafi H, Melnick ER, Fleischman W, Hall MK. Prediction of in-hospital mortality in emergency department patients

with sepsis: a local big data-driven, machine learning approach. Acad Emerg Med. 2016;23(3):269–78.
27. Zhou Z, Srinivasan R. Detecting the state of drowsiness induced by propofol by spatial filter and machine learning algorithm on EEG recording. 2021. Preprint at bioRxiv:2021.07.12.452077.
28. Mizuguchi T, Sawamura S. Machine learning-based causal models for predicting the response of individual patients to dexamethasone treatment as prophylactic antiemetic. Sci Rep. 2023;13(1):1–10.
29. Peng J, Gorham TJ, Meyer BD. Predicting dental general anesthesia use among children with behavioral health conditions. JDR Clin Transl Res. 2024;10(1):7–15.
30. Lee H-C, Ryu H-G, Chung E-J, Jung C-W. Prediction of bispectral index during target-controlled infusion of propofol and remifentanil: a deep learning approach. Anesthesiology. 2018;128(3):492–501.
31. Hashimoto DA, Witkowski E, Gao L, Meireles O, Rosman G. Artificial intelligence in anesthesiology: current techniques, clinical applications, and limitations. Anesthesiology. 2020;132(2):379–94.
32. Kang AR, Lee J, Jung W, Lee M, Park SY, Woo J, Kim SH. Development of a prediction model for hypotension after induction of anesthesia using machine learning. PLoS One. 2020;15(4):0231172.
33. Kim JH, Kim H, Jang JS, Hwang SM, Lim SY, Lee JJ, Kwon YS. Development and validation of a difficult laryngoscopy prediction model using machine learning of neck circumference and thyromental height. BMC Anesthesiol. 2021;21(1):125.
34. Lee H-C, Park Y, Yoon SB, Yang SM, Park D, Jung C-W. VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients. Sci. Data. 2022;9(1):1–10.
35. Breiman L.Random forests. Mach Learn. 2001;45(1):5–32.
36. Das A. Logistic regression. In: Maggino F. (eds) Encyclopedia of Quality of Life and WellBeing Research. Netherlands: Springer; 2014. p. 3680–82.
37. Webb GI. Naïve Bayes. In: Sammut C, Webb GI (eds) Encyclopedia of Machine Learning. US: Springer; 2011. p. 713–14.
38. Schapire RE. Explaining AdaBoost. In: Schölkopf B, Luo Z, Vovk V (eds) Empirical Inference. Berlin Heidelberg: Springer; 2013. p. 37–52.
39. Friedman JH.Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29(5):1189–232.
40. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM; 2016. p. 785–94.
41. Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. 2014. Preprint at arXiv:1412.3555IF: NA NA NA.
42. Du Y, Wang J, Feng W, Pan S, Qin T, Xu R, Wang C: Adarnn: adaptive learning and forecasting of time series. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. Queensland: ACM; 2021. p. 402–11.
43. Zhou T, Ma Z, Wen Q, Wang X, Sun L, Jin R. FEDformer: frequency enhanced decomposed transformer for long-term series forecasting. 2022. Preprint at arXiv:1710.10903IF: NA NA NA.
44. Zhang Y, Yan J. Crossformer: transformer utilizing cross-dimension dependency for multivariate time series forecasting. In: The Eleventh International Conference on Learning Representations. 2023.

Publisher's Note