IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 June 2024, accepted 16 August 2024, date of publication 26 August 2024, date of current version 4 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3449704

RESEARCH ARTICLE

XAI-VSDoA: An Explainable AI-Based Scheme Using Vital Signs to Assess Depth of Anesthesia

NEERAJ KUMAR SHARMA[1], SAKEENA SHAHID[2], SUBODH KUMAR[3], SANJEEV SHARMA[4], NAVEEN KUMAR[5], TANYA GUPTA[5], AND RAKESH KUMAR GUPTA[6]

[1]Department of Computer Science, Ram Lal Anand College, University of Delhi, Delhi 110021, India
[2]Department of Computer Science, Sri Guru Tegh Bahadur Khalsa College, University of Delhi, Delhi 110021, India
[3]Department of Data Science and Analytics, Central University of Rajasthan, Ajmer, Rajasthan 305817, India
[4]Department of Anesthesia, Atal Bihari Vajpayee Institute of Medical Sciences and Dr. Ram Manohar Lohia Hospital, Delhi 110001, India
[5]Department of Computer Science, University of Delhi, Delhi 110021, India
[6]Department of Microbiology, Ram Lal Anand College, University of Delhi, Delhi 110021, India

Corresponding author: Subodh Kumar (subodh.kumar@curaj.ac.in)

This work was supported by the Indian Council of Medical Research (ICMR), New Delhi, India, under the Extramural Research Program under Project 2021-13397.

ABSTRACT Administration of anesthesia is essential in surgical procedures, ensuring patient unconsciousness and safety. Traditional Depth of Anesthesia (DoA) assessment methods rely heavily on the clinical expertise of anesthesiologists and patient physiological responses, which can vary widely due to age, weight, and ethnicity. This variability poses significant challenges in maintaining appropriate anesthesia levels and making timely decisions in critical situations. To address these challenges, we propose XAI-VSDoA, an explainable AI model using vital signs designed to augment DoA assessment by providing accurate predictions and interpretable insights. In this work, we experimented with various machine learning classifiers, including XGBoost, CatBoost, LightGBM, Random Forest, ResNet, and Feed-forward Neural Networks. Among these, the XGBoost model achieved the highest accuracy, with 99.34% on the University of Queensland dataset and 93.07% on the VitalDB dataset. Statistical testing confirmed that XGBoost outperformed the other models. We employed explainable AI techniques such as LIME and SHAP to identify the top 10 features significantly influencing the model's predictions, ensuring the model's transparency and reliability. These methods consistently highlighted the same influential features, reinforcing the model's interpretability. Our proposed scheme demonstrated exceptional performance using numeric vital signs, with XAI techniques validating the key features. This interpretability boosts confidence in the model, enhancing its utility to augument and support the clininal observations of anethesiologiss in anesthesia management. Our findings underscore the potential of XAI-VSDoA as a valuable tool for clinical use, enhancing patient safety and decision-making in anesthesia.

INDEX TERMS Depth of anesthesia, Bispectral Index, local interpretable model-agnostic explanations (LIME), machine learning, SHapley Additive exPlanations (SHAP), explainable artificial intelligence (XAI), vital signs.

I. INTRODUCTION

The administration of anesthesia during surgery is an essential and crucial aspect of contemporary medical practice. It allows patients to undergo surgical procedures with minimal pain and discomfort while ensuring their safety throughout the operation. Depth of Anesthesia (DoA) is the degree to which a general anesthetic agent suppresses the Central Nervous System (CNS) [1]. Prudent management of anesthesia is crucial as its inappropriate dosages can lead to intraoperative and postoperative complications. An excessive administration of anesthesia can lead to cardiac and respiratory problems, delayed emergence from anesthesia, and an extended period of recuperation [2]. In contrast,

The associate editor coordinating the review of this manuscript and approving it for publication was M. Shamim Kaiser.

administering too little anesthesia can result in the patient regaining consciousness during surgical stimulus, inadequate pain relief, heightened stress levels, and the ability to remember the surgical procedure [3]. This inability to properly adjust the anesthetic depth may result from patient-specific drug requirements, patients' inability to tolerate sufficient anesthesia due to factors like poor cardiac function, masking of physiological indicators by certain medications, and compromised drug delivery systems due to equipment issues or misuse [4]. Consequently, inappropriate DoA can lead to patient discomfort and potential safety concerns. Despite constant monitoring of physiological parameters (vital signs) such as pulse, blood pressure, heart electrical activity, airway pressure, and gas concentrations, instances of awareness have been documented [3]. This signifies that the task is not inherently straightforward.

Anesthesiologists are responsible for assessing and maintaining an appropriate level of anesthesia throughout the perioperative period by adjusting the anaesthetic dose based on physiological parameters (for instance, heart rate and blood pressure) and other factors like lachrymation, movement, and response to verbal stimuli. Traditionally, anesthesiologists have mainly relied on their clinical expertise and measurements of physiological symptoms to ascertain DoA. However, dealing with large volumes of diverse data with multiple modalities is a practical challenge for anesthesiologists.

Fortunately, advancements in medical technology have led to the development of specialized monitoring devices explicitly designed to assess DoA. The most widely used device for assessing DoA is the Bispectral Index (BIS) monitor, which produces a BIS value between 0 to 100 [5]. The value is generated by analyzing electroencephalogram (EEG) signals [6]. A value of 0 signifies a lack of brain activity, while 100 indicates a state of wakefulness. Values below 40 indicate a profound state of hypnosis. BIS values greater than 60 indicate light sedation. BIS values ranging from 40 to 60 indicate appropriate levels of general anesthesia for surgical procedures that maintain a desirable equilibrium between unconsciousness and safety. Studies indicate that BIS monitoring reduces the amount of anesthesia used in surgeries, the likelihood of nausea and vomiting, and recovery room duration by a moderate amount. The BIS monitor is notably expensive for developing countries and poses a significant financial challenge for healthcare facilities and professionals seeking to integrate this technology into their practices. Another aspect contributing to the hesitant use of BIS monitor is the undisclosed nature of its underlying algorithm [7], [8]. Anesthesiologists, reliant on the BIS device to assess DoA, express concerns over their inability to discern the specific criteria influencing the BIS value. This can potentially impact the trust in the reliability of information guiding anaesthesia administration. It is worth mentioning that there are several other DoA monitoring devices available in the market, such as the Patient State Analyser 4000, Score of Neonatal Acute Physiology (SNAP) monitor, Central Function Analyzing Monitor (CFAM), Narcotrend Monitor, Cerebral State Monitor (CSM), Entropy-Module, Auditory Evoked Potential (AEP) monitor, and Index of Consciousness (IoC) monitor. However, similar to the BIS monitor, these devices are hard to adopt as a standard device because of their setup and maintenance cost, further fueled by the lack of explanations for the values displayed [9].

Artificial Intelligence (AI) has emerged as a practical approach to address the limitations of existing monitors for assessing the DoA. By leveraging large quantities of collected data, Machine Learning (ML) algorithms can identify patterns and connections that contribute to precise DoA evaluation. These algorithms can incorporate diverse inputs, including vital signs such as heart rate, blood pressure, blood oxygen saturation, respiratory rate, and bio-signals such as electroencephalogram (EEG) and electrocardiogram (ECG) signals with which they develop predictive models for DoA monitoring [10].

In the past, several research works have assessed the DoA using AI. Sadrawi et al. [11] employed artificial neural networks (ANNs) to estimate the DoA using multiple physiological parameters (or vital signs) as inputs. They utilized empirical mode decomposition (EMD) to separate the electroencephalography (EEG) signal from the noise. Following this, they trained their models using these refined signals alongside the average values of essential vital signs such as electromyography (EMG) and heart rate. Using the ANN model, they achieved a Mean Absolute Error (MAE) of 6.54 with a standard deviation of 6.69, thus outperforming the BIS monitoring system, which had an MAE of 12.31 with a standard deviation of 13.06. Zhan et al. [12] employed Deep Neural Network (DNN) to distinguish between different anesthesia states using the features derived from Heart Rate Variability (HRV). They extracted 4 features from HRV to be used as input and the assessment of consciousness level from clinical experts as output. Their trained DNN achieved an accuracy of 90.1%. Liu et al. [13] utilized a boosting framework algorithm for training a series of weak learners into strong learners by assigning different weights based on their classification accuracy. They used four types of clinical monitoring data, including EMG, end-tidal carbon dioxide partial pressure, remifentanil dosage, and flow rate as input and the expert's predicted levels of consciousness as the target to train different models such as decision tree, k-nearest neighbour, and support vector machine. Their scheme achieved a mean-squared error (MSE) of 0.06, a mean absolute error (MAE) of 0.16 and a higher R-square value of 0.94. Liu et al. [14] introduced the similarity and distribution Index (SDI) for assessing anaesthesia depth using heart rate variability (HRV). The SDI derived from 32-second HRV segments were modelled using Artificial Neural Networks. An ensemble ANN provided an effective depth of anesthesia assessment, and the SDI, computed from ECG recordings, robustly correlated with expert anaesthesiologists' evaluations. Subramanian et al. [15] examine autonomic dynamics during propofol sedation to differentiate induction from emergence, evaluate historical data's impact, and discern post-sedation baseline from pre-sedation. Data from eleven volunteers, comprising HRV indices, were analyzed. The results of logistic regression, LASSO for pruning, and 10-fold cross-validation show an AUC (Area Under the Curve) of 0.706 between induction and emergence. Later, Subramanian et al. [16] delved into autonomic changes during loss and regain of consciousness (LOC and ROC) in general anesthesia. They used multimodal autonomic indices: heart rate variability (HRV), blood pressure (BP), and electrodermal activity (EDA) in 9 volunteers who were under propofol sedation. They employed trained logistic regression models with LASSO regularization and leave-one-subject-out cross-validation. Their results indicate effective differentiation of pre- and post-LOC/ROC periods with an AUC of approximately 0.8. Table 1 summarises research works that have employed vital signs to assess DoA.

Several alternative input data have been studied in the existing literature that do not incorporate vital signs as primary inputs and opt for using artificial intelligence algorithms in evaluating DoA. Chowdhury et al. [17] conducted a study that employed deep learning techniques to predict the DoA by analyzing electrocardiogram (ECG) and photoplethysmography (PPG) signals. They converted the ECG and PPG signals into heat maps. Further, taking these heatmaps as input and the corresponding BIS value as the target, they trained various deep learning models, including VGG19, AlexNet, and 6, 8, and 10-layered Convolution Neural Networks (CNNs). Their experimental results reveal that their 10-layered CNN model achieved the highest accuracy of 86%. Bahador et al. [18] proposed a multimodal spatio-temporal-spectral information fusion-based technique to enhance the accuracy and reliability of deep learning models. They fused EEG time-frequency data with ECG signal data and utilized pre-trained deep learning models such as SqueezeNet, GoogLeNet, Inceptionv3, ResNet18, and AlexNet to determine the DoA. Their experimental results reveal that their SqueezeNet model achieved the best precision value of 94.14%. Liu et al. [19] employed the Short-Time Fourier Transform to convert the EEG signal into spectrograms. Further, taking spectrograms as input and the BIS value as output, they trained three CNN models: CifarNet, AlexNet, and VGGNet. Their scheme using VGGNet achieved an accuracy of 93.50%. Chen et al. [20] proposed a method for monitoring anesthetized patients during surgery, focusing on characterizing the state of consciousness using EEG analysis and compared the performance of two frequency analysis techniques, namely Empirical Mode Decomposition (EMD) combined with Fast Fourier Transform (FFT), and Hilbert-Huang Transform (HHT). The results indicate agreement between the two techniques. Gonzalez et al. [21] addressed the issue of adapting opioid infusion during anesthesia by evaluating the Analgesia Nociception Index (ANI) as a guidance variable. Using machine learning classifiers trained on data from 17 patients undergoing cholecystectomy, the inclusion of minimum ANI values significantly improved predictive accuracy, demonstrating its potential to outperform traditional signs and anticipate dose changes for preventing hemodynamic events. Wang et al. [22] proposed a model known as KRDGB-CNN, which combined several machine learning techniques, including K-Nearest Neighbors (KNN), Random Forest, Decision Tree, Gaussian Naive Bayes, and Backpropagation Neural Network (BP). These components were integrated within a convolutional neural network (CNN) framework to form the decision layers. Their model achieved an accuracy rate of 92.2%. Similarly, Anand et al. [23] investigated the relationship between EEG signals and the bi-spectral index over time. They extracted the time-domain features and trained various machine learning models to achieve the highest accuracy of 83%.

Using machine learning algorithms to monitor the depth of anesthesia provides a multifaceted approach that outperforms traditional procedures in multiple key aspects. First, machine learning is highly effective in dealing with the intricacies of anesthesia management such as the use of multiple drugs with varying pharmacokinetics. Machine learning can be used to understand the effect of the combination of EEG, heart rate, blood oxygen saturation, and blood pressure (and other physiological parameters) data to comprehensively understand anesthetic depth. Second, in contrast to conventional techniques that depend on explicit rules, machine learning adopts a data-centric approach, acquiring knowledge of patterns and relationships without pre-established thresholds. Third, as high-dimensional datasets become increasingly accessible, machine learning models become scalable and robust. However, machine learning is limited by the lack of explanations due to the opaque nature of complex machine learning models. This raises concerns among healthcare practitioners, particularly anesthesiologists. As a result, Explainable Artificial Intelligence (XAI) methodologies have been introduced to provide insights into the reasoning behind model predictions and enhance confidence in them.

A. MOTIVATION AND CONTRIBUTION

Existing research shows positive results in DoA assessment, but the lack of explanations behind complex machine learning model outcomes leads to scepticism among anesthesiologists. The research gaps leading to the motivation of this can be summarized as below:

- **High Cost and Limited Availability:** Costly and proprietary devices like BIS are not accessible in most medical centres, creating a significant gap in the availability of effective DoA monitoring tools.
- **Intractable Data Analysis:** Analyzing a large number of physiological parameters simultaneously using different monitoring devices can be intractable even for experienced anesthesiologists.
- **Complexity of Algorithms:** The complexity of the underlying algorithms makes it challenging to

TABLE 1. Literature review for DoA monitoring using physiological parameters as inputs.

Reference	Type of Input	Target	Technique Used	Results	Pros and Cons
[11]	EEG signal and Vital signs (Heart rate, Electromyography, pulse, systolic blood pressure, diastolic blood pressure), and Signal Quality Index	Averaged value of anaesthesia level from 5 doctors	Artificial Neural Networks	Mean Absolute Error (MAE) of 6.54 with a standard deviation of 6.69	*Pros:* Multivariable approach to assessing the DoA using multiple physiological signals. *Cons:* Requires sophisticated signal processing
[12]	Heart-rate variability derived features	Averaged value of anaesthesia level from 5 doctors	Deep Neural Networks	Accuracy of 90.1%	*Pros:* Utilized a combination of time-domain and frequency-domain features. *Cons:* Only four HRV-derived features were explored.
[14]	HRV similarity in segments of ECG data	Averaged value of anaesthesia level from 5 doctors	Artificial Neural Network	MAE of 6.314 with a standard deviation of 3.12	*Pros:* Results statistically validated through multiple tests. *Cons:* Cases with extremely low correlation with their index were discarded, which could indicate potential weakness in the model's applicability.
[13]	Electromyography, End-tidal Carbon Dioxide (ETCO2), remifentanil dosage, and flow rate	BIS values	Boosting frameworks (Gradient boosting Regressor and AdaBoost Regressor)	Mean-Squared Error (MSE) of 0.06, a mean absolute error (MAE) of 0.16 and a higher R-square value of 0.94	*Pros:* Thorough comparison with other popular ML algorithms, highlighting superior performance of the boosting framework. *Cons:* Boosting models can be difficult to interpret.
[15]	Features derived from ECG, electrodermal activity (EDA), Blood Pressure	Propofol concentration	Logistic Regression	AUC of 0.706 between induction and emergence	*Pros:* The study shows that there are lasting effects of anesthesia on autonomic function. *Cons:* Conducted on a small sample of eleven healthy volunteers.
[16]	Heart rate variability (HRV), blood pressure (BP), and electrodermal activity (EDA)	LOC (Loss of Consciousness) and ROC (Regain of Consciousness) annotations	Logistic regression models with LASSO regularization	AUC of 0.8	*Pros:* The findings suggest that autonomic biomarkers could serve as more precise indicators for LOC and ROC during anesthesia in low-resource settings. *Cons:* Conducted on a small sample of nine healthy volunteers, which may limit the generalizability of the findings.

understand the decision-making processes of these models, hindering the trust of the end-users, especially in critical medical applications like DoA assessment.

- **Need for Explainable AI Methods:** The need for explainable AI methods is crucial for gaining acceptance and trust in clinical settings. Addressing this issue is vital for integrating machine learning models into medical practice.

This work aims to fill these gaps by proposing an Explainable Artificial Intelligence (XAI)-based model for DoA estimation, with key contributions outlined below.

- We propose an Explainable Artificial Intelligence (XAI)-based machine learning framework, XAI-VSDoA, designed for determining the depth of anesthesia (DoA) using physiological parameters (or vital signs) like heart rate, arterial pressure wave, respiratory rate, and inspiratory CO_2, and others as listed in Table 2.
- Utilizing SHAP and LIME techniques, we analyze our machine learning models to reveal insights into their decisions and highlight the top 10 physiological parameters that greatly influence the model's output.
- We developed an optimized model by employing only the top-ranked features from SHAP and LIME. This model with fewer features produces results comparable to the model that uses the entire dataset.
- We evaluate the performance of the proposed model using two independent and widely used open datasets. This dual assessment enhances the reliability of our findings, as both datasets consistently demonstrated positive results, affirming the efficacy of the proposed model.
- We employ statistical analysis to provide evidence that the model proposed in the framework excels compared to other models in terms of performance. This rigorous analysis confirms our findings' robustness and underscores the proposed model's superiority.

We compared our work with other machine learning-based approaches available in the literature and found that our framework excels in accuracy and other performance metrics. By integrating explainable AI methods, we aim to offer clinicians transparent and understandable explanations of the outcomes of ML models. We intend for the proposed scheme to act as a clinical decision-support system that augments the decision-making abilities of anesthesiologists.

B. ROAD MAP

This paper is organized as follows: Section II describes the preliminary concepts. Section III elucidates the description of the datasets and preprocessing techniques applied to the dataset and presents the proposed framework. Section IV discusses the experimental details. Following that, Section V constitutes the results and comparison with other ML-based approaches. Based on the findings of this work, section VI concludes this paper.

II. PRELIMINARIES

This section briefly discusses the techniques employed in this study, including feed-forward neural networks, ensemble techniques, ResNet, oversampling techniques, and the eXplainable AI (XAI) methods, namely, LIME and SHAP.

A. FEED-FORWARD NEURAL NETWORK

Feed-forward neural networks are computational models that comprise interconnected nodes known as neurons arranged into layers. Every individual neuron receives input signals, computes, and transmits the outcome to the neurons in the subsequent layer. Feed-forward neural networks can learn to recognize complex patterns and relationships in the data due to the presence of activation functions (like ReLU, sigmoid, or tanh) to introduce non-linearity, enabling them to solve various tasks, such as classification, regression, and feature extraction. The learning in these networks is accomplished by using the loss function (also known as the cost or objective function) that evaluates how well the model's predictions match the actual target values. Backpropagation is generally used to minimise the cost function [24], [25].

B. GRADIENT BOOSTED DECISION TREES

Gradient Boosted Decision Trees (GBDT) is an ensemble learning method that combines decision trees and gradient boosting to create a robust and powerful predictive model [26]. The process begins by constructing an initial decision tree with low predictive power, which serves as the first model. Subsequently, additional trees are constructed to rectify the mistakes made by the preceding models. During each iteration, the algorithm gives more weight to misclassified data points or residuals and trains the next tree to reduce the error further. The gradient in GBDT refers to using the gradient (slope) of the loss function to update the model's parameters, minimizing the loss during each iteration. GBDT continues to add trees until a stopping criterion is met, such as a specified number of trees or when the performance improvement becomes negligible.

1) EXTREME GRADIENT BOOSTING

Extreme Gradient Boosting (XGBoost) [27] is an example of a state-of-the-art GBDT algorithm. XGBoost enhances traditional gradient boosting by incorporating L1 and L2 regularization techniques. XGBoost enhances the gradient boosting framework by incorporating several advanced features, such as regularization, which helps mitigate overfitting, and a sparsity-aware algorithm for handling missing data. It operates by constructing an ensemble of weak learners, typically decision trees, where each successive tree attempts to correct the errors of its predecessors. This iterative process results in a robust model with high predictive accuracy. XGBoost is distinguished by its efficiency and scalability, and it can easily handle large datasets and high-dimensional data. It supports parallel and distributed computing, thereby reducing training time significantly. Moreover, its ability to incorporate L1 (Lasso) and L2 (Ridge) regularization provides additional control over the model complexity. XGBoost's flexibility in handling various objective functions and evaluation metrics makes it a versatile tool in academic research and industry applications. Its effectiveness has been demonstrated in numerous data science competitions, making it a preferred choice for many practitioners aiming to achieve superior predictive performance.

2) CATEGORICAL BOOSTING CLASSIFIER

Categorical Boosting Classifier (CatBoost) is a robust gradient-boosting algorithm that stands out for efficiently handling categorical features [28]. CatBoost eliminates the need for extensive preprocessing like one-hot encoding, thus natively supporting categorical features. This capability significantly reduces preprocessing time and complexity, enabling more straightforward model development. Additionally, CatBoost incorporates L2 (ridge) regularization to prevent overfitting and control model complexity, enhancing generalization performance. CatBoost optimizes training efficiency by supporting a wide range of loss functions, making it versatile for various tasks, including regression, classification, and ranking. Its a novel approach to handling categorical features through an ordered boosting process, which reduces the risk of overfitting and introduces unbiased gradient estimation. Furthermore, CatBoost includes advanced techniques such as symmetric tree structures and oblivious trees, which simplify the model and enhance interpretability without compromising performance. CatBoost's final prediction is a weighted sum of individual tree predictions, where each tree is trained to correct the errors of the preceding ones. This iterative refinement results in a robust ensemble model capable of achieving high predictive accuracy.

3) LIGHT GRADIENT BOOSTING MACHINE CLASSIFIER

The Light Gradient Boosting Machine Classifier (LightGBM) classifier [29] is a gradient-boosting framework widely recognized for its speed, efficiency, and predictive accuracy. It constructs a robust predictive model through a series of decision trees. It utilizes a histogram-based technique to categorize data points into histograms, leading to a substantial decrease in memory consumption and a faster training process. Moreover, it employs a leaf-wise tree growth strategy, where it chooses the leaf that results in the most significant reduction in the loss function at each split. This approach leads to trees with less depth and more rapid convergence. LightGBM can directly handle categorical features, reducing the need for preprocessing. Additionally, it provides L1 (Lasso) and L2 (Ridge) regularization techniques to address overfitting.

C. RANDOM FOREST CLASSIFIER

The Random Forest classifier is an ensemble machine learning algorithm. It creates an ensemble of decision trees, with each tree independently constructed through a process that includes bootstrapping - randomly selecting subsets of the training data with replacement, and random feature selection - using only a subset of features at each node for making splits [30]. These strategies introduce diversity among the trees, mitigating overfitting and enhancing the model's robustness. The aggregation of predictions from individual trees is achieved through a voting mechanism for classification tasks, resulting in a final prediction that benefits from the collective insights of the ensemble.

D. RESIDUAL NEURAL NETWORK

Residual Neural Network (ResNet), is a deep learning architecture designed to overcome the challenges associated with training neural networks with many layers. It accomplishes this by integrating residual blocks. These blocks help the network learn residual functions across layers, enhancing its ability to optimize and acquire complex representations [31]. Residual blocks use skip connections to solve the vanishing gradient problem. The residual block learns the difference between the input and the desired output instead of the layer's input-output mapping. The residual is added to the input to get the block output.

While ResNet's primary success lies in computer vision tasks, researchers have explored its potential in other domains, including tabular data. When adapting ResNet for tabular data, 1-D convolutional layers (Conv1D) are often utilized to process the sequential or structured nature of the data. Conv1D layers enable the model to capture local patterns and interactions between neighbouring features within each sample.

E. SMOTE WITH BOOSTING

The Synthetic Minority Over-sampling TEchnique With Boosting (SMOTEWB) [32] aims to address the limitations of the SMOTE algorithm by combining a noise detection method and SMOTE. SMOTE fundamentally generates synthetic data between extant positive observations by identifying voids within the feature space [33]. Nonetheless, this approach faces two principal constraints. Firstly, it may produce synthetic data within the domain of the majority class, giving rise to spurious noise patterns absent in the natural distribution. Secondly, it employs a fixed number of connections between a sample and its neighbours, neglecting the qualitative attributes of these neighbours when fabricating synthetic data. In response, SMOTEWB integrates a noise detection method to identify noisy observations in the positive class and adjusts the number of neighbours for each observation. This ensures synthetic data is created in areas with non-noisy negative observations. Simultaneously, the boosting technique adjusts observation weights during synthetic data generation, addressing class imbalance and improving classifier performance.

Synthetic samples, denoted as x', are created using the SMOTE algorithm's formula:

$$x' = x + (B - x) \times A \tag{1}$$

where x is an existing positive observation, B is a randomly chosen nearest neighbour, and A is a uniformly distributed random value between 0 and 1.

F. NORMALIZATION AND SCALING FEATURES

Min-Max scaling, also referred to as Min-Max normalization, is a data preprocessing technique that rescales numeric features to fit within a predetermined range, typically ranging from 0 to 1. The objective is to standardize all the characteristics to avoid the dominance of any particular feature due to its greater magnitude. The formula for Min-Max scaling is:

$$X_{\text{scaled}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

where X is the original value of the feature. $X_{\min}$ is the minimum value of the feature in the dataset. $X_{\max}$ is the maximum value of the feature in the dataset.

G. EXPLAINABLE ARTIFICIAL INTELLIGENCE

Explainable Artificial Intelligence (XAI) is a crucial concept that addresses the need for transparency and understanding in AI systems. XAI aims to shed light on the black-box nature of many AI algorithms, making it possible to comprehend how AI arrives at its conclusions and predictions [34]. This transparency is essential in fields like medicine, where the decisions made by AI systems directly impact patient diagnosis and treatment. For instance, XAI allows doctors to understand why an AI predicts a patient's risk for a medical condition, improving trust and informed decision-making. In the literature on anesthesia classification discussed above, we noted that several ML and deep learning methods had been proposed for predicting DoA but lacked interpretability and explainability. To overcome these limitations, we have integrated XAI with our models.

1) LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATIONS

LIME is a tool used in the field of XAI to provide interpretable explanations for machine learning models, especially in cases where complex models make it difficult to understand why a particular prediction was made. LIME aims to approximate the decision boundary of a black-box model locally and generate explanations that are understandable to humans [35].

Initially, LIME perturbs the input data point of interest x, introducing slight variations to create a dataset of similar instances. Subsequently, it applies a straightforward and easily understandable model (typically linear regression or decision trees) to this perturbed dataset. This simplified model provides an approximation of the behaviour of the complex model near the original data point. LIME assigns weights to the perturbed instances according to their proximity to the original data point. This ensures that instances closer to the point of interest carry more weight in model fitting. The explanations produced by the LIME are obtained by the following:

$$\xi(x) = argmin_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{2}$$

where $\xi(x)$ symbolizes the interpretable explanation generated by LIME for the specific input instance x. $\mathcal{L}(f, g, \pi_x)$ measures the dissimilarity between the predictions of the complex model f and the interpretable model g for the perturbed instances π_x. Essentially, it quantifies how well g approximates f in the local context. $\Omega(g)$ is a regularization term that penalizes the complexity of the interpretable model g.

2) SHapley ADDITIVE exPLANATIONS

SHapley Additive exPlanations (SHAP) is another framework for explaining the predictions made by machine learning models. It is based on Shapley values, which come from cooperative game theory [36]. In the context of model interpretability, it assigns a Shapley value, denoted as ϕ, to each feature (or input) to quantify its contribution to a model's prediction. The Shapley value can be calculated using the following formula:

$$\phi_i(f) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} (f(S \cup \{i\}) - f(S)) \tag{3}$$

where, $\phi_i(f)$ represent the Shapley value for feature i, N is the set of all features, S is a subset of features that does not include feature i, $f(S)$ is the model's prediction when using the feature set S and $(f(S \cup \{i\}) - f(S))$ is the model's prediction when including feature i in the feature set.

Kernel SHAP is an extension of SHAP that makes the computation of Shapley values more efficient, especially for complex models like deep neural networks. Instead of calculating Shapley values for all possible combinations of feature values, Kernel SHAP approximates these values using a weighted average of predictions for a subset of combinations. It uses a kernel function to weigh the importance of different combinations, making the calculation more tractable and computationally feasible, particularly for high-dimensional datasets.

III. MATERIALS AND METHODS

This section provides the materials and methods employed in our research. It encompasses a detailed description of the dataset used, the data preprocessing techniques applied, and an in-depth presentation of our proposed framework.

A. DATASET DESCRIPTION

This research study employed two datasets: **The University of Queensland Vital Signs Dataset** and the **VitalDB Dataset**. Employing multiple datasets allows for a more robust analysis by providing a broader range of data points and enhancing the generalizability of the findings. The University of Queensland Vital Signs Dataset [37] contains records of vital signs, including parameters like heart rate, pulse rate, blood pressure (systolic, diastolic, and mean), and blood oxygen saturation. These recordings were collected from 32 patients who were given anesthesia at the Royal Adelaide Hospital in Australia. Out of these cases, 25 patients were given general anesthesia, 3 patients were given spinal anesthesia, and 4 patients were given sedatives. The duration of these records varied from 13 minutes to 5 hours, with an average duration of 105 minutes. This dataset predominantly includes data from various monitoring devices, such as the electrocardiograph, pulse oximeter, capnograph, noninvasive arterial blood pressure monitor, airway flow, and pressure monitor. In a few instances, additional data from monitoring devices like the Y-piece spirometer, electroencephalogram monitor, and arterial blood pressure monitor were also included. This database includes a total of 65 parameters.

The VitalDB Dataset [38] is an openly available dataset specifically created to facilitate machine learning research focused on monitoring the vital signs of patients undergoing surgery. The data was collected from patients undergoing non-cardiac surgeries, such as general, thoracic, urologic, and gynecologic procedures, at Seoul National University Hospital in Seoul, Republic of Korea. The dataset comprises comprehensive and detailed information from 6,388 patients, including high-resolution multi-parameter data. This data includes waveform and numeric data representing various intraoperative monitoring parameters, perioperative clinical factors, and time-series laboratory results. The dataset comprises 196 parameters for intraoperative monitoring, 73 for perioperative clinical data, and 34 for time-series laboratory results.

B. DATA PREPROCESSING

The initial phase of this study involves a comprehensive analysis of the available datasets. Initially, from The University of Queensland Vital Signs dataset, data about five patients who had undergone general anesthesia, namely cases 22, 28, 29, 30, and 31, were chosen for investigation as the target BIS value is only available for these patients. The original dataset was recorded at a high temporal resolution of 10 milliseconds. However, it has been observed that the numerical values of vital signs do not change so frequently. So, in order to reduce computational complexity, we downsampled the data to a 2-second internal, comprising 16,773 records. Further, we substituted the dataset's missing values with the corresponding feature's median value. As numerical values of the vital signs relate to the physiological parameters that indicate crucial trends and patterns, we have focused on the vital signs in this study. Based on our discussion with the anesthesiologists, we selected 22 parameters for further investigation. These 22 dataset parameters have been listed in Table 2 along with their description.

In the VitalDB dataset, the data tracks had varying sampling frequencies due to the different monitoring devices used. We utilized data from three distinct monitoring devices — BIS, Solar8000, and Primus. These devices have sampling rates of 1 second, 2 seconds, and 7 seconds, respectively. We opted for a standardized sampling rate of 14 seconds to ensure uniformity in the dataset. We used data from 200 patients for a total of 67,577 records. Following this, expert recommendations guided our feature selection process, which resulted in 23 features from an initial pool of 158 (excluding time-series data). The details of the parameters and their description are given in Table 2. Additionally, missing values were imputed with the median patient value based on medical insights and recommendations.

Feature normalization was performed on both datasets to rescale the feature values within a range of 0-1. For this purpose, a MinMax scaler was utilized to calculate each feature's minimum and maximum values and proportionally scale the feature values to fit within the desired range. This normalization process ensures that all features are on a comparable scale, facilitating fair comparisons and preventing the dominance of certain features solely based on their numerical magnitude. Min-Max scaling was preferred over Standard Scaling due to the absence of a normal distribution in our features, which is assumed by Standard Scaling. Therefore, Min-Max scaling is more appropriate for our specific case.

Both datasets exhibit substantial class imbalance. In the University of Queensland Vital Signs dataset, the instances denoting 'OK Anesthesia', representing appropriate dosages of anesthesia, and 'Light Anesthesia', signifying a mild dosage of anesthesia, are notably fewer than those indicating 'Deep Anesthesia', which signifies higher dosages of anesthesia. Likewise, in the VitalDB dataset, records associated with 'Light Anesthesia' are significantly less prevalent than other classes. To rectify this imbalance, a modified variant of the Synthetic Minority Oversampling Technique (SMOTE) known as SMOTE with Boosting (SMOTEWB) has been employed [32]. Unlike traditional SMOTE, SMOTEWB overcomes limitations by combining a noise detection method based on boosting. This method identifies potential noisy instances and determines the right number of neighbours for each observation, thus reducing the generation of noisy synthetic data.

C. PROPOSED FRAMEWORK (XAI-VSDoA)

In this paper, we propose XAI-VSDoA framework comprising three distinct steps - preprocessing data, training an XGBoost classifier on the input data, and finally, generating explanations using LIME and SHAP. The process is described in Fig. 1. This flowchart illustrates the step-by-step process for a model designed to categorize the depth of anesthesia into three levels using datasets from the University of Queensland Vital Signs dataset and the VitalDB dataset. The preprocessing phase involves four steps, beginning with feature selection guided by a senior anesthesiologist, followed by the removal of rows with null values and signs with a Signal Quality Index (SQI) below 40 to ensure data quality. Subsequently, the data is normalized to bring features to a consistent scale, and the Synthetic Minority Over-sampling Technique with Boosting (SMOTEWB) is applied to address the class imbalance.

Moving forward, the classifier training stage involved rigorous experimentation with a variety of machine learning classifiers. Through this comprehensive evaluation process, XGBoost emerged as the most effective model. This conclusion was substantiated through cross-validation, which demonstrated XGBoost's exceptional performance, achieving an accuracy of 99.34%. Additionally, we conducted a thorough analysis of various performance metrics and performed statistical testing to confirm the reliability of our findings. Following successful training, explanations for the model's predictions are generated using LIME and SHAP. LIME provides local interpretability by fitting a simple model to individual predictions, while SHAP assigns values to features to indicate their contributions to predictions. The final step involves analysis of the generated explanations to ascertain their coherence and gain insights into the factors influencing the model's decision-making process. As shown in Table 3, the anesthesia level (the target variable) takes three distinct values corresponding to the three classes: Deep Anesthesia, characterized by BIS values less than 40; OK Anesthesia, corresponding to BIS values ranging from 40 to 60; and Light Anesthesia, denoted by BIS values exceeding 60. This categorization has been adopted by [19], [39], and [40] among others in automated DoA monitoring.

The choice of XGBoost for this framework stems from comprehensive experimentation with a diverse array of machine-learning classifiers. This inclusive exploration involved ensemble models such as XGBoost, CatBoost, LightGBM, and Random Forest, along with diverse feed-forward neural networks and a ResNet model utilizing 1D filters. Input features for these models were drawn from the University of Queensland Vital Signs Dataset and the VitalDB dataset. The output of the classifiers categorized anesthesia into light, deep, or OK states.

To evaluate model performance and ascertain generalizability, 10-fold cross-validation was used. Rigorous testing using various performance metrics, including accuracy, precision, recall, AUC, and F1-score, was conducted to gauge the models' efficacy. In the domain of anesthesia, machine learning models' explainability is crucial for several reasons. A precise DoA prediction by an ML model directly influences patient well-being, emphasizing

TABLE 2. Physiological parameters chosen for this study along with their description.

University of Queensland Vital Signs Dataset		VitalDB Dataset	
Physiological Parameter	**Description**	**Physiological Parameter**	**Description**
HR	Heart rate derived from the ECG sensor	BIS/EMG	Electromyography power
ST-II	ST segment index derived from the ECG sensor	Solar8000/HR	Heart rate
Pulse	Pulse rate (heart rate) derived from pulse oximetry	Solar8000/Pleth_spo2	Percutaneous oxygen saturation
SpO2	Blood oxygen saturation derived from pulse oximetry	Solar8000/ST_II	ST segment in lead II
Perf	Perfusion index derived from pulse oximetry	Solar8000/VENT_INSP_TM	Inspiratory time (from ventilator)
etCO2	Blood oxygen saturation derived from pulse oximetry	Solar8000/ART_DBP	Diastolic arterial pressure
imCO2	Inspired minimum CO2 measured using sidestream capnography	Solar8000/ART_MBP	Mean arterial pressure
awRR	Respiratory rate derived from capnography	Solar8000/ART_SBP	Systolic arterial pressure
NBP(Sys)	Systolic blood pressure measured with a non-invasive blood pressure cuff	Primus/Compliance	Airway compliance
NBP(Dia)	Diastolic blood pressure measured with a non-invasive blood pressure cuff	Primus/ETCO2	End-tidal CO2
NBP(Mean)	Mean blood pressure measured with a non-invasive blood pressure cuff	Primus/FEO2	Fraction of expired O2
NBP(Pulse)	Pulse (heart) rate measured using the non-invasive blood pressure cuff	Primus/FIO2	Fraction of inspired O2
NBP(Time Remaining)	Time remaining in seconds until the next non-invasive blood pressure sample is taken	Primus/INCO2	Inspiratory CO2
etISO	End-tidal isoflurane gas concentration measured from the airway	Primus/MAC	Minimum alveolar concentration of volatile
etSEV	End-tidal sevoflurane gas concentration measured from the airway	Primus/MAWP_MBAR	Mean airway pressure
MAC	Minimum alveolar concentration of the current anesthetic gas being used (desflurane/isoflurane/sevoflurane)	Primus/MV	Minute volume
etO2	End-tidal oxygen concentration measured from the airway	Primus/PAMB_MBAR	Ambient pressure
inO2	Inspired oxygen concentration measured from the airway.	Primus/PEEP_MBAR	Positive end-expiratory pressure (PEEP)
EMG	Electromyography (EMG) indicator from the BIS monitor.	Primus/PIP_MBAR	Peak inspiratory pressure
Tidal Volume	Tidal volume of the previous breath measured by the ventilator	Primus/PPLAT_MBAR	Plateau pressure
Minute Volume	Current minute volume measured by the ventilator	Primus/RR_CO2	Respiratory rate based on capnography
RR	Current respiratory rate measured by the ventilator	Primus/TV	Tidal volume
		Primus/EXP_SEVO	Expiratory sevoflurane pressure

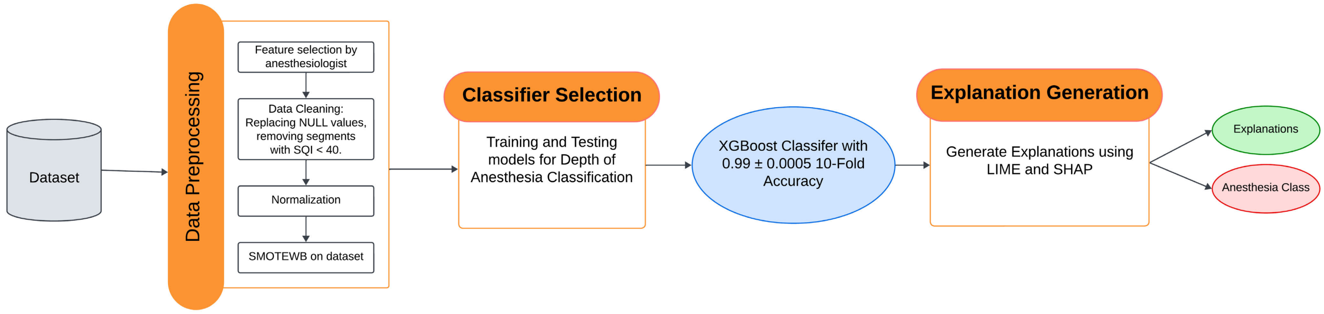

Dataset
Data Preprocessing
Feature selection by anesthesiologist
Data Cleaning: Replacing NULL values, removing segments with SQI < 40.
Normalization
SMOTEWB on dataset
Classifier Selection
Training and Testing models for Depth of Anesthesia Classification
XGBoost Classifer with 0.99 ± 0.0005 10-Fold Accuracy
Explanation Generation
Generate Explanations using LIME and SHAP
Explanations
Anesthesia Class

FIGURE 1. Proposed Framework for assessing the Depth of Anesthesia (DoA). First, the dataset is preprocessed, followed by classifier training. The best model is used for the generation of explanations using XAI techniques - LIME and SHAP.

the need for transparent reasoning to foster trust among anesthesiologists.

Algorithm 1 presents the step-by-step procedure used to train the deep learning model and generate the vital signs that significantly influence the trained model to determine the state of depth of anesthesia. It takes as input a dataset D, where each instance x_i includes input features and corresponding labels y_i, representing the depth of anesthesia categorized as AO (OK Anesthesia), AL (Light Anesthesia), or AD (Deep Anesthesia). Key parameters include k for the number of folds in cross-validation, $k\_topFeatures$ indicating the top features to consider, and

TABLE 3. Categorization of BIS values into anesthesia states.

BIS Range	Anesthesia State	Class Label
$0 \leq \text{BIS} < 40$	Deep Anesthesia	Class 2
$40 \leq \text{BIS} < 60$	OK Anesthesia	Class 1
$60 \leq \text{BIS} \leq 100$	Light Anesthesia	Class 0

XMethods, a list of explainable methods. The output includes the trained model (*Model_best*) and a set of common vital sign attributes (vs_{common}) obtained from various XAI methods. The algorithm proceeds by preprocessing the dataset and addressing missing values and outliers. Subsequently, it employs k-fold cross-validation to train a model (*Model_best*) on vital sign features. Following model training, the algorithm iterates through selected explainable methods, such as LIME and SHAP, to interpret the trained model and extract vital sign features (vs_{lime}, vs_{shap}). It then computes vs_{common} by determining the intersection of features highlighted by LIME and SHAP. Finally, the algorithm analyzes the performance of *Model_best* using the common set of vital sign attributes (vs_{common}). The overarching objective is to enhance the interpretability of the model, facilitating a clearer understanding of the vital sign features influencing the depth of anesthesia predictions.

IV. EXPERIMENTAL DETAILS

We have evaluated the proposed scheme's performance on both the University of Queensland Vital Signs Dataset and VitalDB datasets. For experimentation, we used Google Colab Pro with a GPU A100 accelerator comprising 83GB of system RAM and 40GB of GPU.

A. ARCHITECTURE OF DEEP LEARNING MODELS

In this work, we employed two distinct types of deep learning models: the feed-forward neural network and the Residual Network (ResNet). We experimented with various neural network architectures with 4, 5, 6, and 8 layers, each with different configurations of neurons in each layer. Fig. 2 illustrates the 6-layered and 8-layered feed-forward neural networks implemented within the study. These networks contain a traditional stack of dense layers, as depicted in the figure. Further, we also trained the ResNet model from scratch, which consists of multiple residual blocks and integrates convolutional layers with batch normalization. This ResNet architecture with 1D convolution layers follows a structure similar to the standard ResNet but utilizes 1D convolutions instead of 2D convolutions. The final layer of the model includes global average pooling to reduce dimensions and a dense layer for classification. Fig. 3 depicts the architecture of the ResNet model.

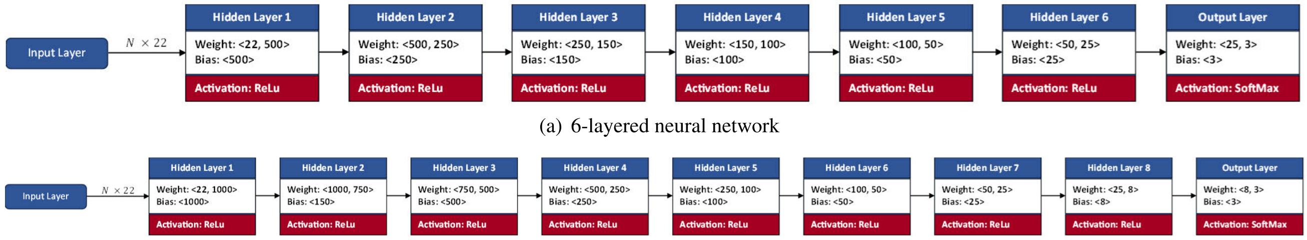
These models employ the categorical cross-entropy loss function, accommodating three output categories. Each model underwent 100 training epochs, allowing ample iterations for learning and optimization. TensorFlow, a widely utilized deep learning framework, served as the backend for the training process due to its robustness and flexibility in model implementation. Subsequently, ReLu activation was applied to the output layers of the models that assign classification probabilities to the input belonging to one of three classes, i.e., Deep Anesthesia (Class-2), OK Anesthesia (Class-1), and Light Anesthesia (Class-0).

Algorithm 1 XAI-VSDoA Framework for Assessing the of Depth of Anesthesia

```
Input: $D$: Dataset $D = \{x_i, y_i\}_{i=1}^{n}$, where $x_i \in \mathbb{R}^d$ is
the input feature, $d$ is the dimension of feature space,
and
        $y \in \{AO, AL, AD\}$ is the target.
        $k$: Number of equal-sized folds or groups the
        dataset
        $D$ is divided into
        $k\_topFeatures$: Top features (Set to 10 in this
        work)
        $XMethods$: List of explainable methods
Output: Trained model ($Model\_best$), frequently
occurring vital features
Procedure:
    1) Preprocess the dataset $D$ via handling missing
    values, removing outliers, attribute selection, and
    scaling.
    2) Apply $k$-fold cross validation taking the vital
    sign features $x$ as input and the corresponding
  class
    labels $y$ as the target to get the trained model
  $Model\_best$.
        $Model\_best \leftarrow modelTraining(D)$        (4)
    3) For all explainable methods
    $(m) \in XMethods$ do:
    a) if $m ==$ LIME do:
        $vs_{lime} \leftarrow LIME(Model\_best, D, k\_topFeatures)$        (5)
    b) if $m ==$ SHAP do:
        $vs_{shap} \leftarrow SHAP(Model\_best, D, k\_topFeatures)$        (6)
    c) Compute the common vital sign attributes
    obtained using different explainable AI methods
    as
        $vs_{common} \leftarrow \bigcap (vs_{lime}, vs_{shap})$        (7)
end For
    4) Analyse the performance of $Model$_best using
    $vs_{common}$ as inputs.
```

B. HYPERPARAMETER TUNING

To determine the optimal hyperparameters for diverse machine learning models, we utilized the Grid Search



(a) 6-layered neural network

(b) 8-layered neural network

FIGURE 2. Architecture of different feed-forward neural networks.

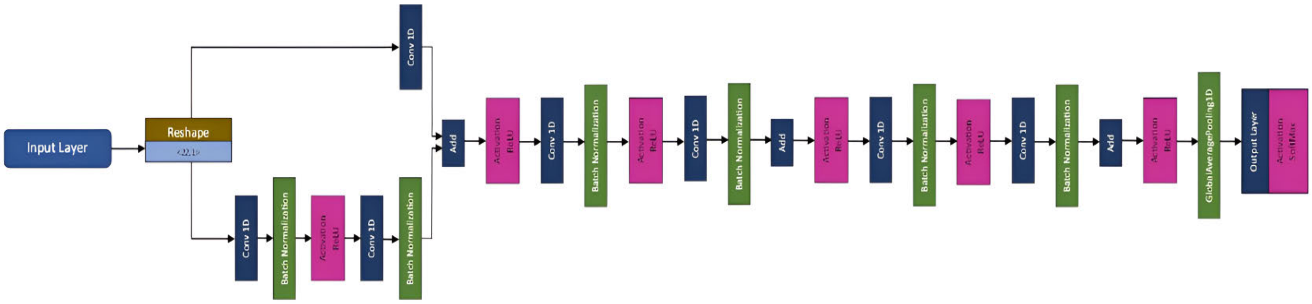


FIGURE 3. Architecture of ResNet model (with 1D convolution layers).

Cross Validation (GridSearchCV) technique. This methodology systematically examines a grid of hyperparameter combinations, employing cross-validation to assess the performance of each combination. For deep learning models, hyperparameter tuning was performed using the *HParams* dashboard accessible through TensorBoard. The results have been summarised in Table 4.

V. RESULTS AND DISCUSSION

This work aimed to assess the depth of anesthesia based on vital sign values. As per the commonly observed practice in the literature, we categorised the depth of anesthesia into three distinct classes: Light Anesthesia, OK Anesthesia, and Deep Anesthesia. The results are presented for the ensemble classifiers and deep learning models to determine the depth of anesthesia. To evaluate the models, several metrics were considered, including accuracy, precision, recall, specificity, and F1 scores. These metrics provide a comprehensive understanding of the model's performance in terms of overall correctness, precision, and recall balance. Furthermore, confusion matrices were generated for each trained model, providing visualizations that aid in interpreting the results and understanding the distribution of predictions across the different categories.

A. PERFORMANCE EVALUATION

The models were trained and evaluated using two distinct datasets: the University of Queensland Vital Signs Dataset and the VitalDB Dataset. Among the ensemble classifiers utilized, the XGBoost model achieved an accuracy of 99.34% on the University of Queensland Vital Signs Dataset and 93.07% on the VitalDB Dataset. The CatBoost model closely followed with accuracies of 99.39% and 92.59% on the University of Queensland Vital Signs and VitalDB datasets, respectively. Additionally, the LightGBM and Random Forest classifiers showed accuracies of 96.66% and 98.18%, respectively, on the University of Queensland Vital Signs Dataset. On the VitalDB Dataset, their accuracies were recorded as 92.34% and 88.20%, respectively. Incorporating neural networks into the analysis, the implemented 8-layered neural network achieved accuracies of 98.40% and 93.08% on the University of Queensland Vital Signs and VitalDB Datasets, respectively. Furthermore, the 1-D ResNet model displayed favourable results, attaining an accuracy of 99.20% on the University of Queensland Vital Signs Dataset and 92.80% on the VitalDB Dataset. In summary, Table 5 and 6 show that the XGBoost Classifier either outperforms or gives comparable performance to other classifiers.

Further, the confusion matrices were also generated for all models. These matrices offer a graphical representation elucidating the counts of True Positives, True Negatives, False Positives, and False Negatives for each classification category. Fig. 4 and 5 showcase the confusion matrices of the XGBoost model, which yielded the most favourable outcomes. These matrices indicate minimal misclassifications of Light Anesthesia examples in the other two classes. This is advantageous, particularly since one of our primary goals is to avoid recall due to low anesthesia dosage. The notable misclassification in Fig. 4 and 5 primarily occurred between the OK Anesthesia and Deep Anesthesia

TABLE 4. Details of tuned hyperparameters.

Model	Hyperparameter	Values/Ranges tested	Optimal value based on experimentation
XGBoost	learning_rate	0.0001 - 0.2	0.01
	gamma	0.1 - 2	0.1
	subsample ratio	{0.6, 0.8, 1}	0.8
	max_depth	{5, 7, 9, 10, 12}	12
	n_estimators	{100, 150, 200}	150
CatBoost	learning_rate	0.1 - 0.5	0.4
	max_depth	{5, 7, 9, 10, 12}	10
	iterations	{500, 1000, 1500}	1000
	bagging_temperature	{0.4, 0.6, 0.8}	0.6
	border_count	{16, 32, 64, 128}	64
LightGBM	learning_rate	0.01 - 0.2	0.1
	max_depth	{7, 11, 12, 15}	15
	subsample	20 - 40	0.8
Random Forest	criterion	{gini, entropy}	entropy
	n_estimators	{100, 150, 200}	150
	max_depth	5 - 15	12
Neural Networks	batch_size	{32, 64, 128, 256}	64
	activation	{tanh, ReLu}	ReLu
	optimizer	{SGD, RMSprop, Adam, AdamW, Adadelta, Nadam}	Adam

TABLE 5. Comparative accuracies of various models implemented on the University of Queensland Vital Signs dataset.

Models Implemented	Accuracy(%)
XGBClassifier	99.34 ± 0.0005
CatBoost	99.39 ± 0.0005
LightGBM	96.66 ± 0.083
Random Forest	98.18 ± 0.001
6-layered NN	98.36 ± 0.013
8-layered NN	98.40 ± 0.002
ResNet (1D variant)	99.20 ± 0.002

TABLE 6. Comparative accuracies of various models implemented on the VitalDB dataset.

Models Implemented	Accuracy(%)
XGBClassifier	93.07 ± 0.001
CatBoost	92.59 ± 0.003
LightGBM	92.34 ± 0.002
Random Forest	88.20 ± 0.002
6-layered NN	92.81 ± 0.016
8-layered NN	93.08 ± 0.017
ResNet (1D variant)	92.80 ± 0.015

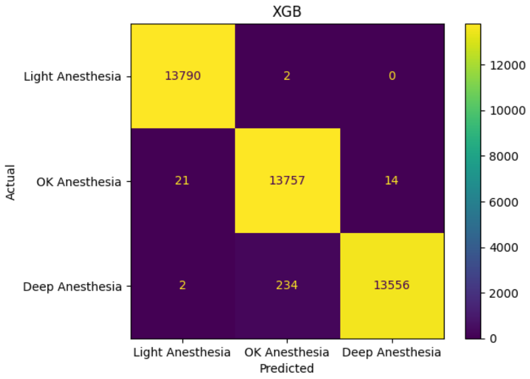


FIGURE 4. Confusion matrix illustrating the performance of the XGBoost model on the University of Queensland Vital Signs dataset.

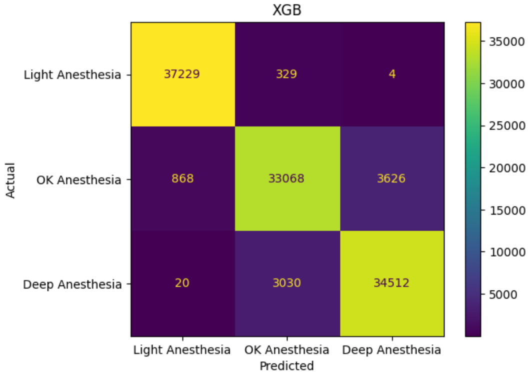


FIGURE 5. Confusion matrix illustrating the performance of the XGBoost model on the VitalDB dataset.

classes, with instances of deep anesthesia being erroneously labelled as OK. This phenomenon can be ascribed to the human tendency to err on the side of caution, potentially choosing slightly higher anesthesia dosages to mitigate the risk of patients waking up during surgery. Consequently, the anesthesia levels might have fallen within the upper range of the deep category (0-40).

Additionally, a detailed analysis of performance metrics was conducted for the XGBoost classifier to gain deeper insights into its efficacy. These metrics encompass a range of statistical measures evaluating the classifier's accuracy, precision, recall, and F1 score. Table 7 compares the performance metrics observed in the University of Queensland Vital Signs and VitalDB datasets, elucidating the classifier's performance across different datasets and highlighting its consistency or variations in predictive capabilities. It may be noted that on both datasets XGBoost can achieve precision, recall, and F1 score greater than 92%, thus establishing a high performance by the model in classifying the examples into three classes.

In the context of applications like depth of anesthesia, we advocate for prioritizing the F1 score as the primary metric. This choice is grounded in the unique characteristics of the F1 score, which balances precision and recall. It proves particularly crucial in scenarios where achieving a harmonious equilibrium between minimizing false positives and false negatives is paramount. In the context of depth of anesthesia, ensuring both the accurate identification of instances requiring intervention (high recall) and the avoidance of unnecessary interventions (high precision) is

TABLE 7. Performance evaluation of XGBoost classifier: comparative analysis between the University of Queensland Vital Signs dataset.

Dataset used	Accuracy	Precision	Recall	F1-Score
The University of Queensland Vital Signs Dataset	99.34%	99.35%	99.34%	99.34%
VitalDB Dataset	93.07%	93.04%	93.07%	93.04%

of utmost importance. Nevertheless, we have disclosed all performance metrics, including accuracy, precision, recall, and F1-score, to foster transparency and contribute to the future development of this research.

1) STATISTICAL ANALYSIS

From Table 5 and 6, we may note that on both the University of Queensland Vital Signs and VitalDB datasets, the XGBoost classifier performs better than all the other classifiers under consideration. However, to capture any stochasticity of the system and statistically evaluate the significance of the difference between XGBoost and other classifiers, we repeated our entire experiment 40 times with random seed values. For all the 40 different iterations, we performed 10-fold cross-validation on all the classifiers. We performed a z-test to statistically evaluate the significance as the number of test samples is more than 30, therefore as per the central limit theorem, we have considered the samples to be approximately normally distributed [41]. Further, we have used the value of the level of significance (α) equal to 0.05.

Let us consider the null hypothesis ($H_0 : \mu_1 = \mu_2$), positing that there exists no significant disparity between the mean values (μ) of the XGBoost classifier and the other classifiers being assessed. Conversely, the alternative hypothesis ($H_1 : \mu_1 > \mu_2$) suggests the superiority of XGBoost over the others. Let z be the test statistics and $z_\alpha = 1.645$ be the critical value. If the value of z lies in the critical region ($z > z_\alpha$) or the p-value is less than α, then we reject the null hypothesis. Otherwise, we conclude that the performance of XGBoost is superior, and the difference is big enough to be statistically significant.

For the University of Queensland Vital Signs dataset, on comparing XGBoost with CatBoost, LGBM, Random Forest, MLP-6layer, and MLP-8layer, the value of test statistics z is -5.76, 2.176892, 98.190808, 881.854, 318.316, and -11.256, respectively. Similarly, on comparing XGBoost with CatBoost, LGBM, Random Forest, MLP-6layer, and MLP-8layer, the p-values are $8e^{-9}$, 0.0294886, 0, 0, 0, and 0, respectively. As in all the cases, the value of test statistics z is not in the 95% region of acceptance [-1.959964: 1.959964], and the p-value is less than the level of significance $\alpha = 0.05$. Therefore, we reject the null hypothesis H_0 and conclude that the performance of XGBoost is superior to other classifiers, and the difference is big enough to be statistically significant. We also performed the statistical evaluation on the VitalDB dataset and found similar results. Therefore, we conclude that the performance of XGBoost is superior to other classifiers on both datasets.

B. INTERPRETATION OF THE MODELS

Although XGBoost provides the highest accuracy, being a black-box model, it does not provide insights into how it came to a decision. So, we deploy XAI models LIME and SHAP to understand how they arrive at the outcomes in terms of the input parameters that serve as the physiological parameters (vital signs). To instil confidence, it is imperative to scrutinize the model's predictive behaviour on a per-instance, per-class basis. This examination allows us to streamline the feature set (set of vital signs), thereby expediting model training and potentially enhancing prediction accuracy by identifying the pivotal elements influencing prediction outcomes. In the context of establishing clinical confidence in predicting DoA, it becomes crucial to elucidate the specific roles played by each of the features (vital signs) employed. Utilizing XAI methods aids in investigating the contributions of vital signs to the predictive process.

Our study utilized LIME and SHAP methodologies in conjunction with trained models to gain insight into their performance and the significance of vital signs in assessing DoA. Our study exclusively presents outcomes derived solely from our top-performing model, XGBoost, which is applied across both datasets. Furthermore, we deliberated upon the feature importance scoring mechanism integrated within XGBoost, elucidating the importance of various vital signs that identify the most pivotal features contributing to the prediction process.

Fig. 6(a) and 6(b) illustrate the feature importance scores obtained through the XGBoost mechanism for both the University of Queensland Vital Signs and VitalDB datasets. The feature importance scores showcased in these figures illuminate the pivotal role played by individual features in the predictive capacity of the XGBoost algorithm across these datasets.

Fig. 7 and 8 show the occurrence of physiological parameters in the top 10 features list, categorized by class, across 15 randomly selected examples for the University of Queensland Vital Signs Dataset and the VitalDB Dataset, respectively. The x-axis represents different physiological parameters, while the y-axis indicates the count of occurrences for each parameter. Examining the results for the University of Queensland Vital Signs Dataset, specifically for class 0 (Light Anesthesia), the LIME explanation highlights vital signs such as EMG, inO2, and HR as the foremost

TABLE 8. Statistical analysis.

$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 > \mu_2$
Region Of Acceptance: $[-1.959964 : 1.959964]$

Sr. No.	Comparison	Critical value	p-value	Decision
1	XGBoost vs CatBoost	-5.76	8.34738e-9	Reject null
2	XGBoost vs LGBM	2.176892	0.0294886	Reject null
3	XGBoost vs Random Forest	98.190808	0	Reject null
4	XGBoost vs 6layer FFN	881.854738	0	Reject null
5	XGBoost vs 8layer FFN	318.316158	0	Reject null
6	XGBoost vs ResNet	-11.256491	0	Reject null

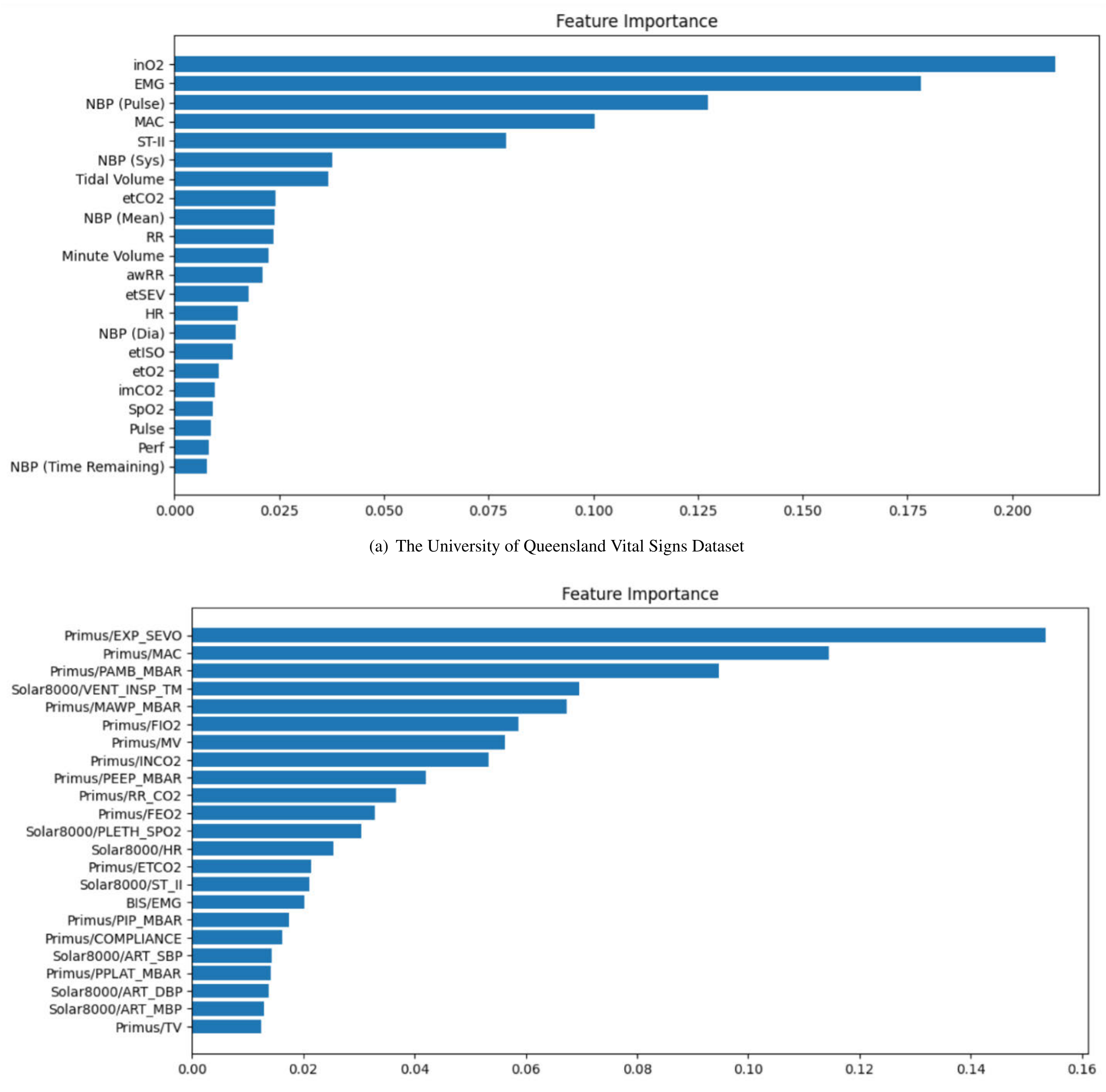
contributors to that particular accurate prediction. Important features for other classes can be read similarly. In the case of heart rate (HR), Class 0 exhibits the highest count, followed by Class 1, and Class 2 has the lowest count. A similar trend is observed for the ST segment (ST-II) and pulse, where Class 0 consistently shows higher counts. Oxygen saturation (SpO2) also follows this pattern, with Class 0 leading in count compared to the other classes. Parameters such as end-tidal CO2 (PetCO2) have a noticeable presence in Class 1 and Class 2, whereas Class 0 has a significantly lower count. Airway respiratory rate (awRR) is more prevalent in Class 0. Non-invasive blood pressure (NBP) measurements, including systolic, diastolic, and mean, show varied distributions; for instance, NBP (Mean) is more pronounced in Class 1. End-tidal isoflurane (etISO) counts are higher in Class 1 and Class 2, reflecting a distinct pattern compared to other parameters. Minimum alveolar concentration (MAC) is more prevalent in Classes 1 and 2. Parameters such as etCO2, iNo2, EMG, tidal volume, minute volume, and respiratory rate (RR) exhibit varying counts across the three classes. Class 0 tends to have higher counts in several physiological parameters, with notable differences in how NBP, etISO, and sSEC are distributed among the classes.

Fig. 9 and 10 depict the explanation provided by the LIME explainer when applied to the University of Queensland Vital Signs and the VitalDB datasets, respectively. These figures illustrate the accurate predictions made by the XGBoost classifier for all three classes. Further, the top 10 physiological parameters significantly contributing to specific predictions are also represented. The chart in 9 explains the prediction probabilities and feature contributions for classifying data into three classes (Class 0, Class 1, and Class 2) for the University of Queensland Vital Signs dataset. The prediction probabilities show that this instance is classified with 100% certainty as Class 0. The chart then details how various features contribute to this classification. For Class 0, features like EMG (greater than 0.26), etISO (less than or equal to 0.43), and Minute Volume (greater than 0.60) are significant contributors. These features increase the likelihood of the instance being classified as Class 0. Conversely, features contributing to Class 1 and Class 2, such as MAC and inO2, are not prominent in this instance. The table at the bottom lists the actual values of these features, which the model uses to predict. Understanding these contributions helps interpret why the model classified this instance as Class 0, providing insights into the key physiological parameters influencing the decision. The remaining plots can be read similarly for other classes.

Notably, the SHAP explainer confirms these findings, as demonstrated in Fig. 11(a). Fig. 11(a) presents a comprehensive overview of the feature importance generated by the SHAP explainer, showcasing the features and their corresponding Shapley values in descending order. Similarly, for the VitalDB dataset, Fig. 10 exhibits the LIME explanation, revealing the key physiological parameters influencing predictions for the vitalDB Dataset. For instance, in class 0, the pivotal parameters or features contributing to the prediction include BIS/EMG, Solar8000/HR, and Primus/MAC. Notably, the SHAP summary plot in Fig. 11(b) substantiates these results, illustrating the identical set of features deemed significant for class 0 predictions.

The examination for the other two classes reveals similar results. For class 1 (OK anesthesia), MAC, tidal Volume, etCO2, and Pulse were highlighted by LIME explanations as the most important for randomly chosen 15 examples of the University of Queensland Vital Signs Dataset. For the vitalDB dataset, BIS/EMG, Solar8000/ST_II, Primus/PAMB_MBAR, and Primus/PEEP_MBAR were considered important. For class 2 (Deep Anesthesia), LIME highlighted etISO, inO2, and EMG as the most important for the University of Queensland Vital Signs Dataset. Similarly, BIS/EMG, Primus/inCO2, and Primus/etCO2 were chosen as the most important for the VitalDB dataset. Notably, these results agreed with the explanations supplied by LIME, thereby enhancing the trust in the predictions of the XGBoost model.

For the University of Queensland Vital Signs Dataset, the top 10 features as given by LIME are EMG, inO2, MAC, Tidal Volume, NBP(Pulse), Minute Volume, NBP(Time Remaining), ST-II, etISO, and etCO2. SHAP enlists HR, inO2, EMG, MAC, Tidal Volume, etISO, Pulse, etCO2, NBP(Pulse), and etISO as the most important top 10 features. It is clear that out of these top 10 most important features identified by the explainable techniques, 8 are common. Similarly, for the VitalDB Dataset, LIME considered Primus/MAC, Solar8000/HR, BIS/EMG, Primus/MV, Primus/INCO2,



(a) The University of Queensland Vital Signs Dataset

(b) The VitalDB Dataset

FIGURE 6. Feature Importance Scores from the XGBoost Model. The plot displays the feature importance scores obtained from the XGBoost machine learning model. Each bar represents the importance of a specific feature in predicting the target variable. Higher bar heights indicate greater importance of the corresponding feature in the predictive process.

Primus/EXP_SEVO, Primus/ETCO2, Primus/PEEP_MBAR, Primus/ST-II, and Solar8000/ART_SBP as the most important top 10 features. SHAP gave similar results as it considered Primus/EXP_SEVO, Primus/PAMB_MBAR, BIS/EMG, Solar8000/HR, Primus/MAC, Primus/MV, Primus/COMPLIANCE, Primus/FE02, Solar8000/ART_SBP, and Primus/ST-II as top 10 features contributing most towards the model's predictions. It can be seen in this case as well that 8 out of the 10 features highlighted as important by LIME and SHAP are common.

Further, we experimented with these subsets of the University of Queensland Vital Signs dataset and the VitalDB dataset, utilizing the 8 crucial features highlighted by explainable methods. The results have been summarised in Table 9. While the accuracy is slightly lower due to the utilization of a subset of features rather than the entire

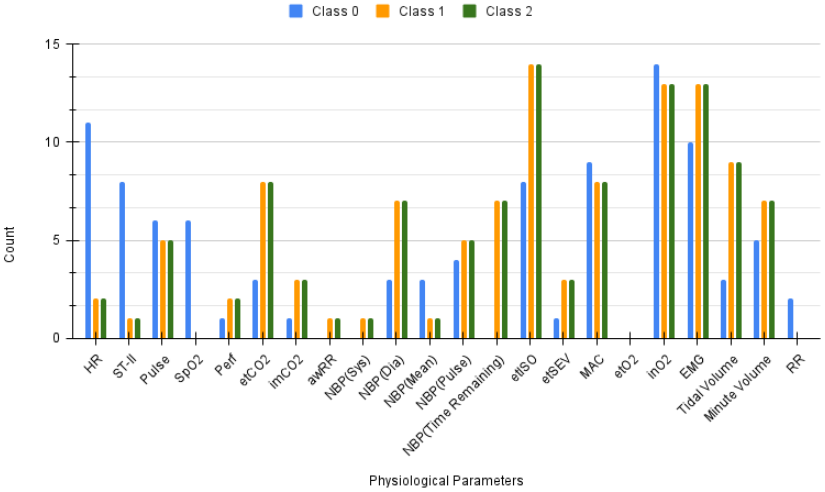


FIGURE 7. **Presence of physiological parameters in the top 10 features list, categorized by class, across 15 randomly selected examples for the University of Queensland Vital Signs dataset.**

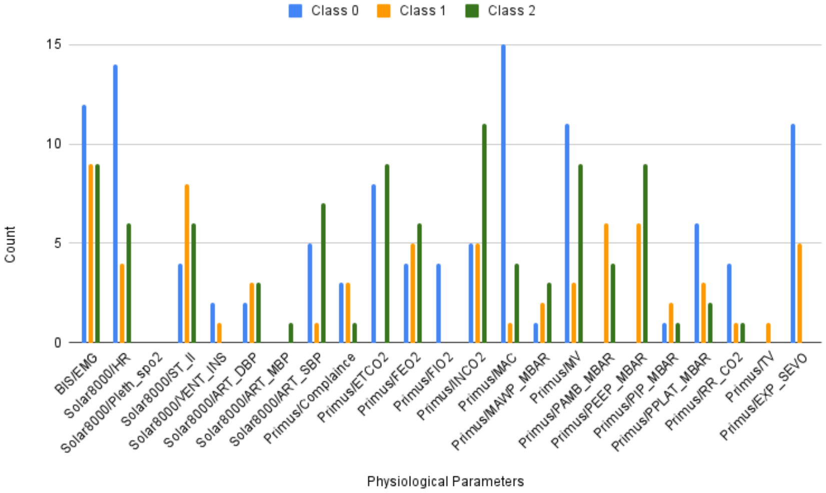


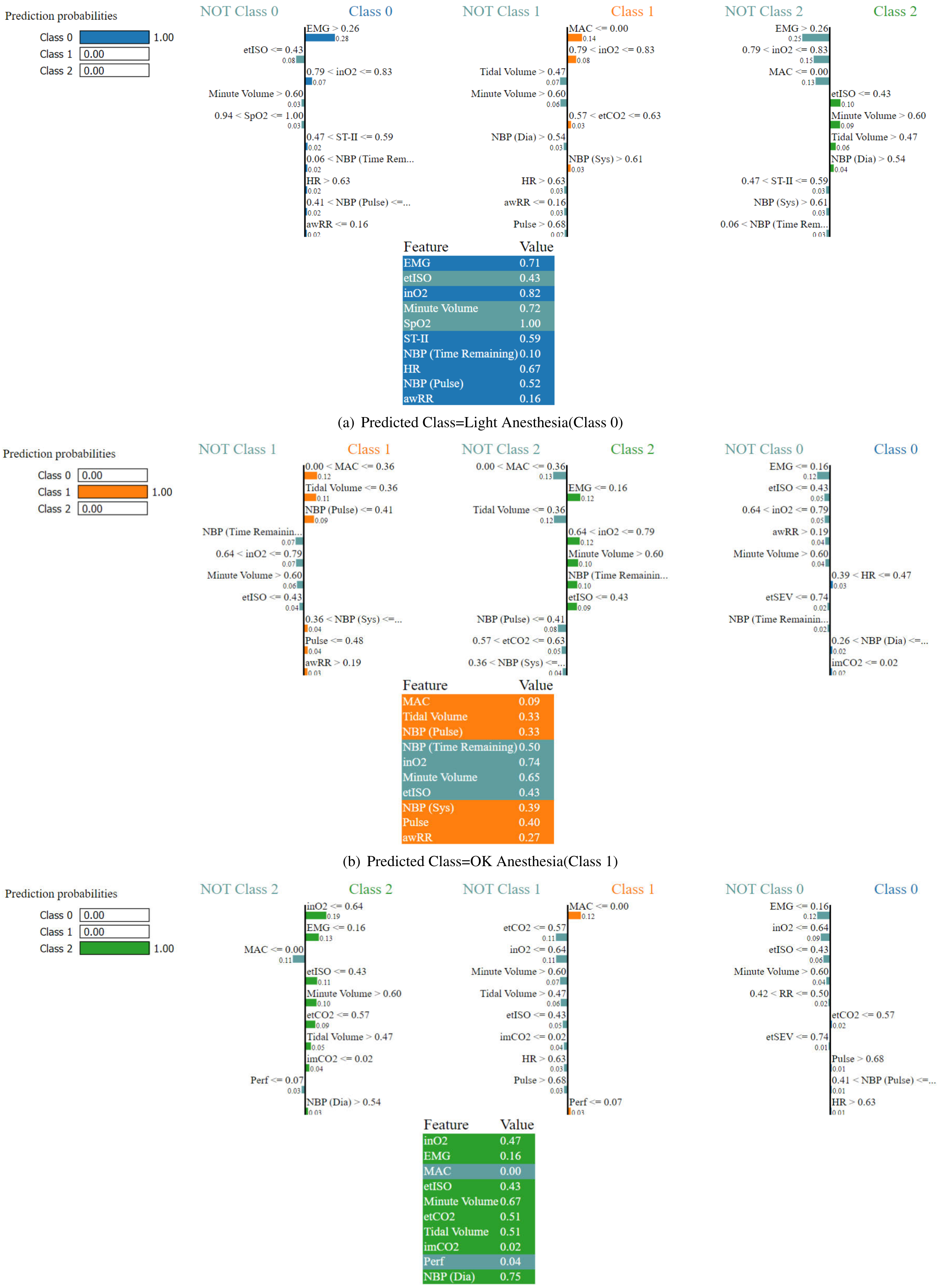
FIGURE 8. **Presence of physiological parameters in the top 10 features list, categorized by class, across 15 randomly selected examples for VitalDB dataset.**

University of Queensland Vital Signs dataset or the VitalDB dataset, it remains comparable.

In conclusion, the LIME and SHAP explainers provide consistent and interpretable insights into the crucial features influencing the XGBoost classifier's accurate predictions for both the University of Queensland Vital Signs and the VitalDB dataset. These findings enhance comprehension of the model's decision-making process and emphasize the consistency of feature importance across various datasets.

C. COMPARATIVE STUDY

This section has conducted a comparative analysis between our proposed framework and relevant prior studies within the anesthesiology domain. Our comparative assessment extends to state-of-the-art schemes focusing on accuracy and AUC scores. For instance, Zhan et al. [12] employed ECG signals, utilizing the Discrete Wavelet Transform for analyzing Heart Rate Variability (HRV) power, thus achieving an accuracy of 90.1%. Dubost et al. [42] employed a Hidden Markov Model (HMM) using physiological variables, achieving an accuracy of 52.87% in identifying consciousness states. Liu et al. [14] proposed the similarity and distribution index (SDI) based on HRV, showcasing a favorable assessment of depth of anesthesia compared to Bispectral Index (BIS), and achieving an AUC score of 0.95. Moreover, Sadrawi et al. [11] integrated EEG signals with mean vital signs, employing an Artificial Neural Network, resulting in an AUC score of 0.96. They further conducted a sensitivity analysis to evaluate feature importance. Tables 10 and 11 present a concise overview of the methodologies utilized for comparison. We have compared our model's



Feature	Value
EMG	0.71
etISO	0.43
inO2	0.82
Minute Volume	0.72
SpO2	1.00
ST-II	0.59
NBP (Time Remaining)	0.10
HR	0.67
NBP (Pulse)	0.52
awRR	0.16

(a) Predicted Class=Light Anesthesia(Class 0)

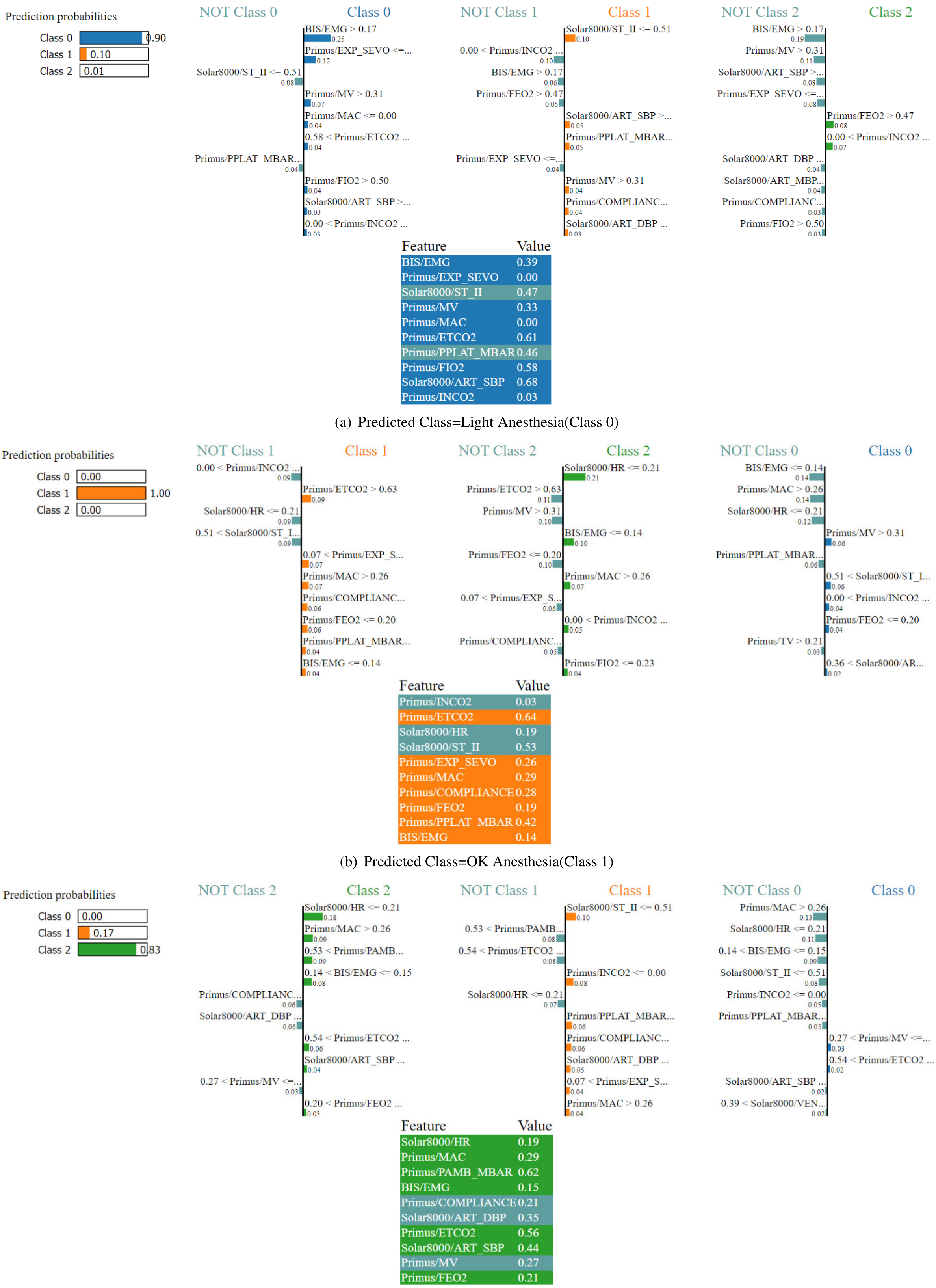
Feature	Value
MAC	0.09
Tidal Volume	0.33
NBP (Pulse)	0.33
NBP (Time Remaining)	0.50
inO2	0.74
Minute Volume	0.65
etISO	0.43
NBP (Sys)	0.39
Pulse	0.40
awRR	0.27

(b) Predicted Class=OK Anesthesia(Class 1)

Feature	Value
inO2	0.47
EMG	0.16
MAC	0.00
etISO	0.43
Minute Volume	0.67
etCO2	0.51
Tidal Volume	0.51
imCO2	0.02
Perf	0.04
NBP (Dia)	0.75

(c) Predicted Class=Deep Anesthesia(Class 2)

FIGURE 9. Visualization of the specific impact of vital signs on the categorization of an individual test instance using LIME alongside the XGBoost classifier, applied to the University of Queensland Vital Signs dataset.



(a) Predicted Class=Light Anesthesia(Class 0)

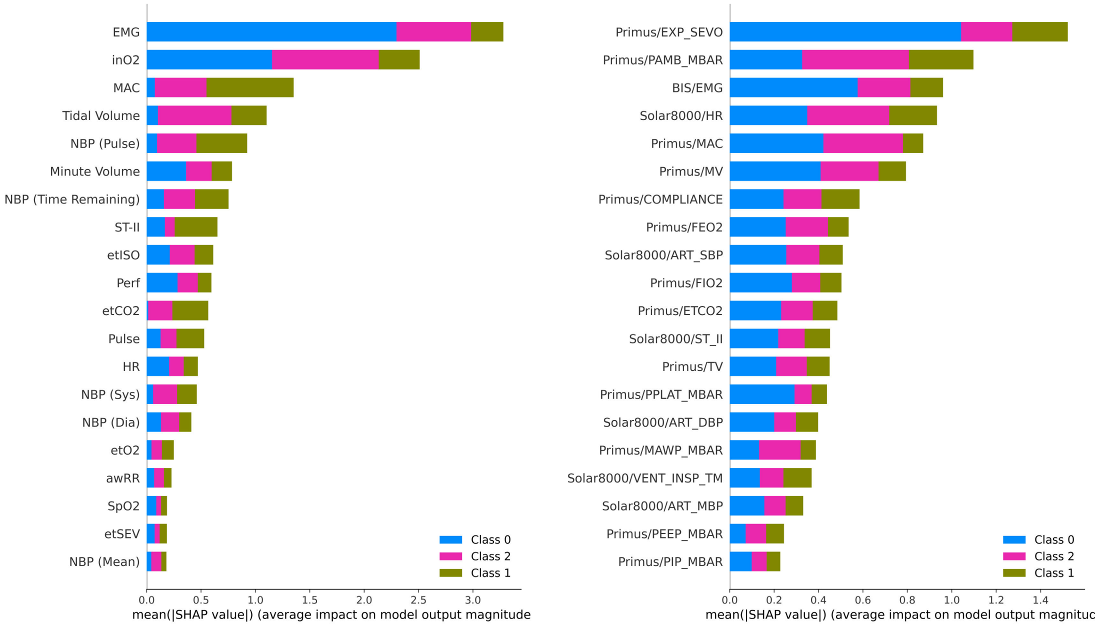
(b) Predicted Class=OK Anesthesia(Class 1)

(c) Predicted Class=Deep Anesthesia(Class 2)

FIGURE 10. Visualization of the specific impact of vital signs on the classification of an individual test instance using LIME alongside the XGBoost classifier, applied to the VitalDB dataset.

TABLE 9. Performance evaluation of XGBoost classifier: using only 8 features highlighted as important by LIME and SHAP.

Dataset used	Accuracy	Precision	Recall	F1-Score
The University of Queensland Vital Signs Dataset	98.84	98.85	98.84	98.84
VitalDB Dataset	91.59	91.55	91.59	91.52



(a) The University of Queensland Vital Signs Dataset

(b) The VitalDB Dataset

FIGURE 11. SHAP summary plot illustrating the importance of various physiological markets in the proposed model, visualizing the impact of each feature (vital sign) on the model's output predictions.

performance with other works in the literature, utilizing metrics such as AUC (Area Under the Curve) and accuracy wherever available to ensure a comprehensive evaluation of its effectiveness. These tables highlight the remarkable performance of our proposed method, achieving an accuracy of 98.02%, which depicts the superior performance achieved by the proposed scheme in determining the DoA. Notably, our approach solely relies on numeric vital signs such as Heart Rate (HR), Non-Invasive Blood Pressure (NBP), pulse, and other readily available metrics typically found in hospital settings.

Table 10 compares the proposed scheme's accuracy with other state-of-the-art methods. Zhan et al. [12] used heart rate variability (HRV) and a Deep Neural Network (DNN), achieving an accuracy of 90.1%. Dubost et al. [42] employed various vital signs, including heart rate, mean blood pressure, respiratory rate, and AA inspiratory concentration, with a Hidden Markov Model, resulting in an accuracy of 52.87%. In contrast, the proposed scheme, which utilizes vital signs such as heart rate, blood pressure, and respiratory rate with an XGBoost Classifier, achieved a significantly higher accuracy of 99.34%, indicating superior performance in classification accuracy compared to the other methods.

Table 11 compares the proposed work with state-of-the-art schemes regarding AUC (Area Under the Curve) scores. The AUC score is a performance metric for classification models, representing the area under the ROC (Receiver Operating Characteristic) curve. It measures the model's ability to distinguish between classes, with a higher AUC indicating better performance. Liu et al. [14] utilized heart rate variability (HRV) and an Artificial Neural Network (ANN) method, achieving an AUC score of 0.952. Subramanian et al. [15] also focused on HRV but employed Logistic Regression, resulting in an AUC score of 0.825. Sadrawi et al. [11] used vital signs such as heart rate, pulse, and blood pressure with ANN, attaining an AUC score of 0.96. Another work by Subramanian et al. [16] included heart rate variability, electrodermal activity, and blood pressure with Logistic Regression, yielding an AUC score of 0.80. In comparison, the proposed scheme, which uses vital signs including heart rate, blood pressure, and respiratory rate with an XGBoost Classifier, achieved a superior AUC score of 0.996, indicating a higher performance in classification accuracy.

Our methodology is exclusively reliant on easily accessible vital signs, which significantly enhances cost-efficiency and accessibility within clinical settings. Furthermore, the

TABLE 10. Comparison of proposed work with state-of-the-art schemes (in terms of accuracy).

Works	Features Used	Methods Used	Accuracy
[12]	Heart rate variability (HRV)	DNN	90.1%
[42]	Different vital signs (Heart Rate (HR), Mean Blood Pressure (MeanBP) Respiratory Rate (RR), and AA Inspiratory Concentration (AAFi))	Hidden Markov Model	52.87%
Proposed Scheme	Vital Signs (such as HR, BP, RR, etc.)	XGBoost Classifier	99.34%

TABLE 11. Comparison of proposed work with state-of-the-art schemes (in terms of AUC score).

Works	Features Used	Methods Used	AUC Score
[14]	Heart rate variability (HRV)	ANN	0.952
[15]	Heart rate variability (HRV)	Logistic Regression	0.825
[11]	Different Vital signs (such as heart rate (HR), pulse, systolic blood pressure (SBP), diastolic blood pressure (DBP), electromyography (EMG))	ANN	0.96
[16]	Heart rate variability, electrodermal activity, blood pressure	Logistic regression	0.80
Proposed Scheme	Vital Signs (such as HR, BP, RR, etc.)	XGBoost Classifier	0.996

integration of XAI increases the reliability and trustworthiness of our model. XAI empowers healthcare professionals by elucidating influential vital signs or physiological parameters (vital signs) in decision-making, augmenting clinical decision-making processes and amplifying the transparency and accountability of the model's predictions. We employed two separate datasets: the University of Queensland Vital Signs Dataset and the VitalDB Dataset. This approach was chosen for several important reasons. First, conducting the experiments on multiple datasets enhances the generalizability of the results, as it allows for the validation of findings across different populations and clinical environments. Second, by running the experiments independently on both datasets, we minimize the risk of dataset-specific biases, thereby increasing the reliability and validity of the study's conclusions.

In the proposed work, we tried our level best to find as many datasets as possible from different ethnicities. However, based on their availability, we were able to use only two datasets. One is the University of Queensland Vital Signs Dataset based in Australia, and the other is the VitalDB dataset based in the Republic of Korea. The use of datasets from two different ethnicities potentially increases the generalizability of the proposed work across patients from different populations.

VI. CONCLUSION

Traditional methods for monitoring the Depth of Anesthesia (DoA) have demonstrated efficacy in clinical settings; however, these are not standardized as patients' responses to anesthetic drugs vary with age, weight, ethnicity, and other factors. On the other hand, machine learning and deep learning models have become powerful tools for DoA assessment, offering improved precision and predictive capabilities, but often suffer from opacity, hindering interpretability and undermining trust and comprehension.

In this work, we proposed XAI-VSDoA—an Explainable AI-based scheme using vital signs for the assessment of the depth of anesthesia. This interpretability boosts confidence in the model, supporting critical reflection on anesthesia management. Our findings underscore the potential of XAI-VSDoA as a valuable tool for clinical use, enhancing patient safety and decision-making in anesthesia. The integration of XAI will augment anesthesiologists' clinical decision-making and enhance trust in the system. Our study entailed diverse experiments encompassing machine learning and deep learning methodologies, yielding noteworthy results. Leveraging two publicly available datasets, we achieved favourable outcomes, with the models exhibiting exceptional performance using numeric vital signs as inputs. Specifically, the XGBoost classifier achieved the highest accuracy of 99.34% with the University of Queensland Vital Signs dataset and 93.07% with the VitalDB dataset. To address the requirement for interpretability, we utilized XAI techniques—LIME and SHAP—which facilitated the identification of the top 10 features significantly influencing the model's predictions. These techniques consistently identified the same set of influential features, bolstering the reliability of interpretation and providing a robust understanding of model predictions. When re-evaluated using the top-ranked physiological parameters as suggested by these XAI techniques, the proposed model, XAI-VSDoA, showed comparable performance, highlighting the importance of the selected features. This interpretability reinforces confidence when the model aligns with established norms and encourages critical reflection when presenting alternative features.

A limitation of this study is the limited sample size and lack of ethnic diversity in the data, potentially decreasing the generalizability of our proposed technique. While using two datasets mitigates this issue to some extent, a more comprehensive dataset encompassing diverse patient demographics is necessary. In future, in addition to vital signs, we intend to augment the proposed model by integrating other biosignals (such as ECG, EEG, and PPG) and additional clinical information. This augmentation aims to enrich the dataset, providing a more nuanced understanding of the factors influencing the DoA. Further, we intend to use more diverse datasets from other ethnicities also to enhance the generalizability of the proposed scheme.

REFERENCES

[1] D. Rani and S. Harsoor, ''Depth of general anaesthesia monitors,'' *Indian J. Anaesthesia*, vol. 56, no. 5, pp. 437–441, 2012, doi: 10.4103/0019-5049.103956. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3530997/

[2] N. K. Sharma, S. Shahid, S. Kumar, S. Sharma, R. K. Gupta, and N. Kumar, ''Predicting depth of anesthesia using EEG signals and deep convolution network,'' in *Proc. 3rd Int. Conf. Artif. Intell. Mach. Learn. Syst.*, May 2024, pp. 1–8, doi: 10.1145/3639856.3639863.

[3] B. A. Orser, C. D. Mazer, and A. J. Baker, ''Awareness during anesthesia,'' *CMAJ*, vol. 178, no. 2, pp. 185–188, Jan. 2008, doi: 10.1503/cmaj.071761. [Online]. Available: https://www.cmaj.ca/content/178/2/185

[4] P. S. Myles, ''Prevention of awareness during anaesthesia,'' *Best Pract. Res. Clin. Anaesthesiol.*, vol. 21, no. 3, pp. 345–355, Sep. 2007.

[5] C. Rosow and P. J. Manberg, ''Bispectral index monitoring,'' *Anesthesiology Clinics North Amer.*, vol. 19, no. 4, pp. 947–966, 2001.

[6] I. J. Rampil, ''A primer for EEG signal processing in anesthesia,'' *Anesthesiology*, vol. 89, no. 4, pp. 980–1002, Oct. 1998.

[7] H.-C. Lee, H.-G. Ryu, Y. Park, S. B. Yoon, S. M. Yang, H.-W. Oh, and C.-W. Jung, ''Data driven investigation of bispectral index algorithm,'' *Sci. Rep.*, vol. 9, no. 1, p. 13769, Sep. 2019, doi: 10.1038/s41598-019-50391-x. [Online]. Available: https://www.nature.com/articles/s41598-019-50391-x

[8] S. Mathur, J. Patel, S. Goldstein, J. M. Hendrix, and A. Jain, ''Bispectral index,'' in *StatPearls*. Treasure Island, FL, USA: StatPearls, 2023. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK539809/

[9] T.-N. Li and Y. Li, ''Depth of anaesthesia monitors and the latest algorithms,'' *Asian Pacific J. Tropical Med.*, vol. 7, no. 6, pp. 429–437, Jun. 2014, doi: 10.1016/S1995-7645(14)60070-5. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1995764514600705

[10] C. W. Connor, ''Artificial intelligence and machine learning in anesthesiology,'' *Anesthesiology*, vol. 131, no. 6, pp. 1346–1359, Dec. 2019, doi: 10.1097/ALN.0000000000002694. [Online]. Available: https://pubs.asahq.org/anesthesiology/article/131/6/1346/922/Artificial-Intelligence-and-Machine-Learning-in

[11] M. Sadrawi, S.-Z. Fan, M. F. Abbod, K.-K. Jen, and J.-S. Shieh, ''Computational depth of anesthesia via multiple vital signs based on artificial neural networks,'' *BioMed Res. Int.*, vol. 2015, pp. 1–13, Oct. 2015, doi: 10.1155/2015/536863. [Online]. Available: https://www.hindawi.com/journals/bmri/2015/536863/

[12] J. Zhan, Z.-X. Wu, Z.-X. Duan, G.-Y. Yang, Z.-Y. Du, X.-H. Bao, and H. Li, ''Heart rate variability-derived features based on deep neural network for distinguishing different anaesthesia states,'' *BMC Anesthesiol.*, vol. 21, no. 1, pp. 1–11, Mar. 2021.

[13] Y. Liu, P. Lei, Y. Wang, J. Zhou, J. Zhang, and H. Cao, ''Boosting framework via clinical monitoring data to predict the depth of anesthesia,'' *Technol. Health Care*, vol. 30, pp. 493–500, Feb. 2022.

[14] Q. Liu, L. Ma, R.-C. Chiu, S.-Z. Fan, M. F. Abbod, and J.-S. Shieh, ''HRV-derived data similarity and distribution index based on ensemble neural network for measuring depth of anaesthesia,'' *PeerJ*, vol. 5, p. e4067, Nov. 2017.

[15] S. Subramanian, R. Barbieri, P. L. Purdon, and E. N. Brown, ''Analyzing transitions in anesthesia by multimodal characterization of autonomic state,'' in *Proc. 11th Conf. Eur. Study Group Cardiovascular Oscillations (ESGCO)*, Jul. 2020, pp. 1–2.

[16] S. Subramanian, R. Barbieri, P. L. Purdon, and E. N. Brown, ''Detecting loss and regain of consciousness during propofol anesthesia using multimodal indices of autonomic state,'' in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBC)*, Jul. 2020, pp. 824–827.

[17] M. R. Chowdhury, R. Madanu, M. F. Abbod, S.-Z. Fan, and J.-S. Shieh, ''Deep learning via ECG and PPG signals for prediction of depth of anesthesia,'' *Biomed. Signal Process. Control*, vol. 68, Jul. 2021, Art. no. 102663.

[18] N. Bahador, J. Jokelainen, S. Mustola, and J. Kortelainen, ''Multimodal spatio-temporal–spectral fusion for deep learning applications in physiological time series processing: A case study in monitoring the depth of anesthesia,'' *Inf. Fusion*, vol. 73, pp. 125–143, Sep. 2021.

[19] Q. Liu, J. Cai, S.-Z. Fan, M. F. Abbod, J.-S. Shieh, Y. Kung, and L. Lin, ''Spectrum analysis of EEG signals using CNN to model patient's consciousness level based on anesthesiologists' experience,'' *IEEE Access*, vol. 7, pp. 53731–53742, 2019, doi: 10.1109/ACCESS.2019.2912273.

[20] S.-J. Chen, C.-J. Peng, Y.-C. Chen, Y.-R. Hwang, Y.-S. Lai, S.-Z. Fan, and K.-K. Jen, ''Comparison of FFT and marginal spectra of EEG using empirical mode decomposition to monitor anesthesia,'' *Comput. Methods Programs Biomed.*, vol. 137, pp. 77–85, Dec. 2016, doi: 10.1016/j.cmpb.2016.08.024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260716302905

[21] J. M. Gonzalez-Cava, R. Arnay, A. León, M. Martín, J. A. Reboso, J. L. Calvo-Rolle, and J. A. Mendez-Perez, ''Machine learning based method for the evaluation of the analgesia nociception index in the assessment of general anesthesia,'' *Comput. Biol. Med.*, vol. 118, Mar. 2020, Art. no. 103645, doi: 10.1016/j.compbiomed.2020.103645. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482520300391

[22] M. Wang, F. Zhu, C. Hou, D. Huo, Y. Lei, Q. Long, and X. Luo, ''Depth classification algorithm of anesthesia based on model fusion,'' *Multimedia Tools Appl.*, pp. 1–17, Mar. 2024.

[23] R. V. Anand, M. F. Abbod, S.-Z. Fan, and J.-S. Shieh, ''Depth analysis of anesthesia using EEG signals via time series feature extraction and machine learning,'' *Sci*, vol. 5, no. 2, p. 19, May 2023.

[24] N. K. Sharma, S. Kumar, and N. Kumar, ''HGSmark: An efficient ECG watermarking scheme using hunger games search and Bayesian regularization BPNN,'' *Biomed. Signal Process. Control*, vol. 83, May 2023, Art. no. 104633.

[25] S. Kumar, N. K. Sharma, and N. Kumar, ''WSOmark: An adaptive dual-purpose color image watermarking using white shark optimizer and Levenberg–Marquardt BPNN,'' *Expert Syst. Appl.*, vol. 226, Sep. 2023, Art. no. 120137.

[26] A. Anghel, N. Papandreou, T. Parnell, A. De Palma, and H. Pozidis, ''Benchmarking and optimization of gradient boosting decision tree algorithms,'' 2018, *arXiv:1809.04559*.

[27] T. Chen and C. Guestrin, ''XGBoost: A scalable tree boosting system,'' in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785. [Online]. Available: https://dl.acm.org/doi/10.1145/2939672.2939785

[28] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, ''CatBoost: Unbiased boosting with categorical features,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.

[29] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, ''LightGBM: A highly efficient gradient boosting decision tree,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–9.

[30] L. Breiman, ''Random forests,'' *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/a:1010933404324.

[31] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[32] F. Sağlam and M. A. Cengiz, ''A novel SMOTE-based resampling technique trough noise detection and the boosting procedure,'' *Expert Syst. Appl.*, vol. 200, Aug. 2022, Art. no. 117023.

[33] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, ''SMOTE: Synthetic minority over-sampling technique,'' *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[34] P. Gohel, P. Singh, and M. Mohanty, ''Explainable AI: Current status and future directions,'' 2021, *arXiv:2107.07045*.

[35] M. T. Ribeiro, S. Singh, and C. Guestrin, ''Why should I trust you?'' in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1135–1144.

[36] S. M. Lundberg and S.-I. Lee, ''A unified approach to interpreting model predictions,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 4768–4777.

[37] D. Liu, M. Görges, and S. A. Jenkins, ''University of Queensland vital signs dataset: Development of an accessible repository of anesthesia patient monitoring data for research,'' *Anesthesia Analgesia*, vol. 114, no. 3, pp. 584–589, 2012.

[38] H.-C. Lee, Y. Park, S. B. Yoon, S. M. Yang, D. Park, and C.-W. Jung, "VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients," *Sci. Data*, vol. 9, no. 1, p. 279, Jun. 2022.

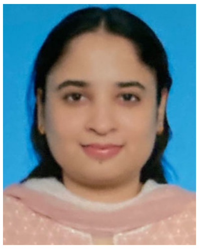
[39] R. Madanu, F. Rahman, M. F. Abbod, S.-Z. Fan, and J.-S. Shieh, "Depth of anesthesia prediction via EEG signals using convolutional neural network and ensemble empirical mode decomposition," *Math. Biosciences Eng.*, vol. 18, no. 5, pp. 5047–5068, 2021.

[40] Q. Wang, F. Liu, G. Wan, and Y. Chen, "Inference of brain states under anesthesia with meta learning based deep learning models," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 1081–1091, 2022, doi: 10.1109/TNSRE.2022.3166517.

[41] R. E. Walpole, R. H. Myers, S. L. Myers, and K. Ye, *Probability and Statistics for Engineers and Scientists*, vol. 5. New York, NY, USA: Macmillan, 1993.

[42] C. Dubost, P. Humbert, L. Oudre, C. Labourdette, N. Vayatis, and P.-P. Vidal, "Quantitative assessment of consciousness during anesthesia without EEG data," *J. Clin. Monitor. Comput.*, vol. 35, no. 5, pp. 993–1005, Oct. 2021.

NEERAJ KUMAR SHARMA received the Ph.D. degree in trust and reputation systems from the Department of Computer Science, University of Delhi. He is a Professor with the Department of Computer Science, Ram Lal Anand College, University of Delhi. He possesses teaching experience of around 21 years at the graduate and undergraduate level. His current research interests include applying machine learning, deep learning, and metaheuristic algorithms in the medical domain; digital watermarking; trusted AI; and trust and reputation systems in e-commerce.



SAKEENA SHAHID received the B.Sc. (Hons.) and M.Sc. degrees in computer science from the University of Delhi, India, where she is currently pursuing the Ph.D. degree with the Department of Computer Science, in the area of assessment of depth of anesthesia using machine learning and explainable AI. She is also an Assistant Professor with the Department of Computer Science, Sri Guru Tegh Bahadur Khalsa College, University of Delhi. Her research interests include the application of machine learning and deep learning in the medical domain. She is a member of ACM.

SUBODH KUMAR received the B.Sc. (Hons.) and M.Sc. degrees in computer science from the University of Delhi, India, and the Ph.D. degree in digital watermarking using machine learning and soft computing techniques. He is currently an Assistant Professor with the Department of Data Science and Analytics, Central University of Rajasthan, Rajasthan, India. His research interests include signal and image processing, machine learning, and soft computing techniques. He is a member of the IEEE Signal Processing Society and ACM.

SANJEEV SHARMA received the M.B.B.S. degree from the B. J. Medical College, Pune, and the M.D. degree in anesthesiology from the Maulana Azad Medical College, Delhi. He is currently a Professor with the Department of Anesthesiology, Atal Bihari Vajpayee Institute of Medical Sciences and Dr. Ram Manohar Lohia Hospital, New Delhi. He possesses around 23 years of teaching, research, and professional experience in anesthesia and critical care.

NAVEEN KUMAR received the Ph.D. degree in computer science from the Indian Institute of Technology (IIT) Delhi, New Delhi, India. He is a Professor with the Department of Computer Science, University of Delhi, New Delhi. He possesses around 41 years of teaching experience at the graduate level with the Department of Computer Science, University of Delhi, since 1983. His research interests include explainable AI, data mining, evolutionary algorithms, parallel computing, and the application of machine learning in the medical domain. He is a member of the Computer Society of India, the Institute of Electronics and Telecommunication Engineers, and ACM.

TANYA GUPTA received the degree in computer science from the Acharya Narendra Dev College, University of Delhi, in 2021, and the master's degree in computer science from the Department of Computer Science, University of Delhi. Her research interests include machine learning, data science, and deep learning.

RAKESH KUMAR GUPTA received the Ph.D. degree in microbiology from the National Diary Research Institute, India, and the Postdoctoral degree from the Center for Environmental Biotechnology, Department of Microbiology, University of Tennessee, Knoxville, TN, USA. He is a Professor and a Principal with the Ram Lal Anand College, University of Delhi. He possesses around 34 years of teaching experience at the undergraduate level. His current research interests include applied microbiology, environmental microbiome, molecular biology, and recombinant DNA technology.