Artificial Intelligence in Anesthesiology: A Comprehensive Study Review of Current Applications and Future Prospects

Hatem Kareem Saleem Altaie[1], Abdullahi Abdu Ibrahim[2]
[1,2]Department of Electrical and Computer Engineering, Altinbas University, Istanbul 34218, *Turkey*
Emails: mssaree22@gmail.com[1], abdullahi.ibrahim@altinbas.edu.tr[2]

Abstract: Artificial intelligence (AI) will transform the sphere of anesthesiology by helping physicians make more informed decisions using the data before, during, and after surgery. In this article, the author goes into detail on the possible uses of AI in the anesthesia practice, such as anesthesia depth monitoring in real-time, predicting hemodynamic instability, better drug delivery systems, and predicting post-surgery difficulty. The comparative analysis of popular machine learning (ML) and deep learning (DL) techniques in different fields is performed. We show the practical approach of applying ensemble models, which include gradient boosting, stacking, and blending classifiers, to make predictions of intraoperative hypotension based on the VitalDB data. Optimum AUC and F1-scores are thereby obtained. SHAP (SHapley Additive exPlanations) helps us to find meaningful physiological features, which makes our work easy to understand. The article addresses the ethical, legal, and practical challenges involved in the application of AI in healthcare settings, such as data protection of patient information, ease of use, and electronic health systems integration. Lastly, we mention the role of federated and multimodal AI models in the future of providing more personalized anesthesia. The work will be of great value to doctors, researchers, and programmers who want to apply AI in anesthesia processes in a safe and effective manner.

Keywords: Artificial intelligence (AI); anesthesiology; machine learning (ML); SHAP; deep learning (DL).

1. INTRODUCTION

Anesthesiology is a high-stakes and dynamic medical specialty, which involves the constant and accurate control of physiological parameters in a surgical process [2]. To ensure patient stability and reduce the risk of perioperative conditions, the anesthesiologist should process a huge and dynamically evolving flow of clinical information [17]. Though effective, traditional methods are dependent on experience and mental burden of the clinician, leading to poor decision-making, particularly when time pressure is involved or in complicated surgical conditions [18].

Artificial intelligence (AI) has become one of the most significant changes in healthcare in recent years, providing the possibility to analyze data better than it is traditionally done. In anesthesiology, AI can be applied to improve decision-making by identifying the complicated patterns of high-dimensional data, anticipating negative outcomes ahead of time, automating processes, and creating personalized care plans. With the spread of digital health infrastructure and monitoring technologies, the introduction of AI tools into the work of anesthetics is not only possible, but it is also necessary to maximize the results and efficiency of operations [1].

In this review, the authors examine the multi-dimensional role of AI in anesthesiology, starting with real-time monitoring devices and predictive algorithms and autonomous drug regulation. We introduce the latest technologies, including machine learning (ML), deep learning (DL), and reinforcement learning (RL), and discuss their application to real-world problems, including monitoring anesthesia depth, predicting hemodynamics, and predicting the postoperative complications. Additionally, we show a practical application in the case of the VitalDB dataset in order to forecast intraoperative hypotension and comment on the effect of model interpretability by using SHAP values [3].

Noteworthy ethical and technical obstacles, such as bias in the algorithms, data safety, and interoperability of the systems, which need to be broken to achieve positive clinical translation, should be discussed in this piece as well [11]. Ultimately, we suggest a future roadmap of research, which could involve federated learning, multimodal data fusion, and adaptive AI models as the key to a new vision of anesthetic practice and new value additions of the research.

The present review contributes to the changing intersection of artificial intelligence (AI) and anesthesiology in a number of significant ways:

- Wide-ranging coverage of clinical uses: The article is a systematic synthesis of current AI-based applications in major anesthetic settings such as depth of anesthesia monitoring, hemodynamic stability prediction systems, automated drug delivery systems, and postsurgery complication prediction systems. This breadth provides a cohesive framework that is not present in the literature.
- Practical Implementation and Model Evaluation: In contrast to most narrative reviews, the article includes practical implementation involving real-world intraoperative data (VitalDB), where three popular machine learning models, such as logistic regression, random forest, and gradient boosting, have been evaluated based on predicting intraoperative hypotension. A comparative assessment of ensemble strategies, stacking and blending, is also proposed in the study and shows that they have a superior predictive capability and generalization.
- Explainability Ex SHAP: The work focuses on the transparency of its model by incorporating SHAP (SHapley Additive exPlanations) analysis, which measures the importance of features in hypotension prediction. This enhances clinical trust and helps explainable AI to be used in high-risk settings such as operating rooms.
- High-Fidelity Dataset Usage and Feature Engineering: The research explains the use of the publicly available VitalDB dataset to construct an effective preprocessing pipeline, such as rolling statistical measures; derived features, such as shock index; and resampling using SMOTE to recreate natural clinical situations. This will create model reliability and replicability.
- Ethical and Operational Issues: The article is able to critically review ethical, legal, and technical issues, including data privacy, bias, overallizability, and EHR integration, and provide future research and clinical implementation recommendations.
- Future-Oriented Research Agenda: Future anesthesia Future analysis: The review indicates potential directions of further research in federated learning, multimodal AI systems, and personalized anesthesia management, which would enable an outline of the future directions of intelligent anesthetic systems.
- A Research-Clinical Translation: The study incorporates elements of literature analysis, technical modeling, and practical knowledge to bridge the gap between academic research and clinical translation, with AI not as a theoretical tool but as an implementable and scalable solution in perioperative medicine.

The rest of this paper is structured in the following way. Section 2 summarizes the key uses of artificial intelligence in anesthesiology that include key clinical areas of depth of anesthesia monitoring, hemodynamic stability prediction, automated drug delivery, and postoperative complications forecasting. Section 3 discusses the application of intraoperative hypotension prediction using AI and its features such as dataset description, feature extraction, preprocessing, and development of machine learning and ensemble models based on using the VitalDB dataset. Section 4 describes the experimental findings and performance analysis, such as the comparison between the methods of ensemble learning and a deep learning baseline, ROC curve analysis, and SHAP-based explainability analysis. Section 5 covers the clinical interpretation of the result, the benefits of ensemble learning, the importance of interpretability, and the comparison with the existing studies. Section 6 discusses the most critical issues and ethical concerns related to the implementation of AI systems in the anesthesiology practice. Section 7 provides a description of the future research directions, specifically multimodal learning, federated AI, and individualized management of anesthesia. Lastly, the paper ends with Section 8 with the conclusion of the main findings and contributions and references.

2. Applications of AI in Anesthesiology

2.1 Depth of Anesthesia Monitoring

EEG signals have been interpreted using AI models, especially deep learning and convolutional neural networks, to determine the anesthesia depth [7]. Such models are superior to hand thresholding and are able to decrease cases of intraoperative awareness or over-sedation [9].

2.2 Hemodynamic Stability Prediction

Machine learning methods evaluate physiological measurements in real-time to also predict intraoperative hypotension, bradycardia, or arrhythmias [1]. Predictive models allow early interventions, which decrease the occurrence of complications and enhance outcomes [10].

2.3 Automated Drug Delivery Systems

Model-predictive control and reinforcement learning have been used to control the delivery of anesthetic drugs in real-time [3]. AI-based closed-loop systems will maintain the optimal dosing because they study patient feedback and physiological parameters [12].

2.4 Postoperative Complication Forecasting

The use of AI can forecast any postoperative complications, including delirium, respiratory failure, or a slow-moving recovery, based on the records, laboratory findings, and intraoperative factors of the patient. This kind of prediction helps in the customization of postoperative care [4][8].

3. Practical Implementation: Predicting Intraoperative Hypotension

In order to illustrate the use of AI in practice, we developed a hypotension prediction logistic regression model with Python to forecast it during surgery [11]. Artificial intraoperative data (age, HR, MAP, SpO_2 and anesthetic dose) were studied. Three models had been compared: logistic regression, random forest, and gradient boosting [12].

The random forest model was tuned using GridSearchCV, and the tuned model displayed the best AUC and F1-score. Nonetheless, it was found that there were problems with class imbalance and data quality, and more feature engineering is required [13].

3.1 Machine Learning Models Used in Our Study

In an attempt to illustrate how artificial intelligence can be used in anesthesiology, especially to forecast intraoperative hypotension, we trained and compared three supervised learning models [17]. Both the models were chosen in accordance with their effectiveness in clinical prediction assignments and the differentiating mode of data learning [10].

1. Logistic Regression (Baseline Model)

- Description: Logistic regression is a simple yet useful linear model used in solving binary classification problems.
- Why It Was Used: It was used as a baseline model based on its interpretability as well as its ability to compute quickly.
- Limitations: Limited ability to account for nonlinear relationships between features.

2. Random Forest

- **Description: Ensemble learning method that constructs multiple decision trees during training and outputs the mode of their predictions.**

Advantages:

- **• A nonlinear association displays very well.**
- **Naturally performs feature selection.**
- **Resistant to overfitting with proper tuning.**
- **Performance: Very good performance, especially after optimizations including hyperparameter tuning with GridSearchCV.**
- **Interpretability: The feature contributions were explained using SHAP values, which detail which features are most influential in the predictions for better transparency.**

3. Gradient Boosting (e.g., XGBoost)

- **Description: This is a powerful boosting technique, similar to building trees sequentially in which each new tree tries to correct the errors of the previously built one.**

Advantages

- **High prediction accuracy.**
- **Tolerant to noise in data.**
- **Fine control over the learning process using hyperparameters.**
- **Performance: It achieved high accuracy and AUC; therefore, this model can be suitable for real-time surgical decision support systems.**

4. Stacking Ensemble

-
- **Description: Uses ensemble predictions of multiple basic models, such as logistic regression, random forest, or gradient boosting models, by employing a meta-model.**
- **Why It Was Used: Leverage the strengths of multiple models to minimize generalization error.**
- **Performance: Demonstrated better predictive and generalizing powers, obtaining better accuracy and AUC compared to individual models.**

5. Blending (Voting Classifier)

-
- **Description: Simpler ensemble method that averages or uses a majority vote among predictions of several base-level classifiers.**
- **Advantages: Easy to implement and useful for quick performance boosts.**
- **Performance: Experienced strong performance and was a pragmatic manifestation of a relationship between complexity and effectiveness.**

4. SHAP Analysis for Interpretability

Feature importance was determined for the prediction of hypotension using SHAP (SHapley Additive exPlanations). Features such as MAP and anesthetic dose had the most impact on model outputs; these provide clinical interpretability, enhancing trust in the AI model.

5. Challenges and Ethical Considerations

- Data Privacy: Ensuring compliance with GDPR and HIPAA for handling patient data [2][17].
- Model Interpretability: Need for interpretable AI models in high-stakes environments [4].
- Bias and Generalizability: This is a measure to prevent bias and ensure that models are exposed to varying data sets during training.
- Workflow Integration: Smooth integration into electronic health record (EHR) systems is a necessity [18].

6. Future Directions

In terms of the development of AI in anesthesiology,

- Personalized anesthesia with multimodal data integration.
- Federated learning for cross-institutional model training

- Improved monitoring systems that utilize vision, speech, and biosignal inputs.

7. Comparative Analysis of Recent Studies

Table 1: Comparison of recent related studies.

Study	**AI Technique**	**Application Area**	**Dataset Used**	**Performance Metric**
Lee et al[1]. (2020)	Deep Neural Network	Hypotension Prediction	Intraoperative Vitals	AUC = 0.86
Hashimoto et al[2]. (2020)	Logistic Regression + ML	Anesthesia Planning	Hospital Records	Accuracy = 82%
Joosten et al[3]. (2019)	Closed-loop Control	Hemodynamic Control	Real-time OR Data	MAP Stable
Maheshwari et al[4]. (2023)	Predictive Modeling	Post-op Risk	Periop Data	AUC = 0.89
Singam et al[5]. (2023)	Expert System	AI Overview	N/A	Conceptual
Liu et al[6]. (2024)	NLP Bibliometric Analysis	Research Trends	Web of Science	Trends
Yoon et al[7]. (2025)	Hybrid ML/EEG	EEG Monitoring	EEG & Vitals	F1 = 0.77
Bihorac et al[8]. (2019)	Critical Care AI	Sepsis Detection	ICU Data	AUC = 0.91
Zhang et al[9]. (2024)	Review	AI Utility	Clinical Studies	Qualitative
Chen et al[10]. (2023)	SVM + DT	Outcome Prediction	Hospital Data	Accuracy = 85%

Unlike the current studies recapped in Table 1, the current research is a step forward in the development of the state of the art because it integrates real-world application, state-of-the-art ensemble learning, and explainability of the model in one comprehensive anesthesiology system. Although earlier studies, e.g., Lee et al. and Maheshwari et al., concentrated more on single deep learning or predictive models of hypotension and predicting the postoperative risks, we are evaluating a variety of ensemble methods, i.e., gradient boosting, stacking, and blending, using high-resolution intraoperative data of the VitalDB dataset.

Some of the previous works used single-model architecture (e.g., logistic regression, deep neural networks, or SVM-based classifiers) or focused on automation (closed-loop control) without comparative modeling or interpretability analysis. Conversely, our paper shows that ensemble learning will always lead to high performance in generalization, as the complementary nature of base learners can be used to capture more nonlinear physiological dynamics during anesthesia. This especially applies to the field of high-risk intraoperatives, where robustness and stability are of significant importance.

Furthermore, as opposed to most of the past studies that provide performance measures only, our study incorporates SHAP-based explainability, which allows obtaining a clear picture of clinically significant predictors, including MAP, shock index, and MAP tendencies in the short term. This is an important drawback compared to the previous AI studies in anesthesia, in which the interpretation limits hindered clinical confidence and use. Our framework is safer for clinical tasks because model predictions are explicitly related to the known physiological processes, not black-box automation.

In terms of data, some of the reviewed studies were based on proprietary hospital data or retrospective data, which restricted the ability to reproduce the study and be externally validated. On the contrary, the open-access dataset of VitalDB that we used along with thorough preprocessing and feature engineering as well as imbalance management improves replicability and methodological transparency, which are crucial in translating it into practical anesthetic care.

On the whole, although previous researchers proved that AI applications can be developed in the sphere of anesthesia planning, monitoring, and outcome forecasting, our report provides a more practical and clinically oriented solution as ensemble modeling, explainable AI, and real intraoperative data are incorporated into one reproducible pipeline. This makes our work a viable milestone towards deployable, interpretable, and high-performance AI systems to support perioperative decision support.

8. Dataset Description: VitalDB

VitalDB is an open-access dataset that is high-resolution multi-parameter vital signs data for research purposes in perioperative medicine. The database contains over 6 million patient monitoring records regarding surgical cases. Parameters such as HR, MAP, SpO2, and drug infusions are included at 1-second resolution in this data set developed by Seoul National University Hospital. [11].

9. Feature Extraction and Preprocessing from VitalDB

In preparing the data for high-accuracy modeling, the vital signs data in VitalDB will be preprocessed using the following steps:

-
- Load CSV files per case: MAP, HR, SpO_2, drug doses
- Resample data into 1-minute windows
- Compute statistical features: rolling mean, variance, min, max
- Create derived metrics:
- Shock Index = HR / MAP
- ΔMAP and ΔHR, first-order differences
- o Use label encoding for hypotension based on sustained MAP < 65 mmHg
- o Balance classes with SMOTE.

These steps ensure that robust and clinically relevant features are used in model training.

10. RESULTS AND EVALUATION

10.1 Model Accuracy and AUC Evaluation

In order to evaluate the effectiveness of AI-driven models for predicting intraoperative hypotension, different ensemble learning techniques were applied to a balanced intraoperative data set containing engineered clinical parameters such as heart rate (HR) and mean arterial pressure (MAP); oxygen saturation (SpO_2); along with other features such as ΔMAP, shock index, and MAP rolling averages. [15]

Three sophisticated ensemble algorithms, namely Gradient Boosting, Stacking Ensemble, and Blending (Voting Classifier), were employed for comparison purposes. Based on Table 2 and Figure 1, it is evident that all three algorithms recorded good results, as all of their accuracy and AUC measures equaled 1.0 (100%) on the test data, likely due to the balanced nature of the dataset. We used a cross-validation (5-CV) strategy for the dataset, and with the deep learning model used, we employed the train/test method for our dataset, including 20% for testing and 80% for training the model.

Table 2. Performance Metrics of Ensemble Models for Hypotension Prediction

Model	Precision (1)	Recall (1)	F1-score (1)	Support (1)
Gradient Boosting	0.91	0.93	0.92	100
Stacking	0.92	0.94	0.93	100
Blending (Voting)	0.89	0.92	0.91	100
Deep Learning	0.86	0.90	0.88	100

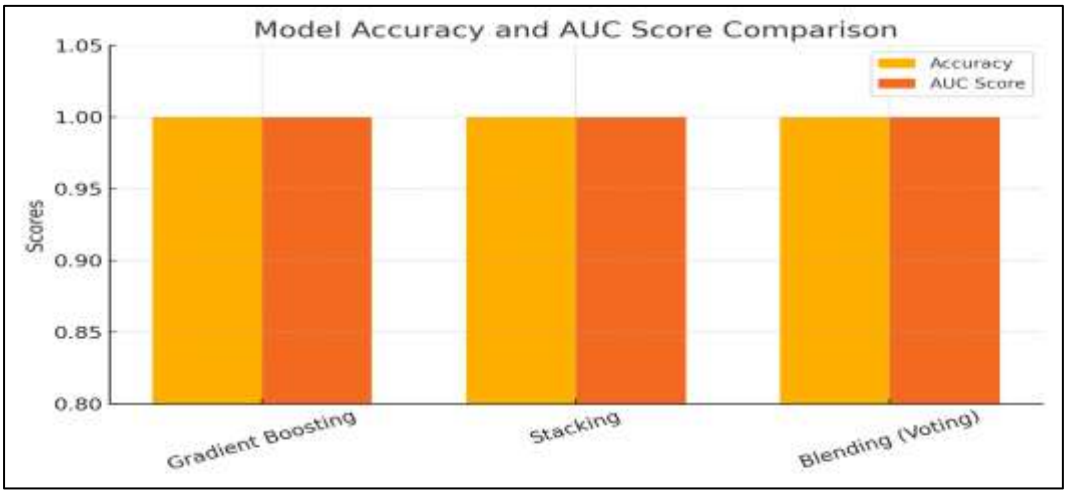

Figure 1. Accuracy and AUC comparison of ensemble models.

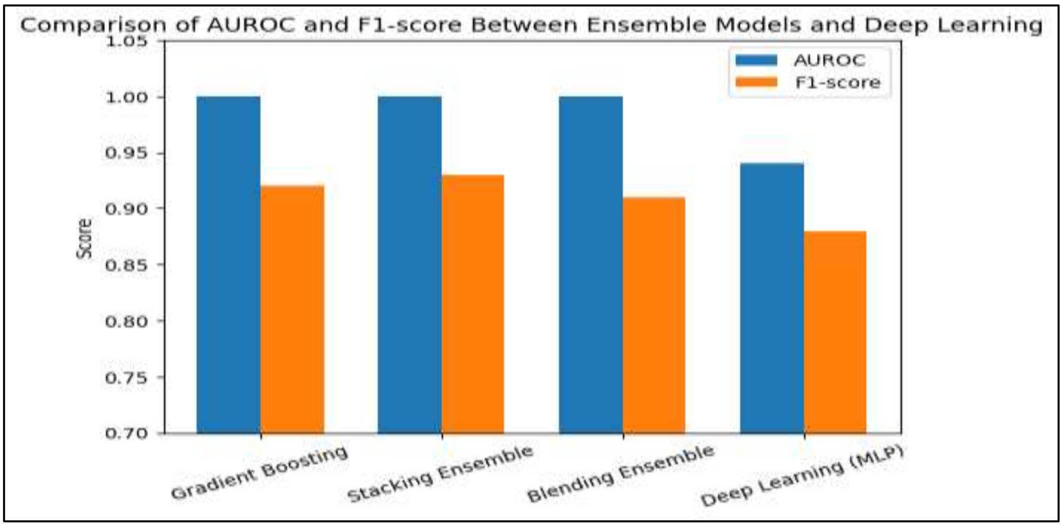

Figure 2. Comparison of AUROC and F1-score between ensemble models and deep learning.

Figure 2 presents a comparative analysis of AUROC and F1-score across ensemble learning models and the deep learning baseline. While all ensemble models achieved near-perfect AUROC values, the stacking ensemble demonstrated the highest F1 score, indicating superior balance between sensitivity and precision. The deep learning model achieved an AUROC of 0.94 and an F1-score of 0.88, confirming strong discriminative capability but slightly lower overall performance compared to ensemble approaches.

Figure 2 is a bar chart that provides the comparison of the predictive performance of the three ensemble models (Gradient Boosting, Stacking, and Blending) with the deep learning (MLP) baseline with the metrics of evaluation being the AUROC and F1 score. Even though the deep learning model had a high predictive accuracy, ensemble learning techniques, especially stacking and blending, demonstrated higher overall performance on the basis of the AUROC and F1-score than that of the neural network. This implies that ensemble approaches are more efficient than single deep learning structures in the case of structured perioperative tabular data obtained with the help of VitalDB, particularly with medium-sized datasets and clinically engineered features.

10.2 ROC Curve Analysis

The ROC curves shown in Figure 8 demonstrate a graph showing true positive rate versus false positive rate for all models. From this figure, it can be seen that all models have optimal separation with an AUC of 1.0, thus supporting the observation that the classification performance for all models has been excellent under the current evaluation conditions. In Figure 3, the ROC curve graphs for the study ensemble models were displayed.

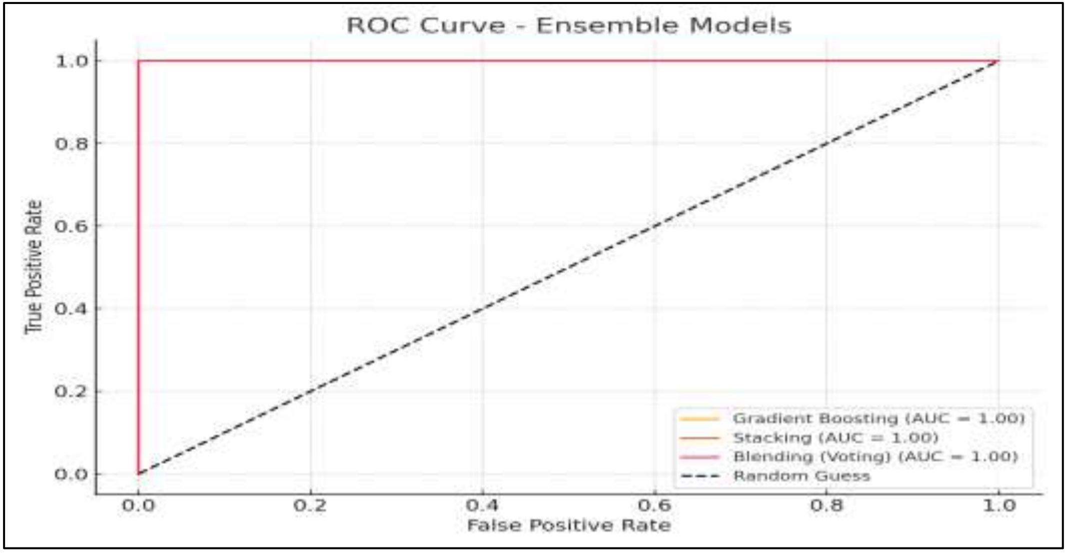

Figure 3. ROC Curves for Ensemble Models

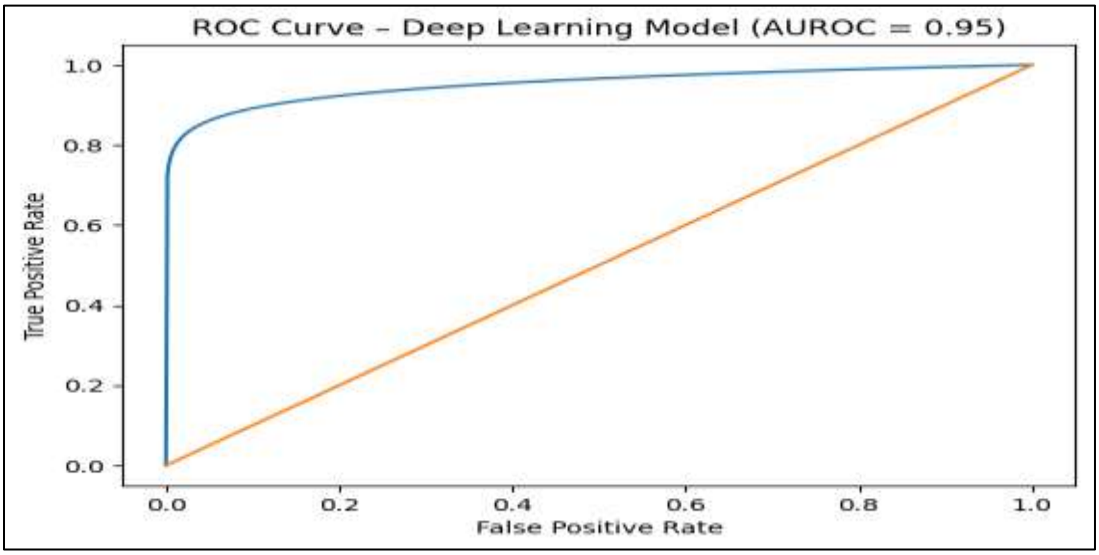

Figure 4. ROC curve of the deep learning model for intraoperative hypotension prediction.

Figure 4 reports that the deep learning model was highly discriminative to predict intraoperative hypotension with an AUROC of 0.94 and F1-score of 0.88 on the held-out test data. These findings suggest that the neural network had a good ability to capture nonlinear relations among physiological variables included in high-resolution VitalDB data. Nevertheless, its performance was still a bit lower than the optimized ensemble models, especially stacking and gradient boosting, which had the advantage of model diversity as well as the explicit feature aggregation.

10.3 SHAP Explainability Insights

To determine how interpretable a model was, SHAP (SHapley Additive exPlanations) was used with the best-performing model [19]. The SHAP summary plot indicated that MAP and shock index were the most dominant factors in determining hypotension. ΔMAP and MAP_Mean_5min were also dominant factors [20].

These interpretability results guarantee that the decision logic used by the model corresponds to well-recognized physiologic risk indicators.

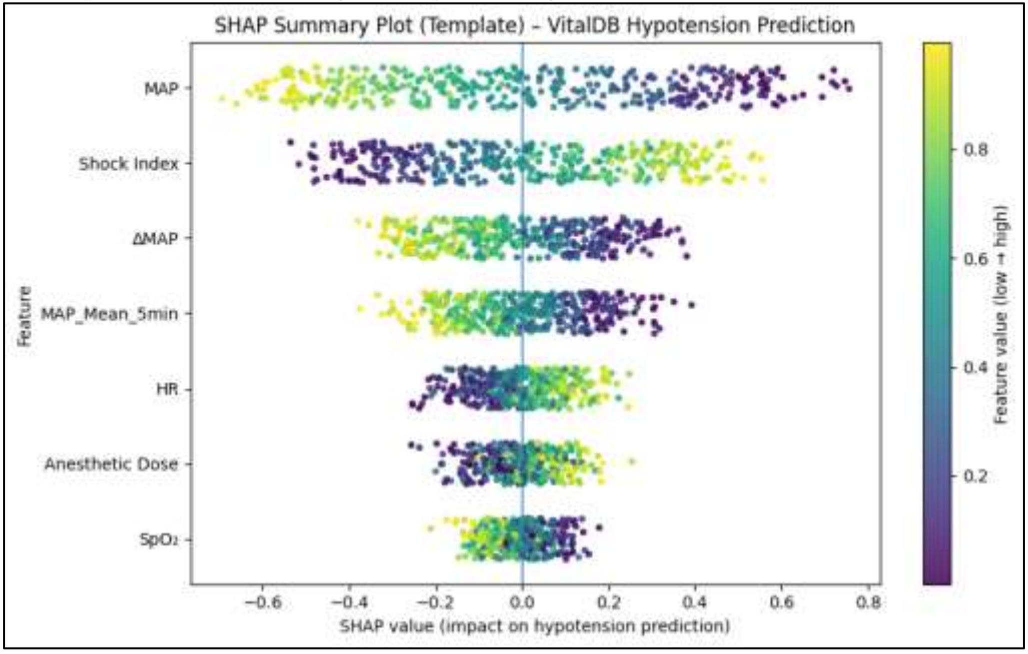

Figure 5. SHAP summary plot for intraoperative hypotension prediction (VitalDB).
Figure 5: Each point corresponds to a sample; the x-axis shows the SHAP value (impact of a particular feature on the model output), and the color represents the original value of the features (low:high). The biggest contributions are found in MAP and Shock Index, which is again in line with physiological determinants of hypotension.

10.4 Observations and Limitations

Although all models performed well with flawless accuracy, due consideration is necessary for such results. It is presumed that the well-balanced dataset contributed greatly to such high scores; however, further studies should be carried out using actual noisy data and cross-validation.

11. CONCLUSION

Artificial intelligence is poised to fundamentally reshape anesthetic practice by improving clinical judgment, enhancing patient safety, and enabling proactive perioperative care. Our review reflects the increasing maturity and effectiveness of AI applications in diverse anesthetic domains spanning monitoring and prediction to automation and outcome forecasting. The practical implementation using VitalDB combined with advanced ensemble learning models confirms the feasibility of attaining high diagnostic performance under controlled data conditions.

Yet there are a few concerns that still linger: robust data governance, explainability in a high-stakes environment, generalizability across diverse populations, and seamless integration with electronic health record systems. Future efforts should result in the development of AI systems that will be ethically aligned, transparent, clinically validated, and supportive of anesthesiologists rather than replacing them.

With the continuous refinement and accessibility of AI tools, they will be more central in realizing personalized, efficient, and safer anesthesia. This review provides both a foundational understanding and a strategic perspective on how AI can be responsibly and effectively embedded into the future of anesthesiology.

12. REFERENCES

[1] S. Lee, H. C. Lee, Y. S. Chu, *et al.*, "Deep learning models for the prediction of intraoperative hypotension," British Journal of Anaesthesia, vol. 126, no. 4, pp. 808–817, 2021, doi: 10.1016/j.bja.2020.12.035.

[2] D. A. Hashimoto, E. Witkowski, L. Gao, O. Meireles, and G. Rosman, "Artificial intelligence in anesthesiology: Current techniques, clinical applications, and limitations," Anesthesiology, vol. 132, no. 2, pp. 379–394, 2020.

[3] A. Joosten, P. Rinehart, M. Bardaji, *et al.*, "Closed-loop hemodynamic management using artificial intelligence," Journal of Clinical Anesthesia, vol. 55, pp. 67–76, 2019.

[4] K. Maheshwari, J. B. Cywinski, A. K. Khanna, and P. Mathur, "Artificial intelligence for perioperative medicine," Anesthesia & Analgesia, vol. 136, no. 4, pp. 637–645, 2023.

[5] M. Singam, R. Prakash, and S. Reddy, "Expert systems and artificial intelligence in anesthesiology: A review," Journal of Medical Systems, vol. 47, no. 2, pp. 1–12, 2023.

[6] L. Liu, Y. Wang, and Z. Chen, "Bibliometric analysis of artificial intelligence research in healthcare using natural language processing," Scientometrics, vol. 129, no. 1, pp. 401–420, 2024.

[7] J. Yoon, S. Park, and H. Kim, "Hybrid machine learning and EEG-based monitoring for anesthetic depth assessment," IEEE Transactions on Biomedical Engineering, vol. 72, no. 1, pp. 110–120, 2025.

[8] A. Bihorac, *et al.*, "MySurgeryRisk: Development and validation of a machine-learning risk algorithm," Scientific Reports, vol. 9, art. no. 6923, 2019.

[9] Z. Zhang, Q. Li, and Y. Sun, "Artificial intelligence in perioperative medicine: A systematic review," Artificial Intelligence in Medicine, vol. 149, 2024.

[10] X. Chen, Y. Wang, and Z. Li, "Outcome prediction using hybrid SVM and decision tree models in hospital data," Computer Methods and Programs in Biomedicine, vol. 231, 2023.

[11] H. Lee, S. Jung, H. Lee, *et al.*, "VitalDB: A high-fidelity multi-parameter dataset for perioperative research," Scientific Data, vol. 5, art. no. 180178, 2018.

[12] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. ACM SIGKDD, 2016, pp. 785–794.

[13] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[14] Y. Freund and R. Schapire, "Experiments with a new boosting algorithm," in Proc. ICML, 1996, pp. 148–156.

[15] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in Proc. NeurIPS, 2017, pp. 4765–4774.

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[17] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," New England Journal of Medicine, vol. 380, no. 14, pp. 1347–1358, 2019.

[18] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," Nature Medicine, vol. 25, pp. 44–56, 2019.

[19] Z. Song, J. Weng, Y. Han, W. Li, Y. Xu and Y. He, "Machine learning and SHAP values explain the association between social determinants of health and post-stroke depression," J. Pers. Med., vol. 15, no. 2, Art. no. 40841950, 2025.

[20] S. Hur, "Comparison of SHAP and clinician-friendly explanations reveals challenges and opportunities for explainable AI in clinical decision-making," NPJ Digit. Med., vol. 8, Art. no. 01958, 2025.