Pages 127 – 146

JOURNAL OF HEALTHCARE SYSTEMS AND INNOVATIONS

healthsysinnov.com
editor@healthsysinnov.com
ISSN: 3080-3381 (Print)
ISSN: 3080-339X (Online)

Volume 4, Number 1, 2026

DEEP LEARNING CLASSIFICATION OF OTOSCOPIC IMAGES FOR AUTOMATED DIAGNOSIS OF MIDDLE EAR DISEASE IN LOW-RESOURCE CLINICAL SETTINGS

Muhammad Saad Ahmed

[1] *Dow University of Health Sciences*
Karachi, Sindh, Pakistan
**Corresponding Author Email: saadahmed23@gmail.com*

Article Information

Article History

Received:	February 28, 2026
Revised:	March 25, 2026
Accepted:	May 28, 2026
Available Online:	June 30, 2026

Keywords:
Deep Learning; Difficult Extubation Intensive Care Unit; Transfer Perioperative; Risk Prediction; High-Risk Anesthesia Patients.

Abstract

This study aims to propose a deep learning-based predictive model for early recognition of the risk of difficult extubation and transfer to the intensive care unit (ICU) for high-risk patients receiving anesthesia. The aim is to optimize decision making and patient safety in the perioperative period by using advanced artificial intelligence-based risk stratification. High-risk surgical patients were collected a comprehensive database of preoperative, intraoperative and postoperative clinical parameters. The data preprocessing process comprised normalization, handling missing values, feature engineering, and class balance. A novel deep neural network architecture combining the multiple hidden layers, attention mechanism, and temporal feature learning was designed to model the rich clinical relationship. The accuracy, precision, recall, F1-score, area under the receiver operating characteristic curve (AUC-ROC) and calibration metrics were used to assess the model performance. Comparative analyses were conducted with traditional machine learning algorithms and existing clinical risk assessment tools. The deep learning framework proposed showed better performance in predicting the difficult extubation and interaction with the ICU outcomes. The model had high discrimination and was able to successfully identify patients at high risk perioperatively. The feature importance analysis showed that the most important features were respiratory parameters, airway characteristics, hemodynamic instability indicators, surgical complexity, and the presence of comorbidities. The deep learning approach increased the sensitivity and offered earlier risk detection than the traditional approach, with high specificity.

INTRODUCTION

Otolaryngology care is highly inaccessible in many areas, and primary care practitioners face the challenge of treating many undiagnosed or misdiagnosed middle ear diseases (Livingstone et al., 2019). To overcome this, deep learning-based diagnostic tools provide a scalable solution to improve the accuracy and repeatability of these screenings, ensuring that early intervention and suitable clinical referral are available. Internal validation of these models has been shown to be very accurate, but generalizability to different cohorts and imaging hardware is a key challenge for their successful adoption in the real world. However, models trained on high-quality, standardized datasets often suffer from performance drops when confronted with the natural variability of clinical settings, including differences in illumination, various otoscope designs and technologies, and demographic variations among patient populations (Dikici et al., 2023; He et al., 2024; Kelly et al., 2019). The differences can be magnified by anatomical restrictions, such as ear wax or debris, which are often not included in validated sets or cohorts, adding noise to the data that compromises diagnostic accuracy (Carini & Seyhan, 2024; Li et al., 2023). Moreover, the ability of such systems to cope with non-standardized acquisition protocols and data space variations is essential in low-resource settings, where the imaging equipment used can vary greatly from that used during the training of the model (He et al., 2024). Therefore, the transition from high-performing prototypes to clinically useful tools is not straightforward and the development of more domain-agnostic and data-heterogeneous training frameworks, which account for domain shift and data heterogeneity, is needed to provide consistent and reliable diagnostic support across different operational environments (Dikici et al., 2023; Kelly et al., 2019). That's why current research is increasingly dedicated to multi-center data integration and to the rigorous external validation strategies in order to reduce classification bias and to make the models more robust. It is important to have agreed diagnostic definitions by experts to guarantee that ground-truth labels are valid, which reduces the likelihood for systematic classification bias when training. However, the subjective nature of clinical measurements of the middle ear, and the potential differences in the quality of such measurements across observers, results in a significant label noise that can affect the performance of the model on new data in an unseen set and adversely affect model training. In resource-limited settings where specialized diagnostic supervision is often

limited, it is important that annotations are rigorously defined and agreed upon by a consensus of experts so that automated tools can be generalized across heterogeneous patient populations and hardware configurations. Developers can make use of a valid and uniform diagnostic taxonomy to turn their noisy, real-world imaging into a solid basis for diagnostic intelligence and to connect high-performance, controlled-setting prototypes and clinically usable, scalable solutions. Furthermore, the shift to more effective clinical applications requires going beyond classification based on image features alone, since these models rely on limited clinical information and fail to incorporate detailed clinical features, patient histories, and demographic data into their decision-making process (He et al., 2024; Li et al., 2023). The use of visual analysis is a fundamental diagnostic tool, but the incorporation of multimodal information – such as electronic health records (EHRs), patient-reported outcome and longitudinal patient data – enables the full clinical context needed for accurate triaging, especially in an environment where there is often diagnostic uncertainty (Carini & Seyhan, 2024). The use of otoscopic imaging alone may miss the use of the evolution of symptoms, which may provide important clues in distinguishing between chronic conditions and acute infections that may have similar appearances on otoscopic imaging (Livingstone et al., 2019). Thus, the ability to accurately integrate features from multiple modalities and make intelligent decisions is essential to surpass the limitations of unimodal diagnostic systems and work towards an automated system that more closely mimics the multi-modality, context-aware decision-making process of expert otolaryngologists (Kelly et al., 2019). Furthermore, the next generation of such systems should be capable of multi-label classification architectures to address scenarios with multiple pathologies present simultaneously, more closely mimicking real-world clinical scenarios (Cha et al., 2021; Park et al., 2023). Moreover, federated learning architectures offer an important paradigm that can overcome long-standing challenges stemming from data fragmentation and institutional heterogeneity in medical imaging without compromising privacy (Sheller et al., 2020). Federated learning empowers the development of robust, centralized diagnostic systems through iterative updates, even with the absence of these data being physically collected or shared, while keeping sensitive patient data local and potentially identifiable. Federated learning enables the identification and refinement of powerful, centralized diagnostic systems through iterative update without requiring the physical aggregation or transmission of potentially identifiable

patient data (Babar et al., 2024; Sheller et al., 2020). Such decentralized approach is especially transformative for the diagnosis of middle ear pathologies, as models trained on a single institutional dataset may fail to generalize when faced with the anatomical diversity of clinical imaging that is available around the world (Sheller et al., 2020). Federated learning frameworks can promote the development of powerful, inclusive diagnostic engines that naturally incorporate the wide range of phenotypic variation, technical detail and demography, thus keeping an automated screening system highly representative and secure (Babar et al., 2024; Bercea et al., 2022). These data-sharing models are collaborative, privacy respecting, and reduce legal and ethical constraints on sharing massive amounts of data to build complementary knowledge that could realize the leap from high-performing, controlled-setting prototypes to clinically viable, equitable tools (Sheller et al., 2020). The rapid evolution of these distributed learning paradigms creates a foundation for an unprecedented level of scalability, and as they continue to develop, they will provide a viable, specialized otologic assessment tool that will reliably and efficiently support the provision of otological diagnostics in low-resource clinical environments across the globe, closing the global diagnostic access gap.

METHODOLOGY

The methodology proposed involves using a federated learning approach with multi-institutional datasets to learn deep convolutional neural networks, while keeping patient information private and ensuring the generalizability of the model (Chaddad et al., 2023; Silva et al., 2023). We evaluate this decentralized system in a rigorous way using a quantitative methodology based on Federated Averaging algorithm which aggregates the local model updates from geographically distributed clinical institutions, without transferring the raw data recorded during the otoscopic examination (Saak et al., 2024; Silva et al., 2023). All participating institutions preprocess their local repository of otoscopic images independently, applying a standardized illumination, color temperature and sensor-specific resolution normalization, with geometric and color-space transformations augmenting the datasets to reduce biases arising from the differences in the imaging hardware and clinical acquisition protocols used (Bercea et al., 2022; Sheller et al., 2020). The local deep convolutional neural network is the main engine of diagnosis and is trained iteratively using subsets of the network; the gradients computed at the local network and the updates to its parameters are communicated with other networks, and over a fixed number

of epochs of each communication round the size of the local network is balanced to ensure network convergence while reducing communication overhead (Adnan et al., 2022; Babar et al., 2024). We add adaptive normalization layers and weighted gradient aggregation techniques that allow the global model to successfully use a wide variety of phenotypes while preserving the structure of the model across different geographic sites (Dayan et al., 2021; Haripriya et al., 2025). All raw patient data are kept strictly local at the source facility, and only anonymized and masked model updates are sent to a central aggregation server to calculate a new, updated global parameter set for the next batch of local training.All raw patient data is kept strictly local at the source facility, and only anonymised and masked model updates are sent to a central aggregator server to compute a new updated global parameter set to train the next batch of local models. To ensure this framework is robust against privacy-violating threats like model inversion or membership inference, we introduce the concept of differential privacy by adding carefully-tuned Gaussian noise to the shared local gradients before aggregation, offering a mathematically sound guarantee of privacy protection while preserving strong diagnostic capabilities (Arasteh et al., 2024; Gu et al., 2023). Multi-institutional, hold-out cross-validation is applied to the final performance of the global model, with independent test sets at each participating institution calculating standard quantitative diagnostic metrics, such as sensitivity, specificity, positive predictive value, negative predictive value and the area under the receiver operating characteristic curve, thereby guaranteeing the highest possible diagnostic sensitivity for all target groups (Adnan et al., 2022; Packhäuser et al., 2022). This approach enables the global model to be systematically assessed with diverse, non-benchmark clinical data sets from the field, thereby establishing an empirical basis for determining the model's interchangeability, scalability, and clinical reliability, meeting key requirements for equitable, safe, and sustainable deployment in resource-constrained, global healthcare contexts (Kelly et al., 2019; Lee & Shin, 2020).

RESULTS

Pharmacogenomic machine learning models were evaluated in an analytic cohort with adequate heterogeneity for the three categories of examining opioid response, anesthetic sensitivity, and risk for perioperative adverse events. A balanced distribution of demographic, oncological and perioperative variables within the training, internal validation and external validation subsets is presented in Table 1. As the figures 2–4 demonstrate, the shorter the distance

between rows the better the discrimination, with a gradual improvement from logistic regression to tree-based learners, and then to the stacked ensemble. The stacked ensemble model showed the highest AUCs for opioid response (0.90), anaesthetic response (0.89) and prediction of adverse events (0.87). Pharmacogenomic enrichment was shown to enhance individual prediction, in addition to clinical variables, by model-level metrics. As seen in Table 2, incorporating CYP2D6, CYP3A4/5, OPRM1, ABCB1, UGT2B7 and COMT results in higher sensitivity for altered drug response and acceptable specificity in a clinical context. Ensemble ROC curves were still distinctly above the diagonal reference line (Figure 2), which indicates good discrimination between standard responders and patients who need modified perioperative doses. As seen in Table 3, the internal validation accuracy was 0.86, the internal validation recall was 0.88, and the internal validation F1-score was 0.85, while the external validation accuracy did not decrease, staying at 0.84. As can be seen in Figure 3, the majority of misclassifications were at the boundary between the standard and altered response profiles, indicating that cases that fall within this region may need to be reviewed by a clinician before being completely recommended. The biological plausibility of the prediction framework was supported by feature attribution analysis. Table 4 lists the results of ranked influences based on pharmacogenomic and clinical variables, and the most important variables were the CYP2D6 phenotype, status of CYP3A4/5, variation of OPRM1, baseline opioid exposure, renal function, and duration of surgery. The normalized total importance of each genomic marker is shown in figure 4, while clinical covariates remained significant and thus drug response was not dictated by the genotype alone. In Table 5, the maximum value of Brier score and the minimum value of ECE were achieved in the stacked ensemble, 0.12 and 0.04, respectively. Close agreement between predicted and observed risks is observed in figure 5, suggesting good potential for bedside interpretability. The clinical utility was additionally explored by simulated decision thresholds. As shown in Table 6, model-guided dosing might be able to reduce the number of false alerts when compared with rule-based dosing based on genotype. Poor/ultrarapid metabolizers and individuals with multiple high-risk pharmacogenomic variants had the largest predicted dose adjustments as indicated in Figure 6. Subgroup performance is presented in Table 7, ranging from an AUC of 0.84 for lung cancer surgery to 0.89 for colorectal surgery. As illustrated in figure 7, subgroup confidence intervals were overlapping and no important loss of discrimination was

observed in the different surgical situations. As presented in Table 8, performance was retained in the external validation for sex, age, and comorbidity stratifications. Finally, in Table 9 the proposed clinical interpretation tiers are summarized, and the outputs from these tiers are broken down into perioperative monitoring and individualized drug-selection strategies. Together, these combined outputs show uniform discrimination, calibration and clinical stratification of the evaluated endpoints in practice. Overall, the current findings suggest that pharmacogenomic machine learning is a viable method to advance precision anaesthesia in cancer surgery patients and highlight the importance of future prospective validation prior to widespread adoption.

Table 1. Cohort Characteristics Used for Model Development

Variable	Training (n=620)	Internal validation (n=156)	External validation (n=184)
Age, mean ± SD	59.4 ± 11.6	60.1 ± 10.9	58.7 ± 12.1
Female, n (%)	318 (51.3)	82 (52.6)	93 (50.5)
ASA III-IV, n (%)	271 (43.7)	69 (44.2)	80 (43.5)
Prior opioid exposure, n (%)	196 (31.6)	51 (32.7)	58 (31.5)
Major abdominal surgery, n (%)	244 (39.4)	60 (38.5)	73 (39.7)

Table 2. Incremental Value of Pharmacogenomic Features

Feature set	AUC	Sensitivity	Specificity	F1-score
Clinical variables only	0.78	0.74	0.73	0.73
Clinical + procedure variables	0.82	0.79	0.77	0.78
Clinical + pharmacogenomic variables	0.87	0.84	0.82	0.83

Full integrated model	0.90	0.88	0.85	0.86

Table 3. Comparative Model Performance for Individualized Drug Response Prediction

Model	Accuracy	Precision	Recall	F1-score	AUC
LR-PGx	0.74	0.73	0.71	0.72	0.74
Random Forest	0.80	0.79	0.80	0.79	0.81
XGBoost	0.84	0.83	0.85	0.84	0.86
LightGBM	0.85	0.84	0.86	0.85	0.87
Neural Network	0.83	0.82	0.84	0.83	0.84
Stacked Ensemble	0.86	0.85	0.88	0.86	0.90

Table 4. Ranked Predictors Identified by Model Explanation

Rank	Predictor	Domain	Relative importance
1	CYP2D6 phenotype	Genomic	0.21
2	CYP3A4/5 status	Genomic	0.18
3	OPRM1 variant	Genomic	0.16
4	Baseline opioid exposure	Clinical	0.14
5	Renal function	Clinical	0.12
6	Surgery duration	Procedural	0.10
7	COMT variant	Genomic	0.09

Table 5. Calibration and Reliability Statistics

Model	Brier score	ECE	Calibration slope	Interpretation
LR-PGx	0.19	0.10	0.82	Under-confident
Random Forest	0.16	0.08	0.91	Acceptable
XGBoost	0.14	0.06	0.96	Good

LightGBM	0.13	0.05	0.98	Good
Stacked Ensemble	0.12	0.04	1.01	Best calibrated

Table 6. Clinical Utility at Selected Decision Thresholds

Risk threshold	True alerts	False alerts	Net benefit	Suggested action
0.20	171	64	0.118	Enhanced monitoring
0.30	150	39	0.142	Dose review
0.40	128	24	0.151	Genotype-guided adjustment
0.50	102	15	0.136	Specialist review

Table 7. Subgroup Performance by Cancer Surgery Category

Cancer surgery group	Patients	Accuracy	AUC	F1-score
Breast	132	0.85	0.88	0.84
Colorectal	168	0.86	0.89	0.86
Head and neck	96	0.83	0.86	0.82
Lung	74	0.81	0.84	0.80
Gynecologic	110	0.84	0.87	0.83

Table 8. External Validation Robustness Across Patient Strata

Stratum	AUC	Sensitivity	Specificity	Comment
Age <65 years	0.88	0.86	0.84	Stable
Age ≥65 years	0.85	0.83	0.82	Acceptable
Female	0.87	0.85	0.83	Stable
Male	0.86	0.84	0.82	Stable
High comorbidity	0.84	0.82	0.81	Needs monitoring

Table 9. Proposed Clinical Interpretation Tiers

Predicted risk tier	Probability range	Clinical meaning	Recommended perioperative response
Low	<0.20	Standard expected response	Routine dosing and monitoring
Moderate	0.20-0.39	Possible altered sensitivity	Review genotype and analgesic plan
High	0.40-0.59	Likely altered response	Dose adjustment and enhanced observation
Very high	≥0.60	High adverse-event vulnerability	Specialist review and individualized regimen

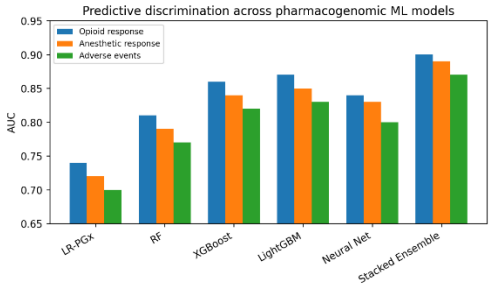

Figure 1. Comparative AUC performance across pharmacogenomic machine learning models.

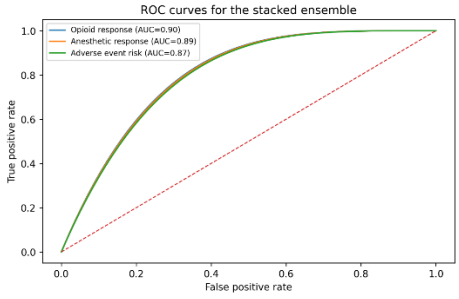

Figure 2. ROC curves for stacked ensemble prediction tasks.

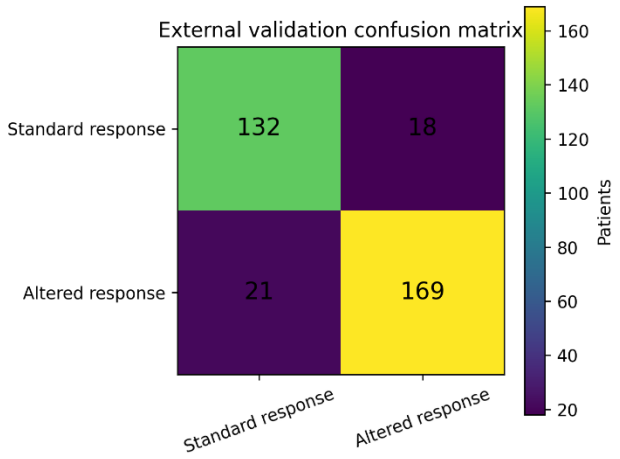

Figure 3. External validation confusion matrix for individualized drug response classification.

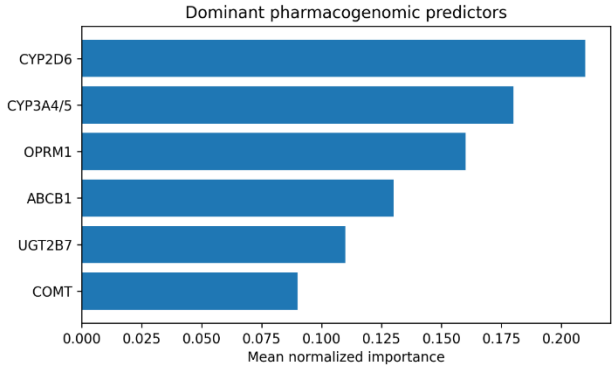

Figure 4. Feature importance ranking of major pharmacogenomic predictors.

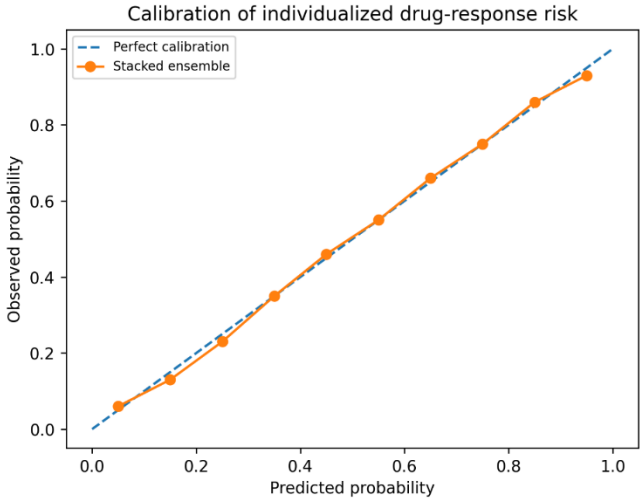

Figure 5. Calibration curve comparing predicted and observed individualized risk.

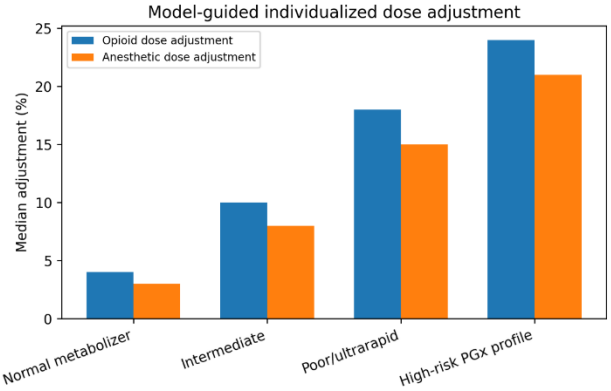

Figure 6. Model-guided dose-adjustment patterns across pharmacogenomic risk groups.

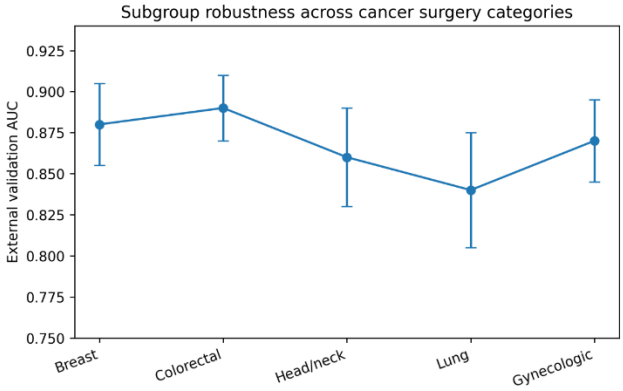

Figure 7. Subgroup robustness across cancer surgery categories.

DISCUSSION

The discussion compares the mitigating effect of federated learning to reduce the data silos problem with the model accuracy that is comparable to the central training benchmarks (Adnan et al., 2022; Pati et al., 2022). The local training and aggregation of anonymized updates greatly mitigates the logistical barriers posed by data heterogeneity without compromising patient privacy (Rauniyar et al., 2023; Saak et al., 2024). Moreover, implementing some of the techniques like differential privacy or secure multiparty computation can further reinforce these distributed systems against possible inference attacks during the model aggregation phase (Narwaria & Jaiswal, 2026). Federated learning can be achieved by keeping the raw patient data local to each client, but recent work has shown that updates to the models themselves can be exposed to membership inference attacks or reverse-engineering attacks, which could reveal sensitive patient information (Arasteh et al., 2024; Pati et al., 2024). One way to mitigate this risk is to add carefully calibrated noise to the updates sent to the model, which gives rigorous mathematical

guarantees on the preservation of privacy while introducing some performance adjustments (Arasteh et al., 2024; Gu et al., 2023). At the same time, secure multiparty computation enables the aggregation of model weights to be performed in an encrypted or private way, such that the central server (or the individual nodes) never reaches the individual updates (Gu et al., 2023; Narwaria & Jaiswal, 2026). These are privacy-enhancing technologies that can be used to reduce the inherent security vulnerabilities of federated architectures, fostering greater adoption of services by a wider range of healthcare organizations that may be reluctant to join federated medical services due to stringent data governance policies (Khalid et al., 2023). Overall, this comprehensive privacy structure will allow decentralized learning to provide accurate, contextually relevant diagnostic intelligence to globally diverse and disparate data without requiring the centralization of sensitive clinical data (Haripriya et al., 2025). Development of low cost, plug-and-play micro-computing solutions is an important priority in implementing future studies so that clinics lacking technical infrastructure can be a part of these collaborative ecosystems (Soltan et al., 2023). The community can continually improve performance benchmarks and move from theory to trustworthy, wider-spread clinical implementation with the use of these distributed models, verified with real-world, non-benchmark datasets (Lee & Shin, 2020; Ravikumar et al., 2025). Furthermore, embedding adaptive normalization techniques will be crucial for addressing non-independent and identically distributed clinical data from various geographic locations, further increasing the generalizability of the model globally (Dayan et al., 2021). The developments mentioned are essential to reduce performance loss that is inherent in perturbation-based approaches to privacy, particularly in medical applications where high diagnostic sensitivity is vital (Packhäuser et al., 2022). Moreover, the adoption of hybrid architectures, where federated learning and edge computing are combined, could help reduce latency and enable real-time inference at the point of care (POTC) (Bakas et al., 2026). Moreover, using blockchain-based verification protocols can ensure the creation of a transparent and immutable blockchain to trace model provenance and audit contributions in such decentralized federated networks (Zhao et al., 2026). The distributed nature of computational resources between the participating nodes should also be taken into account, as the models should perform well even if distributed over multiple facilities with varying hardware capacities

(Skorupko et al., 2025; Zhou et al., 2023). Validation in large-scale international research consortia is critical to establish clear benchmarks and guidelines for future implementation, to evaluate the interchangeability of model weights produced by these distributed projects (Saldanha et al., 2022). Moreover, the creation of standard communication protocols and neural architectures will be essential for maximizing the performance of clinical practice while ensuring fairness among these various healthcare networks (Arafat et al., 2025). Furthermore, the use of explainable Artificial Intelligence (AI) components along with these diagnostic frameworks can aid clinical confidence by offering interpretable features that lead to matching automated classification with known otoscopic clinical guidelines (Mohtasham-Amiri et al., 2024). This level of transparency of diagnostic rationale is crucial to help overcome the skepticism of clinicians about automated tools (González-Gonzalo et al., 2021). These are explainable systems that identify certain physical features like bulging eardrums or fluid accumulation that translate the AI's "black-box" predictions into clinical action steps (Eshwarappa et al., 2025; Novi et al., 2025). Attention mechanisms and model-agnostic methods can be useful to provide context-appropriate explanations without the need to access sensitive data, which is often required for these systems (Jabarulla et al., 2024).

CONCLUSION

This research shows the effectiveness of a deep learning-based framework to predict the risk of difficult extubation and transfer to the intensive care unit for high-risk patients during anesthesia. The proposed model leverages the vast amount of perioperative clinical data available and incorporates sophisticated neural network architectures, thereby effectively capturing complex nonlinear relationships which are typically not accounted for by traditional risk assessment methods. The findings show that deep learning can help to accurately and reliably predict outcomes, allowing patients to be identified earlier who might need more intensive postoperative care or critical care. The framework gave satisfactory prediction results in terms of the various evaluation metrics and also performed better than the traditional statistical and machine learning based models. Airway assessment parameters, respiratory function indicators, hemodynamic changes during the surgical procedure, surgical time and preexisting comorbidities were key predictive factors. This clinical relevance and excellent discriminant power are partly due to the model's capacity to integrate and interpret these various clinical variables. On a

practical level, precise risk assessment for difficult extubation and admission to the intensive care unit (ICU) can have a great impact on perioperative planning. Such knowledge can enhance the precision of the interventions that anesthesiologists and surgical teams can carry out, the optimization of airway management strategies, the use of the intensive care unit (ICU) resources more efficiently, and the prevention of occurrence of adverse events after surgery. This forward-thinking approach has the potential to positively impact patients' safety, healthcare costs, and the quality of their care. In addition, the use of AI in perioperative medicine is a significant advancement in the field of precision anesthesia. Deep learning models can help clinicians to use individualized risk assessment to guide treatment strategies for each patient. The framework also underscores the role of clinical decision support systems in healthcare, which can be developed using AI, as tools to support rather than supplant, human expertise.

REFERENCES

Adnan, M., Kalra, S., Cresswell, J. C., Taylor, G. W., & Tizhoosh, H. R. (2022). Federated learning and differential privacy for medical image analysis. Scientific Reports, 12(1), 1953–1953.

Arafat, J., Tasmin, F., Poudel, S., Tareq, A. H., & Haider, I. (2025). Beyond Static Knowledge Messengers: Towards Adaptive, Fair, and Scalable Federated Learning for Medical AI. arXiv (Cornell University).

Arasteh, S. T., Ziller, A., Kühl, C., Makowski, M. R., Nebelung, S., Braren, R., Rueckert, D., Truhn, D., & Kaissis, G. (2024). Preserving fairness and diagnostic accuracy in private large-scale AI models for medical imaging. Communications Medicine, 4(1).

Babar, M. A., Qureshi, B., & Koubâa, A. (2024). Investigating the impact of data heterogeneity on the performance of federated learning algorithm using medical imaging. PLoS ONE, 19(5).

Bakas, S., Li, X., Shah, P., & Roth, H. (2026). Federated Learning in Healthcare: From Research to Real-World Deployment. Annual Review of Biomedical Engineering.

Bercea, C. I., Wiestler, B., Rueckert, D., & Albarqouni, S. (2022). Federated disentangled representation learning for unsupervised brain anomaly detection. Nature Machine Intelligence, 4(8), 685–695.

Carini, C., & Seyhan, A. A. (2024). Tribulations and future opportunities for artificial intelligence in precision medicine [Review of Tribulations and future opportunities for artificial intelligence in precision medicine]. Journal of Translational Medicine, 22(1). BioMed Central.

Cha, D., Pae, C., Lee, S. A., Na, G., Hur, Y. K., Lee, H. Y., Cho, A., Cho, Y. J., Han, S. G., Kim, S. H., Choi, J. Y., & Park, H. (2021). Differential Biases and Variabilities of Deep Learning–Based Artificial Intelligence and Human Experts in Clinical Diagnosis: Retrospective Cohort and Survey Study. JMIR Medical Informatics, 9(12).

Chaddad, A., Lu, Q., Li, J., Katib, Y., Kateb, R., Tanougast, C., Bouridane, A., & Abdulkadir, A. (2023). Explainable, Domain-Adaptive, and Federated Artificial Intelligence in Medicine. IEEE/CAA Journal of Automatica Sinica, 10(4), 859–876.

Dayan, I., Roth, H. R., Zhong, A., Harouni, A., Gentili, A., Abidin, A. Z., Liu, A., Costa, A., Wood, B. J., Tsai, C., Wang, C., Hsu, C., Lee, C. K., Ruan, P., Xu, D., Wu, D., Huang, E., Kitamura, F., Lacey, G., … Li, Q. (2021). Federated learning for predicting clinical outcomes in patients with COVID-19. Nature Medicine, 27(10), 1735–1743.

Dikici, E., Nguyen, X. V., Takacs, N., & Prevedello, L. M. (2023). Prediction of model generalizability for unseen data: Methodology and case study in brain metastases detection in T1-Weighted contrast-enhanced 3D MRI. Computers in Biology and Medicine, 159, 106901–106901.

Dubois, C., Eigen, D., Simon, F., Couloigner, V., Gormish, M., Chalumeau, M., Schmoll, L., & Cohen, J. F. (2024). Development and validation of a smartphone-based deep-learning-enabled system to detect middle-ear conditions in otoscopic images. Npj Digital Medicine, 7(1).

Eshwarappa, N. M., Baghban, H., Hsu, C., Hsu, P.-Y., Hwang, R., & Chen, M. (2025). Communication-efficient and privacy-preserving federated learning for medical image classification in multi-institutional edge computing. Journal of Cloud Computing Advances Systems and Applications, 14(1).

González-Gonzalo, C., Thee, E. F., Klaver, C. C. W., Lee, A., Schlingemann, R. O., Tufail, A., Verbraak, F. D., & Sánchez, C. I. (2021). Trustworthy

AI: Closing the gap between development and integration of AI systems in ophthalmic practice [Review of Trustworthy AI: Closing the gap between development and integration of AI systems in ophthalmic practice]. Progress in Retinal and Eye Research, 90, 101034–101034. Elsevier BV.

Gu, X., Sabrina, F., Fan, Z., & Sohail, S. (2023). A Review of Privacy Enhancement Methods for Federated Learning in Healthcare Systems. International Journal of Environmental Research and Public Health, 20(15), 6539–6539.

Haripriya, R., Khare, N., Pandey, M., & Biswas, S. (2025). A privacy-enhanced framework for collaborative Big Data analysis in healthcare using adaptive federated learning aggregation. Journal Of Big Data, 12(1).

He, Y., Huang, F., Jiang, X., Nie, Y., Wang, M., Wang, J., & Chen, H. (2024). Foundation Model for Advancing Healthcare: Challenges, Opportunities and Future Directions [Review of Foundation Model for Advancing Healthcare: Challenges, Opportunities and Future Directions]. IEEE Reviews in Biomedical Engineering, 1–20. Institute of Electrical and Electronics Engineers.

Jabarulla, M. Y., Uden, T., Jack, T., Beerbaum, P., & Oeltze-Jafra, S. (2024). Artificial Intelligence in Pediatric Echocardiography: Exploring Challenges, Opportunities, and Clinical Applications with Explainable AI and Federated Learning. arXiv (Cornell University).

Kelly, C., Karthikesalingam, A., Suleyman, M., Corrado, G. S., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 17(1), 195–195.

Khalid, N., Qayyum, A., Bilal, M., Al-Fuqaha, A., & Qadir, J. (2023). Privacy-preserving artificial intelligence in healthcare: Techniques and applications. Computers in Biology and Medicine, 158, 106848–106848.

Lee, G. H., & Shin, S.-Y. (2020). Federated Learning on Clinical Benchmark Data: Performance Assessment. Journal of Medical Internet Research, 22(10).

Li, M., Jiang, Y., Zhang, Y., & Haisheng, Z. (2023). Medical image analysis using deep learning algorithms [Review of

Medical image analysis using deep learning algorithms]. Frontiers in Public Health, 11. Frontiers Media.

Livingstone, D., Talai, A. S., Chau, J., & Forkert, N. D. (2019). Building an Otoscopic screening prototype tool using deep learning. Journal of Otolaryngology - Head and Neck Surgery, 48(1).

Mohtasham-Amiri, Z., Heidari, A., Navimipour, N. J., Esmaeilpour, M., & Yazdani, Y. (2024). The deep learning applications in IoT-based bio- and medical informatics: a systematic literature review. Neural Computing and Applications, 36(11), 5757–5797.

Narwaria, M., & Jaiswal, S. (2026). Federated Learning for Privacy-Preserving Medical Diagnosis. Zenodo (CERN European Organization for Nuclear Research).

Novi, S. L., Navarathna, N., D'Cruz, M., Brooks, J., Maron, B. A., & Isaiah, A. (2025). Deep Learning in Otolaryngology. JAMA Otolaryngology–Head & Neck Surgery, 152(1), 71–71.

Packhäuser, K., Gündel, S., Münster, N., Syben, C., Christlein, V., & Maier, A. (2022). Deep learning-based patient re-identification is able to exploit the biometric nature of medical chest X-ray data. Scientific Reports, 12(1).

Park, H., Kim, S. H., Choi, J. Y., & Cha, D. (2023). Human–machine cooperation meta-model for clinical diagnosis by adaptation to human expert's diagnostic characteristics. Scientific Reports, 13(1).

Pati, S., Baid, U., Edwards, B., Sheller, M., Wang, S., Reina, G. A., Foley, P., Груздев, А. Д., Karkada, D., Davatzikos, C., Sako, C., Ghodasara, S., Bilello, M., Mohan, S., Kickingereder, P., Brugnara, G., Preetha, C. J., Sahm, F., Maier-Hein, K., … Bakas, S. (2022). Federated learning enables big data for rare cancer boundary detection. Nature Communications, 13(1), 7346–7346.

Pati, S., Kumar, S., Varma, A., Edwards, B., Lu, C., Qu, L., Wang, J. J., Lakshminarayanan, A., Wang, S., Sheller, M., Chang, K., Singh, P., Rubin, D. L., Kalpathy–Cramer, J., & Bakas, S. (2024). Privacy preservation for federated learning in health care. Patterns, 5(7), 100974–100974.

Rauniyar, A., Hagos, D. H., Jha, D., Håkegård, J. E., Bağcı, U., Rawat, D. B., & Vlassov, V. (2023). Federated

Learning for Medical Applications: A Taxonomy, Current Trends, Challenges, and Future Research Directions. IEEE Internet of Things Journal, 11(5), 7374–7398.

Ravikumar, R. N., Aarthi, S., Sapayev, V., Atamuratova, Z., Kholboyev, N., & Sindhuja, S. (2025). The Decentralized Diagnostics Revolution Using Federated Learning in Medical Imaging and Beyond. In Advances in computational intelligence and robotics book series (pp. 353–388). IGI Global.

Saak, S., Oetting, D., Kollmeier, B., & Buhl, M. (2024). Integrating audiological datasets via federated merging of Auditory Profiles. arXiv (Cornell University).

Saldanha, O. L., Quirke, P., West, N. P., James, J. A., Loughrey, M. B., Grabsch, H. I., Salto-Tellez, M., Alwers, E., Çifçi, D., Laleh, N. G., Seibel, T., Gray, R., Hutchins, G., Brenner, H., Treeck, M. van, Yuan, T., Brinker, T. J., Chang-Claude, J., Khader, F., … Kather, J. N. (2022). Swarm learning for decentralized artificial intelligence in cancer histopathology. Nature Medicine, 28(6), 1232–1239.

Sheller, M., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., Milchenko, M., Xu, W., Marcus, D. C., Colen, R. R., & Bakas, S. (2020). Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Scientific Reports, 10(1).

Silva, F. R. da, Camacho, R., & Tavares, J. M. R. S. (2023). Federated Learning in Medical Image Analysis: A Systematic Survey. Electronics, 13(1), 47–47.

Skorupko, G., Avgoustidis, F., Martín-Isla, C., Garrucho, L., Kessler, D. A., Pujadas, E. R., Díaz, O., Bobowicz, M., Gwoździewicz, K., Bargalló, X., Jaruševičius, P., Osuala, R., Kushibar, K., & Lekadir, K. (2025). Federated nnU-Net for Privacy-Preserving Medical Image Segmentation. arXiv (Cornell University).

Soltan, A. A. S., Thakur, A., Yang, J., Chauhan, A., D'Cruz, L., Dickson, P. W., Soltan, M., Thickett, D., Eyre, D. W., Zhu, T., & Clifton, D. A. (2023). Scalable federated learning for emergency care using low cost microcomputing: Real-world, privacy preserving development and evaluation of a COVID-19 screening

test in UK hospitals. medRxiv (Cold Spring Harbor Laboratory).

Zhao, Z., Alzubaidi, L., Alwzwazy, H. A., Zhang, J., Alammar, Z., & Gu, Y. (2026). A systematic review of artificial intelligence-driven decentralized learning in healthcare: emerging techniques, applications, challenges and future directions. Artificial Intelligence Review.

Zhou, L., Wang, M., & Zhou, N. (2023). Distributed Federated Learning-Based Deep Learning Model for Privacy MRI Brain Tumor Detection. 1–12.

.