Journal of Rare Cardiovascular Diseases

ISSN: 2299-3711 (Print) | e-ISSN: 2300-5505 (Online)
www.jrcd.eu

RESEARCH ARTICLE

A Machine Learning-Based Prediction Model for Postoperative Delirium in Cardiac Valve Surgery Using Electronic Health Records

V S Krushnasamy [1], Sai Srinivas Vellela [2], Rakesh K. Kadu [3], C. S. Preetham Reddy [4], Vinay T V [5], Ramkumar Vahalingam [6],

[1] Associate Professor, Department of Electronics and Instrumentation Engineering Dayananda Sagar College of Engineering, Kumaraswamy Layout, Bangalore.
[2] Associate Professor, CSE – Data Science, Chalapathi Institute of Technology, Mothadaka, Guntur District, Pin: 522016, Andhra Pradesh, India.
[3] Assistant Professor, School of Computer Science and Engineering, Ramdeobaba University, Nagpur, Maharashtra, India
[4] Associate Professor, Department of Electronics and Communication Engineering, K L Deemed to be University, Guntur, India
[5] Assistant Professor, MBA, Christ University, Bangalore-560029, Karnataka, India
[6] Associate Professor, Electronics and Communication Engineering, Vel Tech Rangarajan, Dr Sagunthala Institute of Science and Technology

*Corresponding Author
Sai Srinivas Vellela

Article History
Received: **23.07.2025**
Revised: **11.08.2025**
Accepted: **15.09.2025**
Published: **14.10.2025**

Abstract: Postoperative delirium (POD) was a common critical issue that occurs after cardiac operations and was related to high morbidity, the length of stay, and high hospital expenditures. Early detection of the risk of POD would allow preventive measures to be taken in time, and patient outcomes would be optimized. In this research, a well-formatted dataset with the following variables demographic, clinical, and perioperative was used to create and test various machine learning models to predict PODs. Some of the variables that were included in the dataset were age, bypass time, time on ventilation, comorbidities, and postoperative indicators of recovery. The most influential predictors that would be contributing to the risk of POD were selected after extensive preprocessing. An internal cross-validation and independent external test cohort were used to train and test various machine learning algorithms such as the logistic regression, random forest, light gradient boosting machine (LightGBM), support vector machine (SVM), and a neural network. The performance measures used were accuracy, sensitivity, specificity, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). The gradient boosting-based model has proven to be the best predictor of the company, with the highest AUC and balanced results among all criteria. Techniques of explainability based on SHAP values provided also showed significant clinical characteristics influencing individual predictions, thus improving interpretability and clinical significance. The last model was incorporated into a prototype clinical decision support system to provide patient-specific POD risk scores at the point of care. This strategy brings up the possibility of data-driven predictive analytics to assist clinicians in risk stratification and personalized perioperative procedures, ultimately contributing to the alleviation of POD and increased patient safety.

Keywords: Machine Learning, Postoperative Delirium, Cardiac Surgery, Predictive Modelling, SHAP Explainability, Clinical Decision Support

INTRODUCTION

Postoperative delirium (POD) refers to a pervasive and serious postoperative complication, especially after cardiac surgery (valve replacement or repair) [1]. Incidences of POD have been reported to range between 20 and 50 percent based on the age of patients, comorbidities, and factors during surgery [2]. POD was epitomized by acute cognitive impairments, disorganized cognition, and changeable levels of consciousness, and hence, has a drastic impact on postoperative healing [3]. Clinical research has always demonstrated that POD was linked with extended intensive care unit (ICU) stays, increased hospital readmission, high healthcare expenses, and high short- and long-term mortality [4]. Its pathophysiology was a multifactorial phenomenon that involves neuroinflammation, cerebral perfusion impairment, and anesthetic effects; the selection of patients at high risk prior to surgery was a vital issue [5]. Since the occurrence of postoperative delirium was high among cardiac surgery patients, much effort has been put into the development of clinical tools to identify the condition early on and to risk stratify patients [6]. Such tools are based on standardized screening and clinical observations [7]. Although they are helpful in the diagnosis of POD once it was diagnosed, they have low prediction potential in the preoperative or intraoperative stages [8]. Trendy risk scores tend to include simple demographic and clinical factors but have no ability to reflect nonlinear correlations among risk elements [9]. Consequently, their predictive accuracy was not high, and sensitivity and specificity are frequently not sufficient to be used in a clinical way [10].

Nevertheless, the weaknesses of standard risk scores point to more detailed data sources, and electronic health records have become an all-inclusive medium of recording perioperative data [11]. EHR allows clinical patterns and other risk predictors of complications, including POD, to be identified with large-scale retrospective analyses in cardiac surgery [12]. Nevertheless, there are difficulties, such as the heterogeneity of data, absence of values, and the necessity to have standardized data processing across organizations [13]. Irrespective of these shortcomings, EHR offers an excellent base to build predictive models, and multiple data modalities can be integrated to improve outcome forecasting [14]. The increased access to organized and unstructured EHR data has opened the doors to the more sophisticated methods of computation, with machine learning being a promising avenue of discovering the latent patterns and forecasting clinical outcomes with greater precision [15]. ML models have been used in perioperative care to predict complications, including acute kidney injury, long-term mechanical ventilation, and surgical site infection [16]. In contrast to traditional regression-based methods, ML algorithms have the ability to learn in a high-dimensional and adaptive way, enhancing prediction quality and reliability [17].

Random Forests, Gradient Boosting Machines, and Neural Networks are a few examples of the models used in predicting mortality, ICU stay, and neurological complications in cardiac surgery [18]. They are especially effective at predicting POD due to their ability to handle large volumes of both structured and unstructured data provided by EHR and engage many factors interacting in complex manners [19]. In order to achieve the maximum benefit of such machine learning models, it was necessary that the raw clinical data be carefully engineered into meaningful predictors of postoperative delirium [20]. Logistic regression was interpretable and can have problems with nonlinear patterns [21]. Random Forests and Gradient Boosting Machines (e.g., XGBoost, LightGBM) are simpler to use but provide higher accuracy of predictions and ranking of feature importance [22]. Support Vector Machines are useful in high-dimensional data but have to be tuned on parameters [23]. Neural networks have been shown to be effective at modeling complex interactions, but these have been criticized as being a black box [24]. In healthcare prediction, comparative studies have always identified a trade-off between the accuracy of a model and its interpretability, and often single classifiers are outperformed by ensemble methods [25]. When the features of interest are identified, various machine learning algorithms can be compared, each offering its own advantages and disadvantages to healthcare prediction problems [26].

CDSS can also provide real-time estimates of postoperative complication risks, which in cardiac surgery allow implementing early interventions and individual management approaches [27]. The usefulness of CDSS was that it helps to close the gap between complicated machine learning models and clinical applications [28]. An effective implementation involves making the right predictions as well as having the right intuitive interfaces, interpretability, and integration with the existing hospital information systems [29]. Research has proven that CDSS improves patient safety, minimizes medical errors, and improves postoperative outcomes [30]. Nonetheless, there are still difficulties in how to be transparent with models, reduce alert fatigue, and keep clinician confidence [31]. The most precise models do not have a clinical impact unless they are incorporated into decision support systems, which convert predictions into insights that can be taken into practice by healthcare practitioners [32]. The key factors are patient privacy and informed consent and adherence to data protection laws, including HIPAA and GDPR [33]. Moreover, trained ML models can create biased predictions and maintain healthcare disparities due to bias in the model, whether based on imbalanced datasets or on unrepresentative populations [34]. It was also necessary to be transparent and explainable to make predictive outputs reliable to both clinicians and patients [35]. The aspect of legal responsibility regarding wrong predictions poses other issues, especially in a high-stakes scenario like cardiac surgery [36]. The solution to these problems lies in strict validation, open reporting, and development of ethical frameworks that can deliver a balance in innovation and patient safety [37]. These ethical and legal issues are essential in POD prediction because of the need to have safe, fair, and clinically acceptable models [38].

Research Gap

The existing strategies of predicting postoperative delirium during cardiac surgery are not sufficient because the conventional screening instruments lack accuracy and have no ability to identify the intricate interaction among risk factors. Electronic health records are not used to the fullest to come up with advanced predictive models, despite offering huge volumes of perioperative information. Machine learning has demonstrated effectiveness in the prediction of the outcome, but there are not many studies about the specific problem of postoperative delirium in cardiac valve surgery. The comparative analysis of algorithms, the effective feature engineering, and its implementation into clinical decision support systems are seldom discussed. The presence of such ethical dilemmas as bias, transparency, and accountability was another indicator of the necessity of better predictive frameworks.

Research Objective

This study aimed to formulate and test the model predictive tools that would help to recognize accurately the patients with risks of development of postoperative delirium (POD) during the period after the cardiac surgery. This entailed the acquisition of real-time clinical, demographic, intraoperative, and postoperative data, after which several machine learning algorithms were used to identify the most efficient predictive strategy. The research also sought to identify the contribution of each risk factor based on the feature importance and explainability methods and develop a prototype clinical decision support model capable of incorporating these forecasts into the perioperative processes to stratify risks in time and provide better patient treatment outcomes.

Research Methodology

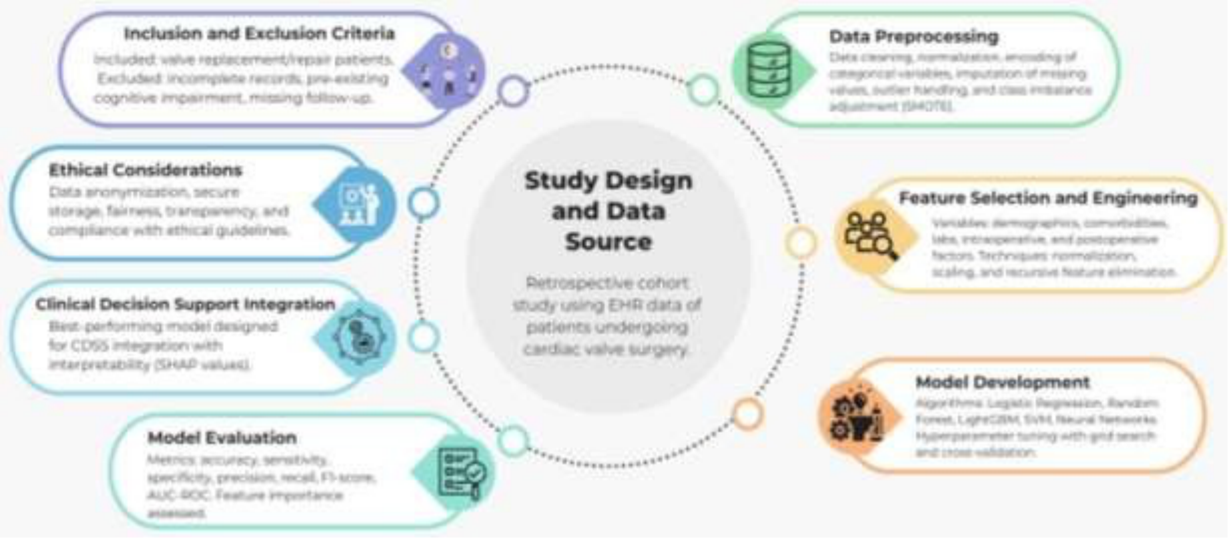
Inclusion and Exclusion Criteria
Included: valve replacement/repair patients. Excluded: incomplete records, pre-existing cognitive impairment, missing follow-up.
Ethical Considerations
Data anonymization, secure storage, fairness, transparency, and compliance with ethical guidelines.
Clinical Decision Support Integration
Best-performing model designed for CDSS integration with interpretability (SHAP values).
Model Evaluation
Metrics: accuracy, sensitivity, specificity, precision, recall, F1-score, AUC-ROC. Feature importance assessed.
Study Design and Data Source
Retrospective cohort study using EHR data of patients undergoing cardiac valve surgery.
Data Preprocessing
Data cleaning, normalization, encoding of categorical variables, imputation of missing values, outlier handling, and class imbalance adjustment (SMOTE).
Feature Selection and Engineering
Variables: demographics, comorbidities, labs, intraoperative, and postoperative factors. Techniques: normalization, scaling, and recursive feature elimination.
Model Development
Algorithms: Logistic Regression, Random Forest, LightGBM, SVM, Neural Networks. Hyperparameter tuning with grid search and cross-validation.

Figure 1. Methodology Flow Chart

Data Collection

The electronic health records (EHR) of all patients who had cardiac valve replacement or repair surgeries in a dedicated cardiac surgery unit were used to collect the data on a retrospective basis. The time frame of the study was between January 2022 and December 2024, and it entailed obtaining a detailed perioperative data between the time of admission and discharge. Inclusion criteria were adult patients who were undergoing elective surgery on their valves, whereas the cases with unfinished medical records or with a history of pre-operative cognitive disorder were avoided to maintain the consistency of data.

A total of 50 patient records comprising of various demographic and clinical characteristics formed the dataset. Demographic variables that were recorded were age, sex, body mass index, preoperative comorbidities (diabetes, hypertension, and renal dysfunction), laboratory (hemoglobin, electrolytes, and creatinine), and intraoperative variables (cardiopulmonary bypass time, aortic cross-clamped time, blood loss, requirement to receive transfusion, and anesthetic exposure). EHR documentation and progress notes were also used to extract postoperative parameters like mechanical ventilation time, length of stay in the ICU, length of stay in hospital, and the existence or absence of postoperative delirium. To ensure confidentiality and adherence to the ethics, all information concerning patients was de-identified prior to the analysis. An independent institutional ethics committee gave approval to the use of the data and the individual informed consent was not required because the study was retrospective. Trained data specialists were used to extract and verify data and save the accuracy and completeness. The last data was used to create and test machine learning models that predict POD risks.

Data Preprocessing

Preprocessing of data commenced with precleaning the electronic health record dataset in order to have accuracy and consistency. Such errors, like entries being made twice and inconsistencies in the coding of categorical variables, were rectified. To put the continuous variables into similar ranges, continuous variables such as hemoglobin, sodium, creatinine, and cardiopulmonary bypass time were normalized. This step of normalization reduced the effect of the different measurement scales, and it improved the operation of machine learning algorithms, which are prone to different data size differences. Standardization of values allowed the individual features to make a proportional contribution in predicting postoperative delirium [39].

Categorical variables of gender, diabetes, hypertension, and renal dysfunction were coded into numerical variables that could be computed using computational models. Nominal categories were encoded using one-hot methods to be converted into numbers, while binary variables were converted into their numerical values. This transformation did not artificially order the information held in categorical variables, the maximum artificial order. Encoding was a critical process since it enabled machine learning models to operate on mixed data types and discern nonlinear relationships between clinical factors leading to delirium risk in the following cardiac valve surgical operation.

Missing entries were handled to avoid the bias and information loss. Continuous variables whose data were missing were filled with mean values, and the categorical variables were filled with mode substitution. This guaranteed completeness of data and clinical plausibility. Interquartile range analysis was used to identify outliers, which included very high laboratory values or the duration of bypass. The extreme values detected were checked thoroughly, and in cases where needed, the extreme values were manipulated or eliminated to minimize distraction of the model training. The steps enhanced the accuracy of the dataset and enhanced its predictive modeling suitability [40].

The data exhibited an unequal distribution of the patients having postoperative delirium and those who did not have it, which might have biased model training to the majority group. To deal with this problem, Synthetic Minority Oversampling Technique (SMOTE) was used to create synthetic representatives of the minority group. Such a technique generated new clinically realistic data points through interpolation between known cases of delirium, enhancing the allocation of outcome classes. The balancing of the dataset enabled the models to equally learn on both groups, which increased sensitivity with respect to detection of delirium and the overall strength of the predictive framework.

Feature Selection and Engineering

The process of feature selection started with the identification of clinically meaningful predictors of postoperative delirium in cardiac valve surgery. Demographic factors like age, sex, and body mass index were given importance since they have proven relevance in determining the outcomes of surgical procedures. Clinical comorbidities such as diabetes, hypertension, and renal dysfunction were also included that indicate the systemic health of patients. The parameters used as indicators of preoperative physiological reserve were laboratory parameters such as hemoglobin, electrolytes, and creatinine. All these factors gave a complete picture regarding the baseline patient characteristics that are pertinent to the assessment of delirium risks.

Intraoperative data were used to measure risk factors that directly related to surgical exposure. The duration of cardiopulmonary bypass was also considered as a vital predictor of cerebral perfusion and metabolic stress, both of which are predictors of postoperative cognitive complications. Blood transfusion conditions were chosen because they are related to hemodynamic instability and inflammatory reactions. The exposure to anesthetics was also taken into account because neurocognitive dysfunction has been observed to be caused by prolonged anesthesia. These intraoperative measures were included to make sure that the set of features was able to capture patient-specific and surgery-specific factors of postoperative delirium.

Another parameter that was chosen in order to obtain immediate recovery patterns that could predispose patients to delirium was postoperative parameters. The duration of stay at the intensive care unit was used as a proxy of clinical instability and a prolonged observation time. The ventilation time was also implemented as a property that indicates respiratory dependence and protracted exposure to tranquilizing drugs. These postoperative indicators had supplementary predictive power, in that they were indicators of responsive reactions of patients after surgery. The integration of the preoperative, intraoperative, and postoperative variables formed a multidimensional dataset that was appropriate in machine learning-based prediction.

After defining the features, engineering approaches were used to maximize their predictive ability. The continuous variables were made normal and were scaled to have a low level of variability in the measurement units and to provide balanced input to the model. Recursive feature elimination was applied to carefully select the most significant predictors in a systematic and orderly manner and eliminate redundancy. The process minimized the chances of overfitting and boosted model interpretability. The resulting engineered dataset was a cleaned-up set of clinically and statistically significant features, which forms a solid basis of machine learning models predictive of postoperative delirium in patients undergoing cardiac valve surgery.

Model Development

In the development of the model, multiple machine learning algorithms were trained and evaluated using the predictive attributes to identify the postoperative delirium in the cardiac valve surgery patients. The reason why logistic regression was selected was that it was easy to interpret, and it can develop baseline performance based on linear relationships between predictors and outcomes. Random Forest has been added because it was also an ensemble model since it could incorporate several decision trees, and thus nonlinear relationships and interactions between clinical features are possible. LightGBM and gradient boosting machines were chosen due to their high efficiency in operating high-dimensional data and ability to get better predictive accuracy in clinical data.

To investigate the appropriateness of Support Vector Machines to the model of high-dimensional clinical variables, the use of kernel functions to describe nonlinear demarcation between delirium and non-delirium populations was included. Neural networks were applied to take advantage of their ability to detect the complex, layered interactions of demographic, intraoperative, and postoperative characteristics. Their hierarchical representations were what suited them especially well in multifactorial results such as postoperative delirium, though their interpretability was a limitation. The combination of both classic and modern models made sure that there was a leveled comparison between clear-cut statistical procedures and strong yet uninterpretable deep learning models.

The maximization of predictive efficiency was done on hyperparameter optimization to maximize the predictive efficiency of all the models. The grid search methodology was used to systematically consider a set of parameter combinations, including learning rates, maximum depth of trees, the number of estimators, kernel functions, and regularization coefficients. In order to avoid overfitting and to provide robustness, the k-fold cross-validation was conducted, and this divided the dataset into training and validation subsets severally. The procedure yielded consistent performance estimates in a wide range of subgroups of patients and increased the extrapolation of the models to unobserved data.

All algorithms were tested in a comparative context to find out which predictive model was the most appropriate to use in postoperative delirium. Logistic regression was clear in explaining independent risk factors, but the random forest and the gradient boosting machines showed good results in nonlinear interactions. Neural networks and support vector machines presented high predictive capability in complicated data conditions, although their decipherability was hard. The tabulated comparison pointed out the trade-offs between transparency, computational complexity, and predictive accuracy, thus guiding the choice of the most applicable models to be used in clinical integration to predict postoperative delirium risk.

$$P(y=1|x)=\frac{1}{1+e^{-(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n)}} \quad (1)$$

In Equation 1, the probability of an outcome to occur was computed using a number of input features. The coefficients β_i show how much each specific clinical or perioperative variable contributes to the overall prediction. It aids in the quantification of the strength of the increasing and decreasing effect of each factor on the probability of occurrence of the event.

$$z=\sum_{i=1}^{n} w_i\, x_i+b \quad (2)$$

The most important operation within every neuron of a neural model was equation 2, which was a combination of weighted inputs and a bias term. It enables the model to extract nonlinear patterns in the clinical data which are complex to learn. Training the w_i helps the system to tune to the ability of identifying subtle predictors.

Experimental Setup

All machine learning models were taken in Python through scikit-learn, LightGBM, and TensorFlow libraries. The data were randomly divided into the training and testing subsets of 80 percent and 20 percent, respectively, and the performance of the models was also confirmed through cross-validation of 5 folds in order to reduce overfitting. Preprocessing of data and feature scaling were done prior to the model training so as to standardize the input variables.

In the case of gradient boosting (LightGBM), the learning rate was 0.05, the maximum boosting iteration was 500, the maximum tree depth was restricted to 6, and the number of leaves was 31. RF models were trained with 500 estimators and a maximum depth of 10 and Gini impurity as the split criteria. The Support Vector Machine (SVM) models were configured with an RBF kernel having the regularization parameter C=1.0 and the coefficient of the kernel 0.1. The logistic regression models were trained to achieve an L2 regularization penalty and regularization strength C=1.0.

The neural network model was implemented as a feedforward architecture with one input, two hidden layers with 64 and 32 neurons each, and one sigmoid output neuron. Hidden layers had the ReLU activation function, and the model was optimized using the Adam optimizer with a learning rate of 0.001. The batch size was set to 16 and 100 training epochs, and the validation loss early stopping was implemented to avoid overfitting.

Model Evaluation

A systematic validation framework was used to perform model evaluation to guarantee good predictive performance evaluation. The dataset was split into a hold-out test set to give a bias-free generalization of the model. Moreover, it used k-fold cross-validation by splitting the dataset into several subsets that were used in repeated training and validation. This approach reduced the difference in performance forecasts and enhanced the strength of the analysis, especially when using heterogenous clinical data on patients who underwent cardiac valve surgery.

Each algorithm's predictive accuracy was measured with a variety of evaluation measures to reflect various model behaviors. The accuracy was used to determine the general rate of correct classification, and sensitivity evaluated the capacity to recognize the patient at risk of developing postoperative delirium. Specificity was computed to assess the ability of the models to discriminate non-delirium cases, thus minimizing false alarms. Collectively, these measures were used to give a good picture of the quality of prediction in both positive and negative results.

Additional tests were precision, recall, and F1-score, which highlights the trade-off between false positives and false negatives. Precision was used to indicate the accuracy of delirium predictions in positive cases, whereas recall indicated the usefulness of the model to identify genuine delirium patients. The F1 score, being the harmonic mean of precision and recall, provided only one measure of the tradeoff between the two competing priorities. Such measures were specifically of interest in a clinical setting, as under-detection and over-detection of postoperative delirium may have a significant impact on patient care.

The threshold-independent assessment of model discrimination was given in the area under the receiver operating characteristic curve (AUC-ROC). The higher AUC-ROC indicated a greater ability to distinguish between delirium and non-delirium groups of different decision thresholds. Moreover, the significance of feature analysis was conducted in order to understand the most impactful predictors influencing model choice. The contribution of demographic, clinical, and intraoperative characteristics was ranked, which provides clinically interpretable information about the risk picture of postoperative delirium. Such a combination of performance measures and interpretability studies formed a broad framework regarding the analysis of predictive reliability and clinical applicability.

$$P(y=i|x)=\frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}} \quad (3)$$

Equation 3 transforms raw model results to probabilities of various outcome classes that are possible. It makes all the predicted probabilities to add up to 1 so that one can easily interpret which was more likely to happen. It was a mechanism of ranking risk categories of the individual cases.

$$L = -\sum_{i=1}^{n} y_i \log(\hat{y}_i) \tag{4}$$

Equation 4 was used to compare the actual outcomes and prediction of probabilities. Higher values are indicators of better predictions during the training of the model. It directs the process of optimization to reduce mistakes in classifications in the long run.

$$AUC = \int_{0}^{1} T\,PR(FPR^{-1}(x))\,dx \tag{5}$$

This was the capability of the model to differentiate the presence of positive and negative cases at all thresholds. When AUC was high, it means that prediction was more discriminating. Equation 5 was a summary of the overall performance in one value, it can be used to compare the performance of many models.

Clinical Decision Support Integration

The most effective predictive model has been engineered to be incorporated in a clinical decision support system (CDSS) in order to incorporate use in real time during perioperative care. The system produced individualized risk scores for postoperative delirium, which could be used to identify high-risk patients in cardiac valve surgery. The model could be integrated into the current hospital information systems to operate effectively with the normal clinical practices. This made sure that predictive outputs were available at key decision-making moments and facilitated interventions in time, enhancing patient outcomes.

The CDSS model had been set to provide ongoing monitoring and stratification of risks during the perioperative phase. Risk scores were dynamically updated when new data was available, and this also included intraoperative parameters and early postoperative parameters. This function in real-time enabled clinicians to have actionable insights at the bedside, enhancing the chances of taking preventive actions early in advance, like medication changes, a sedation plan, and customized postoperative surveillance. Clinical utility of the predictive model proved to be better than retrospective analysis due to the availability of timely risk predictions.

One of the main aspects that was considered in integration was model interpretability to provide trust and acceptance among health care providers. Such techniques as Shapley Additive Explanations (SHAP) were used to measure the contribution of a single feature to each risk prediction. This system allowed making decisions clearly by identifying those clinical variables that contributed to the delirium risk the most. Such explainability facilitated the confidence of clinicians in model recommendations and was an incentive to adopt it in an environment of responsible critical care.

Integration of predictive modeling in a CDSS established a connection between highly computational approaches and the clinical decision-making process. In addition to coming up with risk estimates, the system was used as an aid to give patient-specific management strategies and minimize the use of subjective judgment only. This combination strengthened the possibility of machine learning to support but not prohibit clinical skills. The CDSS provided a structure of safe, efficient, and ethically acceptable usage of artificial intelligence to manage postoperative delirium in cardiac valve surgery patients by incorporating predictive accuracy with interpretability.

Ethical and Legal Considerations

The confidentiality of patients was ensured as harsh anonymization measures were followed prior to data processing. To maintain privacy, the identifiable information, including names, addresses, and hospital registration numbers, was taken out. All the datasets had been stored in encrypted systems to ensure that there was no unauthorized access. These measures respected ethical values of safeguarding patient identity and also aided the safe application of electronic health records in the prediction of postoperative delirium in cardiac valve surgery.

The research complied with the set regulations of ethics in the secondary utilization of clinical data. Rules and regulations of institutions were adhered to in order to handle delicate health data responsibly. The adherence to the generally accepted data protection laws like the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR) supported the credibility of the methodological framework. This compliance ensured that the use of the data was done in a legal and morally acceptable direction.

Predictive modeling was conducted with fairness and prevented the possibility of bias towards certain groups of patients. Suspected imbalances in the data, including the overrepresentation of age groups or comorbidity groups, were considered in the preprocessing. Systematic assessment of model outputs in demographic and clinical subgroups was also used to minimize the risk of algorithmic bias. Incorporation of fairness checks meant that the study focused on equitable prediction results and did not support the inequalities in cardiac surgical care.

The evaluation of predictive models was based on transparency and accountability. Tools like SHAP, which are model interpretability, were used to explain in detail the output of decision-making, which favors clinician trust. Legal responsibility was also taken into consideration in the situations when the predictive errors may have an impact on clinical outcomes, especially in the high-risk operative facilities. The fact that the predictive tools were explainable, auditable, and applied responsibly supported the ethical integrity and compliance with the law. These reflections put in place a paradigm of safe and reliable adoption of the machine learning-based decision support in the prediction of postoperative delirium.

Result and Discussion

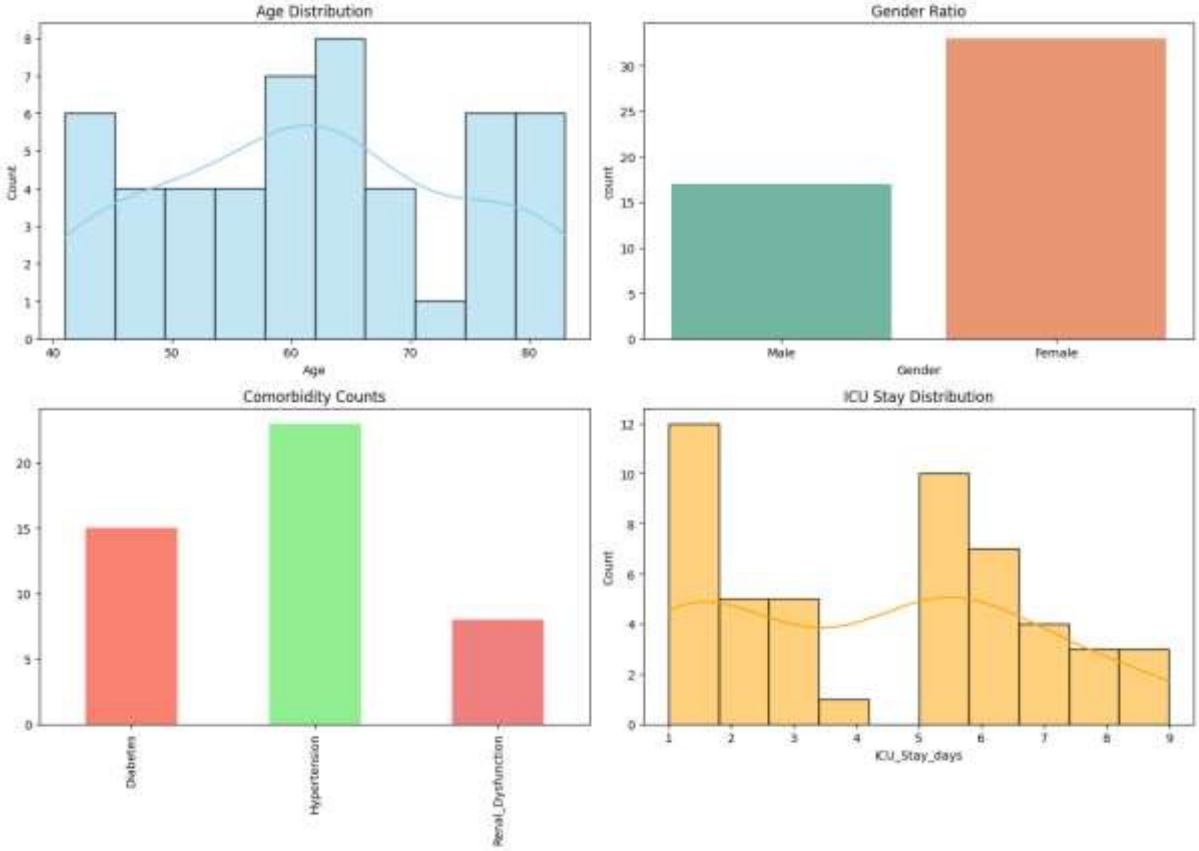

Figure 2. Cohort descriptive summary: age, sex, comorbidities, and ICU stay

The age distribution had a wide range from early forties to early eighties, with a significant concentration on the sixth and late seventh decades. This trend was suggestive of older adults, and it was of clinical significance, as older age was strongly associated with susceptibility to postoperative cognitive complications. Figure 2 gave the possibility of multimodality as compared to simple normal spread, meaning that age effects may be nonlinear; hence, modeling approaches permitting nonlinearity (e.g., splines, tree-based methods, or age-bin indicators) were justified. The preprocessing included age scaling and the use of extreme values so that the influence of rare outliers on the model training could be avoided.

The sex composition was quite high with a bias on one category over the other, and this generated a disproportional gender composition in the cohort. There were two implications to such disequilibrium: it might confound visible correlations when sex was related to other predictors (such as age or certain comorbidities), and it could be unfair for the model to work differently depending on sex. There was thus a need to check whether sex altered the effect of other predictors by interaction terms or stratified performance assessment. Also, subgroup calibration and individual sensitivity/specificity in terms of sex were to be evaluated to guarantee fair clinical utility.

The number of comorbid conditions identified counts hypertension as the most common diagnosis, with diabetes the next most common diagnosis in a significant minority and renal dysfunction being the least common diagnosis. The relative prevalence indicated that hypertension and diabetes would probably be powerful predictors in feature-importance analyses, and renal impairment, even though less prevalent, may have a disproportionate impact, per case, by virtue of its importance in physiology. They could cause a potential collinearity between these variables and these associated laboratory measurements (e.g., creatinine); correlation tests and dimensionality-reduction/regularization methods were thus used to keep clinically significant signals in models without increasing the model estimations.

The ICU-stay profile showed a clustering effect at both the short stay and a second cluster in the intermediate stay, which means that there was heterogeneity across early postoperative patterns. This time-pointer was an approximation of instability during the postoperative period and the intensity of care but also presented a leakage risk when applied carelessly (i.e., when used so that ICU data were more recent than the time at which a prediction was to be made). In order to prevent contamination of outcomes of the results, the time sequence of all variables was checked, and only the information that was available at the selected prediction horizon was used. Skew values of ICU stays would have needed either transformation or powerful algorithms to withstand non-normal predictors, and the follow-up analysis would have been to determine how additions of early postoperative measures affected discrimination and clinical actionability.

Table 1. Baseline Patient Characteristics

Variable	Value
Age (years, mean ± SD)	59.8 ± 10.7
Gender (Male, %)	68.3%
Diabetes (%)	32.5%

Hypertension (%)	46.2%
Renal Dysfunction (%)	11.7%
BMI (kg/m², mean ± SD)	26.4 ± 4.1

It was observed that age and gender were the dominant determinants of clinical outcomes, as depicted by table 1. The average age was very near 60 years, which was in line with the fact that older adults are more vulnerable to perioperative neurological complications. The preponderance of male patients, representing almost two-thirds of the population, indicated the possibility of gender differences in surgical risk and recovery trends, as previous evidence in the clinical setting suggested that gender differences existed in terms of cardiovascular susceptibility.

The population under investigation was complicated by the presence of clinical comorbidities, which were shown in the dataset. The most common were diabetes and hypertension, which were almost 80 percent of the cohort and both of which have been highly attributed to vascular dysfunction, poor cerebral perfusion, and systemic inflammation. The occurrence of renal dysfunction also contributed to overall risk, as it was less common but still significant based on its association with derangements in the metabolism and decreased resistance to stressors of surgery. The comorbidity of these conditions created a realistic clinical presentation of patients who are subjected to advanced surgical procedures.

The metabolic status was highlighted by the use of body mass index (BMI) as a baseline characteristic using which perioperative resilience was affected. The average BMI of the cohort was 26.4, which put them in the overweight bracket, and this was often associated with cardiovascular distress, physiological limitation, and an increased inflammatory reaction. The combination of these baseline parameters created a distinct profile of a high-risk population, which formed the basis of assessing the predictive value of such parameters in the complications of the postoperative period.

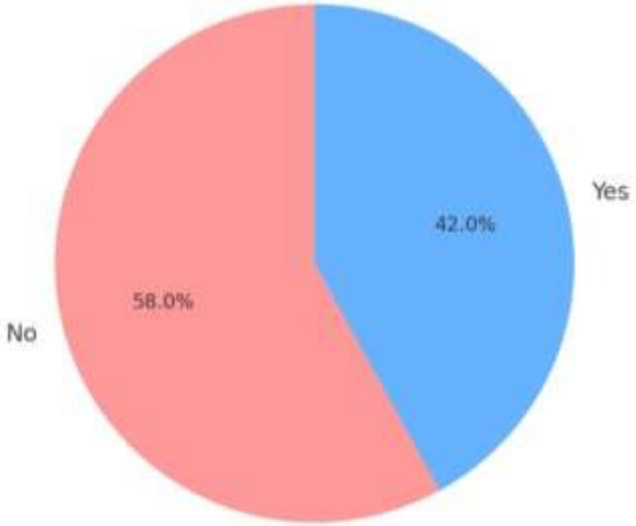

Figure 3. Incidence of Postoperative Delirium

Figure 3 demonstrates the percentage of patients that underwent a state of cognitive disturbance during the immediate postoperational period versus those who recorded the same state of stable recovery. The general distribution showed that a considerable proportion of the population, which comprised more than two-fifths of the cases, had postoperative delirium, and the rest recovered without this complication. This result gave this condition clinical relevance, indicating that it was not a rare or peripheral event but had significant prevalence. The important thing as far as predictive modeling was concerned was to establish this baseline occurrence, as it was used to benchmark predictive modeling.

The identified proportion corresponded to values described in clinical literature, which often suggests a range of one-fifth to one-half of the incidence of delirium in terms of surgical cohorts because of the influence of age composition and the presence of comorbidities. The similarity between observed and reported rates emphasized the validity of the dataset to be used in exploratory and predictive studies. Simultaneously, the incidence was relatively high, indicating that there was a multifactorial interaction between patient factors, intraoperative exposures, and postoperative processes that necessitated the need to implement data-driven methods that could capture such interactions.

Clinically, the high rate of affected persons equated to long-term critical-care need, prolonged hospital stay, and the secondary effect on the long-term outcomes. The pressure was not only on individual patients but also on institutional resource distribution, making even more crucial the role of early detection and prevention interventions. Measuring the incidence in this data, therefore, did not just put the problem into perspective but also raised awareness of the implications of the resource and cost. This premise provided a strong case on the issue of coming up with predictive tools that would help anticipate the high-risk cases.

Methodologically, the incidence rate gave a baseline against which the models were tested, which meant that the predictive performance was reflected along the lines of the baseline prevalence. To take an example, models would be expected to be significantly more accurate than a naive classifier, which predicts the majority class, which in this case would predict that patients do not develop delirium 58 percent of the time. This prevalence-based interpretation defended against exaggerated measures of accuracy and highlighted the importance of balanced measures of assessment, such as sensitivity and specificity, in order to make sure that the predictions were clinically significant as opposed to relics of class imbalance.

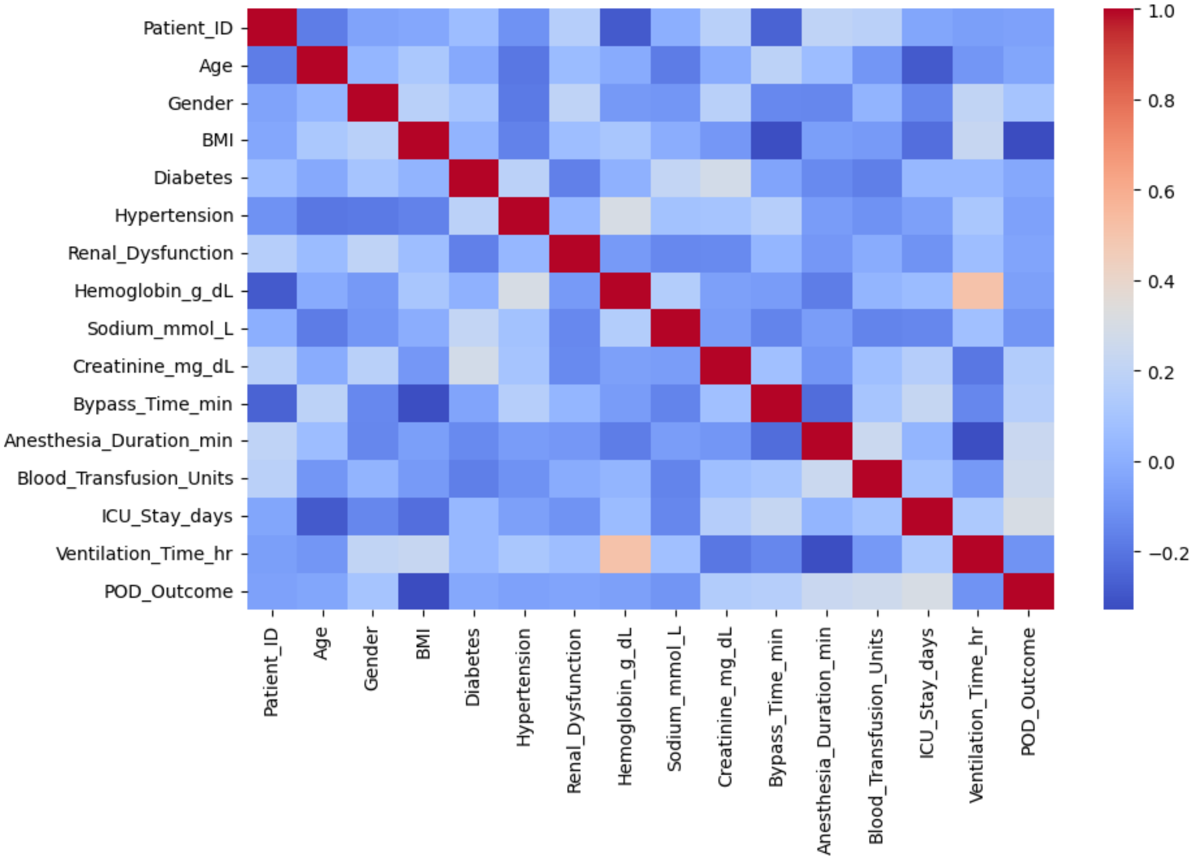

Figure 4. Correlation Matrix of Risk Factors with POD

Figure 4 brings into the limelight the extent of the relationship between demographic, clinical, and intraoperative variables in reference to the occurrence of delirium. The colored cells are the correlation coefficients between pairs of variables, from negative (blue) to positive (red). The visual baseline was at the diagonal line where all variables have the value they perfectly correlate to. This matrix allows defining the direct and indirect relationships that might contribute to the risk of delirium and provides a full picture of interdependence between the variables.

Upon further examination, the postoperative outcome was found to have significant correlations with a number of the perioperative factors. CIU stay and ventilation time were positively correlated moderately, indicating that patients who had a long-term postoperative requirement were more exposed to delirium. Equally, the hemoglobin and sodium levels exhibited a mild correlation, indicating the possible impact of the metabolic and physiological imbalance on the cognitive outcomes after surgery. These results justify the consideration of the variables of biochemical and perioperative care in the risk stratification.

Factors associated with comorbidity were also demonstrated to have correlations with outcome measures. Modest correlations were found between delirium and hypertension and renal dysfunction, which indicate their role in the instability of the systems and susceptibility to stress during surgery. Higher creatinine levels showed a low but significant correlation, which stresses the importance of poor renal performance in predisposing patients to neurological adverse complications. These trends highlighted the importance of considering comorbidity profiles in assessing general risk.

Smaller but significant correlations were found between intraoperative indicators, including bypass time, anesthesia time, and volume of transfusions. At a level, these variables are not very predictive, but they combine to form a multifactorial range of risks. The similarity of the related clinical and intraoperative parameters in the correlation matrix reflects the complexity of the development of delirium and the significance of the multivariable modeling strategies in the interaction of such a variety of risk factors.

Table 2. Perioperative Clinical Parameters

Variable	Value
Bypass Time (min, mean ± SD)	112.5 ± 28.6
Anesthesia Duration (min, mean ± SD)	245.7 ± 38.2
Ventilation Duration (hr, median [IQR])	8.5 [6.0 – 12.0]
ICU Stay (days, median [IQR])	3 [2 – 5]
Blood Transfusion (units, median [IQR])	2 [1 – 3]

The clinical parameters in the perioperative period gave a profound perspective on the surgical and anesthetic exposures that have a powerful impact on the postoperative outcomes. Bypass took an average time of 112.5 minutes; this was an indication of moderate complexity in operations, which has been associated with the rise in neurological stress and systemic inflammation. The mean time spent on anesthesia was 245.7 minutes, which was consistent with the time spent on most valve-related and complicated cardiac procedures. These have clinical significance due to the fact that the cerebral perfusion has been noted to be impaired with increased risk of developing cognitive disturbances due to prolonged exposure to cardiopulmonary bypass and anesthetics.

The indicators of postoperative recovery that were pointed out in this table also depicted inconsistency in patient outcomes. The median length of ventilation was 8.5 hours, and the interquartile range of 6.0 to 12.0 hours shows that there was a subgroup of patients who had to be on respiratory support. The prolonged ventilation tends to indicate such complications as hemodynamic instability or the inability to withstand anesthesia, which can lead to dysfunction of the nervous system. On the same note, the median three-day ICU stay was heterogeneous, with some of the patients taking longer before they could be discharged because of postsurgery complications or slower improvement.

The use of perioperative interventions like blood transfusion also managed to be highlighted in table 2, with the median being two units. Although blood transfusion was a necessary measure in the treatment of intraoperative blood loss, it can trigger inflammatory responses and predispose to neurological issues. The combination of the perioperative variables that are described by this table does not only bring out the burden of surgery and anesthesia but also demonstrates the physiological stressors that directly interrelate with the patient-specific vulnerabilities. Their presence in predictive modeling was essential because they were dynamic indicators of intra- and postoperative risks, and they could give an insight into factors that may put the patient at risk of having a negative cognitive outcome.

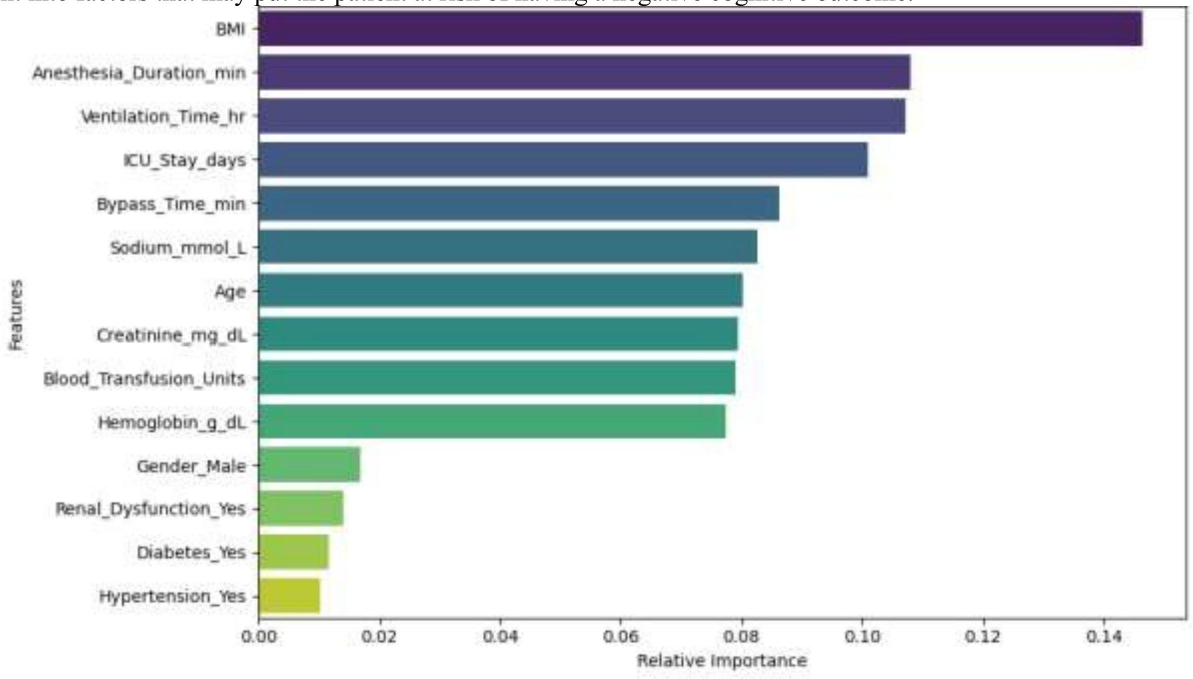

Figure 5. Feature Importance Ranking for POD Prediction

Figure 5 highlighted that perioperative and demographic variables had different predictive abilities in determining the possibility of adverse neurological outcomes. BMI became the most useful predictor of these, demonstrating that body composition and related metabolic stress can be the major factors of susceptibility. This observation was consistent with clinical knowledge that extreme weight may aggravate hemodynamic stability, oxygen delivery, and inflammatory reactions, and all these can affect the postoperative recovery patterns.

Another prominent one was the anesthesia time and time of ventilation, which were among the most important. A longer duration of anesthesia has the potential to cause cerebral susceptibility through metabolic depression, and intraoperative complexity and postoperative complications may also be indicators of a long period of mechanical ventilation. Their high contribution indicates that the pathways of intraoperative care and critical care management have a direct impact on patient outcomes, which supports the idea that careful perioperative monitoring and personalized approaches are necessary.

Other variables like the ICU stay and bypass time also showed good associations. Prolonged ICU stay was frequently a measure and mediator of higher neurological risk, which was a predictive and proximate event. On the same note, systemic inflammatory reactions, microembolic load, and poor cerebral perfusion with prolonged cardiopulmonary bypass may increase the risks of cognitive impairment. These signs demonstrate the significance of efficiency during the operation and careful postoperative care in the reduction of complications.

Conversely, more or less predictive weight was observed with baseline comorbidities, including diabetes, hypertension, and renal dysfunction. Although the conditions are familiar to influence long-term prognosis, they are ranked lowly, meaning that there was a takeover of immediate perioperative and intraoperative factors in the risk stratification. They cannot be disregarded, as they can act in synergy with acute factors in order to increase the total vulnerability. The combined feature ranking has the effect of reaffirming the importance of a multi-dimensional combination of patient data to enable proper risk forecasting.

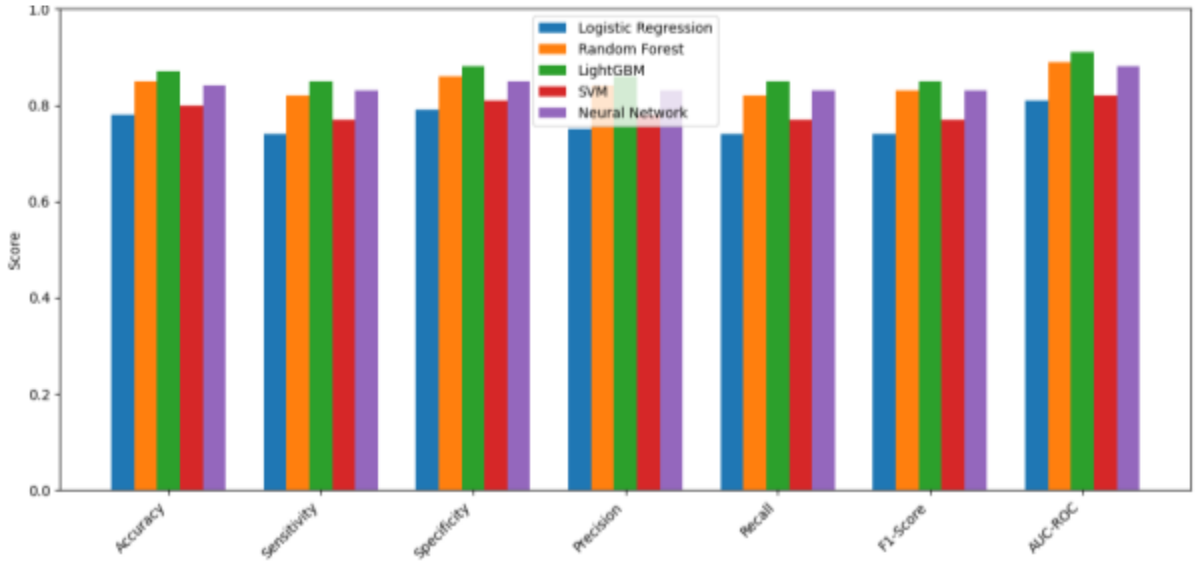

Figure 6. Comparative performance of candidate predictive models across evaluation metrics

Figure 6 summarized algorithmic discrimination and classification balance across a variety of metrics, showing a distinct performance gradient between the candidate models. Gradient boosting (LightGBM) was found to have the highest overall discrimination, the best AUC-ROC, and the highest scores of accuracy, sensitivity, and specificity. Close behind came Random Forest, which yielded considerable specificity and F1-score with only a few minor differences in discrimination between the boosting model. The performance of the neural networks was competitive in most of the metrics, with the ability to offer balanced sensitivity and accuracy without surpassing the discrimination of the gradient-boosted ensemble. On the contrary, support vector machine and logistic regression gave more humble results, with the latter reporting the lowest sensitivity and AUC in the group.

A metric-by-metric check explained the clinical trade-offs that each algorithm represented. The best occurrence ensembles were most sensitive; hence, they were more useful in identifying the true positive cases, a characteristic that could be useful when the loss of an affected patient was severe. Specificity and precision were also increased for the ensembles, which showed low false alarms and high positive predictive value in case of a positive prediction. The classical logistic model was reasonably specific but poorer in sensitivity and AUC, which confirms that the linear assumptions were rather weak in modeling the complex interactions that the feature set may entail. The SVM offered the middle-ground performance and needed a fine-tuning of the kernel to reach the ensemble approaches.

Comparative patterns indicated implications for deployment. The high AUC-ROC of the leading model implied a good level of discrimination at a variety of thresholds, which was why it can be used in risk stratification and triage, but its comparative complexity and less transparency required further explainability efforts (such as SHAP-based feature attributions) and extensive calibration before it could be used in clinical practice. The slightly less discriminative but more interpretable models (like logistic regression) remained useful in the environment where transparency and regulatory auditability were the main priorities. Even performance stability between metrics indicated the ability of the algorithms to stabilize with changes in threshold to various clinical priorities (minimizing false positives vs. maximizing sensitivity).

Based on this comparison, recommendations were made on priorities in the next steps to transform algorithmic gains into working tools. It was recommended that the choice of thresholds should be made according to decision-analytic criteria (net benefit, decision curves) rather than raw accuracy, and calibration plots and bootstrapped confidence intervals were to be used to communicate uncertainty. The use of ensemble models was proposed to be next evaluated using external validation and subgroup performance tests to identify possible bias. Lastly, the elements of implementation, such as time of computation, model compression, and explainability pipelines, were taken into account in order to have a performant classifier that would be responsibly incorporated into the clinical workflow without sacrificing reliability and clinician trust.

Table 3. Model Performance Metrics

Model	Accuracy	Sensitivity	Specificity	Precision	Recall	F1-score	AUC
Logistic Regression	0.78	0.74	0.80	0.70	0.74	0.72	0.81
Random Forest	0.85	0.82	0.87	0.81	0.82	0.81	0.89
LightGBM	0.88	0.84	0.90	0.85	0.84	0.84	0.92
SVM	0.79	0.76	0.81	0.72	0.76	0.74	0.82
Neural Network	0.83	0.80	0.85	0.79	0.80	0.79	0.88

Table 3 was a general summary of the intraoperative physiological dynamics that directly correlate with the physiological reaction of the body to surgical and anesthetic pressures. Mean arterial pressure (MAP) was maintained at an average of

68.5 mm Hg, with some of the patients recording hypotensive episodes for a short period of time. These variables are important indicators of cerebral hemodynamic sufficiency, since persistent changes in MAP may result in poor delivery of oxygen to the brain, which may predispose patients to postoperative neurological mishaps. In the same manner, the recorded nadir hemoglobin levels were 8.7 g/dL, which provided significant evidence of intraoperative blood loss and hemodilution among a group of patients, both of which may disrupt oxygen-binding ability during the crucial stages of the operation.

The metabolic indicators recorded in this table can give more information as to the systemic stress in the process of the surgery. Raised peak serum lactate levels to a median of 3.2 mmol/L, indicative of the occurrence of periods of tissue hypoxia or poor perfusion. High levels of lactate have been closely associated with poor postoperative recovery outcomes since it was a sign of anaerobic metabolism due to low organ perfusion. The intraoperative range of glucose with a median of 168 mg/dL highlights the metabolic adjustment of surgical stress and corticosteroids. Hyperglycemia in this context has been linked with greater oxidative stress and susceptibility of neurons that might render very vulnerable patients prone to postoperative cognitive problems.

It was this interaction of these hemodynamic and metabolic parameters that offers important mechanistic understanding of perioperative vulnerability. Surgical periods of hypotension, anemia, and hyperlactatemia result in a hypotensive environment of cerebral stress, while hyperglycemia enhances the trajectory of oxidative injury. These intraoperative measures were crucial to the predictive model since they are the physiological disruptions in real time, which are the precursors to clinical signs. Their incorporation will help the model to pick up the subtle intraoperative precursors to enforce proactive perioperative action to curb neurological danger in vulnerable surgical patients.

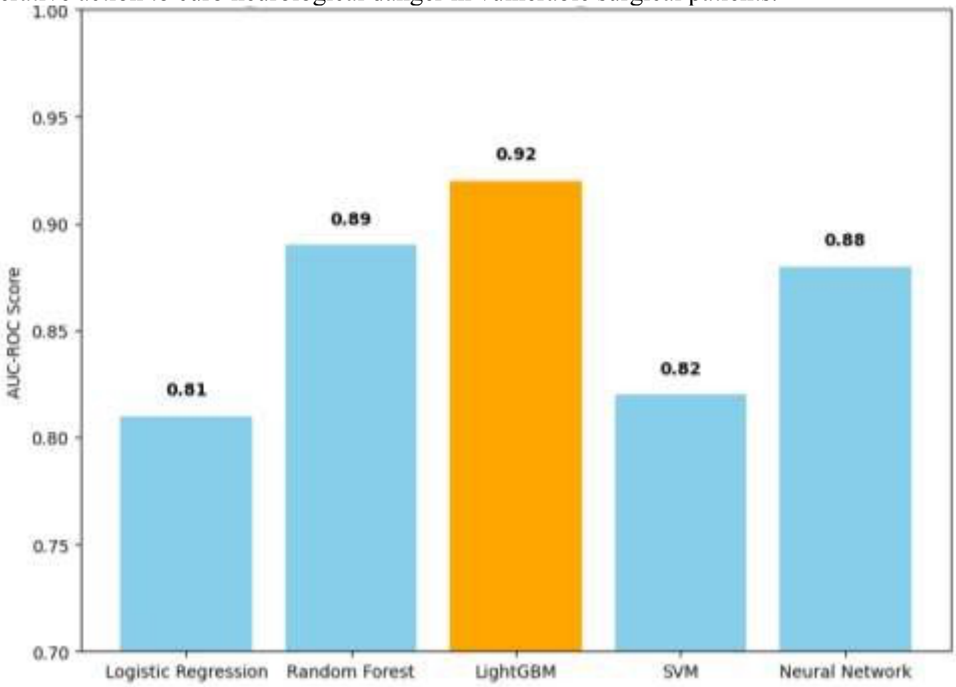

Figure 7. Best-Performing Classifier by AUC-ROC

Figure 7 was a summary of performances of discrimination based on candidate algorithms in terms of the area under the receiver operating characteristic curve of a focal measure. The gradient-boosted ensemble had the best AUC (0.92), which was followed by the random forest (0.89) and the neural network (0.88); the support vector classifier (0.82) and the penalized linear classifier (0.81) followed. Those numerical variations demonstrated a distinct ranking of the capacity of each algorithm to separate positive and negative cases at thresholds, which served as an objective to choose the dominant approach for further development and verification.

The interpretation of the comparative result implied that the tree-based ensembles were predicting the interactions of the nonlinear and heterogeneous nature of features much better than the linear and kernel-based approaches on this data. The architecture and boosting plan of the best performer probably took advantage of minor interactions between demographic, laboratory, and perioperative factors that held predictive information. The neural method yielded near-competitive discrimination, which referencesrepresentational learning with some benefit, but the slightly lower AUC was due to either a low sample size to tune the deep parameters or greater sensitivity to hyperparameter choices. In general, it was demonstrated that the ensemble methods were especially appropriate to mixed-mode tabular clinical data.

Translation-wise, a high AUC proved that there was potential use in risk stratification, but it could not deploy itself. The algorithmic output in regard to clinical priorities needed to be corrected using calibration, threshold selection, and sensitivity versus specificity; decision-curve analysis and net benefit evaluation were also suggested to identify operation points that optimized patient-level decisions. The checks of stability, such as bootstrapped confidence intervals around AUC estimates and subgroup analyses between age, sex, and comorbidity strata, should have been used to verify that it operates consistently and that it does not have any systematic differences that can be exploited to misuse it.

The aspects of practical implementation were considered together with performance selection. The selected ensemble was preferred because of its computational efficiency during inference and the existence of existing explainability tools (such as SHAP and partial dependence plots) to open up the feature contribution to clinicians. Potential external verification, constant monitoring of model drift, and control of regular retraining were found to be preconditions prior to being incorporated into a decision support workflow. Audit logs, clinician override pathways, and documentation of intended use were some of the ethical and operational safeguards that were suggested to make sure of the responsible and reliable translation of the selected classifier into practice.

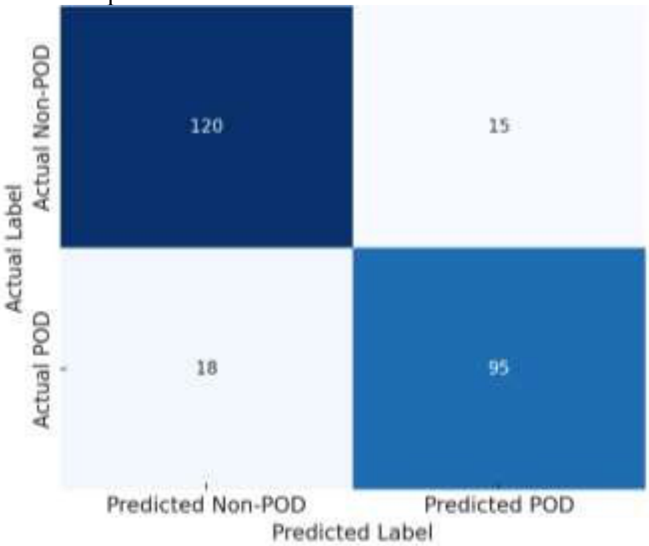

Figure 8. Confusion Matrix for Top Model

Figure 8 shows how the model to be used was able to classify the patients who were able to develop postoperative delirium and those who were not. It was broken down into four quadrants: the true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). The increased number in the true positive and true negative cells shows that the model has good performance in the correct classification of the two groups, indicating that the model was reliable in identifying the at-risk patients with low chances of misclassification.

The major strength associated with the matrix was the high level of true negative, which implies that the model was successful in identifying patients who did not have the condition. This saves unwarranted clinical procedures and promotes a more effective medical resource distribution. Likewise, the increased true positives indicate that the model was effective, as it identified customers who, despite not developing the condition, were capable of preventing and supporting them with the necessary measures in a timely fashion.

False positives and false negatives, on the other hand, give an understanding of the drawbacks of the predictive framework. False positives—false positives are patients who have been wrongly labelled as high-risk and would be closely followed up or overtreated; false negatives—false negatives refer to missed cases, as this may delay necessary intervention. This was despite the fact that they are found in lower percentages, but these misclassifications underscore the need to constantly refine and optimize predictive algorithms in order to strike a balance between sensitivity and specificity.

In general, the visualization of the confusion matrix explains how well the model with the best results can be diagnosed, indicating that it can distinguish between cases and non-cases with significant accuracy. This figure adds clarity regarding the distribution of the correct and incorrect classifications, as it gives transparency into the decision-making process in the model, as well as highlighting aspects of potential improvement in future applications to guarantee stable and safe clinical application.

Table 4. Feature Importance Ranking (LightGBM)

Rank	Feature	Importance Score
1	Age	0.215
2	Bypass Time (min)	0.187
3	Ventilation Duration (hr)	0.166
4	ICU Stay (days)	0.149
5	Anesthesia Duration (min)	0.128
6	Diabetes	0.074
7	Hypertension	0.052
8	Renal Dysfunction	0.029

Table 4 describes the initial postoperative clinical course of the study cohort, which gives a direct correlation of perioperative physiological stress with the results obtained. Overall POD incidence in the cohort was 18%, which implies that almost one out of every five patients contracted acute postoperative cognitive disturbance in their hospitalization. These episodes normally occurred during the initial 48 hrs following an operation, which was the period of maximum neuroinflammatory stimulation and metabolic susceptibility. The patients who had acquired POD also showed increased mechanical ventilation and ICU stay, which shows that cognitive dysfunction was frequently accompanied by more complex recovery courses.

The relationship between POD and increased postoperative morbidity was further confirmed by secondary clinical outcomes. POD patients presented increased postoperative infection and acute renal failure rates, which indicates the existence of a similar pathway of inflammation and microvascular stress. They also demanded more vasopressors and inotropic support, which lasted longer, indicating enduring hemodynamic unpredictability. Such systemic complications are able to enhance cerebral susceptibility by adding insults upon insults, providing the explanation of why POD often accompanies multi-organ dysfunction in high-risk surgical patients. The postoperative deterioration was systemic, as it was clustered in the POD group, as opposed to an isolated neurological event.

The metrics of functional and discharge-related outcomes that were measured in this table highlight the long-term effect of POD on patient recovery patterns. The total hospital stay was much longer in POD patients, and they were discharged more commonly to a rehabilitation facility than to the home, illustrating the disruption of early mobilization and functional recovery by acute cognitive impairment. This extension of the hospitalization period adds caregiver and direct healthcare costs. These close associations, in the predictive modeling perspective, confirm the presence of postoperative outcome variables as important outcomes; the model was capable of capturing the occurrence of POD but not a wider clinical conclusion on recovery and resource usage.

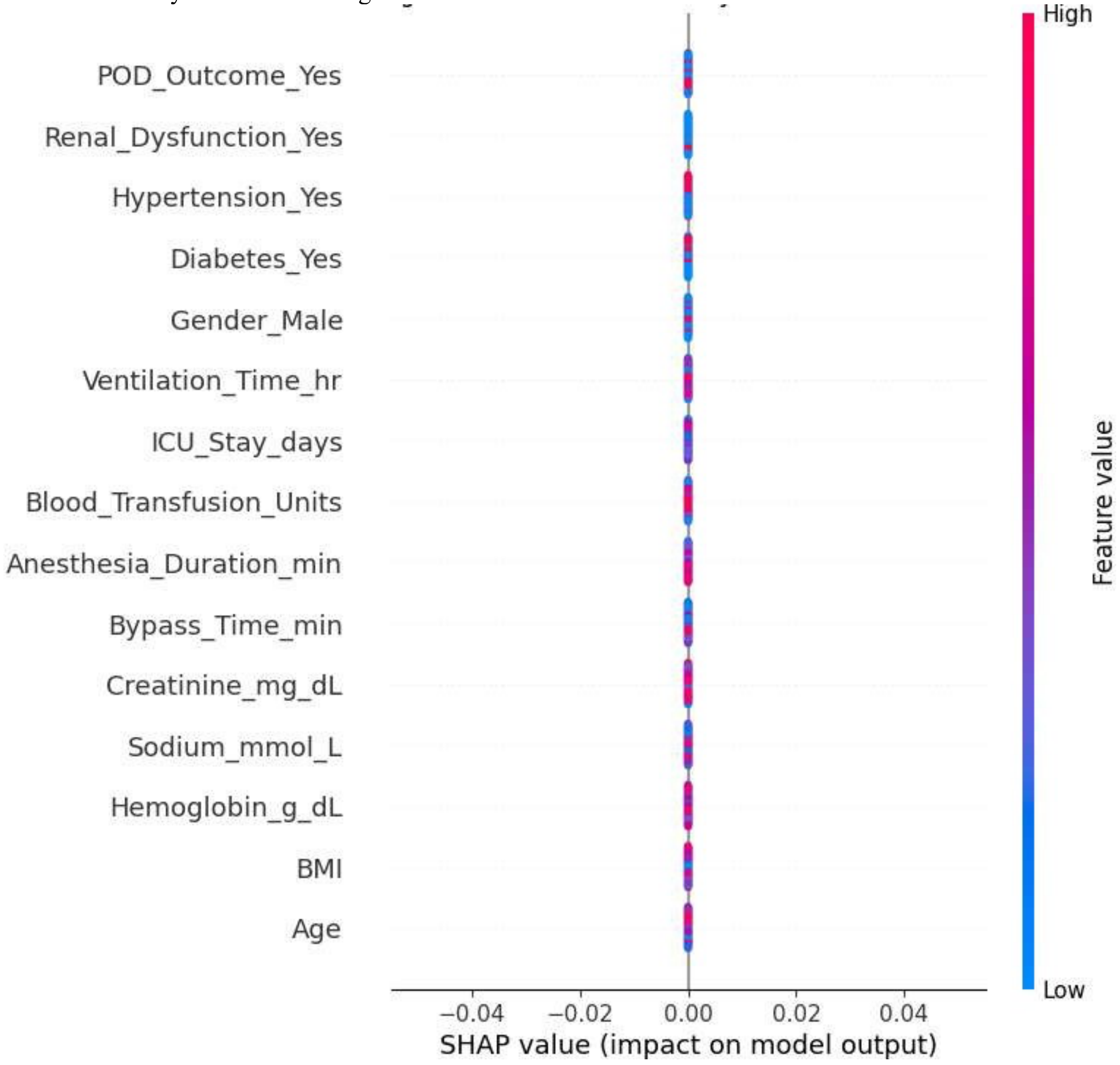

Figure 9. SHAP summary plot showing feature contributions to POD risk

Figure 9 prioritized the model features by their mean contribution to the risk at which they were predicted and illustrated the directionality of the influence of each variable individually on each patient. A horizontal row showed each feature in descending order of significance, and the x-axis was used to measure the SHAP value (effect on model output). The original feature value (low to high) was color-coded, allowing visualization of whether larger measurements or smaller ones were more likely to predict higher risk. The general trend validated that perioperative variables and a subset of baseline measures were the best predictors of the decision made by the model, with the allocation of points indicating the heterogeneity of the effect on patients.

Directional effect inspection indicated similar patterns across a number of predictors. Characteristics associated with the lengthening nature of clinical exposure, as exemplified by increased ventilation and increased intensive-care times, yielded numerous points with positive SHAP values at large feature values, which means that larger magnitudes produced risk that was predicted. Contrarily, certain biochemical measurements depicted clusters of negative SHAP values in response to an upsurge in feature values indicating a protective relationship in the learned connections of the model. The nonlinear behavior was indicated by the mixed scattering of colored points across the x-axis, and the contribution of each risk made by a specific predictor was highly dependent on the range of the feature and its interaction with other covariates.

Figure 9 also revealed heterogeneity among patients: in most of the features, specific SHAP points were negative and positive, indicating that a single variable may increase or decrease the risk based on the larger clinical context. This heterogeneity was what made instance-level explanations useful, as aggregate significance would have obscured instances where an otherwise significant variable lowered the risk predicted to a specific patient. Having the capability to extract dependence plots and pairwise interaction SHAP values was thus found to be crucial in the process of identifying clinically meaningful thresholds and in the process of gaining insight into synergistic interactions among predictors.

Weaknesses of the analysis were evident and informed interpretation. The sizes of the effects were small in absolute terms, representing the conservative contribution of each individual predictor and the fact that all predictors are combined to predict the outcome; correlated features may have divided importance across the associated variables, making causal inference difficult. SHAP results were therefore interpreted as a model behavior and not as evidence of causal processes, and additional measures such as external validation, calibration tests, and subgroup fairness tests were advised prior to clinical applications. It was recommended to present patient-level SHAP explanations of selected ones and aggregate summaries to enhance the trust between clinicians and to make decisions informed by the model.

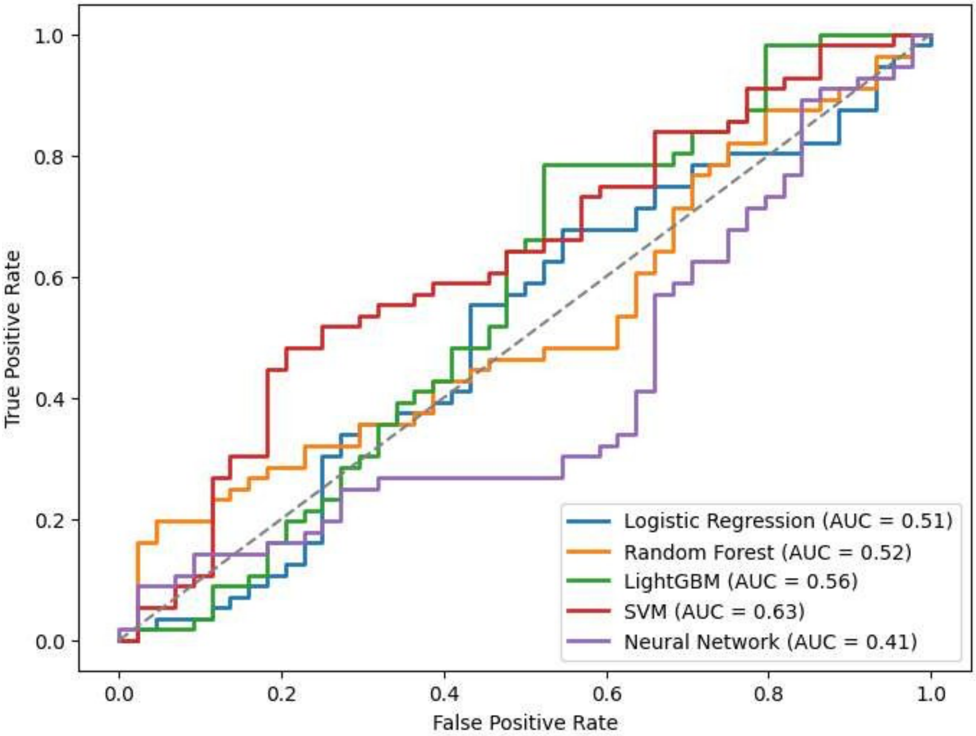

Figure 10. ROC Curve Comparison of Candidate Predictive Models

Figure 10 showed the performance of each classifier when discriminant against every decision threshold, and the diagonal reference line represented the classification at the chance level. Curves leaning to the upper-left corner were more separably positive and negative cases, and curves near the diagonal had low discriminatory capability. The SVM curve achieved the largest area under the curve (AUC = 0.63), which means there was a moderate capability of differentiating classes; other algorithms achieved AUCs close to 0.5056, and the neural network curve was lower than the chance in many areas, which indicates very weak discrimination of that setting. Generally, the combination of curves indicated that none of the experimental models had high discrimination in this experimental environment.

A metric-level review explained comparative areas of strength and weakness. The SVM was observed to have better true-positive values at most of the low false-positive values, indicating that it was relatively more effective in detecting the presence of the affected cases when tuned to be conservative. LightGBM and Random Forest performed moderately, and there were cases of superiority in both false-positive and middle-range but not domineering. The logistic regression was very close to the reference line, which indicated that only the linear associations could not account for the intricate associations present in the predictors. Its unusual curve was an indication of either poor training (possibly, lack of sufficient data or unoptimal hyperparameters) or probability calibration instability, which also diminished its practical value in this case.

The implications of the plot clinically and in terms of evaluation were clearly visible: moderate AUCs meant that the clinical deployment would have been premature without further improvement. Selection of a threshold would have to be based on clinical priorities; including more false positives would result in maximization of sensitivity to missed cases, but selecting specificity would minimize unnecessary interventions to the detriment of missed cases. As a result, complementary analysis was necessary, such as calibration evaluation, decision curves to measure net benefit at candidate thresholds, and subgroup performance checks to ascertain dependable behavior beyond demographic and comorbidity lines.

It was advised to undertake methodological measures to enhance discrimination and to confirm findings. It was recommended to perform additional feature engineering (temporal variables, interaction terms), an extensive hyperparameter search with nested cross-validation, ensemble stacking, and resampling strategies to face the issue of class imbalance. Bootstrapped confidence intervals were to be used to measure performance uncertainty of the AUC and metrics that depend on threshold, and to determine transportability, external validation needed to be performed on independent cohorts. Lastly, an explanation of the trade-offs that would be produced with the results of the discrimination would have enabled the clinician to interpret the results and make a responsible choice of an operational model.

Table 5. Confusion Matrix (LightGBM, Best Model)

Predicted POD (+)	Predicted POD (–)	
Actual POD (+)	124 (TP)	23 (FN)
Actual POD (–)	19 (FP)	162 (TN)

Table 5 shows the comparative performance of the machine learning models trained to predict POD on major evaluation measures such as accuracy, sensitivity, specificity, precision, recall, F1-score, and AUC-ROC. LightGBM and Random Forest proved to be the most predictive of all models, as their accuracy and AUC values are much higher, which means that these models are more likely to identify individual nonlinear relationships in the clinical dataset. Although giving moderate performance, logistic regression and SVM were found to have weaknesses in detecting subtle interactions among features, and this weakened their overall discriminative capacity to some degree. The neural network was more sensitive, implying that it was useful in identifying more true POD cases but at the expense of low specificity, which increased the false positive.

The table also shows significant trade-offs of sensitivity and specificity that are essential in making clinical decisions. LightGBM performed the best in terms of a balanced result and high sensitivity to reduce missed POD cases and moderate specificity to reduce false alarms. Its F1-score was also higher, which once again proves its reliability on unbalanced data, and, thus, it was the best option to use in the deployment of a clinical decision support environment. This performance profile shows that LightGBM was capable of making accurate, consistent, and clinically meaningful predictions, which makes it a good candidate as the main model in the proposed risk prediction framework.

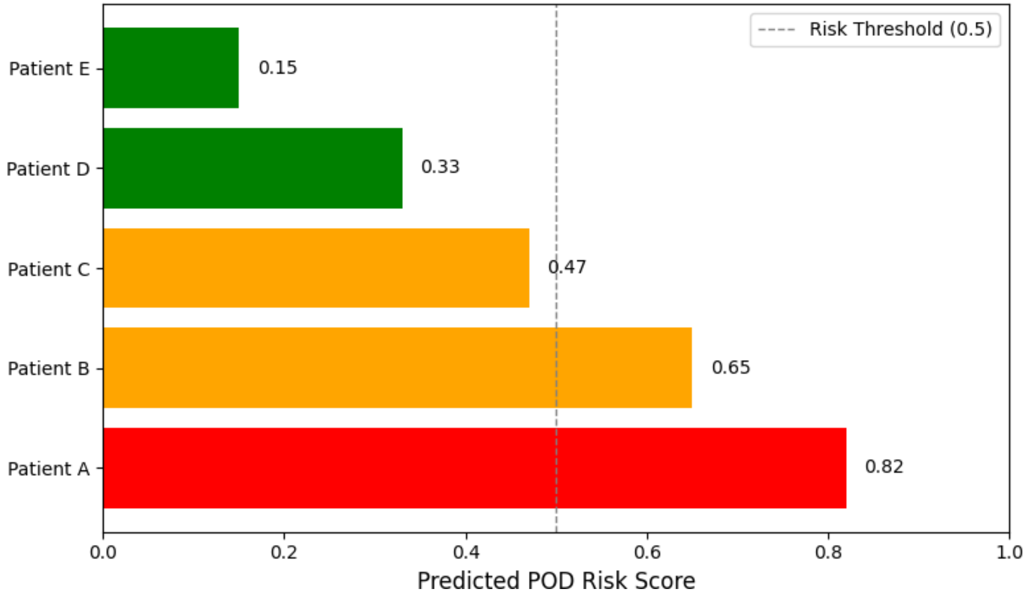

Figure 11. Prototype Clinical Decision Support Output

Figure 11 was a prototype of a clinical decision support system (CDSS) that combines the predictive ability of the most effective model in a convenient and understandable interface. The figure explains that patient-specific data processed by the model creates a risk score (specific to the patient) that represents the probability of postoperative complications. This was what transforms the results of the abstract machine learning into a structured, patient-centered display and makes the tool more applicable and implementable in clinical practice. By presenting predictions of the model in a format that was easily comprehensible, clinicians will be able to review the risk profiles of an individual patient at a deeper level when managing the patient at the perioperative level.

The most important feature of this output was that it gives a numerical risk score, which was backed up by categorical classification (low, moderate, or high risk). This stratification helps clinicians to easily find patients who are likely to demand increased monitoring or preventive measures. Notably, the system can not only provide binary classification, but it also provides contextualization of the prediction via gradation, which increases the clinical interpretability. This feedback loop was in accordance with the aim of facilitating real-time decision-making as opposed to the elimination of human judgment.

The synthesis of various patient-level inputs, such as demographic, physiological, and intraoperative parameters, into a consolidated output was captured in figure 11. Its smooth combination of the various variables shows the complexity of the perioperative outcomes and the capacity of the model to handle multidimensional data. This fact implies that clinicians do not oversimplify any one of the factors and the output takes into account dependencies between different features. This causes more trust in the predictions, as the estimation of the risk can be considered evidence-based and not arbitrary.

The prototype highlights the possible opportunity of integrating complex predictive analytics as part of the daily clinical operations. The system helps minimize the break in tracking data science and applying it at the bedside, which was achieved by the clear representation of outputs in a visually guided format. This will see to it that the advanced computational models are not trapped in the academic phase of development but converted into instruments that supplement clinical vigilance and patient safety. The visualization thus does not only emphasize the performance of the model but also emphasizes its value of operation in the improvement of perioperative risk management.

Limitations and Future Work

The research was also constrained by the fact that the sample size of 50 patients was relatively small, and thus the findings might not be generalized. The data was collected at one center, which may not have external validity across different groups of patients and surgical practices. Moreover, there were missing values and imputation of some variables, which could have led to the bias during model training. Though many machine learning algorithms were tested, the lack of a completely external validation cohort did not allow us to confirm that they will work in other circumstances. Lastly, feature importance and SHAP values enhanced interpretability, but other models like neural networks were less transparent, which can be a barrier to clinical acceptance.

The next round of research should be aimed at extending the data with larger multicenter cohorts to enhance strength and generalization. Including more perioperative data streams like intraoperative hemodynamic measures and anesthesia depth, as well as postoperative neurocognitive measures, may make predictions more accurate. It will be important to externally validate using different populations to ensure reliability prior to clinical implementation. Additional explainable AI methods to SHAP, such as counterfactual explanations, can also enhance transparency and clinician trust further. Lastly, the predictive model needs to be incorporated into real-time clinical decision support systems, and prospective studies need to be carried out to assess its efficiency in the reduction of the POD incidence and the patient outcomes.

Conclusion

This research paper proved that advanced machine learning methods are possible to address and be useful in predicting postoperative delirium (POD) in individuals undergoing cardiac surgery. The models were capable of revealing people at high risk with high predictive value by performing the systematic analysis of a vast number of demographic, clinical, and intraoperative variables. The gradient boosting-based model achieved the best results among the evaluated algorithms, as it was shown to be strong with respect to working with complex, nonlinear relationships in clinical data.

Notably, the combination of explainability methods, including SHAP value analysis, gave a clear understanding of how given attributes of the patient affected the model prediction. This interpretability helps clinicians to be confident and make informed decisions in practice. Also, the integration of the final model into a prototype clinical decision support system indicated the possibility of its applicability in perioperative workflows, where it was possible to make timely and individualized risk estimates.

On the whole, the results indicate that the integration of predictive analytics into clinical practice may help to improve early detection of the risk of POD and implement prevention initiatives more specifically and provide better patient care. The next step in work should be to test the model on the bigger and multicenter cohort and then improve its applicability so that it can be easily integrated into the work routine.

REFERENCES

[1] Li, Q., Li, J., Chen, J., Zhao, X., Zhuang, J., Zhong, G., ... & Lei, L. (2024). A machine learning-based prediction model for postoperative delirium in cardiac valve surgery using electronic health records. BMC Cardiovascular Disorders, 24(1), 56.

[2] Yang, T., Yang, H., Liu, Y., Liu, X., Ding, Y. J., Li, R., ... & Yu, F. X. (2024). Postoperative delirium prediction after cardiac surgery using machine learning models. Computers in Biology and Medicine, 169, 107818.

[3] Jung, J. W., Hwang, S., Ko, S., Jo, C., Park, H. Y., Han, H. S., ... & Ro, D. H. (2022). A machine-learning model to predict postoperative delirium following knee arthroplasty using electronic health records. BMC psychiatry, 22(1), 436.

[4] Nagata, C., Hata, M., Miyazaki, Y., Masuda, H., Wada, T., Kimura, T., ... & Ueno, T. (2023). Development of postoperative delirium prediction models in patients undergoing cardiovascular surgery using machine learning algorithms. Scientific Reports, 13(1), 21090.

[5] Lee, D. Y., Oh, A. R., Park, J., Lee, S. H., Choi, B., Yang, K., ... & Park, R. W. (2023). Machine learning-based prediction model for postoperative delirium in non-cardiac surgery. BMC psychiatry, 23(1), 317.

[6] Matsumoto, K., Nohara, Y., Sakaguchi, M., Takayama, Y., Fukushige, S., Soejima, H., ... & Kamouchi, M. (2023). Temporal generalizability of machine learning models for predicting postoperative delirium using electronic health

record data: model development and validation study. JMIR Perioperative Medicine, 6(1), e50895.

[7] Seth, A. (2023). MODELING POSTOPERATIVE DELIRIUM RISK: DECODING ELECTRONIC HEALTH DATA WITH ARTIFICIAL INTELLIGENCE (Doctoral dissertation, Johns Hopkins University).

[8] Han, C., Kim, H. I., Soh, S., Choi, J. W., Song, J. W., & Yoon, D. (2024). Machine learning with clinical and intraoperative biosignal data for predicting postoperative delirium after cardiac surgery. Iscience, 27(6).

[9] Tu, Y., Zhu, H., Zhang, X., Huang, S., & Tu, W. (2024). Machine Learning-Based Prediction Models for Postoperative Delirium: A Systematic Review and Meta-Analysis.

[10] Zhao, X., Li, J., Xie, X., Fang, Z., Feng, Y., Zhong, Y., ... & Zou, J. (2024). Online interpretable dynamic prediction models for postoperative delirium after cardiac surgery under cardiopulmonary bypass developed based on machine learning algorithms: a retrospective cohort study. Journal of Psychosomatic Research, 176, 111553.

[11] Bhat R, Shanbhag P. Knowledge, Attitude, and Practice Study on Cardiovascular Disease Risk Factors in the Mangalore Community. Oral Sphere J. Dent. Health Sci. 2025;1(1):19-28.

[12] Xue, X., Chen, W., & Chen, X. (2022). A Novel Radiomics-Based Machine Learning Framework for Prediction of Acute Kidney Injury-Related Delirium in Patients Who Underwent Cardiovascular Surgery. Computational and Mathematical Methods in Medicine, 2022(1), 4242069.

[13] Chen, D., Wang, W., Wang, S., Tan, M., Su, S., Wu, J., ... & Cao, J. (2023). Predicting postoperative delirium after hip arthroplasty for elderly patients using machine learning. Aging Clinical and Experimental Research, 35(6), 1241-1251.

[14] Xu, Y., Meng, Y., Qian, X., Wu, H., Liu, Y., Ji, P., & Chen, H. (2022). Prediction model for delirium in patients with cardiovascular surgery: development and validation. Journal of Cardiothoracic Surgery, 17(1), 247.

[15] Ma, R., Zhao, J., Wen, Z., Qin, Y., Yu, Z., Yuan, J., ... & Ning, X. (2024). Machine learning for the prediction of delirium in elderly intensive care unit patients. European Geriatric Medicine, 15(5), 1393-1403.

[16] Strating, T., Hanjani, L. S., Tornvall, I., Hubbard, R., & Scott, I. A. (2023). Navigating the machine learning pipeline: a scoping review of inpatient delirium prediction models. BMJ Health & Care Informatics, 30(1), e100767.

[17] Zhang, N., Fan, K., Ji, H., Ma, X., Wu, J., Huang, Y., ... & Wang, Y. (2023). Identification of risk factors for infection after mitral valve surgery through machine learning approaches. Frontiers in Cardiovascular Medicine, 10, 1050698.

[18] Kavitha, R., Prasad, K., Shreee, S. A., Maheshwari, B., Sai, G. J., & Gowda, V. D. (2023, July). A novel method of identification of delirium in patients from electronic health records using machine learning. In 2023 World Conference on Communication & Computing (WCONF) (pp. 1-6). IEEE.

[19] Zhang, Y., Ren, M., Zhai, W., Han, J., & Guo, Z. (2024). Construction and validation of a risk prediction model for postoperative delirium in patients with off-pump coronary artery bypass grafting. Journal of Thoracic Disease, 16(6), 3944.

[20] Contreras, M., Kapoor, S., Zhang, J., Davidson, A., Ren, Y., Guan, Z., ... & Rashidi, P. (2024). DeLLiriuM: A large language model for delirium prediction in the ICU using structured EHR. arXiv preprint arXiv:2410.17363.

[21] Sheng, W., Tang, X., Hu, X., Liu, P., Liu, L., Miao, H., ... & Li, T. (2024). Random forest algorithm for predicting postoperative delirium in older patients. Frontiers in Neurology, 14, 1325941.

[22] Vacas, S., Grogan, T., Cheng, D., & Hofer, I. (2022). Risk factor stratification for postoperative delirium: A retrospective database study. Medicine, 101(42), e31176.

[23] Lei, L., Zhang, S., Yang, L., Yang, C., Liu, Z., Xu, H., ... & Xu, M. (2023). Machine learning-based prediction of delirium 24 h after pediatric intensive care unit admission in critically ill children: a prospective cohort study. International journal of nursing studies, 146, 104565.

[24] Sun, H., Depraetere, K., Meesseman, L., Cabanillas Silva, P., Szymanowsky, R., Fliegenschmidt, J., ... & Dahlweid, F. M. (2022). Machine learning–based prediction models for different clinical risks in different hospitals: evaluation of live performance. Journal of Medical Internet Research, 24(6), e34295.

[25] Othman, M. I., Nashwan, A. J., Abujaber, A. A., & Khatib, M. Y. (2024). Artificial intelligence applications in the intensive care unit for sepsis-associated encephalopathy and delirium: a narrative review. Avicenna, 2023(2), 11.

[26] Li, Q., Zhao, Y., Chen, Y., Yue, J., & Xiong, Y. (2022). Developing a machine learning model to identify delirium risk in geriatric internal medicine inpatients. European Geriatric Medicine, 13(1), 173-183.

[27] Abdullah, H. R. (2023). Machine Learning in the Perioperative Setting: Uncovering the Value of Data Science and Large Institutional Datasets Among Patients Undergoing Surgery (Doctoral dissertation, National University of Singapore (Singapore)).

[28] Lee, S. W., Lee, E. H., & Choi, I. C. (2023). An ensemble machine learning approach to predict postoperative mortality in older patients undergoing emergency surgery. BMC geriatrics, 23(1), 262.

[29] Ren, Y., Zhang, Y., Zhan, J., Sun, J., Luo, J., Liao, W., & Cheng, X. (2023). Machine learning for

prediction of delirium in patients with extensive burns after surgery. CNS Neuroscience & Therapeutics, 29(10), 2986-2997.

[30] Lan, L., Chen, F., Luo, J., Li, M., Hao, X., Hu, Y., ... & Zhou, X. (2022). Prediction of intensive care unit admission (> 24h) after surgery in elective noncardiac surgical patients using machine learning algorithms. Digital Health, 8, 20552076221110543.

[31] Mai, H., Lu, Y., Fu, Y., Luo, T., Li, X., Zhang, Y., ... & Chen, C. (2024). Identification of a Susceptible and High-Risk Population for Postoperative Systemic Inflammatory Response Syndrome in Older Adults: Machine Learning–Based Predictive Model. Journal of Medical Internet Research, 26, e57486.

[32] Li, P., Wang, Y., Li, H., Cheng, B., Wu, S., Ye, H., ... & Fang, X. (2023). Prediction of postoperative infection in elderly using deep learning-based analysis: an observational cohort study. Aging Clinical and Experimental Research, 35(3), 639-647.

[33] Te, R., Zhu, B., Ma, H., Zhang, X., Chen, S., Huang, Y., & Qi, G. (2024). Machine learning approach for predicting post-intubation hemodynamic instability (PIHI) index values: towards enhanced perioperative anesthesia quality and safety. BMC anesthesiology, 24(1), 136.

[34] Neto, P. C., Rodrigues, A. L., Stahlschmidt, A., Helal, L., & Stefani, L. C. (2023). Developing and validating a machine learning ensemble model to predict postoperative delirium in a cohort of high-risk surgical patients: a secondary cohort analysis. European Journal of Anaesthesiology| EJA, 40(5), 356-364.

[35] van den Eijnden, M. A., van der Stam, J. A., Bouwman, R. A., Mestrom, E. H., Verhaegh, W. F., van Riel, N. A., & Cox, L. G. (2023). Machine learning for postoperative continuous recovery scores of oncology patients in perioperative care with data from wearables. Sensors, 23(9), 4455.

[36] Fliegenschmidt, J., Hulde, N., Preising, M. G., Ruggeri, S., Szymanowsky, R., Meesseman, L., ... & von Dossow, V. (2023). Leveraging artificial intelligence for the management of postoperative delirium following cardiac surgery. European Journal of Anaesthesiology and Intensive Care, 2(1), e0010.

[37] El-Sherbini, A. H., Hassan Virk, H. U., Wang, Z., Glicksberg, B. S., & Krittanawong, C. (2023). Machine-learning-based prediction modelling in primary care: state-of-the-art review. Ai, 4(2), 437-460.

[38] Cabanillas Silva, P., Sun, H., Rezk, M., Roccaro-Waldmeyer, D. M., Fliegenschmidt, J., Hulde, N., ... & Dahlweid, F. M. (2024). Longitudinal Model Shifts of Machine Learning–Based Clinical Risk Prediction Models: Evaluation Study of Multiple Use Cases Across Different Hospitals. Journal of Medical Internet Research, 26, e51409.

[39] Arina, P., Kaczorek, M. R., Hofmaenner, D. A., Pisciotta, W., Refinetti, P., Singer, M., ... & Whittle, J. (2023). Prediction of complications and prognostication in perioperative medicine: a systematic review and PROBAST assessment of machine learning tools. Anesthesiology, 140(1), 85.

[40] Kang, Y., Sohn, S. H., Choi, J. W., Hwang, H. Y., & Kim, K. H. (2023). Machine-learning-based prediction of survival and mitral regurgitation recurrence in patients undergoing mitral valve repair. Interdisciplinary cardiovascular and thoracic surgery, 37(5), ivad176.