Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Efficacy of Different ML Models in Predicting Post-Surgical Complications in Patients

Soham Saxena

sohamsaxena08@gmail.com

ARTICLE INFO

Received: 01 Dec 2025

Revised: 10 Jan 2026

Accepted: 18 Jan 2026

ABSTRACT

Post-surgical complications significantly affect patient recovery and healthcare outcomes, making early and accurate prediction critical for timely intervention. This study evaluates the performance of eight supervised machine learning models—Artificial Neural Networks (ANN), Deep Neural Networks (DNN), Random Forest, Support Vector Machines (SVM), Logistic Regression, XGBoost, CatBoost, and ensemble learning techniques—for predicting post-surgical complications. A clinical dataset comprising surgical characteristics and post-procedural outcomes was used, incorporating key features such as DRG family, discharge volume, and complication indicators. Models were trained and tested using a 70–30 split, with one-hot encoding and class-weight adjustments to address data imbalance. Performance was assessed using accuracy, precision, recall, F-beta score, and Cohen's Kappa. The results indicate that CatBoost achieved the strongest overall performance, with the highest accuracy, recall, and F-beta score. XGBoost and ensemble models also performed well, particularly in identifying high-risk cases, while neural network models demonstrated comparatively lower performance. These findings highlight the potential of tree-based and ensemble learning approaches, especially CatBoost, for developing reliable clinical decision support systems in postoperative care.

Keywords: Post-surgical Complications, Machine Learning, CatBoost Model, AI, Healthcare Prediction, Scoring Mechanisms, Accuracy Score, Cohen's-Kappa Score

INTRODUCTION

Postoperative complications pose significant challenges in surgical care, leading to increased, prolonged hospital stays and elevated healthcare costs. Precise predictions of health issues and complications allows healthcare professionals to implement timely interventions and optimize resource allocation. However, the multifaceted nature of surgical procedures and patient-specific factors makes predicting postoperative complications a complex task.

Machine learning (ML) is emerging as a powerful resource in healthcare, offering advanced analytical capabilities to identify patterns within large datasets. By leveraging ML models, clinicians can enhance their decision-making processes, leading to improved patient care. A few studies have demonstrated the potential of ML models in predicting post-surgical complications, thereby supporting perioperative care management. At best, these have directionally suggested a positive impact of ML on care management; however, specific model efficacies have not been compared. This research precisely attempts to reduce this gap. To demonstrate, an article published by Mahajan A et al. under the JAMA Network Open Access medical journal states that a study conducted by them showed the accuracy of an "automated machine learning model" in predicting "adverse outcomes using only preoperative variables." Another research published under the National Library of Medicine presented a study of "111888 operations at a large academic medical center." This research explicitly states that machine learning models "can predict postoperative complications." Furthermore, the application of machine learning models has also been largely accepted by physicians in the medical industry. For instance, a

study conducted in the Kingdom of Saudi Arabia (KSA) showed that 76.6% of the physicians “believed that the technology improved efficiency in health care delivery.” The study also highlighted certain limitations which the study suggested “could be addressed through informed consent and staff training.”

As described earlier, in this paper, we employ various ML models like Logistic Regression, Support Vector Machines (SVM), Random Forest, CatBoost, XGBoost, Artificial Neural Networks (ANN), Deep Neural Networks (DNN), and ensemble methods such as Stacking Classifiers to predict postoperative complications. These models were selected to evaluate and compare multiple machine learning approaches on the same postoperative complication dataset. Notably, CatBoost, which has been relatively underexplored in prior postoperative complication studies, demonstrated superior performance in this research, whereas other models, which proved to be effective in other studies, achieved comparatively lower accuracy. To evaluate the performance of these models, we utilize metrics such as Accuracy, Precision, Recall, F-beta scores, and Cohen's Kappa scores. The above metrics were utilized in previous studies that were similar to this one. For example, Recall was emphasized due to the consequences of missed complications. Accuracy alone can be misleading. Multiple metrics are required to ensure a fair evaluation due to class imbalance and the clinical importance of false negatives. Together, these metrics provide a meaningful and statistically reliable evaluation, as established in prior peer-reviewed studies.

Overall, this research explores the predictive capabilities of various ML models in the context of postoperative complications. By identifying models best suited for early detection, we aspire to contribute to the development of proactive healthcare strategies that prioritize early intervention and improve postoperative care outcomes.

DATASET

The data used in this research paper were obtained from a comprehensive clinical dataset containing information about various surgical procedures and their associated post-surgical complications. The dataset is named ‘Surgical complications (Canterbury, NZ, 2014-2018),’ and the data were obtained using the New Zealand Official Information Act from the Canterbury District Health Board. It includes data on 216 distinct surgical procedure categories, with 23 features representing different types of post-surgical complications and discharge statistics. Some of the key features include "Number of Discharges," "Number of Discharges with at least one Complication," and various medical codes such as "T81 Complications of procedures NEC," "J95 Postprocedural respiratory disorders NEC," and "I97 Postproc disorder circulatory system NEC." These features capture the presence and the nature of complications across different surgical cases.

The dataset, though, is relatively small. The limited sample size and region-specific data may restrict the generalizability of the findings and increase sensitivity to class imbalance, potentially affecting the results. Additionally, access to real-world surgical outcome data is inherently challenging due to ethical, legal, and privacy constraints. The dataset used in this study was obtained via an Official Information Act request to the Canterbury District Health Board (CDHB), reflecting the limited availability of comprehensive postoperative complication data for research purposes. However, no missing values or data in any column or row were detected in the dataset, leading to minimal data-cleanup. Furthermore, the data is suitable for machine learning models that aim to predict the likelihood of complications based on the type of surgery and related factors. This particular Kaggle dataset (Surgical Complications, 2014-2018) provided open access to real surgical complication records, which enabled transparent and reproducible research without violating patient confidentiality. Furthermore, it includes both numerical and categorical features that reflect patient events and surgical outcomes, enabling a rich feature set for machine learning models. Also, using a real-world dataset with documented complications strengthens

Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

the practical applicability of the findings. This dataset serves as a valuable medium for studying trends in surgical complications and exploring predictive models for proactive healthcare management.

RESEARCH METHODOLOGY

This research follows a structured methodology that includes dataset preprocessing, model development, and evaluation.

I. **Data Acquisition:** The dataset used for this study consists of 216 rows and 23 columns. Each row represents a surgical category related to postoperative complications. Some examples of these rows are:

a. Number of Discharges

b. Number of Discharges with at least one complication

c. Postprocedural Respiratory Disorders

d. Postprocedural Disorder Circulatory System

The target variable was not pre-defined in the dataset and had to be created using data from the 'Number of discharges with at least one complication' row.

II. **Data Preprocessing:** The target variable had to be identified. One-hot encoding was also utilized for categorical variables.

a. **Target Variable Identification:** The column "% of Discharges with at least one complication" was in a percentage format, so it had to be converted to a binary format to be usable. By using binary classification models, a lambda function stripped the percentage symbol and converted the value into a binary target. For eg, values above 0% were considered complicated cases (1), while 0% cases were marked as a non-complication (0).

b. **Handling Categorical Variables:** The feature, "DRG Family", categorizes the type of surgery. This was the only categorical variable in the dataset, so it had to be handled by one-hot encoding. This approach converts each category into a binary feature column, allowing the models to understand the data without assuming an ordinal relationship. The line of code, "drop_first = True", ensures that one category is dropped to avoid multicollinearity.

III. **Train-Test Split:** The data was split into training and test sets using a 70-30 ratio, where 70% of the dataset was used for training, and 30% was used for testing. Roughly, out of 216 samples, 151 were used for training, while 65 were used for testing. This split was consistently applied in all Python files. This is a commonly used approach in medical machine learning studies, where models are trained from scratch. The dataset consists of 216 surgical procedure categories, making it essential to balance learning capacity with reliable evaluation. A 70 - 30 train-test split was chosen because "a larger training set allows the model to learn more effectively," while a 30% testing set ensures "a fair evaluation of the model's effectiveness." The complexity of predicting post-surgical complications across 23 post-procedural features further supports this choice, as "determining the optimal ratio hinges on factors such as dataset size and complexity." Overall, this split achieves the objective of "allowing the model to learn well while also ensuring it can be tested accurately." (Sivakumar et al., 2024)

IV. **Model Selection and Development:** Eight models were selected and shortlisted as they are commonly used in classification tasks. These models have historically demonstrated strong performance in handling structured and categorical data. These models represent a mix of tree-based, linear, ensemble learning and neural network approaches, allowing for a holistic comparison of their effectiveness in predicting post-surgical complications. These supervised machine learning models were developed by making use of the Scikit-Learn library for binary classification of complication vs. no-complication data:

Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

a. Logistic Regression
b. Random Forest
c. Support Vector Machines (SVM)
d. XGBoost
e. CatBoost
f. Artificial Neural Network (ANN)
g. Deep Neural Network (DNN)
h. Ensemble Learning

In some models, class weight imbalance was also addressed to give higher weight to the minority class. Class imbalance handling is a standard recommendation in clinical ML applications (Shickel et al., 2023).

V. **Evaluation Metrics:** All models were evaluated using standard classification metrics:

a. Accuracy
b. Precision
c. Recall
d. F-beta Score (beta = 1)
e. Cohen's Kappa Score

Confusion matrices were then plotted to visualize the classification outputs, where we could see the number of:

a. False Positives (FP) - The model wrongly predicts a complication when none occurred.
b. False Negatives (FN) - The model does not identify complication cases that actually occurred.
c. True Positives (TP) - The model predicts the presence of a complication correctly.
d. True Negatives (TN) - The model correctly predicts the absence of a complication.

SCORING MECHANISMS

A. ACCURACY SCORE

The accuracy score is used widely as an evaluation metric in medical applications (Hicks et al., 2022). It is a scoring mechanism that measures a model's overall correctness. In other words, it is the ratio of all reasonably accurate predicted instances to the total number of predictions. In this study, accuracy allows for an initial comparison across different machine learning models to identify which algorithm performs better overall. To understand accuracy, we can say that:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

B. PRECISION SCORE

The precision score allows us to evaluate how accurately the model identifies complication cases. This is done to not only ensure that patients who are predicted to be at risk truly have complications, but also to reduce false positive predictions. In medical settings, a high precision score means a lower

Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

number of false positives. In simple words, the model does not unnecessarily classify healthy patients as high-risk. To understand this further, we can form an equation. 'C' represents class, which can be either positive or negative:

$$\text{Precision} = \frac{TC}{TC + FC}$$

C. *RECALL SCORE*

The recall score is also known as sensitivity, and this term has been used interchangeably throughout this study. This scoring metric informs the model's ability to identify patients who actually experienced postoperative complications. Here, recall becomes essential because failing to detect a complication could result in delayed treatment or serious clinical consequences. A high recall score, thus, indicates that the model is effective in capturing most real complication cases. To represent this situation, the following equation can be understood:

$$\text{Recall} = \frac{TP}{TP + FN}$$

D. *F-BETA SCORE (BETA = 1)*

The F-Beta score has been utilized in this paper to enable a balanced evaluation of both precision and recall. Since $\boldsymbol{\beta}$ = 1 gives equal importance to both scoring mechanisms, it was useful in this study to avoid false positives and minimize missed complication cases. This metric allowed for a more holistic comparison of model performance rather than just relying on a single measure. The F-beta score can be shown through the following equation:

$$\text{F-beta} = \frac{2 \times TP}{2 \times TP + FP + FN}$$

E. *COHEN'S KAPPA SCORE*

The Cohen's Kappa score was used to evaluate the level of agreement between the model's predictions and the actual outcomes, while also considering the agreements occurring by chance. This was important in our study because of the imbalance in our dataset. A higher Cohen's Kappa score indicates the model's predictions are better than random guessing, making it a stronger metric that tells us about real-world reliability. If Kappa scores range from 0 to 0.60, it shows a fair or moderate agreement. Any value ranging from 0.60 to 1.00 shows a substantial or almost perfect agreement (GeeksforGeeks, 2025b). The formula for calculating the Cohen's Kappa score can be seen below:

$$\text{Cohen's Kappa} = \frac{2 \times (TP \times TN - FN \times FP)}{(TP + FP) \times (FP + TN) \times (TP + FN) \times (FN + TN)}$$

MACHINE LEARNING MODELS

A. *LOGISTIC REGRESSION*

The logistic regression model applies a sigmoid transformation that maps any input value to a range between 0 and 1 (Brownlee, 2023). This allows us to predict whether post-surgical complications occur (a value of '1' is generated) in a particular situation or don't occur (a value of '0' is generated). The logistic function is as follows:

$$S = b_0 + b_1X_1 + b_2X_2 + \ldots + b_iX_i$$

$$P = \frac{e^S}{1 + e^S}$$

Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

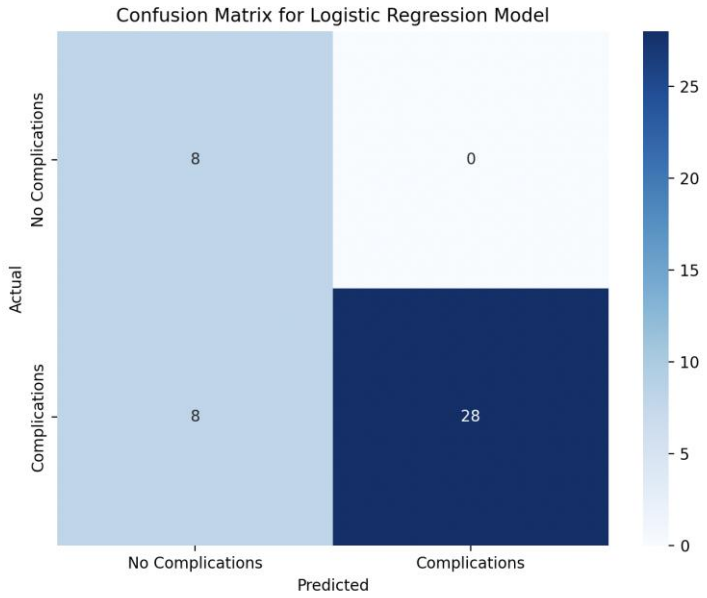
Research Article

Here, the function, 'P', represents the probability of a post-surgical complication for a given patient. b_0 is the intercept (the predicted log-odds when x = 0). b_1, b_2, ..., b_n are the coefficients or weights added to each feature, quantifying how much each feature affects the log-odds of the complication. X_1, X_2, ..., X_i are the patient-specific features used in the model. They can correspond to factors such as age, BMI, surgery duration, and type of surgery, among others. Overall, the model works by calculating a weighted sum of all the features (as defined by the equation 'S') and then applying the logistic function 'P' to map this sum to a probability between 0 and 1. Finally, if the value obtained is greater than or equal to 0.5, a value of 1 is produced, and if the value is less than 0.5, a value of 0 is produced.

To demonstrate the working of the logistic regression model, a representative datapoint from the test set was selected. Using the learned model parameters, the value of S was found to be $S = -0.7918 + \Sigma b_i X_i$, where $\Sigma b_i X_i = 1.58$. The final value of S was thus 0.7882. By applying the logistic function to this value of S, we get:

$$P = \frac{e^{0.7882}}{1 + e^{0.7882}}$$, which is approximately equal to 0.687

Since P > 0.5, the predicted class becomes 1, indicating that the model predicts the occurrence of a post-surgical complication. This matches the actual value in the dataset. The application of the model yields the following confusion matrix:



Source: Author's Data

In this case, a Cohen's Kappa score of 0.56 indicates moderate agreement between the model predictions and actual outcomes. This shows that the logistic regression model is only moderately reliable for predicting post-surgical complications and may not be sufficient for clinical application. To expand on this and to analyse the confusion matrix, the regression model correctly identified 28 cases of postoperative complications and 8 cases without complications. The model produced no false positives, resulting in a precision of 1.0, which is desirable in a clinical context where unnecessary interventions should be minimized. However, the presence of 8 false negatives suggests that some complication cases were not identified or missed. The model's tendency to miss a few true complication cases highlights the need for additional screening processes during clinical integration. In sum, while the model is highly reliable in predicting complications, additional mechanisms like doctor oversight may be required to reduce the risk of missed complication cases.

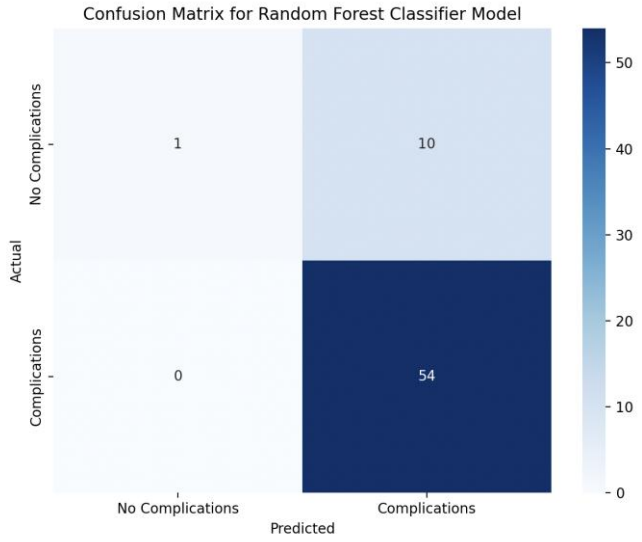
Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

B. RANDOM FOREST

The random forest model has been utilized for both classification and regression problems for many years. The model works by creating several decision trees, each of which looks at a certain part of the dataset. The model then culls out random data from each decision tree and makes a prediction. In the end, the model combines the output from all the decision trees and gives a final prediction, which is agreed to by most trees (GeeksforGeeks, 2025g). Since we have a classification problem, the model chooses a category as a final answer (1 for the occurrence of a post-surgical complication and 0 for none).

We obtain the following confusion matrix upon running the model on our dataset:



Source: Author's Data

Looking at the scoring metrics, it can be said that the model is strong in detecting complications (Recall = 1.00). However, it is not reliable in distinguishing "No Complication" cases, since it mislabels most of them. This also clearly explains why accuracy (85%) looks decent, but Kappa is low.

Furthermore, the confusion matrix shows 54 true positives, meaning the model correctly identified 54 patients who experienced post-surgical complications. This indicates a strong sensitivity toward detecting complications, which is highly critical in clinical applications. As for true negatives, the model only classified 1 patient as without complications, which shows a weaker performance in identifying non-complication cases. Moreover, the model incorrectly predicted complications for 10 patients who did not experience any. This suggests a tendency to over-predict complications, even when they don't occur. While this may not be that big of a problem, it may lead to unnecessary clinical caution. Also, notably, the model made no false-negative predictions, meaning it did not miss any actual complication cases. This adds to the reliability of the model, as such a situation is highly desirable in medical settings. Missing complications can have serious consequences. In totality, we can say that the model is risk-averse. It prioritizes patient safety by minimizing missed complication cases at the expense of some false alarms. During the decision-making process in such settings, this situation is highly preferable compared to others, as false negatives are more dangerous than false positives (Jozef Kapusta et.al.).

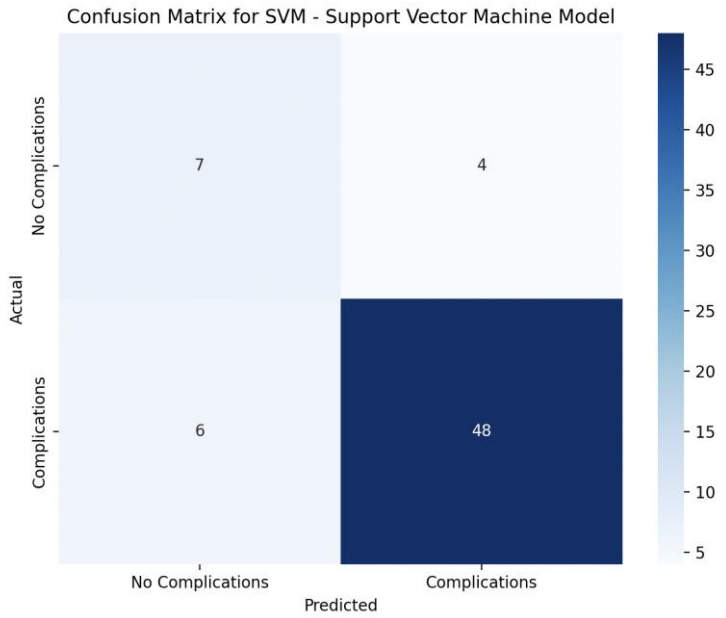
C. SUPPORT VECTOR MACHINES (SVM)

The Support Vector Machine (SVM) model was created by Vladimir N. Vapnik and his colleagues in the 1990s (Kavlakoglu, 2025a). The model is mainly utilized for classification problems such as ours. It

Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

aims to maximize the margin between two categories of data (separate 1s and 0s perfectly). A margin can be defined as the distance between two support vectors. The following confusion matrix is obtained when the model is run on our dataset:



Source: Author's Data

Through the scoring mechanisms, it can be said that the SVM model demonstrates high precision (0.92) and recall (0.89). This indicates that it accurately identifies most complication cases with minimal false alarms, making it a strong model. However, the Cohen's Kappa score (0.49) suggests only moderate agreement, so while accurate, its reliability is not entirely perfect. As for the confusion matrix, the model predicts complication cases far more often than non-complication cases, as seen by the high number of true positives (48) and relatively low number of true negatives (7). Such results were likely to be influenced by the class imbalance in the dataset, where complication cases are more prevalent compared to non-complication cases. The behaviour of the model is also more cautious, which is slightly desirable in these scenarios. Although the model performs well overall, it still fails to identify 6 patients who actually developed postoperative complications. These false negatives represent missed risk cases, and this could lead to delays in monitoring and intervention by doctors. Through this, we can evaluate that even a high-performing ML model cannot fully replace clinical judgment. It must be used as more of a support tool rather than a decision-maker. To conclude, while the SVM model performs well, models like Random Forest, XGBoost, or CatBoost may be preferred due to lower false negatives or lower recall.

D. XGBOOST

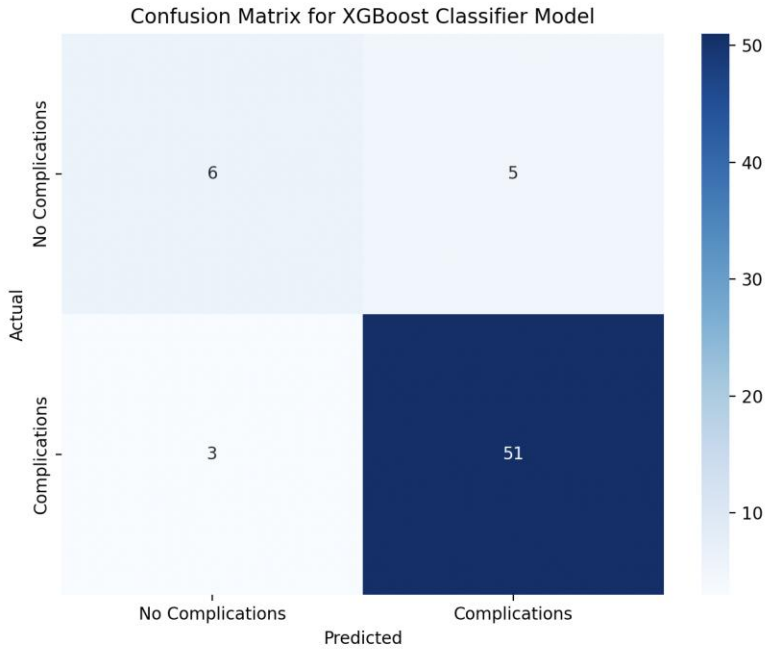
The XGBoost model, like the random forest model, utilizes decision trees for its predictions. However, there are differences to note. While the random forest model utilizes multiple decision trees that cover distinct parts of the dataset, the XGBoost model employs a single tree, trains it, and then predicts its outcome. Next, the errors from that particular outcome are used to train the next tree. This is a continuous cycle where errors are found, and each new tree is trained on the errors of the previous tree. In the end, the sum of all the outcomes is taken. Such an approach minimizes errors in the final prediction, as each tree works on the errors of the previous tree. By continuously correcting its own weaknesses, the model becomes progressively better at handling difficult or misclassified cases. This improves the consistency across its predictions. Additionally, the XGBoost model applies regularization

Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

techniques, which basically prevents the model from overfitting to the training data or the model being too complex. This is very important as overfitting can lead to high training accuracy but poor real-world performance. The model, hence, becomes more robust compared to simpler models, making it well-suited to prediction tasks such as ours.

Here is the confusion matrix generated after running the model on our dataset:



Source: Author's Data

The XGBoost model had a significantly accurate predictive performance with an accuracy score of 87.7%. It also achieved a precision score of 0.911 and a recall score of 0.944, indicating that it correctly identified nearly all complication cases, with only a few false positives. Moreover, a Cohen's Kappa score of 0.528 suggests a moderately reliable classification, despite the slight class imbalance.

To analyse the confusion matrix, the model correctly identifies the majority of patients who experienced complications. This can be clearly seen through the number of true positives (51). This highlights the model's effectiveness in understanding patterns related to post-surgical risks. Furthermore, only 3 actual complication cases were incorrectly classified as non-complication cases. This is crucial as it minimizes the number of false negatives, reducing the risk of missed clinical intervention. Additionally, the model does not excessively favor one class over the other. This is clearly shown by the relatively close values of false positives (5) and false negatives (3). This indicates that the model remains cautious when predicting non-complication cases. The model's focus on such cases is important in a clinical setting. Unnecessary alarms can lead to increased monitoring and resource usage. At the same time, the relatively small number of false negatives shows that the model is effective at identifying patients who are genuinely at risk, which is further supported by its high recall score.

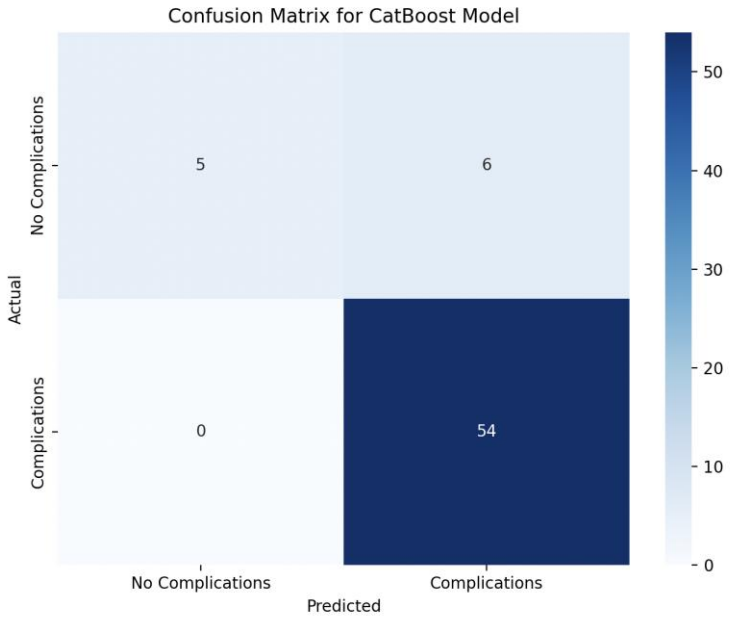
E. CATBOOST

The CatBoost Model is quite similar to the XGBoost model, which utilizes gradient boosting on decision trees and a continuous learning process through errors. However, there are a few differences to note. While XGBoost focuses on minimizing prediction errors through gradient boosting, CatBoost emphasizes stability and reduced overfitting by handling categorical features. This prevents target

Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

leakage (the model uses information during training that wouldn't be available at the time of real-world prediction) through ordered boosting. This makes CatBoost suitable for medical datasets where reliability is critical. The confusion matrix below will help us analyse the effectiveness of our model:



Source: Author's Data

The CatBoost model demonstrated the strongest overall performance compared to all the other evaluated models. It achieved the highest accuracy of 90.1% and the highest F1 score, which showed that both precision and recall were balanced. The model also achieved high recall and Cohen's Kappa scores, showing a strong agreement and greater reliability in real-world decision-making scenarios. The quality of the results highlight the model's ability to generalize well despite the dataset's size or class imbalance.

Regarding the confusion matrix, the model did not miss a single patient who developed post-surgical complications, as evidenced by the absence of false negatives. Along with this, the 54 correctly predicted true positives portray the model's high sensitivity, making it extremely suitable for decision-making in the real world, where early detection is crucial. Furthermore, there were 6 false positives. These patients were incorrectly flagged as high-risk; however, this is generally acceptable in such healthcare contexts where it is safer to over-monitor a patient than to miss a complication entirely. Overall, the CatBoost model was found to be the most effective and clinically reliable model in this study, as shown by its strong performance and high-scoring metrics. These characteristics make it suitable for predicting postoperative complications in healthcare applications, where there is little to no room for error.

F. ARTIFICIAL NEURAL NETWORKS (ANN)

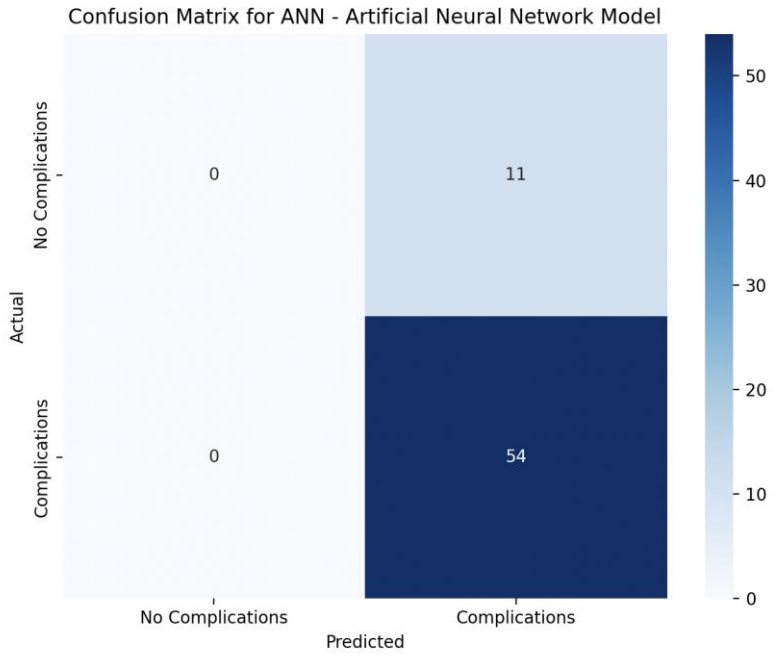
The ANN model relies on layers of interconnected neurons that process data through a network (GeeksforGeeks, 2025a). This network works by learning from the training data to solve a set of problems. The model is continually tuned over time to achieve the most accurate values possible. This tuning is done when the "network converges to a set of weights and biases that minimize error and generalize well to unseen data. (Lee, 2025)" In the end, the model identifies patterns and synthesises data to come to a prediction. The model was utilized in this particular study to evaluate whether a

Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

neural-based approach could better understand the complex, non-linear relationships within the complications dataset. The outcome of predicting post-surgical complications is highly influenced by factors such as the patient's condition or the complexity of the surgery, and these factors may not always follow linear patterns. Here, ANNs would be useful because they learn from interactions between features rather than relying on predefined functions.

Here is the confusion matrix that would help us evaluate the accuracy of the model:



Source: Author's Data

The ANN model achieved a relatively high accuracy of 83.1% and a perfect recall score of 1.000. This shows that it was successful in identifying patients who experienced post-surgical complications. However, such scores can become misleading when seen in isolation. The precision value of 0.831 and a Cohen's Kappa score of 0.000 indicate that the model cannot distinguish between the two categories of data, and that it performs no better than random chance. This further shows that the ANN model considers class imbalance, and is strongly biased towards predicting the majority class, which is complications. While a high recall may seem beneficial, the absence of a strong agreement beyond chance raises concerns about the model's generalizability or application in real-world scenarios. Moreover, this limitation can further be proved through the confusion matrix. The model predicts 54 true positives and 11 false positives, while failing to identify any non-complication cases (true negatives). Such a behaviour justifies the recall score, but also confirms that the cause of such predictions is overfitting. This issue is known to occur in neural networks when they are trained on relatively small or imbalanced datasets (Kavlakoglu, 2025a). Additionally, in real-world settings, such a model would lead to excessive false alarms, unnecessary monitoring, increased workload, or even patient anxiety. Missing a complication is very dangerous, but a model that flags every patient as high-risk is not practically useful. Therefore, despite the high recall score, the ANN demonstrates a restricted medical applicability compared to other models like XGBoost or CatBoost.

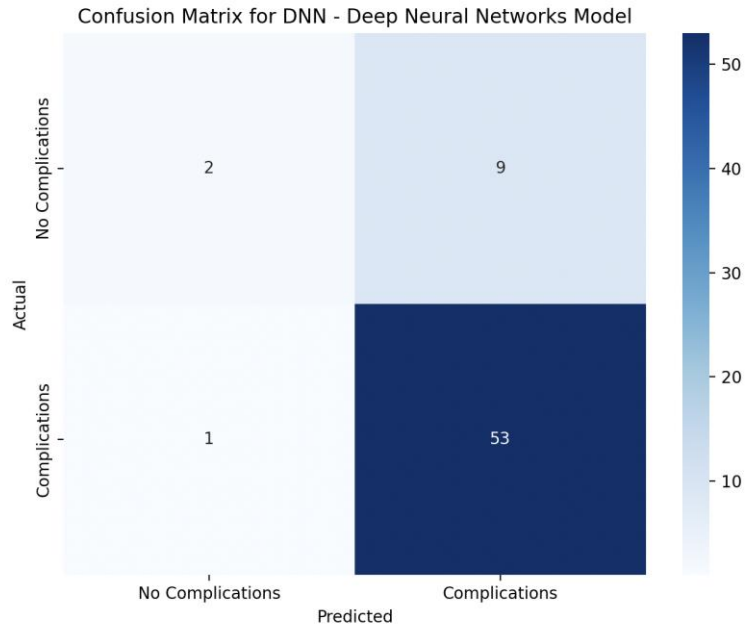
Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

G. *DEEP NEURAL NETWORKS (DNN)*

The Deep Neural Network model is similar to the artificial neural network model. It is also significant to note that the DNN model is an extension of the ANN model ("Deep Learning and Parallel Computing Environment for Bioengineering Systems," 2019). The DNN model consists of more layers that help it analyze complex problems better. Each layer enables the model to comprehend the issue and produce a successful "ideal solution" (Bharath K, 2024). Although the ANN model was utilised earlier in this study, its performance indicated certain limitations when applied to the dataset. The ANN model was ineffective and largely biased toward the majority class. Since it was a shallow neural network, compared to the DNN model, it was perhaps influenced by the dataset's imbalance and failed to capture any complex feature interactions. To address these limitations, a DNN model was tested.

Here is the confusion matrix for the application of the DNN model on the dataset:



Source: Author's Data

The DNN model achieved a strong performance with an accuracy of 0.846, indicating that a large proportion of cases were classified correctly. The model showed a high recall of 0.981, demonstrating its ability to identify patients who experienced post-surgical complications. However, despite these strengths, the Cohen's Kappa score of 0.230 suggests a low agreement beyond chance, suggesting that the model performs well numerically, but the predictions may be influenced by the class imbalance. Furthermore, the confusion matrix reveals that the DNN correctly identified 53 true positives, showing its usefulness in predicting complication cases. On the other hand, it also misclassified 9 non-complication cases as complications, indicating a tendency to over-predict risks. Only 1 false negative was observed, which reflects the model's strong recall performance. Moreover, 2 true negatives were also found, and this highlights that the model was not able to effectively distinguish between complication and non-complication cases. This pattern suggests that while the DNN is highly sensitive, it may be overfitting to complication patterns due to dataset imbalance and small sample size. As a result, although it is effective in detecting complications, the model may not be ideal as a standalone decision-making tool without further tuning or external validation.

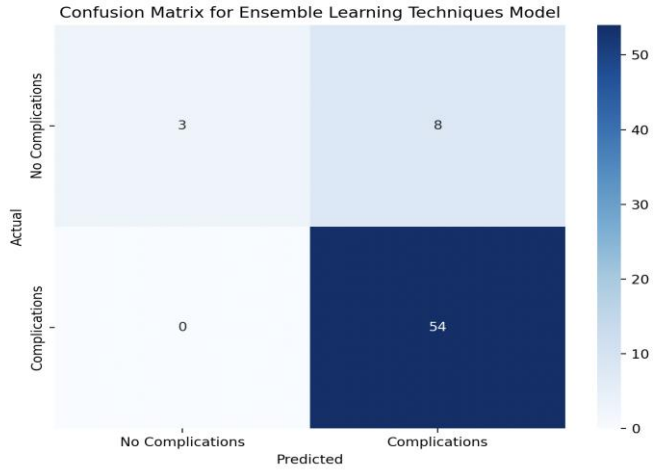
Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

H. ENSEMBLE LEARNING

The Ensemble Learning model combines other machine learning models to create a new model that aims to surpass the accuracy of the other models. In our study, we have used the stacking ensemble learning model or the stacked generalization model. In this study, we have combined the random forest model and the XGBoost model. These models are trained, and then their predictions are used as inputs for a final model, which is the ensemble learning model. This combination of models is utilized to get a more accurate answer or obtain a better performance than any of the individual models (GeeksforGeeks, 2025f). This model was applied to our dataset to check whether a combination of two models gives a higher accuracy than the individual models themselves. However, the accuracy obtained was quite similar to that of the individual models. This can be explained by the fact that both models are tree-based learners and hence capture similar decision patterns within the dataset. The Random Forest model reduced variance through bagging, and the XGBoost model reduced bias through boosting. As a result, each model became highly optimized on its own, so combining the two models did not add any data and kept the accuracy level similar.

Here is the confusion matrix that we obtained after running the model on our dataset:



Source: Author's Data

The ensemble model achieved an accuracy of 0.877, which is the same as that of XGBoost, but slightly higher than Random Forest. This shows that combining both models did not drastically increase the overall accuracy. The model also achieved a recall of 1.000, showing that it successfully identified all patients who experienced postoperative complications. The F-beta score of 0.931 further reflects a strong balance between precision and recall. At the same time, a Cohen's Kappa value of 0.384 suggests a moderate agreement beyond chance. The likely reason for this is the class imbalance and the overlapping predictions from the base models. This can be fixed by utilising a larger dataset or a larger sample size so that each class gets a better representation. Moreover, the confusion matrix shows that the ensemble correctly classified 54 complication cases (true positives), and this resulted in zero false negatives, which is a major advantage. This reflects the strong influence of XGBoost, which was effective at learning complex patterns and minimizing missed positive cases. Simultaneously, the model misclassified 8 non-complication cases as complications, indicating a tendency toward cautious predictions. This behavior is perhaps influenced by the random forest model, which averaged multiple decision trees and favored sensitivity over precision. Lastly, the small number of correctly predicted non-complication cases (3) highlights the class imbalance in the dataset, suggesting the model's limited ability in identifying low-risk patients confidently. Overall, the ensemble model performs similarly to

Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

the XGBoost model; however, it also provides more consistency and robustness to ensure that the detection process is reliable when detecting high-risk patients.

RESULTS

As shown in Table 1, the CatBoost model demonstrated the strongest overall performance among all the models. This can be clearly justified by the consistently high values across all the scoring metrics. The model showed a superior ability in predicting post-surgical complication cases, thereby emerging as one of the most reliable models as a part of this study. The strong performance is easily attributable to the model's ability to effectively handle categorical data while reducing overfitting through ordered boosting.

MODELS	SCORING MECHANISMS				
	Accuracy	**Precision**	**Recall**	**F-beta Score (Beta = 1)**	**Cohen's Kappa**
Logistic Regression	0.800	1.000	0.778	0.875	0.560
Random Forest	0.846	0.843	1.000	0.915	0.142
Support Vector Machines	0.846	0.923	0.888	0.906	0.489
XGBoost	0.877	0.911	0.944	0.927	0.528
CatBoost	0.908	0.900	1.000	0.947	0.581
Artificial Neural Network (ANN)	0.831	0.831	1.000	0.908	0.000
Deep Neural Network (DNN)	0.846	0.855	0.981	0.914	0.230
Ensemble Learning	0.877	0.871	1.000	0.931	0.384

TABLE 1

Following CatBoost, XGBoost, and the ensemble learning model produced strong results. However, the ensemble model did not significantly outperform XGBoost despite combining two models. Next, the random forest and support vector machine models achieved moderate performance. While both models showed a decent accuracy and recall score, their lower Cohen's Kappa scores suggest that they aren't consistent in their predictions and may be influenced by the class imbalance in the dataset. These results suggest that although tree-based and margin-based methods are effective, they may struggle to generalize and may not be applicable in real-world medical situations. Moreover, the neural network-based models, like the ANN and DNN models, demonstrated high sensitivity but weaker overall reliability. Their performance suggests that while neural networks are capable of learning and recognizing difficult patterns, the small dataset size likely restricted their ability to generalize effectively.

Overall, the results indicate that boosting-based models, especially CatBoost, are the most suitable for this classification task. Its ability to manage categorical variables, reduce bias, and understand non-linear relationships makes it highly effective for predicting post-surgical complications. From a real-world healthcare perspective, this becomes extremely important, as models with higher recall and agreement are better suited for clinical decision-making, where missing a complication can have serious implications.

CONCLUSION

This study evaluated multiple supervised machine learning models for predicting post-surgical complications, with the aim of identifying clinically reliable and practically applicable approaches. Among the eight models tested, CatBoost consistently demonstrated superior performance across evaluation metrics, outperforming Random Forest, neural network models, and ensemble combinations. Contrary to expectations based on prior literature, Random Forest showed comparatively weaker results, while tree-based gradient boosting models (CatBoost and XGBoost) proved more robust—highlighting the importance of algorithm selection, particularly in handling categorical clinical data and class imbalance.

From a clinical perspective, the models exhibited strong recall, supporting their utility in early identification of high-risk patients. Although Cohen's Kappa values were moderate, this is likely attributable to dataset imbalance and does not negate the models' practical relevance. The findings align with existing research that emphasizes the effectiveness of boosting and ensemble-based approaches in medical risk prediction tasks.

The results suggest that CatBoost-based predictive systems have strong potential for integration into hospital decision-support frameworks, enabling early intervention, improved resource allocation, and more efficient post-surgical care management. However, broader validation using larger, more diverse datasets and real-time clinical integration remains necessary. Future research should focus on scaling these models, incorporating richer patient-level data, and embedding them into real-world clinical workflows to enhance predictive accuracy and healthcare impact.

BIBLIOGRAPHY

[1] Alhazmi, F. (2025). Understanding physician attitudes toward AI in clinical decision-making. *JMIR Formative Research, 9*, e79730. https://doi.org/10.2196/79730
[2] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD Conference*. https://doi.org/10.1145/2939672.2939785
[3] Cramer, J. (2004). The early origins of the logit model. *Studies in History and Philosophy of Biological and Biomedical Sciences, 35*(4), 613–626. https://doi.org/10.1016/j.shpsc.2004.09.003
[4] Hicks, S. A., et al. (2022). On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports, 12*(1), 5979. https://doi.org/10.1038/s41598-022-09954-8
[5] Park, Y., et al. (2024). Federated learning model for predicting major postoperative complications. *arXiv*. https://arxiv.org/abs/2404.06641
[6] Prokhorenkova, L., et al. (2017). CatBoost: Unbiased boosting with categorical features. *arXiv*. https://arxiv.org/abs/1706.09516
[7] Scarpato, N., et al. (2024). Evaluating explainable machine learning models for clinicians. *Cognitive Computation, 16*(4), 1436–1446. https://doi.org/10.1007/s12559-024-10297-x
[8] Shapey, I. M., & Sultan, M. (2023). Machine learning for prediction of postoperative complications. *Artificial Intelligence Surgery, 3*(1), 1–13. https://doi.org/10.20517/ais.2022.31

[9] Shickel, B., et al. (2023). Dynamic predictions of postoperative complications using explainable deep learning. *Scientific Reports, 13*(1), 1224. https://doi.org/10.1038/s41598-023-27418-5
[10] Sivakumar, M., et al. (2024). Trade-off between training and testing ratios in medical ML. *PeerJ Computer Science, 10*, e2245. https://doi.org/10.7717/peerj-cs.2245
[11] Xue, B., et al. (2021). Machine learning models using perioperative data to predict complications. *JAMA Network Open, 4*(3), e212240. https://doi.org/10.1001/jamanetworkopen.2021.2240
[12] Zamzam, Y. F., et al. (2024). Comparison of CatBoost and Random Forest for medical classification. *JEEEMI, 6*(2), 125–136. https://doi.org/10.35882/jeeemi.v6i2.382
[13] Surgical Complications Dataset (Canterbury, NZ). Kaggle. https://www.kaggle.com/datasets/moshel/surgical-complications-canterbury-nz-20142018