

Use of supervised machine learning algorithms in predicting postoperative mortality in gastrointestinal and HPB surgeries.

Dr. Bhavin Vasavada,

Consultant hepatobiliary and liver transplant surgeon,

Department of hepatobiliary and liver transplant surgery,

Shalby Hospitals,

Ahmedabad.

Email: drbhavin.liversurgeon@gmail.com.

Ethical clearance: obtained from shalby hospital ethics committee.

Abstract:

Aim of the study:

This study aims to evaluate supervised machine learning algorithms in predicting 90 days post-operative mortality in gastrointestinal and HPB surgeries and comparing them with standard logistic regression methods.

Methods:

We evaluated various supervised machine learning classification algorithms like gradient boosting, K-nearest neighbours, random forest, and support vector machines with standard logistic regression methods. We used accuracy and the Receiver operating curve to compare

the methods. 60% of the data were used for training, 20% for validation and 20% for testing. We used JASP 0.16.04 by the University of Amsterdam to run machine learning algorithms and statistical analysis.

Results:

We used data from 504 patients who have undergone gastrointestinal and hepatopancreatic biliary surgery between April 2016 and March 2023. We analyzed algorithms for predicting 90 days post-operative mortality based on features like Major surgeries, Surgeries for malignancies, age, CDC grade of surgeries, Intraoperative hypotension, Open vs Laparoscopic surgeries, ASA grade, Emergency surgeries, Operative time, Intraoperative blood product used, colorectal surgeries, small intestinal surgeries, HPB surgeries, upper gastrointestinal surgeries and hernia. Test accuracies were 96% for gradient boosting, 90 % for K-nearest neighbours, 96% for the random forest, 94% for support vector and Areas under the ROC curve were 0.802 for gradient boosting, 0.489 for K-nearest neighbours, 0.934 for random forest and 0.5 for support vector algorithms. Accuracy and Area under the ROC curve with standard logistic regression method were 94% and 0.757. Features of importance in decreasing order were ASA, operative times, blood products, small bowel surgeries and Age.

Conclusion:

Supervised machine learning algorithms particularly gradient boosting and random forest predicted 90 days post-operative mortality more accurately than logistic regression and such models can be part of the preoperative evaluation in gastrointestinal and HPB surgeries.

Background:

In recent times machine learning has been increasingly used in medical research. In recent times machine learning is found superior to standard statistical methods, particularly regarding complex data structure and helps in studying and making predictions even in nonlinear relationships. Machine learning has many algorithms and is particularly classified into supervised and unsupervised machine learning algorithms. Supervised machine learning methods are used to describe prediction tasks because the goal is to forecast/classify a specific outcome of interest. [1].

There is various type of supervised machine learning algorithms and the support vector algorithm has been applied more frequently but in various studies, the random forest algorithm has shown superior accuracy. [2]

Our study aims to evaluate postoperative mortality using supervised machine learning algorithms in our data and compared them with standard logistic regression statistical analysis to evaluate the usefulness of the model and also evaluate various preoperative and intraoperative features of importance in predicting mortality.

Methods:

We used the data of all the patients operated on for gastrointestinal and hepato-pancreaticobiliary surgery in our institute between April 2016 to March 2023. 60% of the data was used for training, 20% for validation and 20% as a test cohort. We kept 90 days mortality as our target variable. We used various pre-operative and intraoperative features like major surgeries, Surgeries for malignancies, age, CDC grade of surgeries, Intraoperative hypotension, Open vs Laparoscopic surgeries, ASA grade, Emergency surgeries, Operative

time, Intraoperative blood product used, colorectal surgeries, small intestinal surgeries, HPB surgeries, upper gastrointestinal surgeries and a hernia for creating the model.

Definitions:

Mortality:

Nine-day mortality was defined as any cause of mortality in the 90-day postoperative period.

Ninety-day mortality included all the in-hospital mortalities.

.

Intraoperative Hypotension:

Intraoperative hypotension was defined as systolic arterial pressure below 80 mmHg, a decrease in systolic arterial pressure by 20% below baseline, or vasopressor requirement.

Centre of Disease Control Grading:

We also defined surgical wounds according to the Centre for Disease Control as clean (grade 1), clean-contaminated (grade 2), contaminated (grade 3), and dirty (grade 4).

Major and Nonmajor Surgery:

We defined surgeries with literature-proven negligible mortality like laparoscopic cholecystectomy, all hernia surgeries, and laparoscopic appendicectomies as nonmajor surgeries and other surgeries as major surgeries. All emergency surgeries except for the abovementioned surgeries were also defined as major surgeries.

Supervised Machine learning models:

We evaluated various supervised machine learning classification algorithms like gradient boosting, K-nearest neighbours, random forest, and support vector machines with standard logistic regression methods. We used accuracy and the Receiver operating curve to compare the methods. 60% of the data were used for training, 20% for validation and 20% for testing. As the target variable was categorical, we used classical We used JASP 0.16.04 by the University of Amsterdam to run machine learning algorithms and logistic regression analysis was also done using JASP 0.16.04. We also evaluated the confusion matrix, class proportion, evaluation metrics, deviance and out bag improvement plots, relative influences of features and decision boundaries matrix as per the methods applied. We used out-of-box classification error plots and deviance plots to check model accuracy.

Results:

We used data from 504 patients who have undergone gastrointestinal and hepatopancreatic biliary surgery between April 2016 and March 2023. We analyzed algorithms for predicting 90 days post-operative mortality based on features like Major surgeries, Surgeries for malignancies, age, CDC grade of surgeries, Intraoperative hypotension, Open vs Laparoscopic surgeries, ASA grade, Emergency surgeries, Operative time, Intraoperative blood product used, colorectal surgeries, small intestinal surgeries, HPB surgeries, upper gastrointestinal surgeries, and hernia. Test accuracies were 96% for gradient boosting, 90 % for K-nearest neighbours, 96% for the random forest, 94% for support vector and Areas under the ROC curve were 0.857 for gradient boosting, 0.489 for K-nearest neighbours, 0.767 for random forest and 0.5 for support vector algorithms.

As gradient boosting and the random forest were showing the highest accuracies and AUROC (area under ROC curve) we analyzed them in detail.

Gradient boosting algorithm:

As shown in Table 1, 321 patients as training data, 81 for validation and 100 for the test.

Validation accuracy was 97.5% and test accuracy was 96%.

Boosting Classification

Trees	Shrinkage	n(Train)	n(Validation)	n(Test)	Validation Accuracy	Test Accuracy
14	0.100	321	81	100	0.975	0.960

Note. The model is optimized with respect to the *out-of-bag accuracy*.

Evaluation Metrics

	Survived Mortality Average / Total		
Support	96	4	100
Accuracy	0.960	0.960	0.960
Precision (Positive Predictive Value)	0.960	NaN	0.922
Recall (True Positive Rate)	1.000	0.000	0.960
False Positive Rate	1.000	0.000	0.500
False Discovery Rate	0.040	NaN	0.040
F1 Score	0.980	NaN	0.940
Matthews Correlation Coefficient	NaN	NaN	NaN
Area Under Curve (AUC)	0.858	0.855	0.857
Negative Predictive Value	NaN	0.960	0.960
True Negative Rate	0.000	1.000	0.500

Evaluation Metrics

	Survived Mortality Average / Total		
False Negative Rate	0.000	1.000	0.500
False Omission Rate	NaN	0.040	0.040
Threat Score	12.000	0.000	6.000
Statistical Parity	1.000	0.000	1.000

Note. All metrics are calculated for every class against all other classes.

Table 2. Evaluation Metrics.

As shown in evaluation matrices overall accuracy of the model was 96% precision or the positive predictive value was 92% and the Recall or true positive rate was 96%.

ROC curve:

Now any classification algorithms are biased towards groups with larger sample sizes and hence in the table in the mortality cohort some of the values were not accurately calculated and predicted as the overall 90 days mortality rate in the data was 6.2% so. Naturally, there is a huge gap in numbers in both groups one of the methods to circumvent this problem is to see Area Under Curve in both groups which was highly significant with an overall value of 0.857 and 0.858 and 0.855 in survival and mortality groups.

ROC Curves Plot

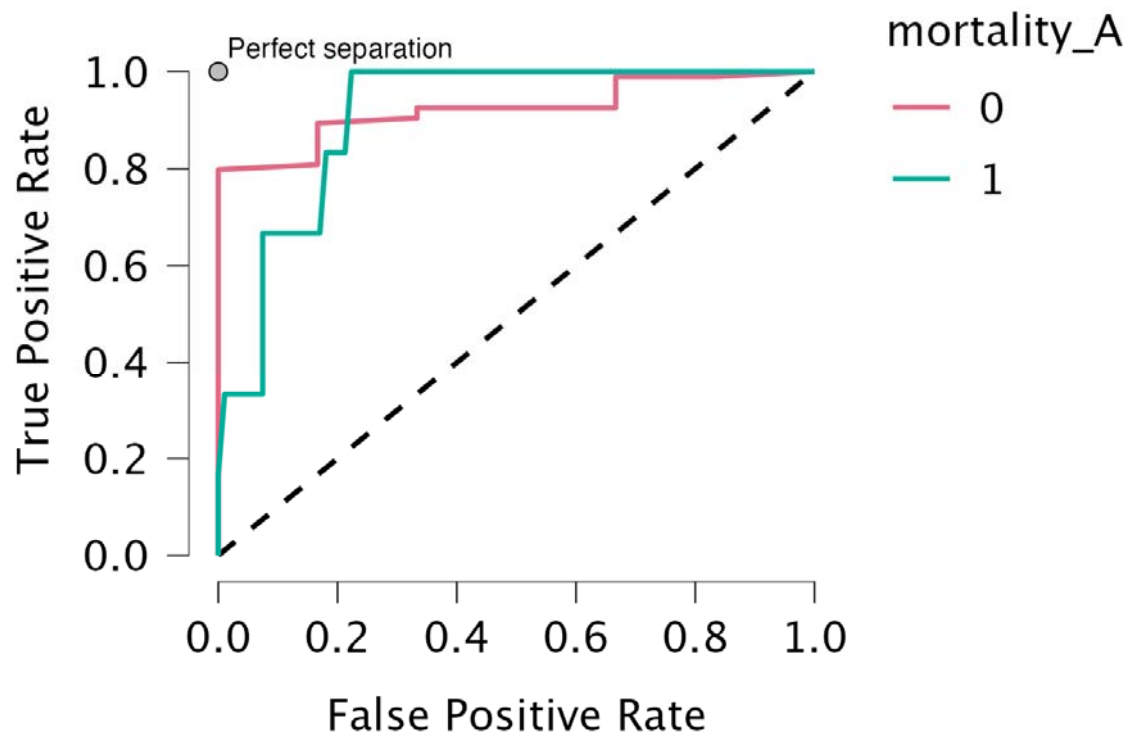
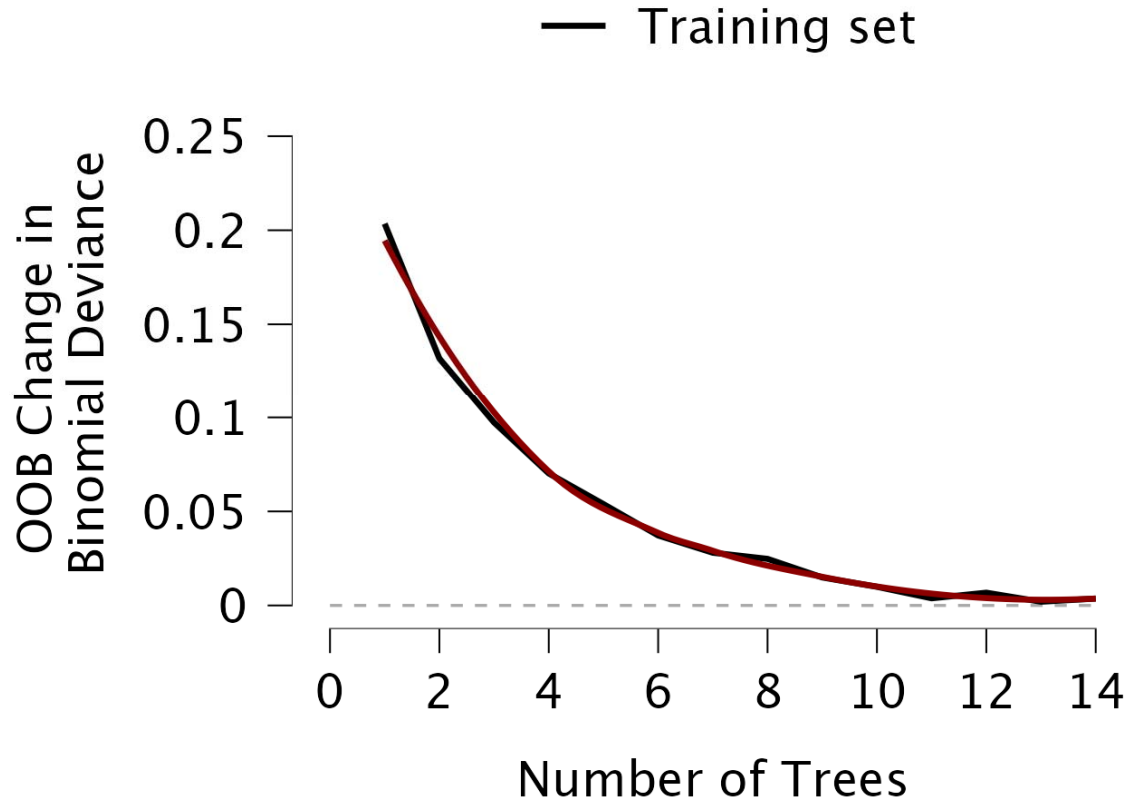


Figure 1. ROC curve of classical gradient boosting.

Out of box error and deviance plot:

Figure 2 shows of box error vs the number of trees plot, as can be seen as the number of trees increased out of box error reduced and was minimal after 14 trees showing the adequacy of the model and adequacy number of trees. The deviance plot also shows a similar trend

Out-of-bag Improvement Plot



Deviance Plot

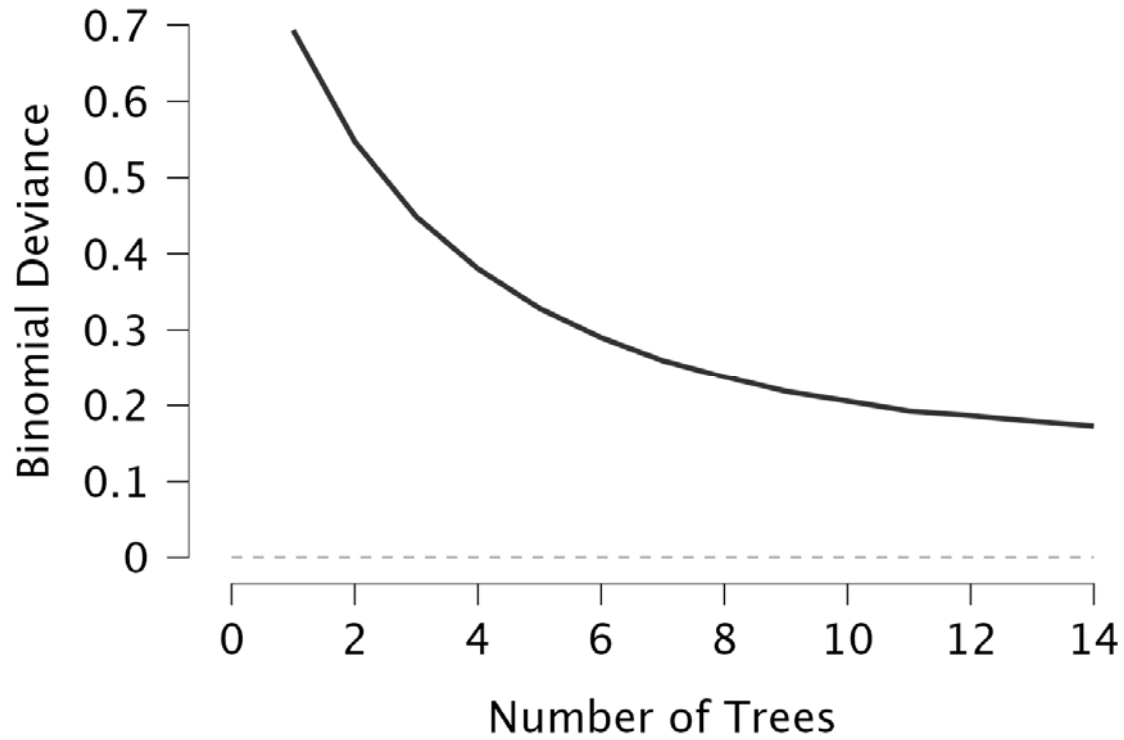


Figure 2. Out of bag improvement plot and deviance plot

Relative influences of features in predicting model:

Relative Influence

	Relative Influence
ASA	59.277
Bloodprodcuts	20.115
Operativetime	13.728
Colorectal	3.714
AGE	3.165
Major	0.000
Malignant	0.000
Gradeofsurgery	0.000

Relative Influence

	Relative Influence
Intraophypotension	0.000
Openlap	0.000
Smallbowel	0.000
UPPERGI	0.000
EMMERGENCY	0.000
HPB	0.000
Hernia	0.000

Table 3. Relative importance of features.

Relative Influence Plot

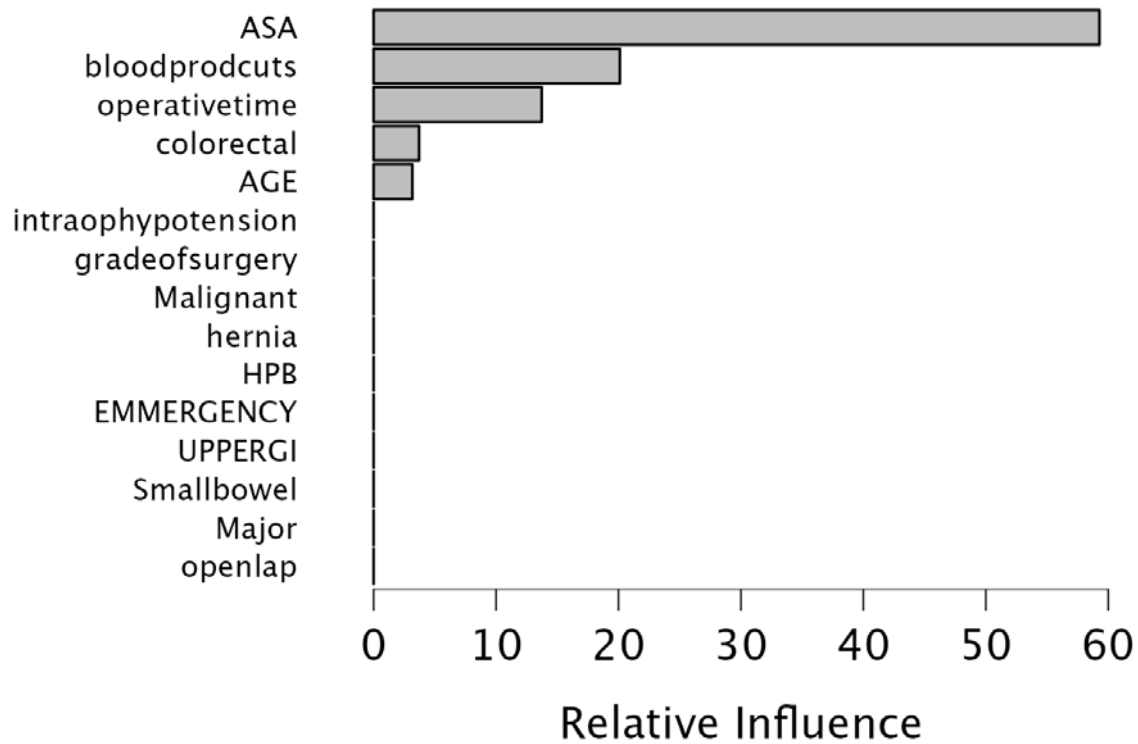


Figure 3. Relative influence of individual features in predicting mortality.

From the relative influence table, it can be seen that ASA grading was the most important feature in predicting mortality (59.27%) followed by blood products (20.11%), operative time (13.72%), colorectal surgery (3.71%) and Age of the patient (3.16%) respectively. It seems the rest of the features almost did not affect 90 days mortality.

Random forest analysis:

We are mentioning in detail the Random Forest analysis algorithm which also showed an overall higher accuracy of 96%, precision of 96.2 and recall of 96.2%. As shown in Table 4, 321 patients' data were used to train the model. 81 in the validation cohort and 100 in the test cohort.

Random Forest Classification

	Features	per				Validation	Test
Trees	split	n(Train)	n(Validation)	n(Test)	Accuracy	Accuracy	
35	3	321	81	100	0.901	0.960	

Note. The model is optimized with respect to the *out-of-bag accuracy*.

Table 4 Training, validity, and test numbers in random forest classification algorithm,

Evaluation Metrics:

Table 5 shows the evaluation matrices of the model, all other indices are acceptable but it shows high false negative rates in the mortality group again due to the model's inherent bias

towards the group with a high sample size and naturally, as overall mortality was 6%, so the model is biased towards survival, which is one of the limitations of these algorithms, however, to evaluate in such cases Area under ROC curve is taken which shows 0.780 in survival group and 0.754 in mortality group with an average of 0.767. The ROC curve is shown in Figure 4.

Table 5. Evaluation Metrics

	Survival Mortality Average / Total		
Support	95	5	100
Accuracy	0.960	0.960	0.960
Precision (Positive Predictive Value)	0.960	1.000	0.962
Recall (True Positive Rate)	1.000	0.200	0.960
False Positive Rate	0.800	0.000	0.400
False Discovery Rate	0.040	0.000	0.020
F1 Score	0.979	0.333	0.947
Matthews Correlation Coefficient	0.438	0.438	0.438
Area Under Curve (AUC)	0.780	0.754	0.767
Negative Predictive Value	1.000	0.960	0.980
True Negative Rate	0.200	1.000	0.600
False Negative Rate	0.000	0.800	0.400
False Omission Rate	0.000	0.040	0.020
Threat Score	11.875	0.250	6.063
Statistical Parity	0.990	0.010	1.000

Note. All metrics are calculated for every class against all other classes.

ROC Curves Plot

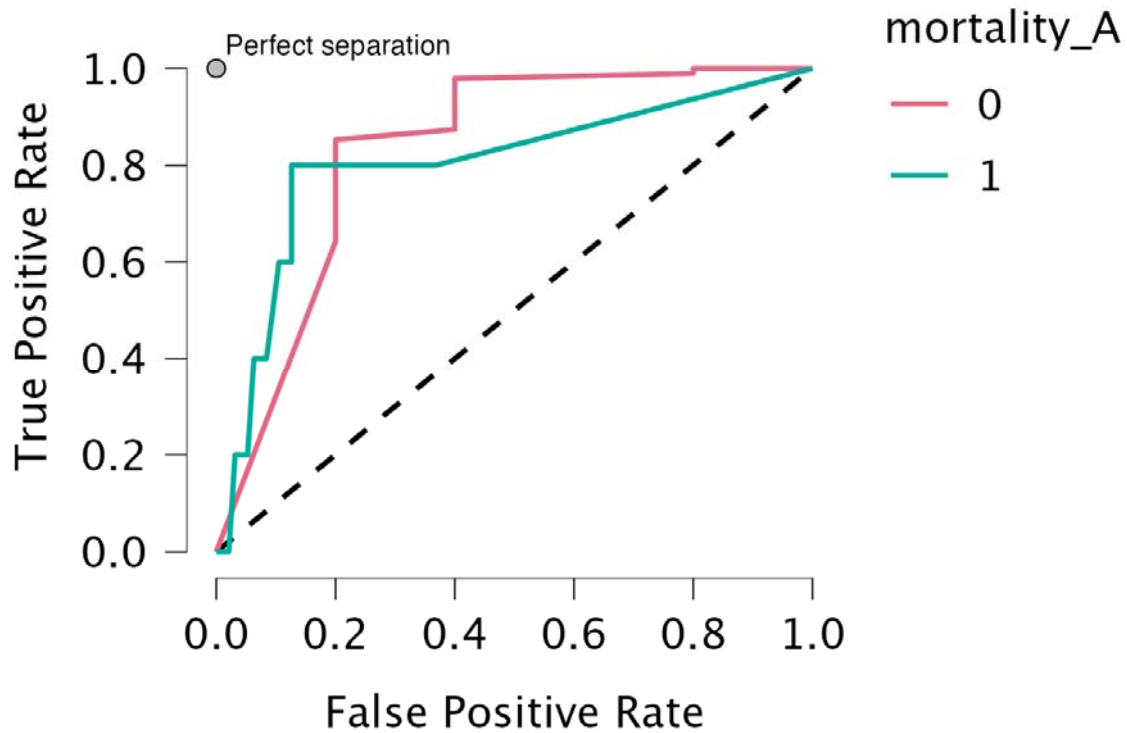


Figure 4. ROC curve red line shows survival group and green mortality group.

Out-of-bag Classification Accuracy Plot:

Figure 5 shows out of bag accuracy plot which shows as the number of trees increase accuracy is increased and around the number of 35 trees accuracy stabilises, which shows 35 trees taken in model are adequate.

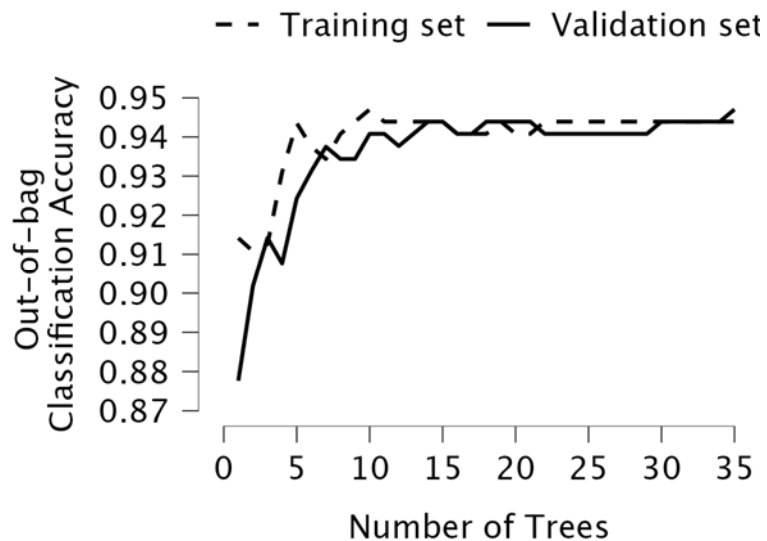


Figure 5. Out of bag accuracy plot

Mean Decrease in Accuracy

In the random forest algorithm features importance is assessed by mean decrease in accuracy, features which show the highest decrease in accuracy are the most important. In this algorithm also ASA grading was showing the highest importance, followed by open surgery, blood products, age operative time, major surgeries, emergency surgeries, small bowel surgeries, grade of surgeries, HPB surgeries and intraoperative hypotension.

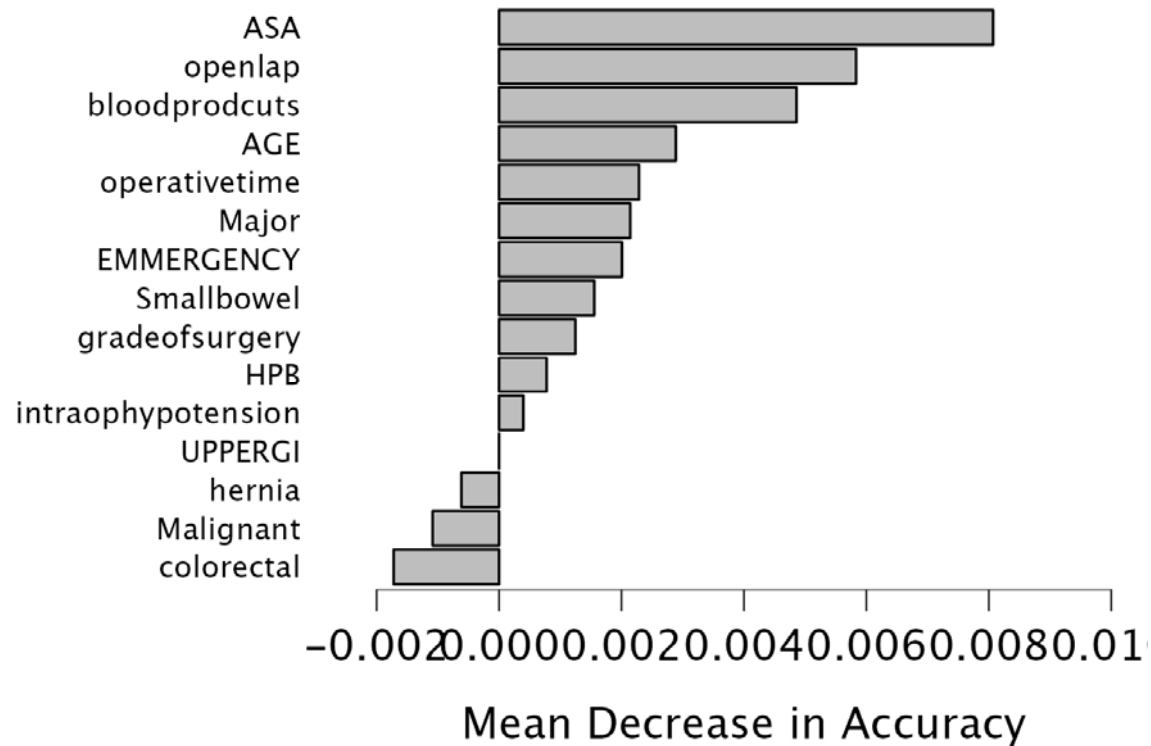


Figure 6. Features importance with mean decrease in accuracy.

Logistic regression:

We compared the accuracy of supervised machine learning models with standardised logistic regression models and their performance diagnostics.

Performance Diagnostics:

Overall accuracy was 95.2% and the area under the curve was 0.703, which showed machine learning models were performing better than standard logistic regression models.

Performance metrics.

Performance metrics

Value

Performance metrics

	Value
Accuracy	0.952
AUC	0.703
Sensitivity	0.419
Specificity	0.987
Precision	0.684
F-measure	0.520

Table 6. Performance metrics of standard logistic regression model.

Performance plots

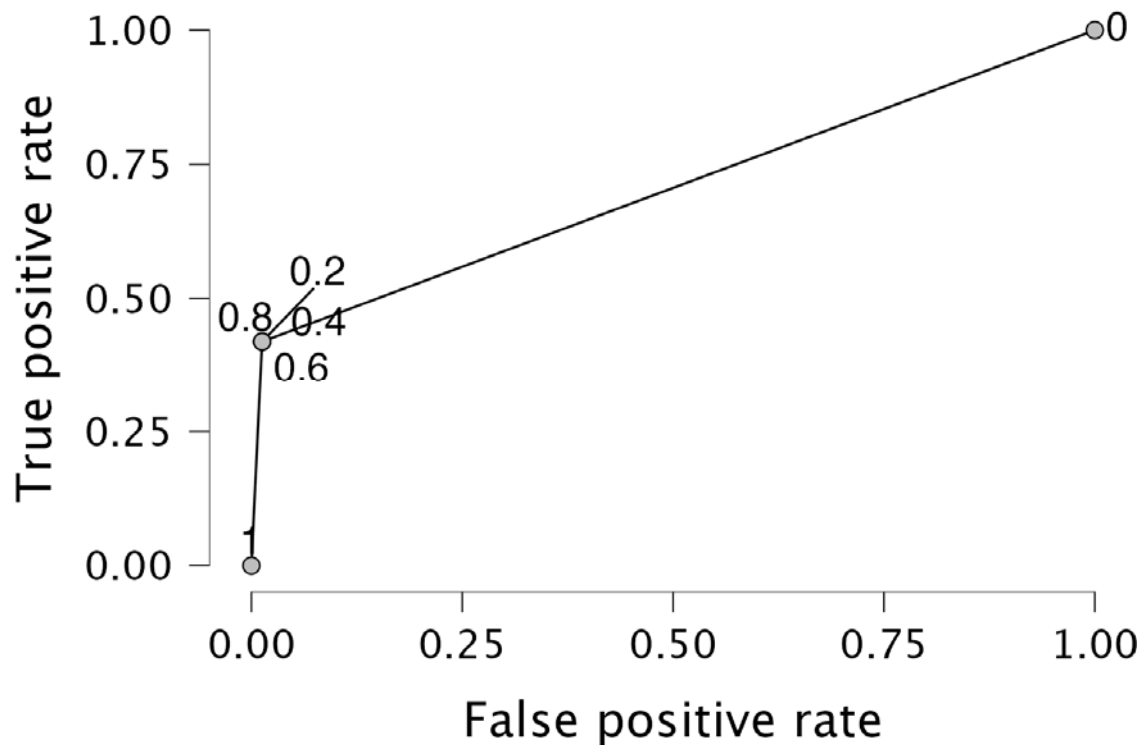


Figure 7. ROC analysis of standard logistic regression model.

Discussion:

Perioperative mortality is one of the most important problems the surgical community must face. Perioperative mortality ranges from 0.1% to as high as 27–30%, depending on the type of surgery [3, 4]. Gastrointestinal and hepatobiliary surgeries are technically demanding procedures and have among the highest perioperative mortality rates. [5,6,7]

Machine learning is a marriage between biostatistics and computer applications, and recently it has gained popularity. There are two kinds of machine learning algorithms supervised and unsupervised. Supervised machine learning algorithms are mainly used for predicting known output or target. [8] There are various scores like (P)POSSUM available to predict postoperative outcomes. [9] However, their accuracies vary according to various centres. One

of the benefits of supervised machine learning models is that they can be trained as per our data and can be used to predict outcomes, based on local factors, patients' profiles etc. and at the same time models created in larger centres can be applied to other centres locally or not locally after checking accuracy.

The primary aim of this study is to create supervised machine learning models based on our data and compare the model with the standard logistic regression method and check for features of importance in predicting the outcomes. We also wanted to evaluate the usefulness of these algorithms in preoperative assessment to predict postoperative outcomes.

As mentioned in the result section we analysed various supervised machine learning algorithms and found that gradient boosting and random forest were predicting the outcomes with the highest accuracy and area under the curve. In gradient boosting American Society of Anaesthesiology, grading was the most important feature, followed by blood products, colorectal surgeries, operative time, and age in decreasing order as shown in Figure 3. In the random forest algorithm, ASA grading was showing the highest importance, followed by open surgery, blood products, age operative time, major surgeries, emergency surgeries, small bowel surgeries, grade of surgeries, HPB surgeries and intraoperative hypotension.

Standard logistic regression showed 95% accuracy compared to 96% accuracy for both gradient boosting and random forest algorithms. It showed an area under the curve of 0.703 compared to 0.857 with gradient boosting and 0.767 for random forest algorithms, which showed supervised machine learning is more or at least as effective as logistic regression.

One of the key limitations of the study is high false negative rates with both the algorithms in the mortality group, but this is the key limitation of classification algorithms when there is a mismatch in class proportion. Overall 90 days mortality was 6.2% so there was a mismatch between survival class and mortality class and hence the finding. As matching the class in

mortality analysis was not possible we evaluated the area under the ROC curve, which is another way to evaluate the model in case of a mismatch in class limitation.

In conclusion, supervised machine learning algorithms are highly accurate and precise in predicting post-operative survival and these models can be part of the routine evaluation in predicting post-operative outcomes.

References:

1. Jiang T, Gradus JL, Rosellini AJ. Supervised Machine Learning: A Brief Primer. *Behav Ther.* 2020;51(5):675-687. doi:10.1016/j.beth.2020.05.002.
2. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med Inform Decis Mak.* 2019;19(1):281. Published 2019 Dec 21. doi:10.1186/s12911-019-1004-8.
3. Ng-Kamstra JS, Arya S, Greenberg SLM, Kotagal M et al (2018) Perioperative mortality rates in low-income and middle-income countries: a systematic review and meta-analysis. *BMJ Glob Health* 3(3):e000810. [https:// doi. org/ 10. 1136/ bmjgh- 2018- 000810](https://doi.org/10.1136/bmjgh-2018-000810)
4. Marino MV, Mituś JW, Vaccarella G et al (2020) Complications profile after robotic pancreatic surgery. *Państwo i Społeczeństwo* 2:51–63. [https:// doi. org/ 10. 34697/ 2451- 0858- pis- 2020-1- 004](https://doi.org/10.34697/2451-0858-pis-2020-1-004)
5. Csikesz NG, Simons JP, Tseng JF et al (2008) Surgical specialization and operative mortality in hepato-pancreatico-biliary (HPB) surgery. *J Gastrointest Surg* 12(9):1534–1539. [https:// doi. org/ 10.1007/ s11605- 008- 0566-z](https://doi.org/10.1007/s11605-008-0566-z)

6. Kneuertz PJ, Pitt HA, Bilimoria KY et al (2012) Risk of morbidity and mortality following hepato-pancreato-biliary surgery. *J Gastrointest Surg* 16(9):1727–1735. <https://doi.org/10.1007/s11605-012-1938-y>.
7. S.rensen LT, Malaki A, Wille-J.rgensen P et al (2007) Risk factors for mortality and postoperative complications after gastrointestinal surgery. *J Gastrointest Surg* 11(7):903–910.
8. Deo RC. Machine Learning in Medicine. *Circulation*. 2015;132(20):1920-1930. doi:10.1161/CIRCULATIONAHA.115.001593
9. Dutton J, Zardab M, De Braal VJF, et al. The accuracy of pre-operative (P)-POSSUM scoring and cardiopulmonary exercise testing in predicting morbidity and mortality after pancreatic and liver surgery: A systematic review. *Ann Med Surg (Lond)*. 2020;62:1-9. Published 2020 Dec 28. doi:10.1016/j.amsu.2020.12.016