International Journal of Surgery 107 (2022) 106954

Contents lists available at ScienceDirect

International Journal of Surgery

journal homepage: www.elsevier.com/locate/ijsu

Retrospective Cohort Study

Study and validation of an explainable machine learning-based mortality prediction following emergency surgery in the elderly: A prospective observational study

Pietro Fransvea [a], Giulia Fransvea [b,c], Piergiuseppe Liuzzi [b,c,*], Gabriele Sganga [a], Andrea Mannini [c], Gianluca Costa [d]

[a] Emergency Surgery and Trauma, Fondazione Policlinico Universitario A. Gemelli IRCCS, Università Cattolica del Sacro Cuore, Largo A. Gemelli 8, Rome, Italy
[b] The BioRobotics Institute, Scuola Superiore Sant'Anna, Viale Rinaldo Piaggio 34, Pontedera, PI, Italy
[c] IRCCS Fondazione Don Carlo Gnocchi ONLUS, Via di Scandicci 269, Firenze, FI, Italy
[d] Surgery Center, Colorectal Surgery Unit – Fondazione Policlinico Campus Bio-Medico, University Hospital of University Campus Bio-Medico of Rome, Rome, Italy

ARTICLE INFO

Keywords:
Death Following Surgery
Machine Learning
Elderly
Explainable models
Emergency care

ABSTRACT

Introduction: The heterogeneity of procedures and the variety of comorbidities of the patients undergoing surgery in an emergency setting makes perioperative risk stratification, planning, and risk mitigation crucial. In this optic, Machine Learning has the capability of deriving data-driven predictions based on multivariate interactions of thousands of instances. Our aim was to cross-validate and test interpretable models for the prediction of post-operative mortality after any surgery in an emergency setting on elderly patients.

Methods: This study is a secondary analysis derived from the FRAILESEL study, a multi-center (N = 29 emergency care units), nationwide, observational prospective study with data collected between 06-2017 and 06–2018 investigating perioperative outcomes of elderly patients (age≥65 years) undergoing emergency surgery. Demographic and clinical data, medical and surgical history, preoperative risk factors, frailty, biochemical blood examination, vital parameters, and operative details were collected and the primary outcome was set to the 30-day mortality.

Results: Of the 2570 included patients (50.66% males, median age 77 [IQR = 13] years) 238 (9.26%) were in the non-survivors group. The best performing solution (MultiLayer Perceptron) resulted in a test accuracy of 94.9% (sensitivity = 92.0%, specificity = 95.2%). Model explanations showed how non-chronic cardiac-related comorbidities reduced activities of daily living, low consciousness levels, high creatinine and low saturation increase the risk of death following surgery.

Conclusions: In this prospective observational study, a robustly cross-validated model resulted in better predictive performance than existing tools and scores in literature. By using only preoperative features and by deriving patient-specific explanations, the model provides crucial information during shared decision-making processes required for risk mitigation procedures.

1. Introduction

Although advances in surgical techniques, anesthetic procedures and postoperative care have all made surgery less hazardous, surgeons are generally more reluctant to operate on elderly because they are perceived to be frail, to have less physiological reserve and more underlying medical conditions. Several factors are thought to be related to the postoperative outcome reassuming fragility of the elderly patient [1]. Recent literature has demonstrated the role of preoperative frailty screening in predicting length of stay, operative risk, and surgical outcomes in elderly. Never as in emergency setting it is paramount to implement the decision-making process and to perform an accurate risk stratification, addressing patients' priority. Recently, medicine has witnessed the emergence of Machine Learning (ML) as a novel tool to analyze large amounts of data. One particular area where ML may add value in health care is in the setting of perioperative risk stratification – i.e., prediction of adverse events (AE), specifically death following surgery (DFS). The latter is crucial for improving shared decision-making

* Corresponding author. IRCCS Fondazione Don Carlo Gnocchi ONLUS, Via di Scandicci 269, Firenze, FI, Italy.
E-mail address: pliuzzi@dongnocchi.it (P. Liuzzi).

https://doi.org/10.1016/j.ijsu.2022.106954
Received 7 June 2022; Received in revised form 7 September 2022; Accepted 3 October 2022
Available online 11 October 2022

among the care team and the patient, perioperative planning, and risk mitigation. Furthermore, given the high heterogeneity of surgery types and the surgery-comorbidities mutual influence, ML perfectly fits the need of deriving data-driven multivariate relationships between patients characteristics and outcomes under different clinical conditions. In particular, comparisons between ML and mortality risk scores have shown notable improvement in predictive accuracies [2–8].

Prediction of DFS is often tackled by developing models targeting specific pathologies or surgery types (e.g. after open repair of abdominal aortic aneurysm [9], radical cystectomy for bladder cancer [10], hip fracture [11], and transcatheter mitral valve repair [12]). Also, level II evidence was provided concerning the improvement of mortality prediction accuracy of advanced ML models with respect to conventional biostatistical analyses [13]. In particular, the prediction of 30-day mortality of neonatal patients after surgery showed how ML can be used with age-related cohorts [14,15]. In cohorts of elderly patients, models targeting AE after specific surgical interventions have been deployed on patients suffering Multiple Organ Dysfunction Syndrome (MODS) [16], undergoing hip fracture [17] or coronary artery bypass surgery grafting [18]. Similarly, in cohorts not based on specifical surgical interventions, Lee et al. targeted the prediction of in-hospital mortality collecting features at the end of surgery [19], while Misic et al. tackled the readmission to the postoperative emergency department by using a limited number of post-operative features [20].

Nevertheless, no study targeted the development of explainable ML models targeting the prediction of mortality after all surgical procedures in emergency setting in elderly patients so far. On this path, we performed a study on a prospectively collected, large multi-center cohort with the aim to develop and cross-validate ML solutions capable of predicting DFS by using data collected up to one day before surgery. Then, the best performing solution was embedded with interpretability methods, in the form of the SHapley Additive Explanation (SHAP). The latter allows the clinicians to understand, patient-by-patient, the predictors' effect onto the prediction. Also, embedding the solution with patient-wise explanations fosters the acceptability and usability of the resulting Clinical Decision Support Tool (CDST) by clinicians, contextualizing the prediction in the optic of the patient clinical condition.

2. Materials and methods

2.1. Research protocol

This study originated from the FRAILESEL (Frailty and Emergency Surgery in the Elderly) study (ClinicalTrials.gov identifier: NCT02825082). The FRAILESEL is a large, nationwide, multicenter (N = 29), prospective study investigating perioperative outcomes of patients (≥65 years) who underwent emergency surgery between 06-2017/06–2018. Patients were recruited following the Helsinki Declaration and enrolled after signing a written consent approved by the Ethical Committee of "Sapienza" University of Rome, Italy (No.4252_2016, ClinicalTrials.gov: NCT02825082, https://clinicaltrials.gov/ct2/show/NCT02825082). This work has been reported in line with the STROCCS criteria [21].

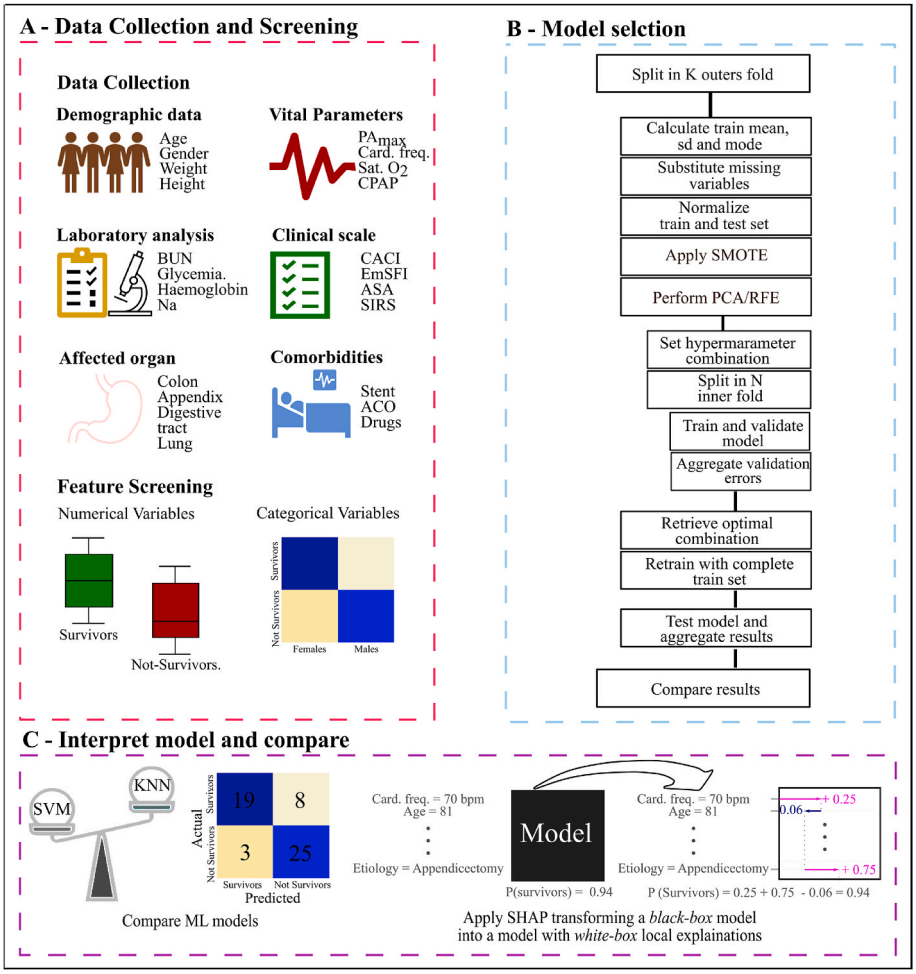

Fig. 1. Pipelines for data collection and feature screening (panel A), model development and validation (panel B) and interpretation (panel C).

2.2. Study population and data collection

The FRAILESEL study investigates over 150 variables exploring 5 domains such as patient demographic and clinical data, preoperative risk factors and operative variables, frailty condition, and postoperative outcome and follow-up (Fig. 1A). Data collected included demographic characteristics, medical and surgical history, common preoperative biochemical blood examination (e.g. C-reactive protein, procalcitonin, arterial blood gas analysis), pathological features, and operative details. Comorbidity was recorded if the condition was being medically treated at the time of admission, or if previous treatment for the condition was in the admission report. Operative procedure performed and surgical diagnoses were classified according to the 9th revision of International Classification of Disease Clinical Modification (ICD-9-CM). The type of surgical approach takes into account open or minimally-invasive procedures, including assisted procedure and conversion to open surgeries. The TNM 8th edition of UICC classification system was adopted for staging malignant tumors and preoperative risk was assessed with anesthesiologist-assigned American Society of Anesthesiologists class. Specific details on collected data and used protocols can be found in Costa et al. [22]. Outcome was categorized as a dichotomized variable (Survivors, S and Non-Survivors, NS).

2.3. Inclusion criteria

All patients >65 years who underwent emergency surgery entered the study. Emergency procedures were defined as unforeseen, non-elective operations according to the NCEPOD Classification of Interventions. All abdominal procedures with ICD-9-CM code numbers ranging from 42.0 to 54.99 were considered eligible. Thoracic procedures (ICD- 9-CM code 32.0–34.99), vascular procedures (ICD-9-CM code 38.0–39.99), gynecological procedures (ICD-9-CM code 55.0–59.99), and urological procedures (ICD-9-CM code 60.0–64.99) were considered eligible for the study if performed by general or emergency surgeon in a general or in a trauma surgery setting. Exclusion criteria were lack of informed consent; patients already hospitalized and scheduled for the same procedure; participation in another trial.

2.4. Features screening

To simulate the same conditions available for the prediction of DFS, only variables taken up to one day before the surgery were retained from the aforementioned database (Fig. 1A). Features with a completion rate lower than 90% were excluded a priori. The features used for outcome predictions were chosen after an univariate screening on SPSS (Vs26, Chicago), by selecting the ones significantly associated with the outcome ($\alpha = 0.05$). Continuous independent variables entered univariate logistic regressions with outcome set to the S/NS groups. When evaluating categorical variables, chi-square tests were applied. Categorical variables were converted in dummy variables before entering ML.

2.5. Model selection

A nested-cross validation approach was implemented [23]. In brief, such approach consists in two k-fold cross-validation loops: an outer loop identifies the test set for each of its folds while the inner loop implements the further split for training and validation. In the outer loop, the dataset was firstly split in train and test set (Fig. 1B). Secondly, inside the train set, hence the inner loop, a k-fold cross-validation was implemented to tune the optimal hyperparameters of the individual models. The number of folds was set to 5. Missing numerical/categorical values in the train and test set were substituted with the mean/mode of the respective train set. Then, the train and test sets' numerical variables were normalized with the train set mean and standard deviation. (Fig. 1B). Subsequently, the training dataset was resampled via the Synthetic Minority Oversampling Technique (SMOTE) [24]. Then, training data entered either Recursive Feature Elimination (RFE, [25]) or a Principal Component Analysis (PCA). Hyperparameter optimization was performed within each train set using *Optuna* [26].

In each optimization trial, hence for each evaluated hyperparameter combination, data were split in the actual train and validation sets implementing the aforementioned inner k-fold cross-validation. The hyperparameter combination maximizing the aggregated k-fold validation accuracy was then chosen for training the final model with all training samples. Such model was then tested with the test outer fold. Such procedure was repeated within all outer folds, test results aggregated, and evaluation metrics calculated. The aforementioned pipeline was repeated with SMOTE + PCA and with SMOTE + RFE. ML models were deployed using Python custom code and the Scikit-Learn [27] library.

2.6. Models and hyper-parameters

Compared models were an Elastic-Net (EN, [28]), a Support Vector Machines (SVM, [29]), a K-nearest neighbors (KNN, [30]), a Decision Tree Classifier (DTC, [31]), and a Multilayer Perceptron (MLP, [32]). EN is a regularized regression which linearly combines the penalties of the Least Absolute Shrinkage and Selection Operator (LASSO) and Ridge regression overcoming respective implementation problems [28]. Hence, Elastic-Net combines LASSO and Ridge modifying the regression parameter estimates as follows:

$$\widehat{\beta} = argmin_{\beta}\left\{\frac{1}{2N_{samples}}||y - X\beta||^2 + \alpha\, l1_{ratio}\, ||\beta||_1 + 0.5\, \alpha(1 - l1_{ratio})\, ||\beta||_2^2\right\}$$

with $l1_{ratio}$ describing the tendency toward a LASSO regularization ($l1_{ratio} \sim 1$) or the Ridge regularization ($l1_{ratio} \sim 0$). Optimized parameters were $l1_{ratio}$ and α.

SVMs result in classifiers based on the margin maximization principle, done via optimization of the box constraint *C*, the kernel function type and the kernel size λ [8]. KNN is an algorithm based on similarity measures between patients sub-cohorts. Optimized parameters are the number of neighbors and the Minkowski distance *p*. Decision trees start from observations about an item (branches of the tree) to infer on the target value (leaves of the tree). At each split of the branches, a node inequality is computed and optimized to maximize information gain while going down the tree. Hyperparameters included the maximum number of features used to compute node equations, the minimal number of samples in a node/leaf, the max depth and a pruning parameter responsible to balance the tree cost-complexity.

2.7. Model explainability

Linear models already allow for interpretability measures, by assigning to each feature the value of its related regression coefficient β_i and calculating the global effect of the feature vector x onto the predictions via the product $\beta \bullet x$. Nevertheless, given the cross-validation implementation, each training will provide a parameter estimate β_N. Accordingly, investigating feature importance by averaging the k coefficients β is possible, but holds two major drawbacks. Firstly, the resulting variability in the parameters estimates can be relevant. Secondly, β estimates are derived from mean trends of the training subset and are not patient-specific. SHAP overcomes these limitations, by determining features contributions to the prediction specifically for the individual subjects, resulting in one value per subject per feature [5,33,34] (Fig. 1C). Lastly, feature importances obtained via SHAP were compared with the test set permutation importances for the best performing model.

3. Results

3.1. Study population

Two thousand five hundred seventy patients were included in the study (50.66% males, median age of 77 years [IQR = 13]) for a total of 238 DFS. Patients had a median weight of 70 Kg [IQR = 20] and a median body-mass index of 25.47 [IQR = 5.04]. Overall, the median time from admission to the surgery was 2 days [IQR = 0] while the post-operative length of stay was found to be 8 days [IQR = 8] making the overall length of stay 10 days long [IQR = 9]. Most DFS were caused by issues with either colon, gallbladder, small bowel or the abdominal wall (86.03% of total surgeries, Fig. 2A). Surgeries resulted in 1715 open, 678 minimally-invasive procedures with a conversion rate of 126 accounting for the 73.5%, 19.7% and 5.0% of all DFSs (Fig. 2B). The entire cohort had a median CACI score of 5 [IQR = 4] with the NS group resulting in a median CACI score of 7 [IQR = 4]. Furthermore, 143, 662, 1383, 339 and 29 patients respectively were in the ASA class 1 through 5.

3.2. Demographics, vital parameters and lab analysis

Univariate logistic regressions showed an association between older age and DFS (Table 1) as well as lower weight and height ($p < 0.05$). Similarly, high levels of glycemia and low levels of hemoglobin were found to associated with DFS ($p < 0.001$). In addition, platelets and white blood cells were higher in the NS group ($p < 0.05$). Also, worse renal function in terms of higher blood urea nitrogen and creatinine levels were associated with DFS ($p < 0.001$).

3.3. Comorbidities, preoperative conditions and clinical scales

The presence of neoplasms and chronic/severe cardiopathy was found to be associated with a worse outcome ($p < 0.001$, Table 2). Myocardial infarction in the previous 6 months and cardiac failure in the previous 30 days resulted fostering the likelihood of DFS. Generally, the prevalence of a cerebrovascular or cardiac disease was significantly higher in the NS group. Also, chronic and acute pulmonary pathologies, cognitive impairments and kidney diseases were associated with DFS.

A laparoscopic approach was found to be strongly decreasing the occurrences of DFSs with respect to other approaches ($p = 0.018$). Higher values on the CACI, ASA, and EmSFI scale were found to be related to DFSs. After dividing the EmSFI score in three previously validated subgroups [35], a significant association was found between being in the 1–3 group and a favorable outcome. Similarly, SIRS, Functional State and 5 items index resulted significantly related with the outcome ($p < 0.001$) as well as consciousness levels, (GCS, Table 3).

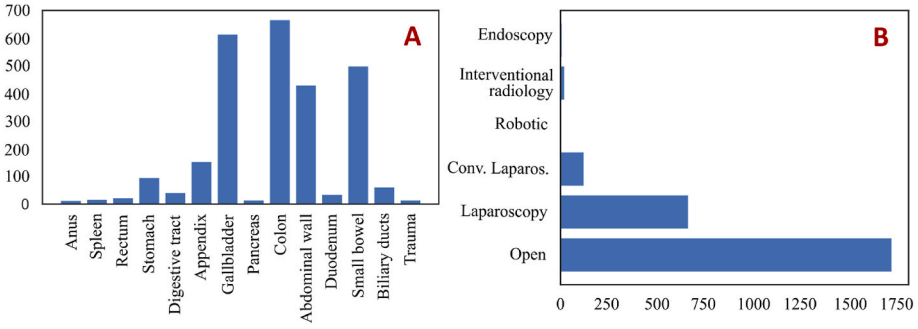

Fig. 2. Main anatomical regions involved within the 2570 primary surgeries (panel A) and related type of approach (panel B). With trauma it is intended all-organ solid injury and abdominal wall and thoracic injury. Also, digestive tract includes patients undergoing digestive tract resection not related to a specific condition such us primary peritonitis, ileocolic resection for intestinal ischemia, perforation due to not recognized condition, foreign body perforation, biliary ileus.

Table 1
Demographics, vital parameters and laboratory analysis descriptive and inferential statistics. Respectively, numerical variables were reported via the median and the interquartile range (in brackets) while categorical ones as count and percentages (in parenthesis). Numerical variables entered a logistic regressions whilst categorical ones a Chi-Square analysis.

	Cohort (n = 2570)	Survivors (n = 2332)	Non-Survivors (n = 238)	Odds Ratio	95% CI	p-value
Demographics						
Age	77 [12]	77 [12]	80 [13]	1.038	1.020–1.056	<0.001
Weight	70 [20]	70 [19]	68 [25]	0.990	0.989–0.9997	0.049
Height	165 [10]	165 [11]	164 [10]	0.982	0.966–0.998	0.033
Age>80	1044 (40.7)	925 (39.7)	119 (50.6)	1.556	1.189–2.036	0.001
Vital Parameters						
Saturation O_2	94 [4]	96 [3]	95 [5]	0.866	0.833–0.901	<0.001
PA_{max}	140 [30]	140 [28]	120 [40]	0.974	0.969–0.980	<0.001
Cardiac freq.	80 [16]	80 [18]	88 [23]	1.021	1.014–1.029	<0.001
Respiratory freq.	16 [4]	16 [4]	16 [4]	1.081	1.036–1.128	<0.001
CPAP	152 (5.9)	118 [5]	34 (14.3)	3.154	2.096–4.739	<0.001
Laboratory analysis						
BUN	33 [30]	32 [28]	48 [37.2]	1.013	1.009–1.016	<0.001
Creatinine	0.97 [0.53]	0.95 [0.49]	1.3 [1.30]	1.569	1.412–1.743	<0.001
Glycemia	125 [57]	124 [54]	136 [78]	1.004	1.002–1.006	<0.001
Haemoglobin	12.8 [3.1]	12.9 [3]	11.8 [3.5]	0.856	0.807–0.907	<0.001
WBC	10.4 [7.4]	10.3 [7.3]	10.9 [7.8]	1.023	1.002–1.044	0.032
INR	1.11 [0.22]	1.1 [0.2]	1.2 [0.36]	1.523	1.258–1.844	<0.001
PLT	240 [118]	238 [116]	253 [172]	1.002	1.001–1.003	<0.001
Na	138 [5]	138 [5]	135 [6]	0.928	0.903–0.954	<0.001

Legend. PA: Arterial Pressure; CPAP: Continuous Positive Airway Pressure; BUN: Blood Urea Nitrogen; WBC: White Blood Cells; INR: International Normalized Ratio; PLT: Platelets.

Table 2
Preoperative comorbidities, organs and type of approach descriptive and inferential statistics. Respectively, numerical variables were reported via the median and the interquartile range (in brackets) while categorical ones as count and percentages (in parenthesis). Numerical variables entered a logistic regressions whilst categorical ones a Chi-Square analysis. Whenever necessary (expected count <5), Fisher exact test was used (indicated with **).

	Cohort (n = 2570)	Survivors (n = 2332)	Non-Survivors (n = 238)	Odds Ratio	95%CI	p-value
Comorbidities						
Malignant neoplasm	707 (27.6)	615 (26.4)	92 (39.1)	1.792	1.357–2.366	<0.001
Chronic Cardiopathy	720 (28.1)	600 (25.8)	120 (51.1)	3.005	2.289–3.945	<0.001
Severe Cardiopathy	139 (5.4)	90 (3.9)	49 (20.9)	6.551	4.486–9.567	<0.001
Moderate Cardiopathy	612 (23.9)	527 (22.6)	85 (36.2)	1.937	1.459–2.571	<0.001
Myocardial infarction (6 m)	72 (2.8)	51 (2.2)	21 (8.8)	4.362	2.576–7.389	<0.001
Cardiac failure [30d]	117 (4.5)	78 (3.3)	39 (16.4)	5.709	3.786–8.608	<0.001
Respiratory failure [30d]	179 (6.9)	141 (6.0)	38 (15.9)	2.977	2.023–4.381	<0.001
Cerebrovascular disease	369 (14.3)	304 (12.9)	65 (27.3)	2.529	1.856–3.446	<0.001
Cardiac disease	727 (28.1)	605 (25.7)	122 (51.3)	3.033	2.315–3.976	<0.001
Peripheral Vasculopathy	395 (15.3)	338 (14.4)	57 (23.9)	1.875	1.362–2.580	<0.001
Percutaneous Cardiac Stent	284 [11]	233 (9.9)	51 (21.4)	2.478	1.767–3.474	<0.001
Kidney diseases	286 (11.1)	227 (9.7)	59 (24.8)	3.083	2.228–4.264	<0.001
Liver diseases	76 (2.9)	52 (2.2)	24 (10.1)	4.978	3.008–8.240	<0.001
Connective tissues pathologies	75 (2.9)	63 (2.7)	12 [5]	1.928	1.024–3.627	0.045
Chronic pulmonary pathologies	458 (17.7)	393 (16.7)	65 (27.3)	1.871	1.379–2.538	<0.001
Acute pulmonary pathologies	68 (2.6)	49 (2.1)	19 (8.0)	4.074	2.356–7.044	<0.001
Cognitive impairment	279 (10.8)	221 (9.4)	58 (24.4)	3.104	2.239–4.304	<0.001
Reduced ADL	724 (28.4)	639 (27.6)	85 (36.2)	1.495	1.121–1.968	0.005
ACO	582 (22.5)	505 (21.5)	77 (32.4)	1.747	1.309–2.332	<0.001
Steroid-immunosuppressants drugs	145 (5.6)	119 (5.1)	26 (10.9)	2.299	1.470–3.595	0.001
CCH/PCI	283 (11.0)	232 (20.0)	51 (21.7)	2.504	1.785–3.513	<0.001
Solid Metastatic Tumor	173 (6.7)	137 (5.9)	36 (15.1)	2.893	1.950–4.293	<0.001
Organ						
Stomach	99 (3.8)	80 (3.4)	19 (8.0)	2.462	1.465–4.137	0.001
Digestive tract	47 (1.8)	24 (1.0)	23 (9.7)	10.368	5.754–18.680	<0.001
Gall bladder	619 (23.9)	589 (25.3)	27 (11.3)	0.380	0.252–0.573	<0.001
Colon	668 (25.8)	568 (24.4)	93 (39.1)	1.980	1.501–2.611	<0.001
Abdominal wall	432 (16.7)	411 (17.6)	19 (8.0)	0.407	0.252–0.658	<0.001
Appendix	154 (6.0)	152 (6.5)	2 (0.8)	0.123	0.030–0.498	<0.001**
Duodenum	37 (1.4)	26 (1.1)	11 (4.6)	4.331	2.113–8.881	<0.001
Type of Approach						
Open	1715 (66.4)	1545 (65.9)	175 (73.5)	1.276	0.952–1.710	0.066
Laparoscopic	678 (26.3)	631 (26.9)	47 (19.7)	0.661	0.473–0.924	0.018
Converted laparoscopic	126 (4.6)	114 (4.9)	12 (5.0)	1.071	0.581–1.973	0.891
Robotic	2 (0.1)	1 (0.0004)	1 (0.4)	9.911	0.618–158.969	0.175**
Interventional Radiology	22 (0.9)	21 (0.9)	1 (0.4)	0.468	0.063–3.494	0.387**
Endoscopy	9 (0.3)	7 (0.3)	2 (0.8)	2.837	0.586–13.732	0.197**

Legend. ADL: Activities of Daily Living; CCH: Chronic Continuous Hypoxia; PCI: Percutaneous Coronary Intervention; ACO: Asthma and COPD Overlap.

Table 3
Clinical scores and scales descriptive and inferential statistics. Respectively, numerical variables were reported via the median and the interquartile range (in brackets) while categorical ones as count and percentages (in parenthesis). Numerical variables entered a logistic regressions whilst categorical ones a Chi-Square analysis.

	Cohort (n = 2570)	Survivors (n = 2332)	Non-Survivors (n = 238)	Odds Ratio	95% CI	p-value
Scores						
CACI	5 [4]	5 [3]	7 [4]	1.273	1.215–1.334	<0.001
ASA	3 [1]	3 [1]	4 [1]	4.639	3.744–5.749	<0.001
Clavien Surgery	0 [0]	0 [0]	0 [1]	1.763	1.570–1.980	<0.001
EmSFI	3 [2]	3 [2]	5 [3]	1.540	1.438–1.649	<0.001
EmSFI subgroups						<0.001
1–3	1619 (63.2)	1543 (66.3)	76 (32.3)	0.245	0.184–0.326	<0.001
4–7	915 (35.7)	770 (33.1)	145 (61.7)	3.199	2.432–4.209	<0.001
8–14	29 (1.1)	15 (0.6)	14 [6]	9.729	4.637–20.415	<0.001
SIRS	783 (30.6)	672 (28.9)	111 (47.2)	2.206	1.682–2.894	<0.001
Funct. State 5-items	1234 (48.1)	1078 (46.3)	156 (66.4)	2.290	1.726–3.038	<0.001
Funct. State	647 (25.2)	540 (23.2)	107 (45.5)	2.768	2.104–3.641	<0.001
GCS	15 [0]	15 [0]	15 [3]	0.776	0.738–0.817	<0.001

Legend. CACI: Charlson Age Comorbidity Index; ASA: American Society of Anesthesiologists score; SIRS: Systemic Inflammatory Response Syndrome; GCS: Glasgow Coma Scale.

3.4. ML models

Best results were obtained with the SMOTE-RFE pipeline compared to the SMOTE-PCA pipeline (Suppl. 1). Validation errors for the EN, SVM, KNN, MLP, DTC models resulted equal to 81.3% (s.d. = 1.3), 82.0% (s.d. = 1.2), 87.4% (s.d. = 0.5), 97.3% (s.d. = 0.02) and 81.3% (s.d. = 0.13). Specifically, the best performing model (MLP) resulted in a test accuracy of 94.9% (sensitivity = 92.0%, specificity = 95.2%, Fig. 3, Suppl. 2).

Embedding the best performing solution with SHAP showed that the most influent variables were found to be related to cardiac comorbidities, in particular, the presence of an acute cardiac disease (Fig. 4). On

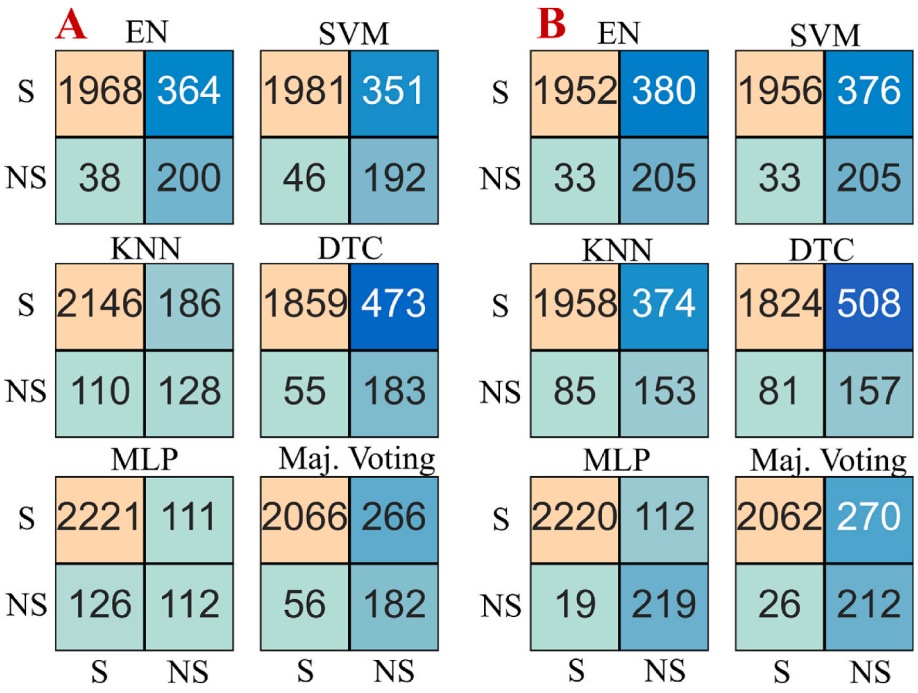

Fig. 3. Confusion matrix for aggregated test results across outer folds. Panel A refers to models with only SMOTE while Panel B includes the addition of RFE (with SMOTE kept on).

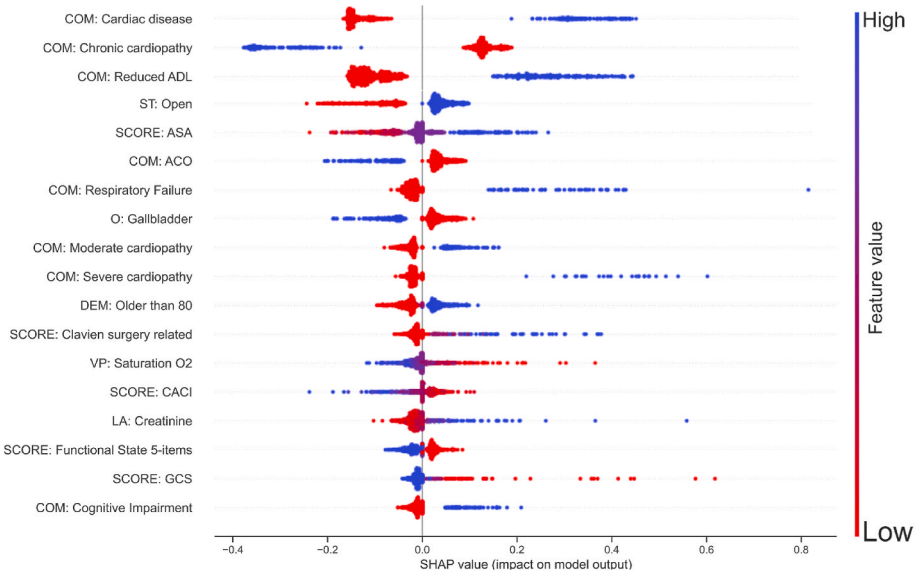

Fig. 4. Contributions of independent variables (y-axis) to the model prediction (x-axis, SHAP value) related to the value of the independent variable coded with colors from low (red) to high (blue). Model predictions are set to 1 for the NS group and to 0 for the S group. Each dot represents one individual patient. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

the other hand, being diagnosed with a prior chronic cardiopathy did not increase the risk of a DFS. Furthermore, severe cardiopathy almost doubled the contribution onto the predictions with respect to a moderate one. Reduced capability of performing Activities of Daily Living was also found to be important predictor of mortality. The latter, when present, increased the likelihood of a DFS more than it decreased it, when absent. Also, the mortality reduction carried by minimally-invasive approaches resulted stronger than the increase of likelihood of DFS after an open surgery. SHAP highlighted how respiratory failure and worse levels on clinical scores contributed to DFS. High creatinine levels increased the chance of DFS, while regular creatinine levels did not influence the model prediction. Among vital parameters, the only factor appearing after the SHAP analysis is the O_2 saturation, with lower levels acting positively on mortality predictions. As expected, low GCS values and presence of cognitive impairments contributed to mortality risk, whilst high GCS values did not impact model predictions.

Comparing SHAP explanations with permutation importance showed how prevalence of cardiac diseases and/or cardiopathies are the predictors most influencing the test accuracy (Fig. 5). Similarly, coma levels (GCS), surgery complexity (Clavien) and physical status were found to be influencing the classification accuracy. Lastly, the type of approach, in particular an open one, the occurrence of a respiratory failure within the previous 30 days and reduced ADLs also influenced the test accuracy.

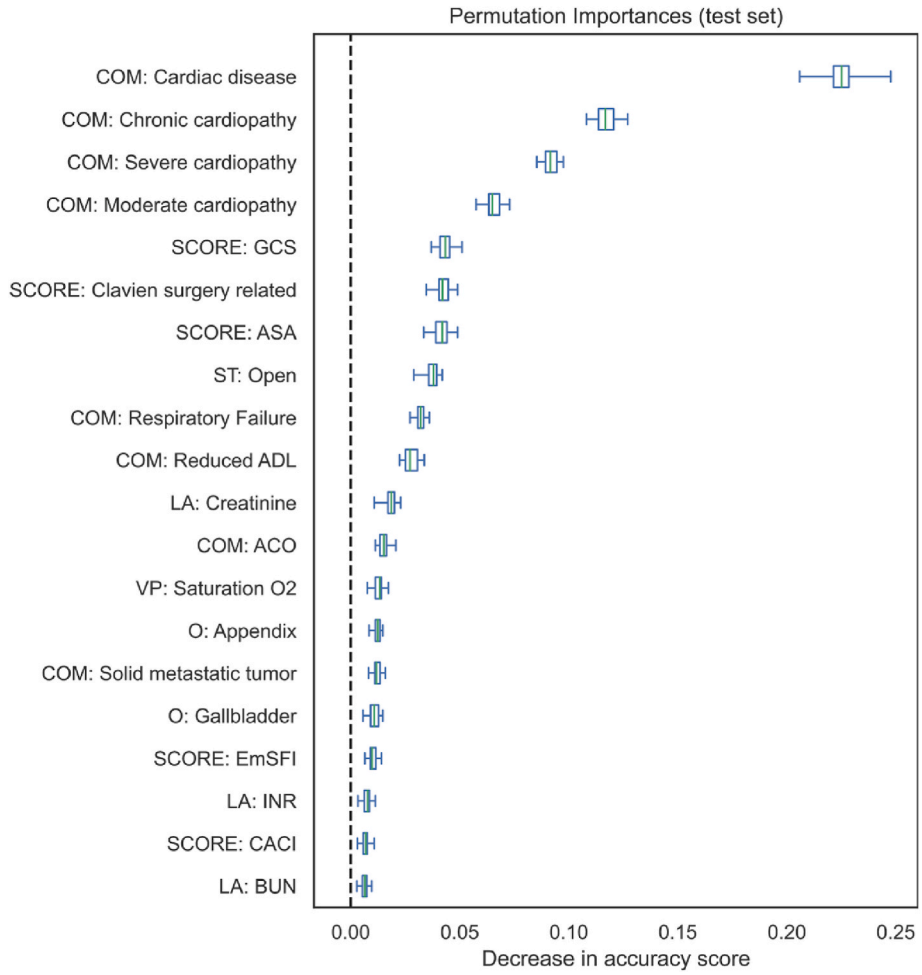

Fig. 5. Decrease in test accuracy given by permutation importance of the best performing solution. Permuted independent variables are plotted on the y-axis and accuracy decrease on the x-axis. Each independent variable is represented with a box-plot indicating the distribution of the accuracy decrease across the number of times that the related feature is permuted (N = 30).

4. Discussions

Prediction is not new to medicine. From risk scores to guide anticoagulation (CHADS2) and the use of cholesterol medications (ASCVD) to risk stratification of patients in the intensive care unit (APACHE), data-driven predictions are routine in medical practice. However, current risk stratification tools provide a somewhat cross-sectional view of patient disease. ML combined with clinical data sources enable us to rapidly generate prediction models for many clinical questions [36,37] performing on par with human physicians [38]. Moreover, ML learns and improve its own performance with time as the model is exposed to more information [39]. Furthermore, even if tackling hospital readmission instead than mortality prediction, it has been showed how such models, inserted in a proper simulation framework, can provide a crucial reduction of the costs by preventing such readmissions. In the same optic, by precisely assessing the patients most likely to die after surgery, preventive measures can be adopted thus reducing such fatalities.

Previous studies on postoperative complications of emergency surgery have mainly used scoring for identifying patients more likely to have complications after surgery [1]. However, these methods are often not sensitive enough for clinical application. Therefore, the CDSTs' potential in identifying risk factors and predicting the surgery complications is worth investigation. To this point, the current study was important because it proposed and internally cross-validated a ML algorithm to predict the risk of DFS based on a large multi-centre, prospectively collected database focused on a population of elderly patient.

Among ML-based surgical risk predictors the Physiologic and Operative Severity Score for the Enumeration of Mortality and Morbidity estimates AEs by using admission and discharge variables [10]. For the purpose of this study, only readily available pre-operative data were included, allowing faster stratification of surgical risk without including second-level examinations (e.g. imaging). Furthermore, knowing already in the preoperative period, which condition bolsters the risk of a DFS, contributes to the discussion on risk-limiting procedures. Such problem has mainly been tackled on specific types of surgeries, however, the dataset used for training our ML algorithms was more complex than others, specifically concerning the type and location of surgeries analysed, compared to illness-specific ML models (e.g. aneurysms [3,9], thoracotomies [13] and open-heart surgeries [40]). Zhong et al., targeting 30-days mortality with a dataset of ~6000 patients reached a sensitivity of 58% and a precision of 58% after only open-heart surgeries. On a dataset of 310 patients undergoing aortic aneurysm repair and a cross-validated MLP a sensitivity-specificity couple of 97%/65.2% was reached in predicting mortality. Similarly, a Naïve Bayes model has been developed for patients undergoing elective surgeries [41] reaching a mortality prediction accuracy of 78%.

To the best of our knowledge, there are no study that targeted the prediction of DFS in elderly patients in an emergency setting. In our database all the procedures carried out under an acute care setting for any disease were included, resulting in multiple combinations of different variables related to the specific disease (e.g. acute cholecystitis, bowel perforation).

Furthermore, the adopted nested cross-validation strategy is the only approach which allows to jointly use a dataset for training and validation of the model, retaining each patient in the test set once. Such procedure tests the model generalization capability, simulating the final translation to clinical practice. Furthermore, by analysing inner loop validation errors (Suppl. 1), different validation sets led to similar validation accuracies, showing robustness of the models to hyperparameter changes. Such extensive optimization allows to find globally optimal, non-overfitting parameters without overfitting the data, visible in the negligible difference between test and validation errors. Lastly, the obtained sensitivity and specificity notably outperform the ones found in literature for what concerns absolute numbers and robustness of validation strategies. In particular, Huang et al. obtained a k-fold cross-validation accuracy of 72.5%, sensitivity of 76% and specificity of 66.6% [18], whilst our best performing model (MLP) resulted in a test accuracy of 94.9% (validation 97.2%, std. = 0.002), sensitivity of 99.2% and specificity of 66.2%. Tedesco et al. obtained with a resampled RF and a train-test split validation strategy a sensitivity of 87.5% and a PPV of 94.7% compared to our PPV of 95.2% [42].

We acknowledge that opaque models as the MLP do not offer insights on the functioning of the black box. An assessment of the importance of each features can be derived by repeatedly and randomly shuffling features across features and evaluating the decrease in accuracy of the model. Such techniques disentangles the relationship between the features and the target, therefore assigning to the model's drop in performance a proxy of the feature importance. Furthermore, this technique can be applied independently on the model type and intrinsic structure, making it easily reproducible and translatable. Also, differently from impurity-based feature importance algorithms (e.g. Gini split index), permutation importance is not biased toward high cardinality features and it is not computed based on intrinsic parameters of the "trained model" itself but on the accuracy computed on an external test set, allowing the computation of feature importance to be done on unseen data. Nevertheless, a disadvantage of such technique is that it assigns importance based on the entire test cohort and it is not patient specific. Thus, we embedded MLP with SHAP explanations. Such technique has already been adopted in the context of general in-hospital mortality [43] and 3-year all-cause mortality in patients with heart failure [44]. The latter, showed, similarly to permutation importance, how cardiac-related comorbidities and reduced ADLs are the features impressing the strongest weight. Also, for example, SHAP highlights how low values of GCS, high values of creatinine and low saturation increase the risk of a DFS, but the opposite does not decrease it. Such directional information could not be retrieved from cohort-based feature importance as Gini index, regression coefficients and permutation importance, since these latter only offer an absolute estimate of the importance of the variable. Besides, SHAP values are computed benchmarking models by variating patient features, making the features contribution to the prediction aware of the data sub-space that the specific patients relies onto. The latter allows clinicians to inspect why a prediction was made, evaluate whether in their opinion such prediction is unbalanced toward a specific clinical issue or is ignoring a crucial medical condition. In this way, the model not only provides a binary answer to a complex question but offers indications on factors that may increase the risks, fostering trust of the operators in using AI-enabled solutions.

5. Conclusions

Clearly, ML is a valuable and increasingly necessary tool for the modern health care system. The ML models developed in this study are robustly validated and result in state-of-the-art performances with the advantage to be easily interpretable from clinicians relying on pre-operative data only. Precisely quantifying the risk of DFS may better inform patient-centred decision-making. It would also direct targeted quality improvement interventions while supporting activities of accountable care organizations that rely on accurate estimates of population risk. Further work will focus on simulating the effectiveness and the efficacy of the improvement brought by the model predictions on the clinical decision process and thus on hospital costs.

Provenance and peer review

Not commissioned, externally peer-reviewed.

Please state whether ethical approval was given, by whom and the relevant Judgement's reference number

Patients were recruited following the Helsinki Declaration and were enrolled after signing a written consent approved by the Ethical Committee of "Sapienza" University of Rome, Italy (approval number 4252_2016). The FRAILESEL study protocol has been register on Clinicaltrials.gov (ClinicalTrials.gov identifier: NCT02825082). The FRAILESEL (Frailty and Emergency Surgery in the Elderly) study (ClinicalTrials.gov identifier: NCT02825082) recruited patients following the Helsinki Declaration after signing a written consent approved by the Ethical Committee of "Sapienza" University of Rome, Italy (No. 4252_2016, ClinicalTrials.gov: NCT02825082, https://clinicaltrials.gov/ct2/show/NCT02825082).

Please state any sources of funding for your research

The study was supported by the "Ricerca Corrente RC2021-RC2022 programs", the 5×Mille funds AF2018: "Data Science in Rehabilitation Medicine" and the 5×Mille funds AF2019: "Study and development of biomedical data science and machine learning methods to support the appropriateness and the decision-making process in rehabilitation medicine" by the Italian Ministry of Health.

Research registration Unique Identifying number (UIN)

1. Name of the registry: Clinical trials.gov.
2. Unique Identifying number or registration ID: NCT02825082.
3. Hyperlink to your specific registration (must be publicly accessible and will be checked): https://clinicaltrials.gov/ct2/show/NCT02825082.

Guarantor

Piergiuseppe Liuzzi and Pietro Fransvea.

Data statement

Data and code used to reproduce results on the paper can be made available upon request from the corresponding author for replication purposes.

CRediT authorship contribution statement

Pietro Fransvea: Conceptualization. **Giulia Fransvea:** Conceptualization, Investigation, Methodology, Formal analysis, Visualization. **Piergiuseppe Liuzzi:** Methodology, Software, Formal analysis, Visualization. **Gabriele Sganga:** Data curation, Investigation. **Andrea Mannini:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **Gianluca Costa:** Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The study was supported by the "Ricerca Corrente RC2021 program", the 5x1000 funds AF2018: "Data Science in Rehabilitation Medicine" and the 5x1000 funds AF2019: "Study and development of biomedical data science and machine learning methods to support the appropriateness and the decision-making process in rehabilitation medicine" by the Italian Ministry of Health.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijsu.2022.106954.

References

[1] G. Eamer, M.J.H. Al-Amoodi, J. Holroyd-Leduc, D.B. Rolfson, L.M. Warkentin, R. G. Khadaroo, Review of risk assessment tools to predict morbidity and mortality in elderly surgical patients, Am. J. Surg. 216 (3) (2018 Sep 1) 585–594.

[2] J. Allyn, N. Allou, P. Augustin, I. Philip, O. Martinet, M. Belghiti, et al., A comparison of a machine learning model with EuroSCORE II in predicting mortality after elective cardiac surgery: a decision curve analysis, PLoS One 12 (1) (2017 Jan 6), e0169772.

[3] Ostberg NP, Zafar MA, Mukherjee SK, Ziganshin BA, Elefteriades JA. A machine learning approach for predicting complications in descending and thoracoabdominal aortic aneurysms. J. Thorac. Cardiovasc. Surg. [Internet]. 2022 Jan 11 [cited 2022 Apr 5]; Available from: https://www.sciencedirect.com/science/article/pii/S0022522322000046.

[4] A. Meyer, D. Zverinski, B. Pfahringer, J. Kempfert, T. Kuehne, S.H. Sündermann, et al., Machine learning for real-time prediction of complications in critical care: a retrospective study, Lancet Respir. Med. 6 (12) (2018 Dec 1) 905–914.

[5] S.M. Lundberg, B. Nair, M.S. Vavilala, M. Horibe, M.J. Eisses, T. Adams, et al., Explainable machine-learning predictions for the prevention of hypoxaemia during surgery, Nat. Biomed. Eng. 2 (10) (2018 Oct) 749–760.

[6] R. Raj, T. Luostarinen, E. Pursiainen, J.P. Posti, R.S.K. Takala, S. Bendel, et al., Machine learning-based dynamic mortality prediction after traumatic brain injury, Sci. Rep. 9 (1) (2019 Nov 27), 17672.

[7] G. Zhang, J. Xu, M. Yu, J. Yuan, F. Chen, A machine learning approach for mortality prediction only using non-invasive parameters, Med. Biol. Eng. Comput. 58 (10) (2020 Oct 1) 2195–2238.

[8] S. Rose, Mortality risk score prediction in an elderly population using machine learning, Am. J. Epidemiol. 177 (5) (2013 Mar 1) 443–452.

[9] A. Monsalve-Torra, D. Ruiz-Fernandez, O. Marin-Alonso, A. Soriano-Payá, J. Camacho-Mackenzie, M. Carreño-Jaimes, Using machine learning methods for predicting inhospital mortality in patients undergoing open repair of abdominal aortic aneurysm, J. Biomed. Inf. 62 (2016 Aug 1) 195–201.

[10] G. Wang, K.M. Lam, Z. Deng, K.S. Choi, Prediction of mortality after radical cystectomy for bladder cancer by machine learning techniques, Comput. Biol. Med. 63 (2015 Aug 1) 124–132.

[11] Y. Li, M. Chen, H. Lv, P. Yin, L. Zhang, P. Tang, A novel machine-learning algorithm for predicting mortality risk after hip fracture surgery, Injury 52 (6) (2021 Jun 1) 1487–1493.

[12] D.F. Hernandez-Suarez, S. Ranka, Y. Kim, A. Latib, J. Wiley, A. Lopez-Candales, et al., Machine-learning-based in-hospital mortality prediction for transcatheter mitral valve repair in the United States, Cardiovasc. Revascularization Med. 22 (2021 Jan 1) 22–28.

[13] U. Benedetto, A. Dimagli, S. Sinha, L. Cocomello, B. Gibbison, M. Caputo, et al., Machine learning improves mortality risk prediction after cardiac surgery: systematic review and meta-analysis [Internet], J. Thorac. Cardiovasc. Surg. (2020 Aug 10) [cited 2022 Apr 5]; Available from: https://www.sciencedirect.com/science/article/pii/S0022522320323576.

[14] J.N. Cooper, P.C. Minneci, K.J. Deans, Postoperative neonatal mortality prediction using superlearning, J. Surg. Res. 221 (2018 Jan 1) 311–319.

[15] Y. Hu, X. Gong, L. Shu, X. Zeng, H. Duan, Q. Luo, et al., Understanding risk factors for postoperative mortality in neonates based on explainable machine learning technology, J. Pediatr. Surg. 56 (12) (2021 Dec 1) 2165–2171.

[16] Liu X, Hu P, Mao Z, Kuo PC, Li P, Liu C, et al. Interpretable machine learning model for early prediction of mortality in elderly patients with multiple organ dysfunction Syndrome (MODS): Multicenter Retrospect. Study Cross Validat. arXiv:200110977 [physics, stat] [Internet]. 2020 Jan 28 [cited 2022 Apr 5]; Available from: http://arxiv.org/abs/2001.10977.

[17] B. Yenidogan, S. Pathak, J. Geerdink, J.H. Hegeman, M. van Keulen, Multimodal machine learning for 30-days post-operative mortality prediction of elderly hip fracture patients, in: 2021 International Conference on Data Mining Workshops (ICDMW), 2021, pp. 508–516.

[18] Y.C. Huang, S.J. Li, M. Chen, T.S. Lee, Y.N. Chien, Machine-learning techniques for feature selection and prediction of mortality in elderly CABG patients, Healthcare 9 (5) (2021 May) 547.

[19] C.K. Lee, I. Hofer, E. Gabel, P. Baldi, M. Cannesson, Development and validation of a deep neural network model for prediction of postoperative in-hospital mortality, Anesthesiology 129 (4) (2018 Oct) 649–662.

[20] V.V. Mišić, E. Gabel, I. Hofer, K. Rajaram, A. Mahajan, Machine learning prediction of postoperative emergency department hospital readmission, Anesthesiology 132 (5) (2020 May) 968–980.

[21] G. Mathew, R. Agha, STROCSS Group, STROCSS 2021: strengthening the reporting of cohort, cross-sectional and case-control studies in surgery, Int. J. Surg. 96 (2021 Dec), 106165.

[22] G. Costa, G. Massa, ERASO (Elderly Risk Assessment for Surgical Outcome) Collaborative Study Group. Frailty and emergency surgery in the elderly: protocol of a prospective, multicenter study in Italy for evaluating perioperative outcome (The FRAILESEL Study), Updates Surg. 70 (1) (2018 Mar) 97–104.

[23] G.C. Cawley, N.L.C. Talbot, On over-fitting in model selection and subsequent selection bias in performance evaluation, J. Mach. Learn. Res. 11 (2010) 2079–2107.

[24] N. Chawla, K. Bowyer, L. Hall, W. Kegelmeyer, SMOTE: synthetic minority oversampling technique, J. Artif. Intell. Res. 16 (2002 Jun) 321–357.

[25] H. Yoon, K. Yang, C. Shahabi, Feature subset selection and feature ranking for multivariate time series, IEEE Trans. Knowl. Data Eng. 17 (9) (2005 Sep) 1186–1198.

[26] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: a next-generation hyperparameter optimization framework, arXiv:190710902 [cs, stat] [Internet]. 2019 Jul 25 [cited 2021 Jan 12]; Available from: http://arxiv.org/abs/1907.10902.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[28] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. Roy. Stat. Soc. Ser. B 67 (2005) 301–320.

[29] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995 Sep) 273–297.

[30] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Trans. Inf. Theor. 13 (1967) 21–27.

[31] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, Chapman and Hall, FL, 1984.

[32] F. Rosenblatt, The perceptron: a probabilistic model for information storage and organization in the brain, Psychol. Rev. 65 (6) (1958) 386–408.

[33] S.M. Lundberg, G. Erion, H. Chen, A. DeGrave, J.M. Prutkin, B. Nair, et al., From local explanations to global understanding with explainable AI for trees, Nat. Mach. Intell. 2 (1) (2020 Jan) 56–67.

[34] Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. Explainable AI for trees: From local explanations to global understanding. arXiv [Internet]. 2019;2(January). Available from: https://doi.org/10.1038/s42256-019-0138-9.

[35] P. Fransvea, V. Fico, V. Cozza, G. Costa, L. Lepre, P. Mercantini, et al., Clinical-pathological features and treatment of acute appendicitis in the very elderly: an interim analysis of the FRAILESEL Italian multicentre prospective study [Internet], Eur. J. Trauma Emerg. Surg. (2021 Mar 18), https://doi.org/10.1007/s00068-021-01645-9 [cited 2022 Apr 5]; Available from:.

[36] F. Shamout, T. Zhu, D.A. Clifton, Machine learning for clinical outcome prediction, IEEE Rev. Biomed. Eng. 14 (2021) 116–126.

[37] D.A. Hashimoto, G. Rosman, D. Rus, O.R. Meireles, Artificial intelligence in surgery: promises and perils, Ann. Surg. 268 (1) (2018 Jul) 70–76.

[38] C. Combi, Editorial from the new Editor-in-Chief: artificial intelligence in medicine and the forthcoming challenges, Artif. Intell. Med. 76 (2017 Feb) 37–39.

[39] V. Gulshan, L. Peng, M. Coram, M.C. Stumpe, D. Wu, A. Narayanaswamy, et al., Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs, JAMA 316 (22) (2016 Dec 13) 2402–2410.

[40] Z. Zhong, X. Yuan, S. Liu, Y. Yang, F. Liu, Machine learning prediction models for prognosis of critically ill patients after open-heart surgery, Sci. Rep. 11 (1) (2021 Dec) 3384.

[41] A.P. Ehlers, S.B. Roy, S. Khor, P. Mandagani, M. Maria, R. Alfonso-Cristancho, et al., Improved risk prediction following surgery using machine learning algorithms, EGEMS (Wash DC) 5 (2) (2017 Apr 20) 3.

[42] S. Tedesco, M. Andrulli, M.Å. Larsson, D. Kelly, A. Alamäki, S. Timmons, et al., Comparison of machine learning techniques for mortality prediction in a prospective cohort of older adults, Int. J. Environ. Res. Publ. Health 18 (23) (2021 Jan), 12806.

[43] E. Stenwig, G. Salvi, P.S. Rossi, N.K. Skjærvold, Comparative analysis of explainable machine learning prediction models for hospital mortality, BMC Med. Res. Methodol. 22 (1) (2022 Dec) 53.

[44] K. Wang, J. Tian, C. Zheng, H. Yang, J. Ren, Y. Liu, et al., Interpretable prediction of 3-year all-cause mortality in patients with heart failure caused by coronary heart disease based on machine learning and SHAP, Comput. Biol. Med. 137 (2021 Oct 1), 104813.