scientific reports

Check for updates

OPEN

A machine learning model for post-operative sepsis prediction in acute surgical patients: a multi-centre, prospective study

Pietro Fransvea[1,5], Piergiuseppe Liuzzi[2]✉, Gianluca Costa[3], Maurizio Sanguinetti[4] & Gabriele Sganga[1]

The aim of this study was to develop and validate a machine learning (ML) solution for predicting sepsis among elderly surgical patients undergoing emergency procedure. Data from over 150 variables were collected across multiple domains. ML models were cross-validated using a nested-cross validation approach. The performance of individual models was assessed based on accuracy, sensitivity, and specificity. A total of 29 medical centres participated in the study, ensuring a diverse and representative sample. The study included patients undergoing emergency procedures across various surgical specialties, encompassing abdominal, thoracic, vascular, gynaecological, and urological surgeries, performed by both general and emergency surgeons in general or trauma surgery settings. Among 2571 enrolled patients, 119 were identified as having sepsis. The Random Forest model demonstrated the highest accuracy of 96.2%, with notable sensitivity and specificity. An ensemble model further improved performance, achieving an accuracy of 96.06%. ML models, show promise in accurately predicting sepsis among elderly surgical patients in emergency settings. These findings underscore the potential of ML in enhancing risk stratification and informing clinical decision-making to improve patient outcomes and post-surgery rehabilitation. Further research and validation studies are warranted to evaluate the real-world applicability and integration of these predictive models into clinical practice.

Sepsis, defined as a life-threatening organ dysfunction caused by a dysregulated host response to infection, is a major cause of death worldwide. A meta-analysis estimated about 31.5 million sepsis and 19.4 million severe sepsis cases occur each year, contributing to 5.3 million deaths[1]. Moreover, the burden of infections in acute care surgery (ACS) is huge. Surgical emergencies alone account for three million admissions per year in the United States (US) with estimated financial costs of USD 28 billion per year. Acute care facilities and ACS patients represent boost sanctuaries for the emergence, development and transmission of infections and multi-resistant organisms. A pivotal study by Moore LJ et al. has identified that the need for emergency surgery is a risk factor for both the development of and death from sepsis and septic shock[2]. In this light, the development of good predictive models of sepsis is pivotal for early identification of patients at higher risk of (or likely to already have) sepsis in order to improve the outcomes. Nonetheless, to date, there is a paucity of epidemiologic data specifically addressing sepsis in the surgical patient. Vogel et al. published a large retrospective evaluation of postoperative sepsis in the United States. Unfortunately, this study was also based on ICD-9 discharge data and only took into consideration elective surgical cases[3]. In emergency settings, it is essential to apply a structured decision-making process and carry out accurate risk stratification to correctly prioritize patients. To this extent, medicine has recently witnessed the emergence of Machine Learning (ML) as a novel tool to analyse 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) and sepsis. The latter is crucial for improving shared decision-making 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'

[1]Fondazione Policlinico Universitario A. Gemelli, IRCCS - Università Cattolica del Sacro Cuore, Rome, Italy. [2]IRCCS Fondazione Don Carlo Gnocchi ONLUS, Firenze Fi, Italy. [3]UOC Chirurgia Colo-rettale, Fondazione Policlinico Universitario Campus Bio-Medico - Link Campus University, Rome, Italy. [4]Fondazione Policlinico Universitario A. Gemelli, UOC Microbiologia, Rome, Italy. [5]UOC Chirurgia d'urgenza e del Trauma, Policlinico Universitario Agostino Gemelli IRCCS, Rome, Italy. ✉email: pliuzzi@dongnocchi.it

characteristics and outcomes under different clinical conditions. To this extent, studies for sepsis prediction using ML are developing rapidly, including the intensive care unit, the emergency department, and general hospital wards. These approaches range from classical algorithms (e.g., regularized logistic regression, random forests, gradient-boosted trees) to deep learning models trained on high-frequency electronic health record data, and in many cases have demonstrated superior discrimination compared with traditional clinical scores such as CURB-65, as well as improved calibration and net benefit in decision-analytic frameworks. In parallel, machine learning has also been applied to surgical populations, for example to predict postoperative sepsis and other complications after specific procedures or in selected cohorts, and to inform perioperative risk stratification and shared decision-making.

However, most existing models focus on general adult or ICU populations, on elective surgery, or on narrow diagnostic groups, and frequently rely on data collected intra-operatively or post-operatively. Evidence specifically addressing elderly patients undergoing emergency surgery across a broad spectrum of procedures remains scarce, despite the fact that this group carries a disproportionate burden of sepsis, frailty and multimorbidity. To our knowledge, no previous study has developed and prospectively evaluated a machine learning model for the prediction of postoperative sepsis in a multi-centre cohort of elderly patients in the acute surgical setting using only routinely available preoperative information.

In this context, we adopted the qSOFA score to define sepsis, as it represents a pragmatic and clinically applicable tool in emergency surgery: it relies exclusively on bedside variables (respiratory rate, mental status, systolic blood pressure), allowing rapid risk stratification even when laboratory data are not yet available and urgent decisions about operative management are required. Although qSOFA is less sensitive than other scoring systems for sepsis screening, it remains a validated prognostic marker for mortality and adverse outcomes in patients with suspected infection outside the ICU and has been used as a bedside indicator of sepsis severity and prognosis in surgical and trauma populations. Within the acute surgical setting, qSOFA therefore offers a standardized, reproducible, and time-efficient method to identify patients at higher risk of deterioration. Addressing this unmet need by developing and validating a preoperative ML model for postoperative sepsis based on such routinely available information is crucial to enable early risk stratification, tailored perioperative planning, a more favourable rehabilitation pathway, and more efficient allocation of limited resources. Nevertheless, no study targeted yet the development of ML models targeting the prediction of sepsis after all surgical procedures in elderly patients in emergency setting. On this path, we performed a study based on a prospectively collected, large multi-centre cohort with the aim to develop and cross-validate ML solutions capable of predicting sepsis by using data collected up to one day before surgery.

Results

Population

Variables from four different domains were collected (i.e., demographic data, clinical and functional evaluations, vital signs, and comorbities/complications, Fig. 1). A full list of the included variables is reported in Supplementary Materials A. Of the 63 included variables, 62 were retained for ML analysis after the missing data analysis with

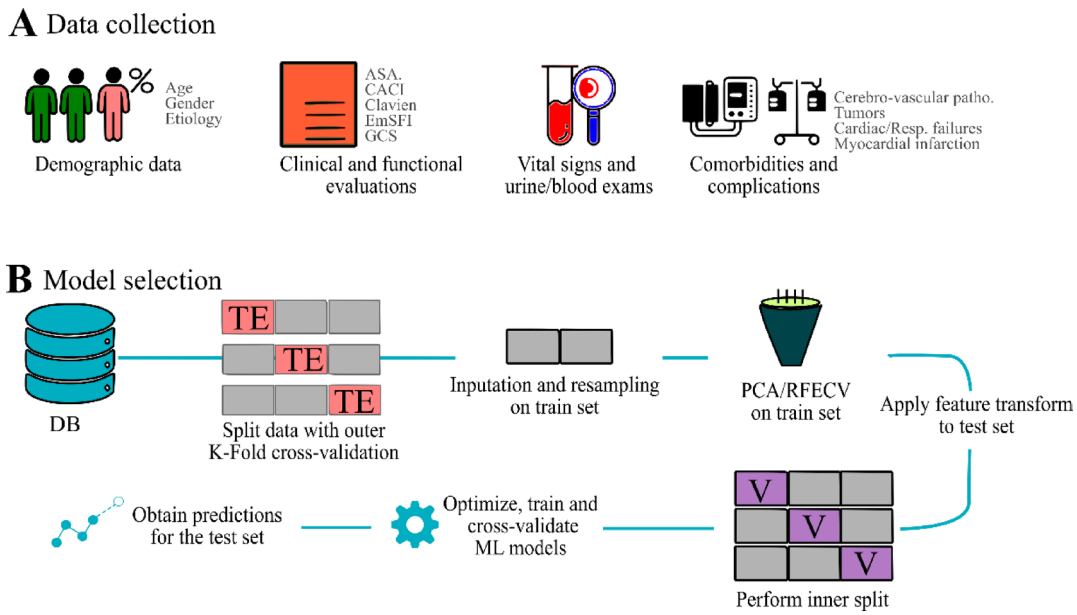

Fig. 1. Study pipeline from data collection (**A**) to model development. Data enters first an outer k-fold cross-validation split where the test folds are defined (**B**). Pre-processing steps (SMOTE, PCA, RFECV) are derived from the train set and applied to both train and test sets. Train sets are then provided to an inner cross-validation split, used to optimize models hyperparameters. Model predictions and their explanations are compared, together with the computation of the *mode* model.

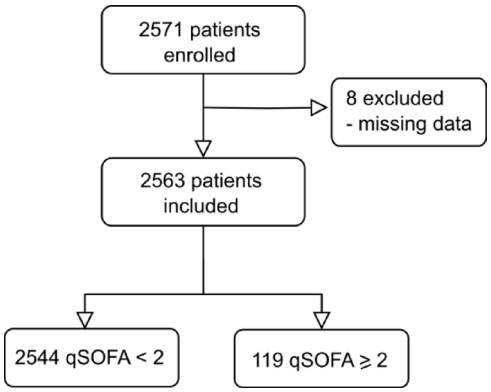

Fig. 2. Study enrollment flowchart.

	qSOFA < 2 (N = 2444)	qSOFA ≥ 2 (N = 119)
Age, *years*	77 [13]	80 [11]
Gender, *M*	1244 (50.9)	57 (47.8)
Length-of-stay, *days*	8 [8]	11 [16]
Clavien *points*	0 [2]	3 [4]
EmSFI, *points*	3 [2]	5 [4]
ASA, *points*	3 [1]	4 [1]
CACI, *points*	5 [3]	6 [5]

Table 1. Descriptive statistics of the enrolled cohort. M: male; EmSFI: Emergency Surgery Frailty Index; SIRS: Systemic Inflammatory Response Syndrome; ASA: American Society of Anesthesiologists; CACI: Charlson Age Comorbidity Index.

	Accuracy	PPV	NPV	F1-score
Baseline	0.954 (0.953–0.955)	0.000 (0.000–0.000)	1.000 (1.000–1.000)	0.000 (0.000–0.000)
KNN	0.933 (0.925–0.941)	0.790 (0.693–0.887)	0.940 (0.928–0.952)	0.521 (0.492–0.550)
EN	0.941 (0.933–0.949)	0.850 (0.701–0.999)	0.945 (0.933–0.957)	0.568 (0.523–0.614)
RF	0.962 (0.954–0.970)	0.766 (0.648–0.883)	0.972 (0.967–0.977)	0.650 (0.570–0.729)
XGB	0.955 (0.942–0.968)	0.791 (0.671–0.911)	0.963 (0.948–0.978)	0.619 (0.549–0.689)
Mode	0.956 (0.948–0.965)	0.774 (0.657–0.891)	0.965 (0.957–0.974)	0.620 (0.553–0.686)

Table 2. Evaluation test metrics for all ML models and the Mode model. The highest value is indicated in bold. Means and 95% Confidence Interval across the folds is reported. EN: Elastic-Net; DTC: Decision Tree Classifier; RF: Random Forest; KNN: K-Nearest Neighbors; XGB: eXtreme Gradient Boosting; PPV: Positive Predictive Value; NPV: Negative Predictive Value.

a median percentage of missing variables within the included ones of 0.97% [IQR = 1.37%] (detailed description of individual features missing data percentage in Supplementary Materials A).

A total of 2571 patients were enrolled in the study, with 8 patients removed due to missing data in the outcome. Among the 2563 patients included, 119 had qSOFA score ≥ 2 (10 patients with qSOFA = 3), thus considered as with sepsis (Fig. 2). Patients had a median age of 78 years old [IQR = 13] with 50.76% of males with a median length of stay of 8 days [IQR = 8] and of 11 days [IQR = 16] for patients with qSOFA < 2 and ≥ 2 respectively. Table 1 describes the main clinical characteristics of the cohort included in this study. No univariate screening or statistical analysis was conducted to avoid any sort of train-test contamination or to introduce any bias in the variable selection process.

Prediction models

We evaluated five supervised learning algorithms—Elastic-Net logistic regression, K-Nearest Neighbors (KNN), Random Forest (RF), gradient-boosted trees (XGBoost), and a majority-class baseline—within a 5 × 5 nested cross-validation framework (Table 2). In addition, we constructed an ensemble model that combined the predictions of the four ML algorithms by majority vote. In each outer split, the average prevalence of

postoperative sepsis was averagely 4.6%. As expected in such an imbalanced setting, the naïve majority-class classifier achieved a high accuracy of 0.954 (95% CI 0.953–0.955) despite having zero sensitivity for sepsis.

All ML models clearly outperformed this baseline in terms of discrimination. Mean test accuracies ranged from 0.933 (0.925–0.941) for KNN to 0.962 (0.954–0.970) for RF (Fig. 3). The highest sensitivity was obtained by Elastic-Net (0.850, 0.701–0.999), while RF achieved the highest specificity (0.972, 0.967–0.977). With respect to ROC-AUC, Elastic-Net, RF, XGBoost, and the ensemble achieved values of 0.948 (0.894–1.000), 0.956 (0.927–0.985), 0.954 (0.923–0.984), and 0.961 (0.937–0.986), respectively, versus 0.500 for the majority-class baseline. The corresponding areas under the precision–recall curve (AuPRC, Fig. 4) were 0.752 (0.649–0.855) for Elastic-Net, 0.665 (0.581–0.750) for RF, 0.700 (0.632–0.768) for XGBoost, and 0.713 (0.620–0.806) for the ensemble, compared with 0.523 (0.522–0.524) for the baseline classifier.

Probabilistic calibration, assessed by Brier scores and calibration plots, further favoured tree-based methods (Fig. 4). RF and XGBoost obtained the lowest Brier scores (both 0.037), followed closely by the ensemble (0.038), whereas KNN and Elastic-Net had slightly higher values (0.052 and 0.071, respectively). Aggregated confusion matrices illustrate the error profiles of each model (Fig. 3): for example, the ensemble model misclassified 113/2,563 patients (86 false positives, 27 false negatives), corresponding to a mean sensitivity of 0.774 and specificity of 0.965. Overall, these results indicate that all ML models, and particularly RF, XGBoost, and the ensemble, provide substantial gains in discrimination and calibration over a naïve majority-class strategy.

Discussion

Predictive models and scores are 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, in recent times, ML promptly responded to the necessity of elaborating large-scale multimodal datasets. Thus, integrating clinical data, vital signs, biochemical examinations, and neurophysiological findings enable us to rapidly generate prediction models for many clinical questions[4,5] performing on par with human physicians[6]. Moreover, ML models can be adapted to specific applications and improve its own performance with time as the model is exposed to more information[7]. Furthermore, 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[8,9]. In this regard, Misic et al. (2020) developed ML models to predict postoperative readmission in surgical patients, and in subsequent work Misic et al. (2021) proposed a simulation-based framework that uses these models to represent patient flow in the hospital and to prioritise high-risk patients for targeted readmission-prevention interventions, thereby demonstrating how ML-based risk prediction can translate into improved resource allocation and cost savings. In the same

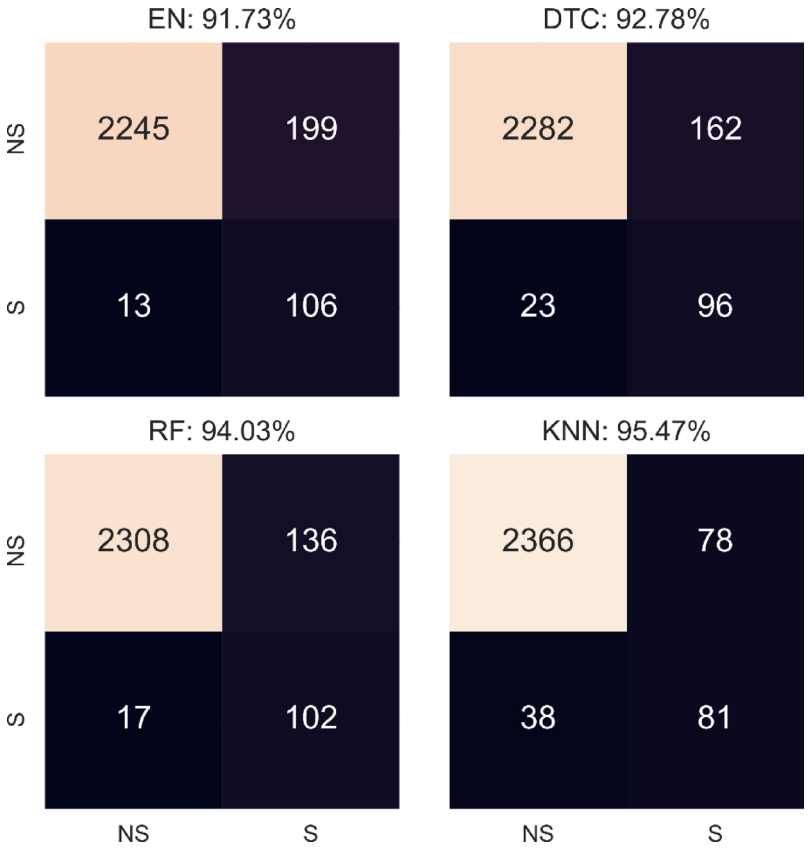

Fig. 3. Confusion matrixes of aggregated test predictions for the four models: EN: Elastic-Net; KNN: K-nearest neighbor; RF: Random Forest; XGB: eXtreme Gradient Boosting.

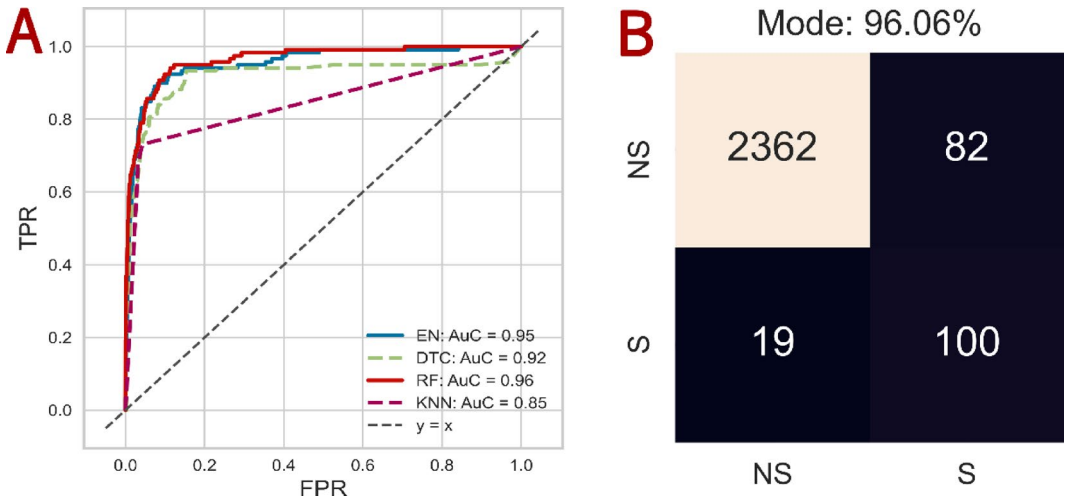

Fig. 4. RoC curve with related AuC for each of the individual models (**A**) and the confusion matrix of the test predictions of the mode model (**B**): EN: Elastic-Net; RF: Random Forest; KNN: K-nearest neighbor; XGB: eXtreme Gradient Boosting. AuC: Area under the Curve; AUPRC: Area Under tbe Precision-Recall Curve; ROC: Receiver-Operator Curve.

optic, by precisely assessing the patients most likely to develop sepsis after surgery, preventive measures can be adopted that can aid surgeons in their decision making regarding the timing and type of intervention, selection and duration of perioperative antibiotics, postoperative disposition, additional monitoring, and need for follow-up surveillance imaging for detection of early infective complication (e.g. abscess), particularly in cases that otherwise would not have raised suspicion in the majority of providers as being high risk for sepsis. To this point, the current study was important because it proposed and internally cross-validated and tested a ML algorithm to identify sepsis ahead of time with excellent individual performance on a large multi-centre, prospectively collected database focused on a population of elderly patient. This model presents an alternative to the all models presented in the present literature because it is a single complex ensemble model derived from comparison of several individual ML models. For the purpose of this study, only readily available pre-operative data were included, allowing faster stratification of sepsis risk without including second-level examinations (e.g., imaging). Furthermore, knowing already in the preoperative period, which condition bolsters the risk of sepsis, contributes to the discussion on the appropriate timing of source control. 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[10]. Moreover, the decision to retain only variables collected before surgery, combined with the exclusion of features with completion rates below 90%, aligns with best practices advocated by Hastie et al.[11,12]. These sources emphasize the crucial role of feature selection in enhancing model interpretability and generalization, underlining the significance of a meticulous approach to data pre-processing. Bunn et al., targeting the risk of sepsis after appendectomy only, with a dataset of 223.214 patients, conclude that Machine learning methods can be used to predict the development of sepsis after appendectomy with moderate accuracy[13]. Similarly, Taylor et al. found that Random Forest outperforms traditional logistic regressions and other traditional clinical tools such as CURB-65 for prediction of sepsis, as captured through 1,697 different International Classification of Disease, version 9, codes with EHR data from 5,278 emergency department visits[14]. In their research Thottakkara et al. evaluated the performance of different ML models for forecasting postoperative sepsis and acute kidney injury (AKI), concluding that generalized additive models and support vector machines had good performance as risk prediction model for postoperative sepsis and AKI[15]. Moreover, the POTTER Study developed an interpretable, accurate, and user-friendly predictor of 30-day outcomes in patients undergoing ACS[16].

To the best of our knowledge, there are no study that specifically targeted the prediction of sepsis in elderly surgical 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). Our *mode* model was found to improve on all the individual ML models tested, reaching an accuracy of 96.06% (sensitivity = 99.20%, specificity = 54.95%) in the prediction of sepsis. Crucial is the high negative predictive value = 84.03% achieved, which shows how ~ 85% of all patients having sepsis were correctly detected before surgery. Importantly, the overall accuracy of our models must be interpreted in light of the strong class imbalance of the cohort, in which a naïve majority-class classifier would already reach an accuracy close to 95%. In this context, discrimination metrics such as AuROC and class-specific measures provide a more meaningful assessment of model performance for clinical decision-making. These properties, together with the favourable Decision Curve Analysis, suggest that the proposed ML solutions could support prioritisation of high-risk patients and targeted allocation of preventive resources, in line with prior work linking higher AuROC values to improved clinical and economic outcomes when such models are used to guide interventions. Among individual models, Random Forest and XGBoost offered the best overall balance between discrimination and calibration, with accuracies of 0.962 and 0.955, ROC-AUCs of

0.956 and 0.954, AuPRCs of 0.665 and 0.700, and Brier scores of 0.037 for both. The ensemble model, obtained by majority vote across all four ML algorithms, further improved global discrimination, reaching an ROC-AUC of 0.961 and AuPRC of 0.713, with accuracy 0.956, sensitivity 0.774 and specificity 0.965. Importantly, at the operating threshold used in our analysis this model misclassified only 27 of 119 patients with sepsis (false negatives) and provided a negative predictive value close to 99%, which is particularly relevant when the primary clinical aim is to safely rule out sepsis and avoid missing high-risk patients. Furthermore, the adoption of a nested-cross validation approach, rarely done in literature[17], adds robustness to the study's methodology. This two-layer k-fold validation strategy, incorporating an outer loop for test set identification and an inner loop for hyperparameter tuning, not only guards against overfitting but also ensures a comprehensive evaluation of the model generalizability. Pre-processing steps, including k-Nearest Neighbours imputation, numerical variable normalization, and Synthetic Minority Oversampling Technique (SMOTE) application, refining the dataset and preparing it for optimal model performance were always derived from the train set and applied to the validation/test sets, to avoid any form of double dipping. Furthermore, the comparison of multiple ML models with hierarchical complexity (e.g., Elastic-Net, KNN, DTC, RF), and use of proper evaluation metrics for unbalanced classification aligns with the comprehensive approach recommended in literature[18,19]. The *mode* approach, integrating posteriors from multiple models, aligns with the ensemble learning principles discussed by Rokach et al.[20], showcasing the potential of combining diverse ML algorithms to enhance predictive performance.

In conclusion, sepsis remains a significant global health concern, particularly in the context of acute care surgery (ACS), where it poses a substantial burden on healthcare systems and contributes to significant morbidity and mortality. Despite its importance, there is a lack of comprehensive epidemiological data specifically addressing sepsis in surgical patients, especially in the emergency setting. Machine learning (ML) has emerged as a promising tool in healthcare, offering the potential to analyze large datasets and develop predictive models for adverse outcomes, including sepsis. Our results demonstrated the effectiveness of ML models in predicting sepsis, with the K-nearest neighbors (KNN) model performing best in terms of accuracy. Furthermore, the ensemble model, which integrated predictions from multiple individual models, showed superior performance compared to individual models, with a high accuracy of 96.06% and notable sensitivity and specificity. This study contributes to the growing body of literature on ML applications in healthcare and fills a critical gap in predicting sepsis in elderly surgical patients in the emergency setting. Moving forward, ML models offer the potential to improve risk stratification, inform clinical decision-making, and ultimately improve outcomes for surgical patients at risk of sepsis. Further research and validation studies are warranted to evaluate the real-world utility and integration of these predictive models into clinical practice.

Methods

Research protocol

This study originated from the FRAILESEL (Frailty and Emergency Surgery in the Elderly[21],) study (ClinicalTrials.gov identifier: NCT02825082). The FRAILESEL is a large, nationwide, multicentre ($N = 29$), 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 informed consent approved by the Ethical Committee of "Sapienza" University of Rome, Italy (No. 4252_{2016}, ClinicalTrials.gov: NCT02825082). This work has been reported in line with the STROCCS criteria[22] and all methods were followed in accordance with the declaration of Helsinki.

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[21]. 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. Sepsi was defined according to the third international consensus definitions for sepsis and septic shock (Sepsis-3[23]). Quick Sofa score and Sofa score were taken into account to stratify the risk of infection.

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 tumours and preoperative risk was assessed with anaesthesiologist-assigned American Society of Anaesthesiologists class. Specific details on collected data and used protocols can be found in Costa et al.[21]. Outcome was categorized via the dichotomized qSOFA score ≥ 2/<2.

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), gynaecological 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.

Feature screening and model selection

To simulate the same conditions available for the prediction of sepsis, only variables taken before the surgery were retained from the aforementioned database (Supplementary Materials A for all variables and Costa et al. for a detailed explanation of the collection pipeline[21]. Features with a completion rate lower than 90% were excluded a priori. The list of the included variables and the results of the missing data analysis were included in Supplementary Materials A. Continuous variables were summarized as median and interquartile range (IQR), whereas categorical variables were reported as counts and percentages. No univariable hypothesis-test-based feature screening was performed before model development in order to avoid information leakage and biased variable selection.

A nested-cross validation approach was implemented[17]. 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. Secondly, inside each train set, hence the inner loop, a k-Fold cross-validation was implemented to tune the optimal hyperparameters of the individual models with the number of inner/outer folds set to 5/5. Missing values were imputed using k-Nearest Neighbours ($N_{neighbours} = 10$) derived from the train set and then applied in the test set. Then, the train and test sets' numerical variables were normalized with the train set mean and standard deviation. Subsequently, each inner training dataset was resampled (to cope with the unbalanced classification problem) via the Synthetic Minority Oversampling Technique (SMOTE[24]. Training data first entered a dimensionality reduction step (whitened Principal Component Analysis with explained variance set to 0.999) and then a feature selection one (cross-validated recursive feature selection step with base estimator a decision tree). Each pre-processing step was computed on the train set and then applied on the respective test set, preventing train-test contamination. For each ML model and each training fold, an hyperparameter optimization was performed using *Optuna*[25]. In each optimization trial, hence for each evaluated hyperparameter combination, the outer training data were split in the actual (inner) train and validation sets implementing the aforementioned inner K-Fold cross-validation loop. 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 ML models included were an Elastic-Net (EN,[11]), a K-nearest neighbours (KNN,[26]), a Decision Tree Classifier (DTC,[27]) a Random Forest (RF[27]), and gradient boosted trees (XGBoost, via the related library). Lastly, the *weighted posteriors* model's prediction was computed by computing the mode of the predictions of all the aforementioned models as in Mannini et al.[28]. The list of optimized parameters for each model is provided within Supplementary Material B.

Performance metrics and statistical analysis

Model performance was assessed on the outer test folds using accuracy, sensitivity, specificity, precision, F1-score, area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), and Brier score. For discrimination analysis, AUROC values were reported with 95% confidence intervals computed using the nonparametric DeLong method. Confidence interval bounds were obtained from the estimated standard error of the AUROC and constrained to the admissible range between 0 and 1. Calibration was additionally assessed using calibration plots, and the potential clinical utility of the best-performing model was explored by decision curve analysis. Performance estimates were aggregated across outer folds and reported with 95% confidence intervals. All analyses were performed in Python using scikit-learn, imbalanced-learn, XGBoost, and Optuna. Decision Curve Analysis (DCA) will be used to evaluate the utility of using the best ML solution when compared to the consequences of false positives (i.e., *treat-all*) and of false negatives (i.e., *treat-none*).

Data availability

Code and trained models are available on the following GitHub link ([https://github.com/PeppeDrone/SepsisFRAILESELPredictor](https:/github.com/PeppeDrone/SepsisFRAILESELPredictor)). Raw data can be obtained after request to the corresponding author for research purposes.

Received: 19 June 2024; Accepted: 23 March 2026
Published online: 02 April 2026

References

1. Fleischmann, C. et al. Assessment of global incidence and mortality of hospital-treated sepsis. current estimates and limitations. *Am. J. Respir Crit. Care Med.* **193** (3), 259–272. https://doi.org/10.1164/rccm.201504-0781OC (2016).
2. Moore, L. J., Moore, F. A., Jones, S. L., Xu, J. & Bass, B. L. Sepsis in general surgery: A deadly complication. *Am. J. Surg.* **198** (6), 868–874. https://doi.org/10.1016/j.amjsurg.2009.05.025 (2009).
3. Vogel, T. R., Dombrovskiy, V. Y. & Lowry, S. F. Trends in postoperative sepsis: are we improving outcomes? *Surg. Infect.* **10** (1), 71–78. https://doi.org/10.1089/sur.2008.046 (2009).
4. Shamout, F., Zhu, T. & Clifton, D. A. Machine learning for clinical outcome prediction. *IEEE Rev. Biomed. Eng.* **14** (9134853), 116–126. https://doi.org/10.1109/RBME.2020.3007816 (2021).
5. Hashimoto, D. A., Rosman, G., Rus, D. & Meireles, O. R. Artificial intelligence in surgery: Promises and perils. *Ann. Surg.* **268** (1), 70–76. https://doi.org/10.1097/SLA.0000000000002693 (2018).
6. Ahmad, Z., Rahim, S., Zubair, M. & Abdul-Ghafar, J. Artificial intelligence (AI) in medicine, current applications and future role with special emphasis on its potential and promise in pathology: present and future impact, obstacles including costs and acceptance among pathologists, practical and philosophical considerations. A comprehensive review. *Diagn. Pathol.* **16**, 24. https://doi.org/10.1186/s13000-021-01085-4 (2021).
7. Pathak, A., Pandey, M. & Rautaray, S. Adaptive Model for Dynamic and Temporal Topic Modeling from Big Data using Deep Learning Architecture. *Int. J. Intell. Syst. Appl.* **11**, 13–27. https://doi.org/10.5815/ijisa.2019.06.02 (2019).

8. Langenberger, B., Schulte, T. & Groene, O. The application of machine learning to predict high-cost patients: A performance-comparison of different models using healthcare claims data. *PLOS ONE*. **18** (1), e0279540. https://doi.org/10.1371/journal.pone.0279540 (2023).
9. Atalan, A., Şahin, H. & Atalan, Y. A. Integration of Machine Learning Algorithms and Discrete-Event Simulation for the Cost of Healthcare Resources. *Healthcare* **10** (10), 1920. https://doi.org/10.3390/healthcare10101920 (2022).
10. Fransvea, P. et al. Study and validation of an explainable machine learning-based mortality prediction following emergency surgery in the elderly: A prospective observational study. *Int. J. Surg. Lond. Engl.* **107**, 106954. (2022). https://doi.org/10.1016/j.ijsu.2022.106954
11. Zou, H. & Hastie, T. Regularization and Variable Selection via the Elastic Nets. *J. R Stat. Soc. B.* **67**, 301–320 (2015).
12. Hastie, T., Friedman, J. & Tibshirani, R. *The Elements of Statistical Learning* (Springer, 2001). https://doi.org/10.1007/978-0-387-21606-5
13. Bunn, C. et al. Application of machine learning to the prediction of postoperative sepsis after appendectomy. *Surgery* **169** (3), 671–677. https://doi.org/10.1016/j.surg.2020.07.045 (2021).
14. Taylor, R. A. et al. Prediction of In-hospital Mortality in Emergency Department Patients With Sepsis: A Local Big Data-Driven, Machine Learning Approach. *Acad. Emerg. Med. Off J. Soc. Acad. Emerg. Med.* **23** (3), 269–278. https://doi.org/10.1111/acem.12876 (2016).
15. Thottakkara, P. et al. Application of Machine Learning Techniques to High-Dimensional Clinical Data to Forecast Postoperative Complications. *PLoS ONE*. **11** (5), e0155705. https://doi.org/10.1371/journal.pone.0155705 (2016).
16. El Hechi, M. W. et al. Validation of the Artificial Intelligence-Based Predictive Optimal Trees in Emergency Surgery Risk (POTTER) Calculator in Emergency General Surgery and Emergency Laparotomy Patients. *J. Am. Coll. Surg.* **232** (6), 912–919e1. https://doi.org/10.1016/j.jamcollsurg.2021.02.009 (2021).
17. Cawley, G. C. & Talbot, N. L. C. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *J. Mach. Learn. Res.* **11** (70), 2079–2107 (2010).
18. Caruana, R., Munson, A. & Niculescu-Mizil, A. Getting the Most Out of Ensemble Selection. In: *Sixth International Conference on Data Mining (ICDM'06)*. :828–833. (2006). https://doi.org/10.1109/ICDM.2006.76
19. Dietterich, T. G. Ensemble Methods in Machine Learning. In: *Multiple Classifier Systems. Lecture Notes in Computer Science*. Springer; :1–15. https://doi.org/10.1007/3-540-45014-9_1 (2000).
20. Rokach, L. Ensemble-based classifiers. *Artif. Intell. Rev.* **33** (1), 1–39. https://doi.org/10.1007/s10462-009-9124-7 (2010).
21. Costa, G., Massa, G. & 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). *Updat Surg.* **70** (1), 97–104. https://doi.org/10.1007/s13304-018-0511-y (2018).
22. Mathew, G. et al. Strengthening the reporting of cohort, cross-sectional and case-control studies in surgery. *Int. J. Surg. Lond. Engl.* **96**, 106165. (2021). https://doi.org/10.1016/j.ijsu.2021.106165
23. Singer, M. et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA* **315** (8), 801–810. https://doi.org/10.1001/jama.2016.0287 (2016).
24. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357. https://doi.org/10.1613/jair.953 (2002).
25. Akiba, T. et al. A Next-generation Hyperparameter Optimization Framework. In: *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. (2019).
26. Cover, T. & Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory*. **13** (1), 21–27. https://doi.org/10.1109/TIT.1967.1053964 (1967).
27. Leo, B. *Classification And Regression Trees* (1st Ed) (Routledge, 1994).
28. Mannini, A. et al. Data-driven prediction of decannulation probability and timing in patients with severe acquired brain injury. *Comput. Methods Programs Biomed.* **209**, 106345. https://doi.org/10.1016/j.cmpb.2021.106345 (2021).

Acknowledgements

The study was funded by the Italian Ministry of Health under the "Ricerca Corrente" program and by the 5 × 1000 Funds via the SeeABI project of the IRCCS Fondazione Don Gnocchi, as well as under the NEUROREHAB4EU THCS 2023 funded by the Italian Ministry of Health and by the WAVE project funded by the Italian Ministry of Defense.

Author contributions

P.F., M.S., and G.C. defined the data collection protocol and participated in data collection; P.F. conceptualized the study; P.L. developed the models and performed formal analysis; P.L. and P.F. wrote the manuscript (draft), G.S provided supervision and all authors read and edited the final version of the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-46040-9.

Correspondence and requests for materials should be addressed to P.L.