2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Perioperative Risk Assessment in Pancreatic Surgery Using Machine Learning

Bjarne Pfitzner*[1], Jonas Chromik*[1], Rachel Brabender[2], Eric Fischer[2], Alexander Kromer[2], Axel Winter[3], Simon Moosburner[3], Igor M. Sauer[3], Thomas Malinka[3], Johann Pratschke[3], Bert Arnrich[1], and Max M. Maurer[3]

Abstract— **Pancreatic surgery is associated with a high risk for postoperative complications and death of patients. Complications occur in a variable interval after the procedure. Often, a patient has already left the ICU and is not properly monitored anymore when the complication occurs. Risk stratification models can assist in identifying patients at risk in order to keep these patients in ICU for longer. This, in turn, helps to identify complications earlier and increase survival rates. We trained multiple machine learning models on pre-, intra- and short term postoperative data from patients who underwent pancreatic resection at the Department of Surgery, Campus Charité Mitte | Campus Virchow-Klinikum, Charité – Universitätsmedizin Berlin. The presented models achieve an area under the precision-recall curve (AUPRC) of up to 0.51 for predicting patient death and 0.53 for predicting a specific major complication. Overall, we found that a classical logistic regression model performs best for the investigated classification tasks. As more patient data becomes available throughout the perioperative stay, the performance of the risk stratification model improves and should therefore repeatedly be computed.**

I. INTRODUCTION

Despite advances in surgical techniques, pancreas resections are still associated with considerable rates of postoperative complications. According to recent studies, more than one out of four patients experiences at least one major complication within the perioperative stay and, subsequently, overall mortality remains as high as 7% [1]. Of particular relevance are post-operative pancreatic fistulas (POPF), i.e. microlesions of the pancreatic organ surface, that may trigger intraabdominal infections and often require additional interventional treatment. Moreover, post-pancreatectomy haemorrhage (PPH), i.e. internal bleeding, occurs as an acute and life-threatening event. Identifying patients at risk is of crucial importance to initiate therapeutic measures at an early stage and prevent patients' conditions from deteriorating.

Typical complications, like POPF or PPH, occur within variable intervals after the actual surgical procedure [2]. At this time, the patient has usually already left the intensive care unit (ICU) with its superior monitoring capabilities and has been transferred to the normal surgical ward with only limited surveillance infrastructure. It is therefore difficult to detect severe complications in a timely manner under the present circumstances, although early detection is crucial for efficient treatment [3]. The aim of this project is to evaluate approaches to assist the treating physicians in identifying patients at risk for postoperative complications.

Considering their potential in pattern recognition, machine learning models appear to be a particularly promising approach to using patient data for event forecasting. To predict the specific endpoints listed in Section III-A, we trained different models as described in Section III-B based on pre-, intra-, and postoperative static patient data. This dataset was furthermore augmented by postoperative dynamic data of four continuously measured vital parameters recorded on the ICU over up to 72 hours after surgery.

II. RELATED WORK

The use of machine learning for surgical risk assessment is not a novel development, however, the majority of studies only take preoperative risk factors into account. In a notable recent work, Chiew et al. [4] try to predict postoperative mortality and prolonged ICU stay using different machine learning models. An important finding is that the area under the receiver operating characteristic curve (AUROC) is not a good metric for judging machine learning models in this area since the class-balance is skewed towards having more samples in the negative class than in the positive one. By predicting all samples as part of the negative class, a decent AUROC is achieved while having a very low sensitivity thus creating no meaningful prediction. The authors recommend using the area under the precision-recall curve (AUPRC) instead, which does not take the overwhelming number of true negatives into account, thus being a better display of a classifier's performance in the presence of highly imbalanced data [5]. The authors report an AUPRC of 0.23 for the patient mortality endpoint.

Another notable work is DyCRS [6], a Hidden Markov Model adaptation for predicting postoperative complications after elective colectomy surgeries. Contrary to Chiew et al., DyCRS uses static patient data such as gender and age in combination with dynamic postoperative features such as heart rate and blood pressure and reports an AUPRC of 0.52.

*These authors contributed equally to this work.
[1]Hasso Plattner Institute, University of Potsdam, Germany
firstname.lastname@hpi.de
[2]Hasso Plattner Institute, University of Potsdam, Germany
firstname.lastname@student.hpi.de
[3]Department of Surgery, Campus Charité Mitte | Campus Virchow-Klinikum, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, 13353 Berlin, Germany
firstname.lastname@charite.de

This research was partly funded by the Federal Ministry of Education and Research of Germany in the framework of KI-LAB-ITSE (01IS19066).

This work is licensed under a Creative Commons Attribution 3.0 License.
For more information, see http://creativecommons.org/licenses/by/3.0/

III. MATERIALS & METHODS

With the approval from the Charité's Ethics Committee (Approval ID: EA2/035/14), we used perioperative patient data from the Department of Surgery, Campus Virchow Klinikum, Charité – Universitätsmedizin Berlin to train multiple machine learning models. In the following, we describe the data and models in detail.

A. Materials

The dataset consists of 521 cases of pancreatic resections and contains static (or tabular) and dynamic (or time-series) data.

Static Data: The static part of the data can be split into 20 pre-, 15 intra- and 8 postoperative features. The preoperative information includes standard patient data such as age (63.9 ± 12.5), sex (44.7% female), weight ($74.9\mathrm{kg} \pm 15.3\mathrm{kg}$) and height ($1.72\mathrm{m} \pm 0.10\mathrm{m}$), as well as comorbidities. Examples of intraoperative data are the type and duration of the surgery and blood loss. Postoperative information is aggregated in two well-known scores reproducing patient condition, namely the second version of the *Simplified Acute Physiology Score* (SAPS II) [7] and the *Therapeutic Intervention Scoring System* (TISS-10) [8]. These scores are calculated directly after ICU admission and every day at 6 a.m. thereafter for a maximum of three days. SAPS II is a score for quantifying the severity of the disease and the morbidity of the patient. TISS-10 quantifies the amount of care needed by a patient, taking into account 10 therapeutic intervention items the patient might receive.

Dynamic Data: The continuous vital parameters are collected for a maximum of 72 hours during patient surveillance on the ICU. Most variables are measured approximately once every 30 minutes. Vital parameters included in the dataset are heart rate (HR), sytolic blood pressure ($\mathrm{BP_{sys}}$) and body temperature (T), as well as information about the volume of urine output ($\mathrm{V_U}$).

Endpoints: For each patient, we used machine learning analysis to predict the following endpoints or labels:

1) Death of the patient (DoP)
2) Re-admission to the ICU (ReA)
3) Post-operative pancreatic fistula (POPF)
4) Post-pancreatectomy haemorrhage (PPH)

All four endpoints are highly imbalanced with a fraction of 6.7% (DoP), 20.9% (ReA), 16.5% (POPF), 10.2% (PPH) representing the positive class. It is important to note, that the labels do not provide any information regarding the point in time of the complication or event. It may have occurred shortly after surgery or any time throughout the inpatient course.

B. Methods

Our aim was to predict the endpoints DoP, ReA, POPF, and PPH. Since the data correspond to different points in time during the inpatient stay, we further investigated whether predictions at an earlier stage – i.e. right after surgery or even preoperatively – are reasonable. Overall, five different machine learning models were applied:

1) Logistic regression (LR) with L2 regularisation,
2) Decision tree (DT) with a maximum depth of 4 and Gini impurity as split criterion,
3) Support vector machine (SVM) with a radial basis function kernel,
4) Gradient boosting machine (GBM) with exponential loss and 100 estimators,
5) Combination of feed-forward neural network (NN) and Gated recurrent unit (GRU).

In contrast to the first four models, the last one utilises separate inputs and processing of static and dynamic data. It consists of an NN for the static data, with two hidden layers having 12 and 8 neurons, respectively, and a single GRU with dropout for the dynamic data. Their last layers are then concatenated and followed by a final dense output layer. Both components are L1-regularised and use the ReLu activation function.

Data Preprocessing: The static part of the data was preprocessed depending on the feature types. Binary features can be given to the models directly, whereas categorical features were one-hot-encoded and numerical features were normalised between 0 and 1.

For the time-series data on heart rate, sytolic blood pressure, and body temperature, two preprocessing steps were required: missing value handling and time interval alignment. first, missing values were inferred using the longitudinal imputation method 'last observation carried forwards'. Then, since the time intervals between measurements were not always of the same length, the values were aligned to exactly 30 minutes using cubic spline interpolation. The volume of urine output had to be preprocessed separately, as it was not measured regularly, but rather at specific instances, i.e. when the discharge bag had to be emptied. Thus it was converted from absolute volume to relative volume per time interval (also 30 minutes).

For the combined NN and GRU model, the lengths of time-series data were aligned across patients by masking shorter sequences with -1 values at the end.

Feature Extraction: From the resulting time-series, we extracted mean value, variance, minimum, maximum, and linear slope as features to be added to the easier models that cannot deal with time-series-data directly. The combined NN and GRU model is able to work with the dynamic data directly, without the necessity to use extracted features.

Feature Selection: A correlation analysis of the static data showed that some features are highly correlated ($\rho > 0.7$), namely the day of the week, the kind of resection performed, and the expertise of the conducting surgeon quantified by the number of procedures performed in his/her career. To avoid unstable models, we selected the single feature that showed the highest predictive potential in a simple LR baseline model only trained on static data, discarding the other correlated features.

To reduce the complexity of the prediction task and thereby improve the models' performances, we moreover performed a feature selection for all models except NN and NN+GRU. As feature selection method, we chose a

stepwise feature selection by cross validation as described in [9] using leave-one-subject-out cross-validated AUPRC as performance metric. This was performed for all models, endpoints and points in time during the perioperative stay.

Model Training and Evaluation: For model evaluation, we performed leave-one-subject-out cross-validation within each analysis and report the AUPRC, as well as the precision at a fixed recall value of 0.6. We want to evaluate whether the models' performances increases as the patients progresses through the perioperative stay. A later stage in this perioperative stay corresponds to more features being available to the model. Hence, we first trained all models using only preoperative features to assess how well complications can be predicted using only basic patient information and comorbidities. Then, we added intraoperative features to assess the prediction performance with data available at the end of the surgery. After that, we additionally included postoperative, static features, and finally postoperative times-series data to determine the importance of vital parameters for the models.

IV. RESULTS

All baseline models were implemented in Python using the scikit-learn[1] framework and default parameters, whereas the deep model was implemented using Keras[2].

For all endpoints, the models improve with an increasing amount of available data as depicted in Fig. 1, which illustrates the AUPRC of all models evaluated at different stages during the inpatient stay. Considering each endpoint separately, however, different models achieve the highest AUPRC. Notably, DT performed specifically well when adding postoperative and dynamic data. In contrast, LR showed the lowest variance across all models and stage defined databases revealing the most stable prediction performance. Overall, it seems more difficult to predict the two specific complications POPF and PPH as compared to the other two more general endpoints DoP and ReA.

Taking a closer look at the commonly used LR model, Fig. 2 shows the AUPRC for the different endpoints based on the full static and dynamic dataset. The coresponding best F1-Scores for the endpoints DoP, ReA, POPF and PPH are 0.54, 0.44, 0.37 and 0.35, respectively.

For the experiments having all input features available, the selected features on average consist of $40\% \pm 11$pp pre-, $23\% \pm 11$pp intra-, $13\% \pm 10$pp postoperative, and $24\% \pm 13$pp dynamic features, indicating the importance of comorbidities.

Finally, Table I shows the precision of all models for a fixed recall value of 0.6 on the complete dataset including all available features. The best result for each endpoint is indicated by a bold font style.

V. DISCUSSION

Our results reveal two main findings: Firstly, for varying endpoints and feature sets, different models show the best

[1] www.scikit-learn.org
[2] www.keras.io

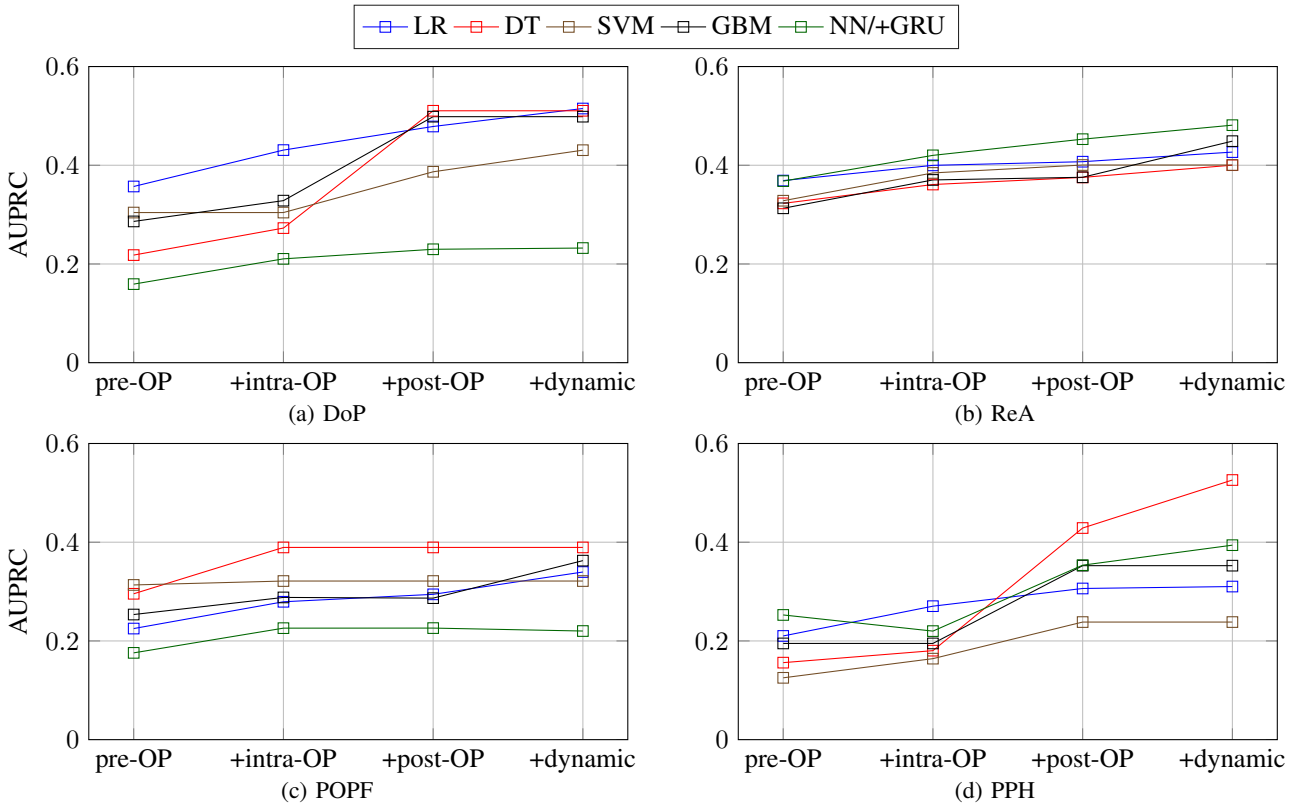

Fig. 1: AUPRC of all models for all examined endpoints. The x-axis depicts the amount of data available to the models starting with only preoperative data and gradually adding intra- and postoperative data, as well as dynamic data.

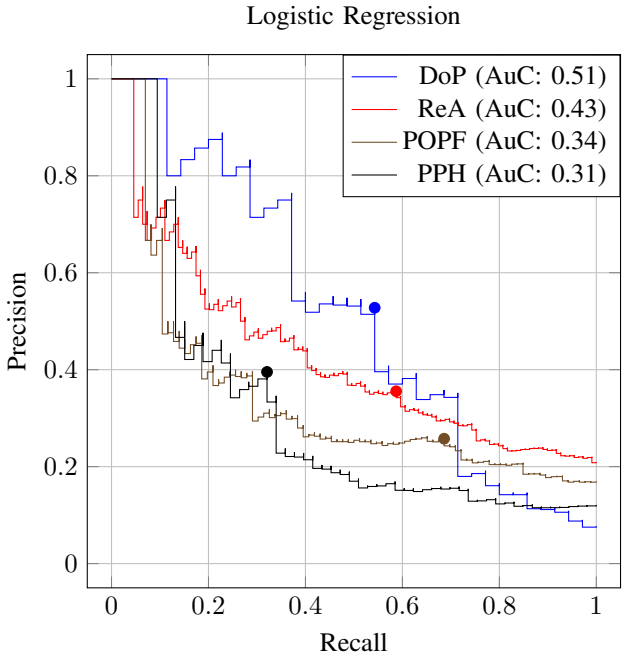

Fig. 2: Precision-Recall curves for the LR model trained on the dataset representing the complete perioperative patient stays (after feature selection). The locations of the highest F1-Scores are indicated by the dots.

TABLE I: Precision values at a fixed recall value of 0.6 based on the complete dataset.

Model	DoP	ReA	POPF	PPH
LR	**0.38**	**0.33**	0.25	0.15
DT	0.37	0.18	0.19	0.11
SVM	0.27	0.29	0.25	0.15
GBM	**0.38**	0.30	**0.30**	**0.19**
NN+GRU	0.24	0.30	0.22	0.09

performance. This motivates an evaluation of numerous types of models for risk prediction tasks. The inferiority of the more complex NN/+GRU model (except for ReA) may arise from the dataset encompassing too few cases and being imbalanced. More research and particularly larger datasets are needed in order to evaluate whether prediction performance can be further increased with models of higher complexity.

Secondly, AUPRC generally increases with more available features per case. This manifests a demand for closer patient monitoring and elaborate data collection in order to enable a better prediction of complications on a per-patient basis. It also shows that machine learning-based risk stratification should be repeatedly performed throughout the inpatient stay in order to increase prediction confidence.

Comparing the different endpoints, ReA revealed the most stable prediction characteristics as indicated by the lowest variance across all models. This could be due to the smallest class imbalance compared to the other endpoints or the significance of the collected data for ReA. A very imbalanced dataset is challenging for machine learning classifiers, which is again aggravated by the low amount of data overall.

Even though our results compare well with other AUPRC reported in aforementioned related work, the yet overall modest precision and recall results may so far only serve as an indicator to support medical professionals in identifying patients with elevated risk for complications or death. However, those patients might subsequently be taken under enhanced surveillance after surgery, either by extending their stay on the ICU or by establishing advanced monitoring capacities at surgical wards, e.g. by equipping them with wearable devices. On the other hand, identifying patients with only low risk is equally important in everyday clinical practice. ICU capacity is scarce and costly, thus being able to discharge patients early based on a low risk profile for complications or low re-admission probability is of high use.

Future work should be concerned with the collection and analysis of more patient data which could improve the classification metrics and reduce the impact of class imbalances. The availability of larger datasets could also allow for deeper, more complex models, which have shown a better performance for many use-cases. Finally, a promising research area for expanding the available amount of data is the application of federated learning, which would not only increase the dataset but enable anonymity of shared data.

References

[1] P. Baum, J. Diers, S. Lichthardt, C. Kastner, N. Schlegel, C.-T. Germer, and A. Wiegering, "Mortality and complications following visceral surgery—a nationwide analysis based on the diagnostic categories used in German hospital invoicing data," *Deutsches Aerzteblatt Online*, Nov. 2019. [Online]. Available: https://www.aerzteblatt.de/10.3238/arztebl.2019.0739

[2] N. Hyman, T. L. Manchester, T. Osler, B. Burns, and P. A. Cataldo, "Anastomotic Leaks After Intestinal Anastomosis: It's Later Than You Think," *Annals of Surgery*, vol. 245, no. 2, pp. 254–258, Feb. 2007. [Online]. Available: http://journals.lww.com/00000658-200702000-00014

[3] N. A. Hirst, J. P. Tiernan, P. A. Millner, and D. G. Jayne, "Systematic review of methods to predict and detect anastomotic leakage in colorectal surgery," *Colorectal Disease*, vol. 16, no. 2, pp. 95–109, Feb. 2014. [Online]. Available: http://doi.wiley.com/10.1111/codi.12411

[4] C. J. Chiew, N. Liu, T. H. Wong, Y. E. Sim, and H. R. Abdullah, "Utilizing Machine Learning Methods for Preoperative Prediction of Postsurgical Mortality and Intensive Care Unit Admission," *Annals of Surgery*, vol. 272, no. 6, pp. 1133–1139, Dec. 2020.

[5] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets," *PloS one*, vol. 10, no. 3, p. e0118432, 2015.

[6] W. Wang, H. Zhao, H. Zhuang, N. Shah, and R. Padman, "DyCRS: Dynamic Interpretable Postoperative Complication Risk Scoring," in *Proceedings of The Web Conference 2020*, ser. WWW '20. New York, NY, USA: Association for Computing Machinery, Apr. 2020, pp. 1839–1850.

[7] J.-R. Le Gall, S. Lemeshow, and F. Saulnier, "A New Simplified Acute Physiology Score (SAPS II) Based on a European/North American Multicenter Study," *JAMA*, vol. 270, no. 24, pp. 2957–2963, 12 1993. [Online]. Available: https://doi.org/10.1001/jama.1993.03510240069035

[8] D. J. Cullen, J. M. Civetta, B. A. Briggs, and L. C. Ferrara, "Therapeutic intervention scoring system: a method for quantitative comparison of patient care," *Critical care medicine*, vol. 2, no. 2, pp. 57–60, 1974, 4832281[pmid]. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/4832281

[9] K. Tanaka, T. Kurita, F. Meyer, L. Berthouze, and T. Kawabe, "Stepwise feature selection by cross validation for eeg-based brain computer interface," in *The 2006 IEEE International Joint Conference on Neural Network Proceedings*. IEEE, 2006, pp. 4672–4677.