The Signature-Based Model for Early Detection of Sepsis.

James Morrill[1, 3] — morrill@maths.ox.ac.uk
A. Kormilitzin[1, 2], Alejo Nevado-Holgado [2], S. Swaminathan[3],
S. Howison[1], T. Lyons[1]

Abstract

Overview

- Optimal feature selection leads to enhanced efficiency and accuracy when developing both supervised and unsupervised machine-learning models.
- We propose a new **signature-based regression model** to automatically identify a patient's risk of sepsis based on physiological data streams.
- We use a gradient boosting machine learning algorithm with features at the current time-points and the signature features extracted from the time-series to model the longitudinal effects of sepsis.

Model performance

- The proposed method achieved a **0.433 utility score** in the official phase of the challenge.
- The signature method shows a systematic and competitive approach to model sepsis by learning from health data streams.

Sepsis Labels

We choose a labelling of the cases to take into account information about the utility function.

- For a given time $t = T$ denote U_0 to be the utility score gained from predicting zero, and U_1 for predicting one.
- Provide the label $U_1 - U_0$ at the given time-point.
- Use these labels to train a regression model.

Time	Utility score 0	Utility score 1	Label
1	0.00	-0.05	-0.05
2	0.00	-0.05	-0.05
3	-0.22	0.89	1.11
4	-0.44	0.78	1.22
5	-0.67	0.67	1.33
6	-0.89	0.56	1.44

Table 1. Example of the case labelling given the utility score from predicting 0 and 1.

This labelling is preferable since it gives **larger absolute values for samples that lead to a larger absolute utility score** and are thus more important to label correctly.

Hand-Crafted Features

Aside from the values of each of the variables at the current time-point, **we augment our feature space with some additional features** derived from the current time-point and from recent time-series information. These are listed below in Table 1.

Feature Name	Description
ShockIndex	Ratio of heart rate to systolic blood pressure (HR/SBP)
BUN/CR	Bilirubin / creatinine ratio
PartialSOFA	Partial recreation of the SOFA score from the components found in the challenge data
SOFA deterioration	Binary marker labelled as 1 if PartialSOFA has deteriorated by 2 points in the previous 24 hours
Min Vitals	Min value each vital sign takes over some look-back window
Max Vitals	Max value each vital sign takes over some look-back window
Num measurements	The number of measurements taken of each lab value over some look-back window.

Table 2. Hand-crafted features derived from the data to augment our feature space with. The separating horizontal line splits the features that are computed from the current time-point from those that require prior time-series information.

Signature Features

We utilise the 'signature transformation' in our feature selection process. This transformation summarises important information about a path allowing us to generate good candidate features to be fed into a machine learning algorithm.

The path signature A path of finite length can be described by the mapping $X : [a, b] \to \mathbb{R}^d$ where each co-ordinate path $(X_t^1, ..., X_t^n)$ is a real valued path $X^i : [a, b] \to \mathbb{R}$. We then define

$$S(X)_{a,t}^{i_1,\dots,i_k} = \int_{a<t_k<t} \dots \int_{a<t_k<t_2} \mathrm{d}X_{t_1}^{i_1} \dots \mathrm{d}X_{t_k}^{i_k}, \qquad (1)$$

which is a real number and called the k-fold iterated integral of X along the indexes $i_1, ..., i_k$. Given this we define the signature of the path X as

$$S(X)_{a,b} = (1, S(X)_{a,t}^{1}, ..., S(X)_{a,t}^{d}, S(X)_{a,t}^{1,1}, S(X)_{a,t}^{1,2}, ...). \qquad (2)$$

This sequence of real numbers encodes the information of the path. The entries of the signature provide good candidates for the path features to be used in machine learning. We will use the values of the truncated signature over one breath as our features.

Summary

- Provides a hierarchical method to summarise the longitudinal information about a path.
- Theoretically, any continuous function on a path can be approximated arbitrarily well by a linear function of the signature terms.
- Reduces the need for more ad-hoc feature engineering.

Algorithm Features

- We take the signature of different feature paths in time-series truncated to some order.
- The resulting values are fed into our learning algorithms as features.
- Transformations can be applied to the paths before their signatures are computed, thus helping to uncover additional information at lower orders of the signature.

Model Training and Hyperparameters

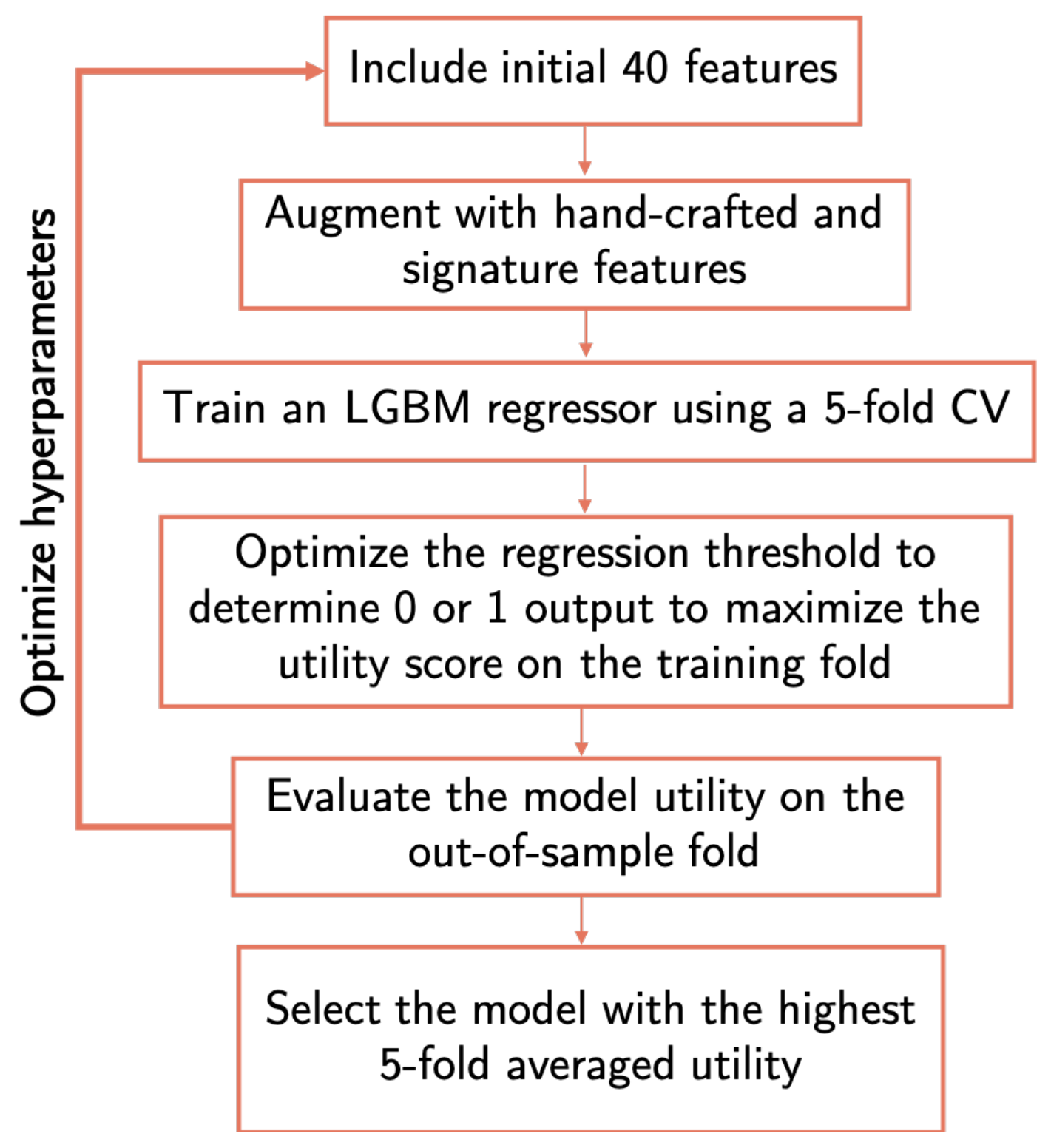
Hyperparameters Below is a list of the optimised hyperparameters.

Parameter	Final Value
Num measurement look-back	8
Min/max look-back	6
Sig look-back	7
Sig columns	All features
Sig order	3
Sig leadlag	True

Sig = Signature

Table 3. A list of basic hyperparameters optimised in model training.

Note that the look-back windows are all close to the six-hour value. Six hours being the common time over which sepsis is assumed to be detectable.



Include initial 40 features

Augment with hand-crafted and signature features

Train an LGBM regressor using a 5-fold CV

Optimize the regression threshold to determine 0 or 1 output to maximize the utility score on the training fold

Evaluate the model utility on the out-of-sample fold

Select the model with the highest 5-fold averaged utility

Figure 1. Model training and hyperparameter optimization process.

Results

CV results

We give the 5-fold CV results of the top-scoring algorithm on the training set below.

	Fold 1	2	3	4	5	Average (std)
Utility score	0.432	0.434	0.437	0.448	0.400	0.430 (0.018)

Table 4. Scores of the submitted algorithm on each cross-validation fold of the training data.

Usefulness of signatures

Table 5 gives the averaged CV score with different feature subsets. Inclusion of the signature values gives good improvement on the overall utility score: **the signature transformation is successfully uncovering relevant information from the time-series that can be used to discriminate cases of sepsis.**

Features	Averaged utility score
Time only	0.282
Original 40 features only	0.389
Hand-crafted features included	0.418
Hand-crafted features and signatures included	0.430

Table 5. Scores from models trained on different subsets of features.

Making Early Predictions

Informing physician decisions

- We want to provide **clinically actionable information to physicians** to help make decisions in hospitals.
- Given regression output, **we select a threshold such that once exceeded, the patient is marked 'at risk' of developing sepsis**.
- The **AUC ROC value** for marking such people as septic **is 0.868**.

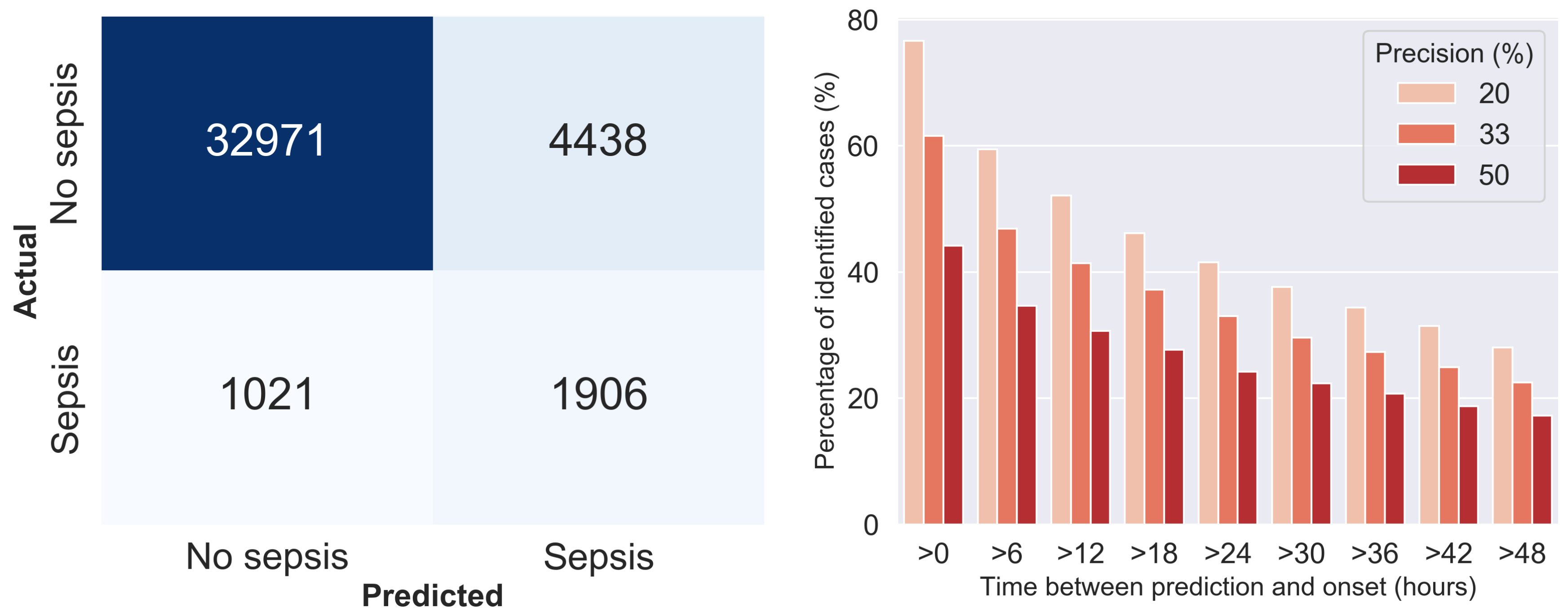


Figure 2. Confusion matrix displaying the number of people predicted as likely to get sepsis compared with those who actually end up with sepsis with the threshold tuned to 33% specificity (left). Proportion of sepsis cases predicted correctly in different time windows tuned to different precision levels (right).

Analysis

- The results suggest that the developed model can be **used as an early screening tool** to determine high risk patients.
- The screening threshold can be chosen to achieve **clinically meaningful sensitivity and specificity**.
- We see from the above right figure that **early detection is usually not in the desired region within 6 hours**. It often occurs much later.

Acknowledgements

ITEREX THERAPEUTICS
AT-HOME DIGITAL THERAPIES

EPSRC
Engineering and Physical Sciences Research Council

[1] Mathematical Institute, University of Oxford, [2] Department of Psychiatry, University of Oxford, [3] Iterex Therapeutics, Manhattan NY,