Explainable Machine-Learning Predictions For The Prevention of Hypoxaemia in Surgery

Journal Club - June 13 2019

Background

- Surgery and anesthesia pose risk of complications and death
- Adverse events are often preventable
- With increased adoption of electronic medical record systems, there is an abundance of data recorded about the patient
- Unfortunately, these data are rarely used to improve patient outcomes

Background

- Hypoxaemia is defined as low arterial blood oxygen (SpO2 <= 92%), and can cause serious harm to patients during surgery
- Real-time data monitors only allow reactive measures; clinicians alone are not able to reliably predict or prevent future events of hypoxaemia

Problem Formulation

Objectives Build a machine-learning system that can predict:

1) At the start of the procedure, the risk of hypoxaemia at any time during the procedure

2) Throughout the procedure, the risk of hypoxaemia in the next 5 minutes

...and provide interpretable results.

Input Features

- **Text data:** Pre-op notes, intra-op procedure notes (bag of words)
- **Binary data:** Procedure codes, OR location, anaesthesia type (general vs regional), gender
- **Numeric data:** continuous monitoring data (hemodynamic and ventilation), lab data, medication dosages
  - Last value, time since last value
  - Exponential decaying weighted average and variance*

*https://stats.stackexchange.com/questions/286640/definition-of-the-function-for-exponentially-decaying-weighted-average/286643

Models

Baseline models:

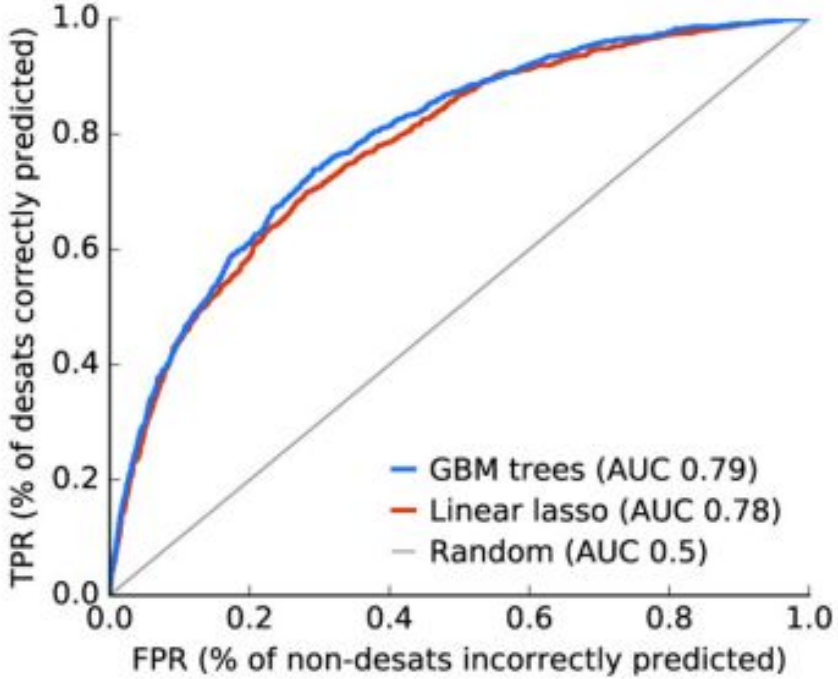
1) Logistic regression with lasso
2) Linear SVM autoregressive model modeling the stream of SpO2 data
3) Parzen window

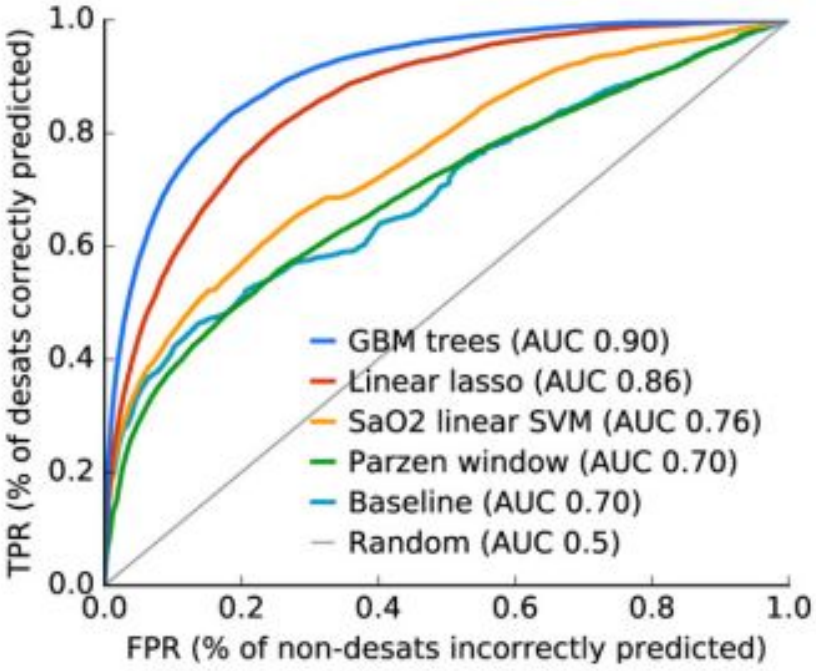
Comparative Models:

1) Xgboost for pre-operative risk
2) Xgboost for real-time risk

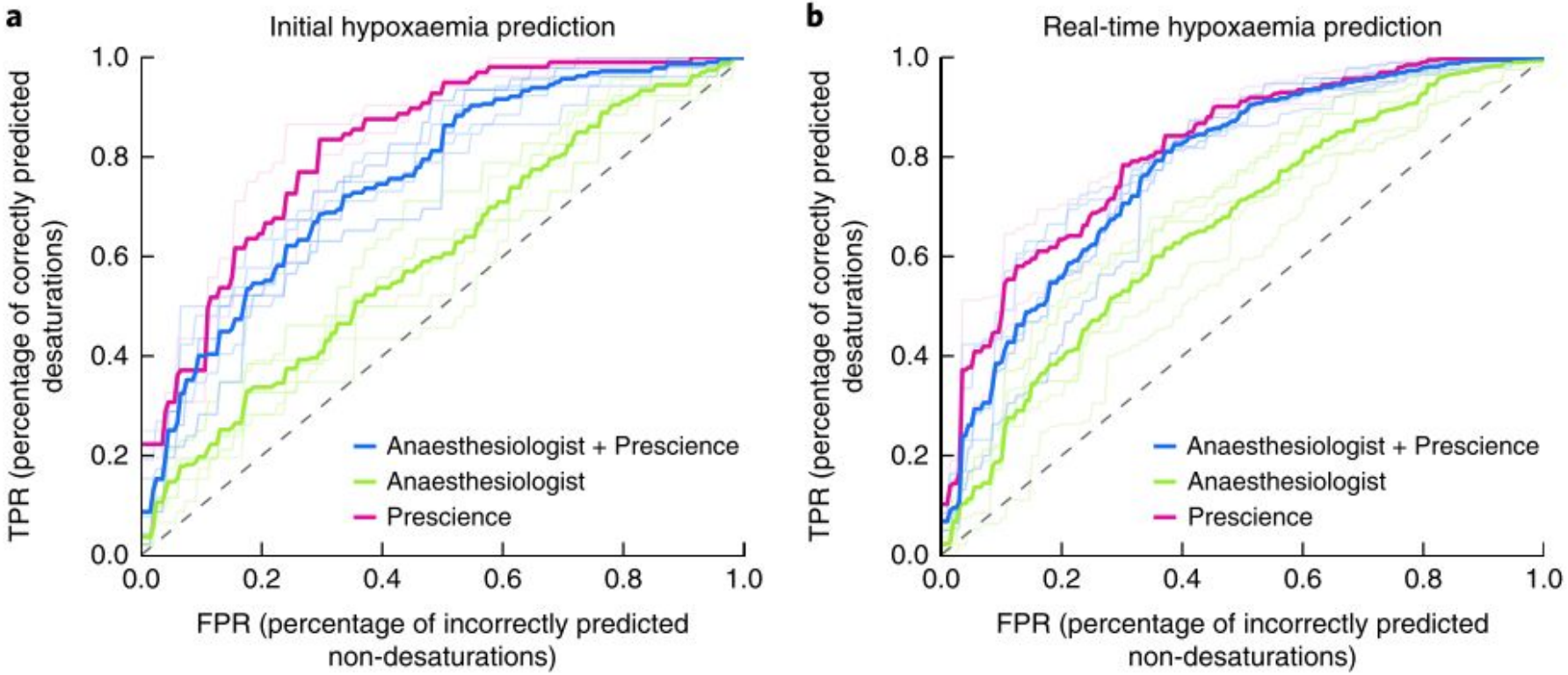
Model Development

- Hyperparameters were tuned using validation set
- Bagging was employed using 50% sub-samples of the training data for each tree
- Real-time models were slightly more complex (max depth 6 vs 4); minimum child weight was increased (10 vs 1) to counter added complexity
- Test set was compressed until method development was completed





AUC comparison of XGBoost and baseline models
(left: pre-op risk model; right: real-time risk model)



Model performance was better than anaesthesiologists (with or without the model guidance)

Model complexity vs Model interpretability

↑ complexity ⇒ ↑ performance + ↓ interpretability

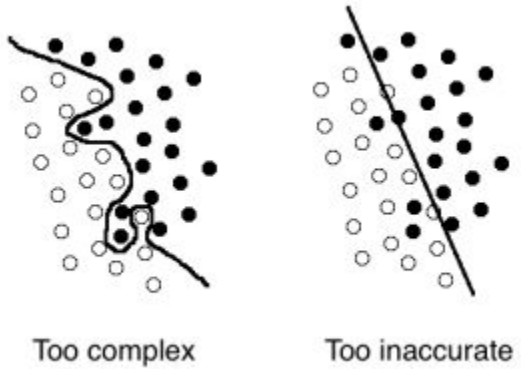


Image source:
https://prateekvjoshi.com/2014/09/07/whats-the-importance-of-hyperparameters-in-machine-learning/

Shapley values

A **model-agnostic** approach to quantify feature importance, with desirable qualities such as **additivity** (AKA local accuracy) and **consistency**

Shapley values - additivity

Interpretation: the additivity assumption states that the model output is the sum of:

1) The expected model output (eg. intercept, bias)
2) The sum of feature attributions from all input predictors

$$f(x) = \phi_0(f, x) + \sum_{i=1}^{M} \phi_i(f, x)$$

Model output

Expected model output over training data

Sum of feature attributions

Shapley value - consistency (aka monotonicity)

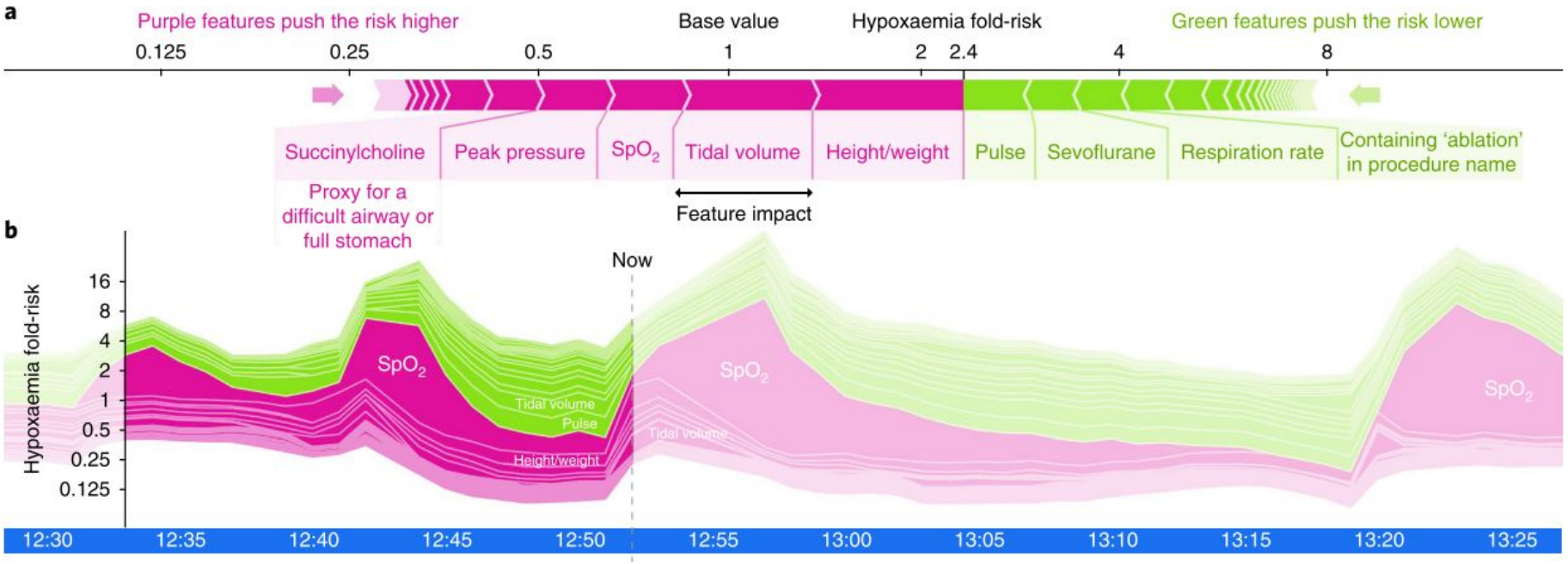
Interpretation: if a predictor is more important in Model A vs Model B, then the importance attribution for that feature should be higher in Model A vs Model B.

$$f^A(S \bigcup \{i\}) - f^A(S) \geq f^B(S \bigcup \{i\}) - f^B(S)$$

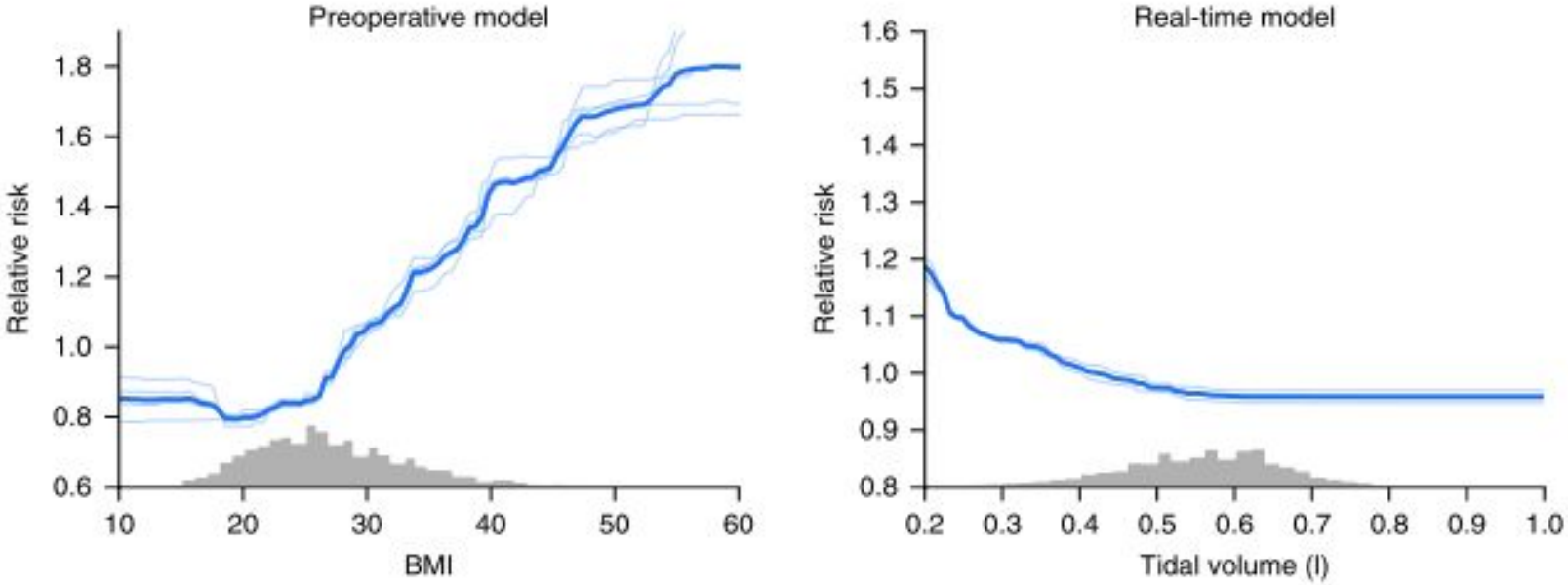
Difference in model A output
when x_i is present vs absent

$$\phi_i(f^A, x) \geq \phi_i(f^B, x)$$

Feature attribution
for x_i in model A



Prediction breakdown: 2.4 fold risk of event in next 5 mins (natural log scale; height/weight and tidal volume contributed most to increased risk)



Partial dependence plot of average feature effects

Discussion

Discussion Questions

1. What are some ways to evaluate feature importance? Do you find them to be useful in practice, why or why not?

Discussion Questions

2. Is model interpretability useful? When should model interpretability be prioritized over performance?

Discussion Questions

3. Besides interpretability, what are some desirable traits for prediction models in production?

Discussion Questions

4. How can we use Shapley values? How should we NOT use Shapley values?

Extra slide - Outcomes definition

1) **Hypoxaemia** was defined as dropping from >=95% to <=92% for in the next five minutes
2) **Not hypoxaemic** meant >=95% for the previous 10 minutes and the next 10 minutes