Towards Trustworthy Cross-patient Model Development

Ali El-Merhi[1,2], **Helena Odenstedt Hergés**[1,2], **Linda Block**[1,2], **Mikael Elam**[3], **Richard Vithal**[1,2], **Jaquette Liljencrantz**[3], **Miroslaw Staron**[4]

[1]Sahlgrenska Academy, Institute of Clinical Sciences, University of Gothenburg, Gothenburg, Sweden,
[2] Department of anesthesia and intensive care, Sahlgrenska university hospital, Gothenburg, Sweden,
[3] Institute of Neuroscience and physiology, Sahlgrenska Academy, University of Gothenburg,
[4] Computer Science and Engineering, Chalmers | University of Gothenburg, Gothenburg, Sweden

Abstract

Machine learning is used in medicine to support physicians in examination, diagnosis, and predicting outcomes. One of the most dynamic area is the usage of patient generated health data from intensive care units. The goal of this paper is to demonstrate how we advance cross-patient ML model development by combining the patient's demographics data with their physiological data. We used a population of patients undergoing Carotid Enderarterectomy (CEA), where we studied differences in model performance and explainability when trained for all patients and one patient at a time. The results show that patients' demographics has a large impact on the performance and explainability and thus trustworthiness. We conclude that we can increase trust in ML models in a cross-patient context, by careful selection of models and patients based on their demographics and the surgical procedure.

1 Introduction

Critical care units increasingly often adopt machine learning to improve the outcome of patients (Maslove, Elbers, and Clermont 2021), (Flechet, Grandas, and Meyfroidt 2016). Critical care patients and patients undergoing surgery are connected to equipment used by physicians to monitor, such signals as ECG, blood pressure or blood oxygenation. These signals can also be used to predict upcoming adverse events and to decrease the time needed to detect these events, thereby reducing the number of invasive tests or increasing the certainty of the diagnosis (Wijnberge et al. 2020), (Carra et al. 2020).

One of the challenges when using ML models in critical care is how to use a model trained on one patient for diagnosing another patient (Komorowski 2019). From the medical perspective, the demographics of the patients can differ greatly, even if they undergo the same procedure or suffer from the same disease. From the technical perspective, the signals collected during the procedures can be collected differently (e.g. variations in how electrodes are attached to the patient). Therefore, in order to trust the ML models, it is important to identify the variability, understand the source of it and then choose the right strategy to handle it.

In practice, in most cases, ML engineers and scientists adopt the strategy to collect more data in the hope to create models which are robust to these variations (Smiti 2020). This means that the models' predictions will be of a better quality, but, at the same time, more data can lead to lower explainability and thus trustworthiness.

In this paper, we set off to investigate how much the results vary from when we train the model on all patients, and when we train the model on one patient at a time. The goal was to explore the variability of models and therefore assess how robust the models are to different patients.

The data was collected from a pipeline where we designed and conducted a clinical study of patients undergoing CEA (Staron et al. 2021), (Block et al. 2020). We used a RandomForest classifier, which has been used in many studies analyzing signals from patients (Qi 2012).

Our results show that there is a large variability between how we train and use the model. For the same model, when trained on the data from all patients, the accuracy was 0.86. However, for the same algorithm, but trained on one patient at a time, the accuracy varied from 0.07 to 0.63. The model performed best for the pair of patients who were the most similar (same gender, similar age, similar initial status). Although the results were better for the model trained on all patients, we could observe that the model is more trustworthy if we can trace the training model to a specific patient rather than a large dataset with high variability.

2 Context – CEA and Collected Signals

Carotid endarterectomy (CEA) is recommended in symptomatic patients with $\geq 50\%$ stenosis and in asymptomatic patients with $\geq 60\%$ stenosis (Eckstein 2018). CEA poses an increased risk of cerebral ischemia due to embolization and hemodynamic derangement during clamping of the carotid arteries. In our centre, all carotid endarterectomies are performed with patients under general anaesthesia. All intraoperative monitoring, including invasive arterial blood pressure (ABP) and neuromonitoring, is connected to patients and monitoring starts prior to induction of anaesthesia. Baseline values before induction of anesthesia are recorded. Noradrenaline infusion is titrated as needed to maintain ABP close to baseline values. After the skin incision and surgical exposure but before plaque removal, the surgeon cross-clamps the common carotid artery , the external carotid artery , and the internal carotid artery (ICA), so that the stenotic bifurcation is excluded from the circulation. At this moment, the surgeon, taking into account information from available neuromonitoring and measured blood pressure in the distal stump of the ICA, evaluates the risk of cerebral ischemia due to hypoperfusion and decides if a shunt is to be placed to maintain cerebral perfusion during clamping. After plaque removal the artery is sutured and the clamp is released to restore blood flow. After emergence from anaesthesia, the patient undergoes frequent bedside neurological examinations.

Since the patients are connected to monitors, we can collect the following signals, and transform them into ML features:

- EEG: frequency bands (alpha, beta, delta, gamma, and theta) for each channel (F3-Cz, F4-Cz, C3-Cz, C4-Cz, P3-Cz, and P4-Cz).
- ECG: Inter-beat intervals (IBI), BPM (Heartbeats per minute), SDNN (Standard deviation of the NN (R-R) intervals), SDSD (standard deviation of successive differences), RMSSD (Root mean square of the successive differences), PNN50 (proportion of NN50[1] divided by the total number of NN (R-R) intervals), PNN20 (The proportion of NN50[2] divided by the total number of NN (R-R) intervals), HR_MAD (median absolute deviation of RR intervals), SD1 (standard deviation perpendicular to identity line), SD2 (standard deviation along identity line), S (Area of ellipse described by SD1 and SD2), SD1/SD2 (SD1/SD2 ratio), Breathing rate (estimated from the ECG signal),
- ABP (Arterial Blood Pressure): ABP mean,
- NIRS (Near-Infrared Spectroscopy): rSO2 left (side of the brain), and rSO2 right (side of the brain), and
- SpO_2 (Oxygen Saturation): SpO_2 value.

The data set contains also a decision class, which is the set of events which we want to recognize/diagnose during CEA. We use the following events:

- Pre-clamp anaesthesia – induction of anaesthesia- when the patient's anesthesia starts.
- Pre-clamp surgery – opening the cartoid artery.
- Clamped artery – when the carotid artery is closed.
- Shunt – when the surgeons make a shunt around the closed artery to restore blood flow (in case the patient's blood oxygenation decreases to a pre-defined level).
- HRV window anesthesia pre-clamp – six breaths during one minute in the ventilator, to establish a baseline for the analysis of HRV signals under controlled conditions. This is done both before the clamping of the artery and after the clamping of the artery.
- Post-operative anesthesia – when the patient is regaining consciousness.
- Post-operative care – when the patient is awake and under post-operative care in the neurointensive care unit.

The dataset in our study consists of 48 feature columns and one class column. It

We experiment with several ML algorithms – RandomForest, SVM, AdaBoost, Decision Trees (CART) and artifical neural networks. We chose RandomForest for our comparisons due to its robustness and ability to handle large data sets (Qi 2012). In this study, we follow the procedure:

1. Collect the data from all patients in one table.
2. Train and validate the classifier based on all data, using 0.33 train-test split.
3. Train the classifier on one patient at a time and validate on all other patients.

This procedure let us compare how the performance of the classifier varies depending on the way in which it is trained and evaluated.

3 Results and interpretation

To understand the data, we use the t-SNE diagram as a means for visualization of all data points (Figure 1). The diagram shows that there are areas where different events/decision classes overlap, but in the majority of cases the data points for specific events are grouped together (e.g. the purple dots, corresponding to the clamping of the artery, forms several distinct areas in the figure).

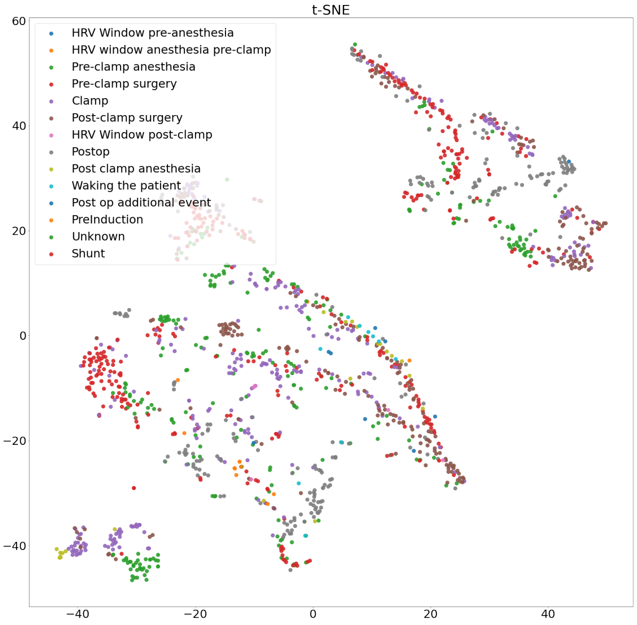


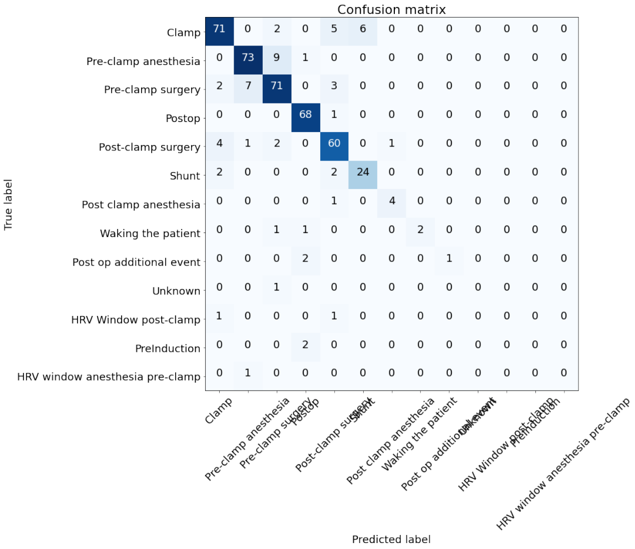
Figure 1: t-SNE diagram to illustrate the data set. Each data point represents one minute of a CEA procedure. Each color represents the type of event/decision class.

We train the classifier with the following parameters: number of trees: 64, number of leaves: 128. The accuracy for that model is 0.86. Figure 2 presents the confusion matrix for this model. The confusion matrix shows that most of

[1] The number of pairs of successive NN (R-R) intervals that differ by more than 50 ms.

[2] The number of pairs of successive NN (R-R) intervals that differ by more than 20 ms.

the errors are for these classes which have the fewest data points (e.g. Pre-induction); this is despite balancing classes using weights.



	Clamp	Pre-clamp anesthesia	Pre-clamp surgery	Postop	Post-clamp surgery	Shunt	Post clamp anesthesia	Waking the patient	Post op additional event	Unknown	HRV Window post-clamp	PreInduction	HRV window anesthesia pre-clamp
Clamp	71	0	2	0	5	6	0	0	0	0	0	0	0
Pre-clamp anesthesia	0	73	9	1	0	0	0	0	0	0	0	0	0
Pre-clamp surgery	2	7	71	0	3	0	0	0	0	0	0	0	0
Postop	0	0	0	68	1	0	0	0	0	0	0	0	0
Post-clamp surgery	4	1	2	0	60	0	1	0	0	0	0	0	0
Shunt	2	0	0	0	2	24	0	0	0	0	0	0	0
Post clamp anesthesia	0	0	0	0	1	0	4	0	0	0	0	0	0
Waking the patient	0	0	1	1	0	0	0	2	0	0	0	0	0
Post op additional event	0	0	0	2	0	0	0	0	1	0	0	0	0
Unknown	0	0	1	0	0	0	0	0	0	0	0	0	0
HRV Window post-clamp	1	0	0	0	1	0	0	0	0	0	0	0	0
PreInduction	0	0	0	2	0	0	0	0	0	0	0	0	0
HRV window anesthesia pre-clamp	0	1	0	0	0	0	0	0	0	0	0	0	0

Figure 2: Confusion matrix for the cross-patient model for the validation data (0.33 of the entire data set).

Figure 3 presents a diagram of accuracy for a cross-patient trained model. Each model has been trained on one patient and applied on another. The lowest accuracy is 0.07 and the highest is 0.63. Although the diagram shows a significant variability in accuracy, we can understand it better if we observe the difference between the feature importance for the cross-patient model and the model for the best pair of patients (C009 and C001).

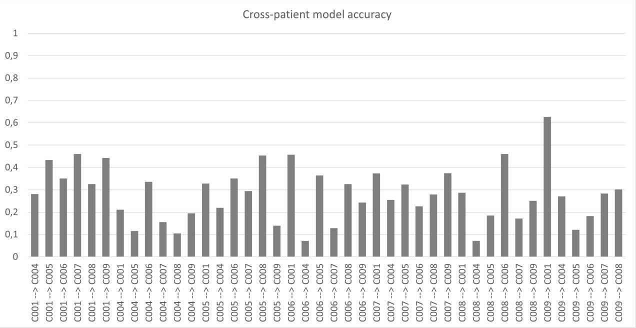


Figure 3: Accuracy of the models trained for each pair of patients. The model is trained on one patient and then validated on another patient (train patient –>validation patient).

From the perspective of explainability and, thus, trust in the model, accuracy is not the best validation metric. In our study, we use feature importance charts as a means to understand how good the models are and whether they can be trusted. For the cross-patient model, Figure 4 presents the feature importance chart for the most important 15 features. The chart shows that it is the arterial blood pressure (ABP) which is the most important feature.

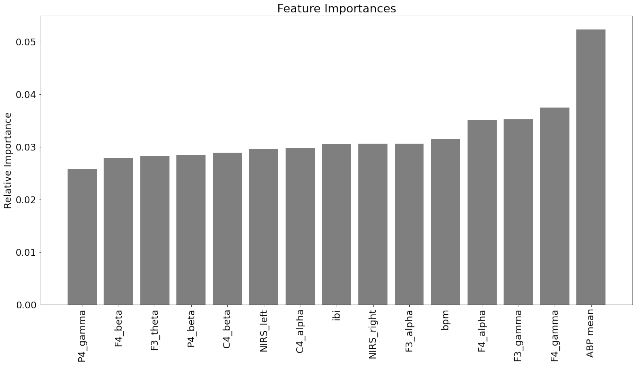


Figure 4: Feature importance for the model trained for all patients. The most important feature is the Arterial Blood Pressure (ABP).

For the model trained on individual patients, however, the feature importance chart is different – see Figure 5. The most important feature is the NIRS signal from the left side of the head. Although both ABP and NIRS can have a good explanation, there is a medical reason for this difference. Patients C001 and C009, are similar in terms of: which artery was clamped (right), condition for admittance (Amaurois fugax dx; Paresthesia left hand vs. Hemiparesis (left); Amaurois fugax sin.), gender and age. Patients C006 and C004 (the worst match) are of different gender and arterial clamp was performed on different sides. As NIRS has been found useful for same-side arterial clamp (Lewis et al. 2018), it confuses the algorithm for different-side arterial clamping, thus rendering this signal untrustworthy in this context.

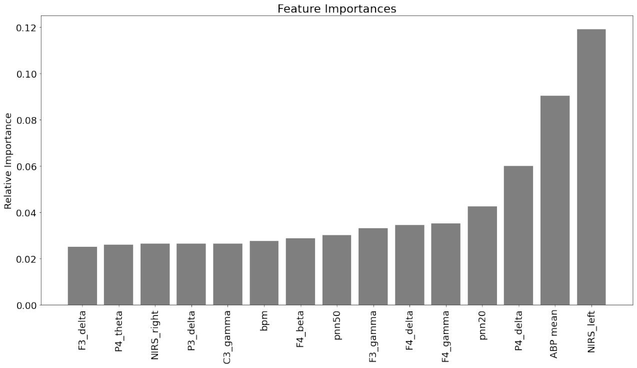


Figure 5: Feature importance for the model trained for patient C009 and applied on the patient C001 (the pair with the highest accuracy or 0.63). The most important feature is the NIRS signal.

Considering both of these charts together, our interpretation is that the similarity of the patients condition is important for the trustworthiness of the model – clamping on the same side strengths the models accuracy, compare to other pairs of patients. We observe also that the number of data points in the data set increases the accuracy of the model in general.

We combine these two approaches: set aside one patient for validation while training the model on all other patients. The results are shown in Figure 6.

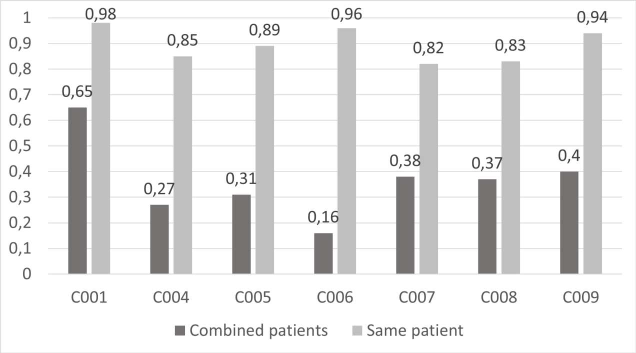


Figure 6: Accuracy of the combined model – one patient is set aside for validation while the model is trained for the data from all other patients; and models trained on the same patient – setting aside 0.33 points for validation.

The highest accuracy is for patient C001 – 0.65. The lowest accuracy is for patient C006 – 0.16. There is an improvement across all patients, although not a large one. For patient C001, we also include the feature importance chart – Figure 7.

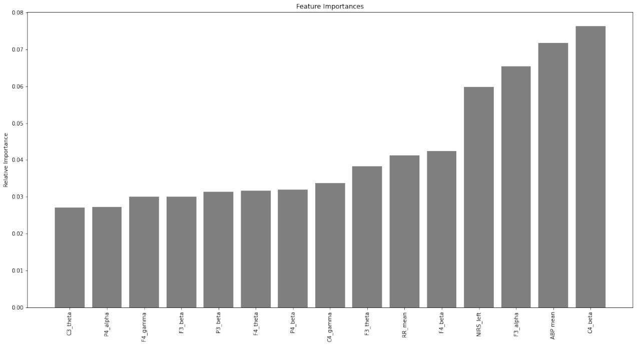


Figure 7: Feature importance chart for the model validated for patient C001.

Our recommendation, therefore, is to decide early whether we want more explainability or higher accuracy. For a reference, the accuracy if we train the model for 0.67 of the data for one patient and validate on the remaining 0.33, the accuracy is above 0.82.

4 Conclusions

In the context of using machine learning models in medical practice, trustworthiness is one of the crucial components in the adoption of this technology in practice. In this paper, we present results of a study on how the performance of a Random Forest classifier changes when using the same data in two different ways – training the model on the entire data set and training it per patient.

The results show that the model trained on all data is more accurate, but has lower explainability and thus trustworthiness. The model trained on a single patient has lower accuracy, but higher explainability. The accuracy depends on the selection of the patients, which is, therefore, a required feature for this algorithm.

Our conclusions are that we need to combine both approaches – selecting the most relevant patients and training the model on a larger data set. We plan to evaluate these conclusions based on new patients groups.

References

Block, L.; El-Merhi, A.; Liljencrantz, J.; Naredi, S.; Staron, M.; and Odenstedt Hergès, H. 2020. Cerebral ischemia detection using artificial intelligence (CIDAI)—A study protocol. *Acta Anaesthesiologica Scandinavica* 64(9): 1335–1342.

Carra, G.; Salluh, J. I.; da Silva Ramos, F. J.; and Meyfroidt, G. 2020. Data-driven ICU management: Using big data and algorithms to improve outcomes. *Journal of Critical Care* .

Eckstein, H.-H. 2018. European Society for Vascular Surgery guidelines on the management of atherosclerotic carotid and vertebral artery disease. *European Journal of Vascular and Endovascular Surgery* 55(1): 1–2.

Flechet, M.; Grandas, F. G.; and Meyfroidt, G. 2016. Informatics in neurocritical care: new ideas for Big Data. *Current Opinion in Critical Care* 22(2): 87–93.

Komorowski, M. 2019. Artificial intelligence in intensive care: are we there yet? *Intensive care medicine* 45(9): 1298–1300.

Lewis, C.; Parulkar, S. D.; Bebawy, J.; Sherwani, S.; and Hogue, C. W. 2018. Cerebral neuromonitoring during cardiac surgery: a critical appraisal with an emphasis on near-infrared spectroscopy. *Journal of cardiothoracic and vascular anesthesia* 32(5): 2313–2322.

Maslove, M. D.; Elbers, P. W.; and Clermont, G. 2021. Artificial intelligence in telemetry: what clinicians should know. *Intensive Care Medicine* (3): 143–151.

Qi, Y. 2012. Random forest for bioinformatics. In *Ensemble machine learning*, 307–323. Springer.

Smiti, A. 2020. When machine learning meets medical world: Current status and future challenges. *Computer Science Review* 37: 100280.

Staron, M.; Hergés, H. O.; Naredi, S.; Block, L.; El-Merhi, A.; Vithal, R.; and Elam, M. 2021. Robust Machine Learning in Critical Care–Software Engineering and Medical Perspectives. *arXiv preprint arXiv:2103.08291* .

Wijnberge, M.; Geerts, B. F.; Hol, L.; Lemmers, N.; Mulder, M. P.; Berge, P.; Schenk, J.; Terwindt, L. E.; Hollmann, M. W.; Vlaar, A. P.; et al. 2020. Effect of a machine learning–derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery: the HYPE randomized clinical trial. *Jama* 323(11): 1052–1060.