Predictive model structure

The trained classifiers are an ensemble of two different models: a logistic regression (LR) and a random forest (RF) (28) model. The decision of building a model combining these two classification methodologies arises by comparing the most popular classification algorithms in biological sciences (Support Vector Machines (SVM), LR, RF, and Multi-Layer Perceptron (MLP)) on 4 different metrics: accuracy, ROC-AUC, sensitivity and specificity.

This analysis allowed us to verify that LR and RF seem to achieve better performance, compared to SVM models in all training sets [(see Table A2, Supplemental Digital Content 7, which contains the performance metrics when model are trained on MIMIC-III training set); (see Table A3, Supplemental Digital Content 8, which contains the performance metrics when models are trained on the eICU-CRD training set); & (see Table A4, Supplemental Digital Content 9, which contains the performance metrics when models are trained on both databases)]. As expected, the ensemble of these two approaches achieves a better performance in almost all the estimated metrics. Most likely, it allows to account for possible non-linear patterns in the decision boundary of the final classifier that LR may not detect. The two components are trained in a joint optimization procedure that fixes the hyperparameters of both algorithms through a Bayesian modeling. (27)

Bayesian optimization is a heuristic method that is capable of achieving results comparable to a grid search in fewer iterations and without the need to explore a massive hyperparameter space. Bayesian statistics help to focus the search routine in each iteration on areas of the hyperparameters space that seems to be more promising with respect to the specified loss function.

The designed algorithm predicts the probability that a certain patient will bleed (or need a transfusion as surrogate marker) in the predefined forecasting window. Usually, the predicted probabilities $\hat{y}$ in binary classification are mapped into deterministic labels by selecting the output label, or class, with higher probability. This step is essential to estimate specificity, sensitivity, accuracy and the confusion matrix (see Figure 7, which contains the confusion matrices for all the classifiers). However, in classification problems with class imbalance the dataset is biased to the dominant or majority class (the most frequent class). This imbalance can affect the learning of any machine learning algorithm and skew the model predictions towards the majority class.

It is verified that this imbalance is more significant in the eICU-CRD, where there are more entries for the "non-bleeding" label (23.31%) and pushing the predictions in favor to this label. In a medical context, this could be undesirable since missing a bleed-event is more *costly* than missing a non-bleeding one. We addressed this problem by searching during the optimization for a decision threshold γ that determines, for example, if a patient with a probability of needing transfusion of 0.31 should be labeled as "bleeding" or "non-bleeding". This allows to boost the model recall and in exchange for some false positives.

The customized loss function estimated during model optimization takes into consideration the F1-score and accuracy as shown in the following equation:

$$loss = 1 - (0.8 * F1\_score + 0.2 * accuracy)$$

By using the above combination of F1-score and accuracy, it is forced the classifiers to jointly maximize precision and the recall of the final model notwithstanding accuracy.

The *base* ensemble model is a voting classifier (*hard voting*) composed of one LR model and one RF model. A label (class 1 or 0) is assigned to the most frequent predicted label (*i.e.* the one which is predicted by at least two classifiers). During the Bayesian optimization, we evaluate 100 voting classifiers on a training set. After this procedure, we kept the best three models and we averaged the predicted probabilities of these three models (*soft voting*). In addition to that, the threshold γ for defining the outcome is chosen during hyperparameter optimization in order to obtain a calibrated model with the best performance.

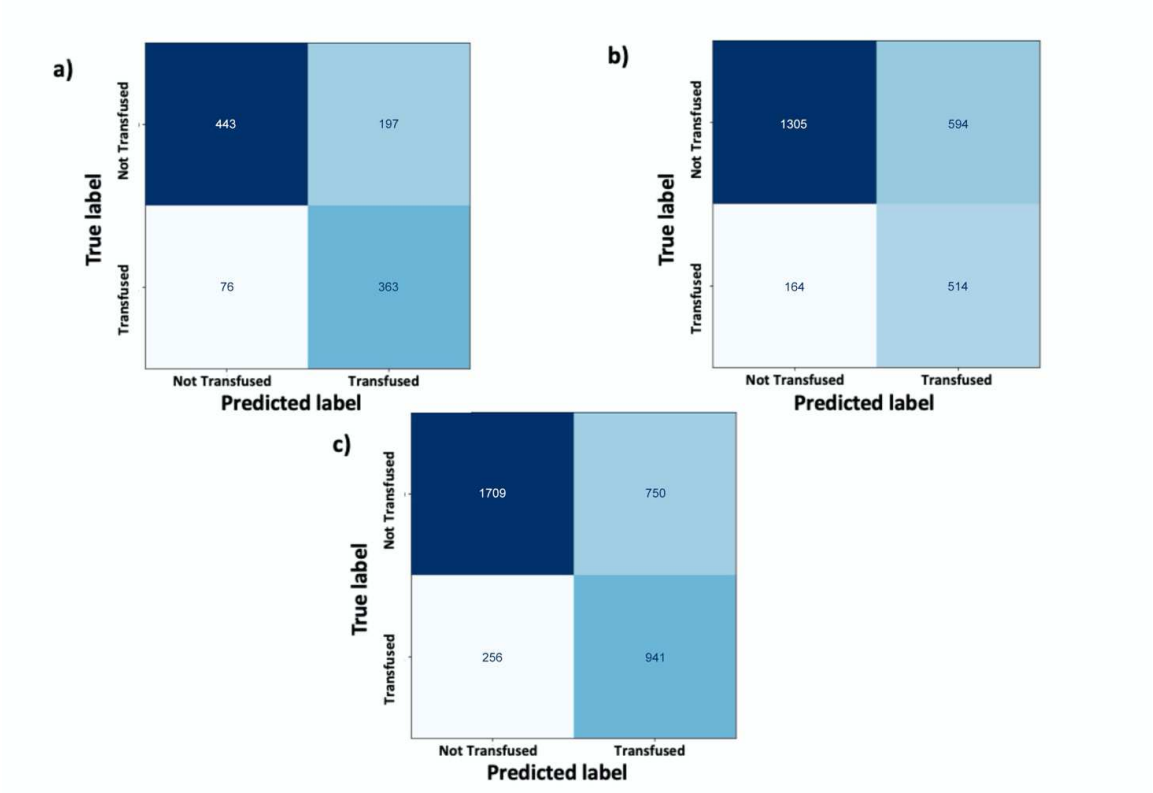

Figure 7 - Confusion matrices obtained when models are trained on a) the MIMIC-III training set, b) the eICU-CRD, and c) on the training set that contains both the MIMIC-III and the eICU-CRD.