OBUS – Multiple Gestation Evaluation Report

Summary

The Multiple Gestation (TWIN) model was evaluated on the extended FAMLI dataset (see [0.1 OBUS Data Description]). Owing to the scarcity of twin data in the FAMLI dataset, the data were split into a holdout (testing) set and a development set, which was further split into five parts for 5-fold cross-validation. All splits were done at the patient level. The 5-fold split of the development set permitted an assessment of model stability. Consistent performance of all 5-fold models on their respective validation folds, and on the testing set, is an indication of a stable model architecture and training process.

Single fold validation result of final model

The results shown here were based on training a model with data from Folds 0, 1, 2, and 4 and tuning with data from Fold3 of the 5-fold cross-validation development set ("Fold3 model"), using the final codebase and dataset v9.4. Figure 1 shows the results of this model on the Fold3 validation data.

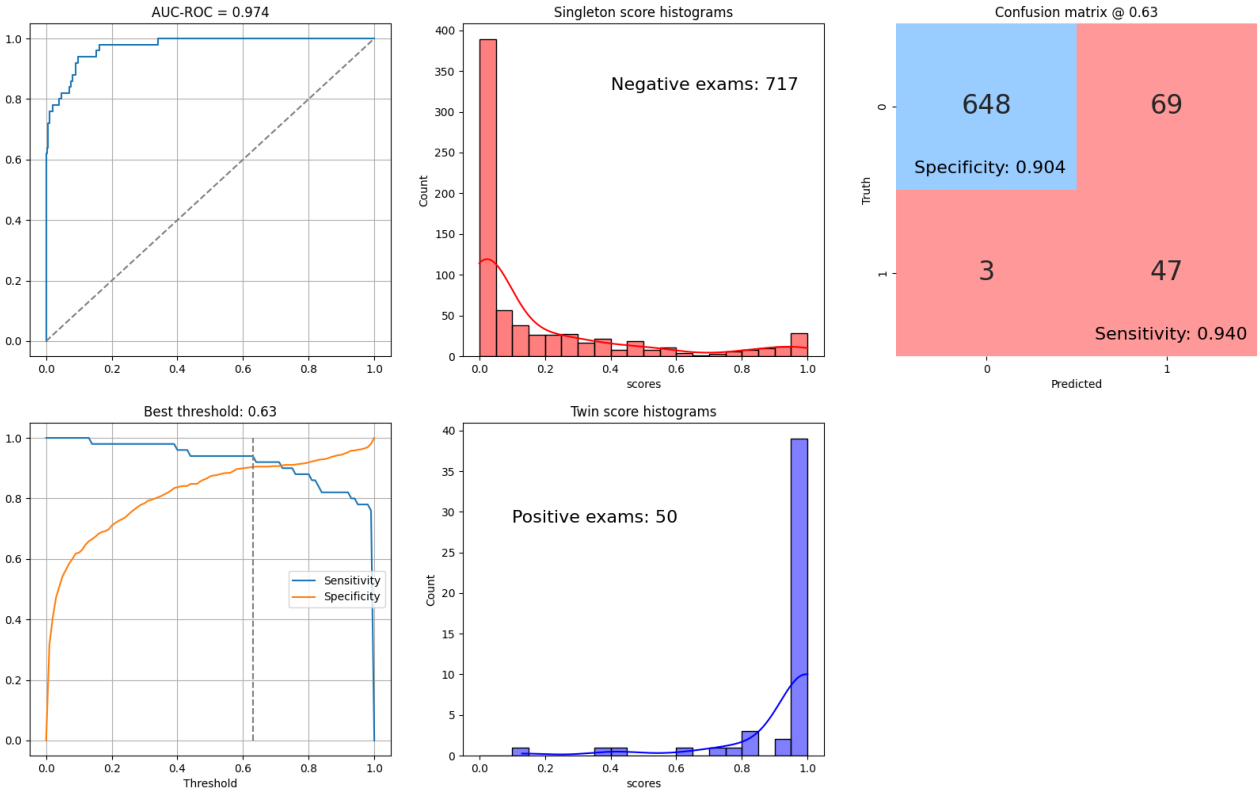

Figure 1. Fold3 model results on the Fold3 validation data.

Cross-validation and ensemble result on earlier model (trained with v9.2 dataset)

The full 5-fold cross-validation experiment was performed on an earlier v9.2 dataset; it was not repeated on v9.4 due to time constraints. Note that the test set between the two datasets did not change. This experiment showed relatively consistent performance between the 5 folds; some variation is expected due to the different exams in each validation set.

Fold	Sensitivity	Specificity	AUC
Fold0	0.839	0.947	0.933
Fold1	0.732	0.862	0.834
Fold2	0.907	0.947	0.970
Fold3	0.941	0.927	0.973
Fold4	0.868	0.960	0.900

Table 1: Validation set Sensitivity, Specificity, and AUC for each of the 5-fold experiments.

In addition to the five models resulting from the five-fold experiment, ensemble models can be used by combining the models. We explored a simple ensemble method of averaging the results from two or more of the 5-fold models, then applying a threshold. While this method would require longer inference time (as multiple models need to be run), it can result in better and/or more stable results.

By manually reviewing the sensitivity on the validation sets and the specificity on the validation and calibration sets, we selected a model consisting of an ensemble of the Fold3 and Fold4 models, and a threshold of 0.2. As a secondary option, in case the 2x inference time of an ensemble is prohibitive, we selected the Fold3 model as the single best model, also with a threshold of 0.2. Importantly, these decisions were made prior to evaluating performance on the test set. Sensitivity and Specificity vs. Threshold plots from the Fold3 model, on both the validation and calibration set, are shown in Figure 2.

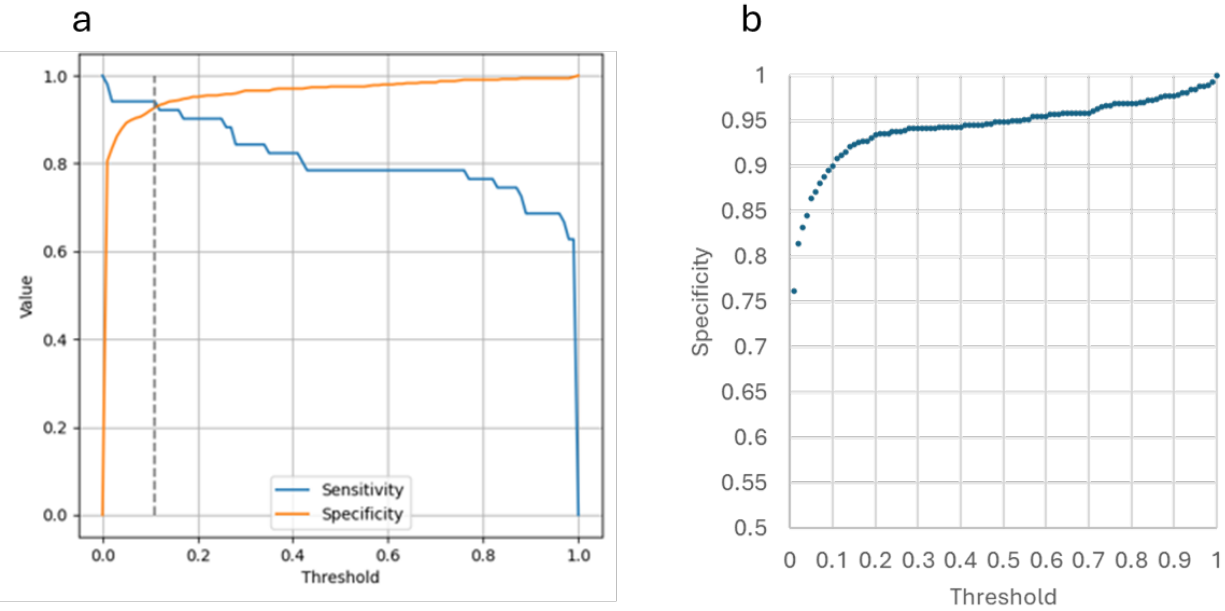

Figure 2: Plots of sensitivity and specificity of the v9.2 Fold3 model on (a) the Fold3 validation set and (b) the calibration set. Notes: The dashed line does not represent the selected threshold of 0.2. The y-axis of plot (b) goes from 0.5 to 1.

After selecting the models and hyperparameters, the models were evaluated using the test set. The first two rows of Table 2 show the performance of the 'selected' models. The additional rows are roughly ordered by how well they performed on the validation and calibration sets. As can be seen, the results of the different folds and ensembles is similar, with AUC varying between 0.930 and 0.961.

Fold(s)	**Threshold**	**AUC**	**Sens**	**Spec**
Ensemble 3+4, average scores	0.2	0.951	89.7%	94.6%
Fold 3	0.2	0.947	87.2%	94.6%
Fold 4	0.2	0.933	87.2%	95.7%
Ensemble 0+3+4, average scores	0.2	0.954	92.3%	94.8%
Ensemble 0+3+4, majority vote	From each fold	NA	84.6%	95.8%
Fold 0	0.1	0.942	82.1%	94.4%
Fold 2	0.25	0.945	87.2%	92.9%
Ensemble 0+2+3+4, average scores	0.2	0.961	92.3%	93.3%
Fold 1	0.3	0.930	87.2%	92.0%

Table 2: Performance of various models from the 5-fold experiment on the test set, with thresholds selected based on validation and calibration set results.