Technical Report 3:
Heart Failure Risk Prediction in the Smurf Society

Brainy Smurf Doctor Smurf Papa Smurf

In this third technical report, we describe how we combine the tabular features with the heart scans. We train a support vector machine (SVM) regressor on the combined dataset and significantly improve the prediction performance compared to our previous approaches relying solely on the tabular dataset.

Data pre-processing Regarding the tabular part, the pre-processing is the same as in technical reports 2 and 3. To integrate heart scans, we first remove image borders to avoid unnecessary noise (we consider 5 pixels-wide borders). We then flatten the cropped images to 1D vectors and finally concatenate them with the tabular features. These steps are illustrated in Figure 1. Both tabular and image features (pixel values) are standardized.

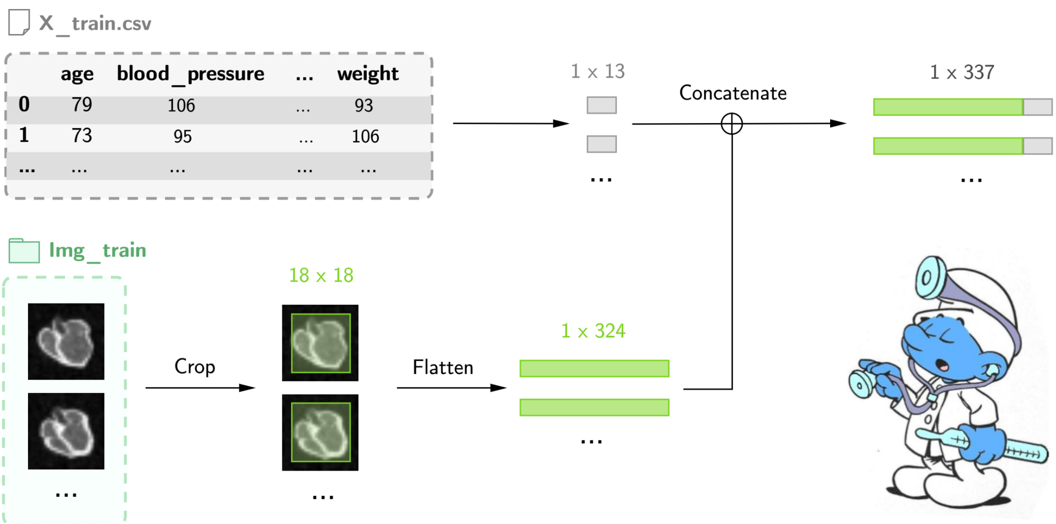

Figure 1: Combination of tabular features and image features.

Feature selection We keep the 9 tabular features that are the most correlated with the target (In technical report 2, we detailed a grid search procedure which outputted 9 as the best value for the number k of selected tabular features). Excluded feature are: "profession", "hemoglobin", "calcium" and "vitamin D". We also use correlation with the target to select most relevant pixels in images. The number p of selected pixels is treated as a hyper-parameter.

Model selection and implementation We use the `SVR` model from `scikit-learn`. We fix the "rbf" kernel as it consistently produced superior results in our previous experiments. In addition to p, we tune the regularization coefficient C, and the margin of tolerance ϵ using a grid search with a 5-fold cross validation . The search space is defined in Table 1.

Hyper-parameter	**Allowed values**
p	2, 4, 8, 16, 32, 64, 128
C	0.01, 0.05, 0.1, 0.5, 1
ϵ	0.01, 0.05, 0.1, 0.5, 1

Table 1: Search space for the grid search.

The best parameters are $\{p\text{: } 32,\ C\text{: } 0.5,\ \epsilon\text{: } 0.01\}$ (note that they are not the same as in technical report 2). The mean validation RSME for the best model is 0.050, with a standard deviation of 0.003.

Results analysis We retrain a model with the best set of hyper-parameters on the whole training set and then obtain a RMSE of 0.049 on the test set. This is in fact better than the validation score but can be explained either by luck, or because we retrain the model on a larger data sample.

We observe a significant improvement when using heart scans, suggesting that these images contain information not captured by other tabular features. To identify this information, we highlight the pixels selected from the heart scans (cf. Figure 2). These selected pixels form distinct regions within the images, although we are not yet certain what these regions represent or how to interpret them fully.

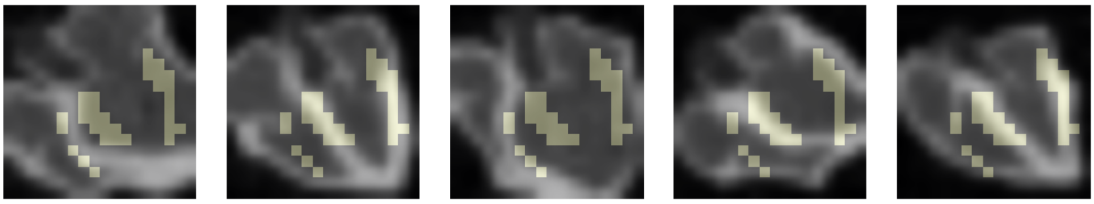

Figure 2: Most important pixels for heart failure prediction according to the proposed feature selection strategy. Selected pixels are highlighted in yellow. Gaussian interpolation is applied to images for visualization purposes.