TabPFNEnsemble: An ensembling approach for TabPFN

universität freiburg

Magnus Buehler [1] Giacomo Mossio [1] Sathya Kamesh Bhethanabhotla [1]

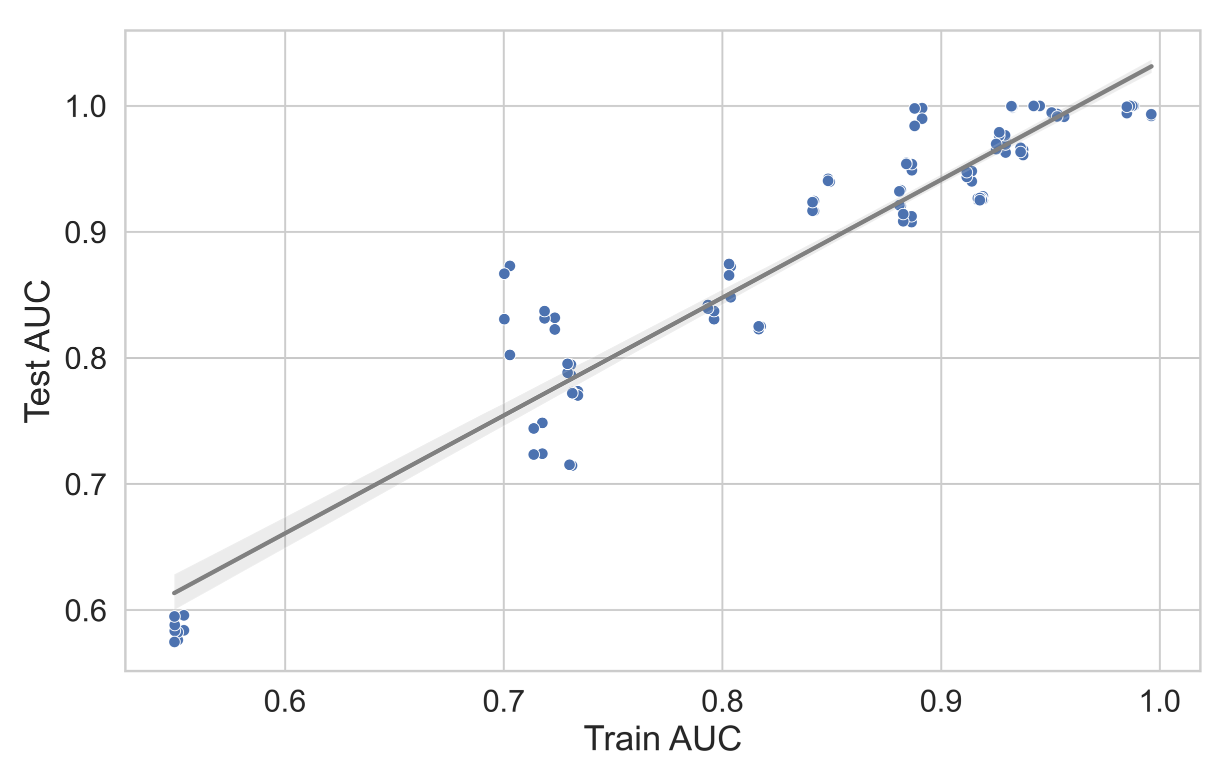
[1]University of Freiburg

Idea

We introduce an ensembling routine for TabPFN over both preprocessing methods and PFNs trained on different prior types and configurations.

Each data transformation and/or PFN is weighted based on its train data AUC performance to compute the final prediction. We evaluate and compare our approach across 30 datasets.

Train AUC vs. Test AUC across datasets

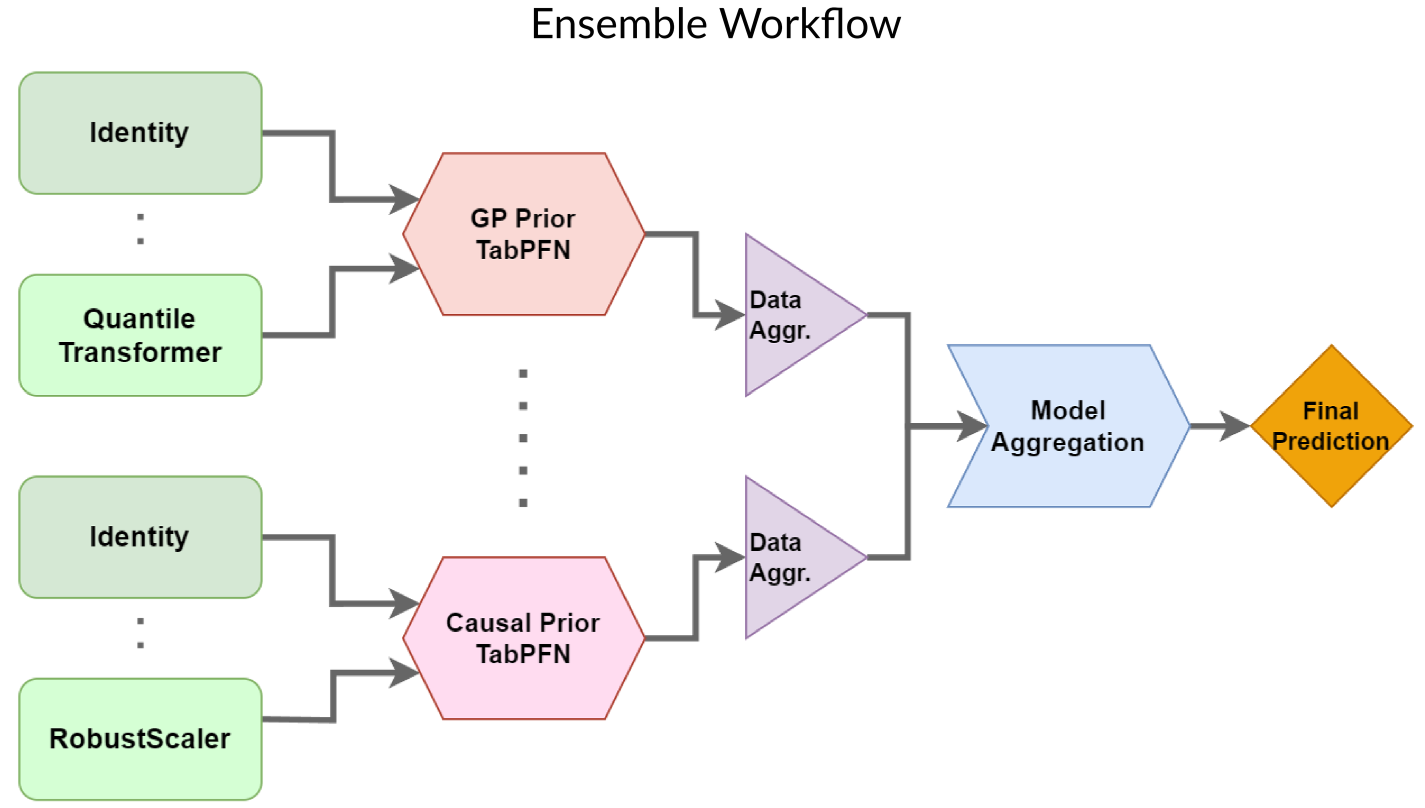


- The plot indicates significant correlations (Pearson correlation coefficient 0.96) in performance, implying that train AUC serves as a reliable estimator for test performance.
- Consequently, the ensemble is constructed such that the AUC over the train data is maximized

Building the Ensemble

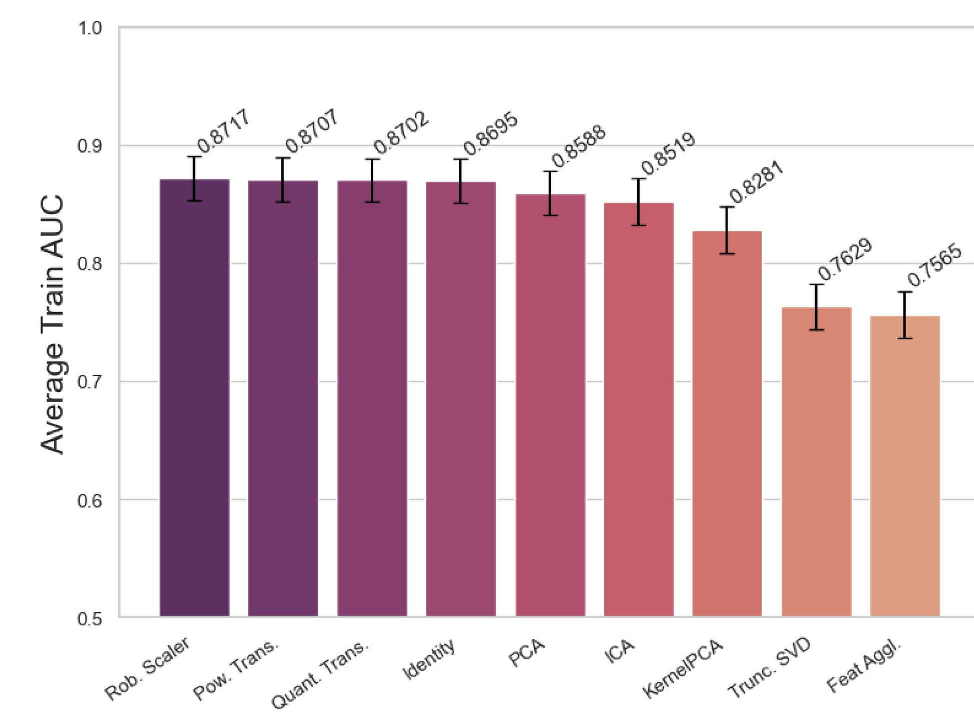
- Different data transformations are utilized as well as multiple PFNs, trained on different prior types.
- Predictions on unseen data are weighted across all transformations and/or across all PFNs using different weighting methods - (weighted average, single best performer)

Ensemble Workflow



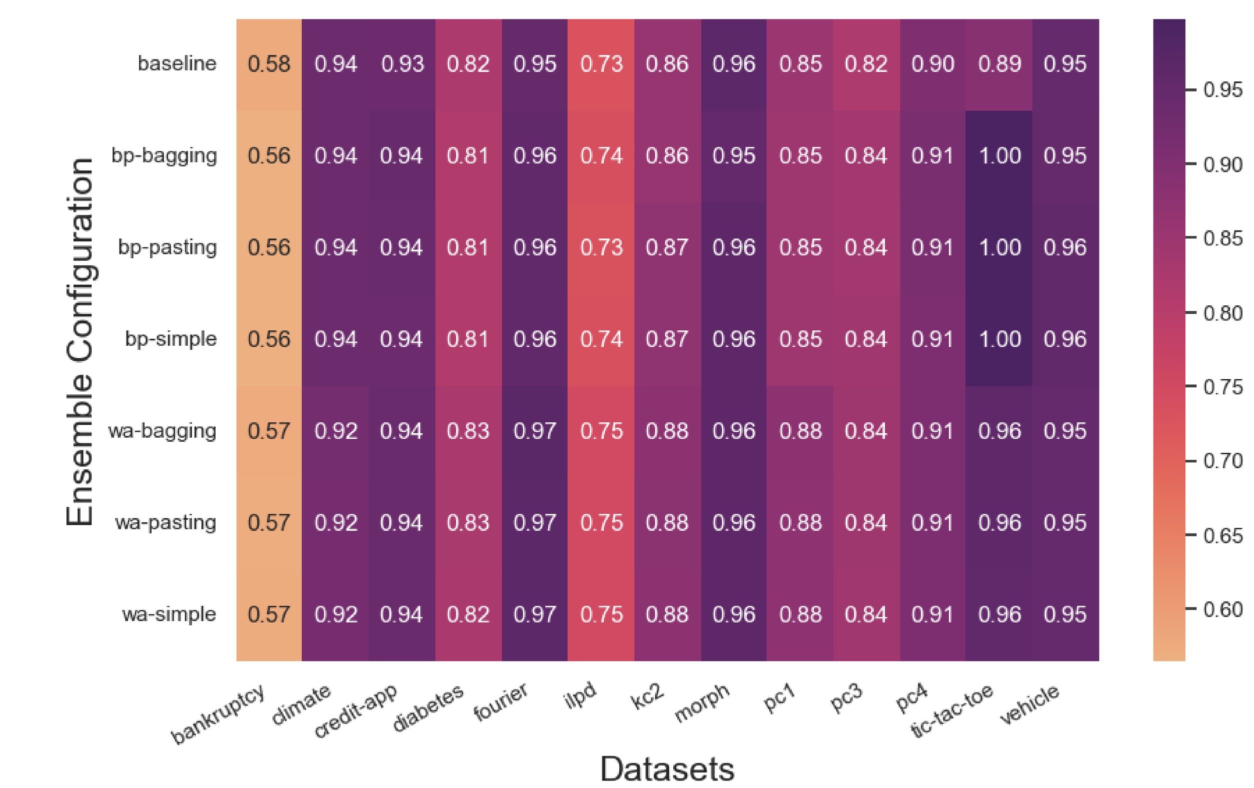
Only Data Ensembling

Average Train AUC - Data Transformations



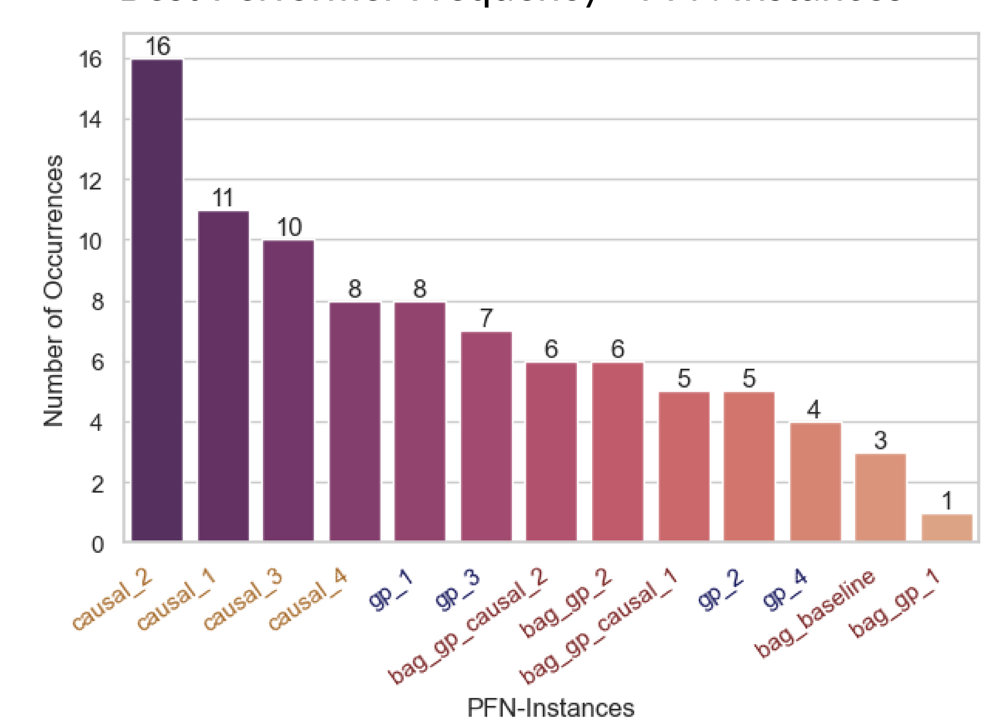
- Only data ensembling uses the baseline PFN in combination with a variety of data transformations.
- Identity, as well as PowerTransformer, QuantileTransformer and RobustScaler yield better results, whereas the performance decreases for FeatureAgglomeration, TruncatedSVD, and KernelPCA.
- Approach exhibits a general improvement over the baseline, with only marginal performance degradation observed for a limited subset of datasets and configurations.
- Over the 30 datasets, the "weighted average" approach is significantly better compared to the baseline on Test data AUC.

Test AUC - Data Ensembling



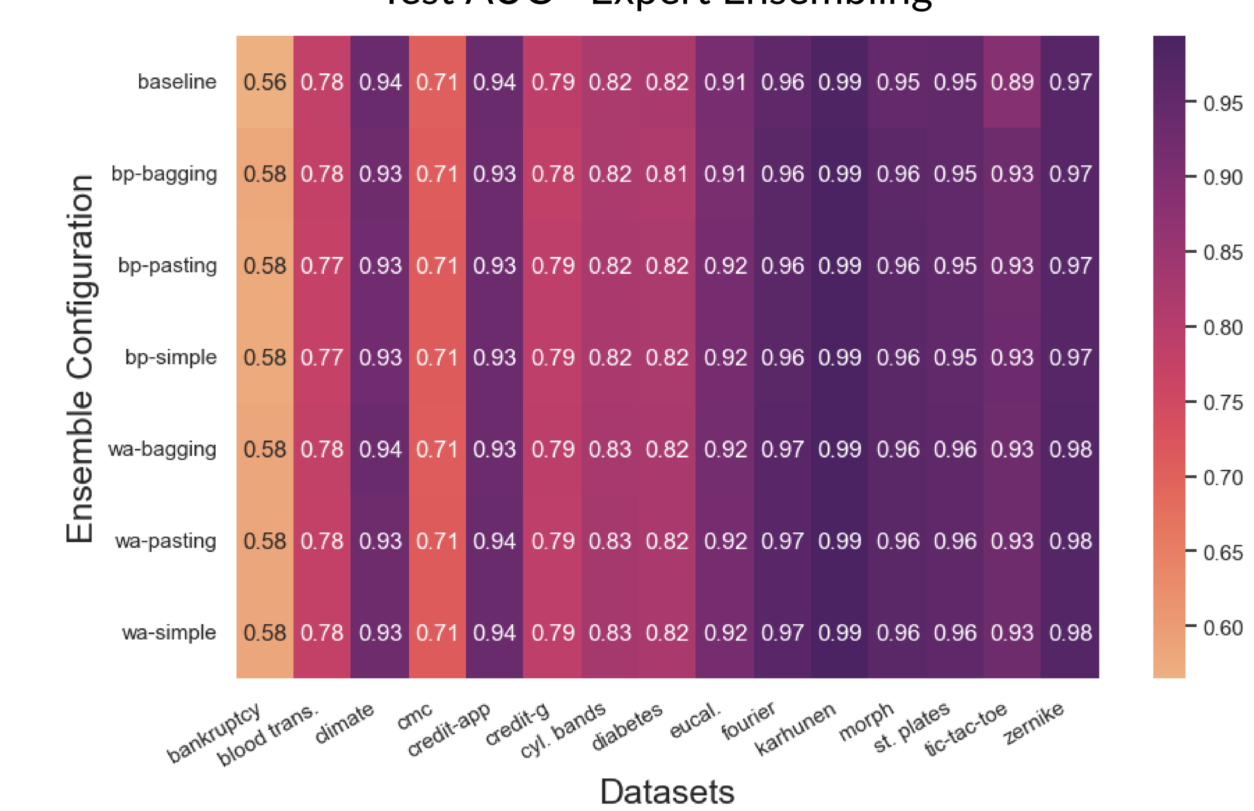
Only Expert Ensembling

Best Performer Frequency - PFN Instances



- Expert ensembling involves 14 PFNs, each trained on different priors, making each PFN an 'expert' for a specific underlying distribution.
- 4 PFNs were trained exclusively on Gaussian Process priors, 4 on structural causal model priors, and 6 PFNs using a combination of both.
- The ensembles with the single best performer configuration pick the SCM-based PFNs the most, whereas the baseline PFN was selected as the best performer only 3 times.
- None of the expert-only ensembling strategies showed significant performance differences in Test AUC compared to the baseline.

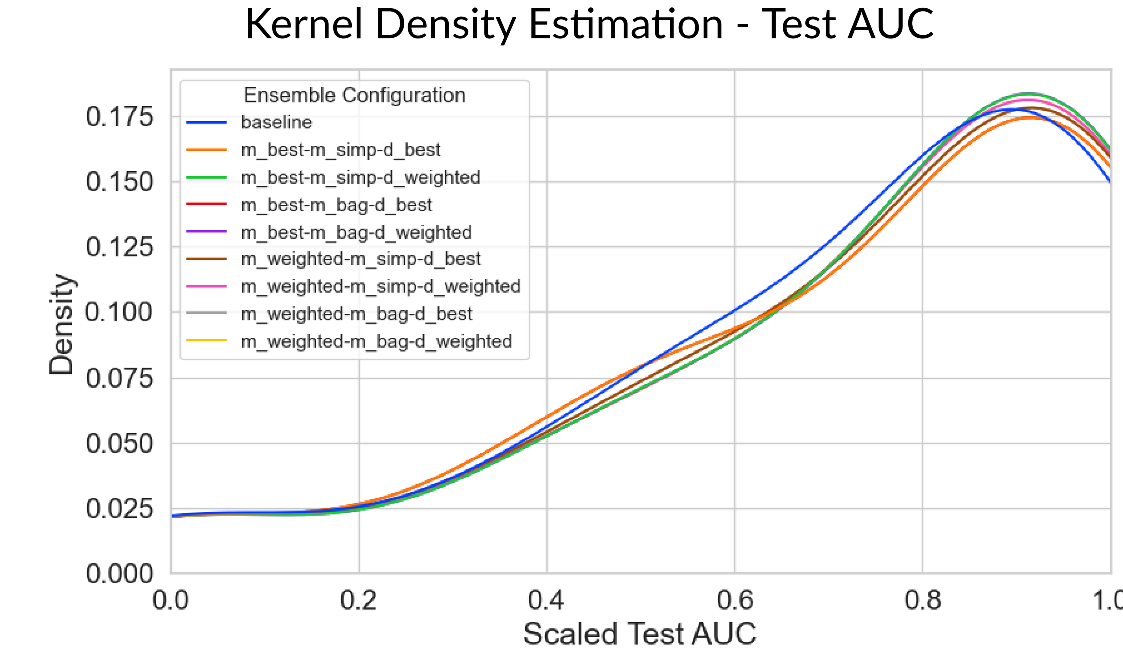
Test AUC - Expert Ensembling



Data And Expert Ensembling

- The combined evaluation method emsembles over data transformations followed by ensembling over multiple PFNs.
- We observe that the configurations with the weighted averages over data transforms yield significantly better AUC scores compared to the baseline.
- The kernel density estimates for some of the Ensemble Test AUCs have a visibly higher mean and are shifted more towards optimal AUC performance.

Kernel Density Estimation - Test AUC



Conclusions and Further Work

- AUC is a very effective metric which can be maximized during the forward pass to yield better generalization performance
- Data preprocessing in combination with weighted averaging yields significantly better results
- Exploring additional weighting techniques and data transformation methods is a promising path forward
- Examine the construction of new prior types

bhethans@tf.uni-freiburg.de, mossiog@tf.uni-freiburg.de, buehlema@tf.uni-freiburg.de