TabSTAR

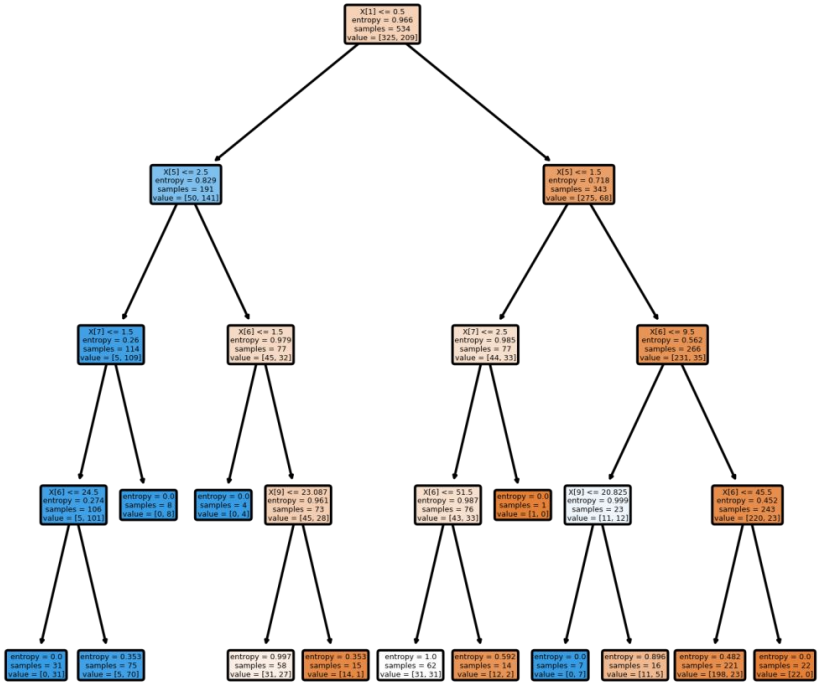
A Tabular Foundation Model for Tabular Data with Text Fields

Alan Arazi Eilam Shapira Roi Reichart

Technion - IIT

NEURAL INFORMATION PROCESSING SYSTEMS 2025

Tabular Data: Deep Learning is NOT all you need?

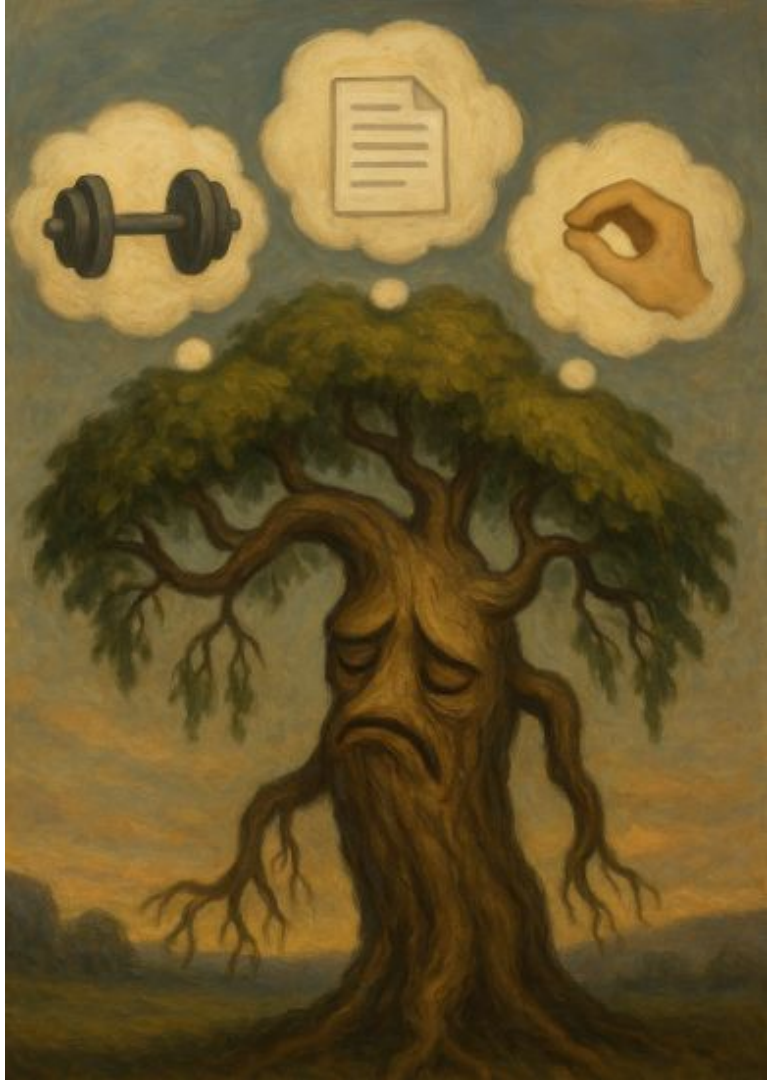


[Tabular Data: Deep Learning Is Not All You Need](#) (Shwartz-Ziv, Information Fusion 2021)

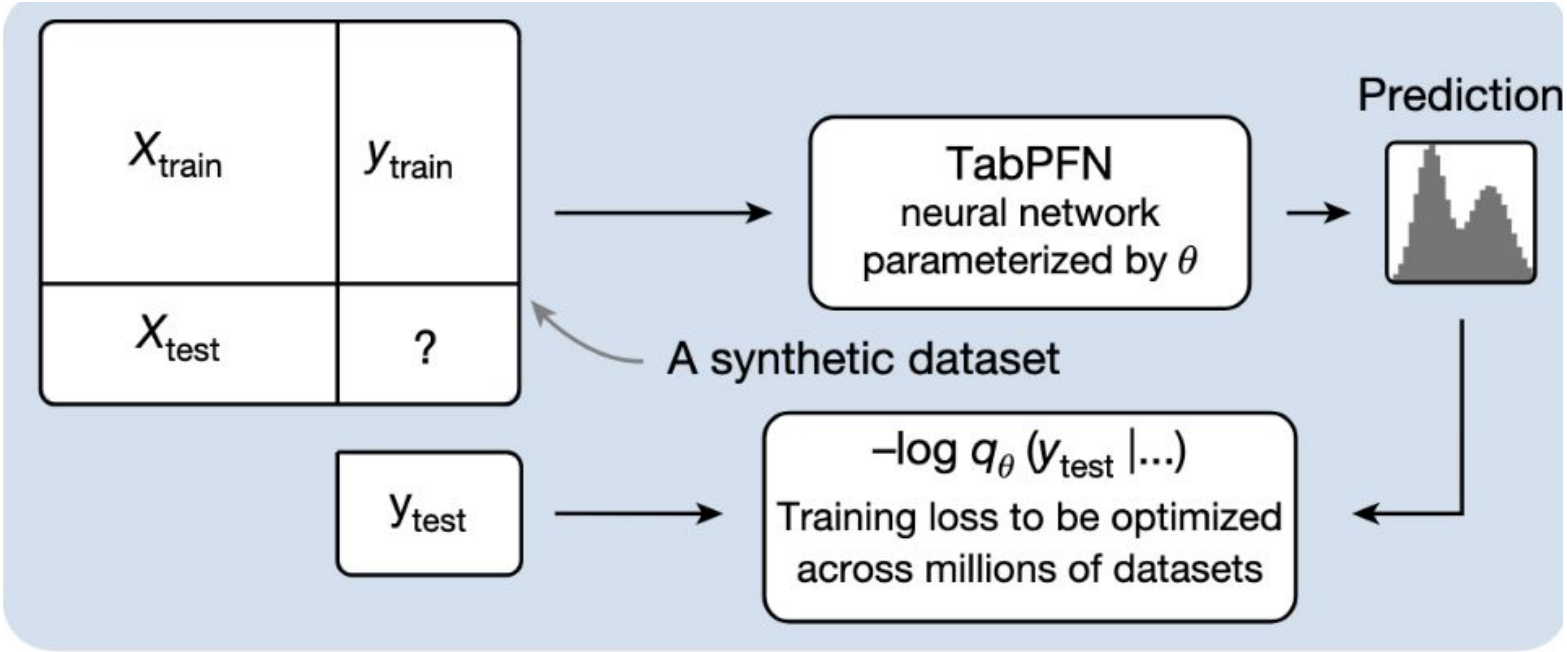
Trees aren’t perfect

- No pretraining

- Don’t support free text features



TabPFN-v2: The Rise of Tabular Foundation Models

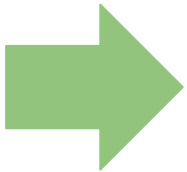


Why Tabular Foundation Models Should Be a Research Priority (van Breugel, ICML 2024)

Accurate predictions on small data with a tabular foundation model (Hollmann, Nature 2025)



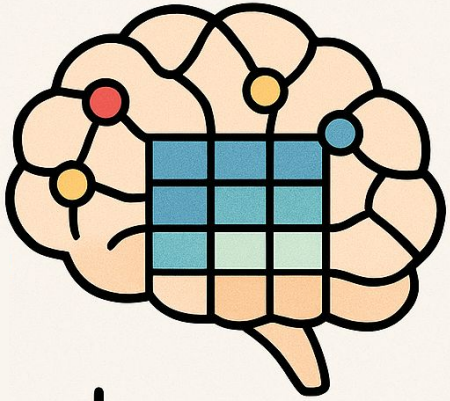

Age	Department	Report
45	Cardiology	Mild chest discomfort…
62	Neurology	Complaints of headache…
38	Oncology	Completed treatment cycle…
55	Neurology	Reports episodes of vertigo…



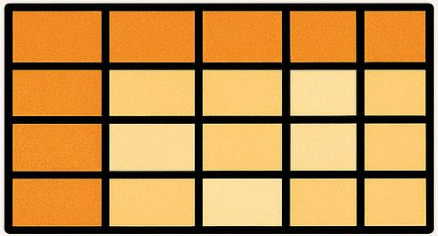
Decision
Released
Hospitalized
Released
Hospitalized

TabSTAR: Semantically Target-Aware Representations

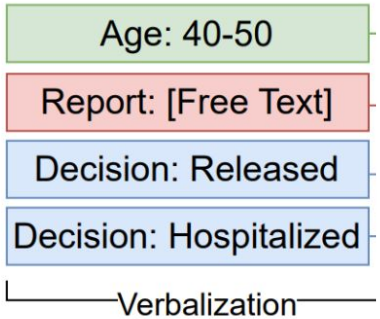
PRETRAINING



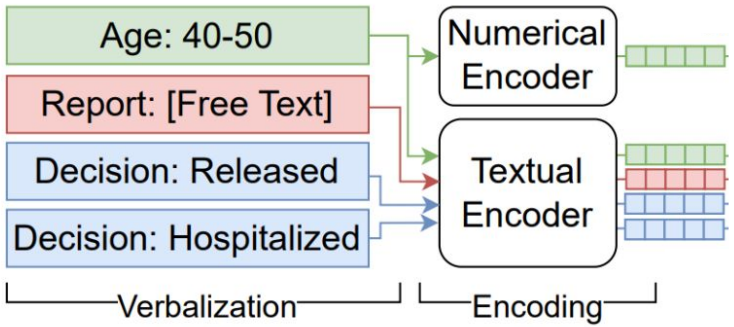
DOWNSTREAM



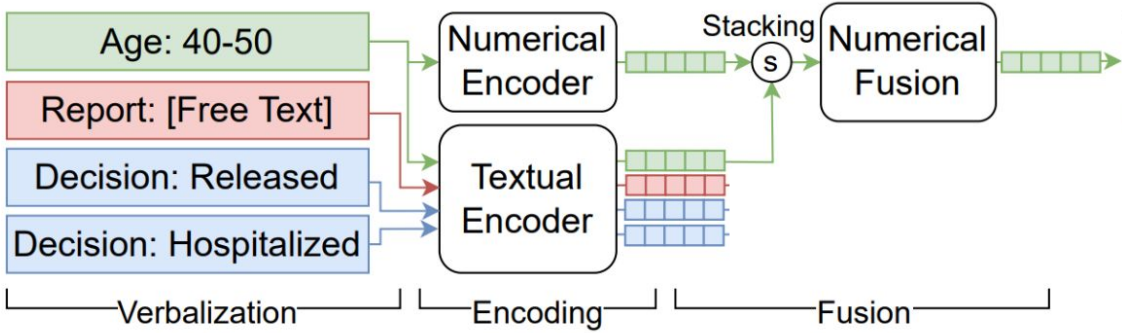
Verbalization 💬



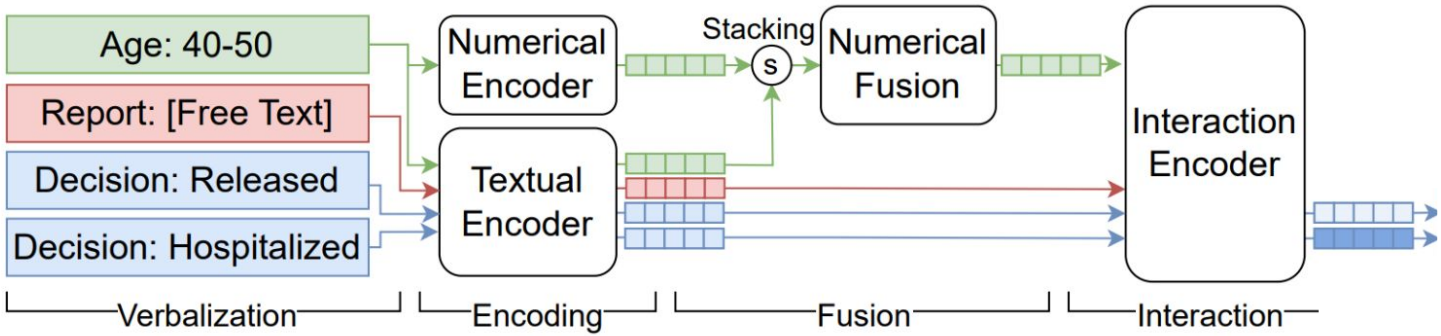
Encoding 🔢


Age: 40-50
Report: [Free Text]
Decision: Released
Decision: Hospitalized
Numerical Encoder
Textual Encoder
Verbalization
Encoding

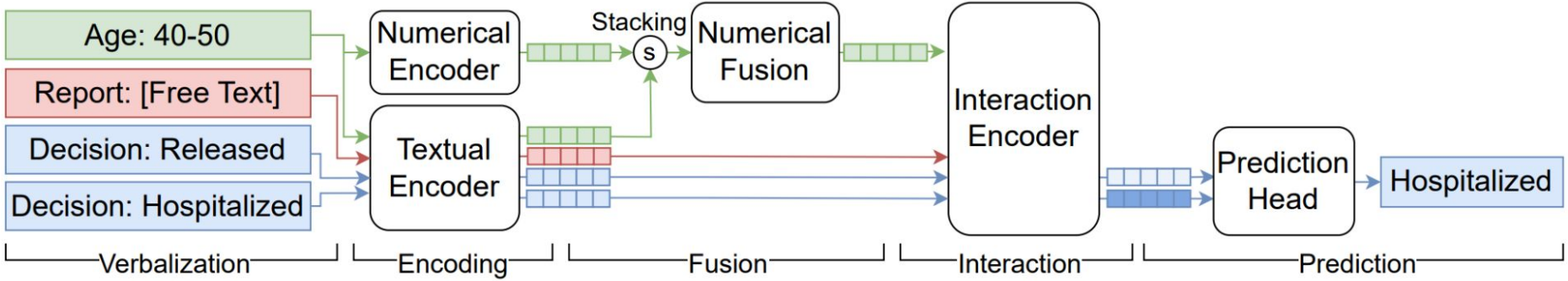
Fusion ☯️

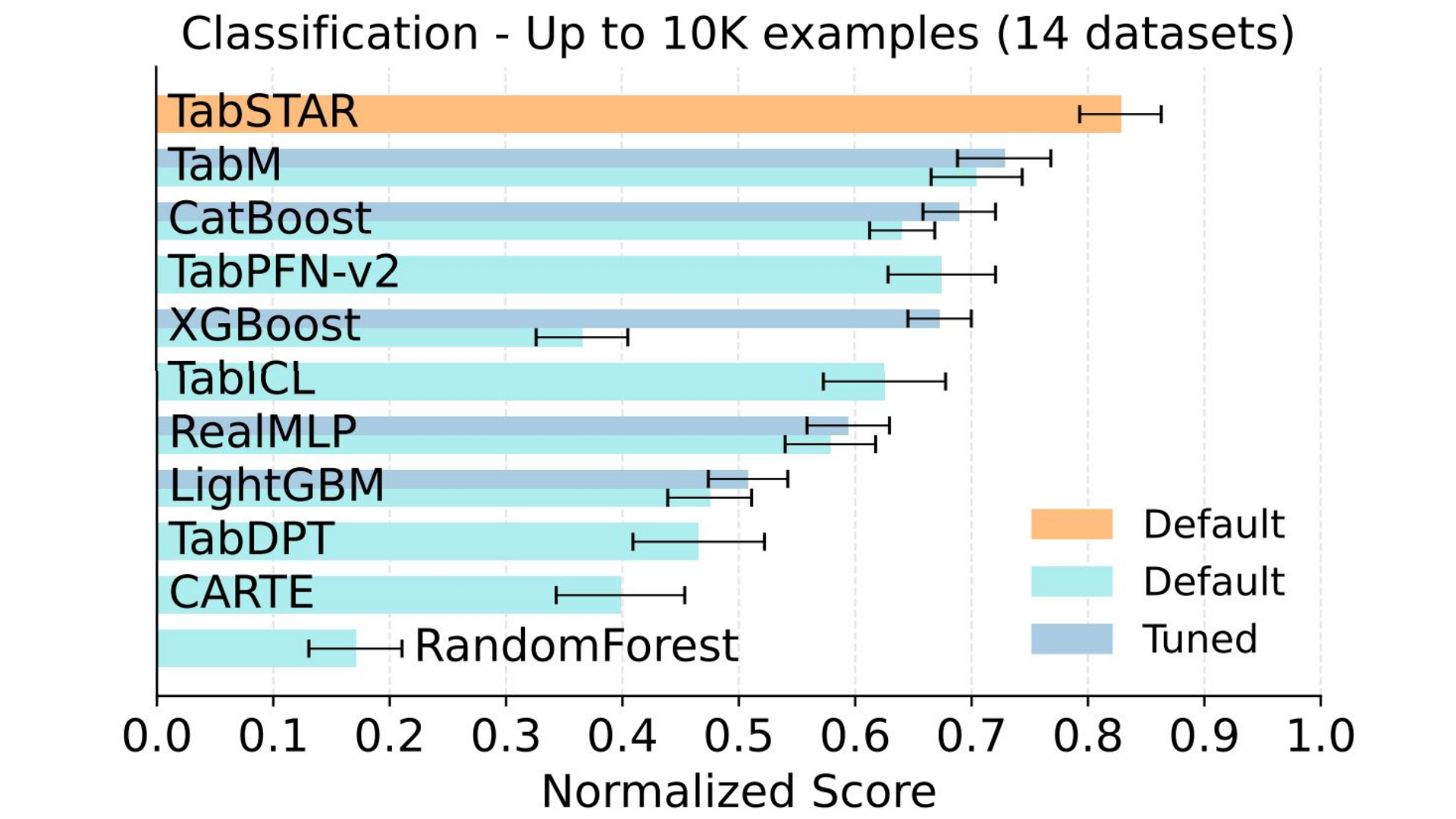

Age: 40-50
Report: [Free Text]
Decision: Released
Decision: Hospitalized
Numerical Encoder
Textual Encoder
Stacking
s
Numerical Fusion
Verbalization
Encoding
Fusion

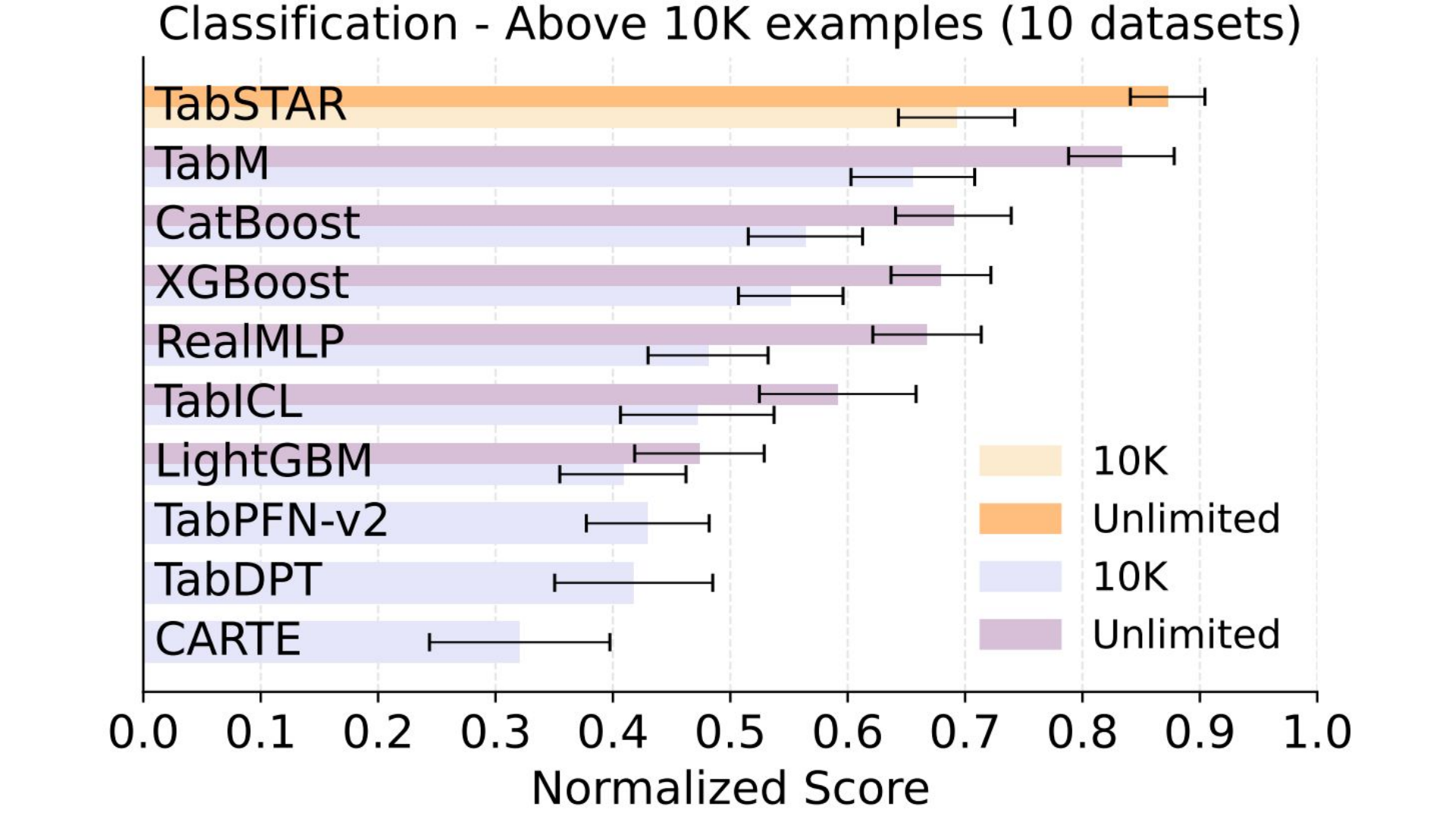
Interaction 🤗



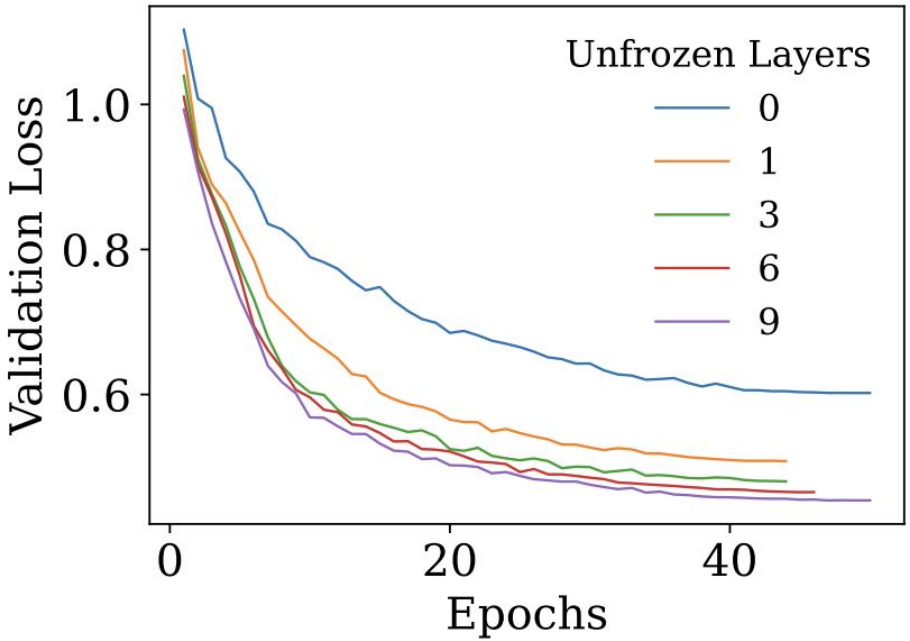
Prediction 🔮

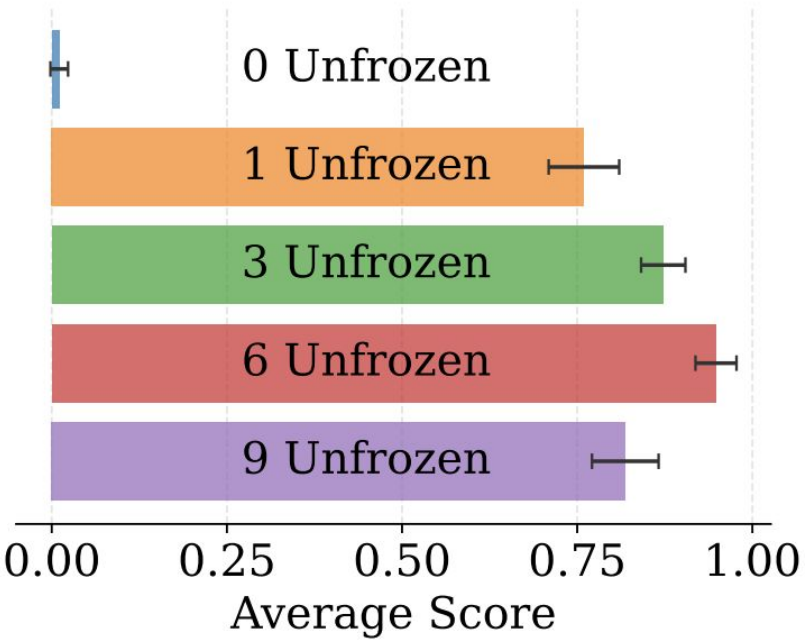

Age: 40-50
Report: [Free Text]
Decision: Released
Decision: Hospitalized
Numerical Encoder
Textual Encoder
Stacking
s
Numerical Fusion
Interaction Encoder
Prediction Head
Hospitalized
Verbalization
Encoding
Fusion
Interaction
Prediction


Classification - Up to 10K examples (14 datasets)
TabSTAR
TabM
CatBoost
TabPFN-v2
XGBoost
TabICL
RealMLP
LightGBM
TabDPT
CARTE
RandomForest
Default
Default
Tuned
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Normalized Score

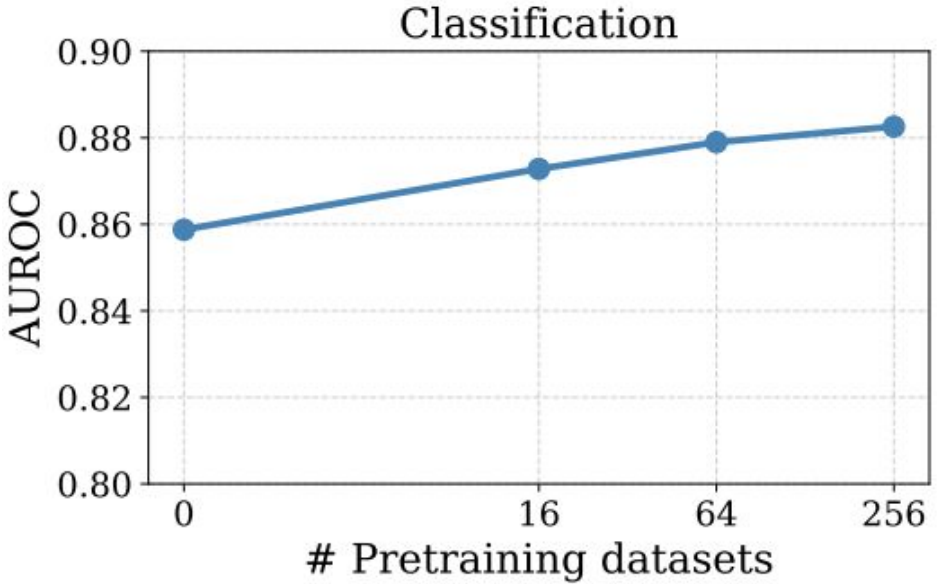

Classification - Above 10K examples (10 datasets)
TabSTAR
TabM
CatBoost
XGBoost
RealMLP
TabICL
LightGBM
TabPFN-v2
TabDPT
CARTE
10K
Unlimited
10K
Unlimited
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Normalized Score

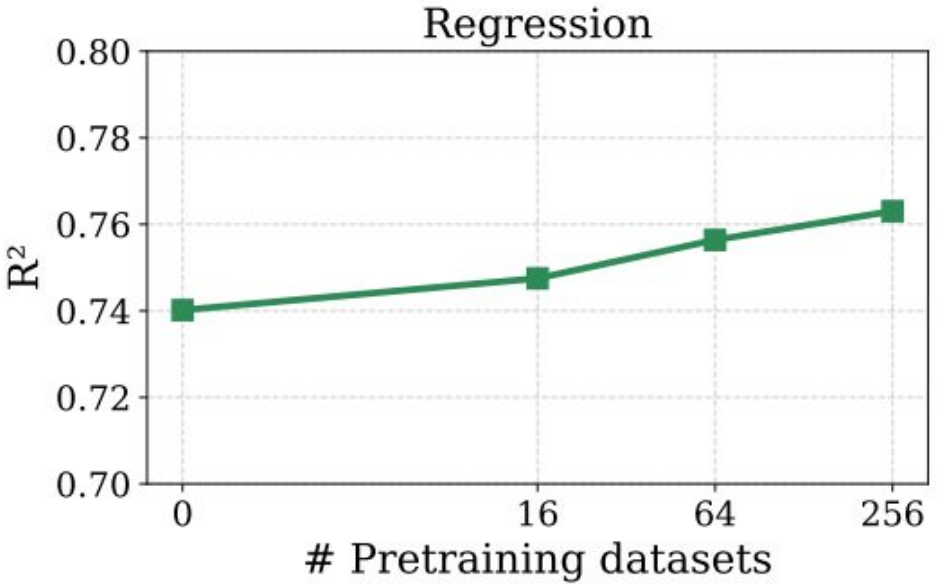
End-to-end learning is the key to TabSTAR's success





Scaling Laws with the number of pretraining datasets





Future Work: TabSTAR-v2?

```python
from tabstar.tabstar_model import TabSTARClassifier

tabstar = TabSTARClassifier()
tabstar.fit(x_train, y_train)
```

https://github.com/alanarazi7/TabSTAR