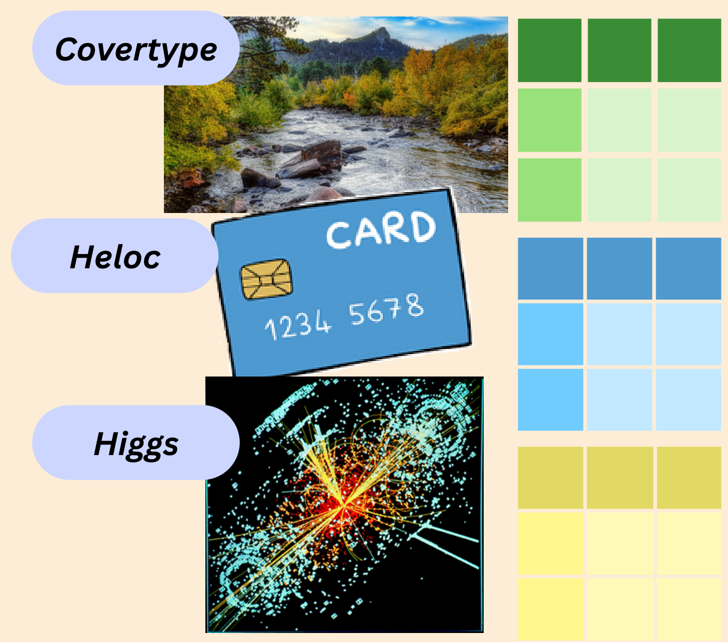
https://github.com/DjamelZair/UnifiedTabularLearning.git

UNIFIED TABULAR LEARNING

Can a multi-task classification model be trained effectively on heterogeneous tabular datasets with substantial inter-dataset size imbalance?

Djamel Zair (11015934)
Lisa van Oosten (13129236)
Rinske Oskamp (15817997)
Group 22
UvA - Applied Machine Learning 2025

3 HETEROGENEOUS TABULAR DATASETS



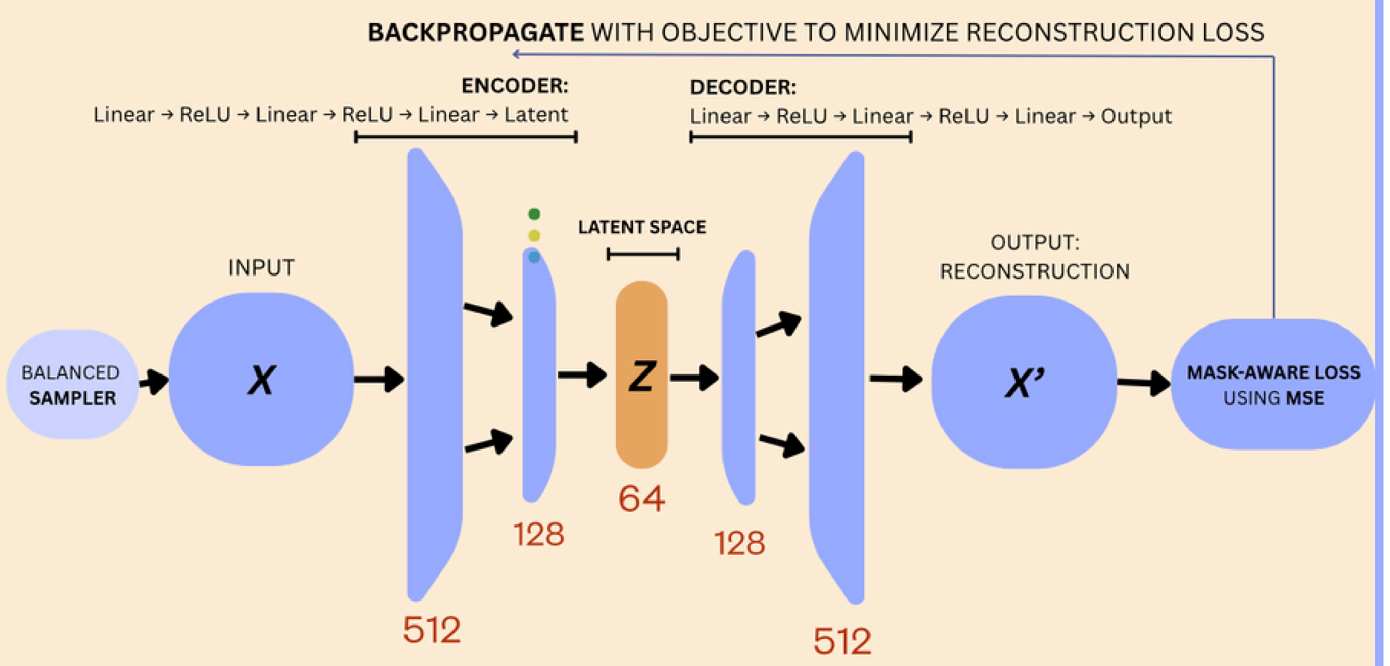
Challenges:
Very different feature spaces
Sample size imbalance
Class imbalance
One classification model

Our solution:
Make shared embeddings such that all datasets share the same representation space, on which a classifier is trained once

OUR PIPELINE:

AUTOENCODER FOR SHARED REPRESENTATIONS

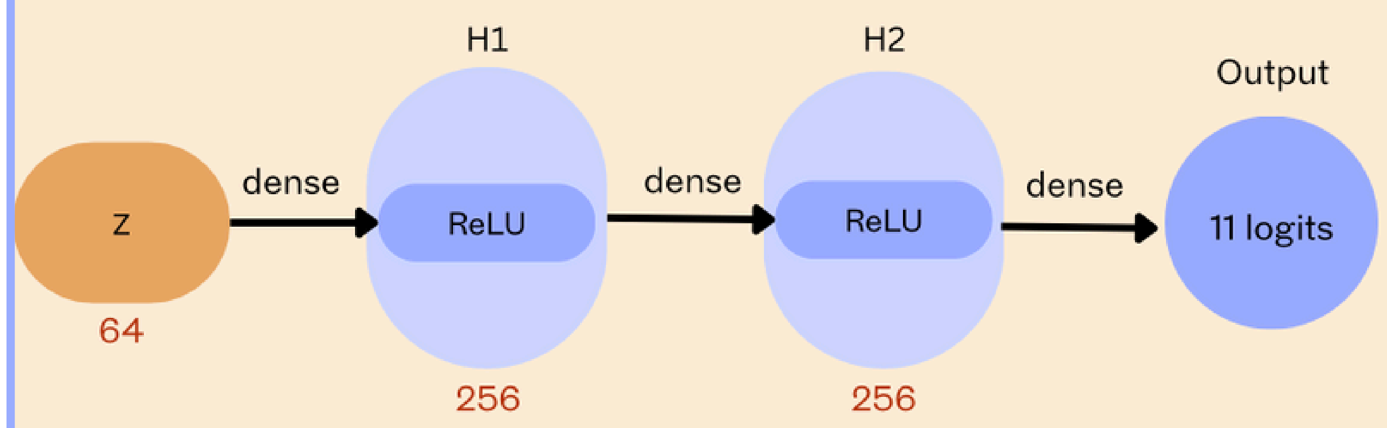
WHY? We wanted to make one shared representation capturing structure across our tables while still letting the model specialize per dataset.



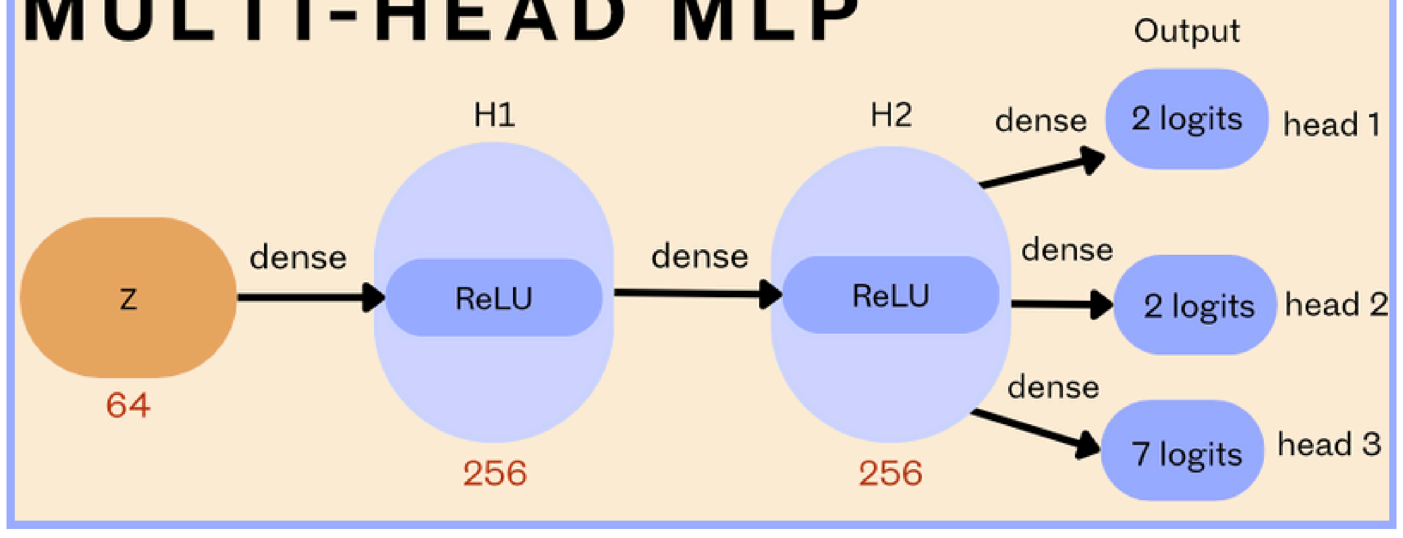
HOW? By making a self-supervised autoencoder to compresses our merged data into a latent vector to learn a unified embedding for classification

CLASSIFICATION *EXPERIMENT*

SINGLE-HEAD MLP



MULTI-HEAD MLP



BASELINE **TABPFN-2.5** (PRIOR LABS)

Tabular Prior-Data Fitted Network. **Foundation model for tabular data.**
Intended use: classification tasks with ≤50 000 samples and ≤2000 features.
The model is pre-trained on millions of synthetic tabular datasets to learn patterns that are common tabular datasets

PERFORMANCE

	Covertype			Heloc			Higgs		
	A	R	P	A	R	P	A	R	P
Baseline	0.660	-	-	0.833	-	-	0.735	-	-
MLP (Single)	0.466	0.719	0.445	0.740	0.735	0.747	**0.740**	0.784	0.762
MLP (Multi)	0.811	0.806	0.741	0.656	0.655	0.655	**0.795**	0.813	0.784

The most important measure differs:
For predicting covertype we want high accuracy, we do not wat to mislabel frauds, and do not want to mis higgs bosons signals

Beating the baseline

TRAINING STRATEGY

- ✓ **Stratisfied Test Split**
- ✓ **Leak-tight test evaluation**
- ✓ **K-Fold CV for HP tuning**
- ✓ **Balanced sampler & class-weighted loss**

Cross-Entropy Loss

HYPERPARAMETERS

- **DROPOUT**
- **WEIGHT DECAY**
- **LEARNING RATE**
- **BATCH SIZE**

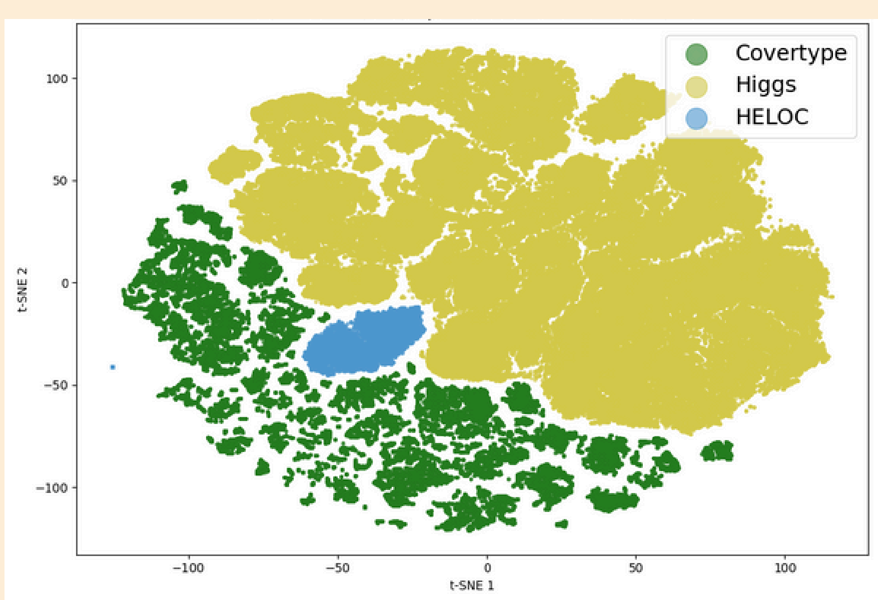
ADAM OPTIMIZED

ANALYSIS OF COMPLEXITY

Aspect	Our Model	TabPFN
Architecture	Deep autoencoder + small MLP classifier	Transformer based probabilistic foundation model
Pre-training cost	None	Extremely high pre-training cost on large GPU clusters
User training cost	Moderate training cost	None
Scaling	Scales normally for HIGGS-sized data	Not made for datasets with more than 50.000 samples
Time Complexity	$T \propto N(E(ae)D+E(mlp)L)$	$T \propto 10^7$

INTERPRETATION

t-SNE Latent Space



Most important features:

Covertype
Geospatial gradients
HIGGS
Physical event geometry
HELOC
Creditworthiness trajectories

CONCLUSION

- ★ **COMPARABLE ACCURACY TO TABPFN**
- ★ **HANDLES MORE THAN 50.000 SAMPLES**
- ★ **POST HOC INTERPRETABILITY**
- ★ **NOT FUTURE DOMAIN-AGNOSTIC**: LIMITED GENERALIZATION

YES!

SOURCES

- Hollmann, N., Müller, S., Eggensperger, K., & Lindauer, M. (2022). TabPFN: A transformer that solves small tabular classification problems in a second. arXiv. https://arxiv.org/abs/2207.01848
- Erickson, N., Purucker, L., Tschalzev, A., Holzmüller, D., Desai, P. M., Salinas, D., & Hutter, F. (2025). TabArena: A living benchmark for machine learning on tabular data. In Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS). https://huggingface.co/spaces/TabArena/leaderboard