Effective Distillation for Tabular Datasets

Inwon Kang [1] Parikshit Ram [2] Yi Zhou [2] Horst Samulowitz [2] Oshani Seneviratne [1]

[1]Rensselaer Polytechnic Institute [2]IBM Research

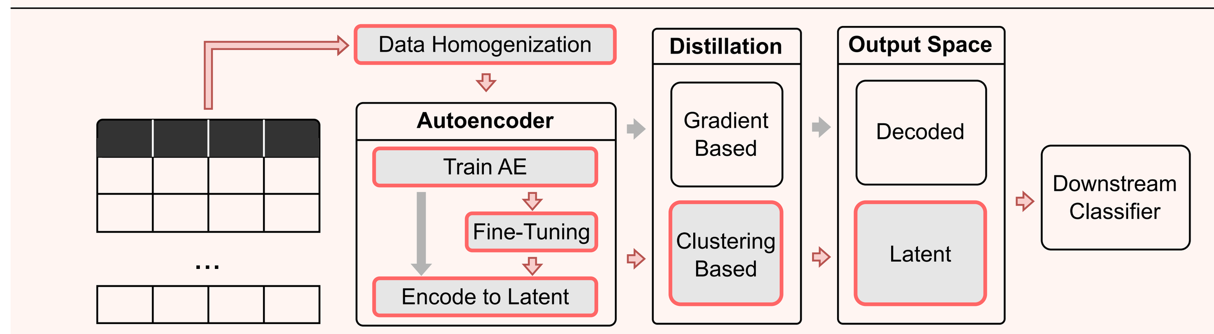
What is Data Distillation?

Data Distillation is **the task of reducing a large dataset into a smaller dataset**. The goal is to have a classifier trained on the smaller dataset perform comparably to a classifier trained on the full dataset. The idea has been proposed and studied for mainly image datasets [1, 2].

Why Tabular Data?

- Tree-based classifiers tend to outperform NN-based models on tabular data [3].
- Tree-based classifiers cannot benefit from incremental training the same way as NN-based models.
- One-hot representation can lead to a blow-up in feature size.

Proposed Approach



- Considered **model-agnostic** pipelines that uses an autoencoder architecture for latent representation of the data.
- Compared the efficacy of different components by measuring the performance of the downstream classifier trained on the distilled dataset.

Naive	Random Sampling
Image Distillation	*Kernel Inducing Points(KIP)* [1]

Table 1. Baselines considered.

Method	Description
Autoencoder	None / Vanilla / Supervised-FT
Distillation Method	K-Means / Agglomerative / KIP
Centroid Method*	Mean / Nearest
Output†	encoded / decoded

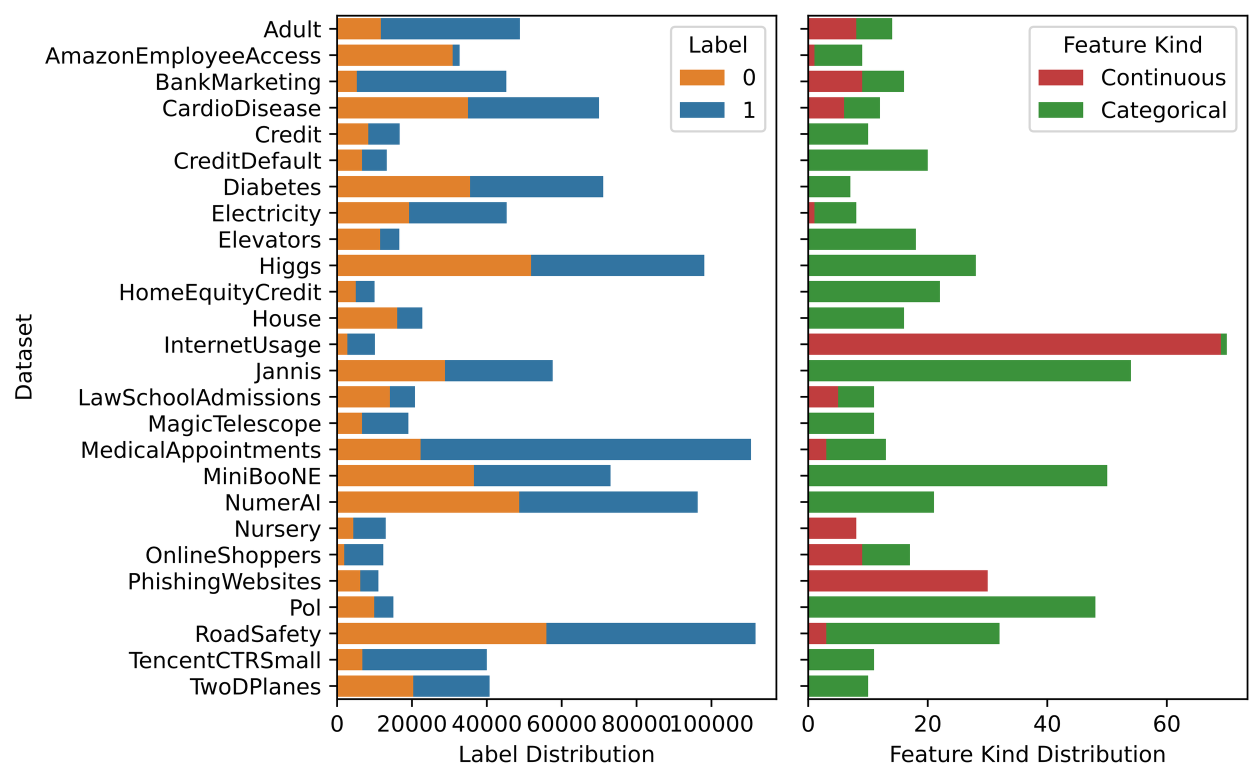
Table 2. Hyperparameters considered for distillation pipelines.
*: Only applicable to clustering-based methods.
†: Only applicable when autoencoder is used.

Experiment Details

- Downstream classifiers: **XGBoost**; **MLP**; **Logisitc Regression**; **Naive Bayes** and **Nearest-Neighbors**.
- Consider distill size $N \in [20, 200]$.
- Random iterations are repeated 5 times.
- Total number of pipelines including baseline: 76.

Datasets

Considered 26 datasets with more than 100,000 rows and 10 columns from OpenML(`openml.org`).



Autoencoders

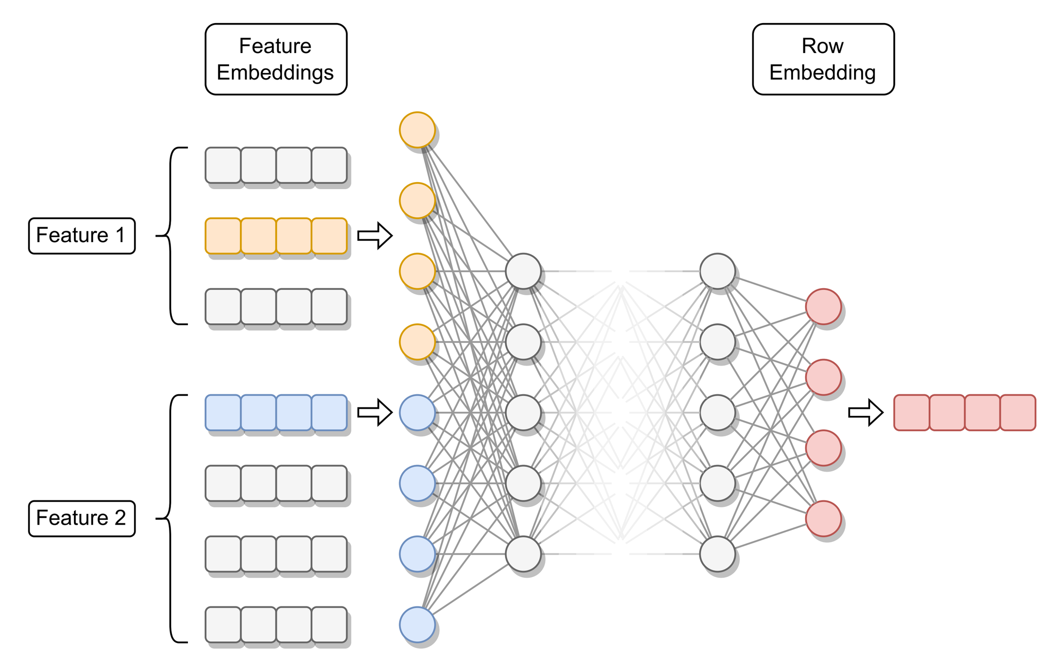


Figure 1. MLP Architecture

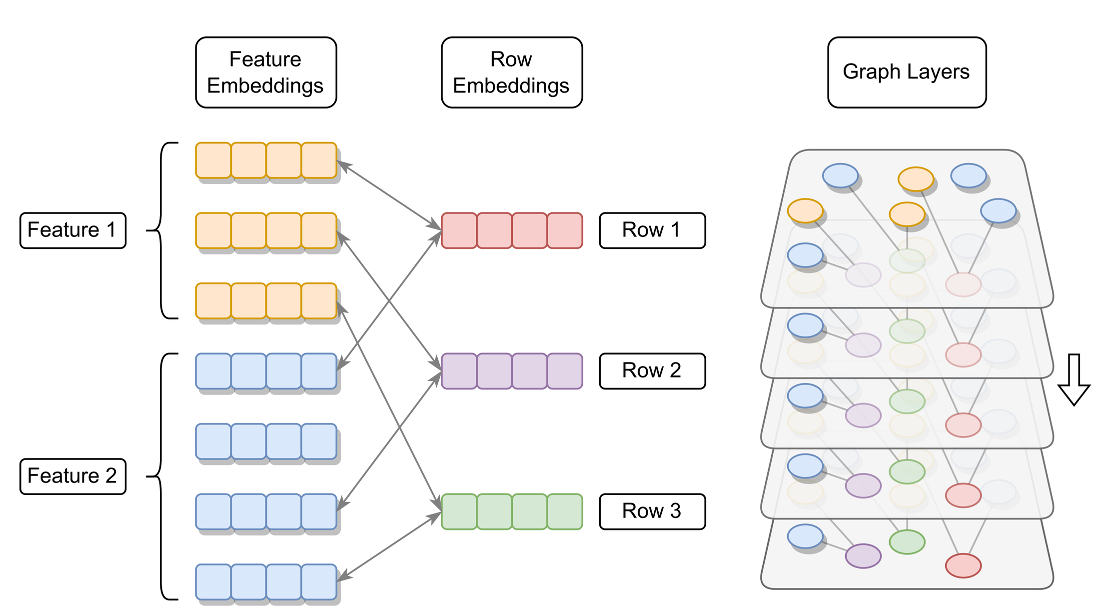


Figure 2. GNN Architecture [4]

Training Objectives:

$$\mathcal{L}_{tabular} = \frac{1}{n}\sum_{i=1}^{n}\left(-\frac{1}{\log(c_i)}\sum_{j=1}^{c_i} x_{i,j}\log(\hat{x}_{i,j})\right) \quad (1)$$

$$\mathcal{L}_{supervised} = \mathcal{L}_{tabular}(x, \hat{x}) + \alpha\mathcal{L}_{ce}(y, \hat{y}) \quad (2)$$

What is the Effect of Supervised Fine-Tuning?

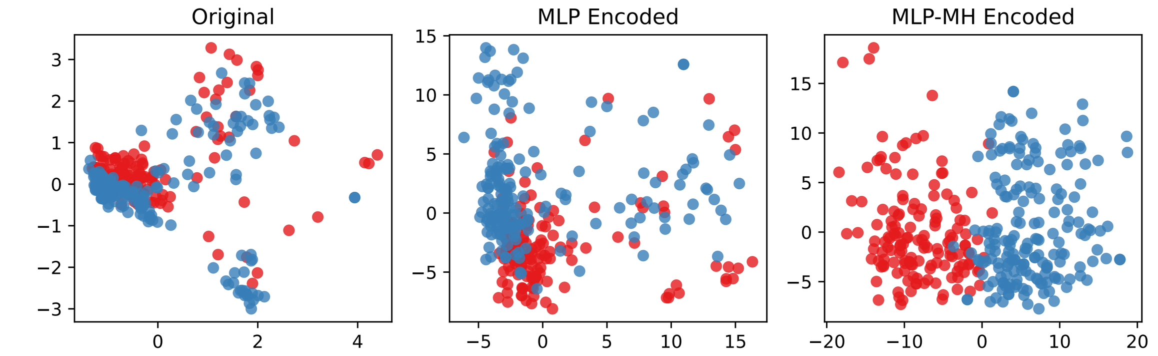


Figure 3. Phishing Website Dataset

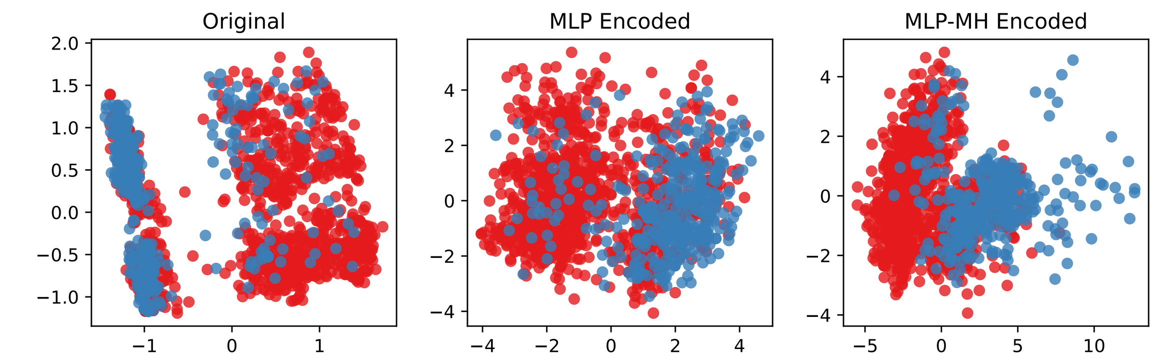


Figure 4. Adult Dataset

Model	Reconstruction	SFT-Reconstruction	SFT-Classification
FFN	**0.9616±0.0768**	0.9570±0.0804	**0.7937±0.1419**
GNN	0.9608±0.0737	**0.9585±0.0773**	0.7909±0.1374

Table 3. Performance comparison of autoencoder architectures.

- SFT does not degrade the reconstruction performance of the decoder.
- SFT results in label-aware encodings in the latent space.

Which Encoder Leads to Better Performance?

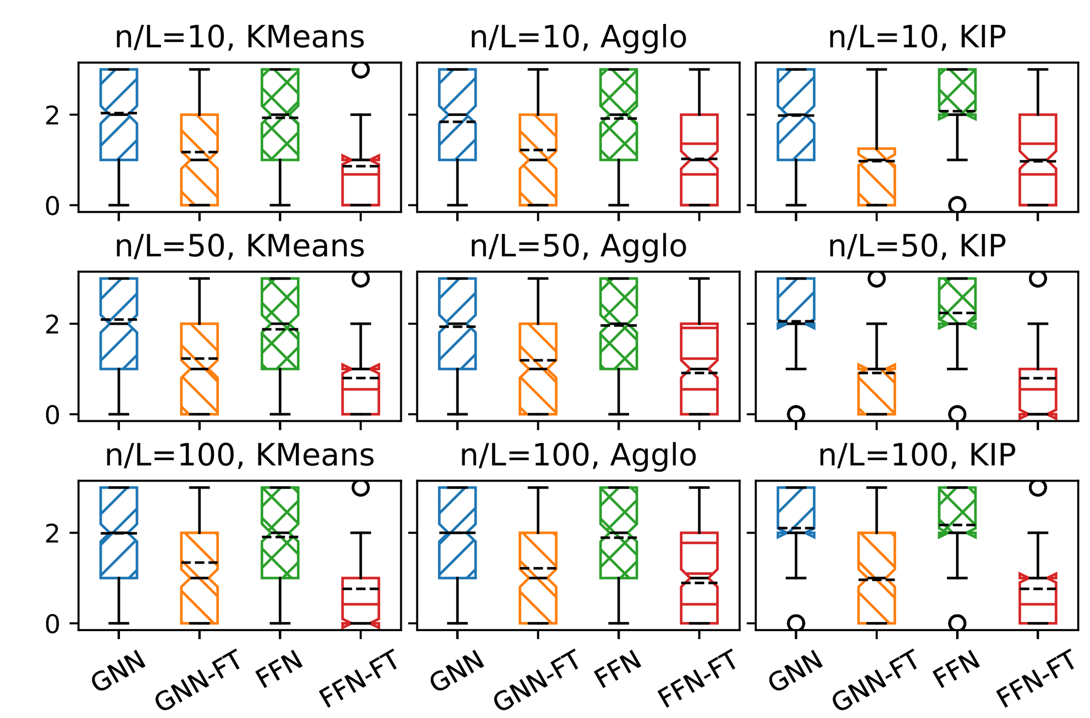


Figure 5. Rank of autoencoders grouped by distillation methods and distill size N.

Model	# Enc. Params ↓	Dec. Params ↓	Clf. Params ↓
FFN	24316 47916 111891	18425 42494 72745	12402 12402 22702
GNN	2832 3880 4904	25711 52645 82795	12402 22702 22702

Table 4. Parameters of autoencoder modules.

- GraphSage outperforms GCN and GAT.
- FFN-FT leads in overall performance, closely followed by GNN-FT.

Which Distillation Method Leads to Better Performance?

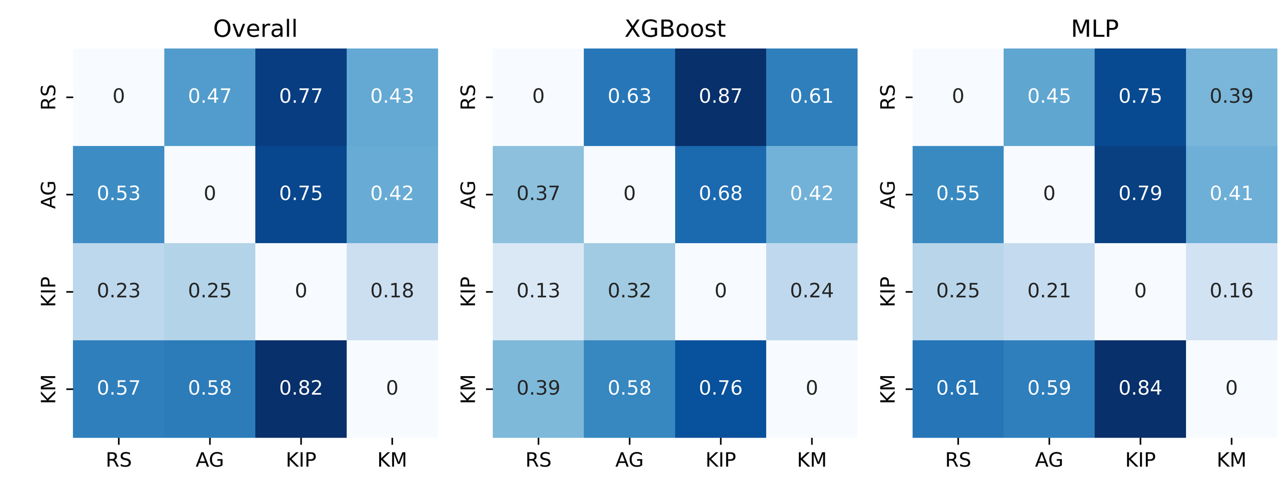


Figure 6. Pairwise comparison of distillation methods. Rows denote *victories*, columns denote *losses*.

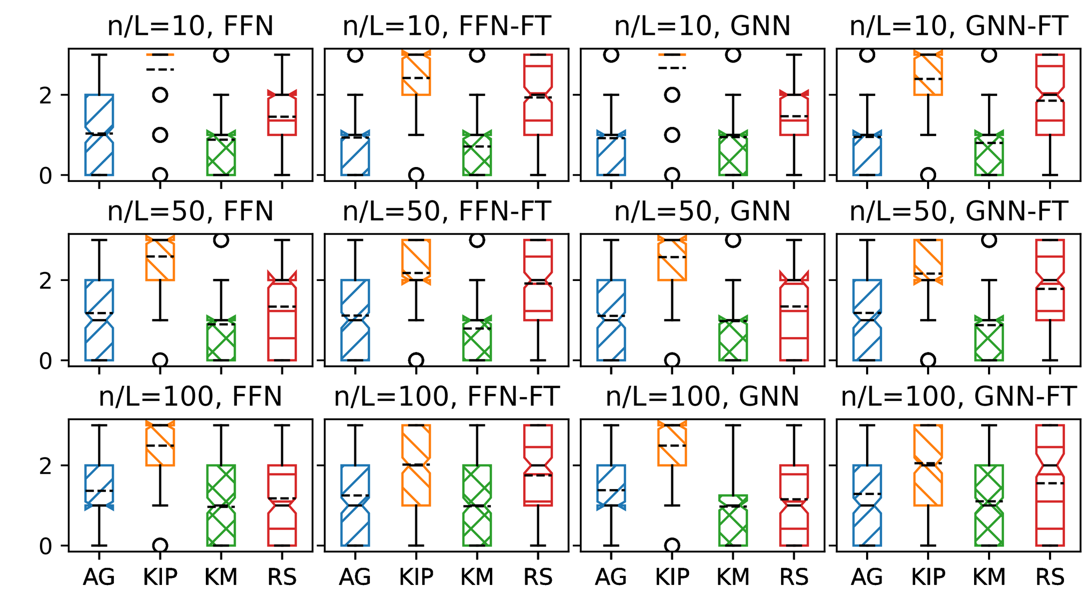


Figure 7. Rank of distill method grouped by distill size N and encoder.

- K-Means has highest tendency to outperform other distillation methods under equal settings.
- Image algorithm (KIP) is outperformed in most cases by every other distillation method.

Conclusion

- Data distillation method for image datasets do not directly translate to tabular datasets.
- K-Means is the most effective distillation method across 26 datasets considered.
- Pipelines using the encoded output of FFN-FT autoencoder with K-Means lead to the best downstream classifier performance.
- GNN-based autoencoders offer the benefit of much smaller parameter size for a small trade-off in performance.

References

[1] Timothy Nguyen, Zhourong Chen, and Jaehoon Lee. Dataset Meta-Learning from Kernel Ridge-Regression, March 2021.

[2] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A. Efros. Dataset Distillation, February 2020.

[3] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on tabular data?, 2022.

[4] Qitian Wu, Chenxiao Yang, and Junchi Yan. Towards Open-World Feature Extrapolation: An Inductive Graph Learning Approach, October 2021.

inwon.kang04@gmail.com