

SubTab: Subsetting Features of Tabular Data for Self-Supervised Representation Learning

Talip Uçar, Ehsan Hajiramezanali, Lindsay Edwards

Respiratory and Immunology, R&D, AstraZeneca

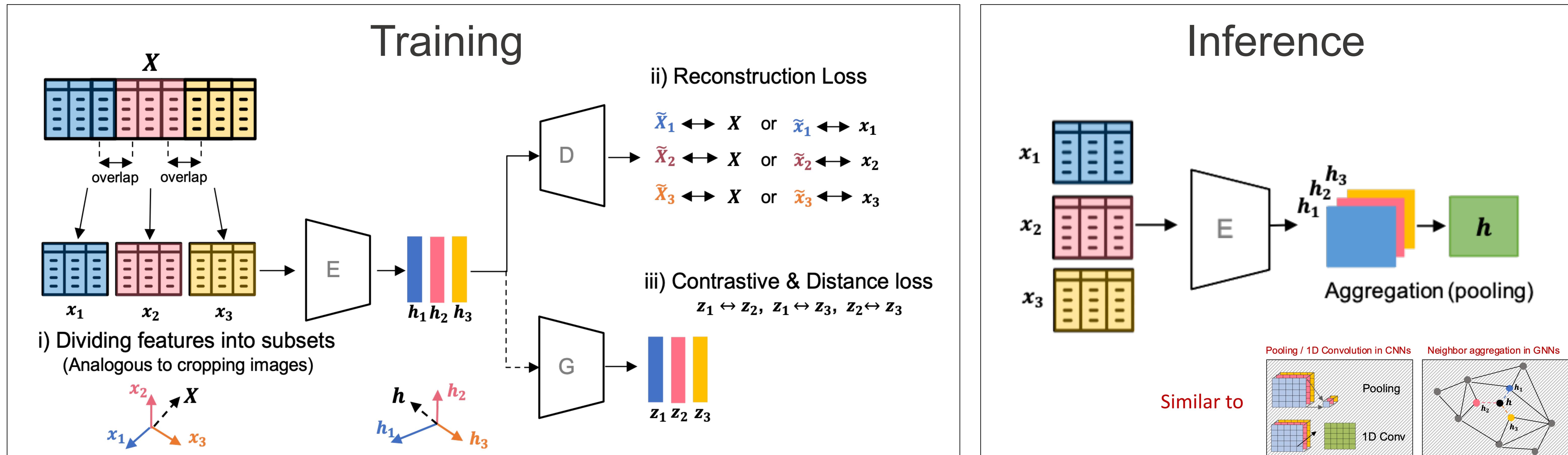
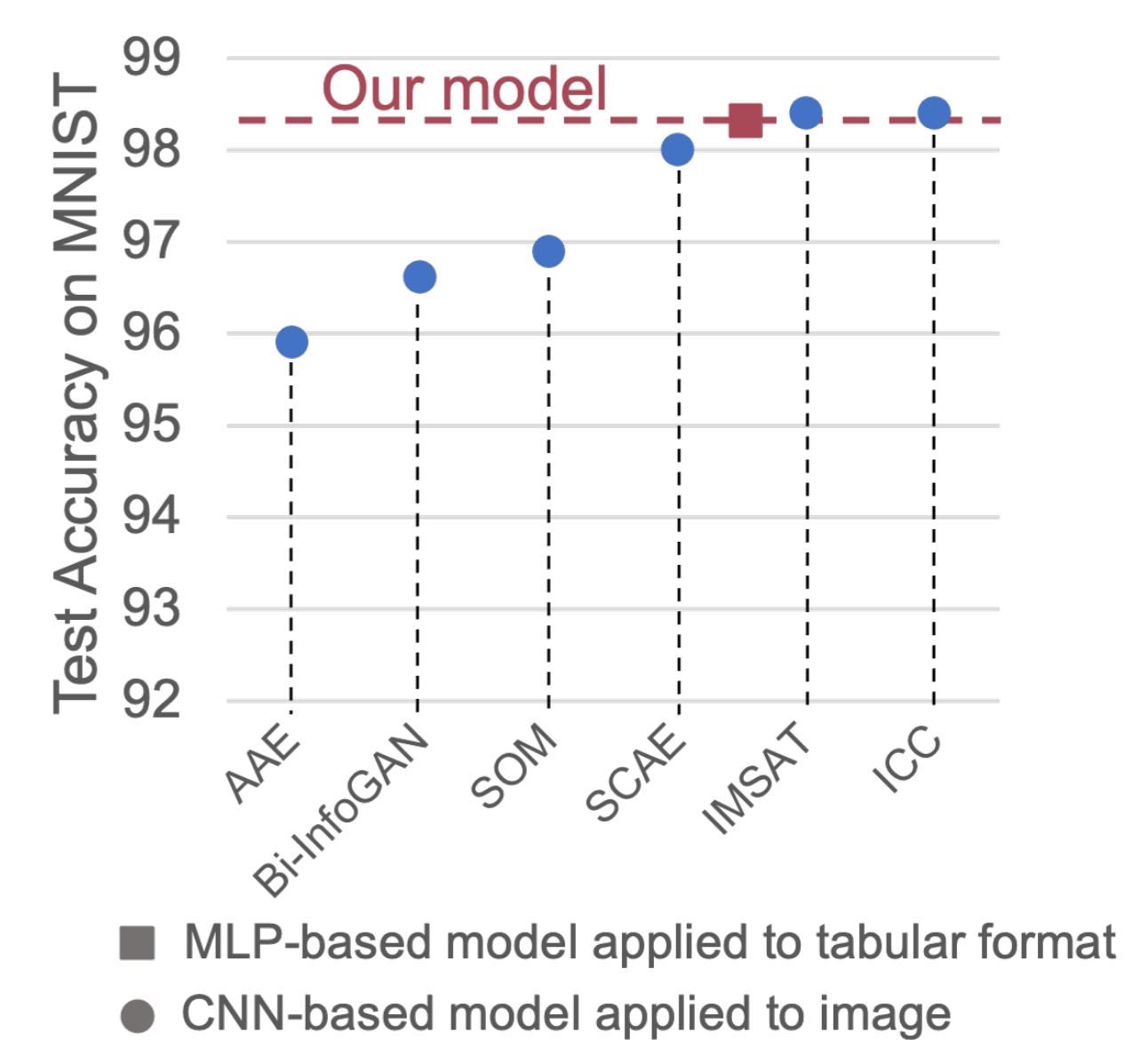
Abstract

Problem Statement

- ❑ Lack of effective data augmentation methods in tabular domain
- ❑ Self-supervised learning in tabular data is understudied:
 - ❑ Lack of effective data augmentation methods
 - ❑ Lack of specialized architectures for tabular data

SubTab

- ❑ Introduces effective methods for tabular data
- ❑ Achieves 98.31%, **the state of the art (SOTA)**, on MNIST data in tabular format and surpasses existing baselines on three other real-world datasets by a significant margin.
- ❑ Makes a simple MLP-based model perform on par with CNN-based SOTA models

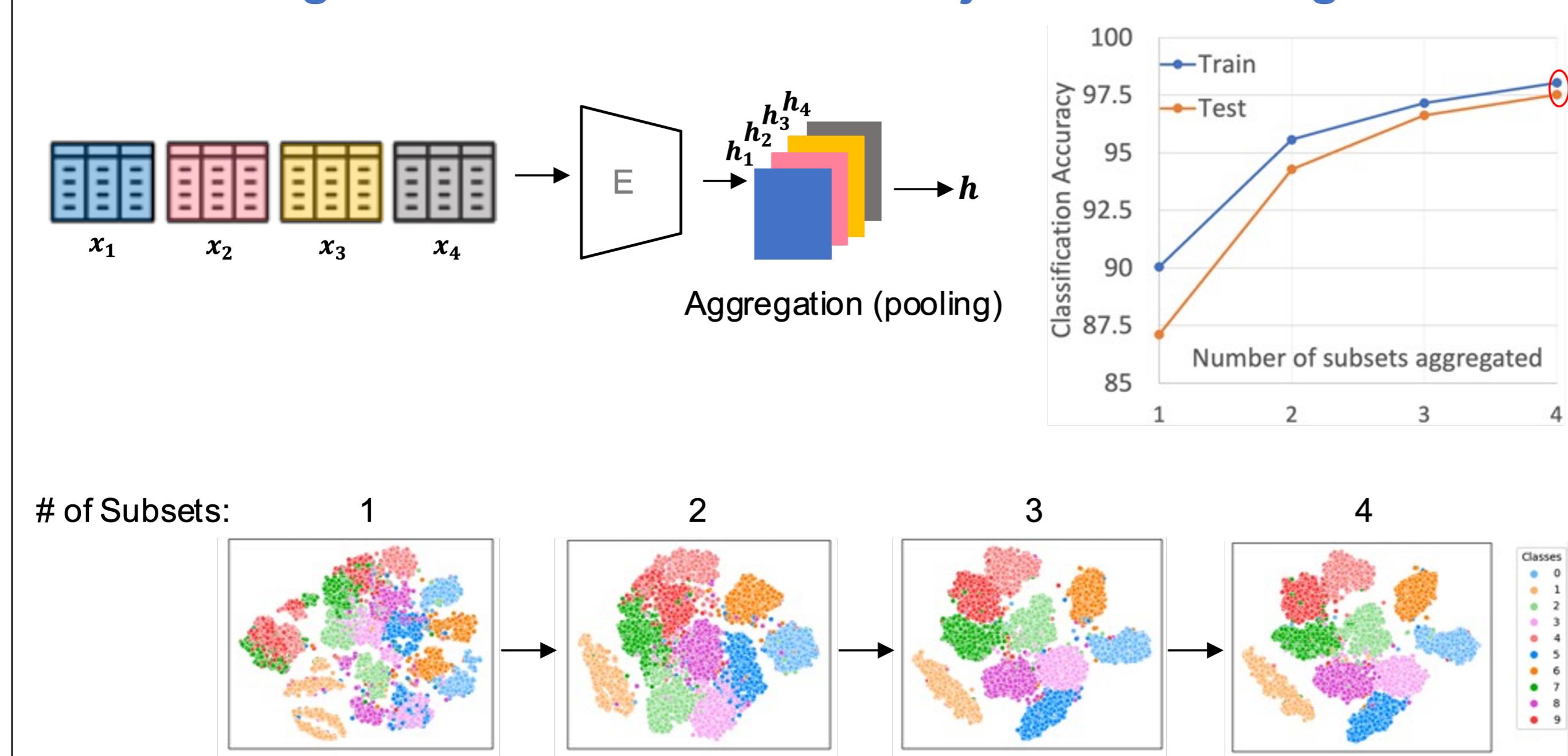


Comparing Self-Supervised and Autoencoder models

Table 1: Accuracy scores for all models for various datasets. The abbreviations in the table; NC: Neighbour columns used, RF: Random features used, G: Gaussian noise used, S: Swap noise used.

Type	Models	MNIST	Income	Blog	Obesity	TCGA
Supervised baseline	Logistic Regression	92.60±0.03	84.68±0.05	84.15±0.12	62.35±4.02	36.98±1.25
	Random Forest	96.96±0.06	84.62±0.07	83.61±0.15	67.45±2.23	61.62±1.02
	XGBoost	98.02±0.086	86.11±0.20	84.29±0.23	64.05±4.52	72.61±1.31
Autoencoder baseline	AE	92.77±0.32	84.67±0.07	84.06±0.24	61.96±3.28	55.16±0.75
	AE w/ Dropout (p=0.2)	94.31±0.28	85.00±0.10	84.18±0.20	62.74±4.38	56.87±2.26
	DAE (RF)	96.30±0.14 (S)	84.37±0.36 (G)	84.12±0.29 (G)	56.43±5.79 (G)	54.31±1.39 (G)
Self-supervised	CAE (NC)	96.39±0.20 (S)	84.24±0.18 (G)	84.3±0.31 (G)	62.26±5.01 (G)	54.20±1.17 (G)
	VIME-self	95.23±0.17 (S)	84.43±0.08 (G)	84.11±0.27 (G)	66.45±4.54 (G)	55.11±1.37 (G)
	SubTab with:					
	Base model (No noise)	97.26±0.2	85.31±0.08	84.29±0.26	68.01±3.07	57.02±1.50
	+Noise	97.47±0.18 (S)	85.34±0.07 (G)	84.47±0.15 (G)	71.13±4.08 (G)	58.25±1.36 (G)
	+Distance loss	97.52±0.14 (S)	85.35±0.06 (G)	84.64±0.19 (G)	69.25±4.19 (G)	58.15±1.56 (G)
	+LatentDim=512	97.86±0.07 (S)				

Measuring information content of the joint embedding



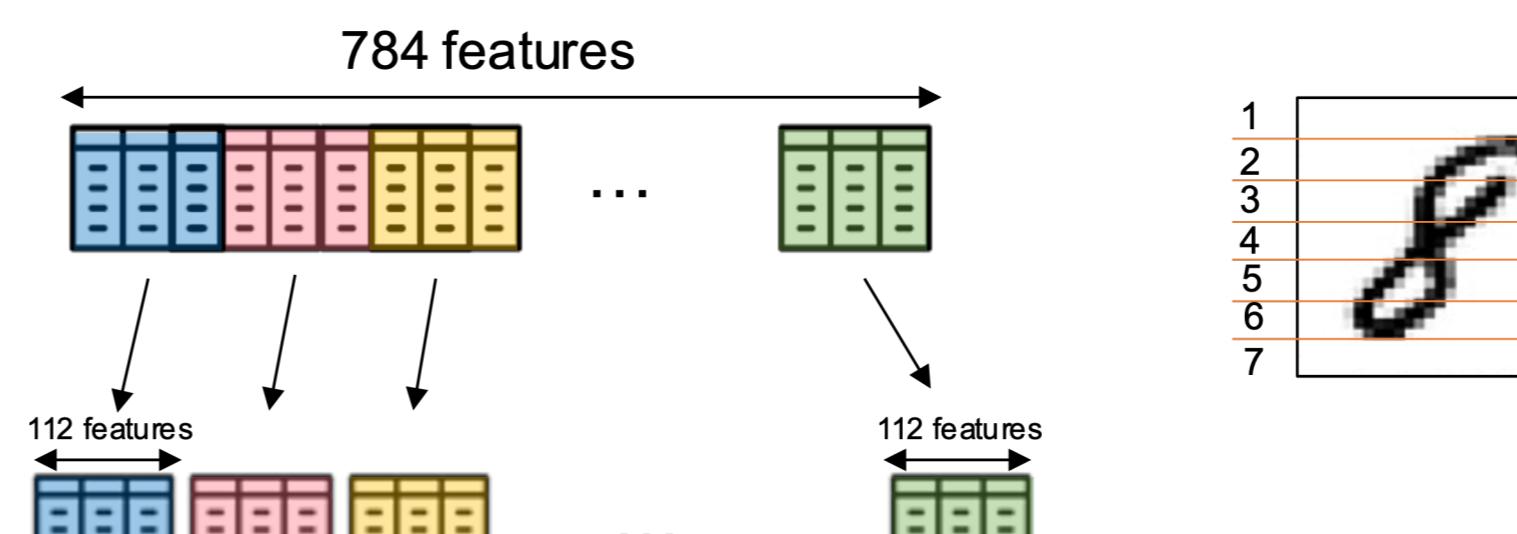
Ablation study - I: Effects of loss terms, aggregation type etc.

RL	CL	Noise	DL	SF	LD	Agg	Test Accuracy
+	-	-	-	-	128	+	97.13
-	+	-	-	-	128	+	97.11
+	+	-	-	-	128	+	97.26
+	+	Zero-out	-	-	128	+	97.25
+	+	Gaussian	-	-	128	+	97.25
+	+	Swap	-	-	128	+	97.47
+	+	Swap	+	-	128	+	97.52
+	+	Swap	+	+	128	+	97.2
+	+	Swap	+	-	512	-	95.92
+	+	Swap	+	-	512	+	97.86

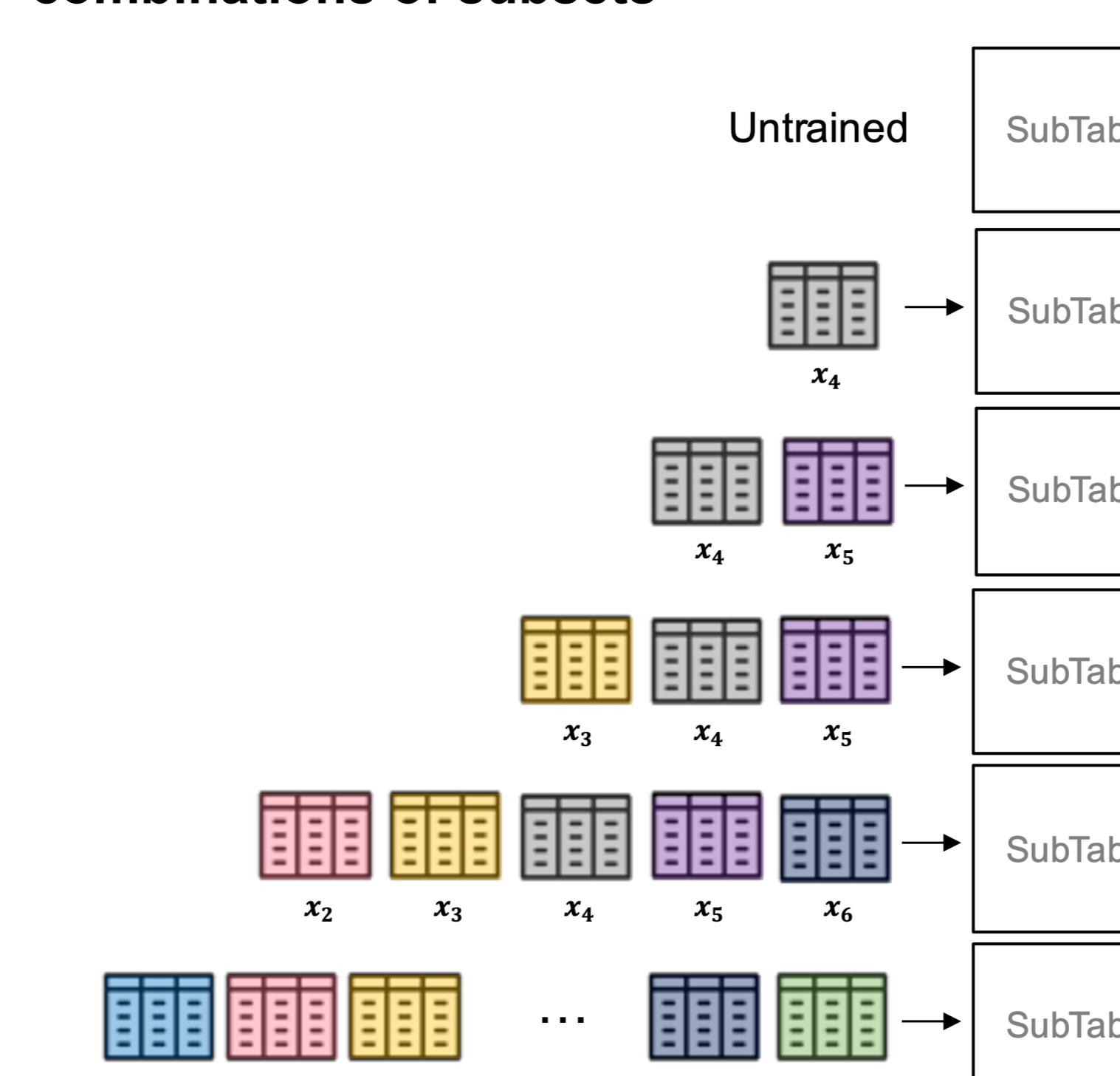
Table 2: Ablation study using MNIST with 4 subsets with 75% overlap. Abbreviations are; RL: Reconstruction Loss, CL: Contrastive Loss, DL: Distance Loss, SF: Shuffled Features, LD: Latent Dim, Agg: Aggregating embeddings.

Experiments: Joint embedding under missing data during training and/or test time

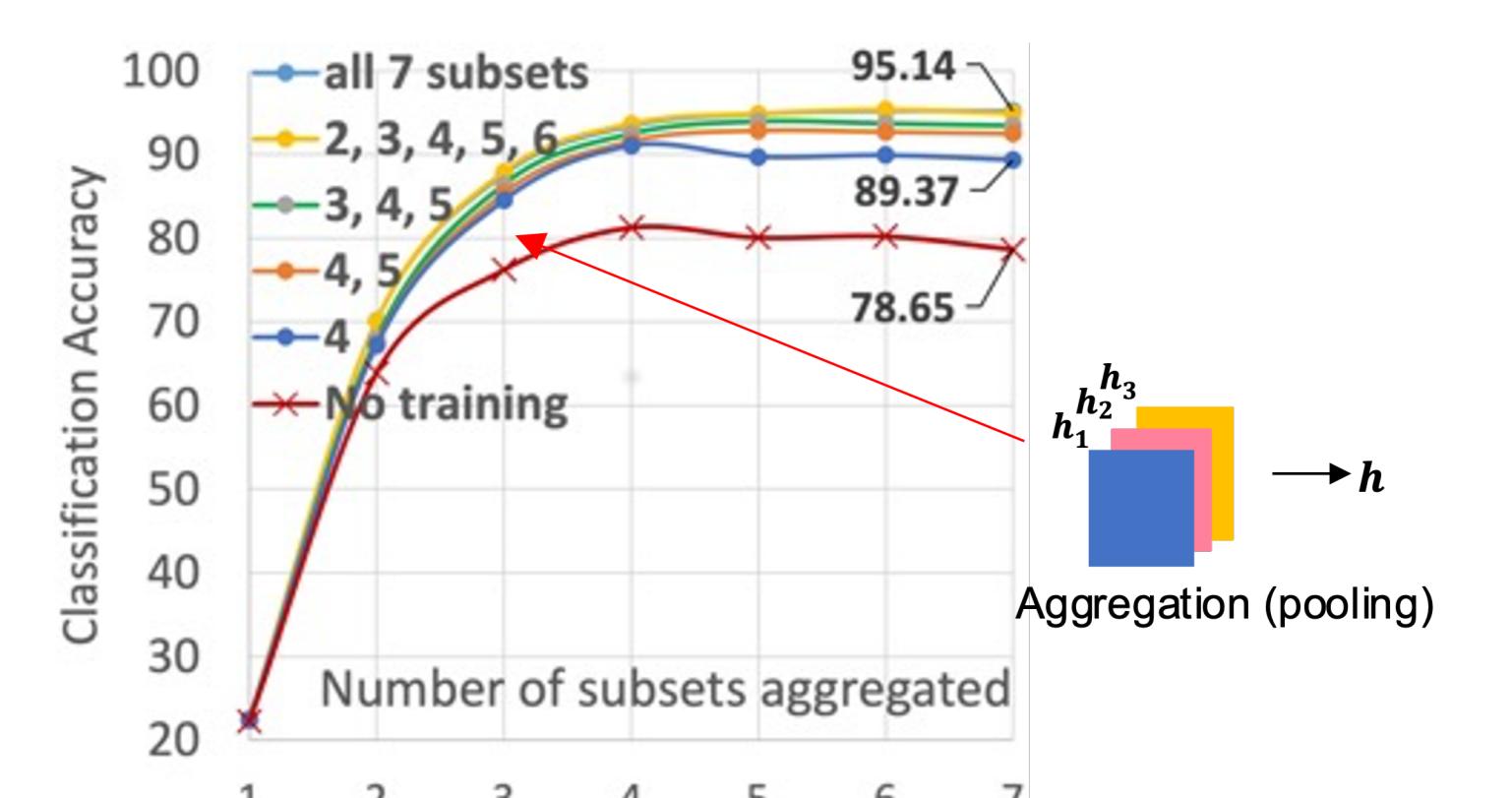
I. Slicing MNIST digits to 7 subsets



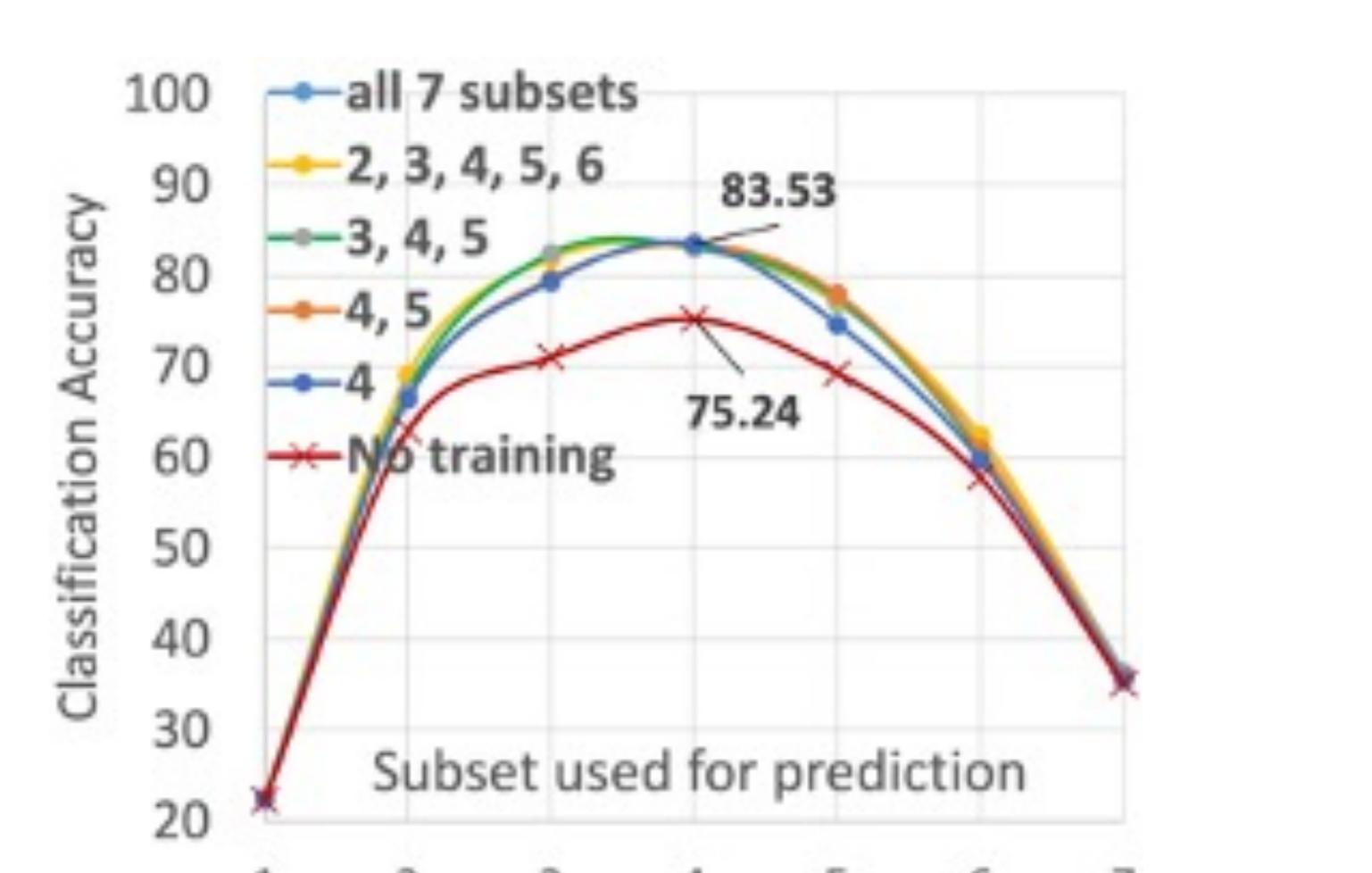
II. Training 5 models using different combinations of subsets



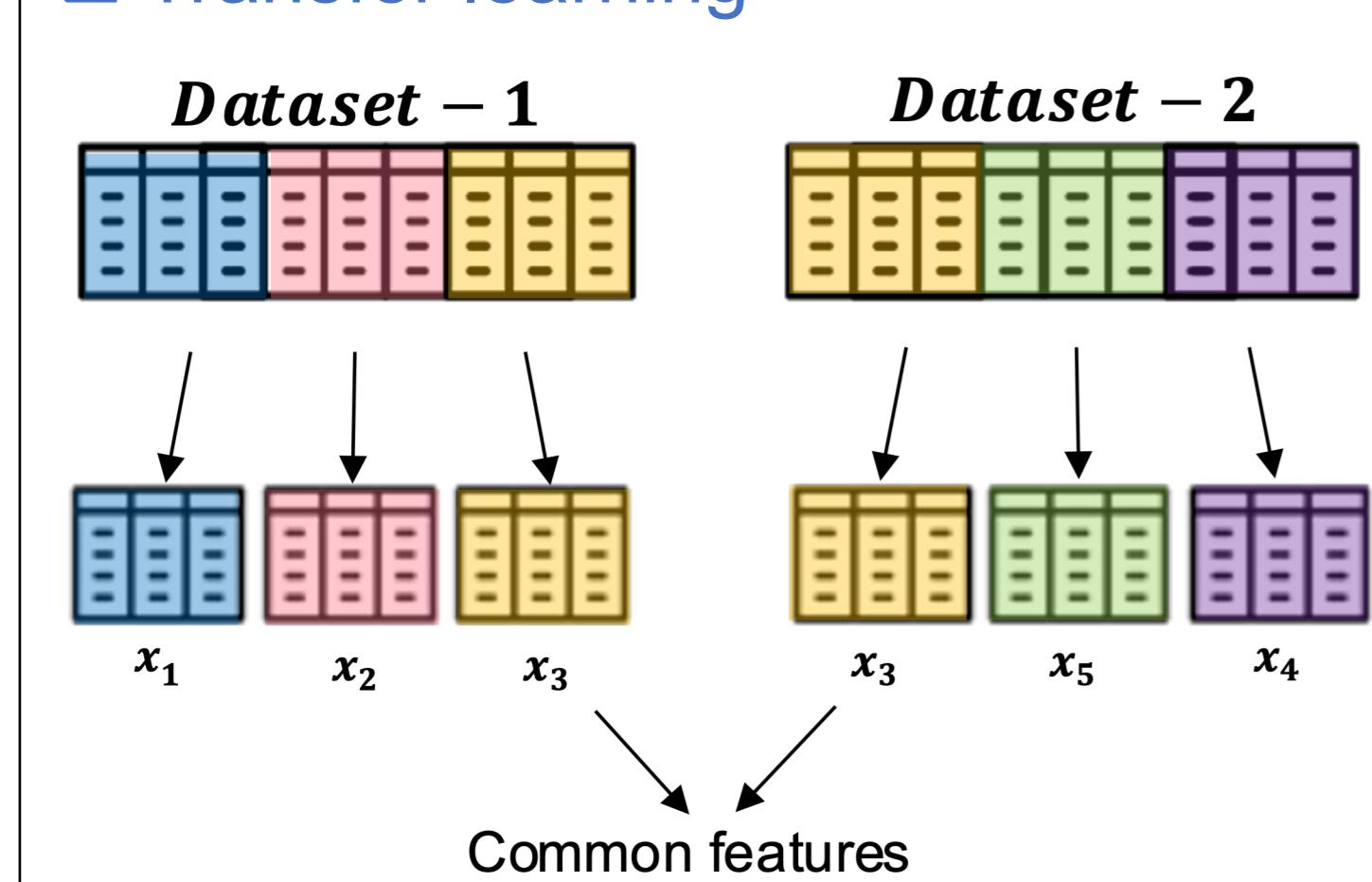
III. Measuring information content of joint embedding obtained from 6 models



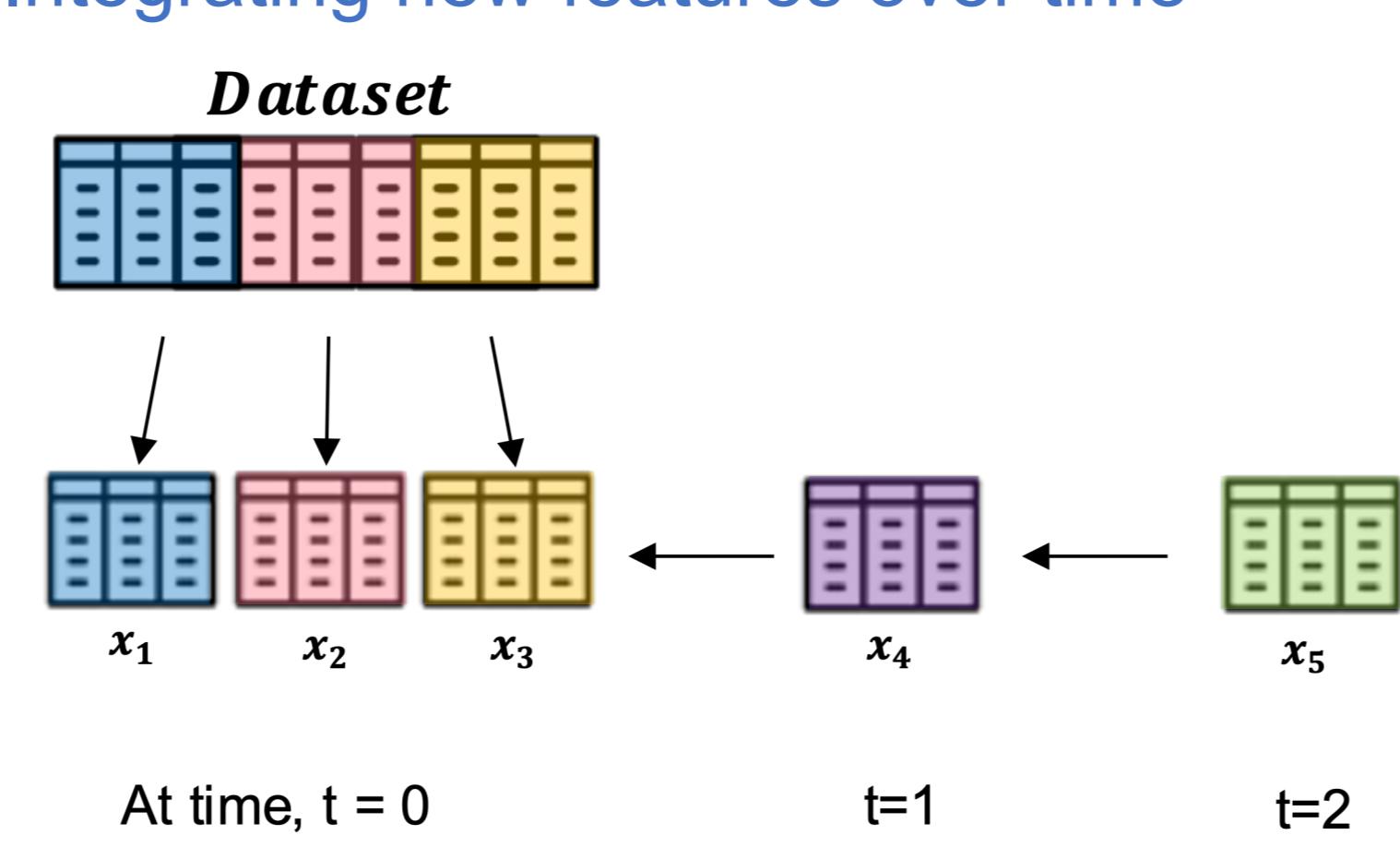
IV. Measuring information content of the individual subsets



Transfer learning



Integrating new features over time



Tabular data with missing features

