AstraZeneca

SubTab

Subsetting Features of **Tab**ular Data for
Self-Supervised Representation Learning

Talip Uçar
Ehsan Hajiramezanali
Lindsay Edwards

Outline

I. Abstract
- ❑ Summary of our work

II. Motivation / Background
- ❑ Comparing data types
- ❑ Challenges in tabular data
  - ❑ Data augmentation
  - ❑ Parameter sharing

III. SubTab
- ❑ Framework
- ❑ Results
- ❑ Applications
- ❑ Summary

Abstract

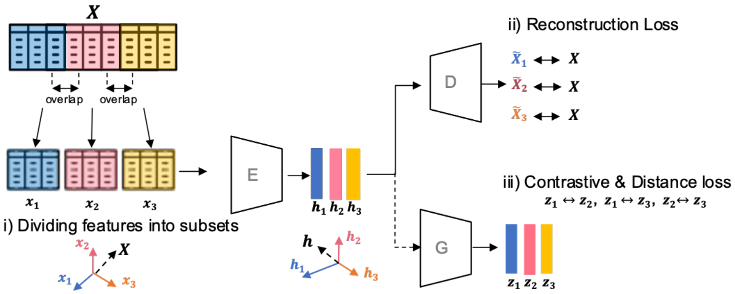
Problem:

- ❑ It is not easy to design effective data augmentation methods in tabular domain
- ❑ Self-supervised representation learning in tabular data is understudied:
  - ❑ Lack of effective data augmentation methods
  - ❑ Lack of specialized architectures for tabular data

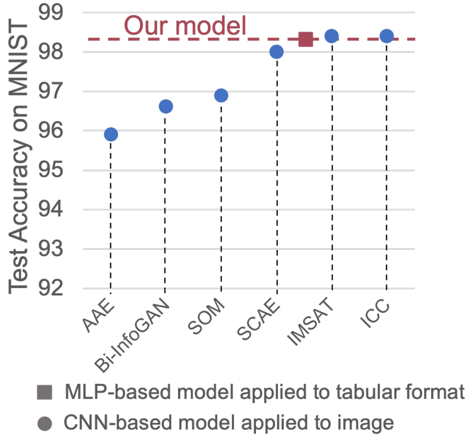
Our work - SubTab:

- ❑ Introduces effective methods for tabular data
- ❑ Achieves 98.31%, the state of the art (SOTA), on MNIST data in tabular format.
- ❑ Makes a simple MLP-based model perform on par with CNN-based SOTA models

SubTab:



Result on MNIST:



Motivation / Background

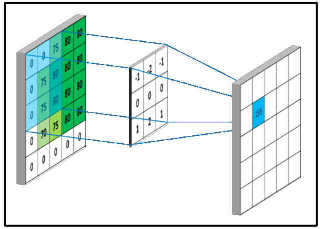
Types of Data

Tabular Data

Age	Gender	BMI	Insulin	…
50	M	26.6	0	…
31	F	33.6	94	…

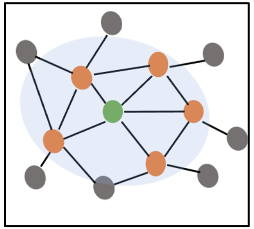
?

Images



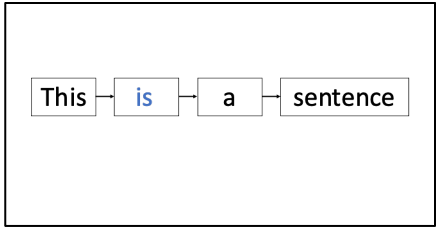
- ❑ Sequence of pixels
- ❑ CNNs*
- ❑ Parameter sharing

Graph



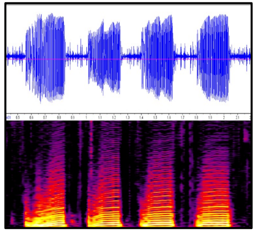
- ❑ Neighboring nodes
- ❑ GNNs
- ❑ Parameter sharing

Text



- ❑ Sequence of words
- ❑ LSTMs*
- ❑ Parameter sharing

Audio



- ❑ Sequence of samples
- ❑ CNNs / LSTMs*
- ❑ Parameter sharing

* There is a trend towards using Transformers
Image source for CNN: https://medium.com/analytics-vidhya/cnn-convolutional-neural-network-8d0a292b4498

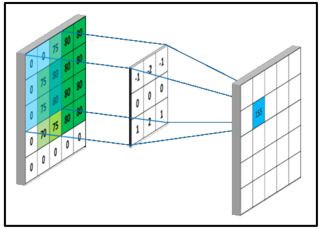
Types of Data

Tabular Data

Age	Gender	BMI	Insulin	…
50	M	26.6	0	…
31	F	33.6	94	…

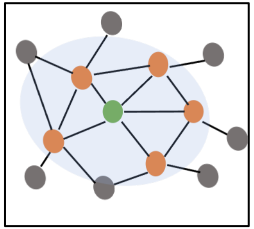
?

Images



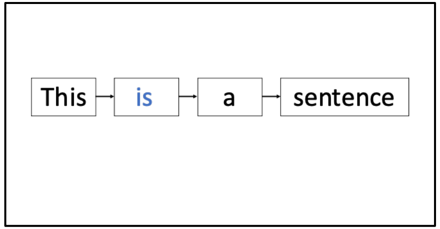
- ❑ Sequence of pixels
- ❑ CNNs*
- ❑ Parameter sharing

Graph



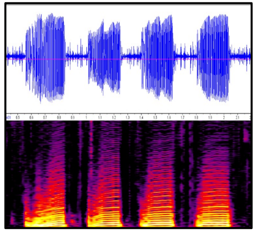
- ❑ Neighboring nodes
- ❑ GNNs
- ❑ Parameter sharing

Text



- ❑ Sequence of words
- ❑ LSTMs*
- ❑ Parameter sharing

Audio



- ❑ Sequence of samples
- ❑ CNNs / LSTMs*
- ❑ Parameter sharing

Common factors:

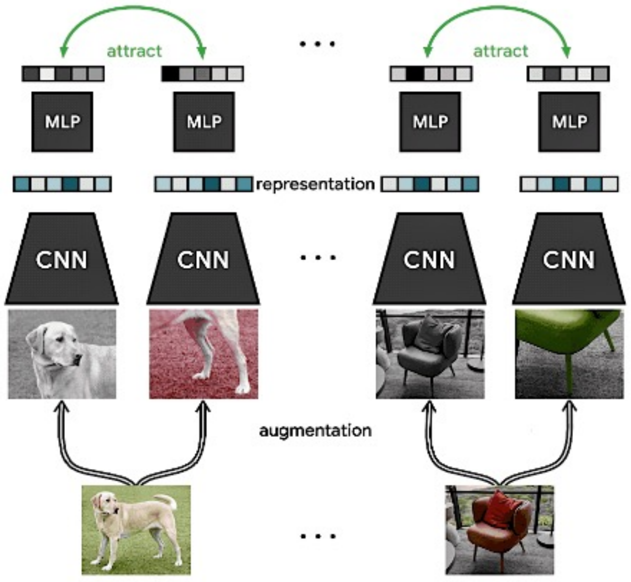
- ❑ Data augmentation can take advantage of structure in the data
- ❑ Parameter sharing through specialized architectures

* There is a trend towards using Transformers
Image source for CNN: https://medium.com/analytics-vidhya/cnn-convolutional-neural-network-8d0a292b4498

Prominent Works in Representation Learning

- ❑ Most prominent works are done in Computer Vision and NLP
- ❑ They take advantage of Data Augmentation

Computer Vision: SimCLR



NLP: BERT

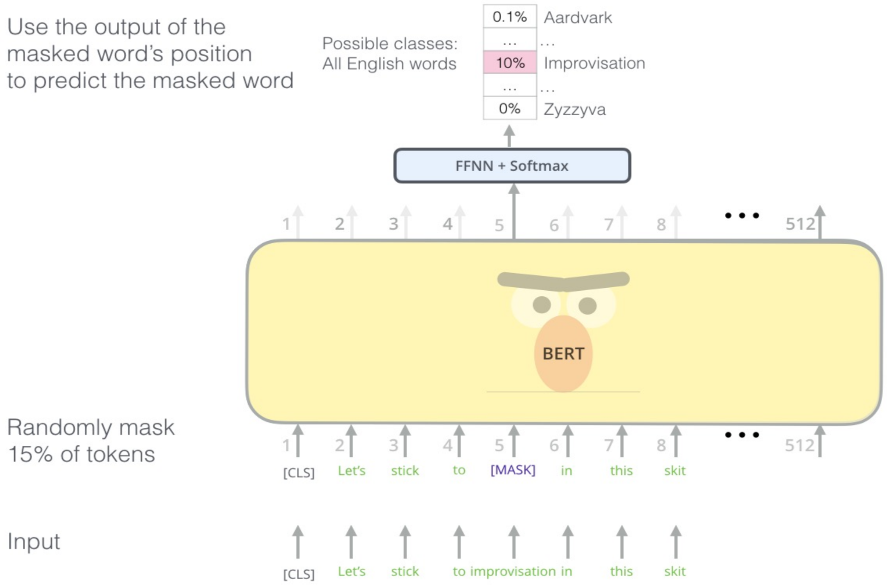
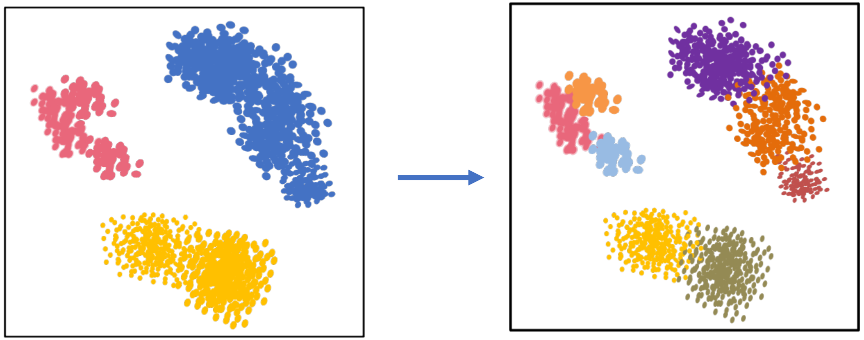


Image sources: https://github.com/google-research/simclr

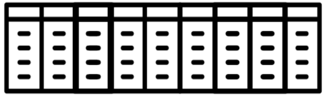
https://jalammar.github.io/illustrated-bert/

Data Augmentation in Tabular Data

Tabular Data



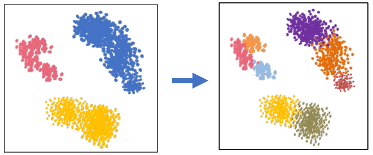
$$X = Patient\ records$$

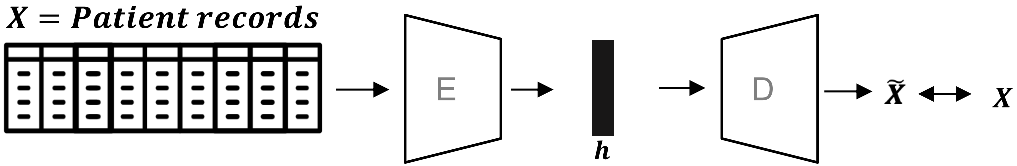


$\equiv$

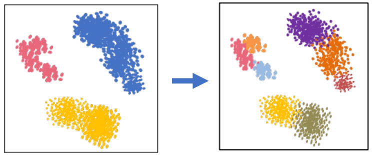
Age	Gender	BMI	Insulin	…
50	M	26.6	0	…
31	F	33.6	94	…

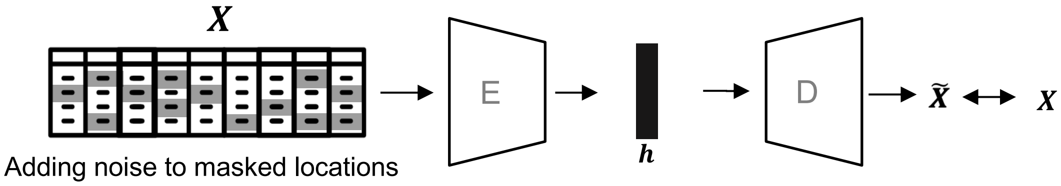
Autoencoder





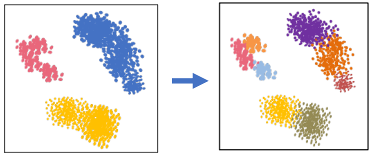
De-noising Autoencoder

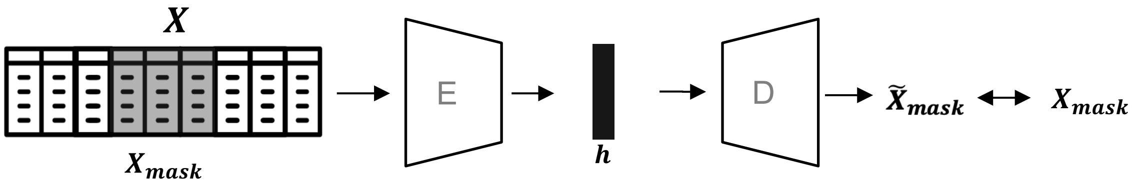




Pascal Vincent et al., “Extracting and Composing Robust Features with Denoising Autoencoders”, ICML 2008

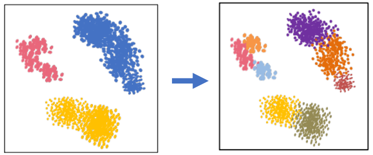
Context Encoder

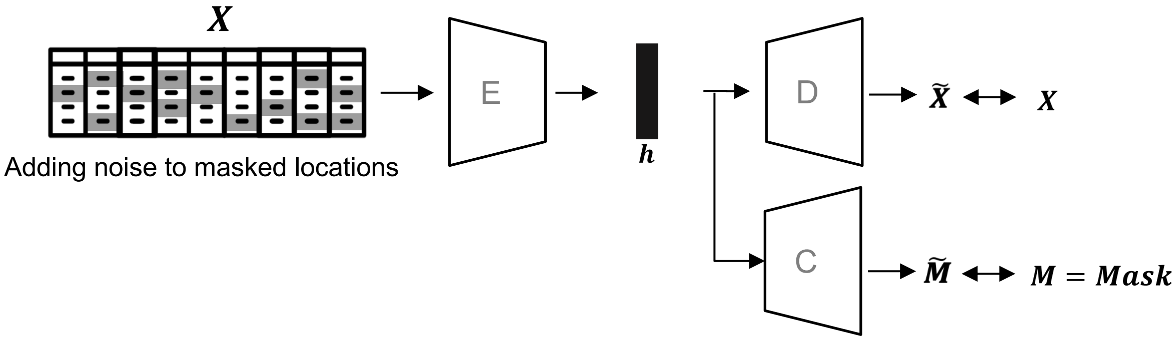




Deepak Pathak et al., “Context Encoders: Feature Learning by Inpainting”, https://arxiv.org/abs/1604.07379

VIME-Self

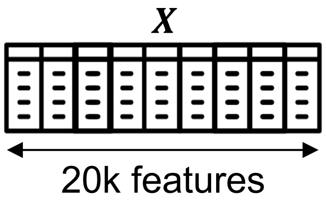


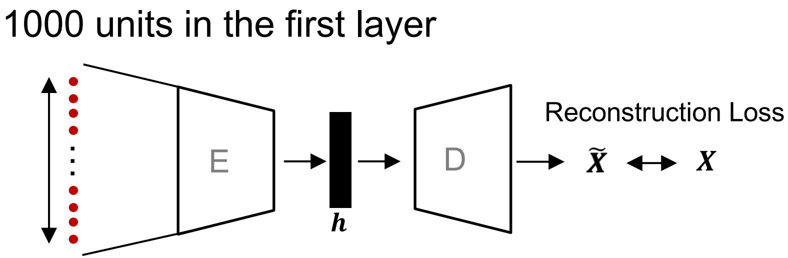


Jinsung Yoon et al., “Extending the success of self-and semi-supervised learning to tabular domain. *Advances in Neural Information Processing Systems*, 33, 2020.”

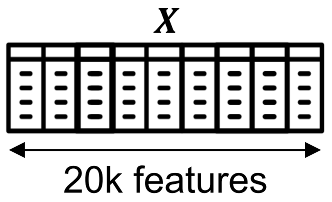
Parameter sharing in Tabular Data

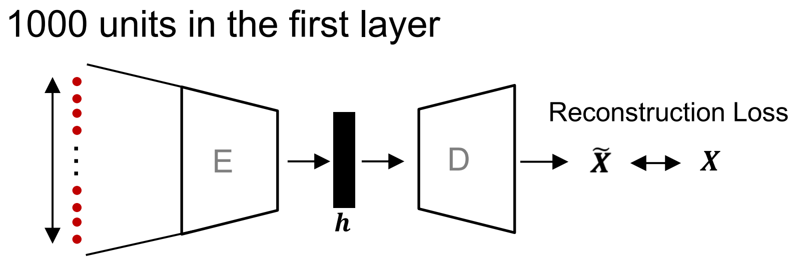
Parameter sharing (or lack of it) in Tabular Data





Parameter sharing (or lack of it) in Tabular Data

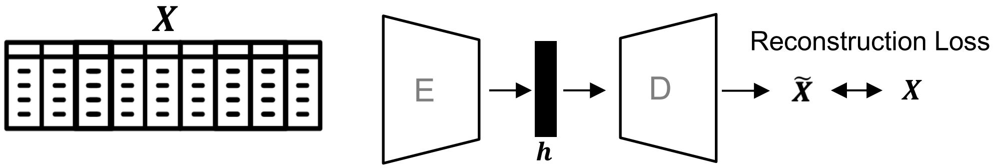




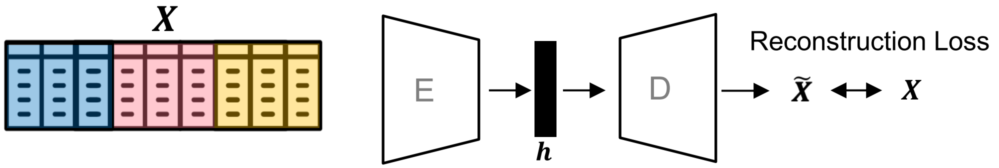
20 million parameters in the first layer alone

SubTab

Self-Supervised Representation Learning in SubTab

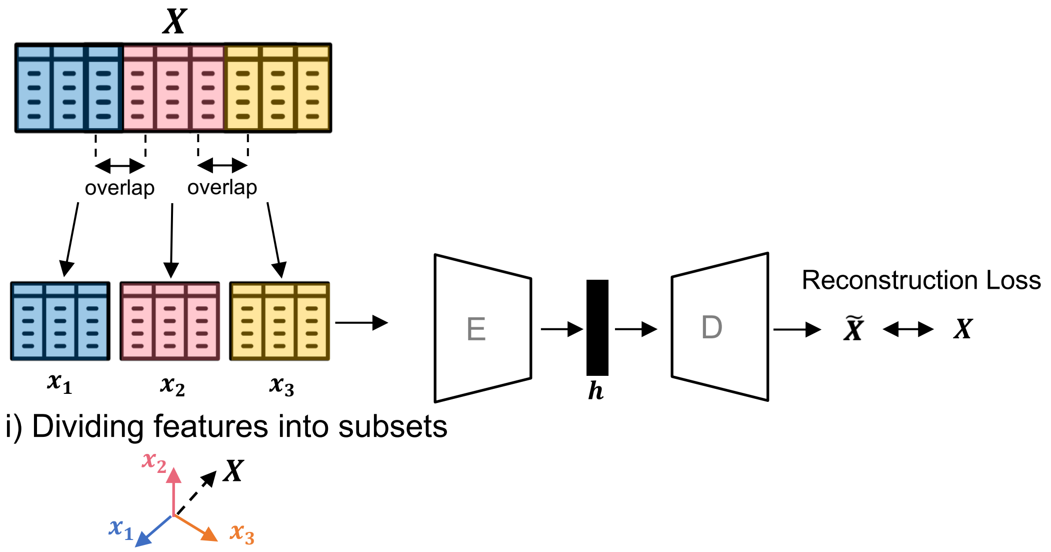


Self-Supervised Representation Learning in SubTab

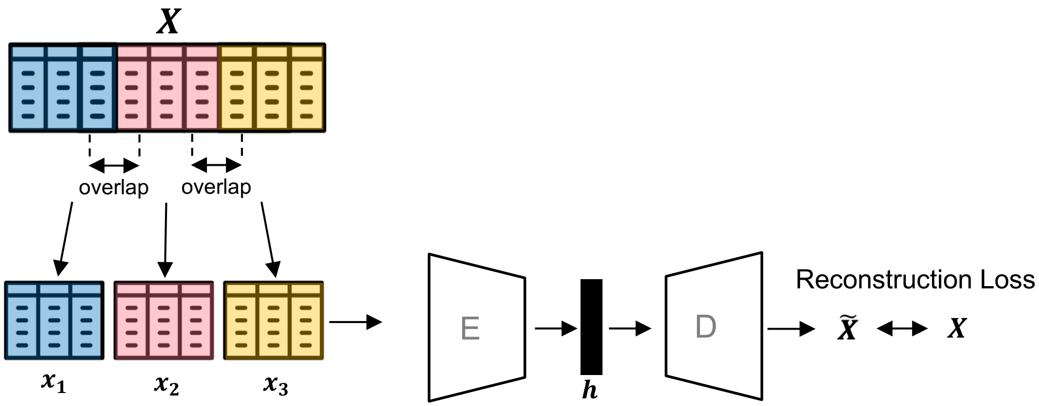


i) Dividing features into subsets

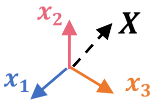
Self-Supervised Representation Learning in SubTab



Self-Supervised Representation Learning in SubTab

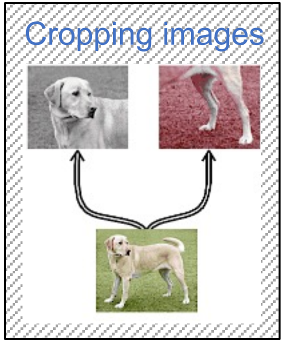


i) Dividing features into subsets

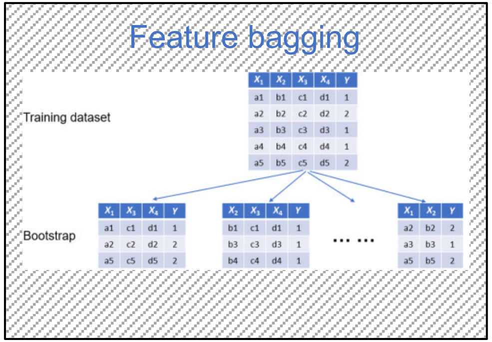


Similar to

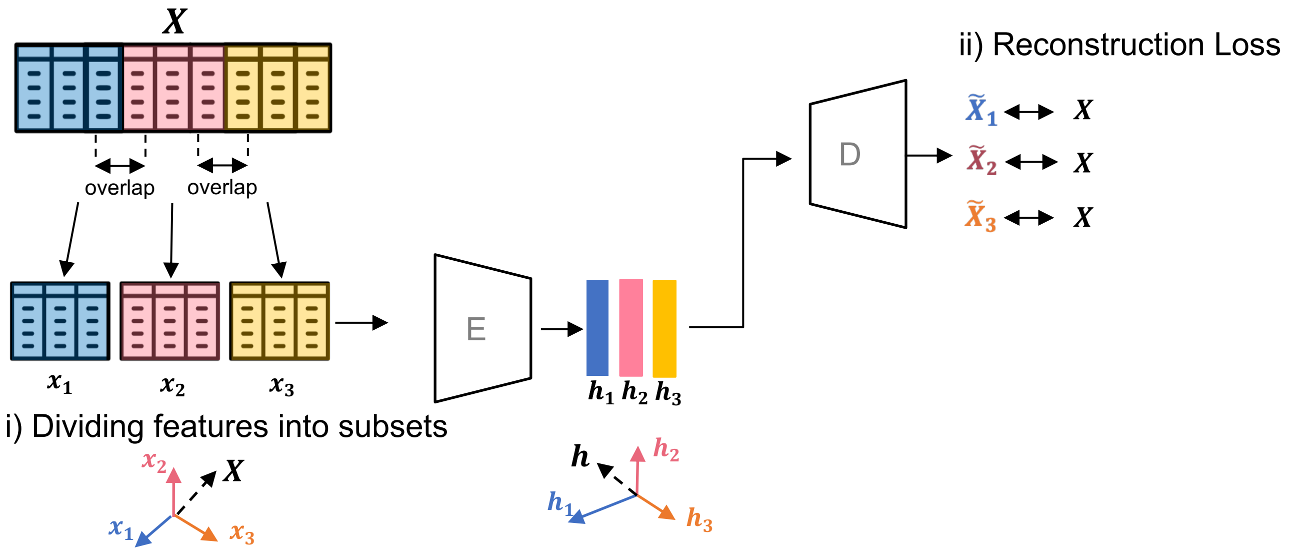
Computer Vision



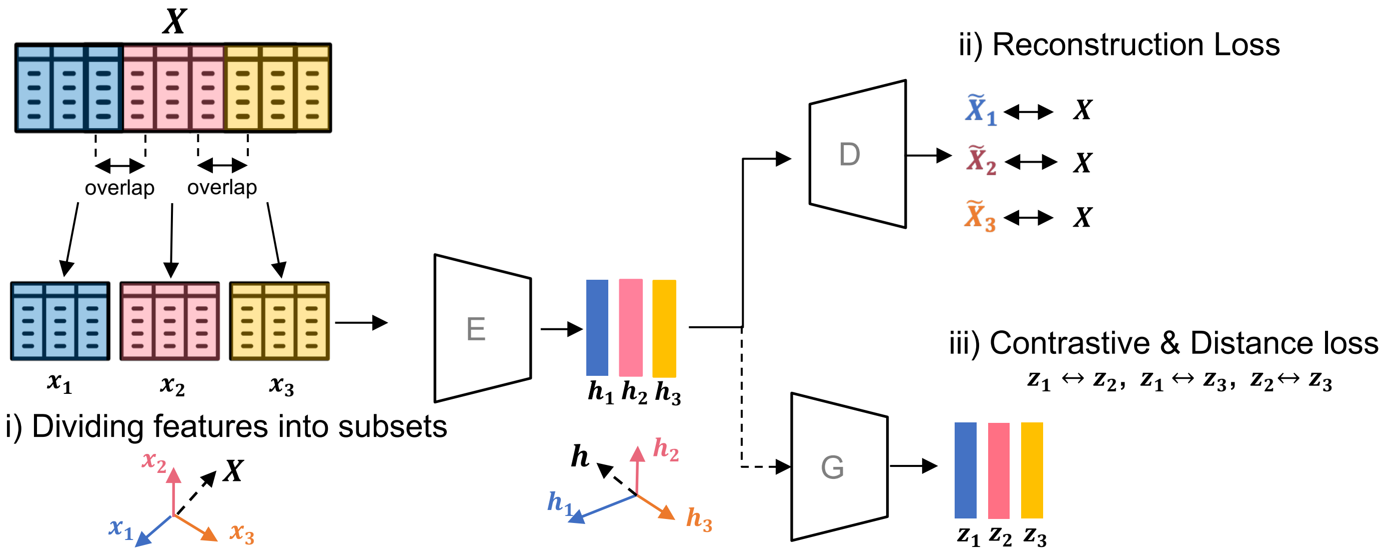
Ensemble Learning



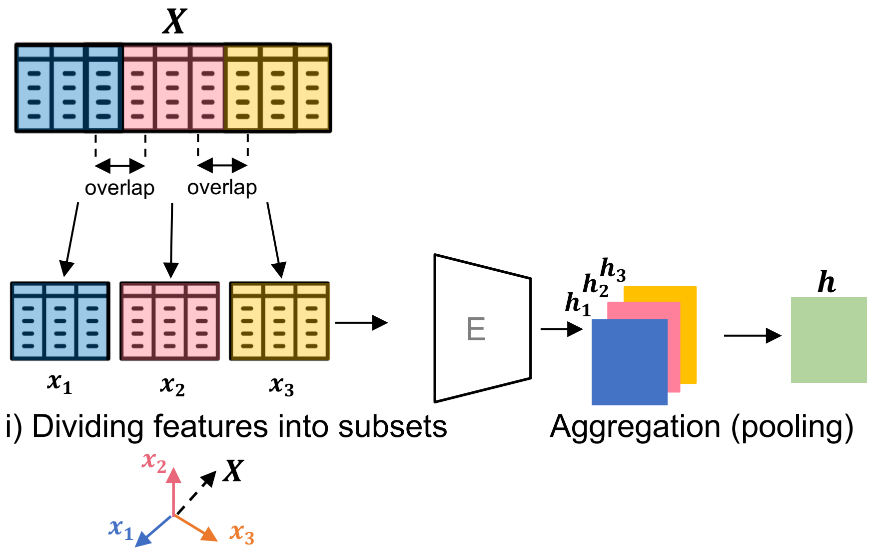
Self-Supervised Representation Learning in SubTab



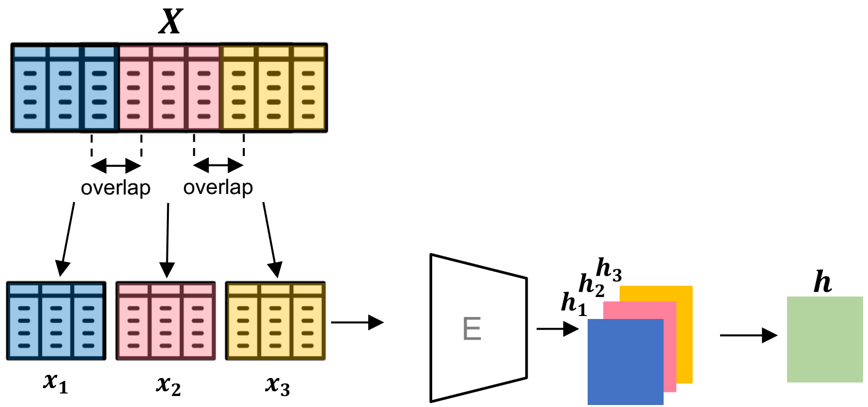
Self-Supervised Representation Learning in SubTab



SubTab at test time

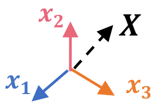


SubTab at test time

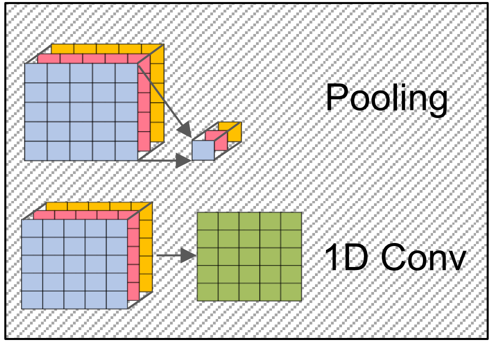


i) Dividing features into subsets

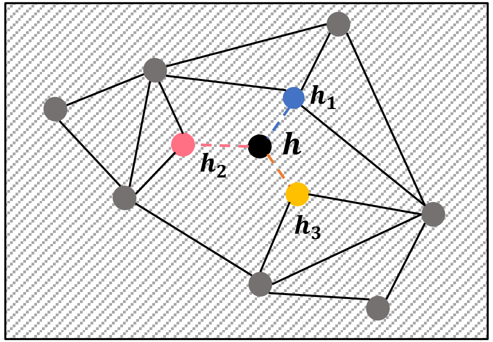
Aggregation (pooling)



Pooling / 1D Convolution in CNNs



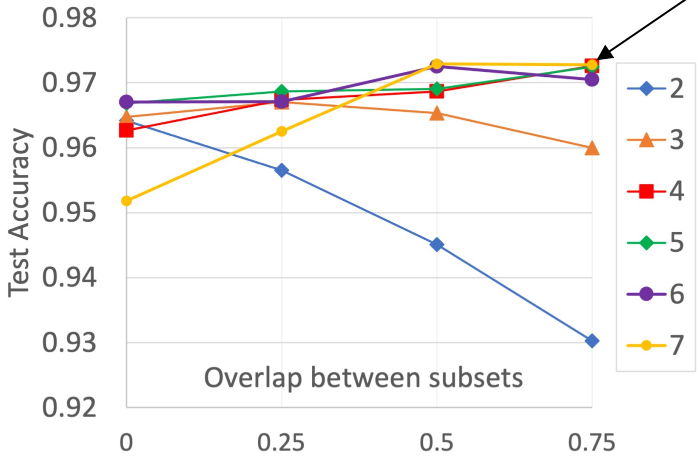
Neighbor aggregation in GNNs



Results & Summary

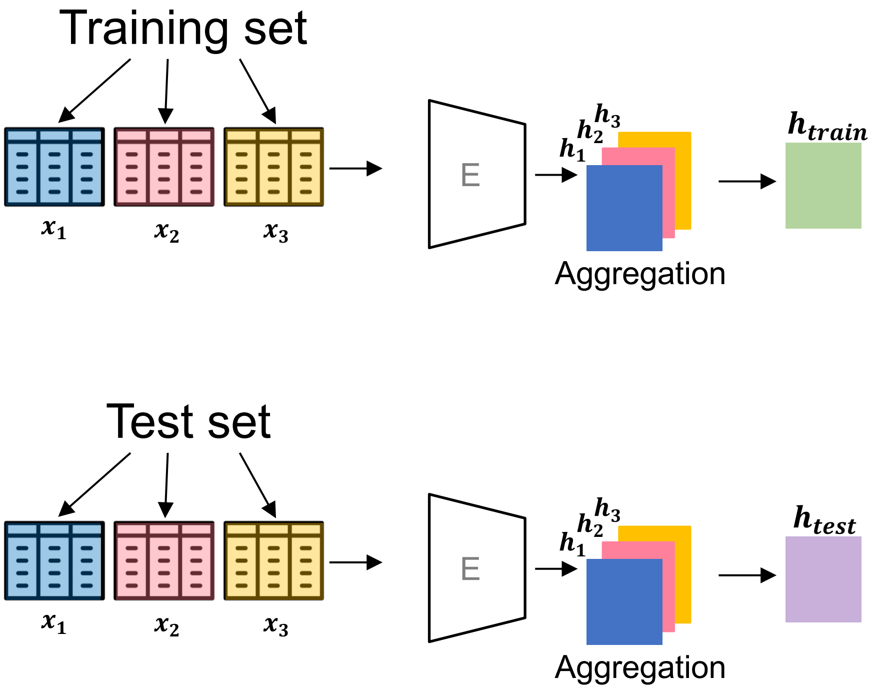
Performance over different number of subsets and overlap for MNIST

We used 4 subsets with 75% overlap



Evaluation

I. Extract embeddings



II. Evaluate representation

Results on 5 datasets

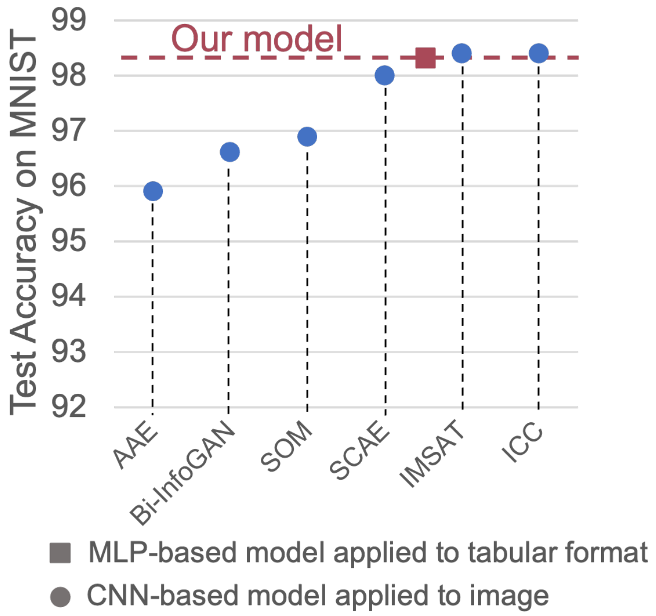
Table 1: Accuracy scores for all models for various datasets. The abbreviations in the table; NC: Neighbour columns used, RF: Random features used, G: Gaussian noise used, S: Swap noise used.

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
Self-supervised	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)
	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)**	-	-	-	-

Shallow vs Deep Architecture

Table 3: Comparing shallow and deep SubTab architectures.

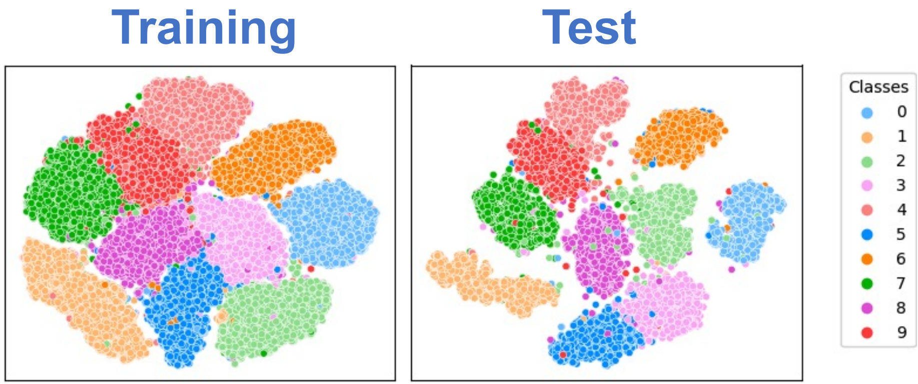
Model	MNIST	Income	Blog	Obesity	TCGA
Deep SubTab	97.86±0.07	85.35±0.06	84.64±0.19	**71.13±4.08**	58.25± 1.36
Shallow SubTab	**98.31±0.06**	85.34±0.03	84.64±0.09	66.88±5.35	**61.41±1.11**



- ❑ Shallow architecture performs better in some datasets
- ❑ MLP-based SubTab performs on par with CNN-based SOTA models.

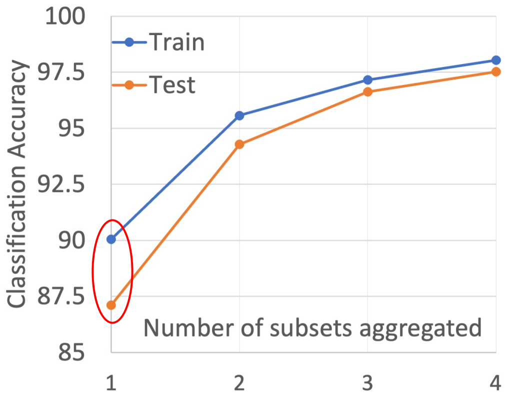
Representation Quality of SubTab

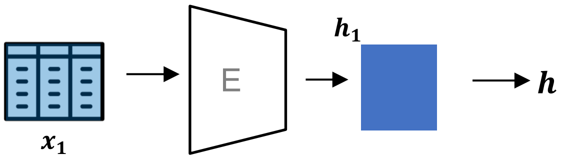
t-SNE plots for training and test set of MNIST for 4 subsets with 75% overlap



Information content in the joint embedding

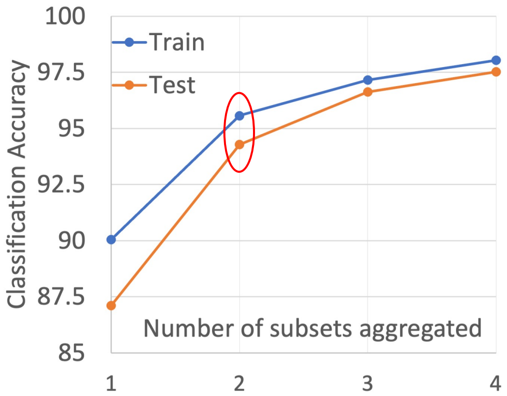
E = Encoder trained on 4 subsets with 75% overlap

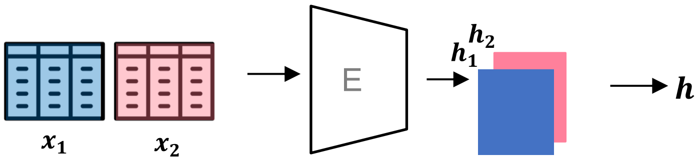




Information content in the joint embedding

E = Encoder trained on 4 subsets with 75% overlap

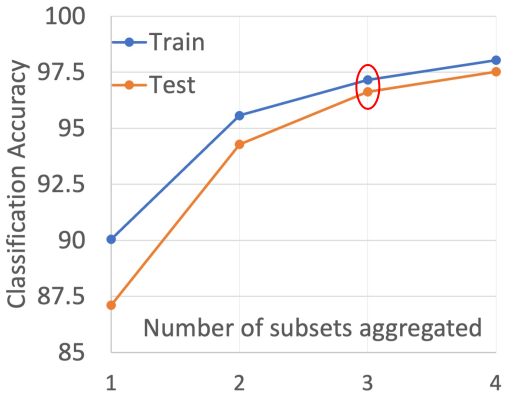


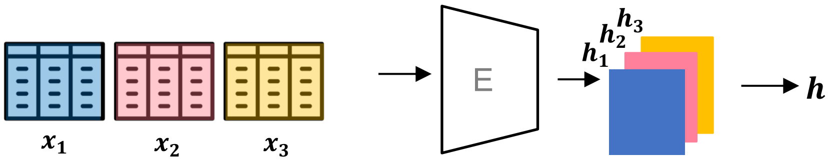


Aggregation (pooling)

Information content in the joint embedding

E = Encoder trained on 4 subsets with 75% overlap

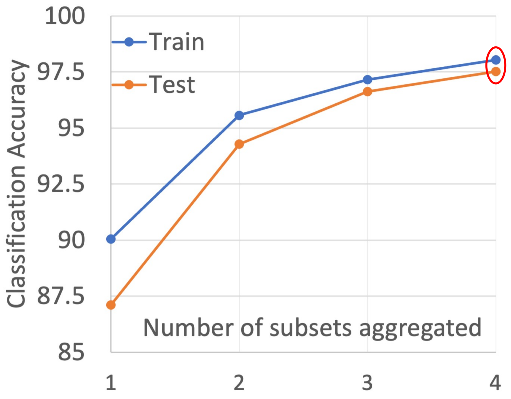


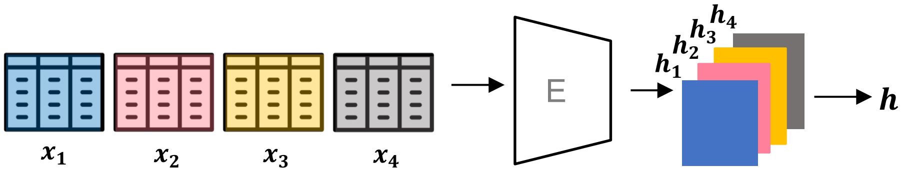


Aggregation (pooling)

Information content in the joint embedding

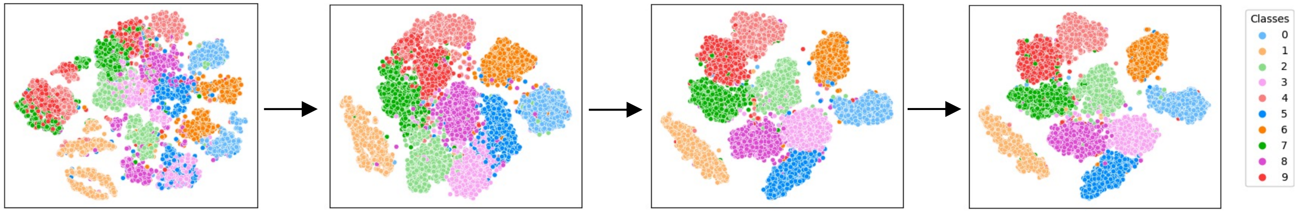
E = Encoder trained on 4 subsets with 75% overlap





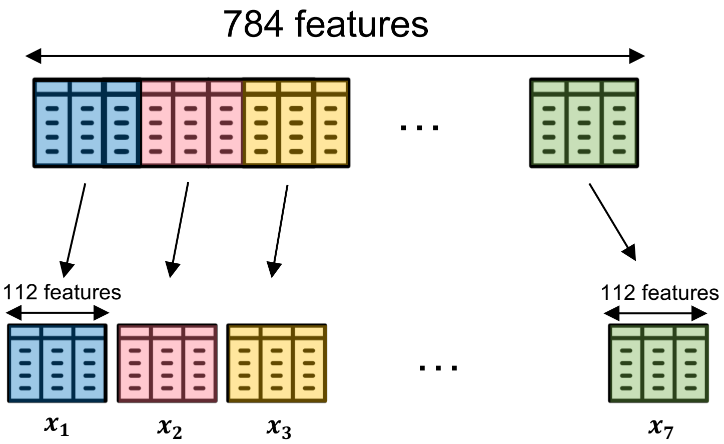
Aggregation (pooling)

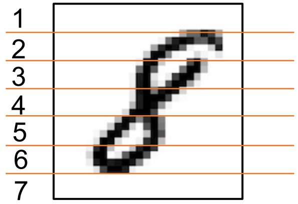
# of Subsets: 1 2 3 4



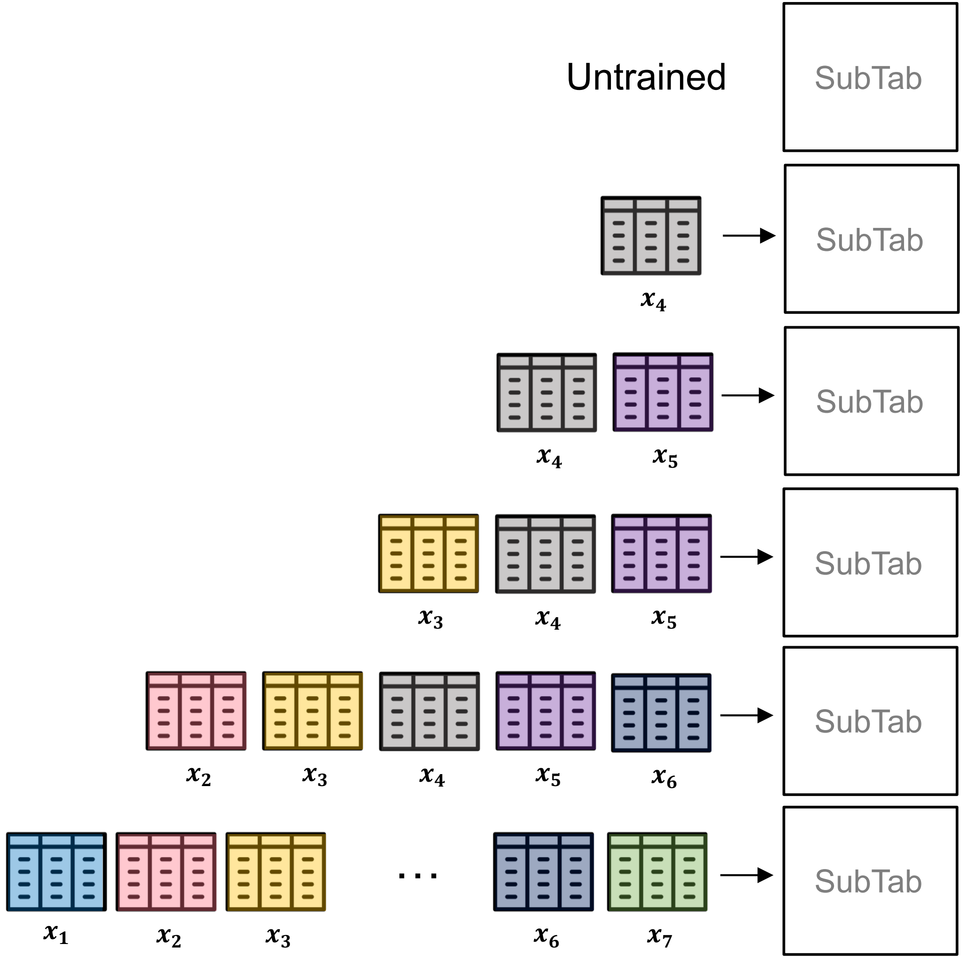
Additional Experiments with MNIST

Slicing MNIST digits to 7 subsets

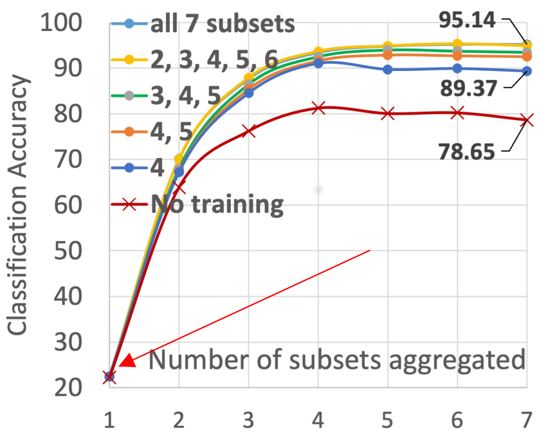


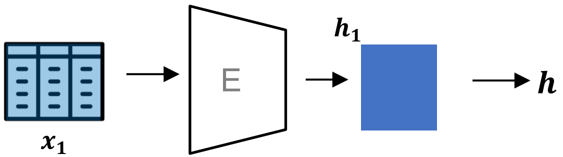


Trained 6 different models

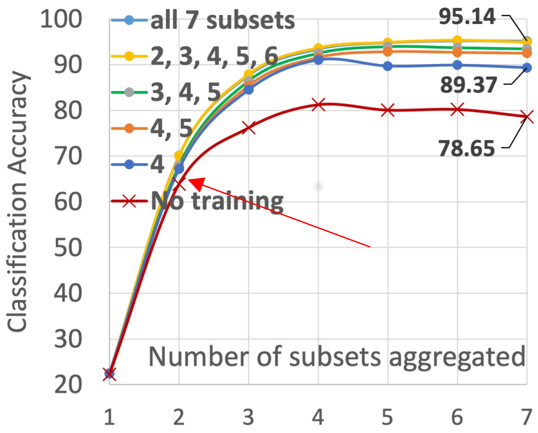


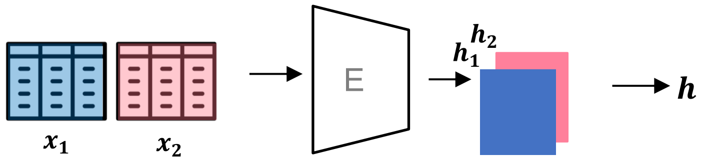
Information content in the joint embedding





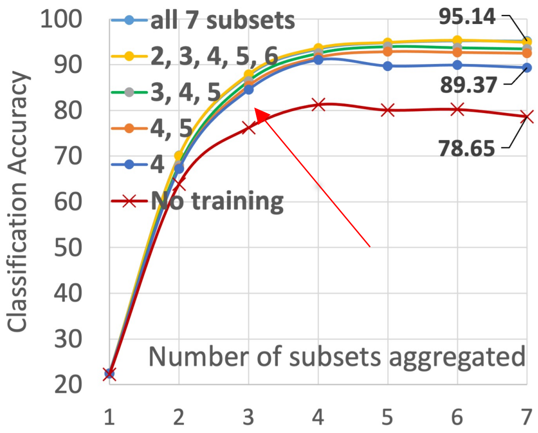
Information content in the joint embedding

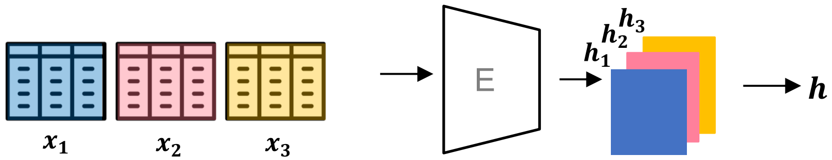




Aggregation (pooling)

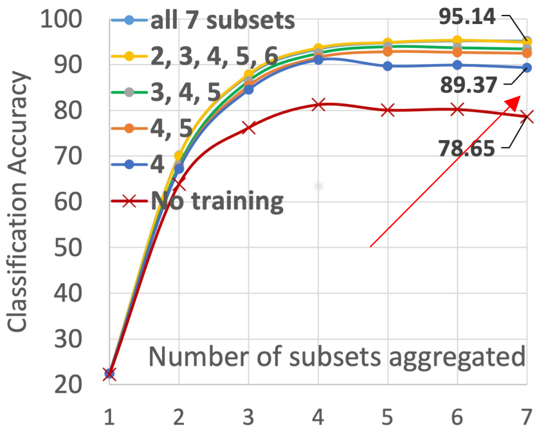
Information content in the joint embedding

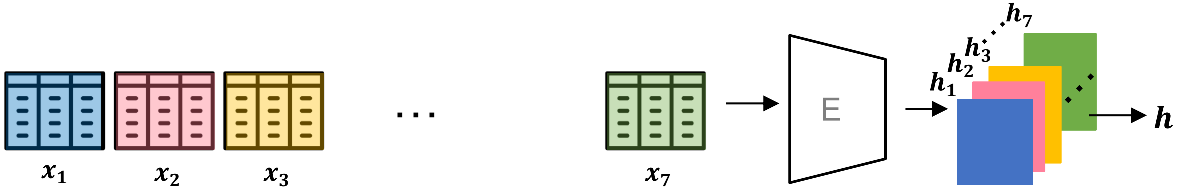




Aggregation (pooling)

Information content in the joint embedding



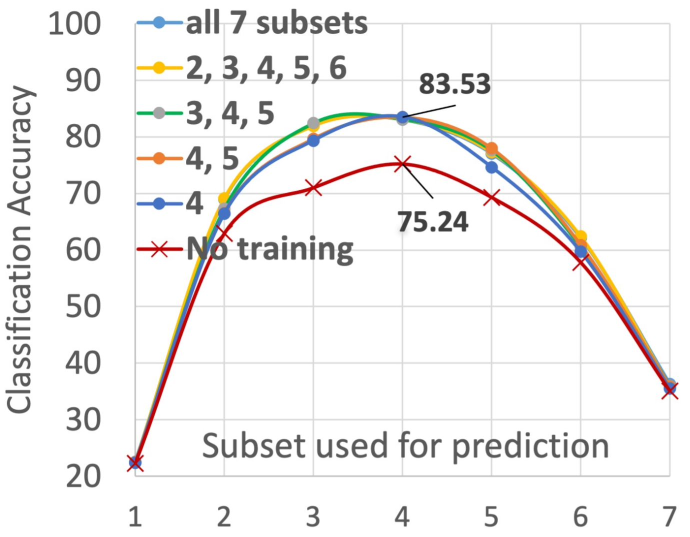


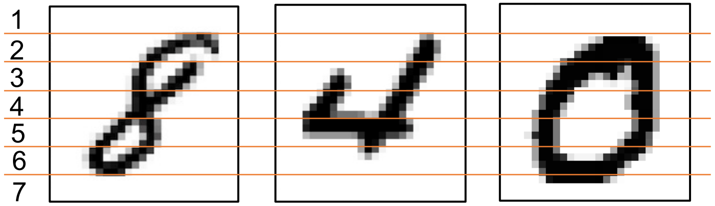
Aggregation (pooling)

- We can have missing features during training and/or test time, and still perform well.

Experiment-2

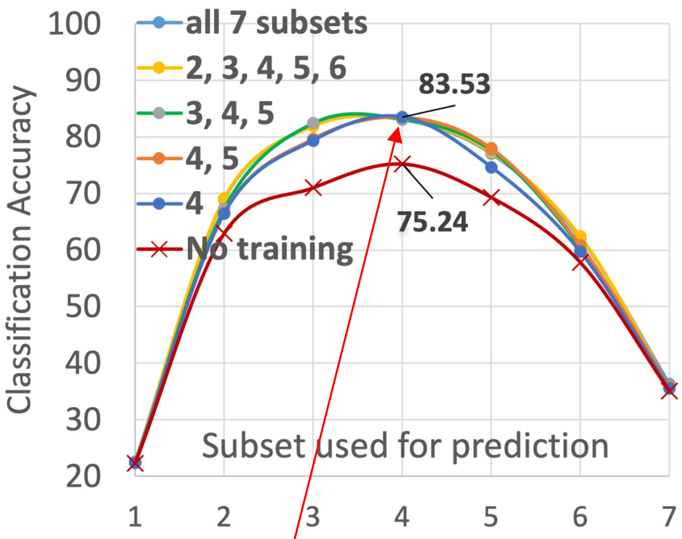
Information content of individual subsets

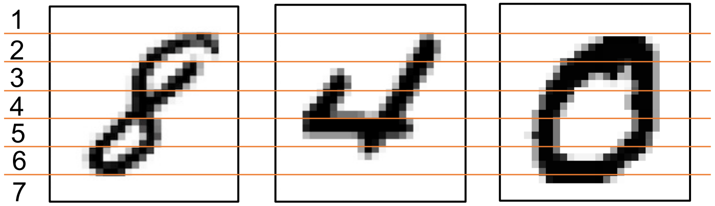




- ❑ We can discover informative subsets
- ❑ We can even use untrained model for discovering them

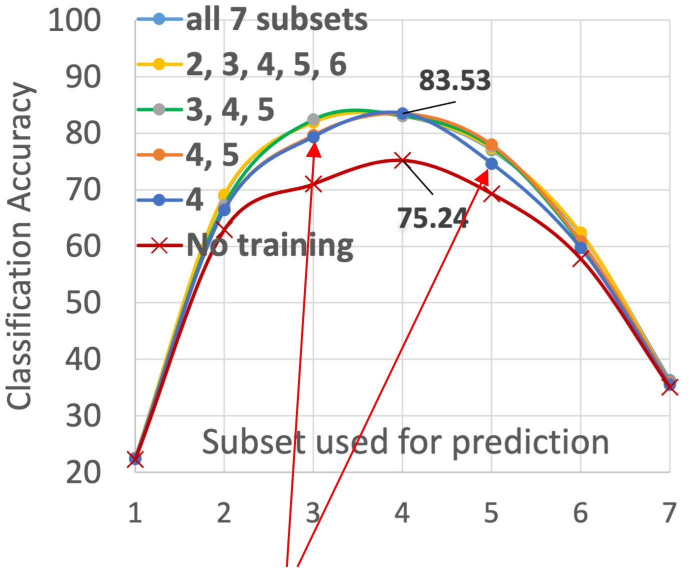
Information content of individual subsets

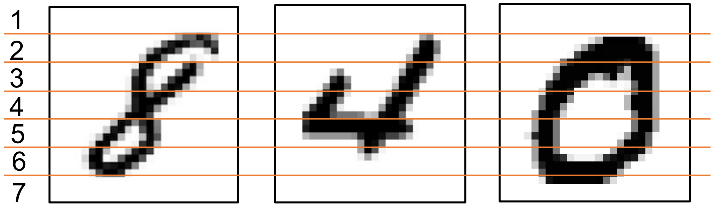




- ❑ Information content of a subset does not depend on other subsets
  - ❑ All models trained on subset 4 has same performance

Information content of individual subsets

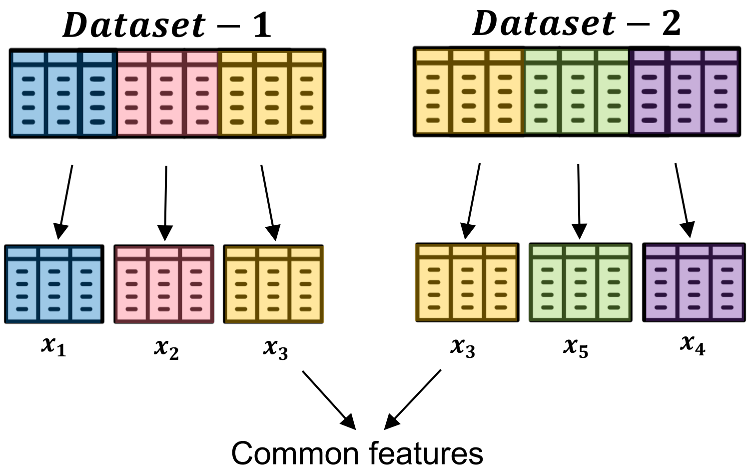




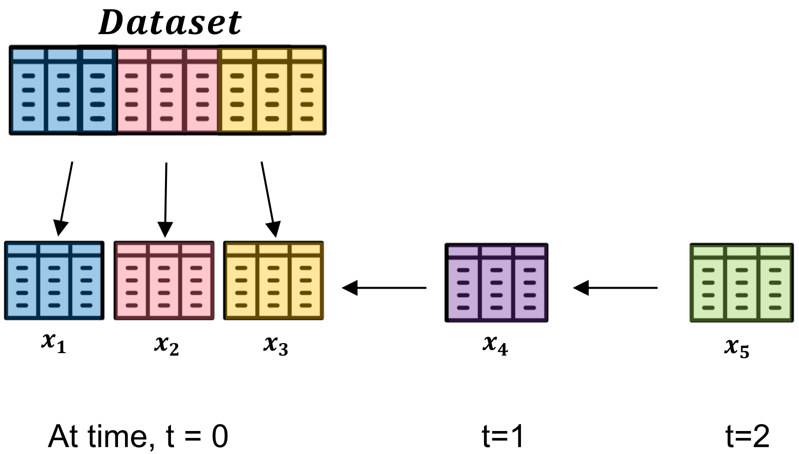
- ❑ Information content of a subset does not depend on other subsets
  - ❑ All models trained on subset 4 has same performance
  - ❑ Same can be seen for subset 3 and 5

Possible applications

- Transfer learning



- Integrating new features over time



- And there are others such as distributed training, multi-modal learning and so on.

Possible applications – Tabular data with missing features

Electronic Health Records

Patient	Age	BMI	…	Gender	Smoke	…	Income	Diet
1	67	23.2	…	F	N	…	16.29	HP
2	45	30	…	M	Y	…	14.64	MP
3	74	22.6	…	?	Y	…	?	MP
4	32	28.1	…	F	Y	…	?	?
5	54	?	…	M	N	…	15.28	LP

Possible applications – Tabular data with missing features

Electronic Health Records

Patient	Age	BMI
1	67	23.2
2	45	30
3	74	22.6
4	32	28.1
5	54	?

…	…
…	…
…	…
…	…
…	…
…	…

Income	Diet
16.29	HP
14.64	MP
?	MP
?	?
15.28	LP

Gender
F
M
?
F
M

Smoke
N
Y
Y
Y
N

Possible applications – Tabular data with missing features

Electronic Health Records

Patient	Age	BMI	Smoke	…	…	Gender	Income	Diet
1	67	23.2	N	…	…	F	16.29	HP
2	45	30	Y	…	…	M	14.64	MP
3	74	22.6	Y	…	…	?	?	MP
4	32	28.1	Y	…	…	F	?	?
5	54	?	N	…	…	M	15.28	LP

Possible applications – Tabular data with missing features

Electronic Health Records

Patient	Age	BMI	Smoke	…	…	Gender	Income	Diet
1	67	23.2	N	…	…	F	16.29	HP
2	45	30	Y	…	…	M	14.64	MP
3	74	22.6	Y	…	…	?	?	MP
4	32	28.1	Y	…	…	F	?	?
5	54	?	N	…	…	M	15.28	LP

Subset 1 … Subset 2

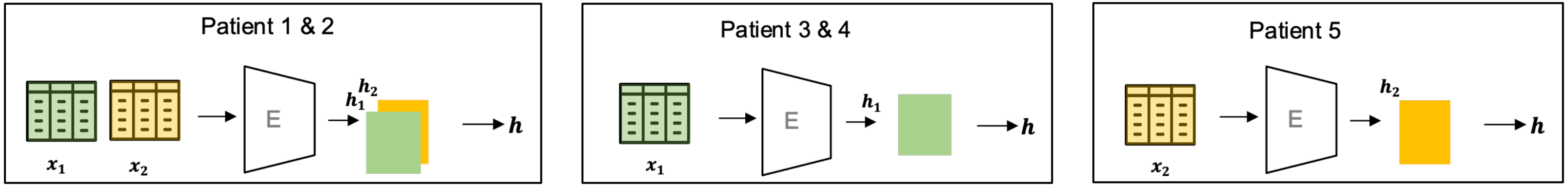
Possible applications – Tabular data with missing features

Electronic Health Records

Patient	Age	BMI	Smoke	…	…	Gender	Income	Diet
1	67	23.2	N	…	…	F	16.29	HP
2	45	30	Y	…	…	M	14.64	MP
3	74	22.6	Y	…	…	?	?	MP
4	32	28.1	Y	…	…	F	?	?
5	54	?	N	…	…	M	15.28	LP

Subset 1 … Subset 2

Personalized modelling



Summary

- We showed a new method for representation learning using tabular data

- But the problem is no way solved:
  - Tabular data comes in many forms
  - There is no single solution that can fit all situations

- We will continue developing methods to address existing challenges

Thanks!

GitHub: https://github.com/AstraZeneca/SubTab