# Why does synthesized data improve multi-sequence classification?

Gijs van Tulder and Marleen de Bruijne

### **Motivation**

Data synthesis methods can be used to predict missing modalities in incomplete datasets. Some people have found that using synthetic data can improve classification accuracy.

### Method 1: Neural networks

Neural networks can be trained to predict one modality from the others. Because a network predicts only one modality, you need multiple networks to predict all modalities.

### Method 2: Restricted Boltzmann machines (RBMs)

With RBMs, a single model can be used to predict any of the sequences.

*P*(**h**|**v**)



But synthesis models mostly transform existing information. Why does data synthesis still help classification?



### **Experiments**

We did experiments with two synthesis methods and two classifiers. We used data from the BRATS 2013 brain tumor segmentation dataset.



Five tissue types



# P(v|h)

## **Explanation 1: Synthesis allows more training data**

Using synthetic data, you can train a single classifier on samples with different missing modalities. This gave better results than training on subsets.

full set with all four modalities

training on subsets

### Conclusions

73.2

Synthetic data can improve classification accuracy when training with incomplete datasets.

We proposed synthesis using RBMs to naturally handle incomplete samples.

Synthetic data can help because it

full cot with synthetic data	
Tull set with synthetic data	
use average patch	70.9
neural network synthesis	71.4
RBM synthesis	70.3
	random forest classification accuracy

59.5 – 69.0

allows training with more samples and can provide useful transformations.

This depends on the type of classifier.

### **Explanation 2: Synthesis models provide new transformations of the data**

Data synthesis models can provide nonlinear transformations of the data, which can be useful for classifiers that cannot do this on their own.

Synthetic data improved the accuracy of our linear SVMs...

68.8 missing one modality no T1 67.9 synthetic T1 from neural network 67.4 synthetic T1 from RBM 69.3 no T1c 58.7 More flexible classifiers can extract the same information from the original data and will not become much better from using synthetic data.

# but synthetic data did not improve our random forests.

all four modalities

			73.2
missing one modality			
no T1			73.0
synthetic T1 from neural network			73.0
synthetic T1 from RBM			73.3
no T1c	61.6		
synthetic T1c from neural network	62.2		
synthetic T1c from RBM	61.5		
no T2			72.9
synthetic T2 from neural network			72.9
synthetic T2 from RBM			72.5
no FLAIR		69.6	
synthetic FLAIR from neural network		69.9	
synthetic FLAIR from RBM		69.9	

The classifiers can also be trained with values from the RBM's hidden layer. This is a non-linear transformation of the input, so it will be more useful to a simple linear classifier than to more flexible classifiers.

Training on the RBM's hidden layer helped the linear SVMs more than the random forests.

### training on images

ar SVM	68.8



random forest 73.2 training on the RBM hidden layer

linear SVM 72.9 random forest 74.2 classification accuracy

random forest classification accuracy

**Biomedical Imaging Group Rotterdam, the Netherlands** www.bigr.nl · g.vantulder@erasmusmc.nl

MICCAI October 2015, München, Germany

This research is financed by the Netherlands Organization for Scientific Research (NWO).



zajns