Machine Learning Algorithms for Low Data Environments

Alan Lu, Department of Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign

Mentors: Neil Getty

Supervisor: Fangfang Xia

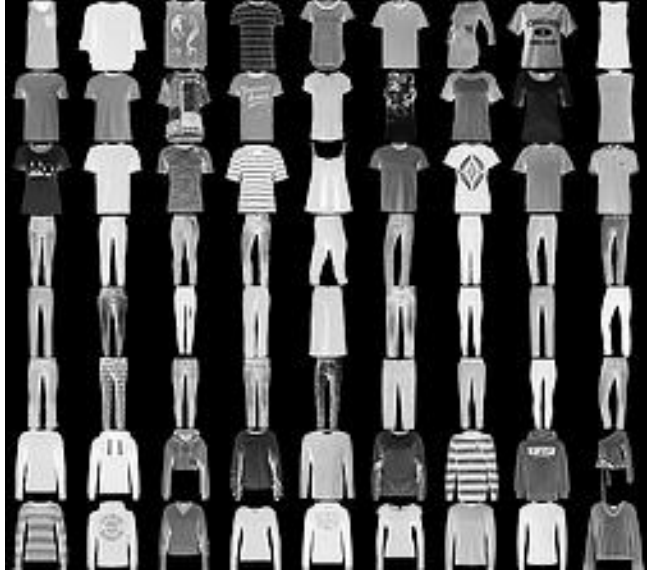
Argonne NATIONAL LABORATORY

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Motivation

- Given a large amount of data, current state-of-the-art machine learning algorithms perform with high accuracy
- In many industries obtaining data is expensive and difficult
- It is necessary to apply machine learning to these industries with small data
  - Cancer patient data
  - Pharmaceutical applications
  - One-shot learning applications - image recognition, drug discovery

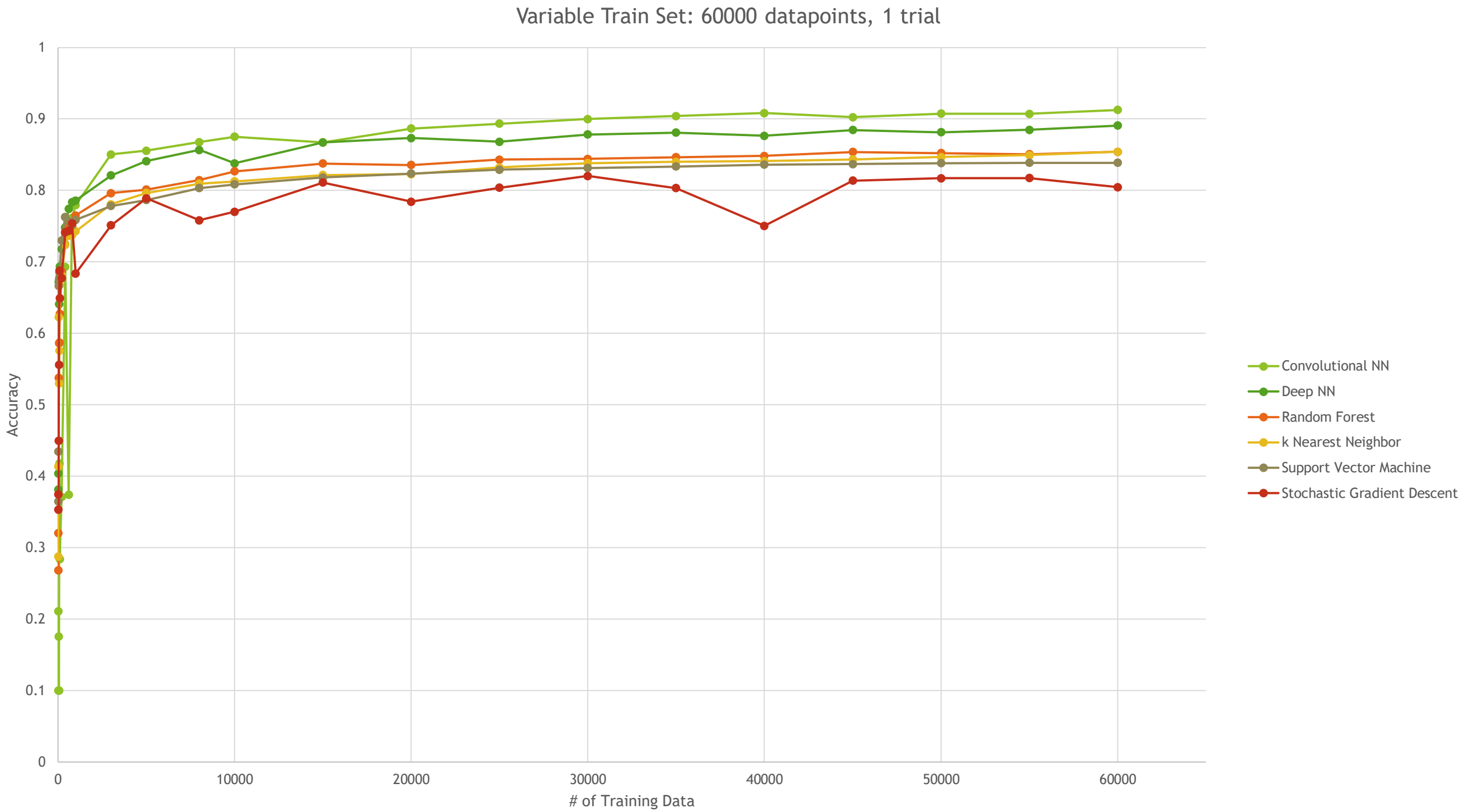
Datasets

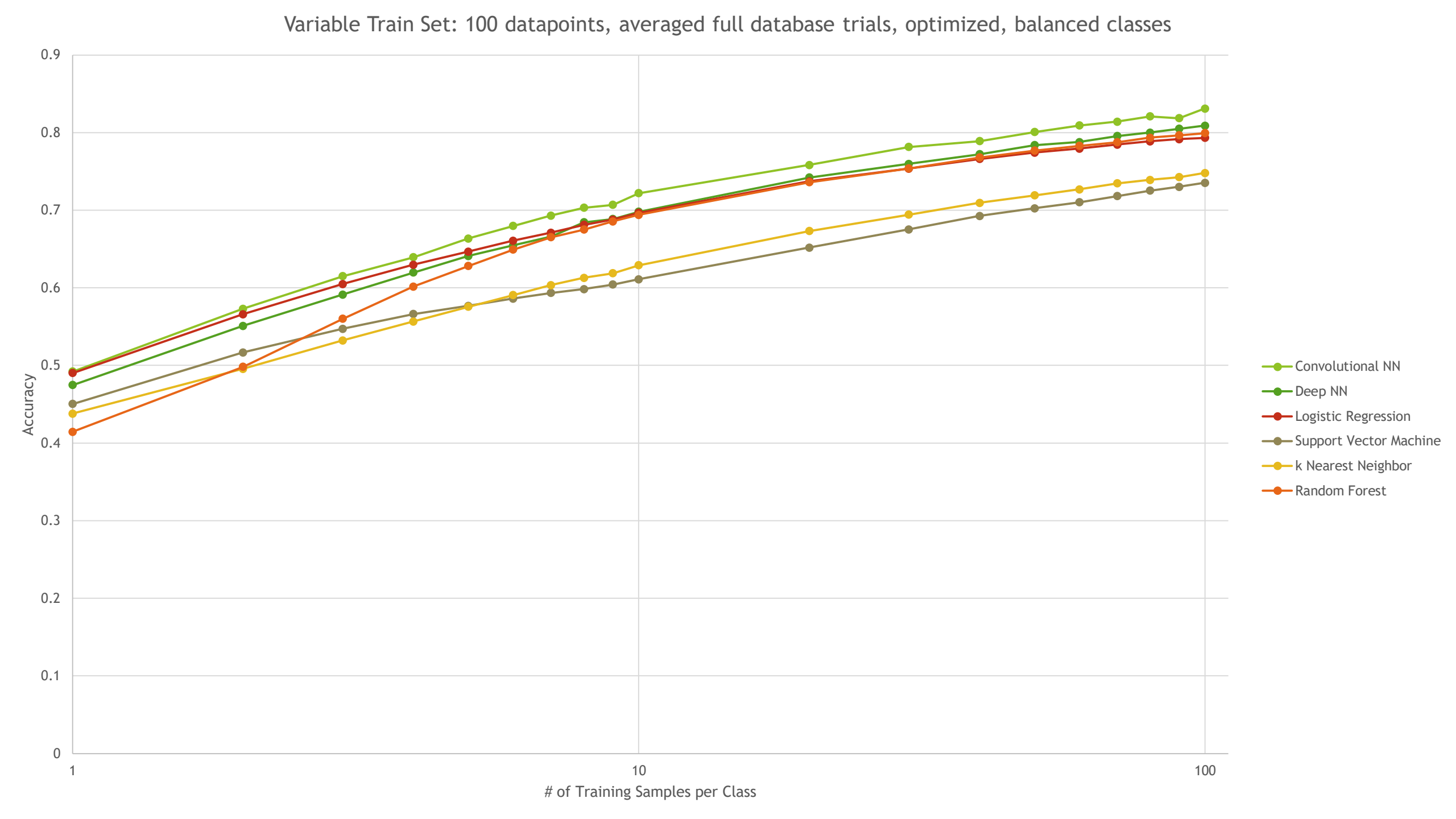


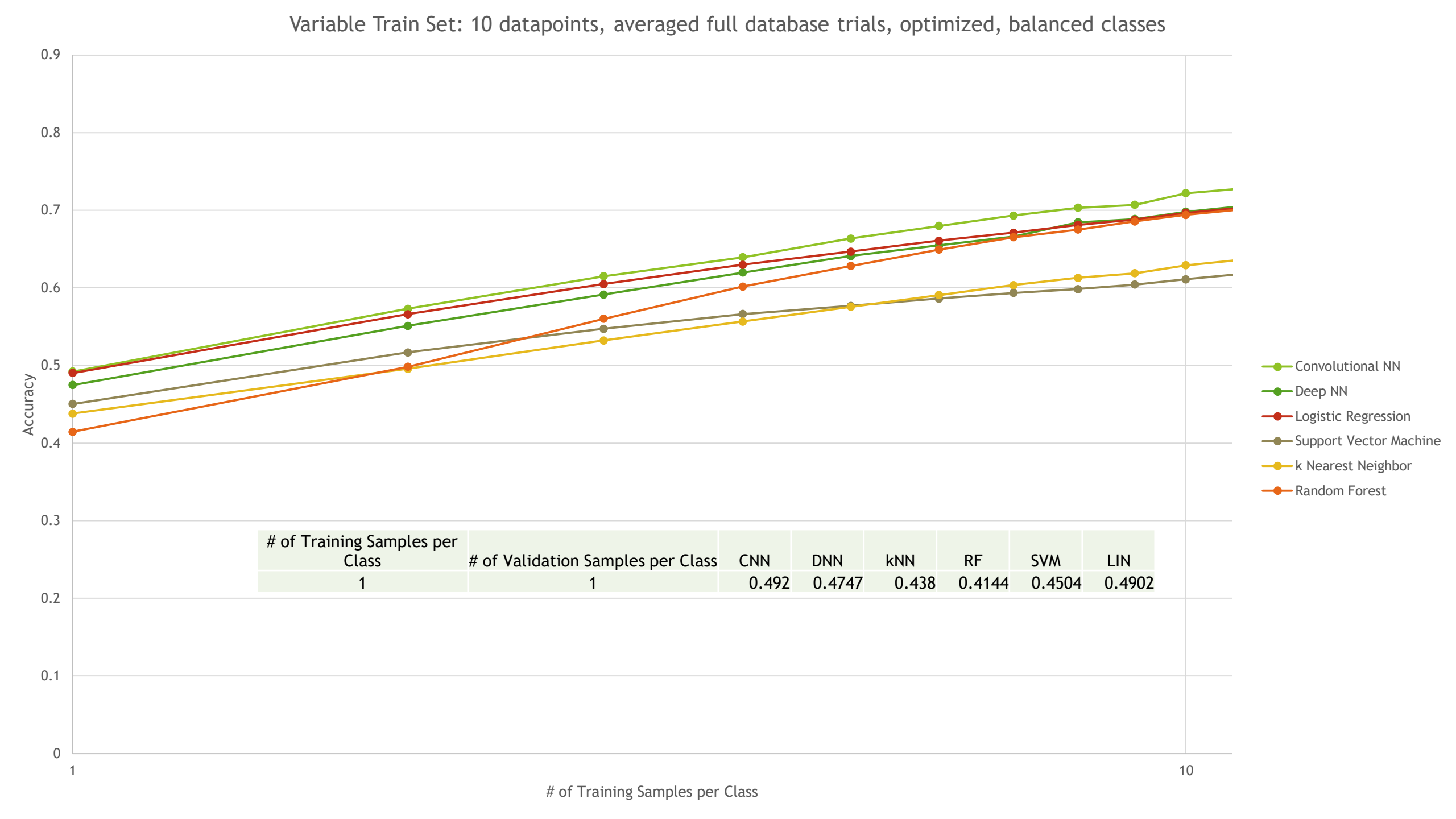
- Fashion-MNIST Dataset
  - 70000 labeled 28x28 images – 60000 training, 10000 testing
  - 10 fashion categories – T-shirt, Dress, Coat, Sandal, etc.
- MNIST Dataset
  - 70000 labeled 28x28 images – 60000 training, 10000 testing
  - 10 handwritten digits
- Omniglot Dataset
  - “Transpose of MNIST”
  - 1623 different handwritten characters from 50 different alphabets – 30 training, 20 testing
  - 20 examples of every character
  - Mongolian, Armenian, Hebrew, (Futurama, Alphabet of the Magi)
- Why these?
  - Clean – no corrupted files
  - Image Datasets – intuitive, CNN-applicable
  - Lots of Documentation, Academic
  - Good for low-shot learning - Omniglot

Algorithms tested

- Convolutional Neural Network
- Deep Neural Network
- Random Forest
- K Nearest-Neighbor
- Support Vector Machine
- Logistic Regression
- ~~Stochastic Gradient Descent~~


Variable Train Set: 60000 datapoints, 1 trial
Accuracy
# of Training Data
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
10000
20000
30000
40000
50000
60000
Convolutional NN
Deep NN
Random Forest
k Nearest Neighbor
Support Vector Machine
Stochastic Gradient Descent


Variable Train Set: 100 datapoints, averaged full database trials, optimized, balanced classes
Accuracy
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
10
100
# of Training Samples per Class
Convolutional NN
Deep NN
Logistic Regression
Support Vector Machine
k Nearest Neighbor
Random Forest



# of Training Samples per Class	# of Validation Samples per Class	CNN	DNN	kNN	RF	SVM	LIN
1	1	0.492	0.4747	0.438	0.4144	0.4504	0.4902

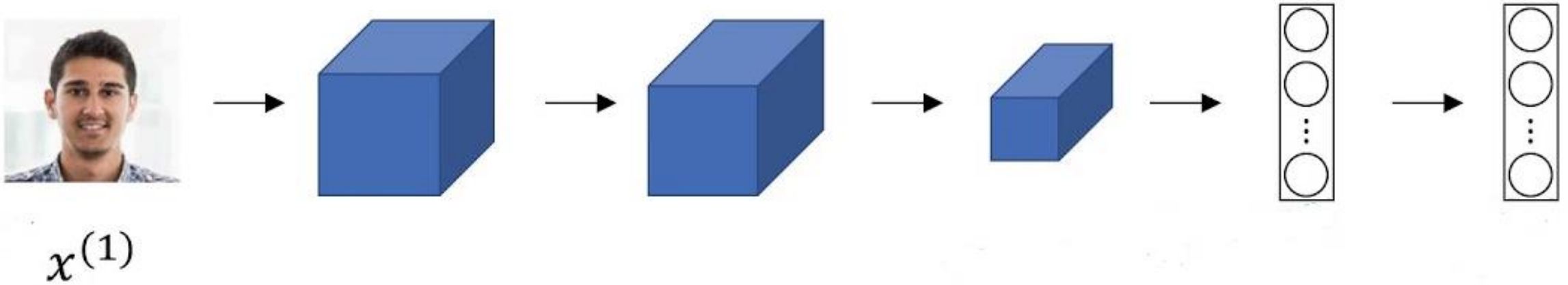
One-Shot Learning

4-way one-shot learning

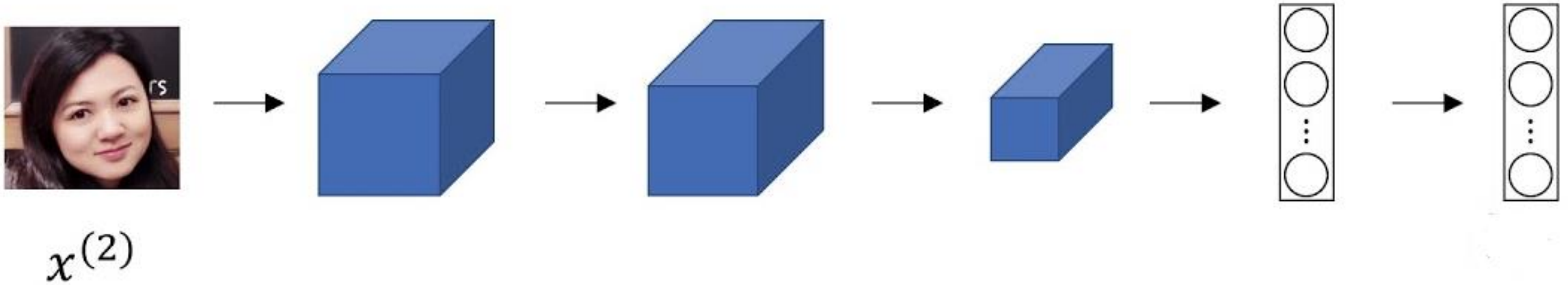
Specialized techniques

- **Siamese Network**
- Triplet Loss
- CapsuleNet

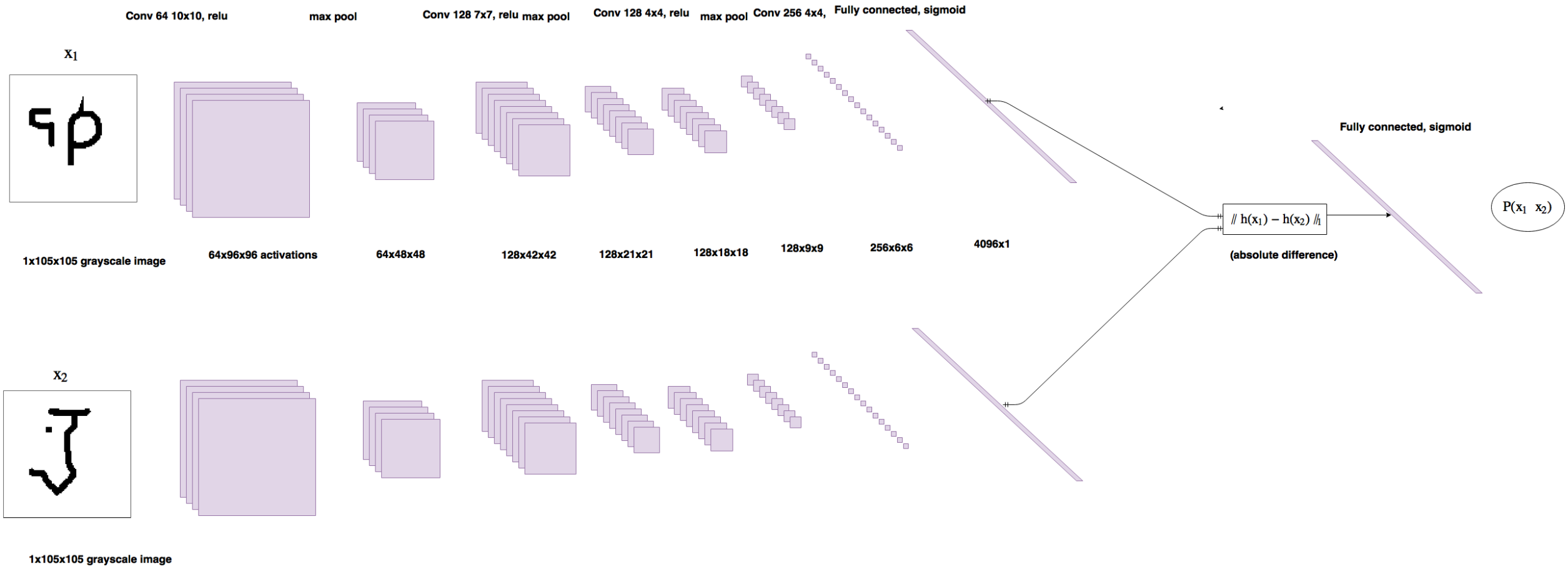
Siamese Neural Network (SNN)



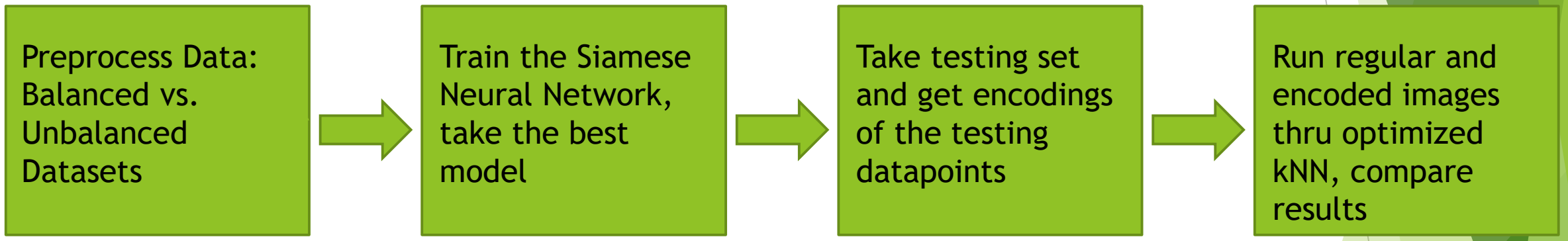
$x^{(1)}$



$x^{(2)}$


Conv 64 10x10, relu
max pool
Conv 128 7x7, relu
max pool
Conv 128 4x4, relu
max pool
Conv 256 4x4,
Fully connected, sigmoid
x_1
x_2
1x105x105 grayscale image
64x96x96 activations
64x48x48
128x42x42
128x21x21
128x18x18
128x9x9
256x6x6
4096x1
$\| h(x_1) - h(x_2) \|_1$
(absolute difference)
Fully connected, sigmoid
$P(x_1, x_2)$
1x105x105 grayscale image

How we use SNNs



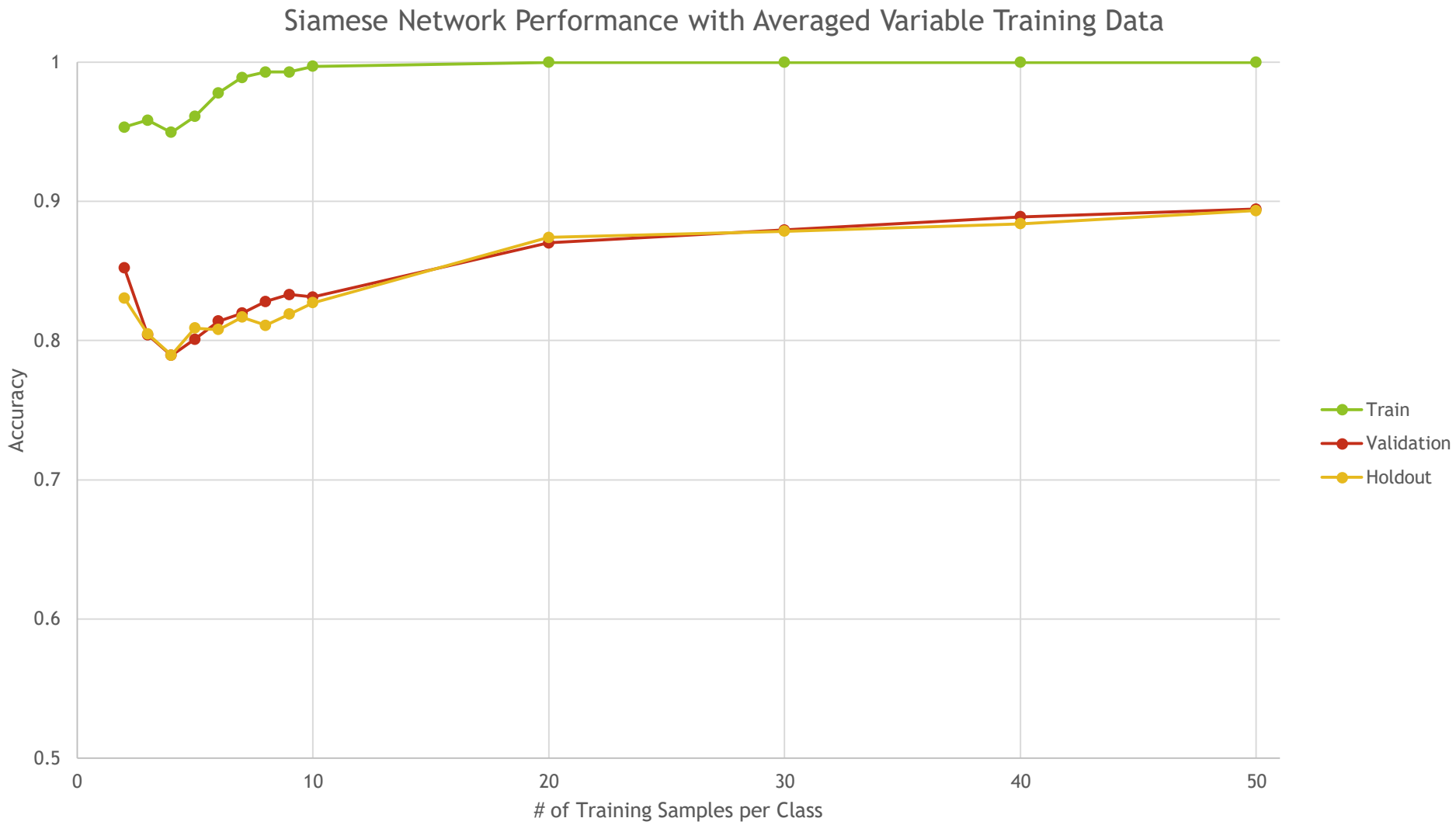
SNN Study 1: Variable Train Set

- Continuation of the first experiment – how much does the Siamese network improve the accuracy for a variable train set?

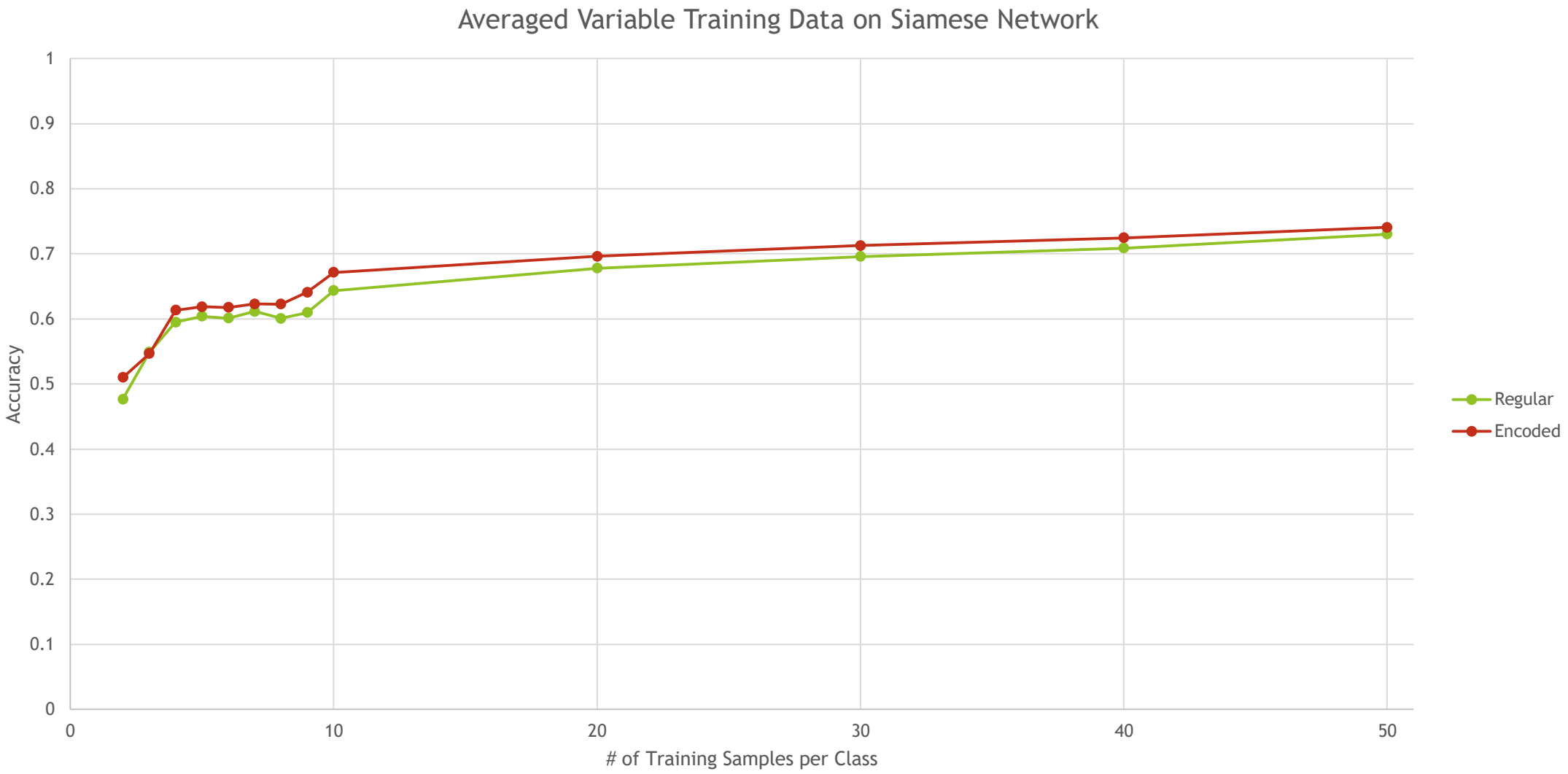
SNN Study 1: Variable Train Set – Network Accuracy

Network Scores:		
Train	Val	Hold
98.67%	86.62%	86.44%

Results for half similar pairs, half different pairs on 50 samples per class



SNN Study 1: Variable Train Set – Testing Accuracy



SNN Study 2: Using entire dataset

- 30000 images to train Siamese Network
  - Subset of 150 images per class
  - $\binom{1500}{2} = 1124250$ pairs compared to $\binom{500}{2} = 124750$ pairs
- 30000 images to test accuracies on an optimized kNN

Balanced split

Network Accuracies:			Testing Accuracies:	
Train	Val	Hold	Regular	Encoded
99.99%	89.9%	88.58%	76.96%	77.18%

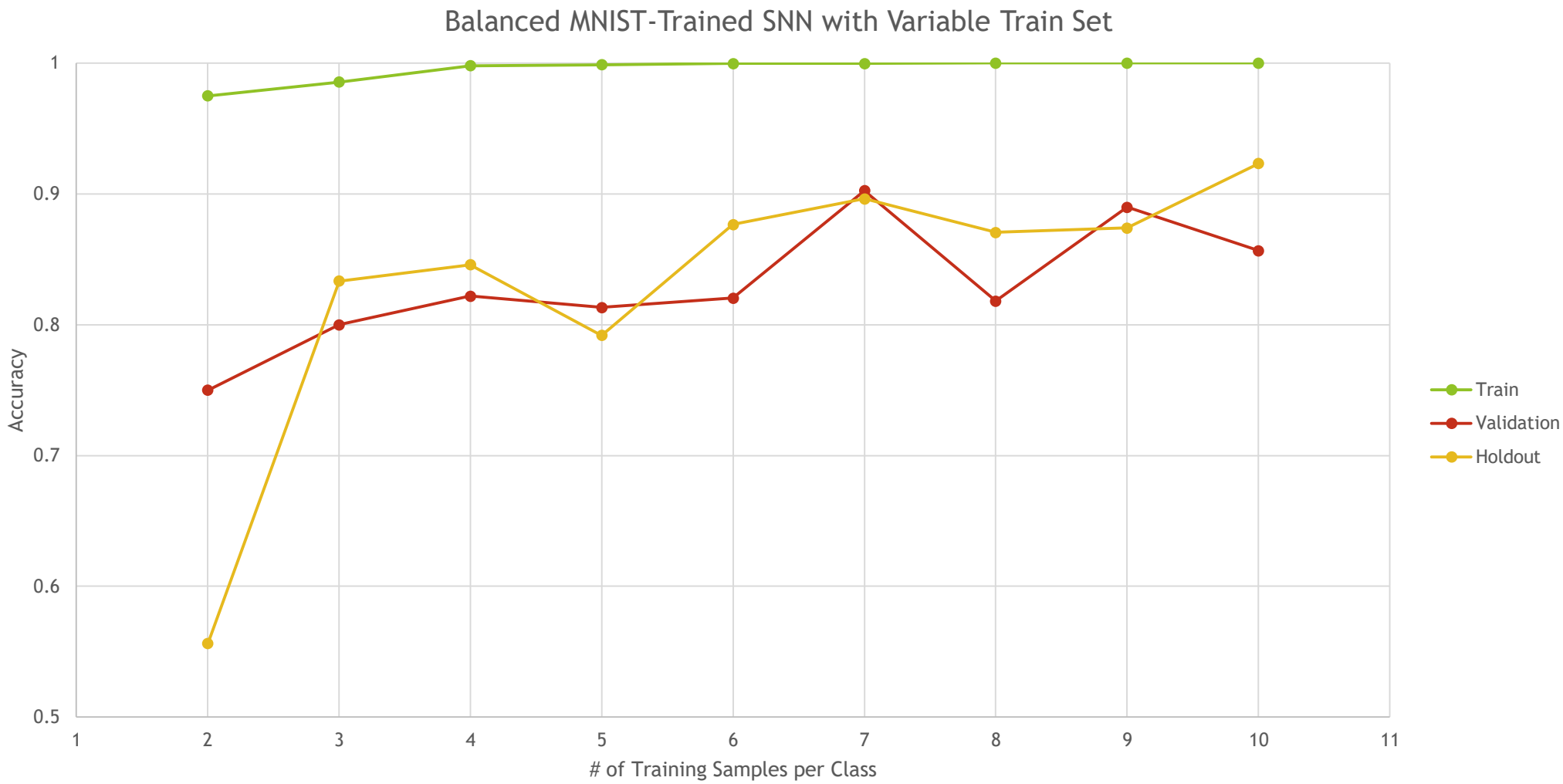
Unbalanced split

Network Accuracies:			Testing Accuracies:	
Train	Val	Hold	Regular	Encoded
99.99%	93.82%	91.73%	76.99%	83.15%

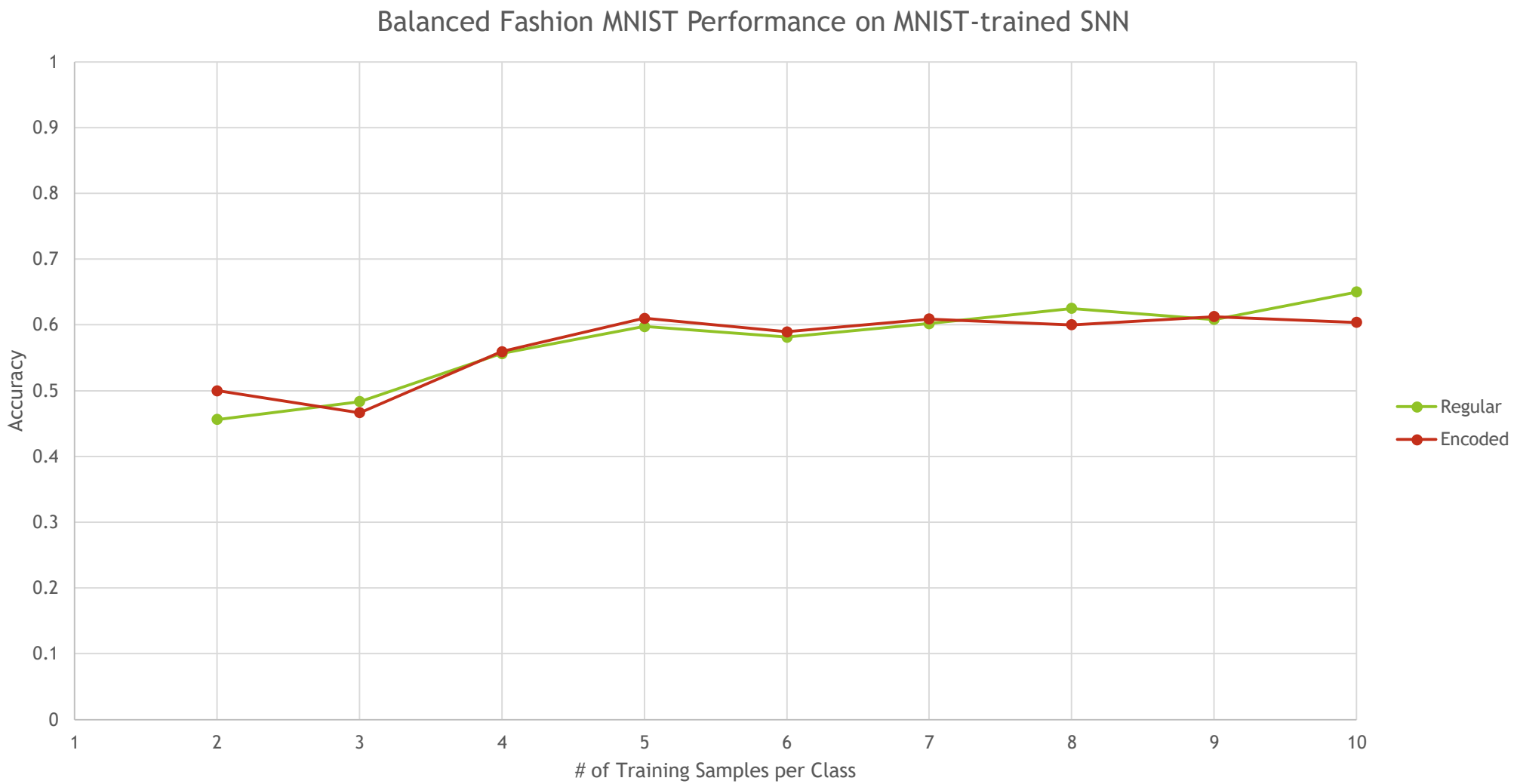
SNN Study 3: Generalization of features

- Paper: Trained model with omniglot, tested on MNIST – 70.3%
- MNIST dataset → Fashion MNIST dataset

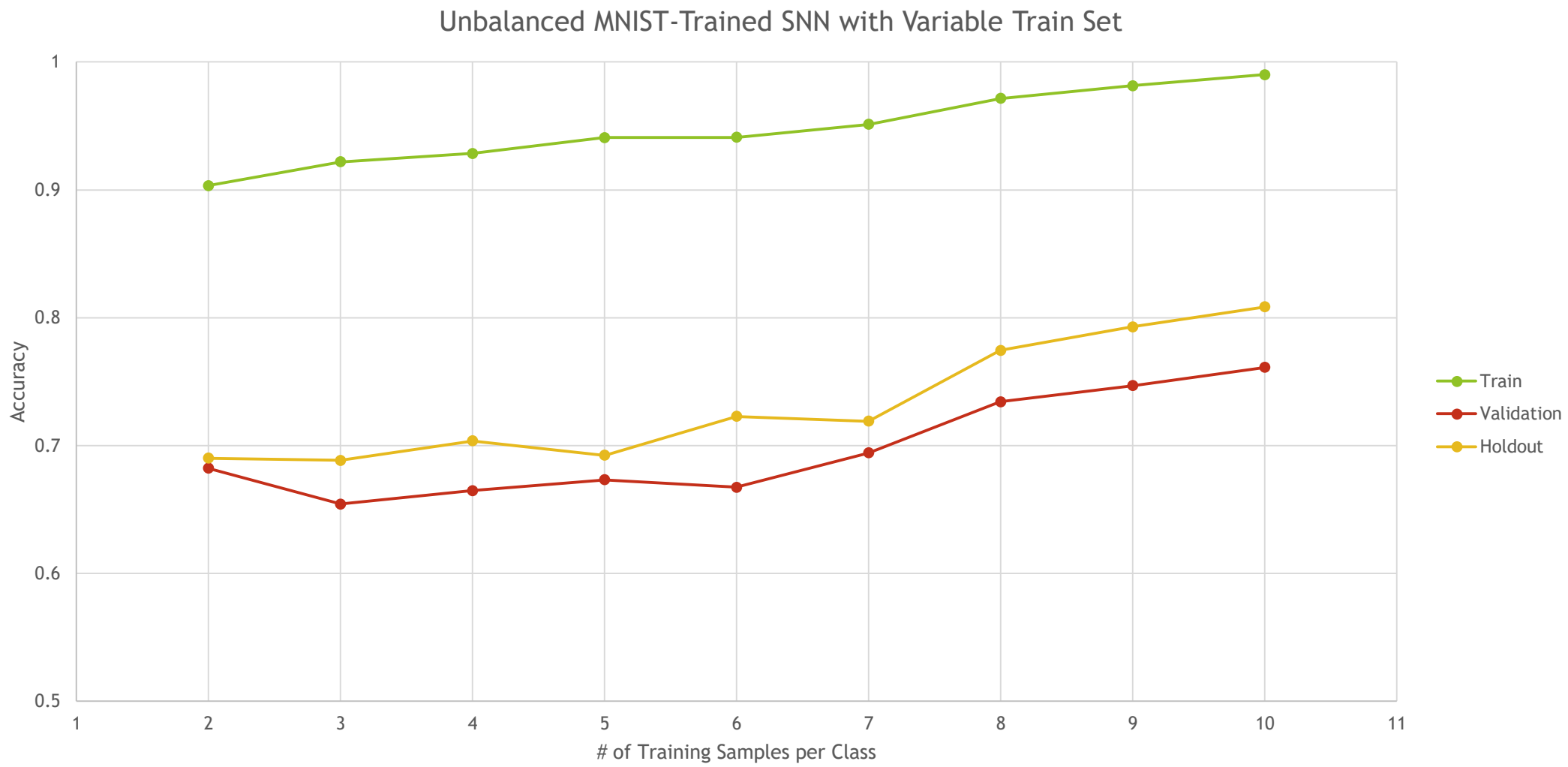
SNN Study 3 - Network Accuracy, Balanced



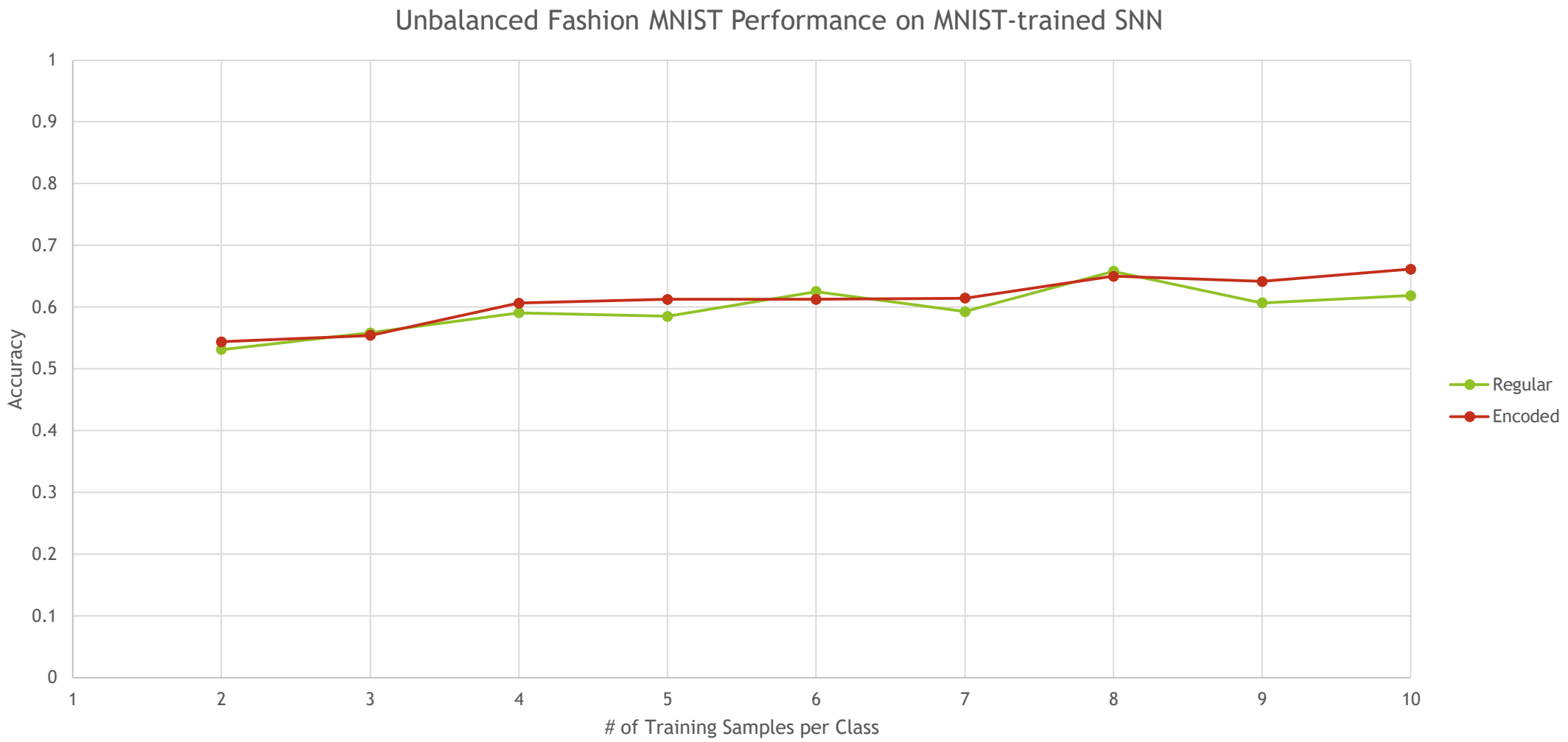
SNN Study 3 - Testing Accuracy, Balanced



SNN Study 3 - Network Accuracy, Unbalanced

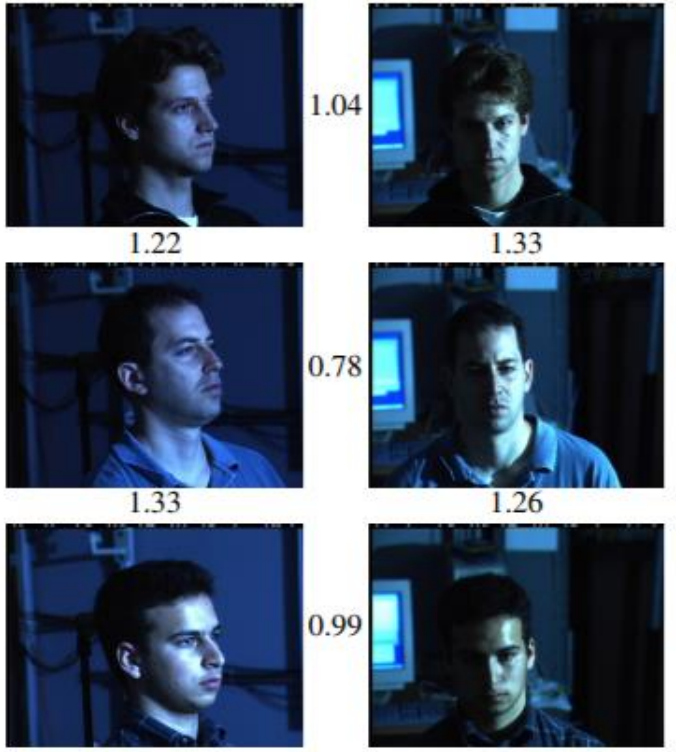


SNN Study 3 - Testing Accuracy, Unbalanced



Future work

- Another SNN Study: Train with 150 pts per class, test on variable, small dataset
- Finetune the models for study 3
- Triplet Loss Algorithm
- CapsuleNet Algorithm
- Apply algorithms on Omniglot Dataset
- Apply algorithms on cancer dataset



Related Papers: One-Shot Learning

- “Low Data Drug Discovery with One-Shot Learning”
- “Active One-Shot Learning”

Acknowledgements

Thanks to Argonne National Labs and the Department of Energy for making this project possible. Additional thanks to Dr. Fangfang Xia and Neil Getty for their continuous expertise and support throughout the internship.