SAND2021-12349PE

Sandia National Laboratories

Exceptional service in the national interest

Classification of Optical Ports Using Machine Learning

Abigail Pribisova

MARTIANS End of Summer Symposium 2021

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. ation under contract DE-NA0003525.

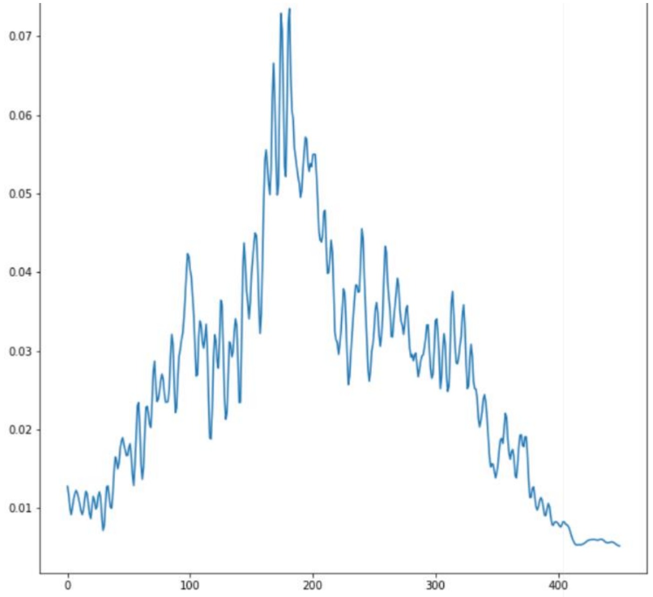
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

U.S. DEPARTMENT OF ENERGY

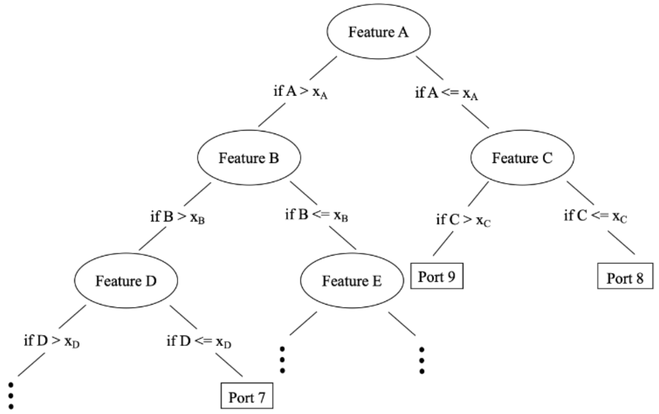
NNSA

Objective

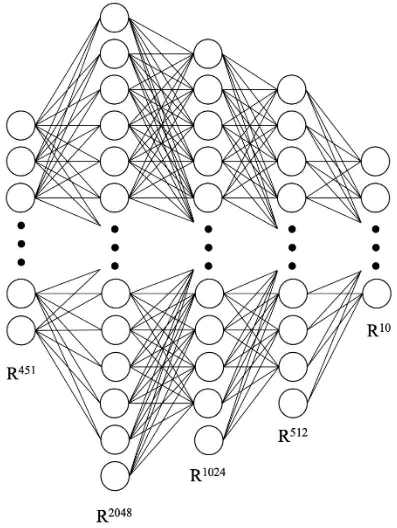
- Predict the optical port from which a particular spectrum was produced
- Compare the effectiveness of decision trees versus fully-connected neural networks



Port 2

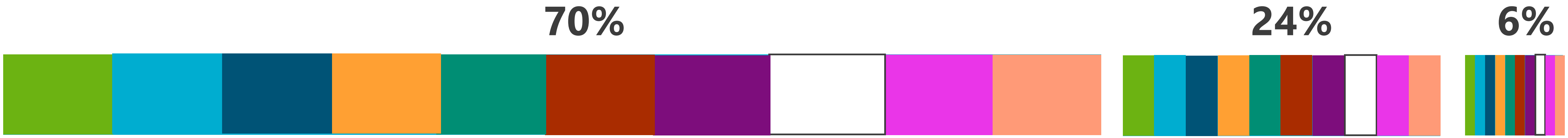


vs.



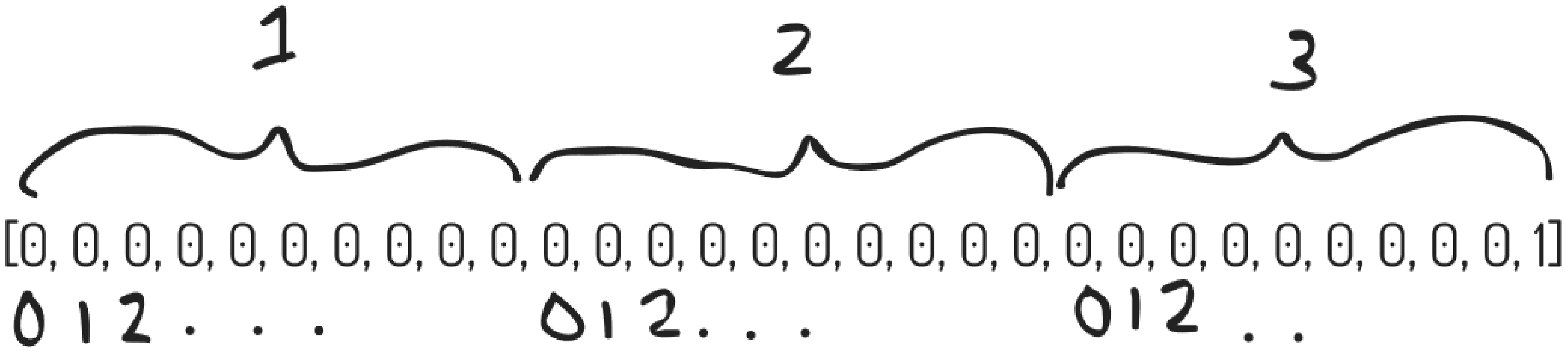
Data Generation

- 750,000 spectrum-port pairs – 250,000 from each optical device
- Spectra = 451-component vectors
- Ports (0-9) = label encoding ([9]) or one-hot vector encoding ([0, 0, 0, 0, 0, 0, 0, 0, 0, 1])

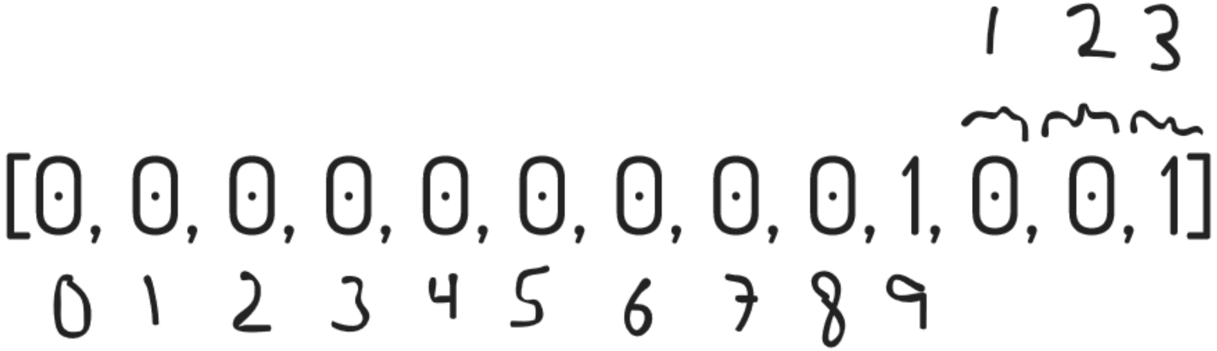


Data Generation cont.

- Multilabel classification -> classify both port number AND optical device
  1. 30-component vector

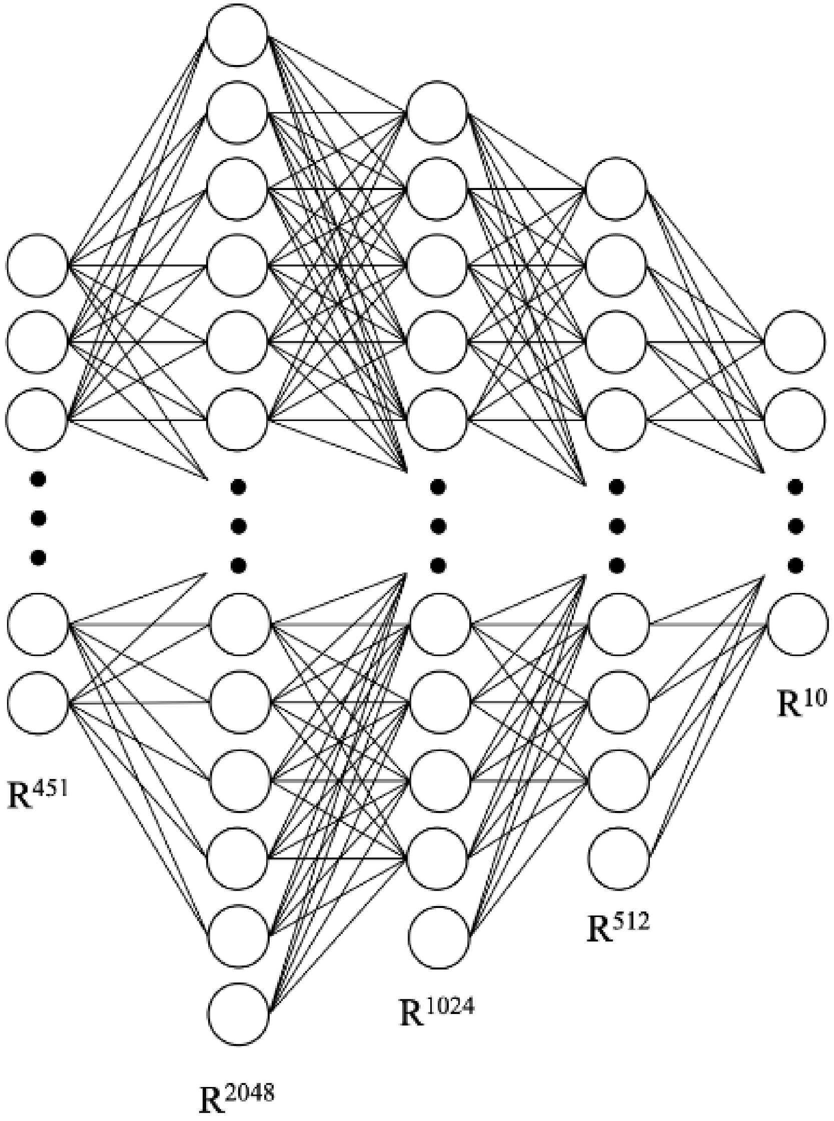


  2. 13-component vector



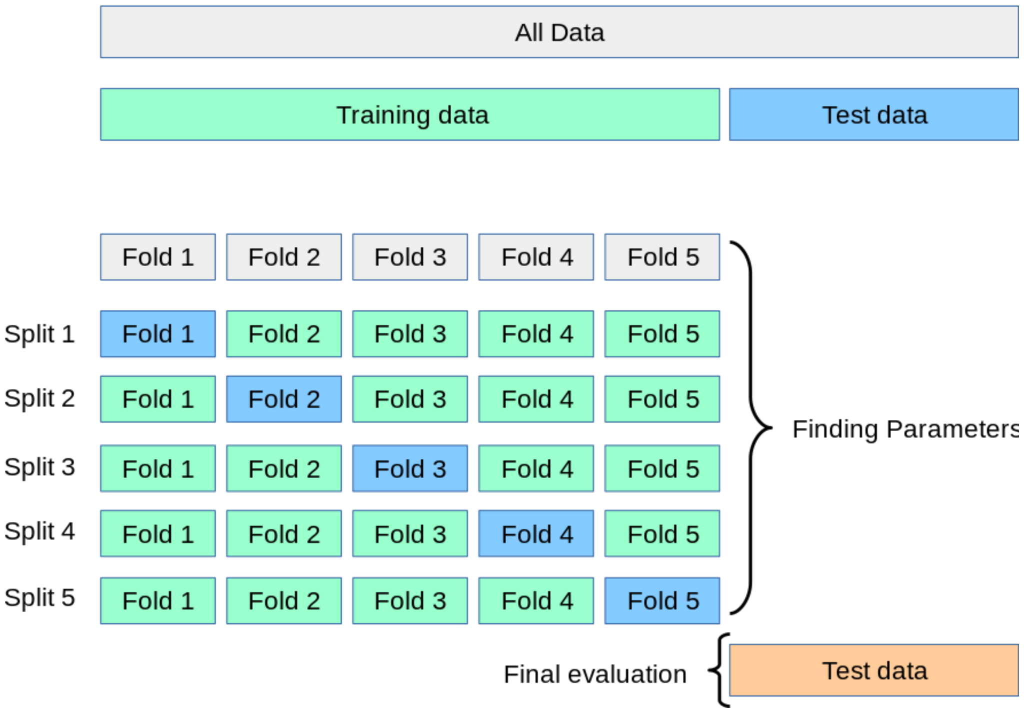
Machine Learning Methods

- Inner activation function: **ReLU**
- Final activation function: **softmax**
- Optimizer: **Adam**
- Loss: **categorical crossentropy**
- Training time: **3 hours on 32 GB GPU**
- BatchNormalization layer
- EarlyStopping function



Machine Learning Methods cont.

- scikit-learn DecisionTreeClassifier
- Criterion: **Gini impurity function**
- Minimum samples: **2**
- Training time: **1 hour on 1 CPU**



All Data

Training data

Test data

Fold 1 Fold 2 Fold 3 Fold 4 Fold 5

Split 1 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5

Split 2 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5

Split 3 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5

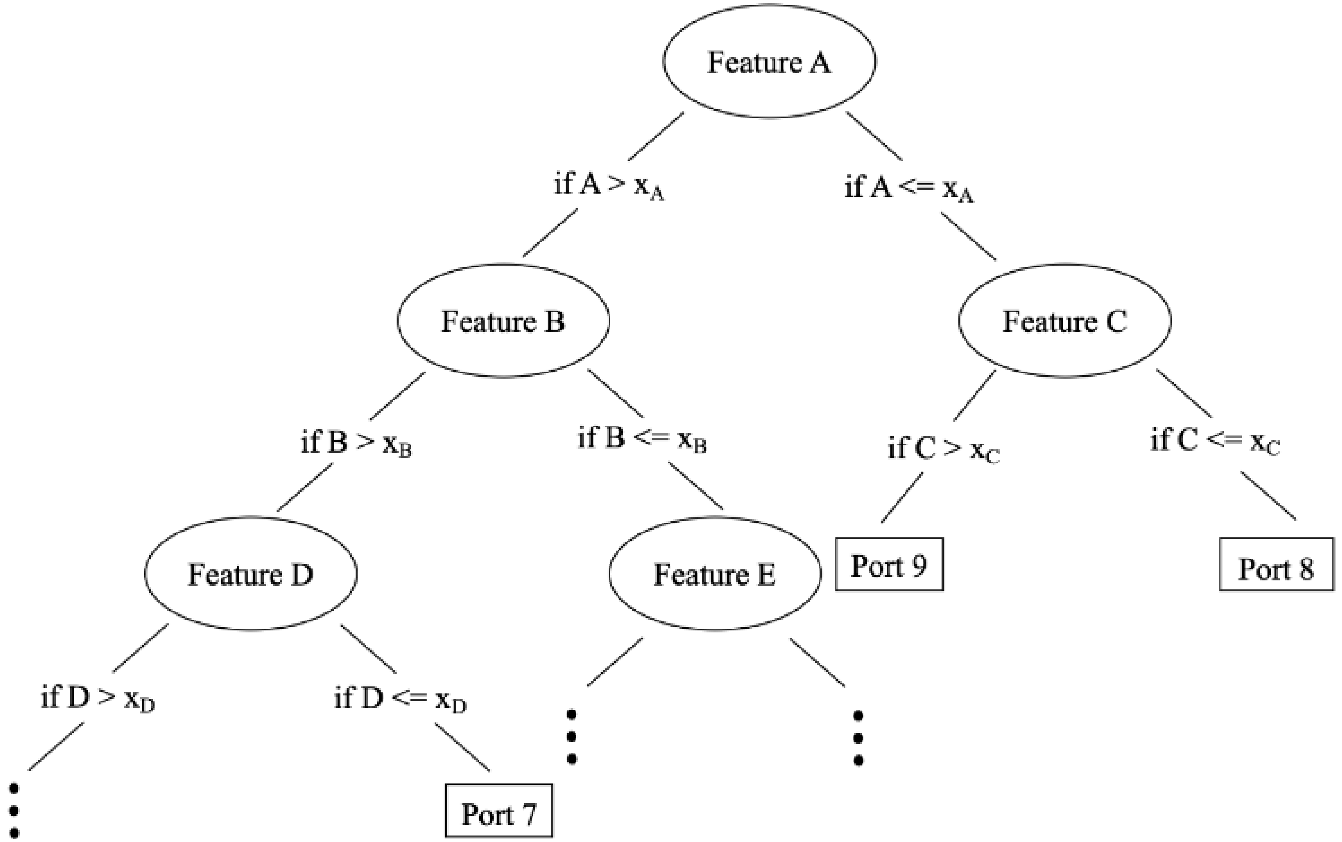
Split 4 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5

Split 5 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5

Finding Parameters

Final evaluation

Test data



Feature A

if A > x_A

if A <= x_A

Feature B

Feature C

if B > x_B

if B <= x_B

if C > x_C

if C <= x_C

Feature D

Feature E

Port 9

Port 8

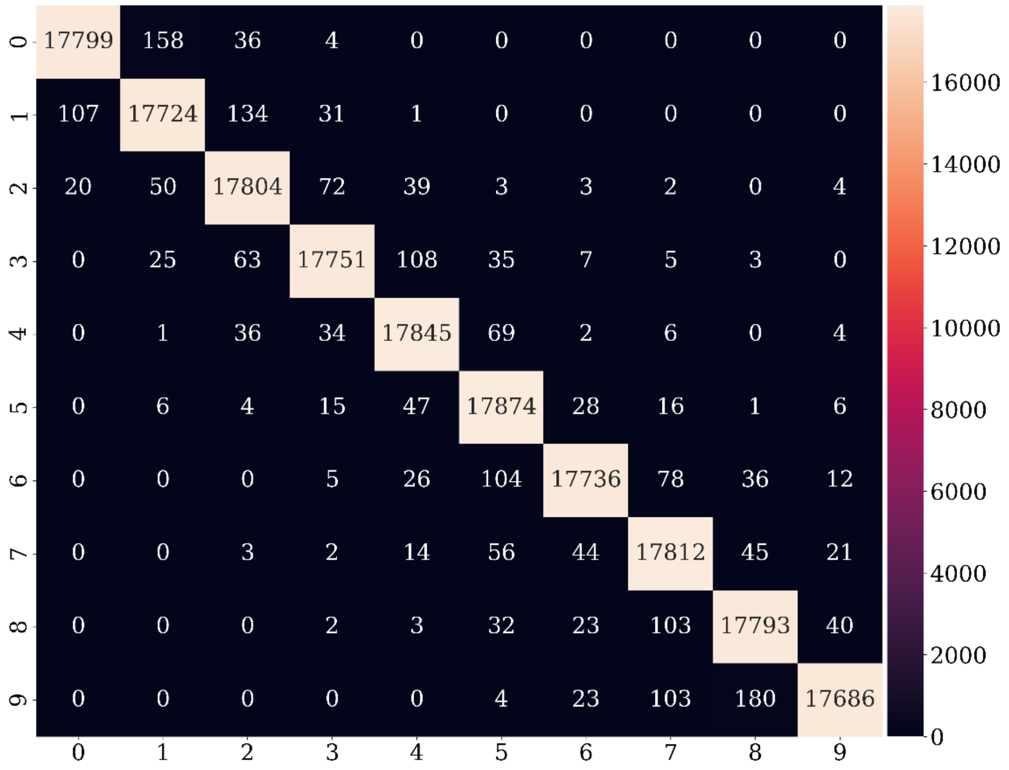
if D > x_D

if D <= x_D

Port 7

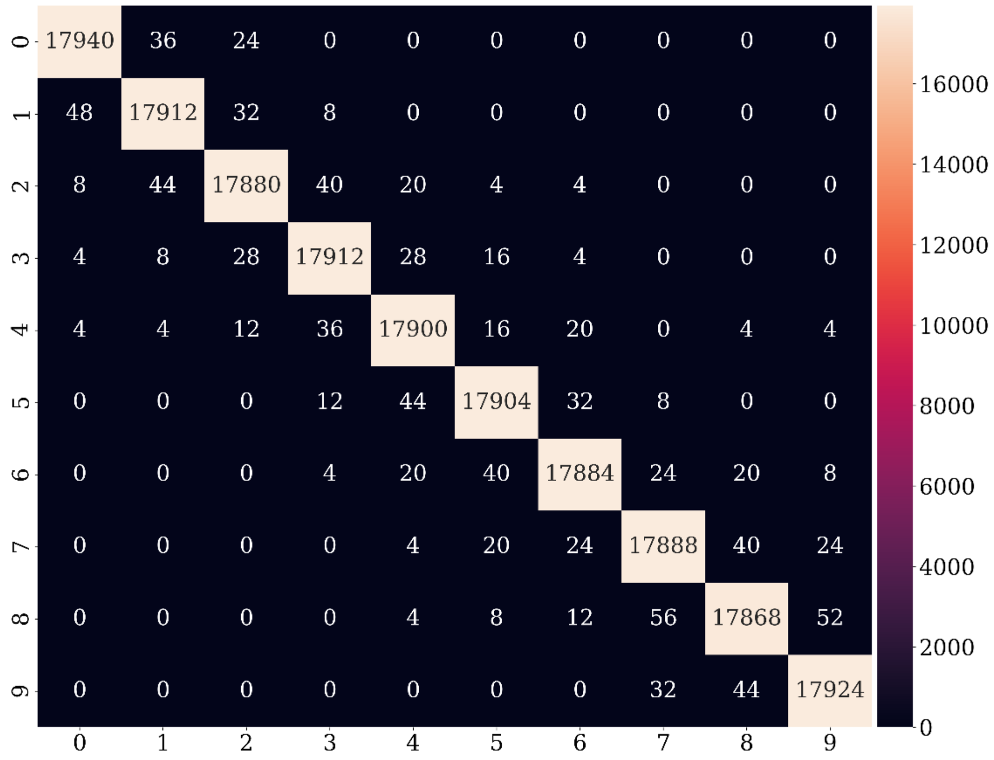
Results

- Neural network: 98.07% (standard deviation: 0.727%)
- Decision tree: 99.43% (standard deviation: 0.0112%)



	0	1	2	3	4	5	6	7	8	9
0	17799	158	36	4	0	0	0	0	0	0
1	107	17724	134	31	1	0	0	0	0	0
2	20	50	17804	72	39	3	3	2	0	4
3	0	25	63	17751	108	35	7	5	3	0
4	0	1	36	34	17845	69	2	6	0	4
5	0	6	4	15	47	17874	28	16	1	6
6	0	0	0	5	26	104	17736	78	36	12
7	0	0	3	2	14	56	44	17812	45	21
8	0	0	0	2	3	32	23	103	17793	40
9	0	0	0	0	0	4	23	103	180	17686

Neural Network

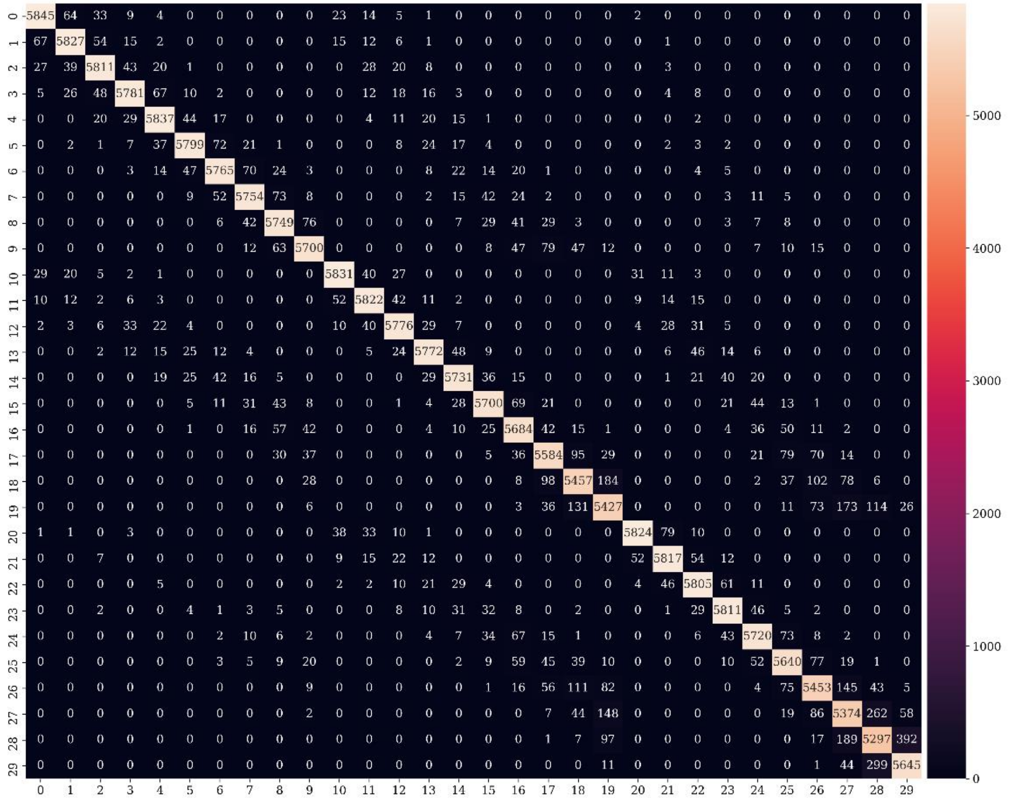


	0	1	2	3	4	5	6	7	8	9
0	17940	36	24	0	0	0	0	0	0	0
1	48	17912	32	8	0	0	0	0	0	0
2	8	44	17880	40	20	4	4	0	0	0
3	4	8	28	17912	28	16	4	0	0	0
4	4	4	12	36	17900	16	20	0	4	4
5	0	0	0	12	44	17904	32	8	0	0
6	0	0	0	4	20	40	17884	24	20	8
7	0	0	0	0	4	20	24	17888	40	24
8	0	0	0	0	4	8	12	56	17868	52
9	0	0	0	0	0	0	0	32	44	17924

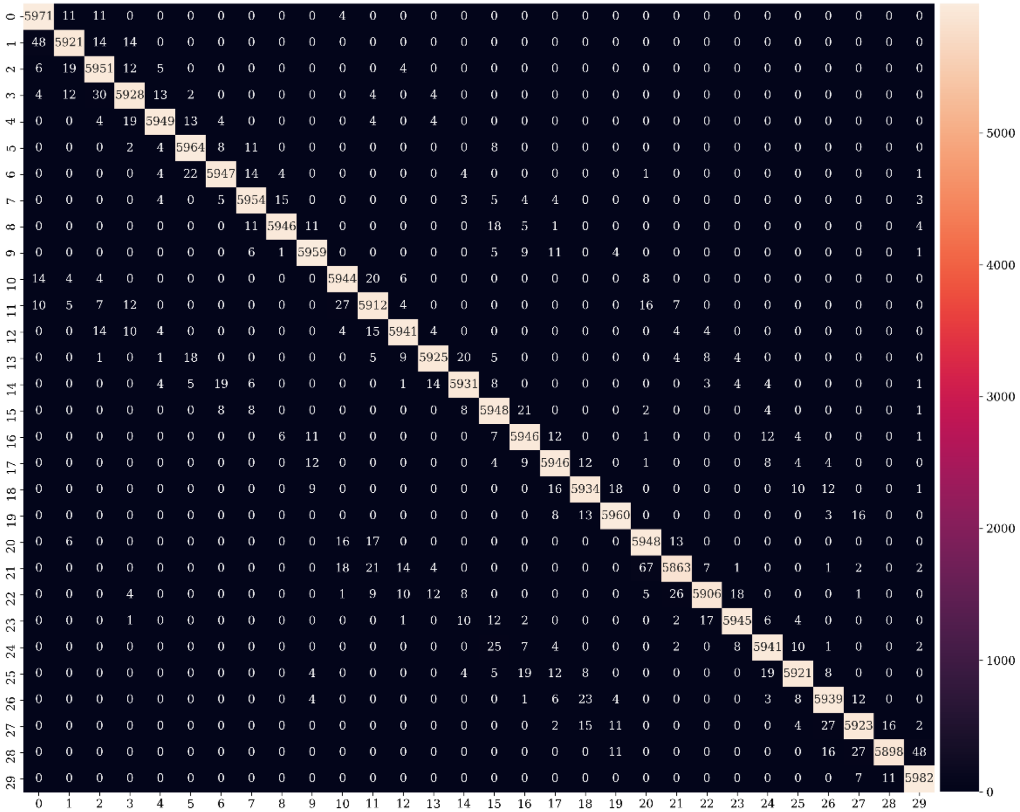
Decision Tree

Results cont.

- Multilabel neural network: 98.99%
- Multilabel decision tree: 95.02%



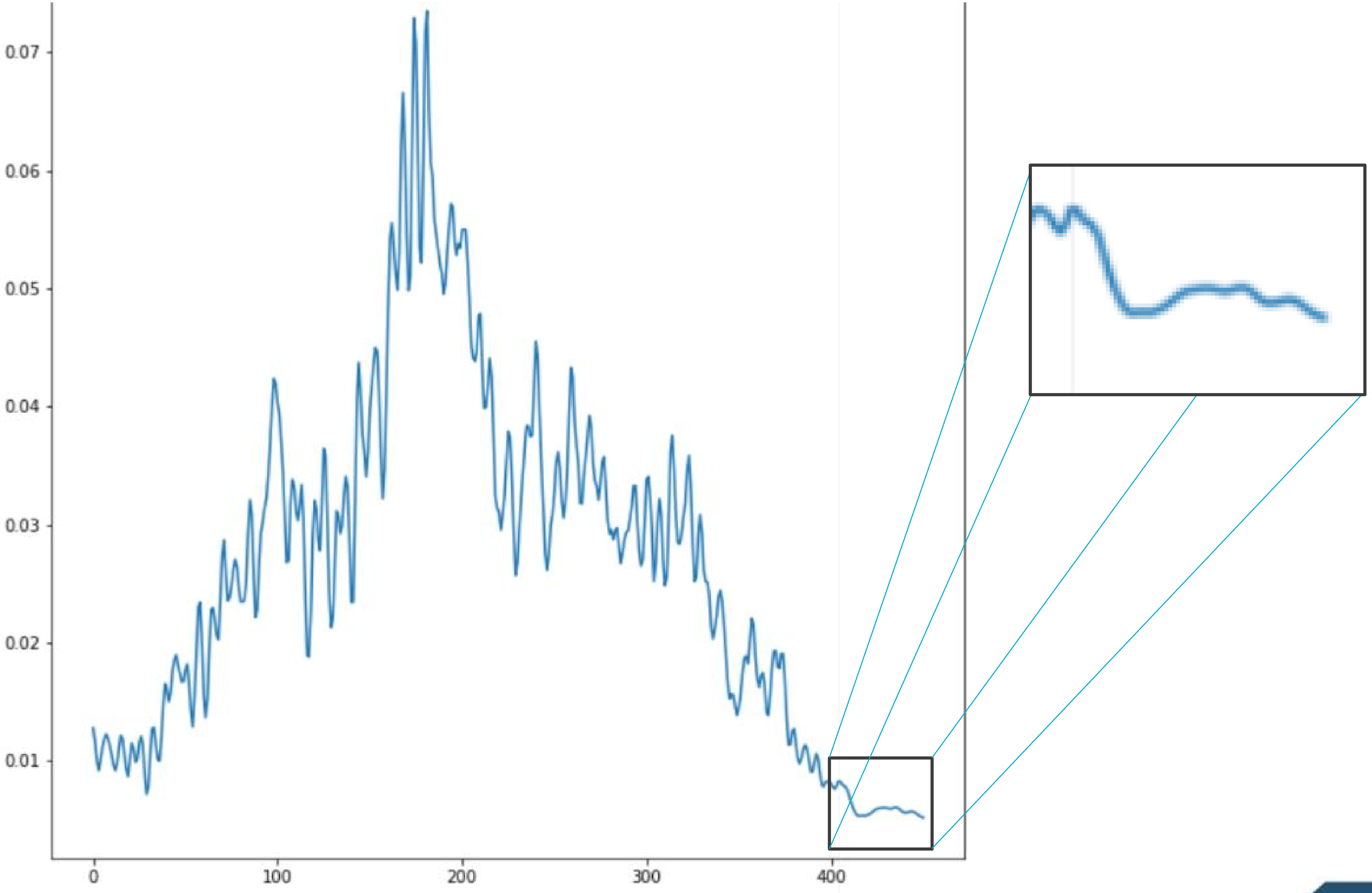
Neural Network



Decision Tree

Results cont.

- Feature importance



Results cont.

Percentage of data / number of data samples used for training and testing (% / samples)	Fully-connected neural network accuracy (%) ~ standard deviation (%) over 10 runs	Decision tree classifier accuracy (%) ~ standard deviation (%) over 10 runs
100 / 750,000	98.07 ~ 0.727	99.43 ~ 0.0112
90 / 675,000	96.81 ~ 1.054	98.88 ~ 0.0137
80 / 600,000	96.25 ~ 0.644	98.29 ~ 0.0157
70 / 525,000	93.36 ~ 1.512	96.54 ~ 0.0194
60 / 450,000	89.96 ~ 2.809	94.32 ~ 0.0395
50 / 375,000	82.60 ~ 2.077	86.98 ~ 0.0564
40 / 300,000	73.61 ~ 2.171	81.03 ~ 0.0659
30 / 225,000	61.73 ~ 0.508	62.56 ~ 0.0781
20 / 150,000	39.73 ~ 1.411	37.65 ~ 0.1391
10 / 75,000	35.82 ~ 6.332	44.04 ~ 0.1682

Conclusion

- Decision trees
  - Less hyperparamter tuning
  - Higher accuracy with less data
  - Better intuition for performance
  - More flexible with output vector format
- Neural networks
  - Less sensitive to length of output vector representation