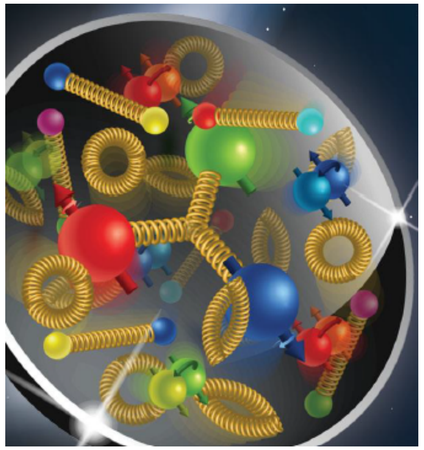
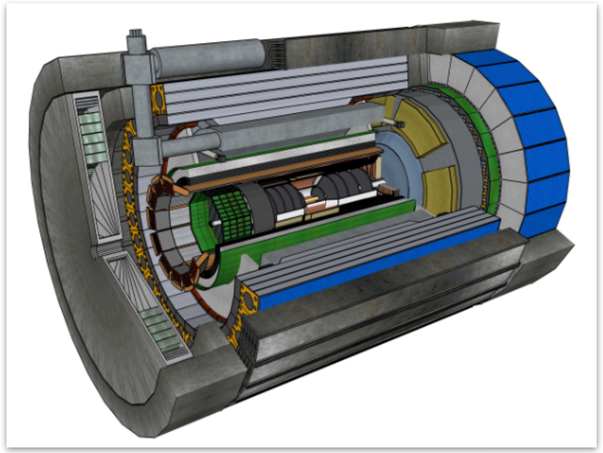
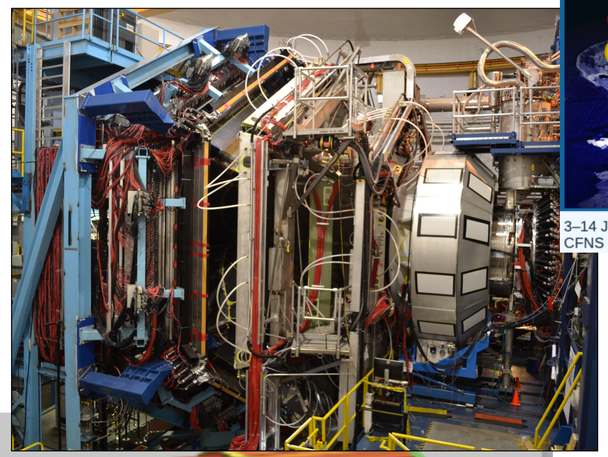
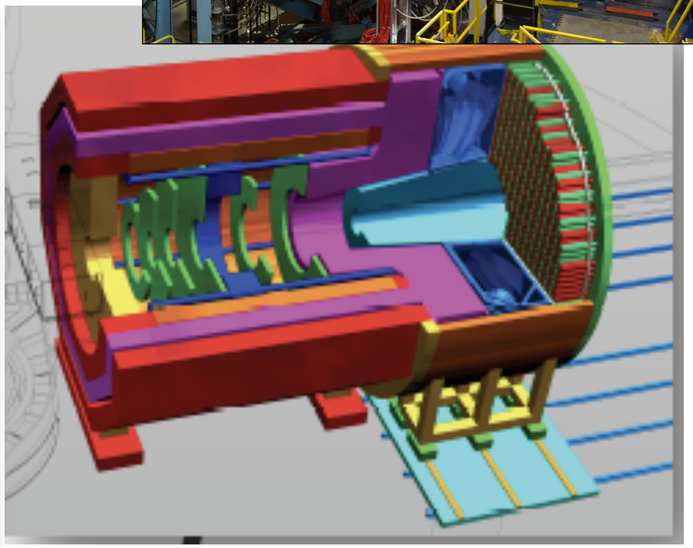








AI-supported methods for Real-time data analysis

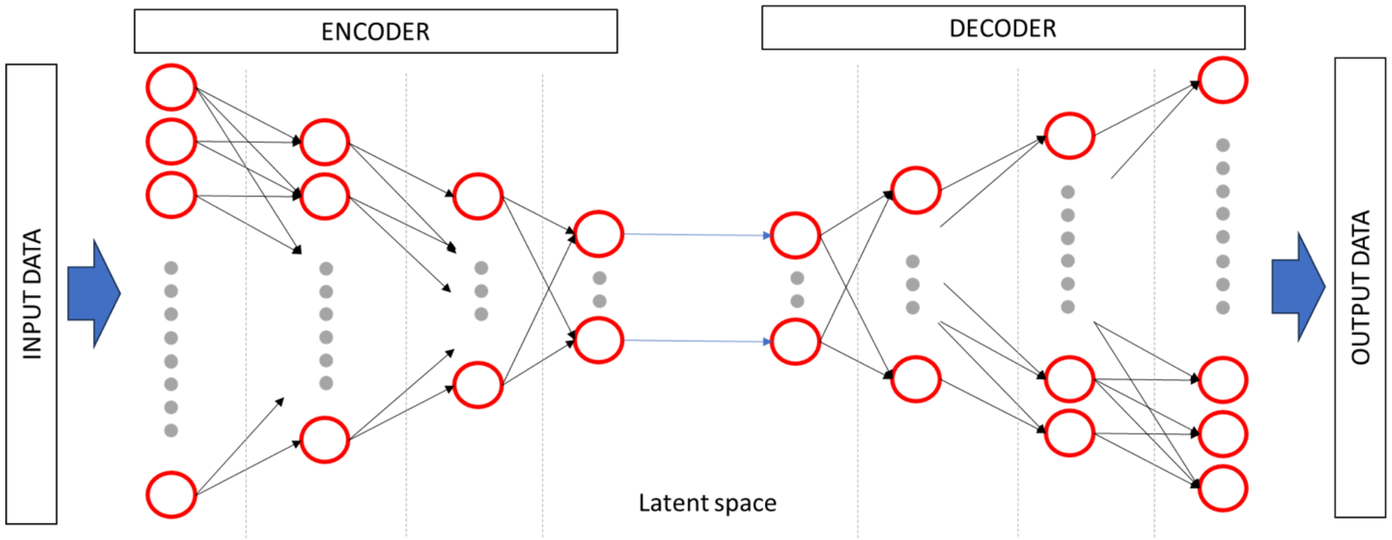
Part III - Autoencoder in action

M.Battaglieri (INFN)

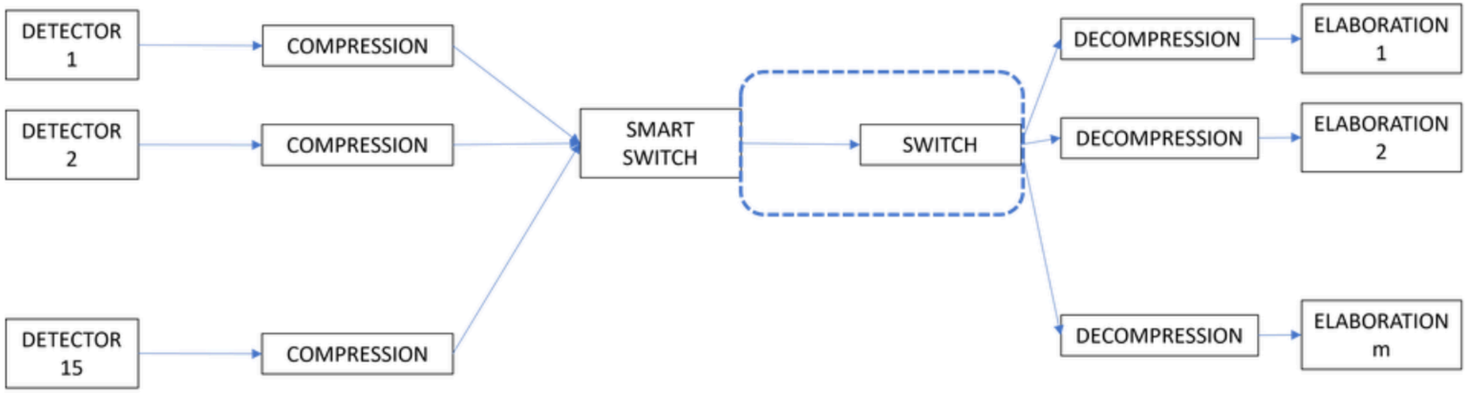
Autoencoder for data reduction

Credit to Fabi Rossi (INFN-GE)
https://colab.research.google.com/drive/1fSif01Wc6wXP_TyQdI3trReSf_4ltYph?usp=sharing#scrollTo=a3ls7urgXlA1

- Jupyter notebook
- COLAB (Google)
- Python
- Keras AI/ML libraries

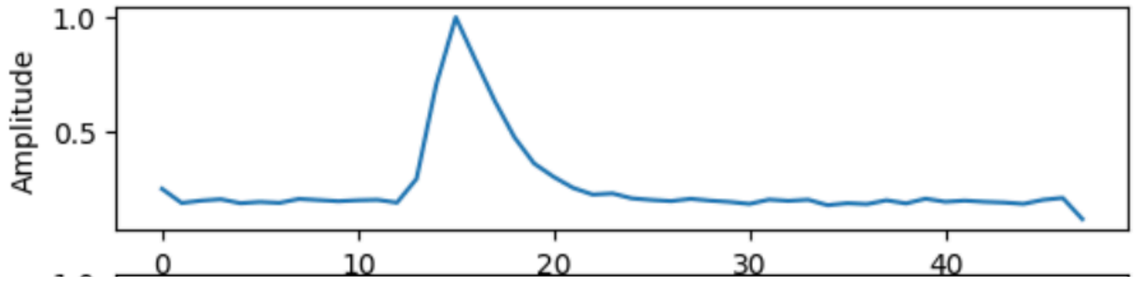


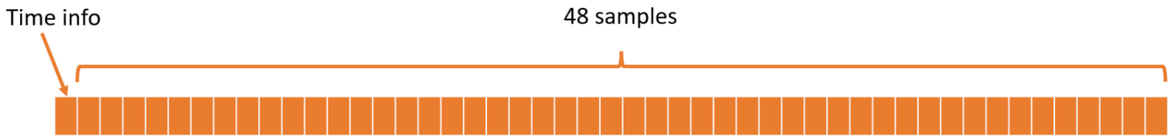
Autoencoder for data reduction

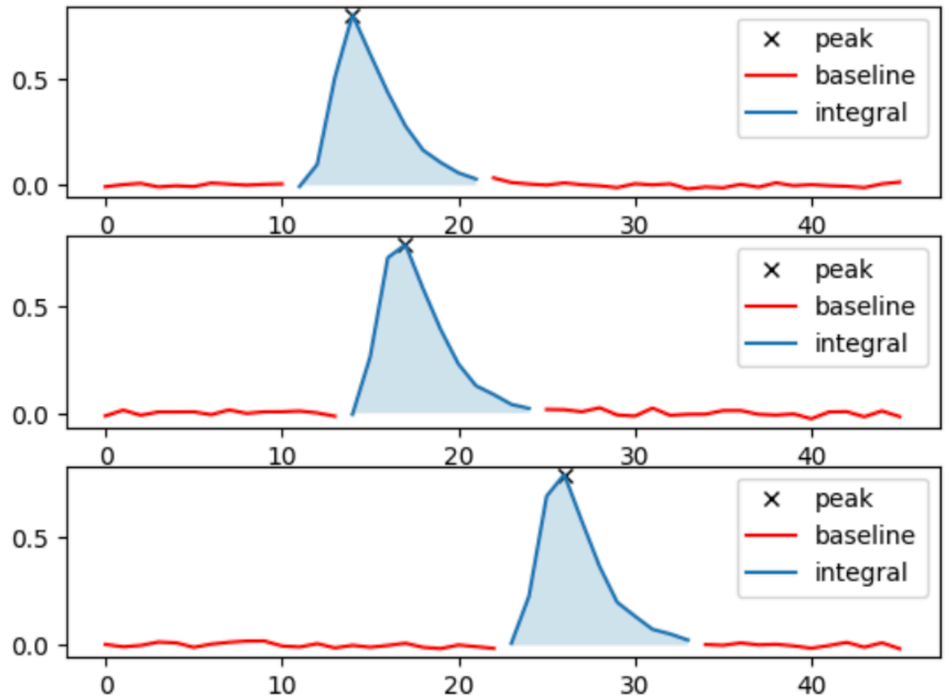


Goal: compress the waveform preserving the information (and eventually retrieve it)

Digitized waveform (48 samples)








```
Integral:[3.12319, 0] Baseline:0.19383583333333
Integral:[3.256660000000001, 0] Baseline:0.2101
Integral:[3.1619400000000004, 0] Baseline:0.200
array([[3.12319, 0.     ],
       [3.25666, 0.     ],
       [3.16194, 0.     ]])
```

Integral and peak position is a possible data reduction (lossy) algorithm

Autoencoder for data reduction

ML NN: FF Autoencoder with dim of latent feature space < dim input layer
(48,96,48,12 - 12, 48, 96, 48)
dim [48] → [12]

- REQUIREMENTS: the NN should be implemented on an FPGA (only integer numbers): as small as possible (regularisation) and weights need to be INT (quantized)
- DATA SET: 25k waveform digitized by a fADC250 at JLab
- Training/Validation/Validation/Test: 48%/15%/12%/25%
- LOSS function: MSE
- ADAM optimizer ($\eta = 10^{-3}$)
- EPOCHS: ~100
- ACTIVATION: ReLU
- WEIGHTS SETTING: training sample, test (during training on validation sample)
- PERFORMANCE: Quality based on comparison of WF integral TRUE/MODEL
- After training, the final weights are transferred to the FPGA for fast inference

```
Model: "Baseline_Model"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_4 (InputLayer)        [(None, 48)]              0

 dense_21 (Dense)            (None, 96)                4704

 dense_22 (Dense)            (None, 48)                4656

 dense_23 (Dense)            (None, 12)                588

 dense_24 (Dense)            (None, 12)                156

 dense_25 (Dense)            (None, 48)                624

 dense_26 (Dense)            (None, 96)                4704

 dense_27 (Dense)            (None, 48)                4656

=================================================================
Total params: 20088 (78.47 KB)
Trainable params: 20088 (78.47 KB)
Non-trainable params: 0 (0.00 Byte)
```

Baseline model
- best results
- Starting point for further optimization

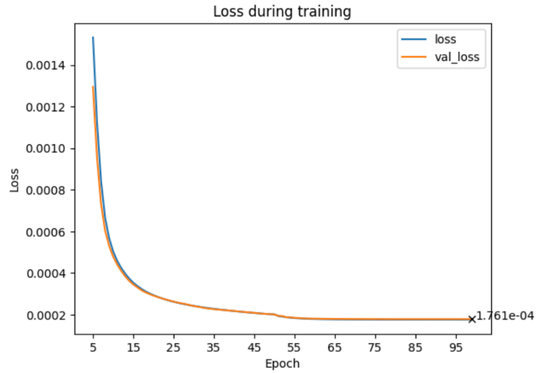
Pruned model
- zero (low values w_i) uppression
- re-trained starting from baseline
- sparsity evaluation

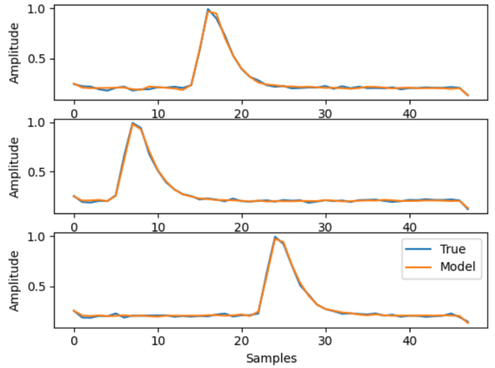
Quantized model
- re-trained starting from pruned (keeping pruned model)

Autoencoder for data reduction

Results

Baseline model

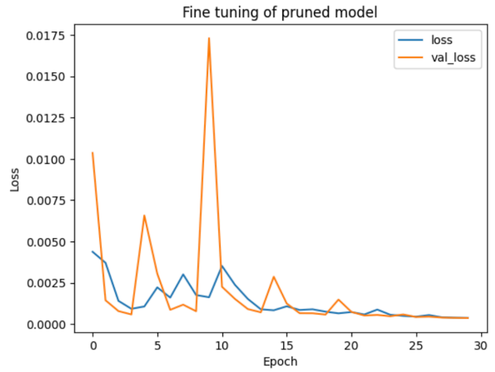


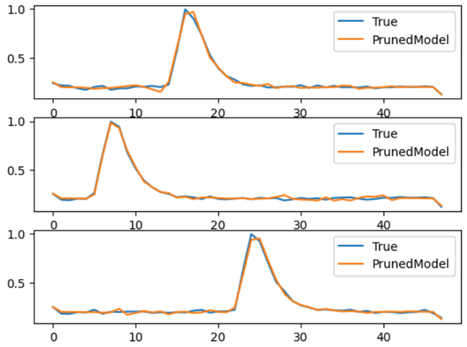


Original	Recon	Error%
3.154	3.151	0.093
3.245	3.224	0.656
3.220	3.225	-0.141

```
dense_21/kernel:0: 79.99% sparsity  (3686/4608)
dense_21/bias:0: 2.08% sparsity  (2/96)
dense_22/kernel:0: 79.99% sparsity  (3686/4608)
dense_22/bias:0: 0.00% sparsity  (0/48)
dense_23/kernel:0: 80.03% sparsity  (461/576)
dense_23/bias:0: 0.00% sparsity  (0/12)
dense_24/kernel:0: 79.86% sparsity  (115/144)
dense_24/bias:0: 8.33% sparsity  (1/12)
dense_25/kernel:0: 80.03% sparsity  (461/576)
dense_25/bias:0: 4.17% sparsity  (2/48)
dense_26/kernel:0: 79.99% sparsity  (3686/4608)
dense_26/bias:0: 0.00% sparsity  (0/96)
dense_27/kernel:0: 79.99% sparsity  (3686/4608)
dense_27/bias:0: 0.00% sparsity  (0/48)
```

Pruned model



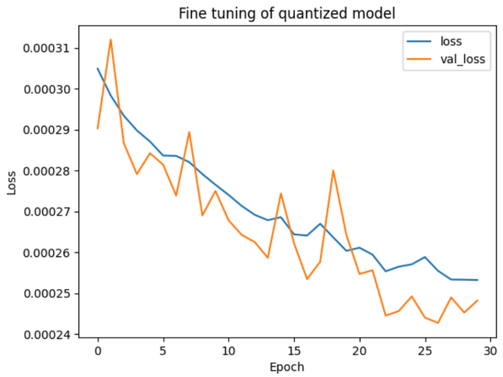


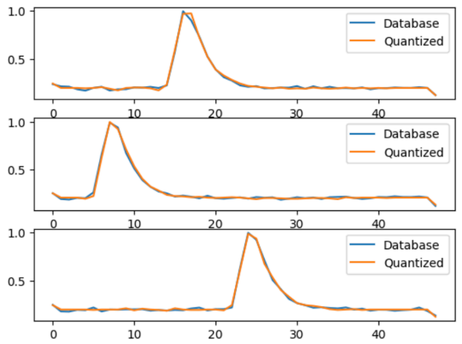
Original	p_Recon	Error%
3.154	3.247	-2.934
3.245	3.207	1.183
3.220	3.229	-0.278

```
Original model
[[ 2.56448984e-05  2.04816848e-01 -2.20985472e-01 ...  1.59392953e-01
   1.78266406e-01 -1.47754893e-01]
 [-1.90256640e-01  3.36886868e-02 -1.01705797e-01 ... -1.12358101e-01
  -2.99690175e-03 -1.91902950e-01]
 [-5.07703274e-02  7.48560280e-02  4.45862524e-02 ... -6.69857115e-02
   7.05242679e-02  7.37352520e-02]
 ...
 [ 1.14938974e-01 -1.74160168e-01 -1.65666074e-01 ... -1.08603626e-01
   1.09323412e-01  1.56354651e-01]
 [ 1.82403296e-01 -1.98742032e-01  1.27638802e-01 ... -9.61249769e-02
   1.73485637e-01  1.09817624e-01]
 [-1.05654188e-01 -1.86228636e-03 -1.85805395e-01 ...  1.30987376e-01
   1.46889716e-01  6.92100003e-02]]

Pruned model
[[-0.          0.22588937 -0.22141045 ...  0.          0.17150576
   0.        ]
 [-0.19826248  0.          0.         ...  0.          0.
  -0.1860043 ]
 [-0.          0.          0.         ...  0.          0.
   0.        ]
 ...
 [-0.          0.          0.         ...  0.          0.
   0.        ]
 [ 0.17366754 -0.1748717   0.         ...  0.          0.
   0.        ]
 [-0.          0.         -0.18634865 ...  0.          0.
   0.        ]]
```

Quantized model





Original	q_Recon	Error%
3.154	3.310	-4.936
3.245	3.248	-0.094
3.220	3.237	-0.538

Autoencoder for data reduction

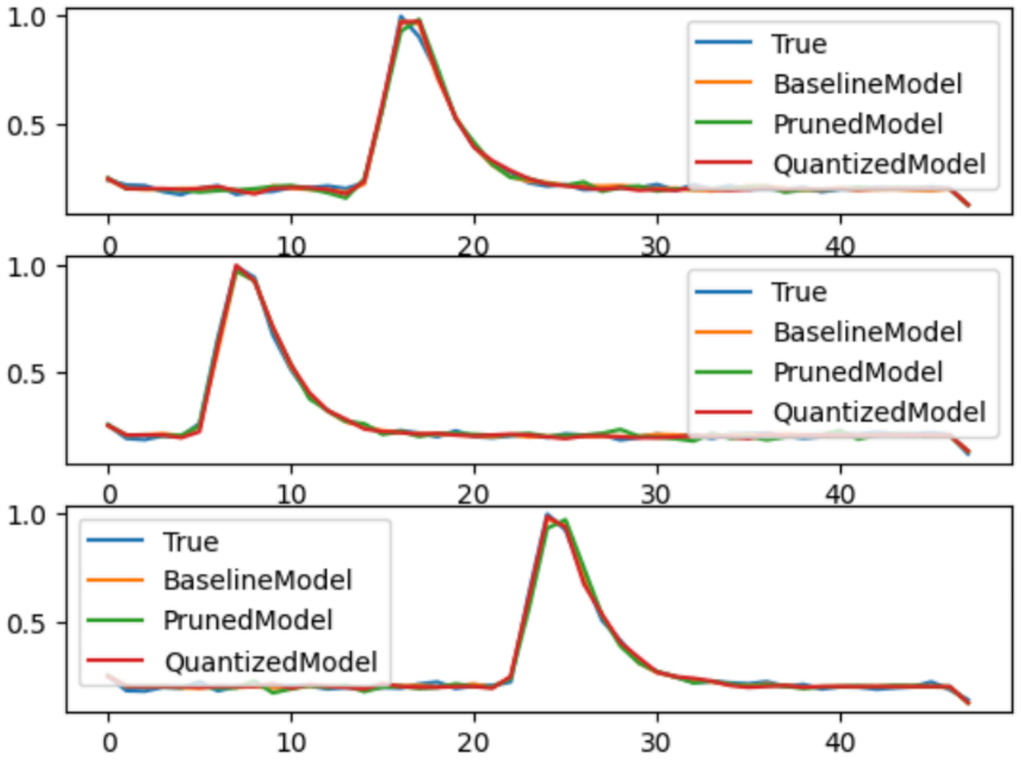
Data preparation

time info (to be removed)

48 samples for each waveform

```
0.32722, 0.24801, 0.18675, 0.19671, 0.20269, 0.18576, 0.19124, 0.18725, 0.20418, 0.19920, 0.19422, 0.19821, 0.20020,
0.18825, 0.29183, 0.70070, 0.99502, 0.80976, 0.62949, 0.47211, 0.35857, 0.30030, 0.25199, 0.22311, 0.22809, 0.20667,
0.19970, 0.19472, 0.20468, 0.19671, 0.19124, 0.18277, 0.20120, 0.19572, 0.20070, 0.17679, 0.18576, 0.18177, 0.19821,
0.18476, 0.20518, 0.19173, 0.19671, 0.19173, 0.18875, 0.18277, 0.20020, 0.20867, 0.11653,
0.38067, 0.25946, 0.19740, 0.22415, 0.20061, 0.21559, 0.21559, 0.21612, 0.20328, 0.22522, 0.20917, 0.21612, 0.21666,
0.21987, 0.21131, 0.19580, 0.20489, 0.47237, 0.93404, 0.99502, 0.78907, 0.59862, 0.43920, 0.33756, 0.29690, 0.25090,
0.23217, 0.22736, 0.22629, 0.21666, 0.23538, 0.20168, 0.19633, 0.23378, 0.20007, 0.20489, 0.20542, 0.22201, 0.22201,
0.20489, 0.20114, 0.20649, 0.18403, 0.21559, 0.21719, 0.19312, 0.22040, 0.19419, 0.13641,
0.56963, 0.23684, 0.20505, 0.19429, 0.20044, 0.21633, 0.21274, 0.19173, 0.20659, 0.21531, 0.22043, 0.22095, 0.19788,
0.19326, 0.20864, 0.18916, 0.20095, 0.19224, 0.20044, 0.21069, 0.19173, 0.18609, 0.20249, 0.19480, 0.18711, 0.21018,
0.42908, 0.89404, 0.99502, 0.77715, 0.56749, 0.40088, 0.33680, 0.27529, 0.25375, 0.22556, 0.20454, 0.20095, 0.21223,
0.20300, 0.20557, 0.19839, 0.18762, 0.19993, 0.21479, 0.19224, 0.21274, 0.18557, 0.13431,
0.66478, 0.26429, 0.20707, 0.22015, 0.19835, 0.22723, 0.21524, 0.23050, 0.22505, 0.23595, 0.19072, 0.21797, 0.20707,
0.21960, 0.22178, 0.21524, 0.19781, 0.21470, 0.22069, 0.20217, 0.20870, 0.19726, 0.22287, 0.21306, 0.20271, 0.21470,
0.21470, 0.19617, 0.21034, 0.21197, 0.29099, 0.67516, 0.99502, 0.96233, 0.75417, 0.51277, 0.41795, 0.32695, 0.28772,
0.25393, 0.22451, 0.24031, 0.21197, 0.22560, 0.22342, 0.21034, 0.21197, 0.21633, 0.13187,
0.44122, 0.24836, 0.20412, 0.20306, 0.21371, 0.21478, 0.19986, 0.22384, 0.21691, 0.18920, 0.20892, 0.21691, 0.20039,
0.23290, 0.20998, 0.22171, 0.23024, 0.20732, 0.20039, 0.21638, 0.51963, 0.99502, 0.93214, 0.79357, 0.54361, 0.44608,
0.35175, 0.27660, 0.24942, 0.22970, 0.22651, 0.20998, 0.21318, 0.20039, 0.19666, 0.23876, 0.20679, 0.21052, 0.20892,
0.22597, 0.21371, 0.20359, 0.21478, 0.20039, 0.20838, 0.18920, 0.20412, 0.20412, 0.12951,
0.64918, 0.27930, 0.20934, 0.20288, 0.21095, 0.21526, 0.20288, 0.20772, 0.20611, 0.20665, 0.21633, 0.22010, 0.20180,
0.20934, 0.20503, 0.20557, 0.21310, 0.21472, 0.22010, 0.22387, 0.21956, 0.22710, 0.19319, 0.22010, 0.20718, 0.21310,
0.20934, 0.20665, 0.21902, 0.24109, 0.53168, 0.95520, 0.99502, 0.75232, 0.57743, 0.40307, 0.34172, 0.28575, 0.25131,
0.23624, 0.22602, 0.23840, 0.20880, 0.21795, 0.20934, 0.21849, 0.19965, 0.21795, 0.12377,
...
```

Autoencoder for data reduction



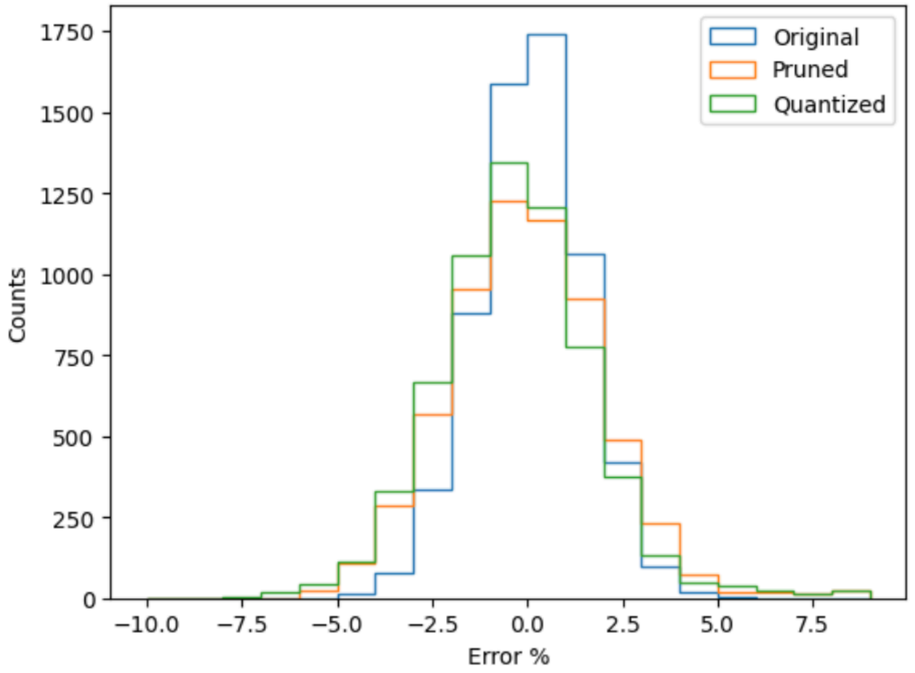

```
Original Recon Error%
3.154    3.151  0.093
3.245    3.224  0.656
3.220    3.225  -0.141
Original p_Recon Error%
3.154    3.247  -2.934
3.245    3.207  1.183
3.220    3.229  -0.278
Original q_Recon Error%
3.154    3.310  -4.936
3.245    3.248  -0.094
3.220    3.237  -0.538
```

Results

```
Comparison between models
##############################
Baseline model:
  Mean: 0.12 Std: 1.41
  Elapsed time: 284 us
  Float model in kb: 82.0

Pruned model:
  Mean: 0.23 Std: 2.77
  Elapsed time: 279 us
  Pruned model in kb: 82.0

Pruned+Quantized model:
  Mean: -0.24 Std: 2.24
  Elapsed time: 396 us
  Quantized model in kb: 26.2
```

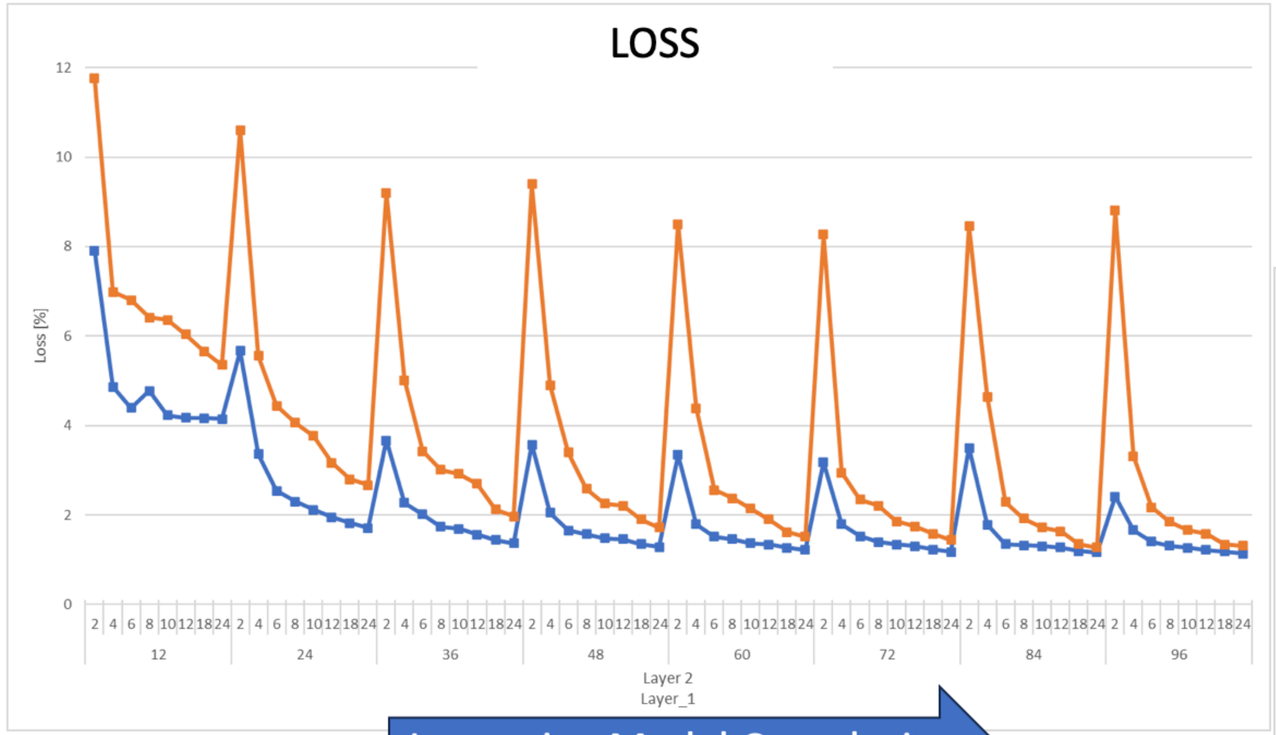


Compression 1:4

Autoencoder for data reduction

- Performance considerations

Signal Compression: Execution time vs Loss



Increasing Model Complexity

Model: RASPBERRY PI4
Quad core Cortex-A72 (ARM v8)
C code without any optimizations