Data Augmentation and Feature Engineering for Machine Learning in Neutron Activation Analysis

K. Brylew[1], T. Szczesniak[1], M. Grodzicka-Kobylka[1], L. Adamowski[1], L. Janiak[1], M. Neuer[2]

[1]National Centre for Nuclear Research, A. Soltana 7, 05-400 Otwock-Swierk, Poland

[2]innoRIID GmbH, Grevenbroich, Germany

NATIONAL CENTRE FOR NUCLEAR RESEARCH ŚWIERK

N-01-127

Abstract

Neutron activation analysis (NAA) is a widely used technique for detecting trace elements in various materials. In recent years, machine learning (ML) algorithms have shown great potential for improving the accuracy and efficiency of NAA. In this work, to achieve optimal results, data augmentation and feature engineering techniques are applied to NAA datasets to improve the quality and quantity of data available for training ML models. We will investigate the effectiveness of various data augmentation and feature engineering techniques in improving the performance of ML models for NAA.

We explore techniques such as feature selection and combination, temporal averaging, and evaluate their impact on the accuracy of NAA models. The results of this study will provide valuable insights into the optimal strategies for data augmentation and feature engineering in NAA, and could potentially lead to more accurate and efficient NAA systems in the future.

1. Sample information

55 samples were prepared made from finely ground metal chips. For each sample composition, spectra were measured for 60 seconds and repeated 60 times, resulting in a one-hour measurement during which the sample was constantly rotating with a period of ~5 s to improve the average uniformity of the material distribution inside the measurement chamber. The sample was irradiated with an intense PuBe neutron source emitting ~2x106 n/s, and data acquisition was performed using a CAEN DT5730 Digitizer (8 Channels, 14 bits, 500 MS/s).

Tab. 1. Mass ranges present in the preparation of calibration samples.

Element	Al	Cr	Cu	Fe	Mg	Mn
Minimum [g]	2765	0	2	3	1	0
Maximum [g]	10039	8	704	1065	68	31

Element	Ni	Pb	Si	Ti	Zn	Total
Minimum [g]	0	0	8	1	1	2800
Maximum [g]	45	25	710	6	46	10716

Fig. 1. Typical quality of the investigated samples

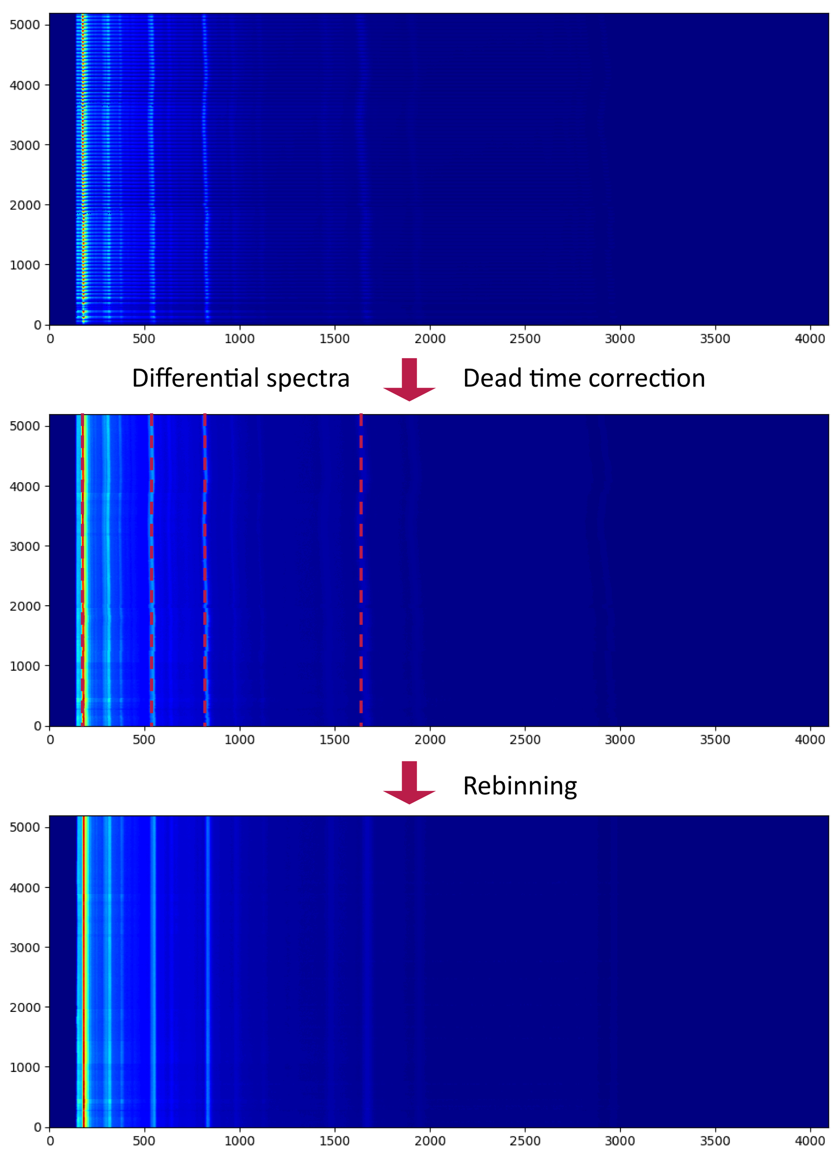
2. Experimental setup — see N-11-152!

The current setup is shown in Fig. 2, where the Large Sample Sensor (LSS) is configured to emulate industrial conditions at aluminum refinery. This configuration can detect gamma rays induced by neutrons in material constantly moving in a vertical pipe. Currently, five different detectors (3x LaBr:Ce, 1x LaBr:Sr, BGO) are installed in the setup (Fig. 3a, 3b).

Fig. 2. Photo of the LSS demonstrator placed in the experimental setup.

Fig. 3. Photo of one of the LaBr3 detectors (a) and BGO (b) used in the experiments.

3. Data preprocessing



The raw data is a series of accumulating histograms of counts in each channel collected every 60s during an hour long measurement. The spectra measured in succession are subtracted from each other. After each minute the information about the dead time of the digitizer to correct for the expected amplitude that was lost during dead time and because of shorter or longer measurement time.

The spectra have to be rebinned to account for variability of electrical gain and other environmental effects. The variability of the sample mass should result in the change of peak amplitude and it cannot be disturbed by effect of shifting peak positions.

For training set and test set common calibration lines are selected.

4. Model training

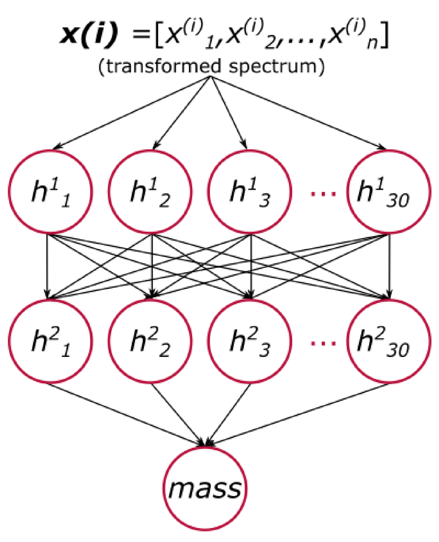
> The dataset is split into test set (15 %) and training set (85 %).

> Input layer depends on the number of channels that varies for each element.

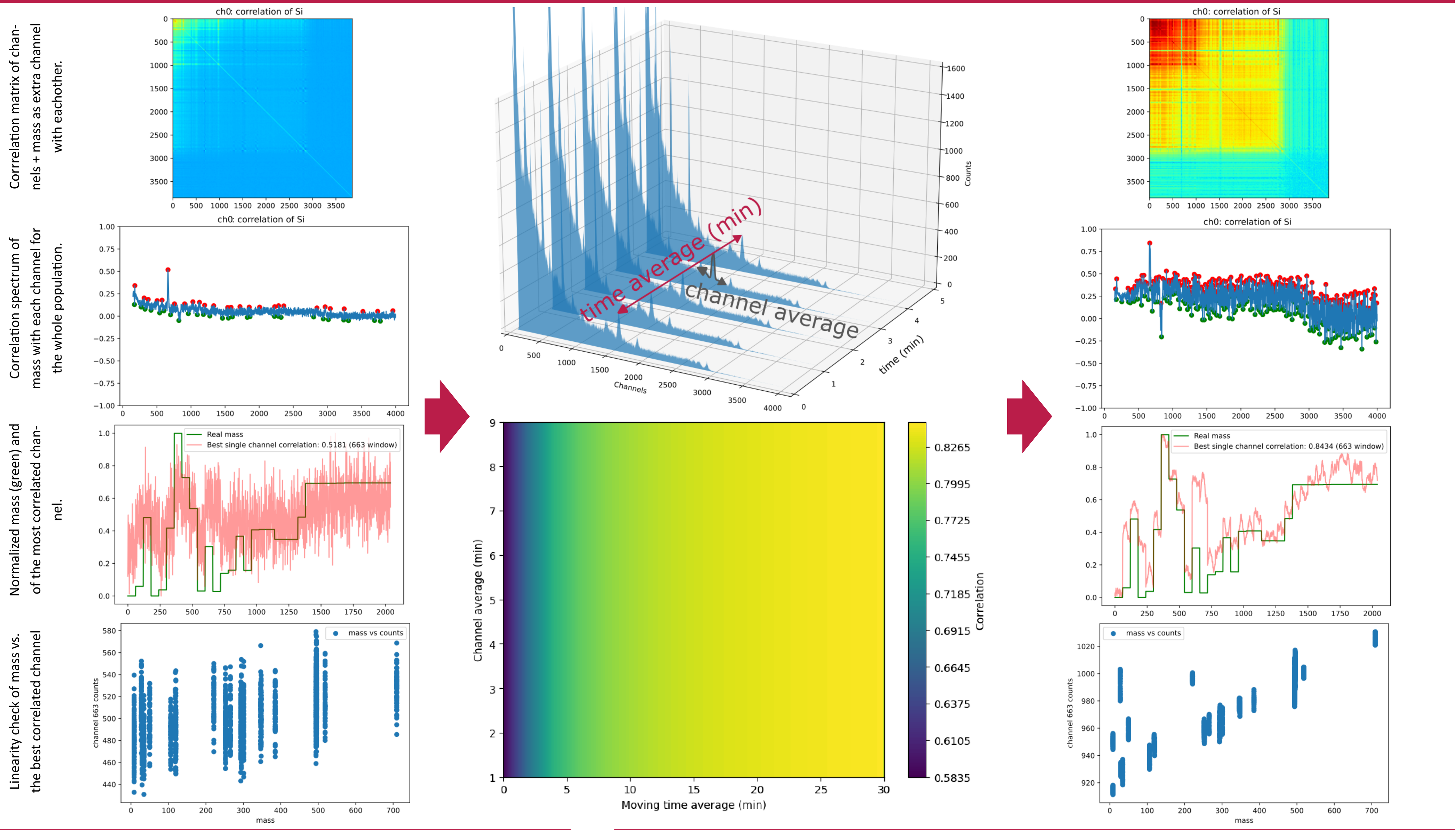
> Two hidden layers with 30 neurons each.

> The deep layers are activated using the hyperbolic tangent (tanh) function and each layer incorporates L1L2 regularization to mitigate overfitting concerns.

> The output layer utilizes a sigmoid activation function to produce a continuous numerical output.



5. Feature extraction



6. Feature joining

The counts from all of the detectors are added to improve the signal to noise ratio. This can be applied in two ways. If we assume that features from detector A and B can be written as a matrices:

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{bmatrix}$$

$$B = \begin{bmatrix} b_{1,1} & b_{1,2} & \cdots & b_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m,1} & b_{m,2} & \cdots & b_{m,n} \end{bmatrix}$$

Where m is the number of instances (spectra), and n is the number of channels (or more generally features). We can join the data from these detectors in two manners, i.e. by adding corresponding elements:

$$A + B = \begin{bmatrix} a_{1,1}+b_{1,1} & a_{1,2}+b_{1,2} & \cdots & a_{1,n}+b_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1}+b_{m,1} & a_{m,2}+b_{m,2} & \cdots & a_{m,n}+b_{m,n} \end{bmatrix},$$

or concatenating the vectors horizontally:

$$(A,B) = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} & b_{1,1} & b_{1,2} & \cdots & b_{1,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} & b_{m,1} & b_{m,2} & \cdots & b_{m,n} \end{bmatrix}$$

In this work we are using five detectors, described as D_0, D_1, D_2, D_3, D_4 and five configurations have been tested:

> D_0

> D_3

> $D_0 + D_1 + D_2 + D_4$

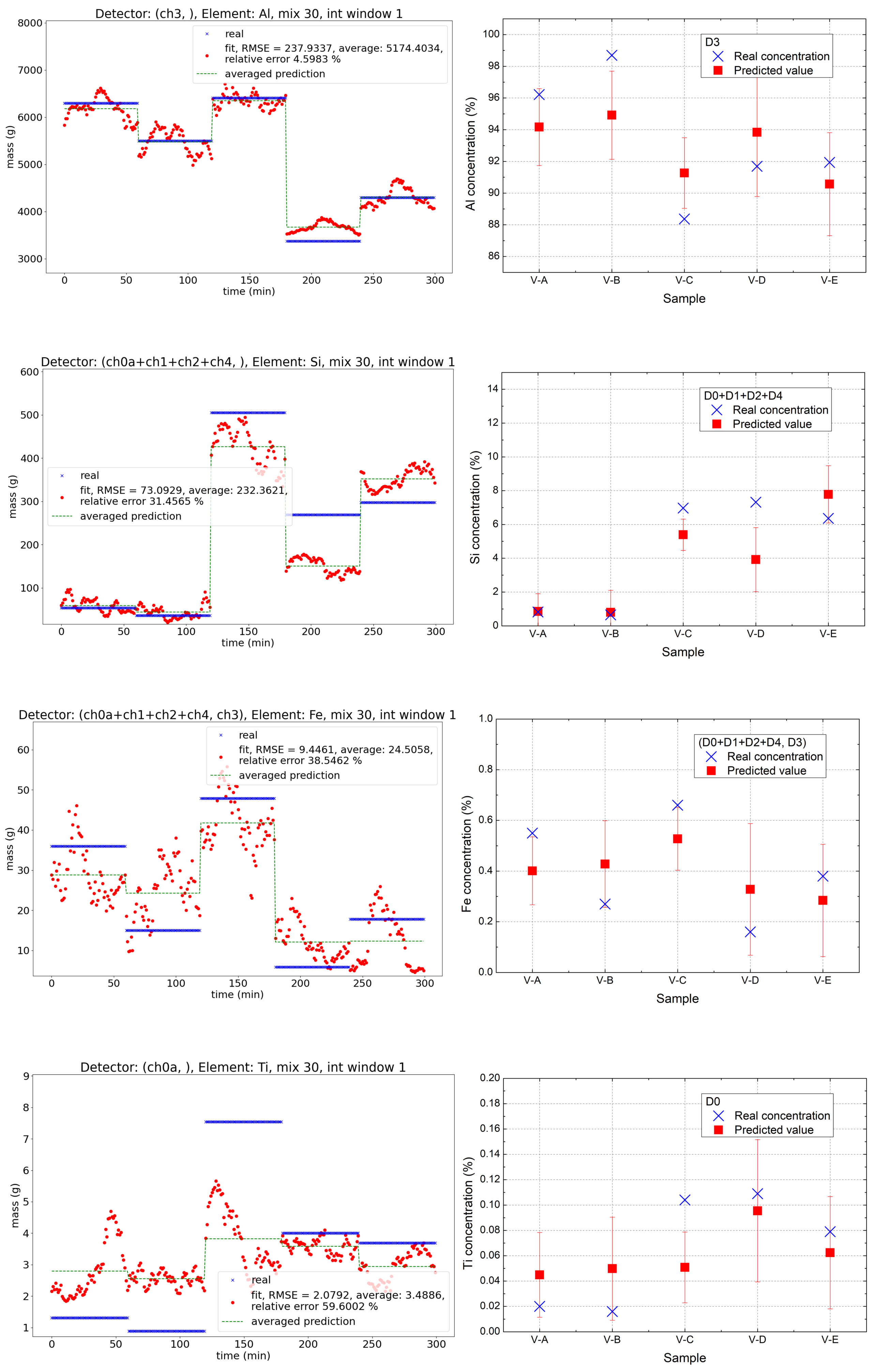
> $D_0 + D_1 + D_2 + D_3 + D_4$

> $(D_0+D_1+D_2+D_4, D_3)$

7. Results

Error estimation based on RMSE (g)

	D0	D3	D0+D1+D2+D4	D0+D1+D2+D3 +D4	(D0+D1+D2+D4, D3)
Al	347.48	237.93	466.54	497.89	508.14
Cr	3.72	3.26	3.47	3.31	2.69
Cu	35.39	83.15	45.61	66.72	42.05
Fe	10.99	11.13	14.36	12.93	9.44
Mg	15.79	18.22	13.55	14.97	18.20
Mn	5.77	6.50	5.99	6.98	4.86
Ni	6.74	11.80	1.18	1.37	1.08
Pb	5.39	6.53	7.26	7.12	4.29
Si	97.50	143.42	73.09	89.74	78.01
Ti	2.07	2.42	2.23	2.13	2.32
Zn	5.02	8.35	7.02	16.20	15.14
RMSE sum	535.86	532.71	640.30	719.36	686.22



8. Conclusions

> Real samples with known compositions are better training material for neural networks. The uniformity of finely chipped metals does not introduce problems with dependance of metal components in sample

> Careful preparation of training data is important since the performed averaging experiments suggest that counts from single channels can be taken as features for input of the neurlal networks

> Positioning of the detectors makes them sensitive for different components, unshielded detectors behind the sample are better stuited for total mass prediction

> Combination of data from multiple detectors makes the setup more sensitive for trace amounts of elements.

> Our results suggest that the limit of detectability in our setup is below 20 g of uniformly distributed mass.

Acknowledgements

This work was supported in part by the European Union's Horizon 2020 research and innovation programme under grant agreement No 869882.

Contact information

> Web: https://www.ncbj.gov.pl/en

> Email: kamil.brylew@ncbj.gov.pl

> Phone: +48 661 420 142

REVaMP

Funded by

Horizon 2020 European Union Funding for Research & Innovation

SPIRE