PROCEEDINGS OF THE YEREVAN STATE UNIVERSITY

Physical and Mathematical Sciences 2026, **60**(2), p. 102–111

P h y s i c s

A NEURAL-NETWORK METHOD FOR COUNTING AND LOCATING OVERLAPPING SCINTILLATION PULSES IN DIGITIZED WAVEFORMS

A. A. HOGHMRTSYAN *

A. Alikhanian National Science Laboratory (AANL),
Yerevan Physics Institute, Armenia

At high counting rates, scintillation pulses overlap in the digitizer window, and when the arrival times are not fixed by an accelerator clock, the classical reconstruction methods begin to lose them. We present a neural-network method that counts the overlapping pulses and measures the arrival time and amplitude of each. The method is studied on simulated 1 *GHz* waveforms containing zero to six pulses. Two small convolutional networks recovered 99.2% of the individual pulses, with a mean timing error of 4.6 *ns* and an amplitude error of 1.1 ADC counts (9% of a typical amplitude); the pulse count was determined correctly in 96.5% of the windows. Under identical conditions and scoring, a matched filter recovered only 46% of the pulses with a 7.5 *ns* timing error, and an iterative CLEAN-type template subtraction 57% with a 5.2 *ns* error. The false-positive rate was 0.4% for the network, 2.6% for CLEAN, and 12% for the matched filter.

https://doi.org/10.46991/PYSUA.2026.60.2.102

MSC2020: 004.85:621.391.

Keywords: pulse pile-up, scintillation detectors, neural networks, waveform reconstruction, convolutional networks, template subtraction.

Introduction. Scintillation detectors are among the most widely used instruments in nuclear and particle physics. A scintillating crystal or organic material converts the energy deposited by an incident particle into visible light. The light is collected by a photomultiplier tube (PMT) or a silicon photomultiplier (SiPM) and converted into an electrical pulse [1]. The temporal profile of the pulse is governed by the scintillation mechanism: the rise time reflects the population of the luminescent states, and the decay time their de-excitation. Decay times span a wide range [2]. The fast component of BaF_2 lasts less than one nanosecond, plastic scintillators a few nanoseconds, and lead tungstate ($PbWO_4$), used, for example, in the CMS electromagnetic calorimeter, about ten nanoseconds. LYSO is characterized by decay times of tens of nanoseconds, and BGO by hundreds.

* E-mail: artur.hoghmrtsyan@cern.ch

Modern readout systems digitize the photodetector signal directly with fast analog-to-digital converters (ADCs) at sampling rates of up to 1 GHz and with 12–14-bit resolution [3]. The complete pulse shape is thus recorded within a readout window of fixed length. For every particle, the waveform contains two physically important quantities. The arrival time of the pulse is needed for time-of-flight and coincidence measurements. The amplitude, or integrated charge, is proportional to the deposited energy.

At high counting rates, this picture is complicated by pulse pile-up. When the mean interval between events becomes comparable to the pulse duration, several pulses fall into the same readout window and add up. For a source of rate R and a window of length T, the probability of at least one additional event is $1 - e^{-RT}$. At $R = 10^5\,s^{-1}$ and $T = 120\,ns$, it amounts to about 1% per window, but it grows rapidly with increasing rate. In high-luminosity collider experiments and in measurements of high-activity sources, pile-up becomes the dominant systematic effect [4]. The summed pulse height no longer corresponds to a single event, so the energy spectrum is distorted. Merged peaks degrade the timing accuracy, and unresolved coincidences are silently lost from the event count.

Many methods have been developed to mitigate pile-up. Analog shaping filters shorten the effective pulse duration at the cost of timing and energy resolution [2]. Digital peak finding in the filtered waveform resolves pulses down to a separation of roughly one rise time. At smaller separations, the composite waveform has fewer maxima than true pulses. In high-energy physics, the standard tool is waveform fitting: the recorded samples are described by one or more copies of a known pulse template. One variant uses optimal-filtering weights, introduced for liquid-argon calorimetry [5]. Another is an explicit multi-template χ^2 fit. An example is the "multifit" reconstruction of the CMS $PbWO_4$ calorimeter, which represents the waveform as a sum of up to ten template amplitudes in order to suppress pile-up from neighbouring collisions [6].

Such fits work particularly well at a collider, because the possible arrival times are known in advance: signals occur only at the accelerator bunch crossings. The amplitudes then become linear parameters, and the fit is fast and stable. If the arrival times are unknown, as for a radioactive source, cosmic rays, or any asynchronous signal, they become nonlinear parameters. The well-known difficulties of multi-pulse fitting then appear. The number of pulses has to be specified before the minimization and revised, if the hypothesis is wrong; the χ^2 surface contains local minima. For two pulses separated by an interval of the order of the rise time, the problem becomes degenerate, and the time and amplitude estimates are strongly correlated. The fit quality also depends directly on the accuracy of the template shape, whereas the actual pulse shape changes with temperature, counting rate, and detector ageing.

A practical alternative, widely used for recovering piled-up pulses, is iterative template subtraction. It is often called a CLEAN-type method, by analogy with Högbom's deconvolution algorithm for radio interferometry [7]. Pulses are found and subtracted from the waveform one after another, starting with the strongest, so the overlap is handled step by step rather than jointly [8]. The algorithm immediately

commits to the strongest candidate. Any error in this estimate, whether caused by noise or by the tail of a neighbouring pulse biasing the amplitude, propagates into all subsequent residuals. Weak pulses on the tails of strong ones are often not detected at all. Such losses are easy to underestimate: the pulses that "survive" are precisely the well-separated ones, which look perfectly reconstructed.

The present work investigates whether a neural network trained on simulated waveforms of exactly known content can better solve the problem with asynchronous arrival times. The tasks are (a) to determine the number of pulses in the window and (b) to estimate the time and amplitude of each of them. The network receives no prior hypothesis about the number of pulses and performs no separate fit of each event. The comparison is made with a matched filter and with a CLEAN-type iterative subtraction on the same data and with identical scoring. For the time and amplitude of well-separated pulses, the network is moderately more accurate. For the detection of overlapping pulses, the difference is qualitative: the network misses 0.8% of the pulses, whereas the classical methods miss 43–54%.

Simulated Waveforms. Training requires a large number of waveforms of exactly known content, so this study is based on simulation. The pulse shape is given by the standard two-exponential function

$$p(t) = A\left(1 - e^{-(t-t_0)/\tau_r}\right) e^{-(t-t_0)/\tau_f}, \tag{1}$$

where t_0 is the onset time and A is the amplitude parameter. The rise time is $\tau_r = 2$ *ns* and the decay time is $\tau_f = 10$ *ns*; such values are typical of a fast scintillator read out by a fast photodetector. The pulse reaches its maximum $\tau_r \ln(1 + \tau_f/\tau_r) \approx 3.6$ *ns* after the onset, at about 58% of the value of A. The full width at half maximum is approximately 12 *ns*. For the amplitudes used, the visible peak height ranges from about 3 to 12 ADC counts.

Each simulated window is 120 *ns* long and is sampled at 1 GHz, i.e. it contains 120 samples with an exact step of 1 *ns*. The pulse multiplicity is drawn uniformly from 0 to 6, so each class is represented by approximately $7,000$ waveforms and the networks receive balanced statistics. Six pulses in a 120 *ns* window correspond to a severe pile-up regime, in which the tails overlap almost everywhere. The onset times are distributed uniformly in $t_0 \in [5, 110]$ *ns*; the margins at both edges exclude truncated pulses. The times are continuous and not tied to the sampling grid. A minimum interval of 2 *ns*, equal to the rise time constant, is imposed between onsets, since at a smaller distance two pulses are, in principle, indistinguishable from a single stronger one. The amplitudes are distributed uniformly in $A \in [5, 20]$ ADC counts. Together with the specified noise level, this corresponds to a peak signal-to-noise ratio from 6 to 58, i.e. from barely distinguishable to clearly visible pulses.

The simulation includes the imperfections of real digitizing hardware. White Gaussian noise is added with a standard deviation $\sigma \in \{0.2, 0.3, 0.4, 0.5\}$ counts, chosen independently for each waveform. The networks therefore cannot adapt to a single noise level and must generalize to the variations caused by temperature, grounding, and detector ageing. A pedestal of 200 counts is added, with a random

variation of ± 5 counts between waveforms, as well as a slow linear drift of up to ± 2 counts across the window. Such values are typical of DC-coupled ADC front-ends, where the operating point is set above zero to accommodate negative noise fluctuations. For each waveform, the exact onset time and amplitude of every pulse are stored; they serve as the training labels and as the true values for all comparisons.

In total, $50,000$ waveforms were generated: $40,000$ for training and $10,000$ for independent validation. All numbers given below refer to the validation sample only. Before reconstruction, the pedestal is removed with a rolling-quantile filter: the 10th percentile in a 31-sample window. The filter follows the noise floor rather than the pulses, so it removes the offset and the drift with practically no distortion of isolated signals.

Methods. The pedestal-subtracted waveform is processed by two convolutional networks. Before being fed to a network, each of the 120 samples is standardized to zero mean and unit variance; the means and variances are determined separately for each sample position over the training set.

Both networks have the same layer sequence. The input of dimension 120×1 passes through two one-dimensional convolutional layers with 64 and 128 filters, a kernel width of 5 samples, stride 1, "same" padding, and ReLU activation. Each convolution is followed by batch normalization. The output is flattened and fed to a fully connected layer of 256 neurons with ReLU activation, batch normalization, and dropout of 0.3, and then to a second fully connected layer of 128 neurons with ReLU activation. Only the output layer differs. Each network contains about $4.0 \cdot 10^6$ trainable parameters, most of which belong to the first fully connected layer. Convolutional filters are natural for this task: they slide along the time axis and respond to the pulse shape regardless of its position, much like a bank of matched filters whose templates are learned from the data rather than specified in advance.

The first network is a classifier. Its output consists of seven softmax neurons that give the probabilities of the multiplicities from 0 to 6. It is trained with the cross-entropy loss and with class weights that balance the multiplicities. The second network is a regressor. It has 14 linear outputs, interpreted as seven (t_0, A) pairs, one slot for each possible pulse; unused slots are trained to return zero. The regression targets are scaled by fixed physical constants: the times by the window length of $120\,ns$ and the amplitudes by the maximum of 20 counts, so that both quantities lie in $[0, 1]$. Fixed scales keep the zero padding of unused slots exactly equal to zero. A statistical normalization of the targets, by contrast, shifts the padding value and, in our tests, degraded the timing accuracy by a factor of four.

The regression loss is the Huber loss with a threshold of 1 in scaled units, applied to all 14 outputs. The timing components carry twice the weight of the amplitude components, since time is the more difficult and more valuable quantity. Active slots enter with full weight, and the unused zero slots with a small weight of 0.05. If a window contains, for example, three pulses, there is no natural rule for assigning them to particular output slots. Therefore, during training, the network outputs are matched to the true pulses with the Hungarian assignment algorithm [9].

The network is not penalized for finding the correct pulses in a different order, an approach borrowed from multi-speaker speech separation [10]. In additional tests, the simple convention of sorting the true pulses by arrival time gave a comparable result, so the choice between these options is not critical.

What is critical is that the unused slots remain weakly pulled towards zero in the loss function. If they are excluded from the loss entirely, they drift to arbitrary pulse-like values. It is then impossible to determine at the analysis stage, which outputs correspond to real signals. This is precisely why the small padding weight of 0.05 is used.

Both networks are trained with the Adam optimizer, with an initial learning rate of 10^{-3} and a batch size of 128, on the $40,000$ training waveforms. The Hungarian matching of the regressor is recomputed for every training batch in the scaled target space. The learning rate is halved whenever the validation loss does not improve for three epochs, down to a minimum of 10^{-6}; training stops after six epochs without improvement. The classifier is trained for up to 30 epochs and the regressor for up to 40. On a 16-core desktop CPU, this takes approximately 16 and 54 *min*, respectively; no GPU is required.

At the analysis stage, the two networks are combined as follows: the classifier determines the multiplicity $\hat{k}$, after which the $\hat{k}$ output slots of the regressor with the largest amplitudes are taken as the reconstructed pulses.

For comparison, two classical methods were implemented on the same waveforms. The matched filter correlates the waveform with the known pulse template; all sufficiently high and sufficiently separated maxima of the correlation are accepted as pulses. The iterative subtraction is the CLEAN-type algorithm of Refs. [7, 8], run for at most six iterations with the same detection threshold as the matched filter. It is the standard classical technique for recovering piled-up pulses with unknown arrival times and the closest counterpart of the task performed by the network. Both methods use the exact pulse template from the simulation, i.e. they operate without any pulse-shape modelling error; this condition, if anything, favours the classical methods.

All three methods are evaluated identically. The reconstructed pulses of each window are matched one-to-one to the true pulses by the Hungarian algorithm. Matched pairs yield the time and amplitude errors. Unmatched true pulses are counted as missed, and unmatched reconstructed pulses as fakes. This set-based protocol is independent of the output ordering and puts the learned and classical methods on an equal footing.

Results on the validation sample. "Count" is the fraction of windows with correctly determined multiplicity; the time and amplitude errors are mean absolute errors over matched pulses

Method	Count	Time, *ns*	Ampl., Counts	Missed
Matched filter	23.9%	7.5	3.0	54%
Iterative subtraction	32.3%	5.2	3.3	43%
Neural network	96.5%	4.6	1.1	0.8%

Results. Table summarizes the comparison on the 10,000 validation waveforms, which contain approximately 29,700 true pulses.

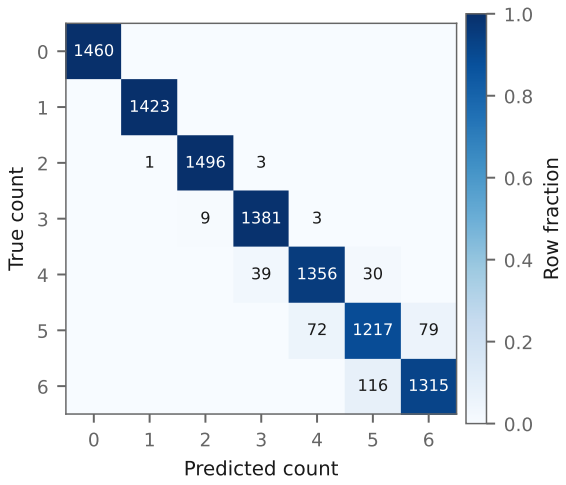


Fig. 1. Multiplicity determination on the validation sample. The rare errors are almost exclusively ± 1 and occur at five or six pulses per window, where two pulses can be nearly simultaneous.

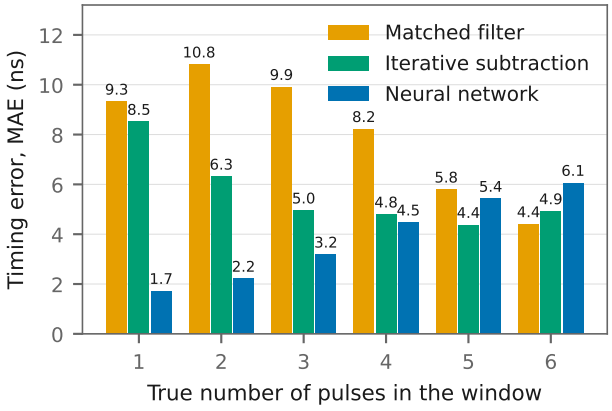


Fig. 2. Timing error of matched pulses versus the true number of pulses in the window. The apparent improvement of the classical methods at high multiplicity is due to their discarding the difficult pulses (43–54% missed overall); the network retains about 99% of the pulses.

Consider first the number of pulses. The network determines the multiplicity correctly in 96.5% of the windows. Its rare errors are almost always off by one and occur at multiplicities of five and six (Fig. 1), where the arrival times of two pulses can genuinely be too close. Within the $2\,ns$ rise time, no method can separate them, even in principle. The classical methods return the correct number in only about a quarter to a third of the windows: when pulses overlap substantially, they merge them.

The second, and decisive, figure of merit is the miss rate. The iterative subtraction loses 43% of the true pulses, the matched filter 54%, and the network only 0.8%. The corresponding fake rates are 2.6%, 12%, and 0.4%. Fig. 2 shows why the errors of the classical methods on the pulses they find look deceptively good at high multiplicity. The difficult pulses are simply absent from their output, so the statistics are computed over an increasingly favourable subset. Any measurement that depends on efficiency, including absolute rates, dead-time corrections, and coincidence counting, inherits these losses directly.

The third figure of merit is the time and amplitude of the reconstructed pulses. For isolated pulses, the network achieves a mean timing error of $1.7\,ns$, against 8–9 ns for the classical methods. The classical result is limited by the discrete correlation maximum, which shifts by several samples as the amplitude approaches the noise level. The network error grows smoothly with multiplicity, up to $6.1\,ns$ at six pulses per window, and with decreasing amplitude. This is as expected: the timing information is carried by the few statistically significant samples of the rising edge, and the limit is set by the noise. The amplitude accuracy of 1.1 counts, about 9% of a typical pulse, is roughly three times better than that of either classical method. The main reason is that an overlapping pulse sits on the tail of its neighbour, which the network has implicitly learned to subtract.

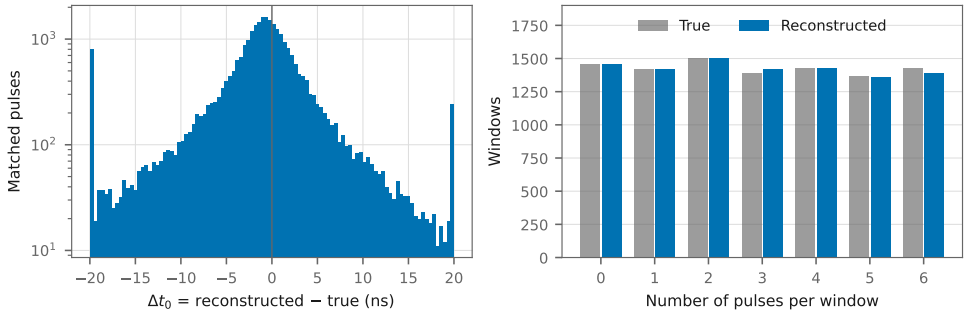


Fig. 3. Left: distribution of the timing residual Δt_0 for all matched pulses of the validation sample; the scale is logarithmic, and values beyond $\pm 20\,ns$ are collected in the edge bins. Right: true and reconstructed multiplicity spectra for the same sample.

The left panel of Fig. 3 shows the distribution of the timing residual Δt_0 for all matched pulses. Its core is narrow and nearly centered, and the median absolute error is $2.3\,ns$. The mean absolute error of $4.6\,ns$ is inflated by the visible tails, which are associated with close pulse pairs whose mutual assignment is ambiguous. The slight asymmetry towards negative Δt_0 has a simple origin: when two pulses partially merge, the reconstructed signal is shifted towards the rising edge of the composite structure. The right panel compares the reconstructed and true multiplicity spectra. In every bin, the reconstructed spectrum reproduces the true one to within about 2%. This is directly relevant to rate measurements and dead-time corrections, where only the spectrum is used, not the window-by-window assignment.

Summary and Outlook. Two small convolutional networks, trained exclusively on simulated waveforms, correctly determine the multiplicity of overlapping scintillation pulses in 96.5% of the readout windows. They reconstruct the arrival times and amplitudes with a mean accuracy of 4.6*ns* and about 9%, while missing fewer than 1% of the pulses. Compared with the matched filter and the commonly used iterative template subtraction, the gain in the accuracy of the parameters of detected pulses is moderate, but the advantage in detection is fundamental. The network requires no prior hypothesis about the number of pulses and performs no per-event fit. This distinguishes it from the multi-template fits of collider experiments, where the arrival times are tied to the bunch-crossing clock [6]. Here the arrival times are free, and it is precisely in this regime that classical fitting is most vulnerable.

This work was supported by the Science Committee of the MESCS RA, in the frames of the research projects No. 23AA-1C023 and No. 21AG-1C028.

Received 29.09.2026
Reviewed 21.10.2026
Accepted 26.10.2026

REFERENCES

1. Knoll G.F. *Radiation Detection and Measurement* (4th ed.). Wiley (2010).
2. Spieler H. *Semiconductor Detector Systems*. Oxford University Press (2005).
3. IAEA. *Instrumentation for Digital Nuclear Spectroscopy. IAEA-TECDOC-1706*. International Atomic Energy Agency (2013).
4. Lee M., Lee D., et al. Pulse Pileup Correction Method for Gamma-ray Spectroscopy in High Radiation Fields. *Nucl. Eng. Technol.* **52** (5) (2020), 1029–1035. https://doi.org/10.1016/j.net.2019.12.003
5. Cleland W.E., Stern E.G. Signal Processing Considerations for Liquid Ionization Calorimeters in a High Rate Environment. *Nucl. Instrum. Meth. A* **338** (2–3) (1994), 467–497. https://doi.org/10.1016/0168-9002(94)91332-3
6. CMS Collaboration. Reconstruction of Signal Amplitudes in the CMS Electromagnetic Calorimeter in the Presence of Overlapping Proton-proton Interactions. *JINST* **15** (2020), P10002. https://doi.org/10.1088/1748-0221/15/10/P10002
7. Högbom J.A. Aperture Synthesis with a Non-regular Distribution of Interferometer Baselines. *Astron. Astrophys. Suppl.* **15** (1974), 417–426.
8. Guo W., Gardner R. P., Mayo C. W. A Study of the Real-time Deconvolution of Digitized Waveforms with Pulse Pile up for Digital Radiation Spectroscopy. *Nucl. Instrum. Meth. A* **544** (3) (2005), 668–678. https://doi.org/10.1016/j.nima.2004.12.036

9. Kuhn H. W. The Hungarian Method for the Assignment Problem. *Naval Research Logistics Quarterly* **2** (1955), 83–97.
10. Yu D., Kolbæk M., et al. Permutation Invariant Training of Deep Models for Speaker-independent Multi-talker Speech Separation. *Proc. ICASSP* (2017). https://doi.org/10.1109/ICASSP.2017.7952154

Ա. Ա. ՀՈՂՄՐՑՅԱՆ

ԹՎԱՅՆԱՑՎԱԾ ՕՍՑԻԼՈԳՐԱՄՆԵՐՈՒՄ ՀԱՄԸՆԿՆՈՂ ՍՑԻՆՏԻԼՅԱՑԻՈՆ ԻՄՊՈՒԼՍՆԵՐԻ ՀԱՇՎՄԱՆ ԵՎ ՏԵՂԱՅՆԱՑՄԱՆ ՆԵՅՐՈՑԱՆՑԱՅԻՆ ՄԵԹՈԴ

Հաշվման բարձր արագությունների դեպքում սցինտիլյացիոն իմպուլսները համընկնում են ընթերցման պատուհանում, և երբ դրանց ժամանման ժամանակները սահմանված չեն արագացուցչի փակփով, վերականգնման դասական մեթոդները սկսում են կորցնել դրանք: Առաջարկվում է նեյրոնային ցանցերի վրա հիմնված մեթոդ, որը հաշվում է համընկնող իմպուլսները և չափում դրանցից յուրաքանչյուրի ժամանման ժամանակն ու ամպլիտուդը: Մեթոդը հետազոտվել է 1 ԳՀց հաճախությամբ թվայնացված մոդելավորված օսցիլոգրամների վրա, որոնք պարունակում են զրոյից մինչև վեց իմպուլս: Երկու փոքր փաթույթային ցանցերը վերականգնում են առանձին իմպուլսների 99,2%-ը՝ 4,6 *նվ* միջին ժամանակային սխալով և ԱՖՓ-ի 1,1 միավոր ամպլիտուդային սխալով, ինչը կազմում է բնորոշ ամպլիտուդի 9%-ը: Իմպուլսների քանակը ճիշտ է որոշվում պատուհանների 96,5%-ում: Նույն պայմաններում և գնահատման նույն եղանակի դեպքում համաձայնեցված ֆիլտրը վերականգնում է իմպուլսների միայն 46%-ը՝ 7,5 *նվ* ժամանակային սխալով, իսկ էտալոնային իմպուլսի իտերացիոն հանումը (CLEAN տիպի մեթոդ)՝ 57%-ը՝ 5,2 *նվ* սխալով: Կեղծ հայտնաբերումների մասնաբաժինը նեյրոնային ցանցի համար կազմում է 0,4%, CLEAN-ի համար՝ 2,6%, իսկ համաձայնեցված ֆիլտրի համար՝ 12% :

А. А. ОГМРЦЯН

НЕЙРОСЕТЕВОЙ МЕТОД ПОДСЧЕТА И ЛОКАЛИЗАЦИИ ПЕРЕКРЫВАЮЩИХСЯ СЦИНТИЛЛЯЦИОННЫХ ИМПУЛЬСОВ В ОЦИФРОВАННЫХ ОСЦИЛЛОГРАММАХ

При высокой скорости счета сцинтилляционные импульсы перекрываются в окне считывания, и, когда времена прихода не заданы тактом ускорителя, классические методы реконструкции начинают их терять.

Предложен нейросетевой метод, который подсчитывает перекрывающиеся импульсы и измеряет время прихода и амплитуду каждого из них. Метод исследован на моделированных осциллограммах, оцифрованных с частотой 1 *ГГц* и содержащих от 0 до шести импульсов. Две небольшие сверточные сети восстанавливают 99,2% отдельных импульсов со средней ошибкой времени 4,6 *нс* и ошибкой амплитуды 1,1 отсчета АЦП, что составляет 9% типичной амплитуды; кратность правильно определяется в 96,5% окон. В идентичных условиях и при одинаковом способе оценки согласованный фильтр восстанавливает лишь 46% импульсов с временной ошибкой 7,5 *нс*, а итерационное вычитание эталонного импульса (метод типа CLEAN) – 57% импульсов с ошибкой 5,2 *нс*. Доля ложных срабатываний составляет 0,4% для нейронной сети, 2,6% для CLEAN и 12% для согласованной фильтрации.