Engineering Innovations
ISSN: 2813-1002, Vol. 19, pp 9-21
doi:10.4028/p-lS9Zoj

Submitted: 2025-11-29
Revised: 2026-03-05
Accepted: 2026-03-20
Online: 2026-06-11

AI Tools for Plasma Diagnostics by X-Ray Imaging and Spectroscopy in the PANDORA Project Frame

Bianca Peri[1,2,a*], Eugenia Naselli[1,b], Giorgio Finocchiaro[1,3,c], Bharat Mishra[1,d], Angelo Pidatella[1,e], Richard Rácz[4,f], Sandor Biri[4,g], David Mascali[1,h]

[1]INFN, Laboratori Nazionali del Sud LNS, Catania, Italy

[2]University "La Sapienza", Rome, Italy

[3]University of Catania UniCt, Catania, Italy

[4]Institute for Nuclear Research (Atomki), Bem square 18/c, 4026 Debrecen, Hungary

[a*]bianca.peri@lns.infn.it, [b]eugenia.naselli@lns.infn.it, [c]giorgio.finocchiaro@lns.infn.it, [d]bharat.mishra@lns.infn.it, [e]pidatella@lns.infn.it, [f]rracz@atomki.hu, [g]biri@atomki.hu, [h]davidmascali@lns.infn.it

Keywords: Plasma Diagnostic, X-ray Spectroscopy, X-ray Imaging, AI tools, Neural Networks

Abstract. Magnetized plasmas in compact traps offer a unique environment for fundamental research. PANDORA (Plasma for Astrophysics Nuclear Decay Observations and Radiation for Archeometry) is a multidisciplinary project focused on the study of β decays in plasmas, using a novel facility that replicates stellar-like conditions.
A plasma diagnostics system based on a soft X-ray pinhole camera has been designed and implemented, with an innovative algorithm for Single-Photon Counting (SPhC) and High Dynamic Range (HDR) analysis. This enables space-resolved X-ray spectroscopy and the determination of magneto-plasma properties such as local thermodynamic parameters (in terms of electron density and temperature) and confinement dynamics.
This work presents results from an AI-based model in MATLAB designed to optimize the above-mentioned algorithm. Using *k-means* clustering, events with similar features were grouped to identify those that differentiate real event from spurious ones. A labeled dataset is then used to train a neural network to minimize pile-up, accelerating the recovery of high-resolution spectra and improving soft X-ray emission analysis.
This contribution details the current neural network development stage and its first applications to experimental data acquired during an experimental campaign carried out at the ATOMKI Laboratory.

Introduction

The PANDORA project, currently under development at INFN-LNS in Catania, is designed to explore β-decay properties from a pool of radioisotopes for the first time in a plasma environment, employing an ECR (Electron Cyclotron Resonance) plasma trap capable of recreating stellar-like conditions [1]. PANDORA will also be an experimental platform for further studies on plasma optical features and astrophysical processes relevant to multi-messenger observations, including effects like opacity [2]. Additionally, it will enable a wide range of applications, including materials science, accelerator physics and ion source technology. Since plasmas inherently radiate across the full electromagnetic spectrum, a comprehensive multi-diagnostic system will be integrated into PANDORA to map these emissions in detail and retrieve information on plasma properties [3].
Among other techniques, X-ray plasma diagnostics were developed over the years, allowing for the volumetric characterization of ECR plasmas. The volume-integrated X-ray diagnostics in the soft-X domain (from 0.5 keV to 30 keV) were adopted to estimate the spectral temperatures in different plasma conditions and various ECR setups versus RF power, gas pressure, and the magnetic field profile [4, 5, 6].
Particular attention is devoted to the soft X-ray domain, as it is linked to warm electron populations essential for driving ionization dynamics and shaping the plasma charge state distribution [1,7].

The work on X-ray imaging techniques and space-resolved spectroscopy was initiated by our team around 20 years ago [8]. The same techniques have also been used in other plasma-based facilities, enabling the morphological study of different kinds of magnetically confined plasmas, evidencing their overall structure and the spatial distribution of the warmer electrons. This plays a main role in leading to X-ray emission via characteristic lines and bremsstrahlung radiation [9,10].
With a view to data evaluation, a novel algorithm for X-ray imaging in a Single-Photon Counting mode has been developed, enabling highly effective, space-resolved plasma spectroscopy in the soft X-ray range and providing new insight into magnetized plasma properties, including local thermodynamic conditions, confinement dynamics, and structural features. The diagnostic setup is based on an X-ray pinhole camera system [7, 11, 12]. Subsequent work focuses, for the first time, on the design and refinement of an AI-assisted diagnostic pipeline, including the optimization of clustering and neural network models, and on its implementation [13].
Using multiple datasets, each photon detection event (i.e., a photon striking the CCD) was analysed with respect to its geometrical and intensity-related characteristics. By applying a *k-means* clustering–based AI tool, groups of events with shared features were identified, revealing the parameters that distinguish genuine signals from spurious ones (i.e., pile-up events). For further details on k-means please refer to the section: “Unsupervised Learning: k-means Clustering”. This process allowed the creation of a labeled dataset for training a neural network capable of detecting and reducing pile-up artifacts.
The goal of this approach is to extract plasma emission spectra in a fraction of the usual time, while enhancing both energy and spatial resolution, maximizing the signal-to-noise ratio, and delivering unprecedented precision in characterizing soft X-ray fluorescence and bremsstrahlung emissions from such plasmas. The following sections detail the architecture of the AI-assisted diagnostic pipeline, the optimization of clustering and neural network models, and their application to experimental measurements for validation and first analysis on new plasma emitted X-rays datasets.
This work contributes to the growing integration of AI-based analysis methods across various research fields. These approaches are becoming increasingly relevant not only in plasma physics and diagnostics, but also in accelerator physics and ion-source development and medical applications. These techniques can enable rapid plasma diagnostics, as demonstrated here, continuous ion source tuning through machine-learning-based monitoring and corrective adjustments for long-term stability, and early failure detection [14, 15]. Significant contributions are also made by the comprehensive review by Dopp et al. [16] on data-driven methodologies, as well as the more detailed study by Kirchen et al. [17], focused on shot-to-shot instability analysis proving the potential of data-driven techniques for the optimization of beam-loading conditions in high-power plasma systems.

Experimental Set-up

The experimental campaign was conducted on the 2nd generation *ECR Ion Source (ECRIS)* at Atomki Laboratory (Debrecen, Hungary) [18, 19], in November 2024. A sketch and a picture of the plasma diagnostic experimental setup are respectively shown in Fig. 1.a) and Fig.1.b).

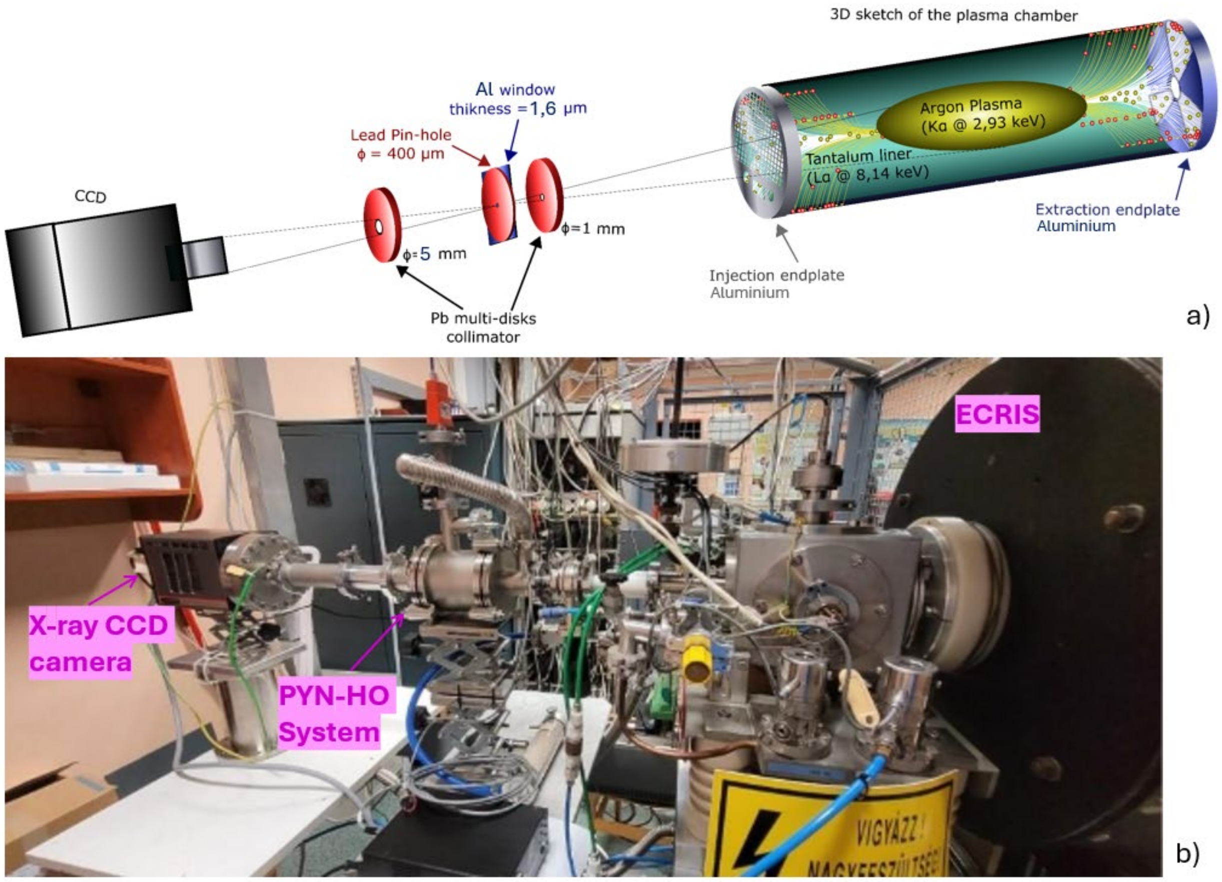

Fig. 1. a) Sketch of the experimental plasma diagnostic setup based on the pinhole CCD camera system. The Sophia Princeton CCD camera, coupled to a pinhole system, is currently integrated into the PYN-HO prototype. It includes a 400-μm-diameter lead pinhole, a multi-disk assembly of lead collimators, and an aluminum window used to shield the camera from UV–VIS radiation. The plasma chamber is also shown, featuring an argon plasma, the magnetic branch lines (where electrons are deconfined and impact on the plasma chamber walls), and the injection and extraction end plates. b) Picture of the experimental setup showing the CCD camera aligned with the pinhole system used to observe plasma emission within the ECRIS at ATOMKI laboratories (Debrecen, Hungary).

It is based on a pinhole X-ray camera system [20] properly designed for spatially, spectrally, and temporally resolved measurements of the plasma emission [21, 11, 22]. The setup consists of a multi-disc lead collimator, a 400 μm lead pinhole, an 800 nm thick aluminum window to shield the camera from UV–VIS radiation and a back-illuminated CCD detector (2048x2048 pixels with pixel size 13.5 μm, model *Sophia* by Princeton Instruments) with the quantum efficiency optimized in the window of soft X-rays, 0.6–30 keV energy range [11, 12].
To enable time-resolved studies, a dedicated X-ray PtIr shuttering system (4 ms of opening/closing time) was integrated, allowing millisecond-resolved acquisitions of fast-evolving plasma phenomena. This configuration enables simultaneous monitoring of plasma dynamics in energy, space, and time domains, providing crucial insights into confinement behaviour and instability evolution.
The overall system performance currently achieves an energy resolution of ~236 eV at 8.1 keV, a spatial resolution of about 500 μm, and a temporal resolution in the 5 ms range, depending on the exposure mode [11].
Complementary diagnostics—including HPGe and SDD spectrometers, scintillators, RF probes, spectrum analyzers, and fast photodiodes—were employed to perform multi-scale correlation studies between X-ray bursts, beam current fluctuations, and instability repetition rates [1].
The flexibility of the diagnostic platform allows systematic investigations of key plasma parameters under different operating conditions, including variations in microwave power, excitation frequency, magnetic field topology, and gas composition. In particular, the setup has been used to explore, among

others, the *gas-mixing effect* in ECR plasmas, a well-known phenomenon that enhances the production of high charge states [23]. By controlling the ratio between buffer and working gases, it was possible to reproduce the conditions under which this effect arises and to analyse the associated changes in X-ray emissivity and confinement dynamics.

Method: Hybrid Approach for Photon Event Classification

A hybrid analytical method that combines the efficiency of an advanced imaging algorithm—developed within the SPhC framework [7, 11, 12]—with both unsupervised and supervised machine learning models has been proposed.
The goal is to build an automated workflow capable of distinguishing valid photon events from spurious detections on a CCD sensor, thereby improving the reliability of event characterization. Our method integrates three key stages: i) feature extraction via a connectivity-based imaging tool, ii) clustering through *k-means* for preliminary data labelling procedures, and iii) supervised classification via a feed-forward neural network [13].

Feature Extraction via MATLAB Imaging Tool

The first step employs the MATLAB function *bwconncomp* [24], which identifies connected components on the CCD matrix and extracts a set of geometric and intensity-related features.
Each photon impinging on the CCD camera releases its energy in the silicon, generating a characteristic number of photoelectrons, proportional to its own energy. The Analogue Digital Unit (ADU), corresponding to the information read by the CCD in each pixel, is, therefore, proportional to the product of the number of incident photons by their energies. The Single-Photon-Counting mode is obtained by fixing a very short exposure time t_{exp} for each of the acquired image frames (several tens of milliseconds), minimizing the probability that two (or more) photons hit the same pixel or its nearest neighbours. SPhC t_{exp} was empirically set up in such a way that only a few pixels were illuminated on the full CCD matrix during a single-frame acquisition; a sequential acquisition of thousands of SPhC frames then allows us to gain the statistics necessary to obtain high-quality X-ray fluorescence spectra and images.
Fig. 2 shows a typical image-frame acquired with an exposure time of 5 ms. The detected events on the CCD are highlighted and visually validated.

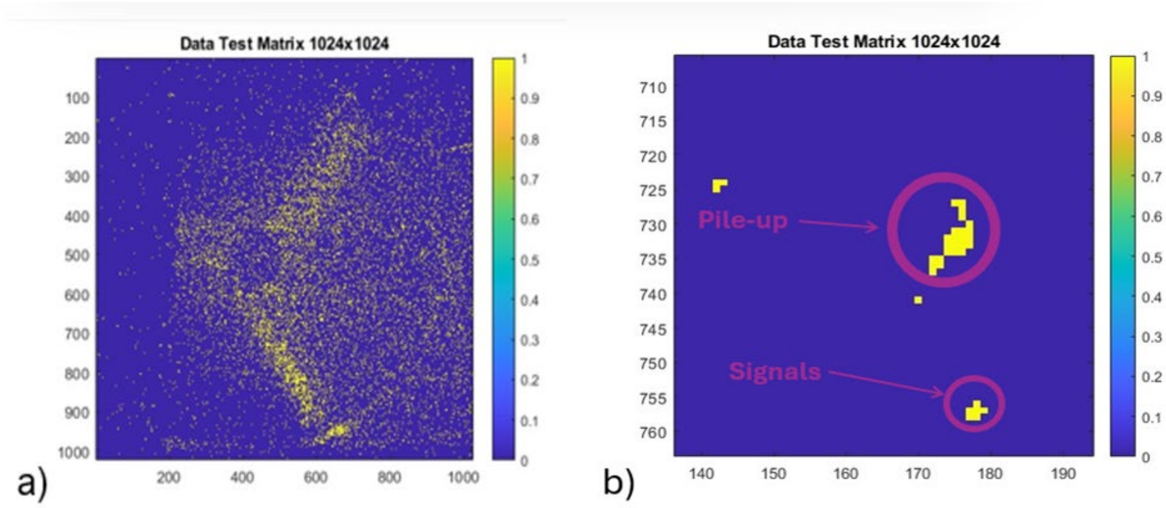

Fig. 2. CCD sensor matrix with detected events highlighted. (a) Binary map of events detected on a single CCD frame, showing the pixel-level detection. (b) Magnified view of the data matrix illustrating individual events and their spatial extent; this close-up facilitates extraction of features such as cluster size (number of pixels per event), local morphology.

Each detected event (i.e., the photon impinging on the CCD which releases its energy in a group of pixels labelled as *cluster*) is described by a group of parameters forming the basis of the structured dataset used for analysis. Among all extracted quantities that define the properties of each cluster, three features play a dominant role:

- Cluster size S – the total area occupied by the event, expressed as the number of connected pixels where the charge of the photon is released.
- Number of local maxima N_{LM} – the number of local maxima present in each cluster, which is a binary indicator of single versus multi-photon occurrence.
- Eccentricity e – the ratio between the major and minor axes of a given cluster, capturing the shape of the detected event.

These features were chosen due to their direct physical interpretability and their ability to effectively disentangle genuine photon detections and background noise.
In the Fig. 2.b) can be observed an example of two events/clusters. In the first case, it is a SPhC event because the charge is released in a single cluster with a single absolute maximum. In the second case, it is a spurious pile-up event because two or more photons release their energy in a cluster characterized by multiple local maxima. The main goal is to recognize and discriminate such event populations in order to optimize the performance of the analysis code by AI-based tools.
Consequently, a dataset is constructed that contains the information for each event, in particular referring to the three main parameters described above: S, N_{LM}, e.

Unsupervised Learning: *K-means* Clustering

K-mean is an unsupervised clustering algorithm that partitions a dataset into k populations by minimizing the within-cluster sum of squared distances [25]. The algorithm starts by initializing k centroids, either randomly or using a heuristic method. Each data point is then assigned to the nearest centroid based on the Euclidean distance metric. The centroids are updated by computing the mean of the points assigned to each cluster. This assignment–update process is repeated iteratively until convergence, i.e. the centroids no longer change significantly or a maximum number of iterations is reached.
The number of populations k was determined heuristically using the Silhouette Method [26], which evaluates how well each point fits within its assigned clusters compared to others.
The configuration that maximizes the Mean Silhouette Score (MSS) is k=2, as can be observed in Fig. 3, meaning *k-means* will identify two different populations within the given dataset.

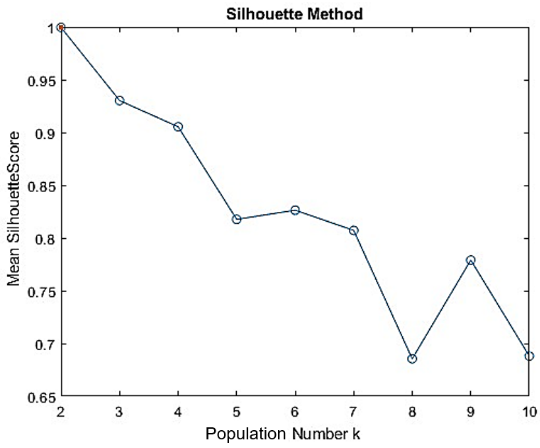

Fig. 3. Plot of the Mean Silhouette Score (MSS) against the number of Population k that *k-means* uses to group the data into.

Through *k-means*, two main populations were identified: one corresponding to small and compact events, and another mainly composed of bigger connected regions with a pronounced elliptical shape. This binary distinction provides the initial labeling needed for the subsequent supervised training stage.
Initially, *k-means* served as a valuable benchmark by rapidly identifying the cluster size threshold that had previously been established through a time-consuming fine-tuning process, involving numerous parameter trials. Subsequently, when provided solely with the parameters considered by the single-photon counting algorithm, *k-means* was able to reproduce the optimal spectrum with

remarkable fidelity, matching the previously achieved resolution (0.231 keV versus the 0.235 keV reference at the tantalum peak) and signal-to-noise ratio (12.05 compare to the 10.79 reference for titanium), but in a fraction of the time, generally under a minute.
As shown in Fig. 4, the application of the *k-means* algorithm was a fundamental benchmark result; in fact, it was able to determine a threshold value for the cluster size at 7 pixels for the CCD camera. In a matter of seconds, the model was able to match the threshold results obtained heuristically in previous studies [7]. Moreover, it's possible to see the statistical distribution of the events into the two categories: "Valid" with 61% events and 39% "Spurious" events.

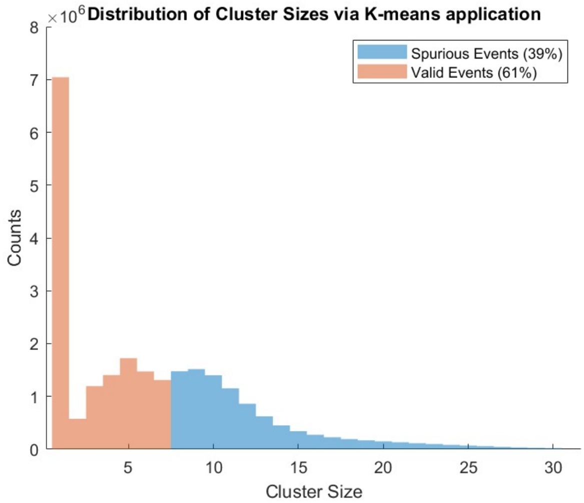

Fig. 4. Plot of counts versus cluster size after applying the *k-means* algorithm. This benchmark result shows that larger events (i.e. the higher the cluster size value) correspond to a multi-photon events occurring within the same region of the CCD.

Supervised Learning: Neural Network Design

With labeled data available, a feed-forward neural network for binary classification was designed and developed.
Before training, every feature was normalized to the range [0,1] using MATLAB's *mapminmax* [27] function to improve convergence and numerical stability.Additionally, class balance was verified to confirm that both categories were adequately represented, avoiding biased learning.
The dataset was randomly split into training (70%), validation (15%), and test (15%) subsets to prevent overfitting.

Network Architecture

The network was implemented using MATLAB's *patternnet* [28]. This feed-forward architecture processes data in a single direction, without feedback loops, and is well-suited for binary classification tasks.
The architecture includes, as shown in Fig. 5:

- Input layer: neurons equal to the number of selected features.
- Hidden layer: 10 neurons with a hyperbolic tangent (tanh) activation function, introducing non-linearity and enabling the network to learn complex relationships.
- Output layer: a single neuron with a sigmoid activation function that produces a probability score for event classification.

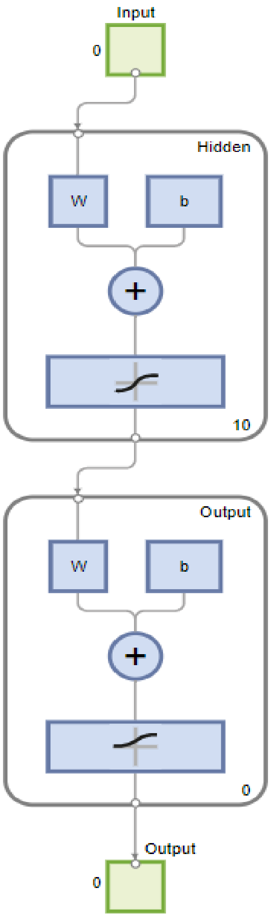

Fig. 5. Architecture of the feed-forward neural network: patternnet encoded in MATLAB with input, hidden, and output layers.

This structure offers an optimal balance between simplicity, computational efficiency, and learning capability.

Results

The following section presents the first obtained results. It begins with the training performance, then moves on to the benchmark results used to ensure the reliability of the network, and finally introduces the first results on a completely new dataset collected during an experimental campaign carried out at Atomki Laboratories in November 2024.

Training Performance

The training process relies on minimizing of the cross-entropy loss function, which measures the distance between predicted and actual class memberships.
Training was conducted over 300 epochs, during which the network demonstrated stable and rapid convergence, as shown in Fig. 6. Training was halted at the 128th epoch based on a validation checkpoint to prevent overfitting.
The cross-entropy loss consistently decreased, and the confusion matrix at the end of training revealed a high accuracy with consistent results across the training, validation, and test subsets.

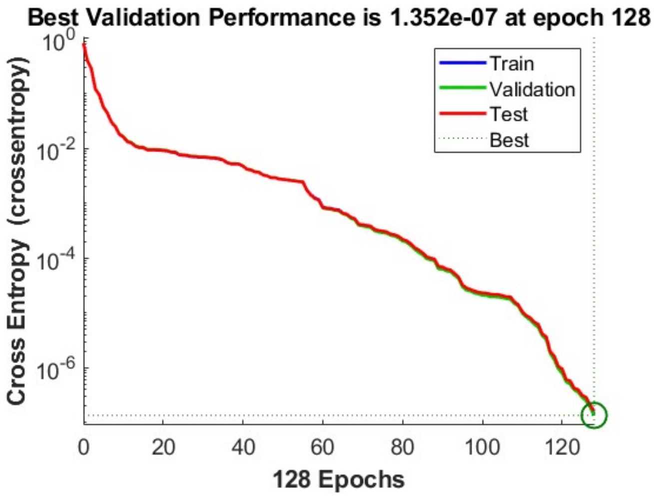

Fig. 6. Loss function plot during training, validation and test phases.

The convergence of the algorithm is further supported by the confusion matrix, which tracks the number of true positives and true negatives throughout the training, validation and test subsets. It provides a direct measure of the network's ability to correctly classify the events. In Fig. 7 the overall confusion matrix is shown.

Fig. 7. Confusion matrix showing the distribution of true positives, true negatives, false positives, and false negatives obtained during the classification task. The matrix illustrates the model's overall performance and its ability to correctly identify each event class. The Output Class represents the guess made by the net, while the target is related to the real class the event belongs to.

The selected parameters for the training process are summarized in Table 1.
The number of neurons in the hidden layer was selected through a trial-and-error procedure, given that an excessively large hidden layer may lead to memorization and overfitting, whereas an insufficient number of neurons prevents the network from learning effectively.
The hyperbolic tangent activation function was adopted for the dense layer as a standard choice for the classification task, while a sigmoid function was used in the output layer to allow the estimation of the probability that each event belongs to one class or the other. Finally, the Scaled Conjugate Gradient (SCG) algorithm was employed for training, as it accelerates and stabilizes convergence by avoiding line search and removing the need for learning-rate tuning. The cross-entropy loss function was used to measure the divergence between predicted and true class distributions, thereby directly optimizing classification performance.

Table 1. Parameters used for the training phase.

Parameters	Value
Neurons in the Hidden Layer	10
Training Algorithm	Scale Conjugate Gradient
Number of Epochs	300
Number of Events	7767611
Activation Function	Hyperbolic Tangent
Activation Function Output	Sigmoid
Minimum Gradient	10^-6
Maximum Fails	10
Loss Function	Cross Entropy

These outcomes confirm both the reliability of the feature selection and the effectiveness of the hybrid workflow. In particular, the unsupervised stage provided robust initial labeling, which allowed the supervised model to learn efficiently even with a relatively small dataset.
The presented hybrid methodology successfully integrates physical feature extraction, unsupervised clustering, and supervised learning within a single pipeline. This approach achieves high classification accuracy while remaining computationally efficient and scalable.
By coupling interpretable physical features with data-driven machine learning models, the method establishes a foundation for future extensions, including multi-class event characterization or online, real-time photon event analysis. The framework demonstrates how a combination of traditional imaging analysis and modern machine learning can advance automated, reliable event detection in experimental physics.

Benchmark Results

The benchmark analysis aimed to reproduce the same results obtained with the Single Photon Counting Algorithm developed at LNS [4, 5, 10], to prove the reliability of the new AI-based approach. The analysis was conducted on the experimental data coming from the campaign carried out at the Atomki Laboratories in 2018. The measurements were conducted with the 2nd generation ECRIS using a basic operation pumping frequency of 13.9 GHz at a power of 200 W. A special design of the plasma chamber (length ~ 210mm, diameter ~ 58mm) was implemented. The chamber lateral wall was covered by a liner of tantalum and made of titanium and aluminum (extraction and injection endplate, respectively). In this way fluxes of electrons escaping the plasma and impinging on the chamber walls can be detected by the fluorescence induced on the metals, which can be easily distinguished from the Ar emission in plasma.
Within the former algorithm, the presence of a threshold in the cluster size was estimated via a semiempirical approach which was hugely computational time consuming, properly balancing statistics and resolutions [7]. The optimize value was S=5.
The data were then analysed using the approach described in the section: “Unsupervised Learning: k-means Clustering” employing of *k-means*. In this case, the determination of the cluster size threshold value is optimized and fully automated, requiring significantly less computational time than the previous semi-empirical approach, while still reaching the same final value: S = 5, as shown in Fig. 8.a). The *k-means* algorithm was able to identify two different populations among the photon events, considering the commonalities between the events features. The results produced the same threshold identification in a matter of minutes. Encouraged by the compatibility between the two algorithms, the spectra obtained within the two were compared.
To evaluate the effectiveness of the cluster-size threshold in distinguishing true SPhC events from spurious ones, the events can be processed to reconstruct the corresponding experimental spectra. In the case where the population corresponds effectively to single-photon events, so that energy information is indeed conserved, one should obtain a well-resolved spectrum in which it is possible to distinguish the characteristic peaks of the system's constituent elements. Referring to the dataset

analysed, the fluorescence lines can be originated from the plasma (Ar) emissions and from the plasma chamber walls (Al, Ti, Ta) ones, due to electrons impacting them and emitting intense radiation.
The benchmark results are displayed in Fig.8.b). The spectra related to the population with cluster size S<=5, analysed by the traditional SPhC algorithm and the *k-means* based one are respectively showed in blue and in green. It is possible to highlight the agreement between the two approaches and that the result shows well-resolved spectra attributable to good single photo-counting events, where each characteristic fluorescence peak is well distinguishable (Ar-Ka, Ti-Ka, Ti-Kb, Ta-La, etc.). Instead, the other population with cluster-size S>5 is mainly affected by spurious events since the spectrum is not well resolved and the characteristic fluorescence lines cannot be distinguishable. The red spectrum shows the spectrum for the entire dataset, without separating the two populations.
It is important to highlight that the spectrum analysed with the new approach is fully compatible with the one obtained with the previous method, which serves as a validation of benchmark measurements.

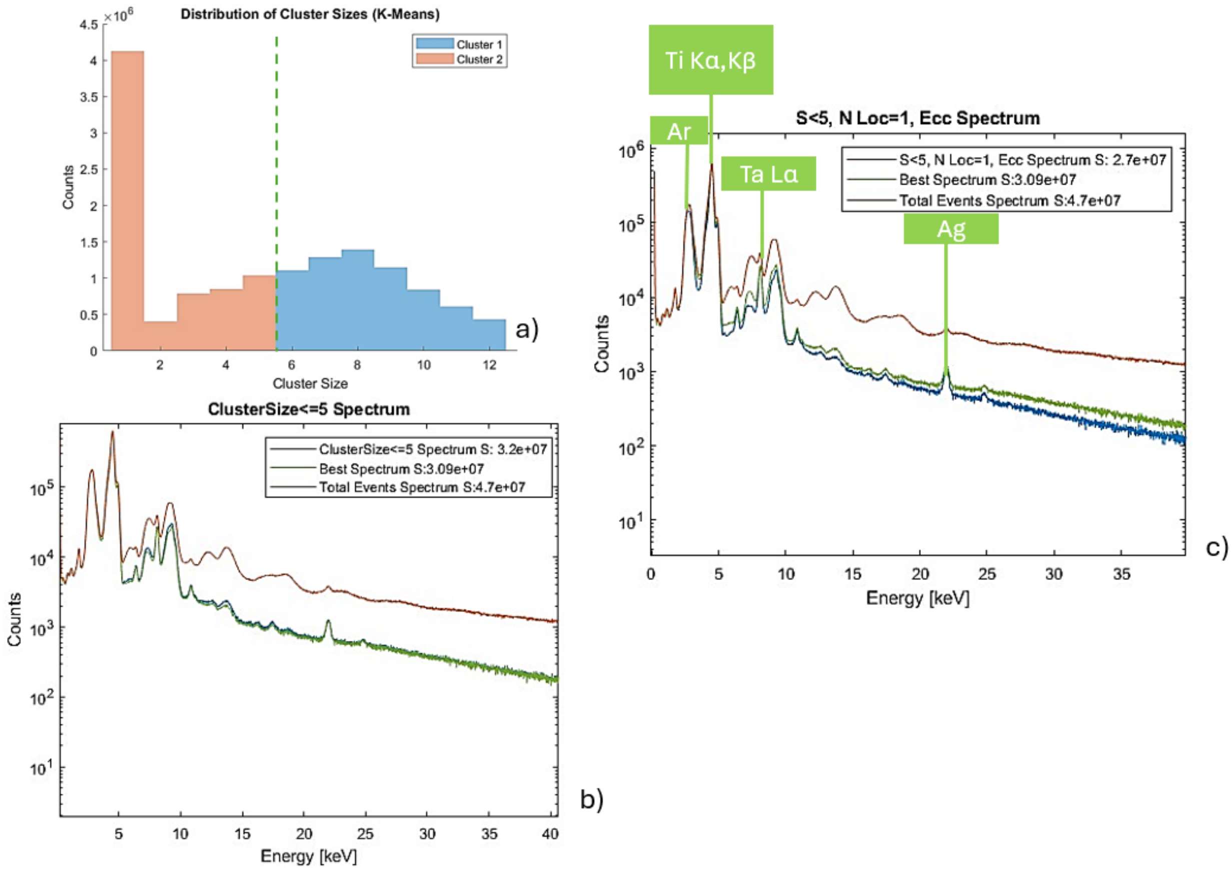

Fig. 8. Panel displaying the benchmark results obtained. In a) the plot features the Area, in terms of number of pixels, against the cluster ID. The two different populations are marked in red and in blue. The green dashed line represents the threshold value of S=5 pixels. In b) plot shows the comparison between different spectra. In red, the spectrum of the entire dataset, without any processing. The green curve represents the best spectrum obtained with the Single Photon Counting algorithm, while the blue plot is the spectrum obtained via application of the *k-means* discrimination algorithm, reducing the space of the parameters to the one employed with the former algorithm. Lastly in c) plot displays the comparison of the spectrum obtained with the *k-means* tool considering multiple features, while the green and the red define in b).

Lastly, multiple parameters in the *k-means* step were added to consider not only the cluster size but also the number of local maxima and the eccentricity.
The spectra are shown in Fig. 8.c) respectively in blue and green for the one analysed by the SPhC algorithm and for the *k-means* based one. The results compared show how the new results are indeed

coherent, showing a spectrum with slightly less statistic, but comparable FWHM and signal to noise ratio, in the tantalum peak.
The main spectral parameters considered in this analysis provide essential information on the physical conditions of the emitting plasma and results are summarized in Table 2. The FWHM (in keV) reflects the broadening of the fluorescence line. The line intensity (in counts) quantifies the emission strength, while the I_{MP}/I_{DP} and I_{MP}/I_{EP} ratios represent, respectively, the ratio between the intensity of the main peak of the Ti (@ 4,51 keV) and of its dimer peak (@ 9,2keV) and the ratio between the intensity of the main peak of the Ta (@ 8,1 keV) and of its escape peak (@ 6,4keV) [7]. Together with the total counts, these parameters form a consistent set of indicators for characterizing the emission spectrum and its underlying dynamics.

Table 2. Summary of the comparison parameters between the two spectra, the one obtained via the SPhC algorithm and the other via *k-means*.

Spectra	**FWHM Ta (keV)**	**Intensity Ti (Counts)**	**I_{MP}/I_{DP} Ti**	**I_{MP}/I_{EP} Ta**	**Total Counts**
SPhC Algorithm	0,25	17,38x10^6	10,79	3,31	3,1 x10^7
K-means	0,25	15,52 x10^6	12,05	3,42	2,7 x10^7

The above-mentioned results were indeed promising, laying the foundation for further development of the technique.

Spectra Retrieval

Once validated, the method was applied to new datasets for more effective and rapid data processing. The two datasets were collected at the ATOMKI Laboratories with the same setup used for the training purposes and described in the experimental set-up section. A basic operation pumping frequency of 14.25 GHz at a power of 205 W was used. Different gas mixtures were tested: i) pure krypton gas and ii) gas mixing set with krypton and neon.
For each configuration, 2000 frames with an exposure time of 160 ms each were acquired and elaborated to obtain the SPhC image and spectrum.
The parameters set for both datasets are summarized in Table 3.

Table 3. Camera and plasma parameters for each configuration. The first refers to the training dataset, where a pure krypton gas was employed, while the second refers to the gas mixing set featuring a krypton and neon plasma.

Dataset	RF Pumping	Net Power	COILS %			Plasma	Pressure
	Freq (GHz)	W	Inj	Mid	Ext	Type of gas	Mbar
Training	14,25	205	98	70	98	Kr	2,1 x10^-6
Test	14,25	204	98	70	98	Kr + Ne	1,2 x10^-6

Once trained, the new dataset is feed into the net. The net itself acts as a classifier, meaning that the incoming data are classified into two categories: “Valid events” and “Spurious events”, based on their features (i.e. cluster size, number of local maxima, etc).
Computationally, the implementation operates efficiently even on non-specialized hardware. Once the dataset events are correctly classified, the spectra of the valid events can be retrieved in under 15 s.
The results were benchmarked by means of the SPhC algorithm currently in use, validating those obtained so far.
The spectra coming from the training dataset and the gas mixing one are shown in Fig. 9. As expected, can be observed a well distinguished peak at low energy coming from the neon presents. Note that a linear energy calibration was performed for the time being.

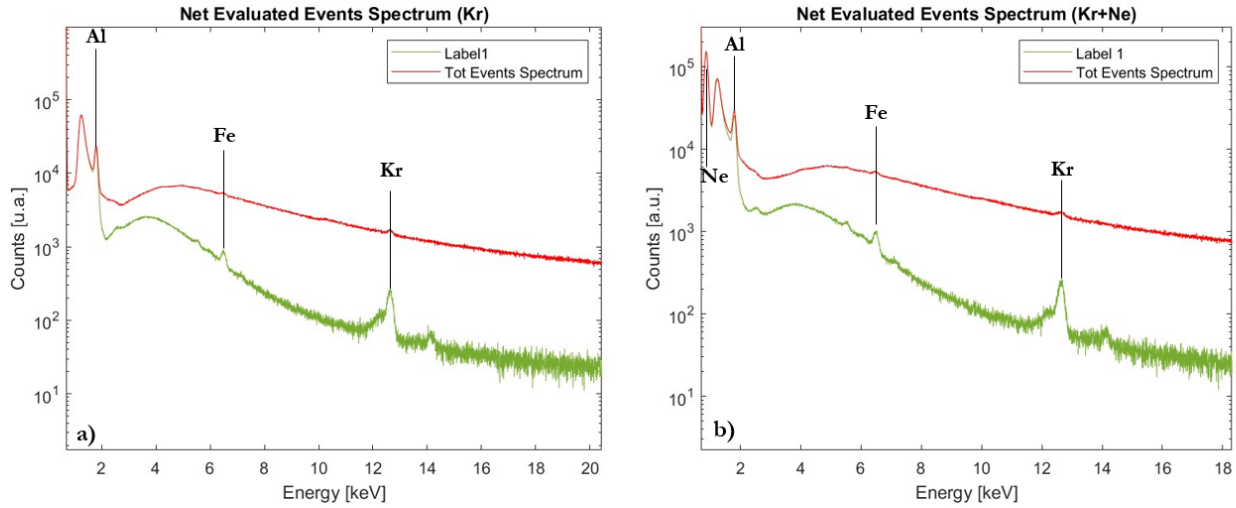

Fig. 9. X-ray spectra evaluated for a) the training set, only krypton plasma and b) gas mixing configuration combining neon and krypton. In both spectra a red curve it's shown that refers to the spectrum of the total events (valid and discarded, before the net discrimination)

The spectra are well resolved if compared to the spectra of the whole dataset events, as expected. The FWHM were evaluated at the krypton peak, and they resulted to be respectively: 0.1967 keV and 0.1983 keV @ 12.4 keV.

Conclusions

In conclusion, a hybrid approach was employed to analyse data acquired from the CCD diagnostic system operating in Single Photon Counting (SPhC) mode. The workflow combines an AI-based unsupervised clustering algorithm *k-means* to generate a labeled dataset, which was then used to train a supervised model. A feed-forward neural network encoded in MATLAB, *patternnet*, was subsequently developed and trained on these labeled data. This achievement served as a benchmark for the integration of machine learning techniques into the analysis pipeline for X-ray spectroscopy in magnetically confined plasmas.

Nevertheless, the current neural network was trained on a specific set-up, which limits its generalization capability, and while the feature extraction system provides a solid foundation for event characterization, the selection and weighting of features could be further optimized. To enhance performance and robustness, further training with diversified and previously unexplored datasets is ongoing. This step represents a crucial milestone toward creating an adaptive data analysis system capable of real-time event classification and pattern recognition in complex plasma environments. Moreover, although the current implementation focuses on offline data analysis, this framework lays the foundation for future adaptations toward online, real-time event monitoring, an objective that could further enhance experimental feedback and efficiency.

Acknowledgments

The authors gratefully acknowledge the support of Italian Ministry of University and Research (MUR) under the grant agreement PRINS2022THRKMK, CUP I53C24002400006.

The authors wish to thank INFN for the support through the project PANDORA_Gr3 funded by third Nat. Sci. Comm. This work has been partially funded by European Union (NextGeneration EU), through the MUR-PNRR project SAMOTHRACE (ECS00000022) and through the MUR-PNRR Project PE0000023-NQSTI.

This project has also received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No. 101057511.

References

[1] D. Mascali et al. Universe 8, 80 (2022)

[2] A. Pidatella et al., Front. Astron. Space Sci, 9 (2022)

[3] D. Mascali et al. Condensed Matter 9(2):28 (2024)

[4] U. Fantz, et al., Preface for the special topic on "ion source diagnostics". Rev. Sci. Instrum. 2022, 93, 101601.

[5] T. Thuillier, et al, Rev. Sci. Instrum. 2022, 93, 021102

[6] D. Nebem, et al, JINST 2019, 14, C02008

[7] E. Naselli et al., Condens. Matter, 7(1), 5 (2022)

[8] R. Racz, et al., Rev. Sci. Instrum. 2016, 87, 02A741.

[9] M. Anton, et al., Plasma Phys. Control. Fusion 1996, 1849, 38

[10] I. Song, et al., Curr. Appl. Phys. 2016, 16, 10

[11] G. Finocchiaro et al., Phys. Plasmas 31, 062506 (2024)

[12] G. Finocchiaro et al., Space and Time-Resolved X-rays Spectroscopy Technique by CCD Pinhole camera, in this proceeding

[13] Peri et al., AI Tools for Plasma Diagnostics by X-ray Imaging and Spectroscopy in the PANDORA Project Frame, EPS 51st Conference on Plasma Physics Proceeding, (2025)

[14] Y. Ong, et al, Advanced algorithms for linear accelerator design and operation. In Proceedings of the 32nd Linear Accelerator Conference (LINAC 2024) (pp. 288–291).

[15] D.S. Todd, et al, Applying Machine Learning Techniques to the operation of the Superconducting ECR ION Source VENUS. In proceeding of the 26th Int. Workshop Electron Cyclotron Resonance Ion Sources ECRIS 2024, Darmstadt, Germany

[16] Döpp et al., High Power Laser Sci. Eng. 11, e55 (2023)

[17] Kirchen et al., Phys. Rev. Lett. 126, 174801 (2021)

[18] S. Biri et al., The Atomki Accelerator Centre, Eur. Phys. J. Plus 136 (2021) 247

[19] S. Biri et al., Rev. Sci. Instrum. 83, 02A341 (2012)

[20] S. Biri et al 2018 JINST **13** C11016]

[21] E Naselli, et al.,2022 JINST **17** C01009.

[22] D. Mascali et al., 2022 Plasma Phys. Control. Fusion **64** 035020

[23] Melin, et al., J. Appl. Phys. 86, 4772–4779 (1999)

[24] *[bwconncomp]* (Version [R2009a]). Natick, Massachusetts: The MathWorks Inc., [2025]

[25] *[k-meansClustering]* (Version [R2006a]). Natick, Massachusetts: The MathWorks Inc., [2025]

[26] *[silhouetteEvaluation]* (Version [R2013b]). Natick, Massachusetts: The MathWorks Inc., [2025]

[27] *[mapminmax]* (Version [R2006a]). Natick, Massachusetts: The MathWorks Inc., [2025]

[28] *[patternnet]* (Version [R2010b]). Natick, Massachusetts: The MathWorks Inc., [2025]