Comparative Assessment of Feature Extraction for Fast Neutron Spectra Prediction with Machine Learning Algorithms using a CVD Diamond Detector

Enrica Belfiore, Pierre-Guy Allinei, Pierre Houedry, Meriem Bahhi, Simon Bartolacci, Adel Saleh, Mehdi Ben Mosbah, Rodolphe Antoni, Abdallah Lyoussi, Jean-Emmanuel Groetz

Comparative Assessment of Feature Extraction for Fast Neutron Spectra Prediction with Machine Learning Algorithms using a CVD Diamond Detector

Enrica Belfiore[1], *Pierre-Guy* Allinei[1], *Pierre* Houedry[2], *Meriem* Bahhi[3], *Simon* Bartolacci[4], *Adel* Saleh[5], *Mehdi* Ben Mosbah[1], *Rodolphe* Antoni[1], *Abdallah* Lyoussi[6] and *Jean-Emmanuel* Groetz[7]

[1] CEA, DES, IRESNE, DTN, Cadarache, Saint-Paul-Lez-Durance, 13108, France
[2] IRISA, École normale supérieure de Rennes, Rennes, 35170, France
[3] Institut de Mathématiques de Bourgogne, Université de Bourgogne, Dijon, 21078, France
[4] Institut Élie Cartan de Lorraine, Université de Lorraine, Nancy, 54012, France
[5] Institut de Mathématique de Marseille, Aix-Marseille University, Marseille, 13009, France
[6] CEA, DES, IRESNE, DER, Cadarache, Saint-Paul-Lez-Durance, 13108, France
[7] Laboratoire Chrono-environnement, Université de Franche-Comté, Besançon, 25030, France

***Abstract* - Machine learning algorithms are recognized as effective tools for addressing challenges in nuclear methodologies and instrumentation, particularly when obtaining experimental data or precise mathematical models is difficult. One notable challenge in fast neutron spectroscopy involves unfolding neutron spectra from solid-state detectors, where conventional deconvolution techniques often suffer from instability. This study uses the PHITS code to simulate the behavior of a CVD diamond detector subjected to various neutron spectra. The simulated data serve as input to a regression model aimed at reconstructing the original spectra. In particular, we investigate the impact of four different signal processing and feature extraction techniques—Fourier transform, nodal basis projection, wavelet decomposition and agglomerative clustering-based compression—on the performance of spectrum reconstruction. These transformations are applied as preprocessing steps, and the resulting features are then used to train a supervised regression model (linear or kernel ridge, depending on the method). Model performance is evaluated using the R^2 metric against reference spectra from the IAEA Compendium database. The most effective approach is then applied to experimental data collected with a CVD diamond detector at the DANAIDES irradiation facility, part of the TOTEM platform in Cadarache. This final test, performed on 14 MeV and AmBe neutron sources, demonstrates the practical feasibility of the proposed methodology for real-world neutron spectrum reconstruction.**

***Keywords* —Solid detectors, PHITS, neutron spectrum prediction.**

1. INTRODUCTION

The application of machine learning (ML) in radiation detection has garnered significant interest in recent years, driven by the need for more efficient, accurate, and autonomous systems. Traditional methods for radiation detection, which often rely on manual analysis and classical statistical techniques, face limitations in terms of processing speed, scalability, and adaptability to complex environments. Machine learning, with its ability to learn patterns from large datasets and improve over time, offers promising solutions to these challenges [1]. Within neutron spectroscopy, ML approaches offer the potential to gather data in scenarios where measurements are challenging with current technologies, thus facilitating the development of innovative measurement techniques. Neutron measurement is crucial for obtaining key nuclear physics data and ensuring real-time control in future reactors. Detectors in this environment must withstand harsh conditions, including intense radiation, high temperatures, and space constraints [2]. Semiconductor detectors have prompted researchers to explore their potential application in fast neutron spectroscopy [3] and they are effectively used in the Joint European Torus (JET) and will be integrated into International Thermonuclear Experimental Reactor (ITER) [4]. However, solid-state detectors face several limitations in neutron spectrum unfolding. These methods often involve matrix inversion and deconvolution procedures, which can lead to oscillatory artefacts and numerical instabilities when over-iterated in pursuit of unattainable resolution. In such cases, regularization or smoothing techniques are typically introduced to stabilize the solution, although these may themselves introduce further uncertainties.

Researchers from CEA Cadarache are exploring new methods to reconstruct the neutron spectrum from data collected by solid-state detectors through ML techniques [5, 6]. The feasibility of this approach was initially demonstrated in a theoretical study using a Tissue Equivalent Proportional Counter gas detector [7]. However, this detector was found to be unsuitable under high neutron fluxes. Consequently, the approach was applied to a Silicon-Carbide (SiC) solid-state detector, which emerged as a promising candidate for these conditions, yielding favorable results in neutron spectrum prediction via ML, supported by theoretical studies in [8]. The challenge encountered with the SiC detector mainly lies in managing the numerous resonances in the cross sections for neutron-silicon interactions. In this work, we explore an alternative detector material: Chemical Vapour Deposition (CVD) diamond. Recent advances in CVD techniques have significantly reduced defect densities, making diamond-based detectors more viable for neutron spectrometry in harsh environments [9, 10, 11]. Using the PHITS Monte Carlo code, we simulate the energy deposition spectra from CVD diamond detectors exposed to various neutron sources. The simulated deposited energy data in the detector are then processed using four distinct signal transformation techniques—Fourier transform, nodal basis projection, wavelet

decomposition, and agglomerative clustering—which serve as feature extraction steps to reduce dimensionality and extract relevant patterns from the raw signals. These features are used to train regression models (either linear or kernel ridge, depending on the case), with the goal of reconstructing the original neutron energy spectrum. Preliminary attempts to apply regression models directly to the raw input data were discarded, as such models suffered from poor generalization due to the high dimensionality and redundancy in the signal, which led to overfitting and unstable predictions. Therefore, the use of feature extraction methods proved essential to reduce complexity, improve robustness, and enable meaningful learning. The model performance is evaluated using the coefficient of determination (R^2) against reference spectra from the IAEA Compendium database. Finally, the best-performing approach is applied to experimental data acquired with a CVD diamond detector exposed to 14 MeV and Americium-Beryllium (AmBe) neutron sources at the DANAIDES irradiation casemate of the TOTEM facility at CEA Cadarache, demonstrating the applicability of the proposed method in real-world conditions. This paper is structured as follows: Section 2 presents a comparison between CVD diamond and SiC detectors in the context of high-flux neutron environments. Section 3 describes the detector model and the PHITS simulations used to generate the data. Section 4 outlines the data processing and machine learning methodology, detailing the signal transformations and regression models employed. Section 5 presents the reconstruction results from both simulated and experimental data. Finally, Section 6 offers conclusions and future perspectives.

2. CVD DIAMOND VS SIC DETECTORS

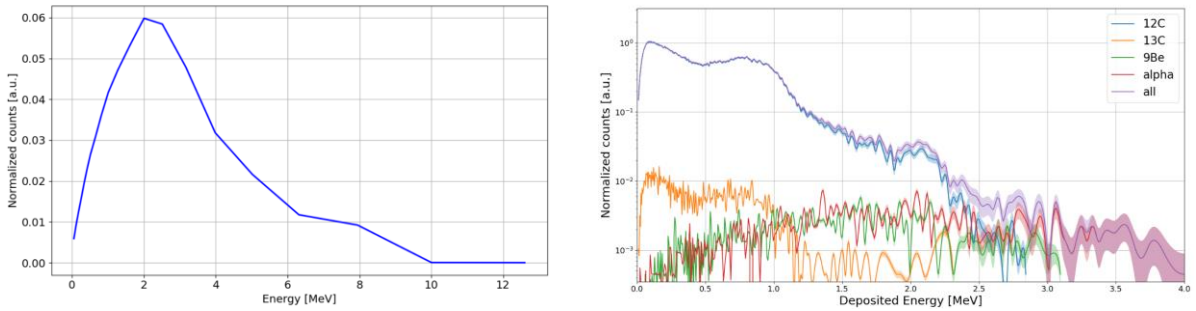
Comparative studies on the performance of CVD diamond and SiC detectors have gained considerable attention in fast neutron spectroscopy [12]. These detectors are among the most commonly used and commercially available semiconductor devices. Their operating principle relies on the collection of charge carriers generated by ionizing particles traversing the semiconductor material. Electrodes implanted within the semiconductor establish an electric drift field, which directs the movement of the ionized charges, leading to the collection of these charges and the generation of a detectable signal [13]. Both sensor types are considered promising candidates for use in harsh environments due to several advantageous properties. In particular, they exhibit high displacement threshold energies—ranging from 20–35 eV for SiC and 40–50 eV for CVD diamond—as well as wide bandgap energies of 3.27 eV and 5.5 eV, respectively. Additionally, they possess excellent thermal conductivities: 4.9 W/cm·K for the former and 22 W/cm·K for the latter. These features make them suitable for high-temperature applications without the need for additional cooling systems, and also render them more radiation resistant compared to other semiconductors such as germanium or gadolinium, etc [14, 15, 16]. In the past, diamond was dismissed as a semiconductor material for detection applications due to its high cost and the presence of impurities and defects in its crystal structure, which could lead to charge carrier issues and polarization effects. However, with advancements in manufacturing techniques such as CVD [17], which significantly reduce defect density, diamond-based sensors have become more accessible for industrial applications. In the analysis in [12], a comparison between the 4H-SiC p+-n diode and a single crystal CVD diamond detector is reported, considering 14 MeV energy neutron flux generated by a D-T neutron generator at the Technical University of Dresden. This study highlights that the selection of the most suitable sensor among these two is heavily influenced by the specific application, as both detectors in this case exhibit well-characterized pulse-height spectra with a clearly defined peak resulting from the $^{12}C(n, \alpha)^{9}Be$ reaction. While the energy resolution of this peak is marginally superior for the 4H-SiC-based detector, the diamond-based detector demonstrates a higher count rate.
In our application, which involves modelling simulations to populate the database used by the ML algorithm for neutron spectrum prediction, the SiC detector appears to be a more promising candidate. This is due to the inclusion of interactions with silicon, which enhance the variability of the system and increase the range of cases that can be addressed. However, previous studies have highlighted that discretizing the energy grid poses greater challenges because it must contend with multiple resonances in the cross sections for Si-n interactions. In this work, we therefore examine the CVD diamond detector for the same application, as it does not require such fine discretization of the energy grid.

3. CVD DIAMOND DETECTOR

To train and validate the machine learning models used in this study, a large dataset was generated via Monte Carlo simulations using the PHITS code [20]. This versatile Monte Carlo particle transport simulation software is capable of accurately modelling particle transport over a wide energy range, employing multiple nuclear reaction models and nuclear data libraries. The simulated detector is a single-crystal CVD diamond sensor with a sensitive volume of 1.4x1.4 mm^2 surface area and 320 μm thickness. The design of the CVD diamond detector was analyzed in [21]. The purpose of the simulation is to produce an energy-dependent dataset that mimics the deposited energy in the detector, obtaining a set of realistic deposited energy patterns corresponding to various incident neutron spectra. These data serve as input for supervised regression models aiming to reconstruct the original neutron energy distribution. The detector was therefore modelled as a parallelepiped with a thickness of 320 μm and a width of 1.4 mm, considering only the sensitive volume of the detector. It is assumed that excluding other parts in the detector design does not influence the output of the simulations since C-n interactions mainly occur within the sensitive volume of the detector. The source geometry is a rectangular solid, positioned in front of the detector. The neutrons are emitted in the z-direction, with an emission angle of 0° (perpendicular to the detector surface, along the +z-axis). Ten batches were set in the simulations, with 50,000,000 particles for each batch. In defining the materials, it was decided to consider a detector made of pure carbon, without doping or impurities. The carbon composition used is 98.94 % ^{12}C and 1.06% ^{13}C, in line with standard natural abundance [22]. All the nuclear data were taken from the JENDL-4.0 nuclear data library [23]. The spectrum was discretized in 600 energy bins, log-equal spaced from 100 keV to 15 MeV. The results of the simulations will provide the amount of energy deposited in the detector at each

point of the energy grid. These energy deposition values will serve as the features in the input vector for the machine learning models, as described in the following section.

To illustrate the purpose of the detector simulation, Figure 1 shows an example of a neutron energy spectrum taken from the IAEA Compendium database (Turkey Point reactor, fuel element at 26 in) and the corresponding deposited energy spectrum obtained from the PHITS simulation of the CVD diamond detector. These input-output pairs represent the data used to train the supervised regression model, which aims to learn the inverse mapping from the deposited energy signal (output of the simulation) to the original neutron spectrum (input of the simulation).



(a) (b)

Fig. 1. (a) Neutron energy spectrum from the IAEA Compendium database (Turkey Point reactor, fuel element at 26 in), (b) Corresponding deposited energy spectrum simulated with PHITS for a CVD diamond detector.

Figure 1a shows the input neutron energy spectrum for the simulation taken from the IAEA Compendium database, corresponding to a measurement at the Turkey Point reactor (fuel element at 26 inches). As evident, the spectrum contains a relatively low number of neutrons with energies exceeding 6 MeV. This energy marks the approximate threshold for the activation of the $^{12}C(n,\alpha)\ ^{9}Be$ reaction, which is one of the few channels leading to heavy charged particle production (alphas and beryllium nuclei) in carbon. As a result, the corresponding deposited energy spectrum obtained through PHITS simulation of the CVD diamond detector (Figure 1b) is dominated by the contribution from neutron elastic scattering with carbon atoms (shown in blue). The contributions from alpha particles (in red) and beryllium nuclei (in green), resulting from the $^{12}C(n,\alpha)\ ^{9}Be$ reaction, are present but significantly smaller due to the limited flux above threshold energy. The total deposited energy spectrum (shown in violet), which includes all particle contributions, is incorporated in the testing database for the machine learning model described in the following section.

4. MACHINE LEARNING APPROACH

This section outlines the methodology adopted to reconstruct the incident neutron spectrum from the simulated energy deposition in a CVD diamond detector using supervised regression techniques. Unlike classical unfolding approaches that rely on detector response functions, this method directly uses the simulated energy deposition profiles as input to train regression models capable of inferring the original neutron spectra. The use of machine learning in nuclear spectroscopy is gaining traction, with several studies exploring the application of regression and deep learning models for neutron and gamma-ray analysis [18, 19].

The training dataset used in this study was generated using the PHITS Monte Carlo code by simulating the response of the detector to 544 synthetic neutron spectra, designed to resemble realistic shapes as described in [8]. For each simulated case, the total energy deposited in the sensitive volume of the detector was recorded, producing spectra segmented into 600 logarithmically spaced bins from 0.1 MeV to 15 MeV. These spectra constitute the explanatory variables (input features) for the regression task.

The target variables are the corresponding neutron spectra used in the simulation, originally segmented into 55 bins across the energy range from 2×10^{-8} MeV to 44 MeV, in accordance with the IAEA Neutron Compendium format [24]. However, many bins in the energy range below 0.05 MeV contain negligible information for the CVD detector, which has low sensitivity in this region. For this reason, only the energy range from 0.05 MeV to 15 MeV, corresponding to 25 bins, was retained for training and evaluation. To address the high dimensionality of the input (600 bins), four signal processing techniques were adopted for feature extraction and dimensionality reduction prior to regression:

- **Fourier Transform**, which captures global spectral trends and periodicities;
- **Nodal basis projection**, which approximates the spectra using orthonormal polynomial bases tailored to spatially distributed signals;
- **Wavelet decomposition**, which captures localized features and discontinuities at multiple resolutions;
- **Agglomerative clustering**, which groups adjacent input bins based on their similarity, effectively reducing dimensionality while preserving local structure and correlations in the deposited energy signal.

These methods were selected to explore complementary perspectives on the input data, offering different trade-offs between localization, smoothness, sparsity, and interpretability in the reduced feature space. While Fourier and nodal decompositions are most suitable for capturing globally smooth patterns, wavelets and agglomerative clustering offer advantages in preserving localized or hierarchical structures. The use of feature extraction is crucial in this context, as it reduces computational complexity,

mitigates the curse of dimensionality, and enhances the model's ability to generalize to unseen data. Initial attempts to apply regression models directly to the raw 600-dimensional input data resulted in poor predictive performance and high computational cost. These outcomes can be attributed to the high dimensionality and redundancy of the input features, which introduce noise and multicollinearity, thereby impairing the learning process and leading to unstable or biased estimators.

All models were trained using a supervised learning framework implemented with scikit-learn [25], and the predicted neutron spectra were post-processed to enforce two physical constraints: non-negativity and normalization (i.e., the integral of each spectrum must equal one). An overview of the complete described reconstruction steps—from simulated neutron source to spectrum estimation—is presented in Figure 2.

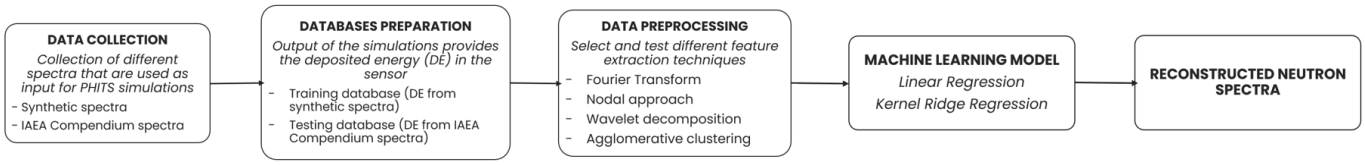


Fig. 2. Steps of the ML methodology for the reconstruction of neutron spectra with CVD diamond sensor.

4.1 Fourier Series

In the context of signal processing, it was considered to explore a Fourier-based approach. The idea involves treating each deposited energy spectrum as a function *x(t)* on [0,1], which is sampled at 600 points (corresponding to the number of energy bins in which the energy grid is discretized in the simulations). Subsequently, the Fourier series of this function can be computed.

For $n \in \mathbb{Z}$, the n-order Fourier series of x is expressed as follows:

$$S_n\big(x(t)\big) = \frac{1}{2}\, a_0(x) + \sum_{k=1}^{n}(a_k(x)\cos(kt2\pi) + b_k(x)\sin(kt2\pi)) \quad (1)$$

where:

- $a_0(x) = \int_0^1 x(t)dt$;
- $b_0(x) = 0$;
- $a_n(x) = 2\int_{[0,1]} x(t)\cos(nt2\pi)\,dt$, for $n > 0$;
- $b_n(x) = 2\int_{[0,1]} x(t)\sin(nt2\pi)\,dt$, for $n > 0$.

With X denoting the deposited energy spectrum data and Y the target neutron spectrum data.

The Fourier coefficients were calculated up to the order 30 for each deposited energy spectrum $x \in X$, and up to the order 5 for each neutron spectrum $y \in Y$. This resulted in two new datasets: $X_{Fourier}$ of size 544x60 and $Y_{Fourier}$ of size 544x10. Subsequently, a multi-output linear regression (with L2-penalization, $\alpha = 10^{-6}$) was trained to map $X_{Fourier} \rightarrow Y_{Fourier}$.

Finally, using the inverse of the Fourier transform equation (1) it is possible to reconstruct the prediction of Y.

4.2 Nodal approach

The second method proposed relies similarly on regarding the deposited energy spectrum as a function $X(f)$, where f represents a energy bin within the feasible energy range $[f_{min}, f_{max}]$. As previously stated, each row of the dataset comprises observations of the deposited energy X at bins $f_{min} = f_1 < f_2 < \cdots < f_N = f_{max}$, where N represents the total number of energy bins, i.e. 600, in the considered CVD diamond detector model. Consequently, each row generates a linear spline interpolation function $X_N(f)$ of X according to:

$$X_N(f) = \sum_{k=1}^{N} X_k \varphi_k(f)\ , \quad (2)$$

where φ_k satisfies the following conditions:

- it is continuous on $[f_{min}, f_{max}]$;
- it is piecewise linear on each $[f_k, f_{k+1}]$;
- it satisfies $\varphi_k = \delta_{kj}$ for all $k, j = 1, \dots, N$.

In equation (2) X_k represents the observed value of $X(f_k)$, i.e. the value in the k-th column, which is assumed to be a good approximation of $X(f_k)$. Instead of applying linear regression on X_N, it was decided to first select a significantly smaller number of bins $f_{min} = \tilde{f}_1 < \cdots < \tilde{f}_n = f_{max}$ with $n \ll N$ (arbitrary chosen), and then to define the vector:

$$X_n = (x_1, \dots, x_n) \coloneqq \Big(X_N(\tilde{f}_1), X_N(\tilde{f}_2), \dots, X_N(\tilde{f}_n)\Big). \quad (3)$$

We denote X_{Nod} as the new dataset of deposited energies of size 544xn, where each row is represented by X_n, and Y remains

unchanged with a size of 544x25. Linear regression (with L2-penalization, $\alpha = 10^{-6}$) is then applied from X_{Nod} to Y. Furthermore, $\tilde{f}_k$ lies between two consecutive bins f_j and f_{j+1}. Defining the parameter $\theta_k \in [0,1]$, it can be expressed as $\tilde{f}_k = \theta_k f_j + (1 - \theta_k) f_{j+1}$. From this formulation and the definition of X_N, it follows:

$$x_k \coloneqq X_N(\tilde{f}_k) = \theta_k X_j + (1 - \theta_k) X_{j+1} \quad , \tag{4}$$

in other words, $X_N(\tilde{f}_k)$ is just a weighted average of the two consecutive values of X_j and X_{j+1}. In our implementation, the bins $\tilde{f}_i$ are equidistant. In addition, instead of taking x_k as in (4), it was simplified with:

$$x_k = \frac{1}{2}(X_j + X_{j+1}) \tag{5}$$

as this did not affect the R^2 error much (as will be shown in the results of the test sets). In addition, a filter has been applied to X_n so that the integral is equal to 1 and that negative values are truncated to 0 similarly to the Fourier basis case.

4.3 Wavelet Decomposition

The Discrete Wavelet Transform is used to decompose the deposited energy into its approximation and detail components at different scales. This decomposition is achieved by applying low-pass and high-pass filters to the energy bins. Given a energy range divided in N bins $f = \{f_0, f_1, \dots f_{N-1}\}$, the goal is to express it in terms of scaling functions and wavelet functions at different levels. This can be written mathematically as:

$$f(t) = \sum_n c_A^1[n]\phi(t-n) + \sum_n c_D^1[n]\psi(t-n) \quad , \tag{6}$$

where:

1. $c_A^1[n]$ are the approximation coefficients,
2. $c_D^1[n]$ are the detail coefficients,
3. $\phi(t)$ is the scaling function (low-pass filter),
4. $\psi(t)$ is the wavelet function (high-pass filter).

For the Haar wavelets [26], the scaling function $\phi(t)$ and wavelet function $\psi(t)$ are defined as follows:

$$\phi(t) = \begin{cases} 1 & \text{if } 0 \le t < 1 \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \psi(t) = \begin{cases} 1 & \text{if } 0 \le t < \frac{1}{2} \\ -1 & \text{if } \frac{1}{2} \le t < 1 \\ 0 & \text{otherwise} \end{cases} . \tag{7}$$

The coefficients are given by:

$$c_A^1[n] = \sum_{m=0}^{1} h[m] f_{2n-m} \quad \text{and} \quad c_D^1[n] = \sum_{m=0}^{1} g[m] f_{2n-m} \quad , \tag{8}$$

where:

$$h = \left[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right] \text{ and } g = \left[-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right] \quad . \tag{9}$$

Finally, a Kernel Ridge Regression (KRR) model was applied between wavelet coefficients for a level equal to 1 and each bin of neutron spectrum ($y \in Y$). The hyperparameters were adjusted with a grid search only on:

1. the regularization parameter α of the RIDGE model.
2. the convolution kernel: linear, gaussian or cosine.

The best results were obtained with α=0.01 and cosine kernel.

4.4 Agglomerative Clustering

This model was chosen to maintain consistency with the study on SiC. The algorithm is based on a regularized linear method (l^2 regularization) associated with a convolution kernel [27]. The overall approach is structured as a two-stage pipeline, combining feature reduction via agglomerative clustering with supervised learning using Kernel Ridge Regression. The 600 energy bins of each deposited energy spectrum are reduced using agglomerative clustering with spatial connectivity constraints. Specifically, each energy bin is treated as a node in a graph, and edges are formed only between adjacent bins, enforcing locality-preserving clustering. This ensures that merged clusters correspond to contiguous energy regions. This hierarchical clustering minimizes a Ward-like objective function:

$$\mathcal{L} = \sum_{c=1}^{n} \sum_{x_i \in C_c} ||x_i - \mu_c||_2^2 \quad , \tag{10}$$

where C_c denotes the cluster c, μ_c its mean, and $n \in \{20,40,60,80,100\}$ is the number of clusters. The cluster centroids (means) are then used as the new features for downstream modeling, effectively reducing the input dimensionality from 600 to n. The KRR model is trained for each of the 25 neutron spectrum bins (i.e., a one-vs-all multi-output regression setup). The KRR model applies l^2-regularization to linear regression in a reproducing kernel Hilbert space, using a convolution kernel for nonlinearity. The prediction function for a single output bin is:

$$\hat{y} = \sum_{i=1}^{m} \alpha_i K(x_i, x) \quad , \tag{11}$$

where $K(\cdot, \cdot)$ is the kernel function and α_i are learned weights. Two types of kernels were explored, the Radial Basis Function (RBF)

$$K(x, x') = \exp\left(-\gamma \left|\left|x - x'\right|\right|^2\right) \tag{12}$$

and the polynomial kernel

$$K(x, x') = (\gamma \langle x, x' \rangle + r)^d \quad , \tag{13}$$

with degree $d \in \{2, 3, 4\}$, and γ, r as tunable parameters.

The hyperparameters adjusted include:

- The number of clusters of explanatory variables;
- The regularization parameter α of the RIDGE model;
- The convolution kernel: polynomial or RBF;
- The γ parameter used in distance calculation via RBF and polynomial kernels;
- The degree of the polynomial kernel: 2, 3, or 4.

An identical model is created for each bin, and then the hyperparameters of each of these models are optimized using a random search grid for the best set of hyperparameters.

A summary of the main characteristics of the proposed approaches is provided in Table 1.

Dimension reduction method	ML model	Input/output dimension	Reduced dimension	Regression details
Fourier basis	Linear Regression	600/25	60	$\alpha = 10^{-6}$
Nodal approach (linear interpolation)	Linear Regression	600/25	25	$\alpha = 10^{-6}$
Wavelet decomposition (Haar DWT)	Kernel Ridge Regression	600/25	300	Cosine kernel, α = 0.01
Agglomerative clustering	Kernel Ridge Regression	600/25	20-100	RBF/ polynomial kernel

Table. 1. Summary of reduction methods and characteristics.

5. RESULTS

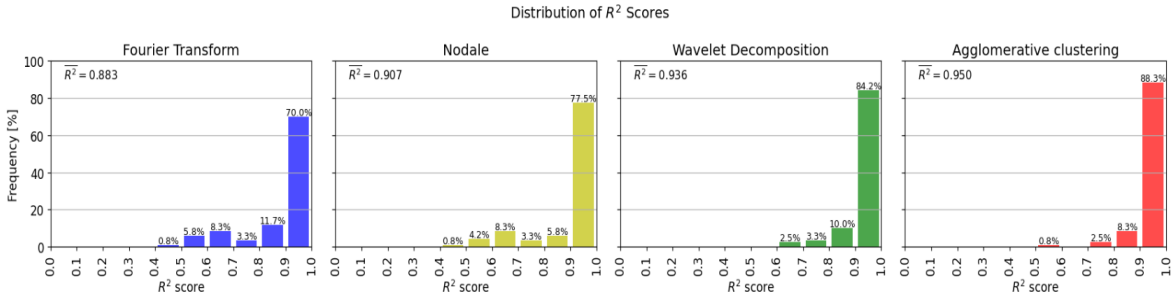
5.1 PREDICTION OF NEUTRON SPECTRA FROM IAEA COMPENDIUM DATABASE

To assess the accuracy in predicting the neutron spectrum, the R^2 metric was selected [28]. It was calculated for each sample and then averaged over the entire training or test subset, as formulated by:

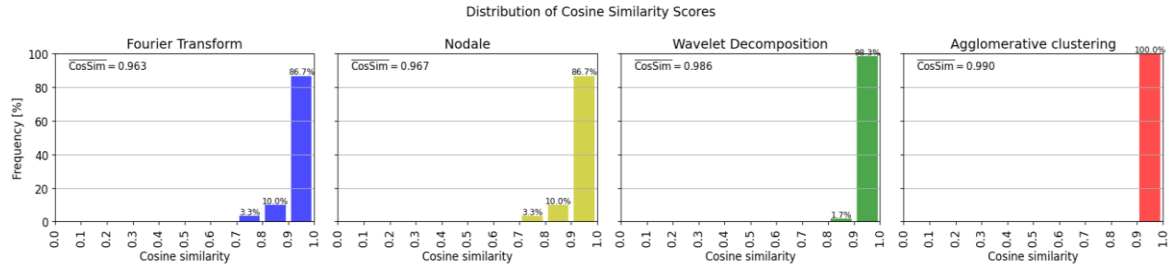
$$R_n^2 = \max\left\{0, \left(1 - \frac{\sum_{i=1}^{m}(t_{n,i} - \hat{t}_{n,i})}{\sum_{i=1}^{m}(t_{n,i} - \bar{t}_{n,i})}\right)\right\} \text{ and } \bar{R}^2 = \frac{1}{N}\sum_{n=1}^{N} R_n^2 \quad , \tag{14}$$

with $t_{n,i}$ and $\hat{t}_{n,i}$ the target and predicted fluence value respectively in the energy bin i for the spectrum n of the training or test subset and N is the total number of spectra analysed in the datasets. In addition to R^2, other metrics have been analysed to better capture the similarity between predicted and simulated spectra. In particular, cosine similarity was evaluated as it directly compares the shape of two spectra independently of their absolute magnitude. This metric is well suited to spectral data as it highlights whether the two vectors point in the same direction. As illustrated in the following figure, the trends observed using cosine similarity closely mirrored those found with R^2: samples with high R^2 also exhibited high cosine similarity, confirming the consistency of the two metrics in assessing prediction accuracy. Therefore, for the sake of clarity, interpretability, and consistency

with previous studies (e.g., the SiC case), R^2 is retained as the primary evaluation metric in the subsequent analysis of the collected results.



(a)



(b)

Fig. 3. R^2 score (a) and cosine similarity score (b) achieved for the test subset for the four described algorithms.

In Figure 3, the average R^2 obtained in predicting the neutron spectra from the test dataset is reported for the four feature extraction methods under study. An R^2 score greater than 0.9 guarantees a high-quality of neutron spectrum reconstruction. For the four algorithms, Fourier presents 70% of predictions in agreement with this criterion, Nodal 78%, Wavelet 84% and the Agglomerative clustering predicts in 88% of the cases a result with the desired R^2 value. Additionally, none of the spectra appear to be predicted with an R^2 lower than 0.4 for any of the four approaches. The relatively lower performance observed with the Fourier transform can be attributed to its intrinsic assumption of periodicity. This limitation arises from the fact that the Fourier series expansion is based on the principle of periodicity, which causes it to fail in representing the last energy bin if it does not align with the first one. These behaviours are visible in Figure 4. The Nodal basis approach yields intermediate performance: better than Fourier but inferior to Wavelet and Agglomerative clustering. This outcome is partially due to the use of a Linear Regression model, which was deemed appropriate in this case because the nodal projection significantly reduces the dimensionality of the input and tends to linearize the relationships between features and target variables. However, linear regression lacks the flexibility to model more complex, nonlinear dependencies, which can limit reconstruction accuracy when such interactions are present in the data. For both Wavelet decomposition and Agglomerative clustering, KRR was adopted as the learning model. These feature extraction methods preserve localized and potentially nonlinear structures in the input data, which justifies the use of KRR—capable of capturing such nonlinearities through the kernel trick. The kernel transforms the input space into a higher-dimensional feature space where linear regression can approximate complex relationships. This combination yields significantly better predictive performance, particularly in cases where the input-output relationship is highly nonlinear. Conversely, the Agglomerative clustering demonstrates poor predictive performance when attempting to reproduce a spectrum characterized by a 3-bump shape. This limitation arises from the absence of spectra with such a form in the training dataset. Despite this omission, the Fourier, Wavelet and Nodal models exhibit superior adaptability to representing the 3-bump spectra, achieving a higher average R^2 score compared to the Agglomerative clustering method for these cases.

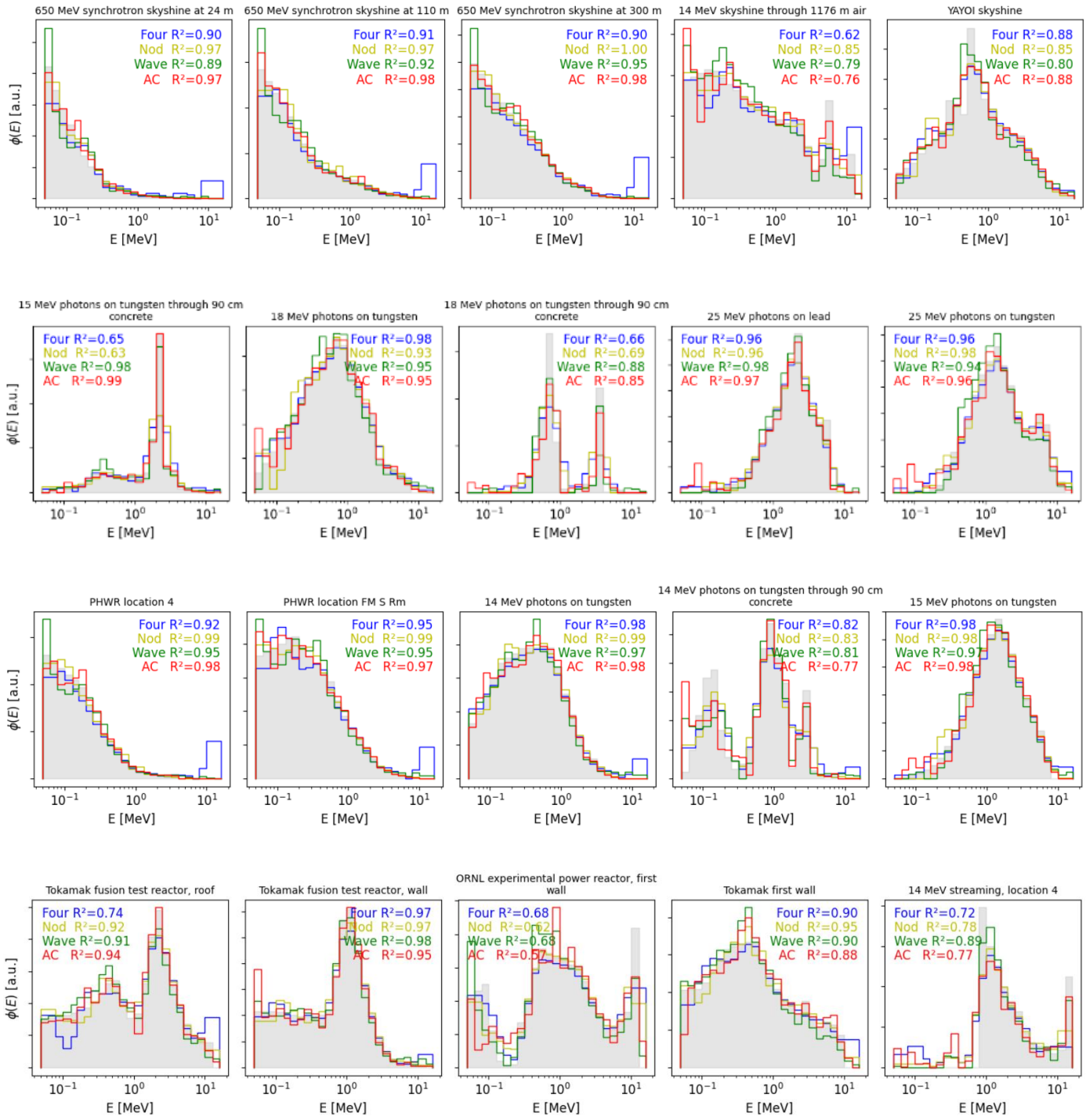


Fig. 4. Comparison of some real spectra taken from the IAEA Compendium database and predicted ones with the four described ML models (abbreviations; Four : Fourier Transform, Nod : Nodale; Wave : Wavelet Decomposition; AC : Agglomerative clustering).

Each model exhibits distinct strengths and limitations in reconstructing the neutron spectra, which opens the possibility of combining their predictions to enhance overall robustness and accuracy. One of the simplest methods is to calculate the median of predictions by bins. The advantage of the median is not to be penalized by extreme values as observed at the ends of the spectra predicted with the extracted features with the Fourier model.

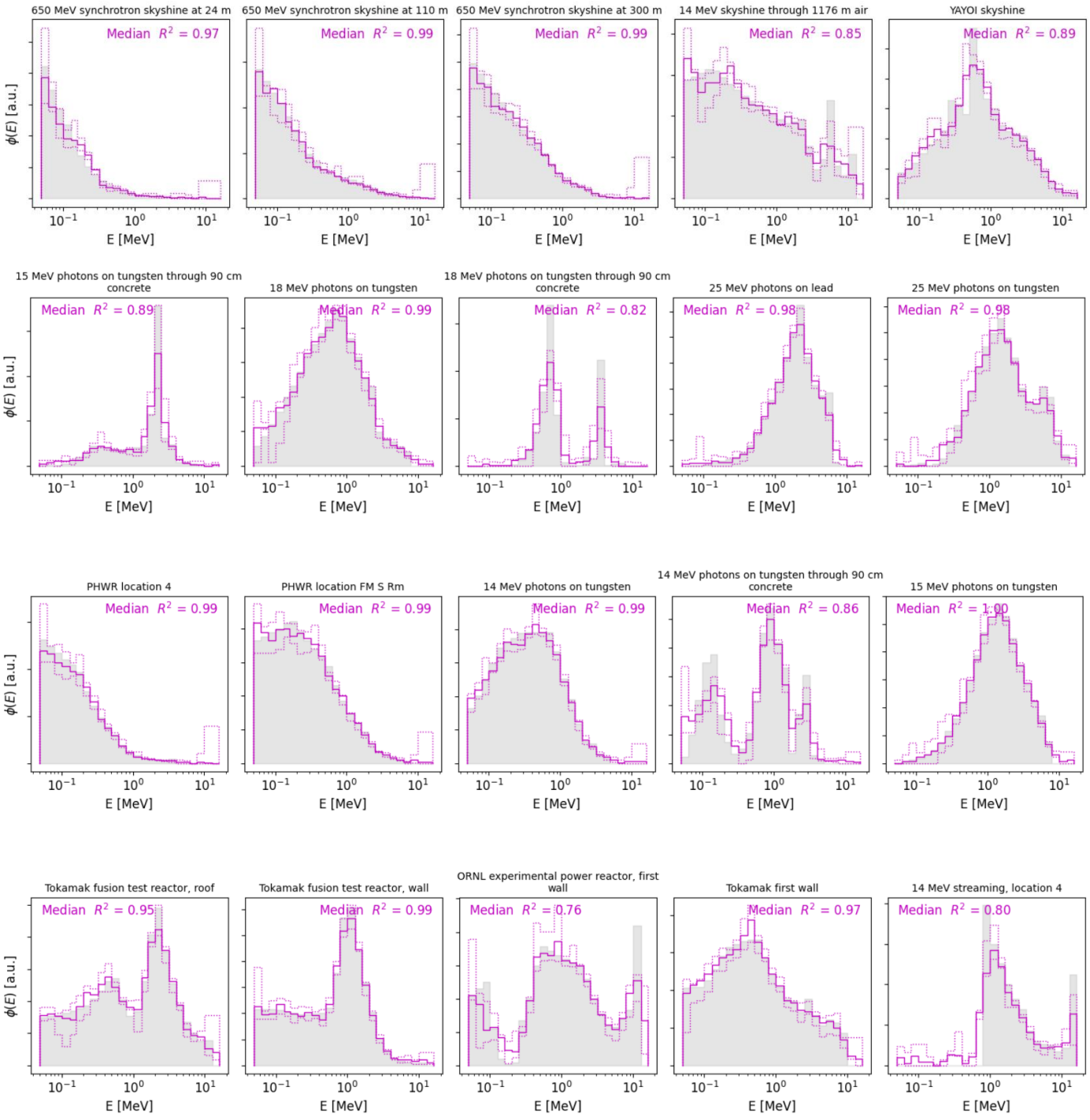


Fig. 5. Comparison of some real spectra taken from the IAEA Compendium database and predicted ones with the median prediction of the four described ML models.

Using the median of the predictions of the four algorithms produces spectra that are mostly better predicted. On the other hand, these predictions are more robust with regard to the individual behavioural differences of each model. It therefore works like a meta-model whose combination function is the median instead of the mean. In Figure 5, the maximum and minimum spectra predicted by the four models for each energy bin are shown with dashed lines. The comparison between the spectra predicted by the median model and the actual spectra from the IAEA Compendium database yields an $\bar{R}^2$ of 0.953, which is slightly higher than that of the Kernel Ridge model with Agglomerative clustering. However, in some cases, the spectra predicted by the median model have a lower R^2 compared to those predicted by Kernel Ridge with Agglomerative clustering. This occurs because the median model, by construction, selects the median prediction from the four models for each energy bin and it does not always correspond

to the best possible prediction. If one of the individual models systematically produces more accurate spectra in a particular region, the median model does not take full advantage of this superior performance but instead blends predictions from all models. Moreover, if multiple models exhibit a similar systematic error—such as a consistent underestimation or overestimation in certain energy bins—the median model will inherit this bias. In contrast, an individual model that happens to be less affected by this bias may produce a higher R^2 for specific spectra. Therefore, it is advisable to compare the estimation of the performance with an appropriate metric of individual models before selecting the median one.

5.2 Prediction of 14 MeV and AmBe source spectra

To evaluate the practical applicability of the proposed methodology, we tested the performance of the Median model on real measurements collected during an experimental campaign at the DANAIDES irradiation casemate within the TOTEM facility at CEA Cadarache [29]. In this facility, the CVD diamond detector was exposed to a 14.1 MeV neutron flux and AmBe neutron source. The D-T neutron generator GENIE16 model from SODERN of this facility was used for this experiment in continuous mode to obtain the 14.1 MeV, with the neutron emission equals to $2x10^8$ ns^{-1} and the Am-Be source with neutron emission $2x10^5$ ns^{-1}.

The sensor was linked to a current amplifier, supplied by Metraware, ensuring optimized integration with the subsequent components of the counting system and it was located 1cm far from the neutron emission point. The bias voltage was applied by a CAEN bias voltage supply controlled via laptop and the detector was operated at -300 V. Finally, the signals were acquired through a Lecroy Waverunner 9254 oscilloscope. A photograph of the experimental apparatus is shown in Figure 6, which includes the CVD diamond detector encapsulated in a protective housing (a), the current preamplifier (b) and the positioning of the sensor close to the neutron generator (c).

(a) (b) (c)

Fig. 6. (a) CVD diamond detector encapsulated in a protective housing, (b) the current preamplifier and (c) the positioning of the sensor close to the neutron generator.

To better illustrate the overall measurement configuration, Figure 7 provides a schematic layout of the experimental setup, including the neutron source, detector positioning, electronics chain (amplifier, HV supply), and acquisition system.

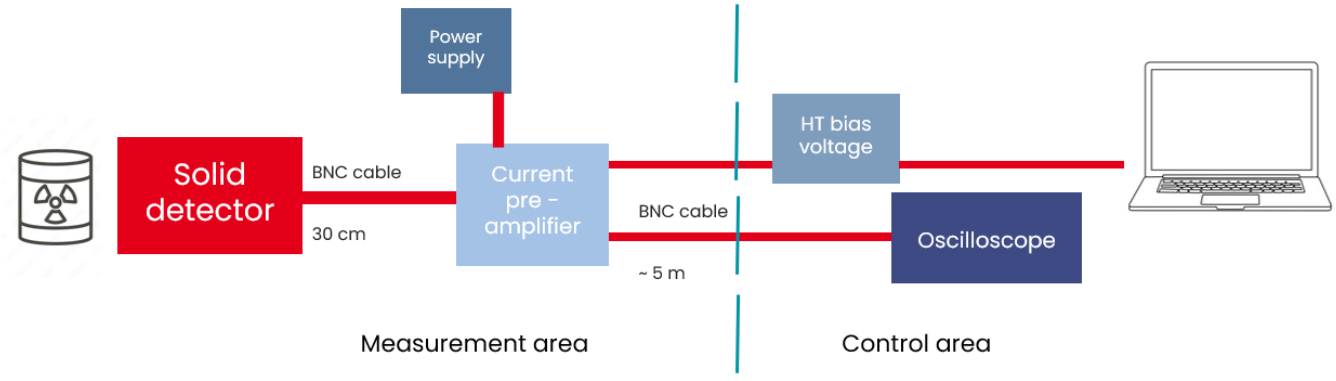
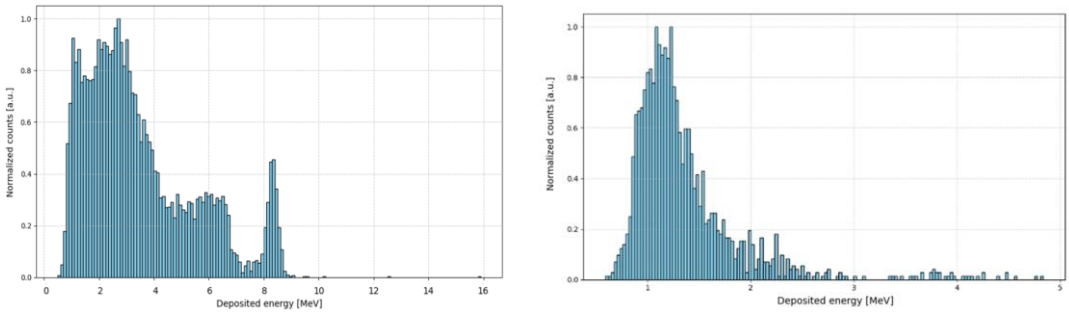


Fig. 7. Experimental setup scheme.

The collected pulses with the measurements with the CVD diamond detector were processed in order to obtain the deposited energy in the detectors, in accordance to [30]. It is important to note that for an accurate prediction of the neutron spectrum using this machine learning approach, a high count rate is not required; rather, a precise characterization of the characteristic peaks in the energy deposition spectrum—as a function of the incident neutron energy—is essential. This is because the counts in the energy deposition spectrum are normalized to their maximum value before being used as input for the ML model. In this experimental campaign, it was observed that 10^4 acquisitions per source (corresponding to approximately 90 minutes for the 14.1 MeV source

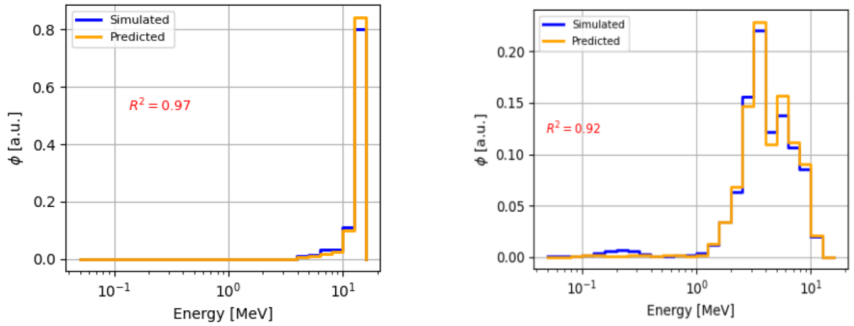
and six hours for the AmBe source) were sufficient to reproduce the characteristic energy deposition peaks, shown in Figure 8. These experimental results are consistent with values reported in the literature. In particular, Figure 8a shows a peak at 8.3 MeV, typically used to assess the resolution of this sensor, arising from the $^{12}C(n,\alpha)^{9}Be$ reaction in which beryllium is produced in its ground state. The bump observed between 4 and 7 MeV is attributed to the production of three α particles, while the higher peak at lower energy is primarily due to scattering. For the AmBe source (Figure 8b), the emitted neutrons have lower energies and do not activate all threshold reactions with carbon atoms; hence, the main peak is predominantly due to scattering reactions, with a smaller contribution from alpha particle production, in agreement with the simulations.



(a) (b)

Fig. 8. Deposited energy in CVD diamond detector from 14.1 MeV (a) and AmBe neurtron source (b).

Normalized deposited energy counts were used as the input for the aforementioned Median model, with the output being the neutron spectra shown in Figure 9. The predicted spectra were compared with reference spectra obtained from PHITS simulations of the DANAIDES irradiation casemate. The comparison reveals excellent agreement, with R^2 values exceeding 0.9 for both spectra. For these experiments, the energy grid discretization adopted for the training and test databases was not optimal, as the selected neutron sources have a much narrower energy range than the grid used. Consequently, adjusting the energy binning according to the specific source to which the detector is exposed could improve the energy discretization, while still respecting the intrinsic energy resolution of CVD diamond sensor [31]. Nonetheless, these results demonstrate a proof of concept regarding the feasibility of this neutron spectrum reconstruction methodology.



(a) (b)

Fig. 9. Predicted and simulated neutron energy spectra of 14.1MeV (a) and AmBe source (b).

6. Conclusion

In this study, a comparative machine learning framework was developed to reconstruct neutron energy spectra from energy deposition data acquired in a CVD diamond detector. The input data were generated through detailed Monte Carlo simulations using the PHITS transport code, modelling the deposited energy in the detector to a set of 544 synthetic neutron spectra and 120 real spectra taken from the IAEA Compendium database. Four distinct feature extraction methods were investigated to reduce the high dimensionality of the raw deposited energy data (600 bins) prior to regression: Fourier Transform, Nodal Basis Projection, Wavelet Decomposition, and Agglomerative Clustering approach. These transformations were combined with two regression models: Linear Regression was used for the Fourier and Nodal projections due to their compact and smooth representation of the signal, while Kernel Ridge Regression was applied to the Wavelet and Agglomerative features to better capture complex nonlinear relationships in the data. The performance of each model was evaluated using the coefficient of determination R^2, revealing that all approaches consistently achieved R^2 scores above 0.7 across the test set. The Kernel Ridge model combined with Agglomerative

Clustering yielded the highest average R^2, with 88% of the spectra reconstructed above the 0.9 threshold. Nevertheless, specific limitations were observed: the Fourier-based model struggled with non-symmetric spectra due to its periodicity assumptions, while the Kernel Ridge model exhibited reduced accuracy in reconstructing spectra with three-peaked ("3-bump") structures, which were underrepresented in the training data. To enhance prediction robustness and generalization, a median-based ensemble model was introduced by computing the bin-wise median across the four individual model outputs. This meta-model showed improved overall R^2 performance ($\overline{R^2}$ = 0.953) and mitigated individual model shortcomings, such as boundary artefacts and localized overfitting. However, the median approach does not always outperform the best individual model on specific spectra, particularly when one model dominates in accuracy for a certain class of inputs. The proposed methodology was further tested through an experimental campaign conducted at the TOTEM facility of CEA Cadarache, where the CVD diamond detector was irradiated with 14.1 MeV and AmBe neutron sources. The reconstructed spectra showed good agreement with reference spectra generated via PHITS simulations under the same experimental conditions, demonstrating the feasibility of applying the method in real-world scenarios. This study lays a proof of concept for applying machine learning to neutron spectroscopy using solid-state detectors. It offers an alternative to classical unfolding methods, avoiding common pitfalls such as instability in matrix inversion and artefacts due to over-regularization. Future developments may include the expansion of the training dataset to cover a wider variety of spectral shapes, refinement of the energy grid discretization, and the integration of advanced learning models such as neural networks or probabilistic method like Gaussian processes, as well as the introduction of the uncertainty quantification. An additional key perspective concerns the incorporation of the total neutron fluence in the reconstruction pipeline: currently, spectra are normalized to unit area, but estimating the absolute fluence would enable a physically meaningful reweighting of the reconstructed spectrum, providing not only spectral shape but also absolute flux values. Moreover, extending the methodology to other detector materials and geometries could broaden its applicability across a wider range of neutron energy regimes.

Acknowledgments

The authors would like to acknowledge the organizers and participants of the SEME (Semaines d'Études Mathématiques et Entreprises) event for their valuable contributions and insightful discussions. Special thanks to AMIES (Agence pour les Mathématiques en Interaction avec l'Entreprise et la Société) and Archimedes Institute from Aix-Marseille University for facilitating this initiative.

References

[1] Q. Huang, Z. Chen, X. Xu, and J. Zhang, "A review of the application of artificial intelligence to nuclear reactors: Where we are and what's next," Heliyon, vol. 9, no. 3, p. e14273, 2023. DOI: 10.1016/j.heliyon.2023.e14273

[2] A. Lyoussi et al., "I_SMART a collaborative project on innovative sensor for material ageing and radiation testing," KIC InnoEnergy, Grenoble, France, 2012. [Online]. Available: https://www.kic-innoenergy.com/projects/i_smart

[3] M. Birk, G. Goldring, and P. Hillman, "Fast neutron spectroscopy with solid state detectors," Nucl. Instrum. Methods, vol. 21, pp. 107–113, 1963. DOI: 10.1016/0029-554X(63)90061-6

[4] J. Gómez-Ros, "Solid state detectors for neutron radiation monitoring in fusion facilities," Radiation Measurements, vol. 71, pp. 57–63, 2014. DOI: 10.1016/j.radmeas.2014.04.006

[5] E. Belfiore et al., "SiC Detector Thickness Optimization for Enhanced Variability Response," in SNA + MC 2024 Conf. Proc., Paris, 2024, pp. 45–50. (in press)

[6] E. Belfiore et al., "Proof of Concept of the Prediction of Neutron Spectra using Solid Detectors via Machine Learning Approach," in IEEE NSS MIC RTSD Conf., Tampa, 2024. DOI: 10.1109/NSSMICRTSD49039.2024.10001

[7] R. Antoni, P.-G. Allinei, and L. Bourgois, "Prediction of fast neutron spectra with a spherical TEPC using a machine-learning algorithm," Nucl. Instrum. Methods Phys. Res. A, vol. 1050, p. 168139, 2023. DOI: 10.1016/j.nima.2023.168139

[8] E. Belfiore et al., "Theoretical analysis of neutron spectra measurement with SiC detectors using a machine learning technique," J. Instrum., vol. 19, no. 1, p. P01007, 2024. DOI: 10.1088/1748-0221/19/01/P01007

[9] Y. Liu et al., "Simultaneous measurement of energy spectrum and fluence of neutrons using a diamond detector," Sci. Rep., vol. 12, no. 1, p. 15325, 2022. DOI: 10.1038/s41598-022-19711-7

[10] E. Griesmayer et al., "Neutron cross section measurements with diamond detectors," EPJ Web Conf., vol. 225, p. 05005, 2019. DOI: 10.1051/epjconf/201922505005

[11] R. Slavickas et al., "Applications of carbon-based diamond detectors: A critical review," Sensors, vol. 22, no. 15, p. 5736, 2022. DOI: 10.3390/s22155736

[12] O. Obraztsova et al., "Comparing the response of a SiC and a sCVD diamond detectors to 14-MeV neutron radiation," IEEE Trans. Nucl. Sci., vol. 65, no. 3, pp. 578–585, 2018. DOI: 10.1109/TNS.2018.2800440

[13] C. Fabjan, "Collider detectors for multi-TeV particles," in Encyclopedia of Physical Science and Technology, 3rd ed., vol. 3, pp. 253–268, 2003.

[14] C. E. Nebel, "CVD diamond: a review on options and reality," Materials, vol. 14, no. 22, p. 7081, 2021. DOI: 10.3390/ma14227081

[15] Z. Cheng et al., "High thermal conductivity in wafer scale cubic silicon carbide crystals," arXiv preprint, arXiv:2207.05292, 2022.

[16] Q. Wahab and M. Willander, “Silicon carbide and diamond for high temperature device applications,” Academia.edu, 2015. [Online]. Available: https://www.academia.edu/20207274
[17] J. Achard, V. Jacques, and A. Tallaire, “CVD diamond single crystals with NV centres: a review of material synthesis and technology for quantum sensing applications,” arXiv preprint, arXiv:1912.09749, 2019.
[18] Z. He et al., “Machine learning in nuclear physics at low and intermediate energies,” Front. Phys., vol. 11, p. 1152892, 2023. DOI: 10.3389/fphy.2023.1152892
[19] M. Kamuda et al., “A comparison of machine learning methods for automated gamma-ray spectroscopy,” J. Radioanal. Nucl. Chem., vol. 324, pp. 1115–1124, 2020. DOI: 10.1007/s10967-020-07125-3
[20] T. Sato et al., “Features of Particle and Heavy Ion Transport code System (PHITS) version 3.02,” J. Nucl. Sci. Technol., vol. 55, no. 6, pp. 684–690, 2018. DOI: 10.1080/00223131.2018.1432933
[21] L. Sobczak, “Dimensionnement d’un détecteur diamant pour la spectrométrie de neutrons rapides à partir d’une mesure micro-dosimétrique,” Internal Technical Report, CEA Cadarache, 2022. (Available upon request)
[22] J. Kelley et al., “Energy levels of light nuclei A = 12,” Nucl. Phys. A, vol. 968, pp. 71–142, 2017. DOI: 10.1016/j.nuclphysa.2017.04.002
[23] K. Shibata et al., “JENDL-4.0: A new library for nuclear science and engineering,” J. Nucl. Sci. Technol., vol. 48, no. 1, pp. 1–30, 2011. DOI: 10.1080/00223131.2011.555445
[24] R. V. Griffith et al., “Compendium of neutron spectra and detector responses for radiation protection purposes,” IAEA Technical Report Series No. 403, 1990.
[25] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011. DOI: 10.5555/1953048.207819
[26] C. K. Chui, An Introduction to Wavelets, San Diego: Academic Press, 1992, pp. 1–272. ISBN: 9780121745841
[27] P. Exterkate, “Model selection in kernel ridge regression,” Comput. Stat. Data Anal., vol. 68, pp. 1–16, 2013. DOI: 10.1016/j.csda.2013.03.005
[28] D. Britto, J. A. Silva, and E. Rocha, “What is R^2 all about?,” Leviathan – Cadernos de Pesquisa Política, vol. 3, no. 1, pp. 60–68, 2011.
[29] J. Roult dit Rouaux et al., “Experimental gamma coincidence spectra recorded in prompt gamma neutron activation analysis,” J. Radioanal. Nucl. Chem., vol. 333, pp. 6577–6592, 2024. DOI: 10.1007/s10967-024-09822-x
[30] Q. Potiron et al., “Modelling of a SiC based detector for the interpretation of 14.1 MeV neutrons measurements,” EPJ Web Conf., vol. 288, p. 04011, 2023. DOI: 10.1051/epjconf/202328804011
[31] C. Weiss and E. Griesmayer, “Fusion neutron diagnostics with CVD diamond detectors,” Fusion Eng. Des., vol. 203, p. 113605, 2024. DOI: 10.1016/j.fusengdes.2023.113605