Machine learning analysis of fission product yields

V. Tsioulos[1,*] and *V.* Prassa[1]

[1]Department of Physics, University of Thessaly, 3rd km Old National Road Lamia - Athens, Lamia, 35100, Fthiotida, Greece

Abstract. Analyzing fission product yields (FPY) is challenging because traditional models, while effective in certain conditions, have limitations in predictive accuracy and handling evolving fission modes. To overcome the limitations, especially in scenarios of limited data availability, machine learning models like gaussian process regression (GPR) and gaussian mixture model (GMM) are used for single-fission yield prediction and uncertainty quantification. The application of machine learning techniques demonstrates their practical utility in areas with constrained data, offering a novel approach for future computational advancements in nuclear physics. Our research aims to identify the most effective method for capturing the distribution of the dataset and extracting high-quality samples. These samples could serve as valuable inputs for more complex probabilistic neural networks like Mixture Density Networks (MDNs).

1 Introduction

Investigating the mechanisms of nuclear fission is highly beneficial for advancing various sectors of nuclear physics, including nuclear structure, nuclear reactions, and nuclear technology [1–3]. Nuclear fission data encompass detailed information on the properties and characteristics of nuclear fission reactions, and their analysis can contribute to a deeper understanding and accurate simulation of the nuclear fission. A key challenge in neutron-induced fission research is the limited availability of data, especially at various neutron energy levels, which necessitates innovative approaches to manage uncertainties. Predominant nuclear data libraries such as ENDF [4], JENDL [5], JEFF [6], and CENDL [7] primarily offer fission yield evaluations at thermal neutron energies around 0.5 MeV, but their coverage is less extensive for fast and 14 MeV high-energy neutrons. This creates a noticeable gap in data, particularly at other energies, limiting access to critical observables like the yields, total kinetic energy of fission fragments, and neutron multiplicity.

Statistics, data science, and machine learning (ML) constitute significant areas of research in modern science. They play an important role in learning how to make predictions from data, and extract key information about physical processes and the underlying scientific laws based on large datasets. ML, a subset of artificial intelligence, is now an extremely useful tool in nuclear physics, mainly due to its exceptional capability in handling large datasets and complex patterns [8]. ML has been primarily applied to nuclear fission data to evaluate fission yields without solely relying on experimental data. This approach involves constructing networks that utilize patterns derived from nuclear data libraries. Significant networks for these tasks include Bayesian Neural Networks (BNNs) [9] and Mixture Density Networks [10, 11]. The complexity of deep learning models poses a challenge due to the substantial data requirement for accurately capturing correlations and predictions. To address this issue, an alternative approach involves developing models that generate synthetic datasets derived from existing data. The method enables the evaluation and comparison of models based on their accuracy in reproducing fission yield distributions. The machine learning models focus on predicting FPY through targeted data analysis, specifically utilizing the Evaluated Nuclear Data LIbrary (JENDL) [12].

Our approach involves generating sample datasets from JENDL using two Gaussian-based networks: the Gaussian Mixture Model and Gaussian Process Regression. GMM's relative ease of use and simplicity in hyperparameter tuning, primarily involving the number of Gaussian components, makes it an effective tool for classifying nuclear data. On the other hand, GPR, despite its complexity and extensive hyperparameter tuning, offers the advantage of modeling uncertainties, making it particularly suitable for predicting single fission yields with high accuracy and reliability. In this work we employ both GMM and GPR to investigate the correlations within nuclear data, utilizing GPR for its predictive accuracy in uncertainties and GMM for its pattern recognition capabilities. Through this comparative analysis, the work aims to enhance the understanding and prediction of fission yields in Nuclear Data Libraries. In this contribution we briefly present our approach and results. Further details can be found in Ref. [11].

2 Machine learning models

2.1 Gaussian Mixture Model

Gaussian Mixture Model is a powerful clustering method based on unsupervised learning, which means that the

*e-mail: vtsioulos@uth.gr

dataset is unlabeled, and the algorithm tries to discern patterns and relationships within the data without any explicit guidance [13]. Unsupervised learning involves training a model using information that is neither classified nor labeled, allowing the algorithm to interpret and act on that information independently [14].

GMM assumes that data points are generated from a mixture of several Gaussian distributions with unknown parameters, where each Gaussian component represents a cluster, and the overall data distribution is modeled as a weighted sum of these Gaussian distributions [15]. Each cluster in GMM is represented by a Gaussian distribution characterized by its mean (μ), covariance (Σ), and weight (π), which indicates the proportion of the dataset that belongs to that cluster [16].

The mean (μ) indicates the center of the Gaussian distribution, the covariance (Σ) describes the shape and orientation of the distribution, and the weight (π) represents the probability of a data point belonging to a particular Gaussian component. GMM employs the Expectation-Maximization (EM) algorithm to estimate the parameters of the Gaussian components, iteratively improving parameter estimates to maximize the likelihood of the data given the model [17]. In the Expectation step (E-step), the algorithm calculates the probability of each data point belonging to each Gaussian component given the current parameter estimates. In the Maximization step (M-step), the algorithm updates the parameters (mean, covariance, and weight) to maximize the likelihood of the data given these probabilities.

Unlike K-means, which assigns each data point to a single cluster, GMM performs soft clustering, assigning each data point a probability of belonging to each cluster, allowing for more nuanced and flexible cluster assignments [18]. The probability density function of a mixture of K Gaussian distributions is given by

$$p(\mathbf{x}|\mathbf{\Theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)$$

where $\mathbf{x}$ is the data point, $\mathbf{\Theta}$ represents the parameters of the mixture model, including means $\{\boldsymbol{\mu}_k\}$, covariances $\{\mathbf{\Sigma}_k\}$, and mixing coefficients $\{\pi_k\}$. π_k is the weight of the k-th Gaussian component, satisfying $\sum_{k=1}^{K} \pi_k = 1$, and $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)$ denotes the Gaussian distribution with mean $\boldsymbol{\mu}_k$ and covariance $\mathbf{\Sigma}_k$ [13].

GMM can model clusters of different shapes, sizes, and densities, making it more versatile than methods like K-means, which assumes spherical clusters [19]. The ability to assign probabilities to cluster memberships allows GMM to handle uncertainty and overlapping clusters more effectively. GMM performs well on high-dimensional data, making it useful in complex applications such as image processing, bioinformatics, and speech recognition [13].

GMM can be used to identify anomalies or outliers in data by modeling the normal data distribution and detecting points that have low probabilities of belonging to any of the Gaussian components [18]. In image processing, GMM can be used to segment images into different regions based on color or texture [20]. GMM is commonly used in the modeling of acoustic features in speech recognition systems, where it helps in clustering phonetic patterns [16].

Despite its advantages, GMM is sensitive to initial parameter settings, and poor initialization can lead to suboptimal clustering results [15]. The EM algorithm can be computationally intensive, especially for large datasets or high-dimensional data. Choosing the right number of Gaussian components is crucial and can significantly impact clustering performance, often determined using criteria such as the Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC) [21].

2.2 Gaussian Process Regression

Gaussian Process Regression is a sophisticated probabilistic algorithm within the ML landscape, categorized under supervised learning. This approach is non-parametric and leverages kernel-based methods, aimed at regression tasks to predict continuous outcomes.

Unlike unsupervised learning, as used in GMM, GPR employs supervised learning. Through supervised learning, GPR models are trained to recognize and associate various patterns in the input data with their respective outputs, thereby gaining a refined understanding of the relationships within the data.

Central to GPR are kernel functions, which specify the covariance between pairs of random variables in the Gaussian process. The choice of kernel significantly influences the model's performance and its ability to generalize from training to unseen data by encoding assumptions about the underlying function. Different kernels capture diverse patterns and relationships in data.

In this analysis, the selection of kernels is deliberate, with each playing a crucial role in predicting the bimodal distribution of fission yields. We start with the simplest, the constant kernel, which assumes a uniform covariance across the input space, useful when combined with other kernels to model functions with varying means:

$$k(x, x') = \sigma^2$$

where σ^2 is the constant variance.

Among the kernels, the radial basis function (RBF) kernel, also known as the squared exponential kernel, stands out for its ability to capture smooth transitions and variations in data:

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2l^2}\right)$$

Here, l is the length-scale parameter.

The rational quadratic kernel extends the RBF by incorporating multiple length-scales, accommodating datasets with both local fluctuations and broader global variations:

$$k(x, x') = \left(1 + \frac{\|x - x'\|^2}{2\alpha l^2}\right)^{-\alpha}$$

where α controls the scale mixture and l is the length-scale.

Lastly, the ExpSineSquared kernel is adept at modeling periodic functions, making it suitable for datasets exhibiting consistent cyclic behavior:

$$k(x, x') = \exp\left(-\frac{2\sin^2\left(\frac{\pi|x-x'|}{p}\right)}{l^2}\right)$$

Here, l is the length-scale and p is the periodicity.

These kernels collectively enhance the model's capability to address the complexities of data dynamics, offering tailored solutions for various data patterns encountered in nuclear physics research.

By implementing GPR for prediction, uncertainty estimation, and sample generation, along with GMM for sample generation through its clustering technique, we effectively created a substantial pool of samples well-suited for training a complex neural network. The accuracy of these samples was assessed by determining the distribution that best aligns with experimental data through a comparative analysis. This comparison utilized Kernel Density Estimation (KDE), a non-parametric method known for its prowess in statistical analysis for estimating the probability density function of a random variable. KDE's utility in smoothing sample data provides a clear visualization of the underlying distribution, which is particularly beneficial when the exact form of the distribution is unknown or difficult to define. By applying KDE in our comparison, we obtained a deeper understanding of how well samples from different methods, such as GPR and GMM, represent the true data distribution, especially for datasets characterized by multi-modal distributions. This approach ensures a thorough and accurate evaluation of the sampling methods' effectiveness in capturing the essential characteristics of the dataset.

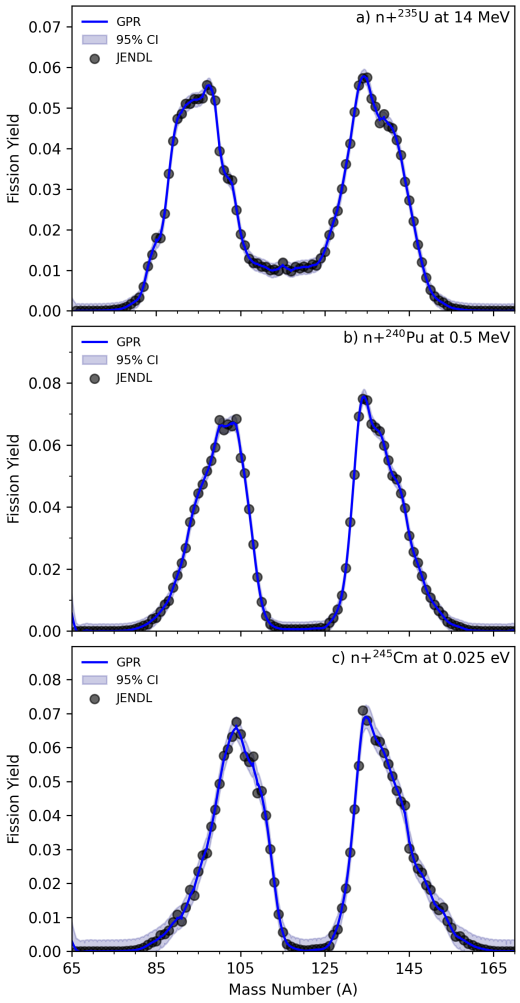

Figure 1. GPR mass yield distributions of induced fission. Experimental data are taken from the JENDL library. The shadowed area corresponds to the 95% confidence interval (CI)

3 Results

In Fig. 1 we demonstrate GPR's precision in managing complex datasets and its effectiveness in producing reliable augmented datasets for the neutron-induced fission mass yields of ^{235}U at 14 MeV, ^{240}Pu at 0.5 MeV, and ^{245}Cm at thermal neutron energies (see details in [11]). The GPR-generated samples exhibit high quality, effectively capturing the overall shape and specific features of the fission yield distributions. GPR's ability to maintain low uncertainties in critical regions, such as the peaks and valleys, is essential for accurate modeling and prediction of fission yields.

GMM and GPR's capabilities are evaluated by comparing their accuracy in reproducing fission yield distributions. Kernel Density Estimation is used for comparative analysis, ensuring thorough evaluation of the sampling methods. In Figs. 2-4 the density distributions of the augmented data of induced fission of ^{235}U at 14 MeV, ^{240}Pu at 0.5 MeV, and ^{245}Cm at thermal neutron energies calculated with the GPR and GMM methods are shown in comparison with the JENDL data. The results show that GPR clearly excelled in capturing the mass yield distributions. The various peaks represent the density of the data points,

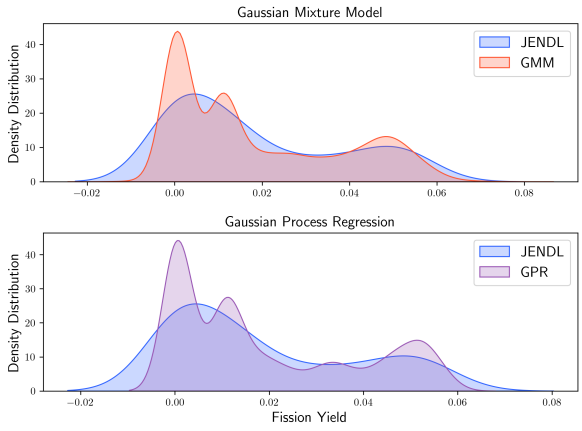

Figure 2. Density distribution of the augmented data of neutron induced fission of ^{235}U at 14 MeV, calculated with the GPR and GMM methods in comparison with the JENDL data

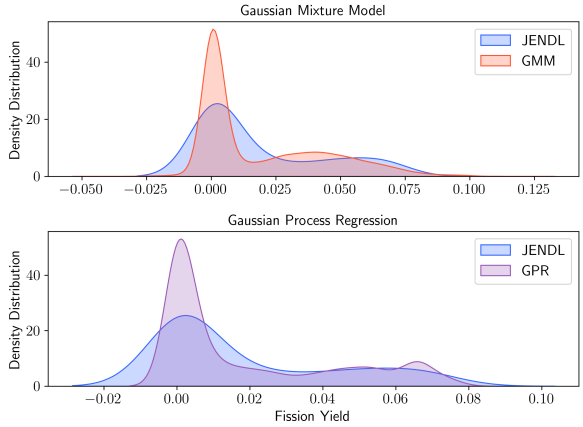

Figure 3. Same as Fig. 2 for ^{240}Pu at 0.5 MeV

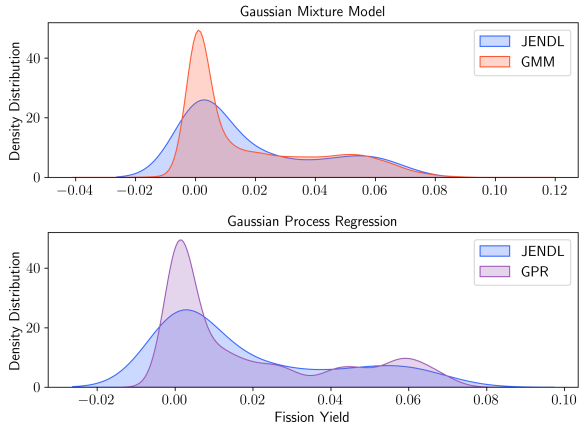

Figure 4. Same as Fig. 2 for ^{245}Cm at thermal neutron energies

with the highest density observed around zero. Despite minor deviations, the achieved precision is more than satisfactory, indicating a robust and reliable dataset. GPR's effectiveness in modeling isotopes underscores its value as a tool for nuclear physics research, capable of generating high-quality datasets for advanced analysis.

4 Summary

This research uses machine learning models, specifically Gaussian Process Regression and Gaussian Mixture Model, to predict fission product yields and manage uncertainties in nuclear physics. These models help overcome limitations of traditional methods, particularly with limited data. GPR excels in maintaining low uncertainties in critical regions, while GMM is effective for pattern recognition. Both models enhance the accuracy and reliability of nuclear fission yield predictions, utilizing data from the Japanese Evaluated Nuclear Data Library.

References

[1] N. Schunck *et al.*, *Reports on Progress in Physics* **79**, 116301 (2016)
[2] K.-H. Schmidt and B. Jurado, Reports on Progress in Physics **81**, 106301 (2018).
[3] M. Bender *et al.*, *Journal of Physics G: Nuclear and Particle Physics* **47**, 113002 (2020)
[4] M.B. Chadwick *et al.*, *Nuclear Data Sheets* **112**, 2887–2996 (2011)
[5] K. Shibata *et al.*, *Journal of Nuclear Science and Technology* **48**, 1 (2011)
[6] Nuclear Energy Agency, *JEFF Nuclear Data Library*, n.d.
[7] Z.G. Ge *et al.*, Journal of the Korean Physical Society **59**, 1052-1056 (2011).
[8] A. Boehnlein *et al.*, *Reviews of Modern Physics* , 94, 031003 (2021)
[9] Z. A. Wang, *et al.*, *Phys. Rev. Lett.* **123**, 122501 (2019)
[10] A. Lovell *et al.*, *arXiv* **abs/2005.03198** (2020)
[11] V. Tsioulos, and V. Prassa, Eur. Phys. J. A (under review) (2024)
[12] A. Dave *et al.*, *arXiv* **abs/2105.14645** (2021)
[13] Bishop, Christopher M., *Pattern Recognition and Machine Learning*, Springer (2006)
[14] G. James *et al.*, *An Introduction to Statistical Learning: With Applications in R* (2013)
[15] G.J. McLachlan *et al.*, *Finite Mixture Models* (2004)
[16] D.A. Reynolds, Douglas A. *Gaussian Mixture Models*, Encyclopedia of Biometrics, , pp. 659–663, Springer (2009)
[17] A.P. Dempster *et al.*, *Maximum Likelihood from Incomplete Data via the EM Algorithm*, Journal of the Royal Statistical Society: Series B (Methodological), 39(1), pp. 1–22, Wiley Online Library (1977)
[18] J.H. Friedman *et al.*, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer (2001)
[19] K.P. Murphy, Kevin P., *Machine Learning: A Probabilistic Perspective*, MIT Press (2012)
[20] Y. Zhong *et al.*, *Image Segmentation Using Gaussian Mixture Models and EM Algorithm*, Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing, 6, pp. 2327–2330, IEEE (2000)
[21] G. Schwarz, *Estimating the Dimension of a Model*, Annals of Statistics, 6(2), pp. 461–464 (1978)