**Proceedings of the INMM 61st Annual Meeting**
**July 12-16, 2020, An International Virtual Experience**

# OPTIMIZATION OF A K-NEAREST NEIGHBORS REGRESSION ALGORITHM FOR IMPROVED PULSE SHAPE DISCRIMINATION OF GAMMA RAYS AND NEUTRONS IN ORGANIC SCINTILLATORS

**Matthew Durbin**
Ken and Mary Alice Lindquist
Department of Nuclear Engineering
The Pennsylvania State University,
University Park, PA, USA

**Marc A. Wonders**
Ken and Mary Alice Lindquist
Department of Nuclear Engineering
The Pennsylvania State University,
University Park, PA, USA

**Marek Flaska**
Ken and Mary Alice Lindquist
Department of Nuclear Engineering
The Pennsylvania State University,
University Park, PA, USA

**Azaree T. Lintereur**
Ken and Mary Alice Lindquist
Department of Nuclear Engineering
The Pennsylvania State University,
University Park, PA, USA

## ABSTRACT

Certain organic scintillators, such as EJ-299 and stilbene, have the ability to accurately distinguish gamma rays and neutrons through the process of pulse shape discrimination (PSD). To improve the PSD performance of multiple organic scintillator-photomultiplier combinations, including two SiPMs and the aforementioned scintillators, a K-Nearest Neighbors (KNN) regression algorithm has been implemented on datasets of mixed gamma ray and neutron pulses. The KNN algorithm works by regressing on a conventionally calculated pulse shape parameter (PSP), leading to more effective discrimination of gamma rays and neutrons. Compared to current machine learning PSD methods, the KNN regression approach has the distinct advantage of being directly comparable to conventional PSD methods by way of a figure-of-merit. Further, it mitigates the bias introduced by pre-labeling pulses as gamma rays or neutrons during the training phases, which is present in classification-based machine-learning PSD. Previous work has shown that this approach improves PSD performance across the detection systems tested. This work aims to further optimize the KNN PSD algorithm at low light output ranges through an investigation of various input features, regression parameters, and feature scaling techniques. The input features and regression parameters are extracted from charge integrals, widths as a function of waveform maxima, and frequency components. The effects of normalization and standardization on input features are also investigated. These optimization efforts are performed on multiple detector-light sensor systems with varying levels of innate PSD performance. This optimization will allow for an algorithm that improves the traditionally achieved PSD performance and lowers the light output threshold at which radiation type can still be effectively distinguished.

# INTRODUCTION

Organic scintillators can be used to detect gamma rays and neutrons, and certain organic scintillators can discriminate between these two types of radiation, as well as heavy charged particles [1]. Examples include EJ-299 and stilbene, and when coupled with fast light sensors such as photomultiplier tubes (PMTs) or silicon photomultipliers (SiPMs), scintillation light can be converted into electrical pulses with temporal information indicative of the detected radiation type. While PMTs have historically been the primary light sensors used, in recent years SiPMs have increasingly been integrated in various detection systems. Their small size, lower voltage requirements, and immunity to magnetic fields, has made them beneficial for use in various radiation detection applications spanning security, nuclear safeguards, and medical fields [2], [3].

Neutron interactions within the organic scintillator will lead to a larger fraction of delayed fluorescence compared to gamma ray interactions. Capturing these differences by way of pulse analysis is a process referred to as pulse shape discrimination (PSD). A common way to accomplish this is to determine a pulse shape parameter (PSP), for which a threshold value is established to discriminate different radiation types. With good PSD capabilities, and an appropriate PSP, a one-dimensional histogram of the PSP values yields a gaussian distribution for each radiation type. There are many PSPs and analysis methods used in the literature, a common example of which is the charge integration-based tail-to-total (TTT) ratio [4]. The TTT is the ratio of the decaying tail portion of the pulse to the total charge integral.

In addition to the traditional charge integration, frequency, and wavelet PSD analysis methods, there has also been interest in machine learning (ML) classification algorithms for PSD purposes [5] [6]. Though works such as these show promise in their radiation distinguishing capabilities, the quantification of these algorithm's performance is partially burdened by the need for classification scoring. Each pulse tested by the ML algorithm must be classified a priori to create a label for training and to determine the algorithm's success, which is thus associated with the classification accuracy of an additional, independent PSD method. Further, there has been minimal work focusing on SiPMs and comparing ML performance in various detector-light sensor combinations has not been emphasized. In prior work, the authors introduced a novel regression-based technique that eliminates dependence on classification accuracy [7]. A K-Nearest Neighbors (KNN) algorithm was trained to regress on a conventionally calculated PSP, with algorithm performance quantified as an improvement in the Figure of Merit (FOM). Preliminary results from this technique demonstrated improvements of 50-200% for the FOM in the 200-700 keVee light output range on a variety of detector-light sensor combinations.

This work aims to optimize the KNN regression approach on SiPM-based detection systems by investigating the effects of different input features, regression parameters, and feature scaling methods. Specifically, pulses are processed for both input features and regression parameters in terms of charge integration, width characterization as a function of maxima, and frequency analysis. The feature scaling techniques investigated include standardization, normalization, and a robust scalar – detailed in the following sections. The optimization efforts aim to improve the FOM of the KNN regression technique, specifically at lower light output ranges.

## METHODS

This work makes use of two SiPMs and two organic scintillators listed in Table 1, making four total detection system combinations.

**Table 1. Detection system components**

| SiPMs | Organic Scintillator |
|---|---|
| Hamamatsu S13360-6075 SiPM | 6x6x6 mm stilbene Crystal |
| SensL 60035 C-series X13 SiPM | 6x6x60 mm EJ-299-33 Plastic |

Measurements of $^{252}$Cf are taken as a mixed gamma ray and neutron source for PSD assessment, and $^{22}$Na is used as a gamma ray calibration source. Pulses are recorded using a CAEN DT5730 digitizer and processed offline, primarily using Python for analysis. Pileup pulses are discarded, and an upper bound on light output is placed at 1000 keVee. This gives approximately 200,000 pulses per detection system combination. KNN implementation and optimization, as well as input feature preprocessing, are done with the scikit-learn library [8].

The KNN regression algorithm works by interpolating the output value of a user specified number of nearest neighbors, K, in the input feature space [9]. Nearest neighbors refer to the training data points with the smallest Euclidean distance to the test point. To bound the choice of K in this work, a value was set to ensure that the algorithm only interpolates over representative features. For example, interpolation of high energy neutron pulses should not begin to interpolate over high energy gamma ray pulses. As a proxy to meet this condition, K is taken as the number of neutron data points above 900 keVee, as approximated by a conventionally calculated PSP threshold. The implemented KNN technique provides a "modified" PSP value for individual pulses based on a regression of conventionally calculated PSP values of similar training pulses.

PSD capabilities of a system are often characterized by the FOM, which parameterizes how well separated the gaussian distributions are in the PSP histogram. The FOM is calculated in Equation 1 [4].

$$FOM = \frac{c_n - c_g}{FWHM_n + FWHM_g} \tag{1}$$

The subscripts of *n* and *g* correspond to neutron and gamma rays, and *c* and *FWHM* are the gaussian centroid position and full width at half maxima, respectively. A three-sigma separation in centroids corresponds to a FOM of 1.27, the value at which the gaussians are considered statistically separated. The FOM is the performance metric used in this work, and is calculated for the KNN-regressed PSP values, as well as the conventional PSP values for reference.

Previous work showed that charge integrals can be used as input features for regressing on a conventionally calculated PSP [7], with ten equally sized charge integrals yielding the best results; further optimization of the algorithm is presented here, with a focus on lower light output ranges.

While the method benefits from the algorithm's ability to pull information effectively from more than just the conventional two component analysis (TTT), all charge integral based features have similar distributions, in which particle types cluster together at low light output. To illustrate this, Figure 1, shows two of these charge integrals plotted against each other for the Hamamatsu 6075-stilbene combination.

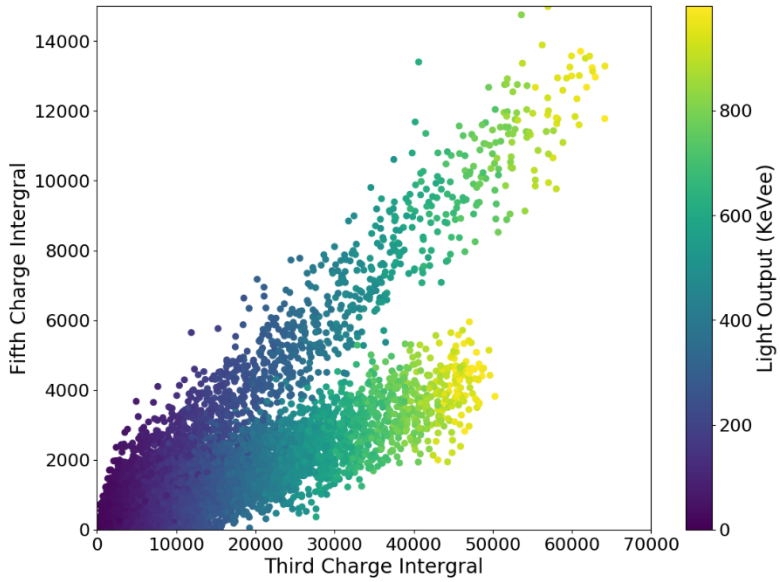


**Figure 1. Scatter plot of two charge integrals, with color indicating light output, as opposed to event density, for the Hamamatsu 6075-stilbene combination.**

Figure 1 shows the third and fifth equally spaced charge integrals, however this general shape is representative of many charge integral comparisons, such as plotting the tail integral against the total. The characteristic distribution of charge integral features will lead to algorithms, KNN or conventional, having good discrimination performance at high light output, but poor performance at low light output. Using the KNN scheme which includes input features, or feature scaling techniques, that lead to data with different distributions could produce information gain and increase discrimination performance at low light output ranges.

Multiple input features for each pulse are used to analyze the pulses in terms of three different domains: charge integrals (C), pulse width as percentages of the maxima (W), and frequency components (F). Charge features are taken as ten equally spaced charge integrals, width features are taken as the full-width at 10%-90% maxima (FW%M) in 10% increments, and amplitudes of the first five features of the real valued-Fourier transform of the entirety of each pulse are taken as the frequency inputs. In addition to testing C, W, and F features separately, combinations are explored denoted by letter combinations (Ex: CW). A visualization of this process is shown in Figure 2.

In addition to extracting features for use as algorithm inputs, a PSP is acquired for each of the three domains (C, W, and F). For the charge domain, it is the previously described TTT. For the width domain, the FW10%M divided by the total recording length (2400 ns) of the pulse is used. For the

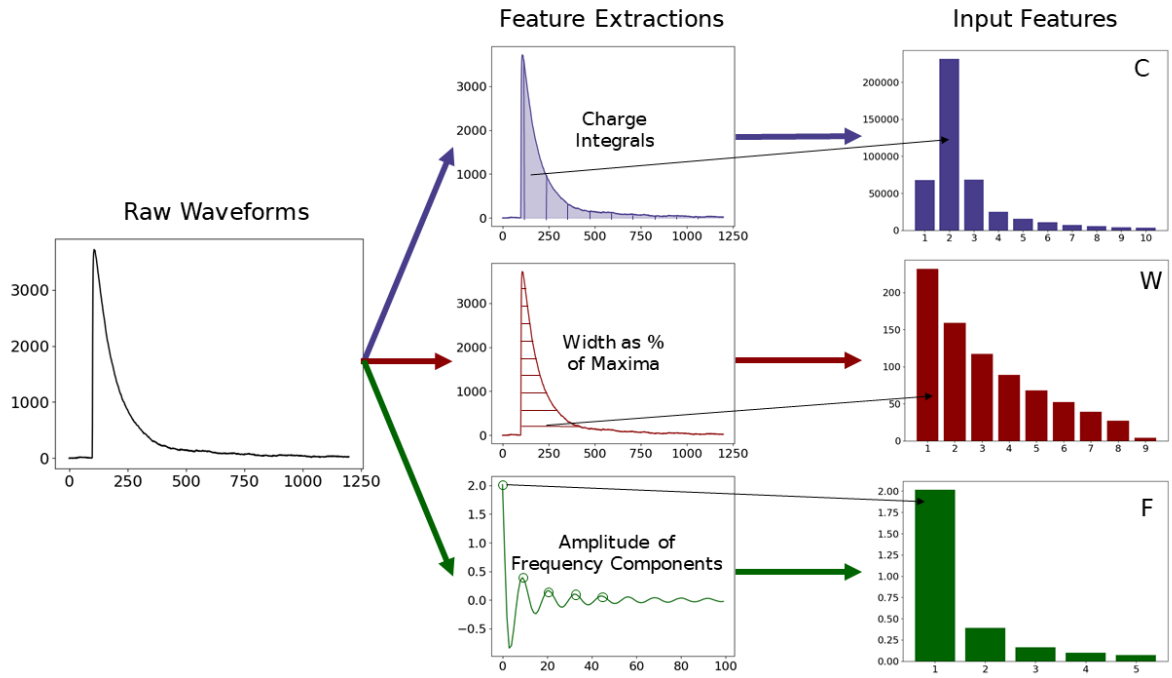


**Figure 2. Feature extraction visualization from raw waveforms. Colored arrows show general flow of the three domains (charge, width, frequency), black arrows give individual examples of extracted input features.**

frequency domain, the amplitude ratio of the first and second FFT real components is used. Each PSP has values between zero and one, yielding similar distributions, and is referred to as C-PSP, W-PSP, and F-PSP throughout this work.

Feature scaling addresses the wide-range of raw input feature values, and thus, gives different input features similar weighting in regression calculations. For the purposes of PSD, it may be possible that certain feature scaling techniques could allow lower energy pulses to be closer to more distinguishable higher energy pulses in the input feature space. In addition to testing the KNN with no feature scaling, a normalization, standardization, and robust scaler are implemented [8]. The normalization technique makes use of scikit-learn's MaxMinScaler, which for each input feature subtracts its minimum value and then divides by its total range. This technique allows the features to maintain their relative value, but on a scale from zero to one. The standardization makes use of scikit-learn's StandardScaler, which subtracts the mean from each feature and divides by the standard deviation. This technique has a mean of zero and is more resilient to outliers compared to the MaxMinScaler. The robust scaler makes use of scikit learn's RobustScaler, which removes the median and scales the data to the interquartile range, further mitigating the effects of outliers.

For each detection system combination, 20% of the total pulses are set aside as a final testing set. The remaining pules are used for training and validation. In the context of ML, a validation set is used to benchmark various optimization efforts or parameters during the training phase. To mitigate the effects or bias of any particular training-validation division, a k-fold cross-validation of five folds is implemented [8]. This technique shuffles and divides the data into five smaller sets, or folds. Each

fold is used as the validation set in turn, with the remaining folds combined to serve as the training set. Validation performance is then quantified in terms of a mean and standard deviation over the folds. Once the algorithm is deemed optimized, all folds are recombined as the training set, and the final performance is given for the testing set. Each combination of input features and feature scaling techniques are studied using this cross-validation method and compared against conventional analysis of the PSPs. Validation and testing results are presented as a function of 200 keVee light output ranges, referred to by their lower bound, to investigate performance as a function of light output. For example, the light output range referred to as "100 keVee" are the pulses with a light output between 100 and 300 keVee. Additionally, a FOM is given for the testing set of each detector system over the light output ranges with lower bounds of 0-500 keVee for the conventional PSP and KNN-regressed PSP values, to emphasize the challenges of lower light output.

## RESULTS

### Validation

Figure 3 shows the C-PSP mean FOM values and their standard deviation over the k-folds for various light output ranges for the Hamamatsu 6075-stilbene combination. Standard deviation is shown as shaded regions, barely visible on the presented scale.

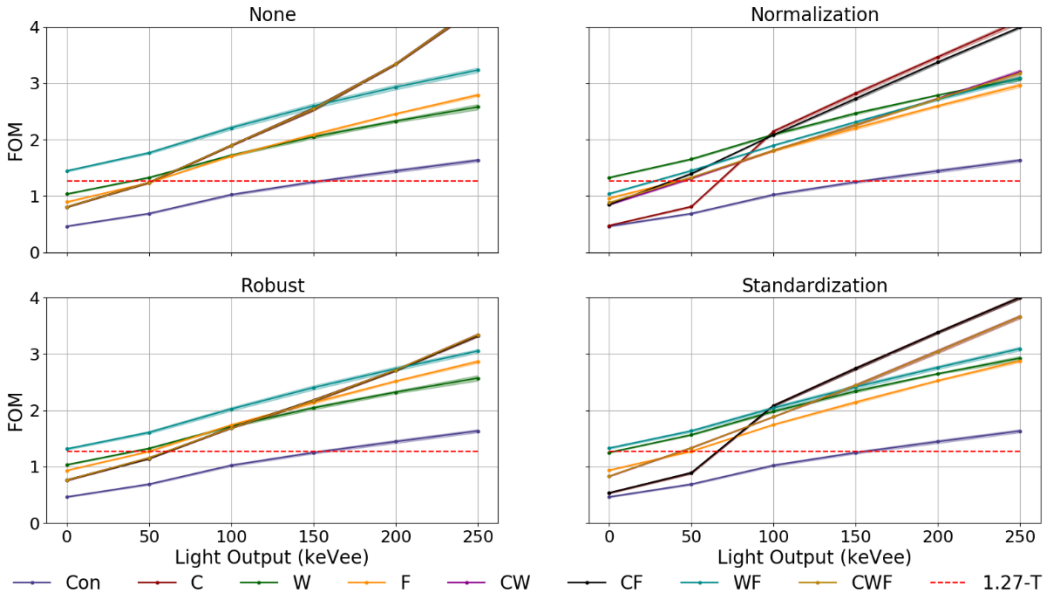


**Figure 3. C-PSP FOM values as a function of the lower bound of 200 keVee light output ranges for the Hamamatsu 6075-stilbene combination. For each feature scaling technique, a subplot shows the FOM values for each of the seven input feature sets, as well as the conventional (Con) values. The 1.27 FOM threshold is shown as a red dashed line.**

The conventional C-PSP FOM meets the 1.27 threshold around the 150 keVee light output range for this detector system, as can be seen in Figure 3. While above this light output the C input features largely perform the best, below this value the WF set of input features have the highest FOM. The success of the C input features at higher energies is not surprising, given the distribution shown in Figure 1. At lower energies, however, it is found that the data of this detection system clustered in the

WF feature space in a pattern more indicative of radiation type than in C feature space, resulting in an increased WF FOM. Similar results are found across detection systems, indicating that optimal input features for PSD performance changes as a function of light output.

Broadly speaking, the various feature scaling techniques do not result in improved performance. While these techniques do change the individual calculations the KNN performs, they do not fundamentally change the particle-cluster structure in the input feature space. An exception to this is that in some cases, the robust and standard scaler techniques aid the CF, CW, and CWF input feature sets. It these cases, feature scales vary by many orders of magnitudes, and scaling techniques allow for more equal weighting of the different feature types. The choice of a regression parameter may be better suited for specific analyses, but the benefits are found to be detection system specific. Notably, Hamamatsu 6075 combinations have better FOM values at low light output ranges with the C-PSP, where SensL 60035C combinations do better with the F-PSP. Table 2 shows the validation results and gives the best combination of input features, scaling techniques, and regression parameters for the four detection systems.

**Table 2. Optimization features from validation phase**

| Detection System | Input Feature Set | Regression Parameter | Feature Scaling |
|---|---|---|---|
| Hamamatsu 6075-stilbene | WF | C-PSP | None |
| Hamamatsu 6075-EJ-299 | CWF | C-PSP | Robust |
| SensL 60035C -stilbene | WF | F-PSP | None |
| SensL 60035C -EJ-299 | WF | F-PSP | None |

Testing

The optimal combinations from the validation phase, shown in Table 2, are used in the testing phase. For each combination, an overall FOM is calculated in the 0-500 keVee output range for the optimized KNN regression technique and an analysis of the conventionally calculated PSP values. Figure 4

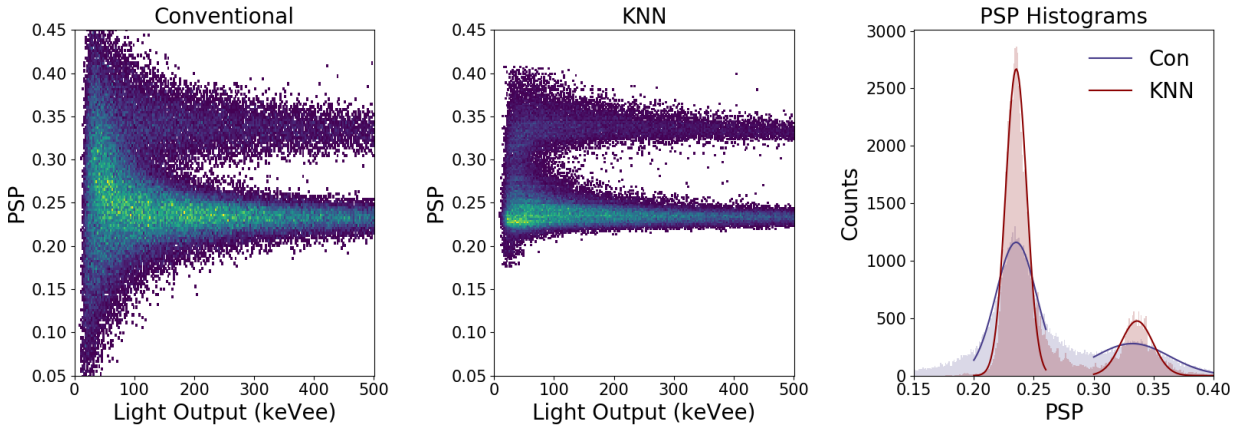


**Figure 4. Testing results for the Hamamatsu 6075-stilbene detection system. Left and center subplots show the PSD plots using the C-PSP for conventionally calculated and KNN-regressed PSP values respectively, and right subplot shows the one-dimensional PSP histogram for both.**

shows the two-dimensional PSD plot, which shows frequency of pulses binned against light output and PSP values, for the conventional and optimized KNN approach, as well as the one-dimensional histogram for both.

From the PSD plots of Figure 4, one can see a notable narrowing of particle features with the KNN technique. In the PSP histogram, the narrowing is also visible, as is the decrease in height of the "valley" feature between particle gaussians, indicative of better discrimination capabilities. Similar plots for the SensL 60035C-EJ-299 system are shown in Figure 5. These two systems are highlighted, as they correspond to the highest and lowest conventional FOM values.

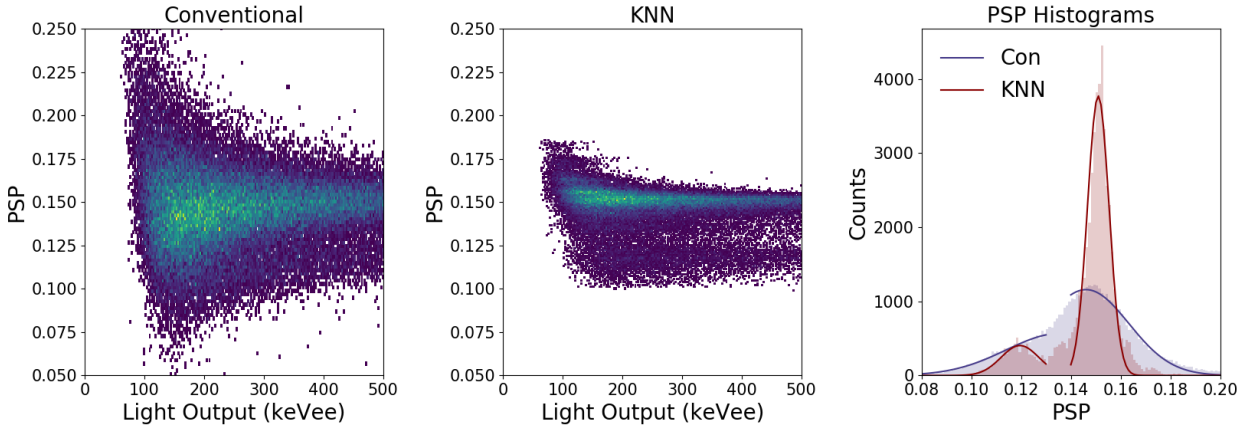


**Figure 5. Testing results for the SensL 60035C-EJ-299 detection system. Left and center subplots show the PSD plots using the F-PSP for conventionally calculated and KNN-regressed PSP values respectively, and right subplot shows the one-dimensional PSP histogram for both.**

While the innate PSD performance is notably lower for this combination, Figure 5 shows that the applied KNN regression technique can still visibly improve the particle discrimination capabilities. In the one-dimensional histogram, the conventional distribution has no distinct features while the KNN distribution reveals notable features. The results of the testing phase for all combinations are shown in Table 3.

**Table 3. Testing results (0-700 keVee light output range)**

| Detection System | FOM – Con | FOM - KNN |
|---|---|---|
| Hamamatsu 6075-stilbene | 0.86 | 1.94 |
| Hamamatsu 6075-EJ-299 | 0.47 | 0.82 |
| SensL 60035C -stilbene | 0.81 | 2.19 |
| SensL 60035C -EJ-299 | Ill defined | 1.12 |

The optimized KNN-regression scheme markedly improves the FOM values, even at this low light output range, for all systems. Stilbene combinations which do not meet the 1.27 FOM threshold conventionally, notably surpass the threshold with the KNN technique.

FOM values as a function of light output ranges of 200 keVee, denoted by their lower bound, are shown in Figure 6. Stilbene combinations meet the 1.27 FOM threshold at all light output ranges with

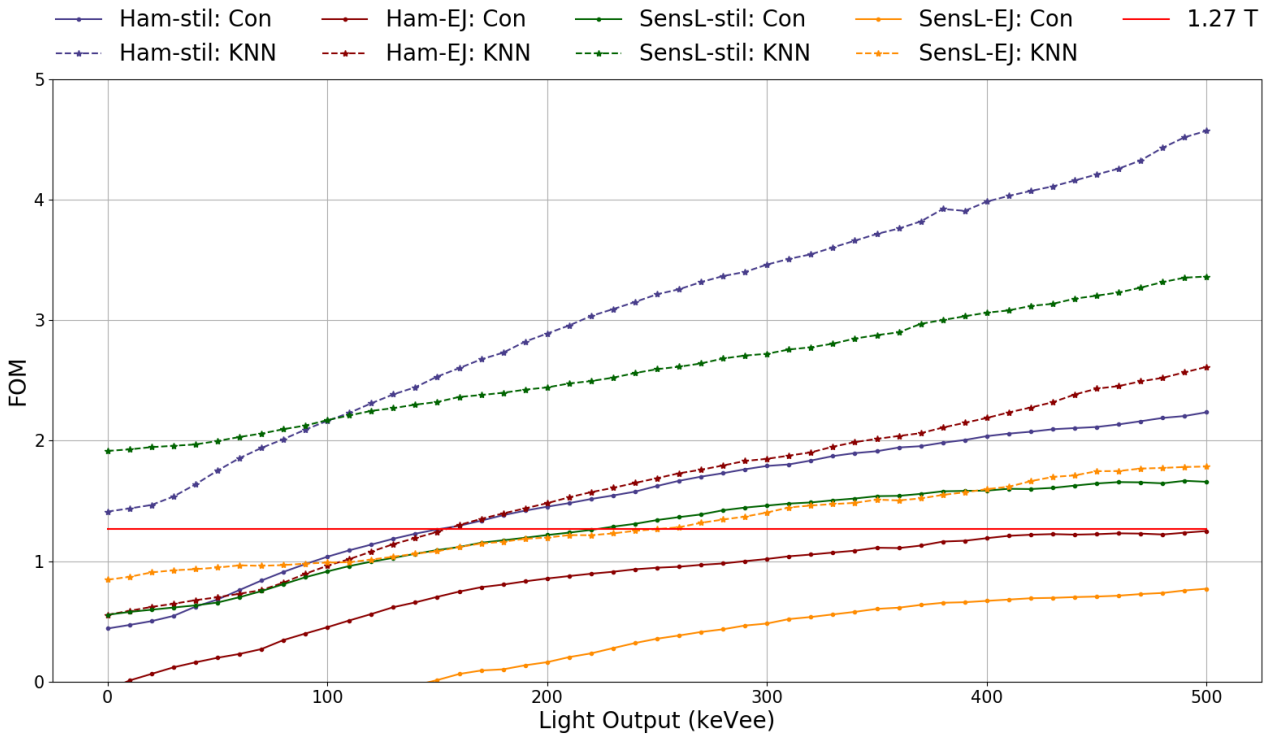


**Figure 6. FOM values of the testing datasets for all detection systems as a function of 200 keVee light output ranges in 10 keVee increments. The solid red line represents the 1.27 FOM threshold.**

the optimized KNN-regression technique, even as low as 0-200 keVee. The Hamamatsu 6075-EJ-299 and SensL 60035C -EJ-299 systems meet the threshold at approximately 170 and 240 keVee, respectively, with the optimized KNN-regression technique, while their conventional FOM's do not meet the threshold over the tested ranges. Figure 6 shows that the optimized KNN-regression technique notably increases the FOM values even at very low light output ranges, and that systems with traditionally lower innate PSD capabilities can have comparable performance with the KNN method to systems with higher conventional PSD performance.

## CONCLUSIONS

A KNN algorithm trained to regress on conventionally calculated PSPs can notably improve FOM values across different detection systems. To optimize this method at lower energies, the effects of input features, regression parameters, and feature scaling techniques are investigated.

While feature scaling techniques do not improve FOM values overall, it is shown that input feature sets that have individual feature scales that vary by many orders of magnitudes can benefit from their implementation, namely the standardization techniques. Various regression parameters are investigated, including ratios of charge integrals, width, and frequency features. While the different regression parameters affect the FOM values, it is determined that the optimal choice is detection system dependent.

FOM values, and thus PSD performance, drastically vary as a function of light output. It is shown that the optimal input features to use for a KNN algorithm depend on the light output range of interest. Notably, input feature sets that include frequency and width features are shown to perform better than other input feature sets that consist of just charge integrals. By extension, this implies that any ML type approach to PSD should be optimized with low light output ranges considered. With the optimal choice of investigated input features, regression parameters, and feature scaling techniques, it is shown the FOM can increase by over a factor of two for all detection systems at the light output range of 0-500 keVee with the KNN-regression technique. Additionally, the light output range at which the 1.27 FOM threshold of statistical separation is notably lowered across detection systems. Given that the KNN is computationally inexpensive and easy to retrain or reoptimize, even in live time, these improvements in PSD performance can lead to improved particle discrimination capabilities of use in a variety of applications.

## REFFERENCES


[1] G. E. Knoll, *Radiation Detectibn and Measurement*, Fourth. New York: John Wiley & Sons, Inc., 2010.

[2] P. Buzhan, B. Dolgoshein, L. Filatov, A. Ilyin, V. Kantzerov, and V. Kaplin, "Silicon photomultiplier and its possible applications," vol. 504, pp. 48–52, 2003, doi: 10.1016/S0168-9002(03)00749-6.

[3] M. A. Wonders, D. L. Chichester, and M. Flaska, "Assessment of Performance of New-Generation Silicon Photomultipliersfor Simultaneous Neutron and Gamma Ray Detection," *IEEE Trans. Nucl. Sci.*, vol. 65, no. 9, pp. 2554–2564, 2018, doi: 10.1109/TNS.2018.2829346.

[4] A. Lintereur, J. Ely, J. Stave, and B. Mcdonald, "Neutron and Gamma Ray Pulse Shape Discrimination with Polyvinyltoluene," 2012.

[5] T. S. Sanderson, C. D. Scott, M. Flaska, J. K. Polack, and S. A. Pozzi, "Machine learning for digital pulse shape discrimination," *IEEE Nucl. Sci. Symp. Conf. Rec.*, pp. 199–202, 2012, doi: 10.1109/NSSMIC.2012.6551092.

[6] C. Fu, A. Di Fulvio, S. D. Clarke, D. Wentzloff, S. A. Pozzi, and H. S. Kim, "Artificial neural network algorithms for pulse shape discrimination and recovery of piled-up pulses in organic scintillators," *Ann. Nucl. Energy*, vol. 120, pp. 410–421, 2018, doi: 10.1016/j.anucene.2018.05.054.

[7] M. Durbin, M. A. Wonders, A. T. Lintereur, and M. Flaska, "Application of a Novel Machine Learning Approach to SiPM-Based Neutron/Gamma Detection and Discrimination," in *2019 IEEE Nuclear Science Symposium and Medical Imaging Conference, NSS/MIC 2019*, 2019, doi: 10.1109/NSS/MIC42101.2019.9059952.

[8] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. 1, pp. 2825–2830, 2011, doi: 10.1007/s13398-014-0173-7.2.

[9] A. Geron, *Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems*. Sebastopol, CA: O'Reilly Media, Inc., 2017.