Center for Open Access in Science ▪ https://www.centerprode.com/ojit.html
Open Journal for Information Technology, 2022, *5*(1), 33-40.
ISSN (Online) 2620-0627 ▪ https://doi.org/10.32591/coas.ojit.0501.03033m

Time-Series Prediction of Gamma-Ray Counts Using XGB Algorithm

Vincent Mutuku

South Eastern Kenya University, School of Science and Computing, Kitui, KENYA

Joshua Mwema

South Eastern Kenya University, Kitui, KENYA
Directorate of Information and Communication Technology

Mutwiri Joseph

Tharaka University College, Science Engineering & Technology, Chuka, KENYA

Received: 11 March 2022 ▪ Revised: 13 June 2022 ▪ Accepted: 25 July 2022

Abstract

Radioactivity is spontaneous and thus not easy to predict when it will occur. The average number of decay events in a given interval can lead to accurate projection of the activity of a sample. The possibility of predicting the number of events that will occur in a given time using machine learning has been investigated. The prediction performance of the Extreme gradient boosted (XGB) regression algorithm was tested on gamma-ray counts for K-40, Pb-212 and Pb-214 photo peaks. The accuracy of the prediction over a six-minute duration was observed to improve at higher peak energies. The best performance was obtained at 1460keV photopeak energy of K-40 while the least is at 239keV peak energy of Pb-212. This could be attributed to higher number of data points at higher peak energies which are broad for NaITi detector hence the model had more features to learn from. High R-squared values in the order of 0.99 and 0.97 for K-40 and Pb-212 peaks respectively suggest model overfitting which is attributed to the small number of detector channels. Although radioactive events are spontaneous in nature and not easy to predict when they will occur, it has been established that the average number of counts during a given period of time can be modelled using the XGB algorithm. A similar study with a NaITi gamma detector of high channel numbers and modelling with other machine learning algorithms would be important to compare the findings of the current study.

Keywords: radioactivity, extreme gradient boost, regression, Gamma-rays, photo-peaks, NaITi.

1. Introduction

Radioactivity is the spontaneous emission of energy and particles from unstable atoms of radioactive material. Naturally occurring radionuclides; Potassium-40, Uranium-238 and Thorium-232 comprise terrestrial radiation (NRCC, 1999) and are used to quantify the radiological safety of a given material. Measurement of gamma rays from these radionuclides is mainly done with NaITi or HpGe gamma-ray spectrometry systems which differ in terms of energy resolution, detection efficiency and mechanism of detection (Hossain, Sharip & Viswanathan,

Correspondence: Vincent Mulwa, South Eastern Kenya University, School of Science and Computing, Kitui, KENYA. E-mail: mulwamutukuu@gmail.com.

2012). For a given range of channels in the detector representing a photopeak, the integral sum of the counts of each energy is a very important parameter for quantifying the radionuclide. Counts per second (cps), i.e., the intensity of the radionuclide is obtained by the ratio of the background-corrected integral sum of counts normally referred to as net area to the live time of the gamma-ray counting. The formation of a peak resulting from γ-ray emissions of a certain radionuclide in the sample is primarily a result of Compton scattering and photoelectric absorption from the incident and scattered photons (James & Christine, 2015). Highly-resolved photo peaks are obtained by longer measurement times by accumulating the radiation absorption and scattering events. However, this is determined by the activity of the sample wherein the intensity of the radionuclide is an indicator of the former, i.e., high-intensity samples imply high activity and take a shorter duration to form peaks. A review of some radiation surveys shows that different researchers use different sample run times, e.g., 27.7hrs (Sharma, Singh, Esakki & Tripath, 2016), 23.8hrs (Asaduzzaman, Mannan, Khandaker, Farook, Elkezza & Amin, 2015), 6.1hrs (Aslam, Gul, Ara & Hussain, 2012), and 5.5hrs (Viruthagiri, Rajamannan & S., 2013). While longer measurement times are recommended, treating all samples as low intensity may unnecessarily lengthen the data collection leading to delayed research output especially in developing countries where research equipment are few compared to a large number of researchers. Given that radioactivity is spontaneous, it's impossible to predict when the next unstable atom would decay and emit a gamma-ray. However, when γ-ray counting starts from the time

$$t = 0\ to\ t1, t2\ and\ t3\ all\ the\ way\ to\ t_n$$

an average of n_t counts are registered by the detector at the end of each duration t_n. Since the half-lives of the radionuclides are long enough (Ebbing & Wentworth, 1995; Connell & Pike, 2005), the activity of each radionuclide remains the same within the measurement time. Thus, the intensity (cps) of the radionuclides within a sample material is characteristic and probabilistic. Based on the scattering angle, the energy counts are registered at one of the three regions, Compton continuum, Compton edge or full energy peak. The capability of XGB to deduce the hidden patterns in the interaction events leading to a full energy peak were examined to predict the number of counts in a given time. Accurate prediction of the number of counts in a range of channels could result in a predicted spectrum which implies that shorter measurement times can be adopted to accurately predict counts over a longer time for rapid research and development.

2. Literature review

Studies related to the current study have been reviewed here to understand the current scope of applications of machine learning in nuclear studies.

Klaus and John (1995) described application of neural networks (NN) in predicting probabilities of nuclear stability and relaxation to ground state. In the study, a feedforward network was implemented where the inputs were nuclide parameters which include the proton and neutron numbers. The dynamics of NN weights were managed by a stochastic back-propagation algorithm coupled by entropy function. Whilst NNs were retrained severally using different architectures leading to different models which performed well, it is difficult to obtain a high-quality performance with a global model in regard to the existing nuclear theory.

Niu, Liang, Sun, Long and Niu (2019) investigated the prediction of nuclear β decay using neural networks. Although some physics theories underlying nuclear β such as Fermi theory of β-decay and dependencies of half-lives which include pairing correlations and decay energies were embedded into a Bayesian NN (BNN), other unclear physics were left for the BNN to learn. To the researchers, the high prediction accuracy achieved is very instrumental in simulations involving the r-process.

Empirical formulas in nuclear decays are normally used with less modification as they are conventionally established for computations (Saxena, Sharma & Prafulla, 2021). A study by Saxena, Sharma and Prafulla (2021) shows that inclusion of machine learning in understanding certain phenomena can help modify the existing formulars thereby improving the precision. The researchers showed that adding asymmetry components predicted the half-lives in α-decay with more precision than the empirical formulas. Machine learning methods used include; XGBoost, rando forest, decision trees and multilayer perceptron NN whose results excellently agreed with experimental decay modes. At the same time, S., Freitas and John (2019) predicted the systematics of α-decay of heavy and superheavy nuclei using artificial neural networks (ANN) by backpropagation algorithm with regularization. The investigation highlighted the strengths and limitations of applying machine learning in studying nuclear events beyond stability.

The two body-bound state of deuteron was studied with a single layer feed-forward NN (Keeble & Rios, 2020). The NN successfully represented the S and D state wave functions. Compared with solutions of diagonalization tools, the study's results show that a 6 hidden node NN can seamlessly represent the ground state wavefunction with binding energy that is 0.1% of the theoretical dimensions. It is postulated that this method can pave way for variational ANN to solve nuclear many body problems.

Most of the studies have investigated half-lives and nuclear stability landscape in α-decay using different machine learning methods. The neural networks are the most used in learning the complex concepts in nuclear physics. While α-decay are important, γrays are also very critical in terms of the health effects they cause as they are ionizing radiation. It is important to apply machine learning techniques to have a deeper understanding of the decay behaviour. The number of gamma decays in a given interval is predicted in this study using XGBoost algorithm.

3. Materials and methods

The investigation was implemented in two phases; experimental data collection and machine learning implementation on the spectrum samples as outline in figure 1. A γ-ray spectrometer system comprising; NaITi γ-ray detector, lead shield and a multichannel analyser software was used in gamma-ray counting and acquisition of the sample spectra. The soil samples were prepared according to Sharma, Singh, Esakki, and Tripath (2016), packed in airtight containers with an Aluminum foil reinforced lid, and stored for 30 days to achieve secular equilibrium (Aslam, Gul, Ara & Hussain, 2012).

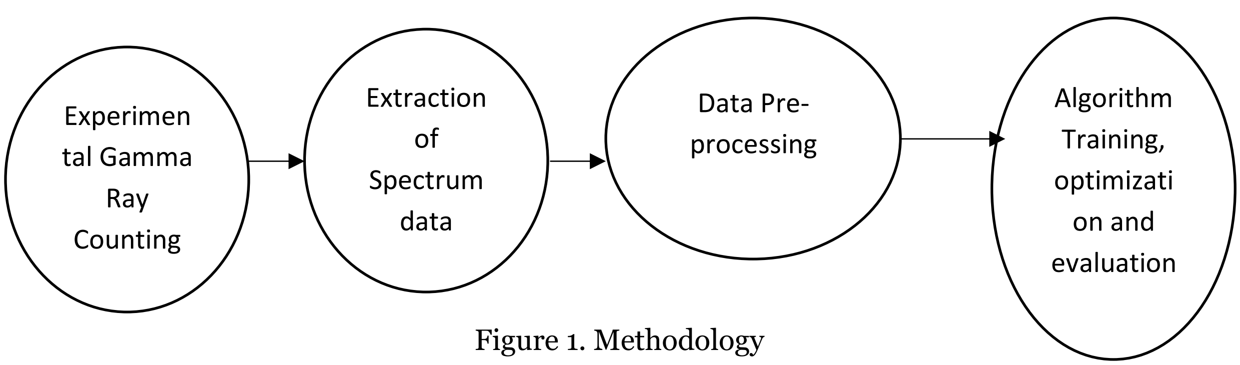

Figure 1. Methodology

Before γ-ray counting, the detector was energy-calibrated to obtain a channel-energy relationship that would help in identifying radionuclide full energy peaks in the spectrum. IAEA certified reference materials were used for both resolution and energy calibration. The sample was placed at the centre of the detector area and the shield covered with its lead lid. A total of 5 spectrums was obtained for measurement times; 4.5hrs, 6.5hrs, 7.5hrs and 7.6hrs. Since the detector has 1024 channels, each spectrum had 1024 instances of gamma-ray counts. As a way of data cleaning, full energy peaks' counts for radionuclides of interest were extracted from the spectrums, i.e., K-40 at 1460keV, Pb-214 at 352keV representing U-238 and Pb-212 at 239keV representing Th-232. Each column in the dataset represented counts of energies at different measurement durations and different channels. Thus, across the row were energy counts on the same channel for different durations. The last column in each dataset was set as the target in the Python program written to implement the XGB regression on the dataset. 80% and 20% of the dataset was used to train and test the model performance respectively. Further, XGB hyperparameter tuning was done to improve the performance of the model after each training. The R-squared value was the main metric to evaluate the model's prediction performance. The optimal model hyperparameters were set as;

Colsample_bytree:0.3

Learning rate: 0.1

Max_depth :5

Alpha:11

N_estimators=3000

The model flow chart is shown in figure 2.

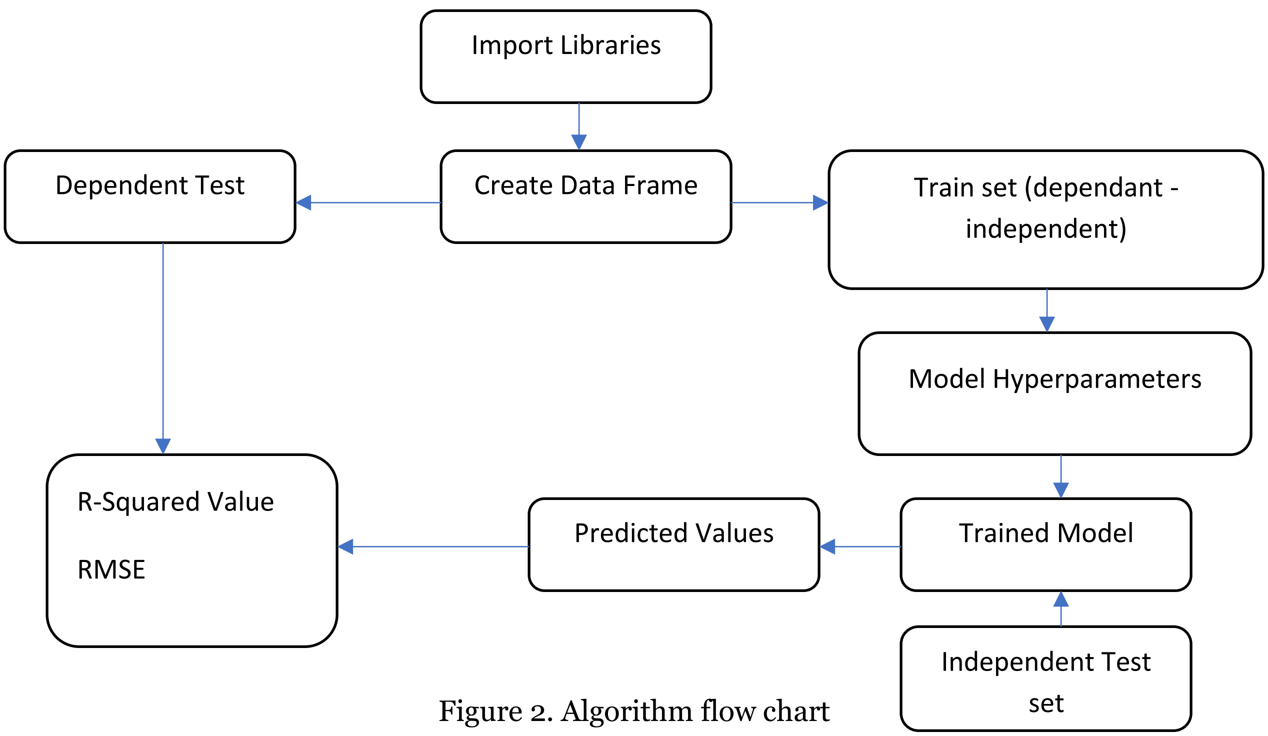

Figure 2. Algorithm flow chart

4. Results and discussion

Peak-wise prediction of the counts shows that the K-40 peak at 0.99 had a higher R-squared value while Pb-212 was the least with 0.92. Resultantly, there is a good agreement between the experimental and the predicted counts. With combined datasets for all the peaks, the R-square is equal to that of K-40. Generally, the R-square value increased with peak energy from 239keV to 1460keV. On the other hand, the RMSE value decreased with increasing peak energy from 59.62 at 239keV to 10.36 at 1460keV. The RMSE value for the combined dataset lies between that of K-40 and Pb-214 where a summary of the model performance is found in tables 1 and 2. The energy resolution, ER, of the detector, was determined according to equation 1 for three gamma-ray energies at their respective full width at half-maximum of the peak height.

$$ER = \frac{FWHM}{(PhotoPeak\ Energy)} \quad (1)$$

The energy resolution reduced from 8.43% at 239keV to 4.38% at 1460keV which compares with (Akkurt, Gunoglu & Arda, 2014) as shown in figures 3 and 4. Among the prediction errors, the greatest is 8.2% for the combined dataset while the least is 0.1% for the K-40 peak. Generally, the maximum and minimum error statistics for the three datasets exhibit a cyclical trend i.e., start slightly high at Pb-212, drop at Pb-214 and K-40, then increase for the combined dataset. This cyclical nature is similar to what is observed in the R-squared value as peak energies increase. On average, the model performed best at the K-40 peak with the lowest average error of 2%. The low energy resolution at 1460keV produced a broad peak for K-40 providing more instances given that larger datasets yield better performance (Althnian, AlSaeed, Al-Baity, Samha, Dris, Alzakari, Abou, Elwafa & Kurdi, 2021). Since the number of channels for the detector is relatively small, 1024, the resulting photopeak datasets were also small in size which explains the overfitting observed in the R-squared values.

Table 1. XGB Regression model performance

Radionuclide	**keV**	**R2**	**RSME**	**Resolution (%)**
Pb-212	239	0.97	59.62	8.43
Pb-214	352	0.92	29.59	5.60
K-40	1460	0.99	10.36	4.38
Datasets combined	na	0.99	24.95	na

Table 2. Prediction Errors

	Prediction Absolute Errors			
	Combined Datasets	**K-40 Peak**	**Pb-212 Peak**	**Pb-214 Peak**
MAX	8.2%	5.0%	5.5%	4.4%
MIN	0.3%	0.1%	1.6%	1.4%
Average	3.5	2.0	3.5	2.5

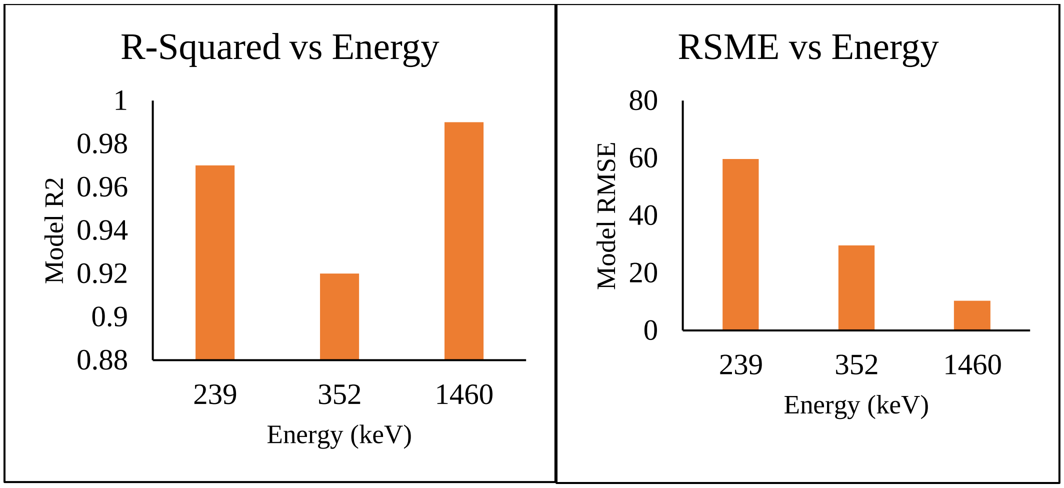

Figure 3. Prediction R-squared AND RMSE Values

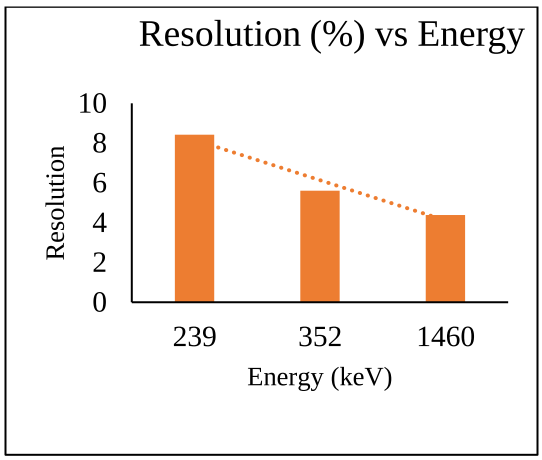

Figure 4. NaITi Energy Resolution (Present data)

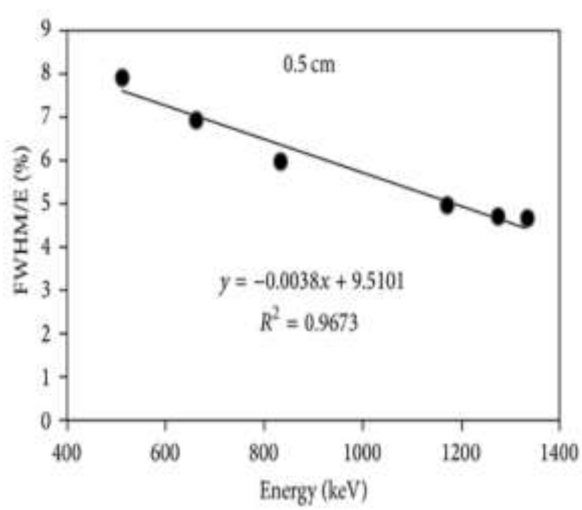

Figure 5. NaITi energy resolution (Akkurt, Gunoglu & Arda, 2014)

5. Conclusions

The prediction performance of the XGB regression algorithm has been evaluated based on experimental and predicted values. The model performs best at higher gamma-ray energies compared to lower ones. The algorithm exhibited excellent fitting capabilities for the gamma-ray counts for 6 mins. It would be important for another study to be done with a similar detector that has a larger number of channels offering larger datasets and investigating overfitting. Additionally, a similar study can be done using a hyper-purity germanium gamma ray spectrometry system to compare the performance of the model between the two systems. Also, other ML algorithms can be tested and their performances compared to the findings of this study. Further research into incorporating machine learning algorithms in scientific works may pave way for the development of more intelligent scientific research software. This may yield rapid research and development across many sectors.

Acknowledgements

This work has been supported by the South Eastern Kenya University, by providing the gamma-ray spectrometry system for the gamma-ray measurements.

The authors declare no competing interests.

References

Akkurt, I., Gunoglu, K., & Arda, S. (2014). Detection efficiency of NaI(Tl) Detector in 511–1332 keV energy range," *Science and Technology of Nuclear Installations,* vol. 2014, 2014.

Althnian, A., AlSaeed, D., Al-Baity, H., Samha, A., Dris, A., Alzakari, N., Abou, A., Elwafa, & Kurdi, H. (2021). Impact of Dataset Size on Classification Performance: An empirical evaluation in the medical domain. *Appl. Sci*., 11, p. 796.

Asaduzzaman, K., Mannan, F., Khandaker, M., Farook, M., Elkezza, A., & Amin, Y. (2015). Assessment of natural radioactivity levels and potential radiological risks of common building materials used in Bangladeshi dwellings. *PLOS ONE*, 10.

Aslam, M., Gul, R., Ara, T., & Hussain, M. (2012). Assessment of radiological hazards of naturally occurring radioactive materials in cement industry. *Radiation Protection Dosimetry*, *3*(151), 483-488.

Connell, C., & Pike, G. (2005). *Encyclopedia of Analytical Science* (Second Edition).

Ebbing, D., & Wentworth (1995). In *Experiments in Introductory Chemistry*, p. 221.

Hossain, I., Sharip, N., & Viswanathan, S. (2012). Efficiency and resolution of HPGe and NaI(Tl) detectors using gamma-ray spectroscopy. *Scientific Research and Essays*, *7*(1), 86-89.

Keeble, J., & Rios, A. (2020). Machine learning the deuteron. *Physics Letters B*.

Klaus, G., & John, C. (1995). Neural networks that learn to predict probabilities: Global models of nuclear stability and decay. *Neural Networks*, 8(2), 291-311.

James, P., & Christine, C. (2015). *Compton scattering of Cs-137 Gamma rays*. Department of Physics and Astronomy, The University of Tennessee, 401 Nielsen Physics Building, Knoxville.

Niu, M., Liang, Z., Sun, H., Long, W., & Niu, F. (2019). Predictions of nuclear β-decay half-lives with machine learning and their impact on r-process nucleosynthesis, *Physics Review C*.

NRCC (1999). *Natural radioactivity and radiation*. National Academies Press (US), Washington (DC).

S., P., Freitas, A., & John, C. (2019). Experiments in machine learning of alpha-decay half-lives. *Nuclear Theory*.

Saxena, G., Sharma, K., & Prafulla, S. (2021). Modified empirical formulas and machine learning for α-decay systematics. *Journal of Physics G: Nuclear and Particle Physics*, *48*(5).

Sharma, N., Singh, J., Esakki, C. A., & Tripath, R. (2016). A study of the natural radioactivity and radon exhalation rate in some cements used in India and its radiological significance. *Journal of Radiation Research and Applied Sciences*, *9*(1), 47-56.

Viruthagiri, G., Rajamannan, B., & S. J. K. (2013). Radioactivity and associated radiation hazards in ceramic raw materials and end products. *Radiation Protection Dosimetry*, *3*(157), 383-391.