Preparing Spectral Data for Machine Learning: A Study of Geological Classification from Aerial Surveys

Jun Woo Chung
Rochester Institute of Technology
jc4303@rit.edu

Alex Sim
Lawrence Berkeley National Laboratory
asim@lbl.gov

Brian J. Quiter
Lawrence Berkeley National Laboratory
bjquiter@lbl.gov

Yuxin Wu
Lawrence Berkeley National Laboratory
ywu3@lbl.gov

Weijie Zhao
Rochester Institute of Technology
wjzvcs@rit.edu

Kesheng Wu
Lawrence Berkeley National Laboratory
kwu@lbl.gov

Abstract

This work aims to identify and remedy distinctive challenges in the preparation of spectral data for machine learning. It does so by conducting a case study that involves matching an airborne gamma-ray spectral survey of the San Francisco Bay area to geological classifications provided by the United States Geological Survey. Our investigation has revealed three key approaches for enhancing accuracy in this task: 1) eliminating extraneous data segments unrelated to the main task, 2) augmenting minority classes to improve class balances, and 3) merging less pertinent classes. By incorporating these methods, we were able to achieve a significant increase in classification accuracy. Specifically, we increased the accuracy from an initial 40.8% to approximately 72.7%. We plan to continue our work to further enhance performance, with the goal of extending the applicability of these methods to other data types and tasks. One potential future application is prospecting for rare earth elements with aerial surveys.

1 Introduction

Spectral measurement devices are vital in scientific research, especially in remote sensing and environmental monitoring [2, 3, 20]. With the rise of machine learning (ML), there is an increasing push to optimize spectral data analysis using these techniques. Although ML has been used in various domains, including coal composition [6] and geological mapping [7], spectral data often require rigorous pre-processing before integration with ML.

This study explores the process of preparing gamma-ray spectral data for an ML model to classify the spectra based on a geological map provided by the United States Geological Survey (USGS) [12]. We delve into the nuances of the spectral data, detailing strategies to enhance the efficacy of the ML model. By employing these methods, the classification accuracy of our model increased from 40.8% to 72.7%. We believe that further enhancements could be achieved through the continued refinement of processing and learning techniques.

Problem statement. Our primary goal is to build a model, M, that takes physical spectral measurements, D_{gs}, and predicts classifications S_c, optimizing metrics such as classification accuracy and F1

Machine Learning and the Physical Sciences Workshop, NeurIPS 2023.

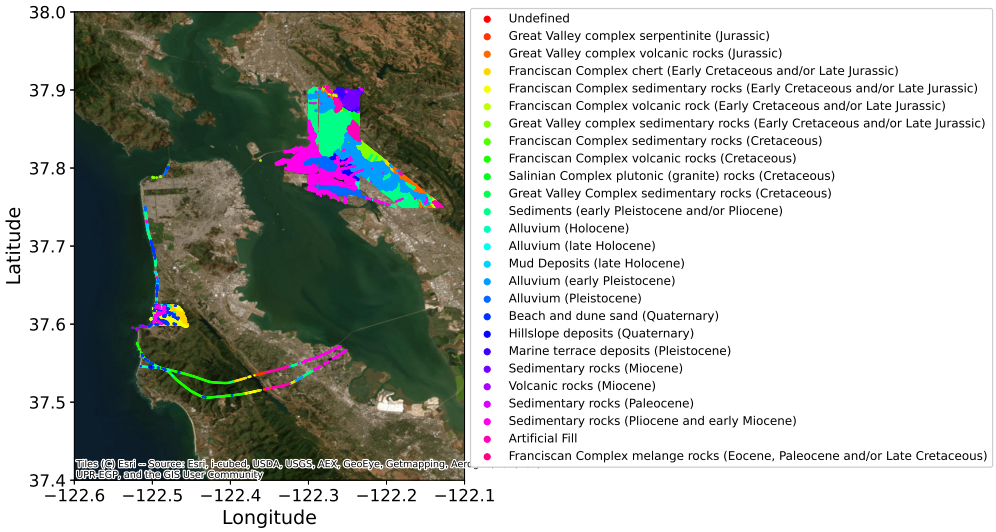

Figure 1: Measured data points and their USGS Geologic Map classifications of the San Francisco Bay area. Oakland (top right) predominantly displays alluvium and Artificial Fill. Measurements over water and the area below the Oakland patch are omitted due to a lack of classification information available w.r.t the USGS database.

score. Challenges stem from intricacies in D_{gs} and and limits of physical representativeness of S_c. Our current focus lies in data preprocessing, with future scope towards advanced learning techniques.

Challenges & approaches. We identify three major challenges: (1) Gamma-ray emission spectra are challenging due to weak ambiguous signals. Visually differentiating these spectra requires specialized knowledge. Furthermore, the discrete nature of detector counts causes low-frequency events to resemble noise. Many ML methods are not effective in handling such noisy data. (2) Classes within the data set are highly imbalanced, a common challenge in many real-world scientific applications [4, 15], which is further complicated by the existence of incoherent classes representing unknown mixtures of multiple classes. (3) The influence of environmental factors such as altitude and atmospheric changes in scientific measurements is unavoidable, and careful data selection is crucial to mitigate the impact of these environmental variations.

Contributions. Our work elucidates the challenges in preparing spectral data for ML, such as data noise and environmental influences. We propose a suite of techniques for spectral data refinement for ML, combining empirical and physical insights, through which we achieved a classification accuracy of 72.7% from an initial accuracy of 40.8%.

2 Methodology

To evaluate the application of machine learning techniques on spectral data, we construct a test problem to classify geological formations with gamma-ray measurements obtained from an aerial survey. This section details the data, the classification task, and its inherent complexities.

2.1 Classification problem

Our dataset encompasses an aerial survey of gamma-ray spectra (Figure 2) in the San Francisco Bay area (Figure 1), along with altitude, time and geolocation. These data were collected from August 27 to 31, 2014 by the Aerial Measuring System (AMS) for the process of developing the Airborne Radiological Enhanced Sensor System (ARES) [20], using a helicopter-mounted detector array using standard methodology [14] (which constitutes flying over the target area in a lawnmower pattern). Each data sample, which captures a one-second spectral measurement, spans 1022 bins of 3 KeV each, with an energy range up to 3 MeV. Each bin accounts for the gamma-ray events detected in that second within the specific energy range (Figure 2).

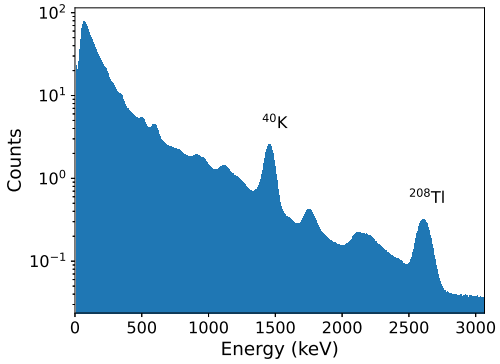

Figure 2: The average 1-s measured spectrum over the full dataset. The effect of gamma-ray down-scattering appears as an exponential decay as a function of energy. Two commonly observed peaks are labeled. The energy bin width is 3 KeV.

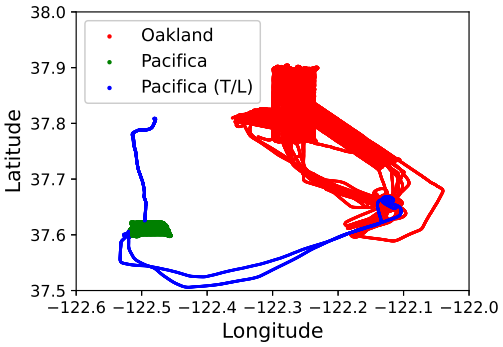

Figure 3: The divisions between Oakland and Pacifica data and between Pacifica and take-off/transit/landing data. The line emanating from the top of Pacifica is an exploratory detour.

We employ the USGS Geologic Map of the San Francisco Bay area, which details geological formations and rock types with the corresponding latitude and longitude coordinates, for classification. [12] However, during training and testing, we excluded these coordinates and time, relying only on the spectrum and altitude. Altitude was included as it has been shown to significantly effect detection of gamma-ray photons [17]. Altitude was kept at 200-300 feet above sea level during measurement as much as possible and no significant correlation between altitude and longitude/latitude/time were found, but unavoidable fluctuations in height are unavoidable due to wind and other factors. Data are randomly split 80-20% for training and testing.

Dataset issues and limitations. We describe the key challenges in our classification task, noting that many of these challenges also impact similar applications.

1. Pervasive skew in input data: the overall exponential decay shown in Figure 2 is present in all spectra, requiring careful feature engineering and selection of modeling techniques [3, 20].

2. Discrete/high-noise nature: The low event counts on the right side of Figure 2 mean that an individual spectrum has seemingly random 0 or 1 discrete gamma-ray events at higher energy, making them act similarly to noise in training and classification.

3. Class imbalance: As in many real-world applications, there is a very high imbalance of samples from different classes (some more than 10,000 records and others as few as 4) limiting the ability of many ML methods to recognize small classes.

4. Class incoherence: Some classes (e.g., "Artificial Fill"), especially deposition-formed non-bedrock ones, lack clear class boundaries and may overlap with others in characteristics.

5. Takeoff/Transit/Landing data: Data from the take-off, transit, and landing phases of the helicopter are misclassified more frequently compared to the main survey data.

2.2 Model and input data

We chose gradient boosting as our classification method for this exploration, driven by the following reasons; The limitations of dataset size (to a minimum of 3,000 with the application of our methods) make gradient boosting a more adaptable choice over deep neural networks due to its flexibility with diverse dataset sizes [24]. Second, the scale-invariant nature of gradient boosting efficiently manages the numerical complexity of gamma-ray spectral data. Its efficacy in geological classification is further supported by prior research [16, 27].

The gradient boosting algorithm used in this case study is implemented via the XGBoost Python library [8] with parameter settings of *learning_rate=0.02* , *max_depth=4*, *n_estimators=4000*, and *early_stopping_rounds=50*. Initial trials with other models, such as 1-dimensional CNN and transformer, showed inferior classification accuracy compared to gradient-boosting variants.

Pipeline	Selection	Pruning	SMOTE	cGAN	Combination	Accuracy (%)	F1 Score
1	X	X	X	X	X	40.8	0.360
2	O	X	X	X	X	54.9	0.500
3	O	O	X	X	X	63.2	0.550
4	O	O	O	X	X	63.4	0.597
5	O	O	X	O	X	65.3	0.627
6	O	O	O	X	O	72.7	0.718

Table 1: Partial summary of the tested configurations. Note that the test datasets are a randomly chosen selection of points within the full dataset, and thus pipeline 1 and 2 involve different test datasets from the others. That said, applying the same test dataset as Pipelines 3-6 (i.e. Pruned Pacifica) yielded accuracies of 42.9% for Pipeline 1 and 61.9% for Pipeline 2, still lower than the more complex pipelines.

3 Experimental evaluation

In our experimental evaluation, our objective is to measure the efficacy of our data preparation methods. Specifically, we are interested in understanding whether data selection and pruning can enhance classification accuracy, exploring strategies and the potential uplift from addressing class imbalance, and finding ways to minimize the negative impacts of incoherent classes on model quality.

Spatial data selection. While the initial gamma-ray spectrum data is comprised of measurements in both Pacifica and Oakland (Figure 3), we focused solely on Pacifica due to the prevalence of geologically inconsistent (and thus less meaningful) classes in Oakland, notably "Artificial Fill", for which there is no guarantee of any consistency in composition. Oakland consists predominantly of such classes, leading to potential spillover misclassifications in both areas. Given these challenges, we opted to exclude Oakland data. Focusing only on Pacifica, which is characterized by consistent bedrock, yielded a 14% boost in test accuracy compared to models trained on the entire dataset (as shown in Table 1, Pipeline 1 vs Pipeline 2).

Data pruning. Data from the gamma-ray detector take-off and landing phases (Figure 3), easily identifiable via timestamps and altitude, were prone to misclassification. Removing these segments led to an 8% rise in classification accuracy (from 54.9% to 63.2% - Table 1, Pipeline 2 vs Pipeline 3). This discrepancy may arise due to inconsistencies in measurement conditions during these phases. For instance, during transit, the helicopter often flew at altitudes around 1000 feet, contrasting with the typical 200-300 feet during the actual survey process.

Data augmentation. As detailed in Item 3 of Section 2.1, a large imbalance is present between classes for the entire data set and there are similar issues for the Pacifica data set alone. To address this, we explored two data augmentation techniques: SMOTE [4] and a cGAN-trained generator [18]. Applying SMOTE, which has seen success in geological machine learning [11, 16], gave a marginal accuracy boost from 63.2% to 63.4%, and more significant F1 score improvement by 0.046 (Figure 4).

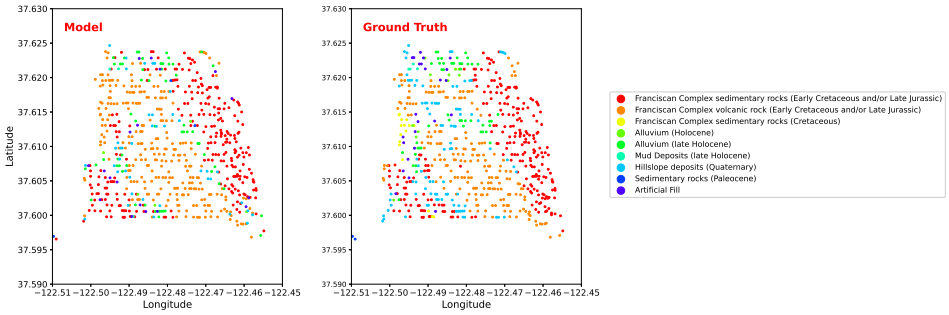

Figure 4: The prediction (left) and ground truth (right) of the gradient boosting model trained on Pacifica data with landing and takeoff data pruning, and SMOTE applied (Pipeline 4 in Table 1).

For the cGAN approach, the spectra data was log-transformed and scaled between (-1, 1), while the altitude data was directly scaled. The two were then merged. Using a 100-dimension cGAN, we achieved a 65.3% accuracy, surpassing the SMOTE results by 1.8%. We believe that there is potential

for further improvement through iterative experimentation. However, the computational resources and effort often required in training effective GANs may be prohibitive for some applications.

Incoherent class consolidation. The classification scheme provided by USGS contains several classes that cannot be expected to have a consistent composition (e.g. 'Alluvium', the deposition of which can involve mechanisms that transport material over great distances [19], or 'Artificial Fill', for which no meaningful assumption of composition can be made) The amalgamation of these geologically incoherent (and thus likely to be less geologically meaningful) classes in Pacifica into a single miscellaneous class (Table 1, Pipeline 6) in the preprocessing step improved model accuracy by about 8%, from 63.4% (Pipeline 4) to 72.7%. This is approximately 2. 3% higher than the accuracy anticipated if these classes were grouped post-prediction from a model trained on the ungrouped data (70.4%). This outcome underscores the utility of class amalgamation when such classes lack relevance to the central research question.

4 Related Work

Spectral data. The application of spectral data to machine learning has seen a heavy focus on optical spectra due to their relative ease of measurement. Machine learning methods based on hyperspectral imaging have been applied in numerous fields from land cover classification [1, 2], mineral prospectivity mapping [21], and biological analysis [10, 22], and x-ray spectra have seen applications in material characterization [5, 26]. Machine learning methods using gamma-ray spectra have been applied to radioisotope detection [13] and radioactive material detection [23, 25].

Geological applications. There has been a recent boom in geological applications of machine learning methods, particularly in the area of mineral prospectivity mapping. However, most current studies heavily involve geochemical data [16, 21, 28] which generally show a more direct correlation than spectral data. If spectral data is used, it is generally in the form of hyperspectral or multispectral satellite imaging data, in part due to the availability of large datasets such as Gaofen-2 [9].

5 Conclusions

In this paper, we describe the challenges in preparing spectral data for machine learning and demonstrate a number of techniques to address these challenges via a case study of classifying geological types based on an aerial survey of gamma-ray spectra. Our results emphasize the importance of data pruning. By omitting non-relevant segments such as the takeoff and landing phases, we significantly boosted model accuracy from 40.8% to 63.2%. This highlights the need for a thorough understanding of the physical context of the data as well as its collection process, and their potential implications and usefulness for model training.

To address class imbalance, we explore data augmentation techniques such as SMOTE and cGANs. Preliminary comparisons suggest that cGANs might offer superior results (65.3% versus 64.5% accuracy), and further exploration may be warranted in tailoring the GAN architecture to improve the augmentation quality. However, SMOTE offers a much simpler and more resource-efficient alternative. We also introduce a strategy to improve accuracy by handling geologically incoherent classes through class consolidation, highlighting the benefits of a context-aware approach in geological machine learning. Similar methods may be applied for other domains if certain classes are less meaningful for the application in question.

Future work will focus on hyperparameter tuning and model architectures to refine the framework for applications such as detecting and predicting rare earth element concentrations via aerial surveys.

Acknowledgements

This work was supported in part by the Office of Advanced Scientific Computing Research, Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231, and used resources of the National Energy Research Scientific Computing Center (NERSC).

References

[1] Muhammad Ahmad, Sidrah Shabbir, Swalpa Kumar Roy, Danfeng Hong, Xin Wu, Jing Yao, Adil Mehmood Khan, Manuel Mazzara, Salvatore Distefano, and Jocelyn Chanussot. Hyperspectral image classification - traditional to deep models: A survey for future prospects. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, 15:968–999, 2022.

[2] Nicolas Audebert, Bertrand Le Saux, and Sebastien Lefevre. Deep learning for classification of hyperspectral data: A comparative review. *IEEE Geoscience and Remote Sensing Magazine*, 7(2):159–173, 2019.

[3] Mark S. Bandstra, Brian J. Quiter, Marco Salathe, Kyle J. Bilton, Joseph C. Curtis, Steven Goldenberg, and Tenzing H. Y. Joshi. Correlations between panoramic imagery and gamma-ray background in an urban area. *IEEE Transactions on Nuclear Science*, 68(12):2818–2834, 2021.

[4] Kevin W. Bowyer, Nitesh V. Chawla, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *CoRR*, abs/1106.1813, 2011.

[5] Matthew R. Carbone, Mehmet Topsakal, Deyu Lu, and Shinjae Yoo. Machine-learning x-ray absorption spectra to quantitative accuracy. *Phys. Rev. Lett.*, 124:156401, 2020.

[6] Zeynep Ceylan and Bilal Sungur. Estimation of coal elemental composition from proximate analysis using machine learning techniques. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, 42(20):2576–2592, 2020.

[7] Snehamoy Chatterjee, Maria Mastalerz, Agnieszka Drobniak, and Cevat Ö. Karacan. Machine learning and data augmentation approach for identification of rare earth element potential in Indiana coals, USA. *International Journal of Coal Geology*, 259:104054, 2022.

[8] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

[9] Gong Cheng, Yuying Ban, Xiaoqing Deng, Huan Li, Hongrui Zhang, Guangqiang Li, Lingyi Liao, and Rehan Khan. Research on quantitative inversion of ion adsorption type rare earth ore based on convolutional neural network. *Frontiers in Earth Science*, 10, 2022.

[10] Viktor Dremin, Zbignevs Marcinkevics, Evgeny Zherebtsov, Alexey Popov, Andris Grabovskis, Hedviga Kronberga, Kristine Geldnere, Alexander Doronin, Igor V. Meglinski, and Alexander Bykov. Skin complications of diabetes mellitus revealed by polarized hyperspectral imaging and machine learning. *IEEE Trans. Medical Imaging*, 40(4):1207–1216, 2021.

[11] Mingjing Fan, Keyan Xiao, Li Sun, and Yang Xu. Metallogenic prediction based on geological-model driven and data-driven multisource information fusion: A case study of gold deposits in Xiong'ershan area, Henan Province, China. *Ore Geology Reviews*, 156:105390, 2023.

[12] Russell W. Graymer, Barry C. Moring, George J. Saucedo, Carl M. Wentworth, Earl E. Brabb, and Keith L. Knudsen. Geologic map of the San Francisco Bay region. *Scientific Investigations Map*, 2918, 2006.

[13] Mark Kamuda, Jacob Stinnett, and Clair Sullivan. Automated isotope identification algorithm using artificial neural networks. *IEEE Transactions on Nuclear Science*, pages 1–1, 2017.

[14] Remote Sensing Laboratory. An aerial radiological survey of the King and Pierce counties, WA-technical report. Technical report, 2011.

[15] Guillaume Lemaitre, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18:17:1–17:5, 2017.

[16] Timothy C.C. Lui, Daniel D. Gregory, Marek Anderson, Well-Shen Lee, and Sharon A. Cowling. Applying machine learning methods to predict geology using soil sample geochemistry. *Applied Computing and Geosciences*, 16:100094, 2022.

[17] Brian R. S. Minty. Fundamentals of airborne gamma-ray spectrometry. *AGSO Journal of Australian Geology and Geophysics*, 17:39–50, 1997.

[18] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014.

[19] Jonathan D. Phillips. The source of alluvium in large rivers of the lower coastal plain of north carolina. *CATENA*, 19(1):59–75, 1992.

[20] Brian J. Quiter, Tenzing H. Y. Joshi, Mark S. Bandstra, and Kai Vetter. CsI(Na) detector array characterization for ARES program. *IEEE Transactions on Nuclear Science*, 63(2):673–678, 2016.

[21] Victor Rodriguez-Galiano, Manuel S. Castillo, Mario Chica, and Mario C. Rivas. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*, 71:804–818, 2015.

[22] Xiao Shang and Laurie A. Chisholm. Classification of Australian native forest species using hyperspectral emote sensing and machine-learning classification algorithms. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, 7(6):2481–2489, 2014.

[23] Shiven Sharma, Colin Bellinger, Nathalie Japkowicz, Rodney Berg, and Kurt Ungar. Anomaly detection in gamma-ray spectra: A machine learning perspective. In *2012 IEEE Symposium on Computational Intelligence for Security and Defence Applications*, pages 1–8, 2012.

[24] Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6:60, 2019.

[25] Adam Varley, Andrew Tyler, Leslie Smith, Paul Dale, and Mike Davies. Mapping the spatial distribution and activity of 226Ra at legacy sites through machine learning interpretation of gamma-ray spectrometry data. *Science of The Total Environment*, 545-546:654–661, 2016.

[26] Shuting Xiang, Peipei Huang, Junying Li, Yang Liu, Nicholas Marcella, Prahlad K. Routh, Gonghu Li, and Anatoly I. Frenkel. Solving the structure of "single-atom" catalysts using machine learning – assisted XANES analysis. *Phys. Chem. Chem. Phys.*, 24:5116–5124, 2022.

[27] Jiangning Yin and Nan Li. Ensemble learning models with a bayesian optimization algorithm for mineral prospectivity mapping. *Ore Geology Reviews*, 145:104916, 2022.

[28] Chaojie Zheng, Feng Yuan, Xianrong Luo, Xiaohui Li, Panfeng Liu, Meilan Wen, Zesu Chen, and Stefano Albanese. Mineral prospectivity mapping based on support vector machine and random forest algorithm – a case study from Ashele copper–zinc deposit, Xinjiang, NW China. *Ore Geology Reviews*, 159:105567, 2023.