Journal of Radioanalytical and Nuclear Chemistry (2023) 332:3259–3271
https://doi.org/10.1007/s10967-023-09024-x

Classification and authentication of materials using prompt gamma ray activation analysis

Nathan A. Mahynski[1] · Jacob I. Monroe[1,2] · David A. Sheen[1] · Rick L. Paul[1] · H. Heather Chen-Mayer[1] · Vincent K. Shen[1]

Received: 7 April 2023 / Accepted: 21 June 2023 / Published online: 12 July 2023

Abstract
Prompt gamma ray activation analysis (PGAA) is a non-destructive nuclear measurement technique that quantifies isotopes present in a sample. Here, we use PGAA spectra to train different types of models to elucidate how discriminating these spectra are for various classes of materials. We trained discriminative models for closed set scenarios, where all possible material classes are known. We also trained class models to address open set conditions, where this enumeration is impossible. After appropriate pre-processing and data treatments, all such models performed nearly perfectly on our dataset, suggesting PGAA spectra may serve as powerful nuclear fingerprints for robust material classification.

Keywords Prompt gamma ray activation analysis · Machine learning · Class modeling · Material classification · Material authentication

Introduction

Prompt gamma ray activation analysis (PGAA) is a non-destructive nuclear measurement method employing gamma ray emission spectroscopy. The spectra generally cover a high-dimensional (multi-channel) energy space occupied by the isotopically characteristic gamma peaks of a material. Conventional spectral analysis of the location and intensity of a pre-determined energy peak can quantify the presence of an isotope in a material, providing a material signature. To obtain multi-elemental information, an expert must perform a peak-by-peak analysis, comparing against tabulated gamma ray emission energies and probabilities, which is a time-consuming and potentially error-prone task.

At the National Institute of Standards and Technology (NIST) Center for Neutron Research (NCNR), PGAA is performed on a nuclear reactor-based instrument, utilizing a high intensity and cold (low energy) neutron beam to irradiate a sample which emits gamma rays, intended for high-precision measurements with metrological quality. In industrial and field settings, as a non-destructive technique, PGAA has been widely used as an online monitoring tool in manufacturing, for example, in cement production [1]. It has also been shown to be capable of detecting explosive materials [2, 3]. There are many situations when such rapid approximate classification of a material is necessary without any prior knowledge about the nature of the material.

In this work, we use this spectrum as a holistic signature and explore how well it can be used to distinguish between a range of different real-world materials, without the need to perform traditional peak-by-peak analysis. Since prompt gamma emissions do not depend on the chemical (electronic) state of a material the spectrum can provide a simplified picture solely based on the atomic composition. Unlike other non-targeted spectroscopic analysis methods, PGAA spectra contain contributions from each isotope present in a sample that undergoes the nuclear activation process. It is therefore natural to assume that such spectra will contain enough information to discriminate between many different classes of materials and complex mixtures.

Here, we explore whether accurate predictive models can be developed for this purpose. To do so, we make retrospective use of PGAA data acquired at the NCNR on various standard reference materials (SRMs) and other common materials to train these models. Given that electronic

✉ Nathan A. Mahynski
nathan.mahynski@nist.gov

[1] Chemical Sciences Division, National Institute of Standards and Technology, Gaithersburg, MD 20899-8320, USA

[2] Ralph E. Martin Department of Chemical Engineering, University of Arkansas, Fayetteville, AR 72701, USA

structure creates chemical identity and that PGAA only measures the nuclear composition of a material, we investigate what classes of materials it can differentiate. For example, if PGAA-based models can distinguish between explosive and non-explosive materials, can they also be used to distinguish between organic fuels?

We employed various machine learning and conventional multivariate classification and authentication tools to answer this question. Conventional machine learning classification models often excel at differentiation between a set of finite, well-sampled categories; these are referred to as discriminative models. However, in this case the categories themselves are not necessarily well-defined. For example, should all organic material be considered a single class, or can it be broken into subcategories such as coal, coke, and oil? Furthermore, these algorithms often require that all possible classes in multiclass classification tasks be sampled (closed set conditions) and trained on, but this is not reasonable when considering an infinite number of different possible materials (open set conditions) that may be encountered during deployment. Open set recognition and detecting if a test sample falls outside a model's training distribution is an active area of research in machine learning, which requires algorithms that generalize well [4–7], and is beyond the scope of this work.

Instead, we focus on more conventional class modeling approaches to build classifiers capable of working in this setting [8, 9]; these multivariate methods build models by observing measurements from a single class and yield a binary prediction that a new sample is, or is not, consistent with that class. This distinguishes the task of authentication from conventional classification. Essentially, they develop acceptance boundaries that are elliptical and do not tessellate the latent space (though they may overlap), which is how many conventional machine learning classifiers operate (cf. Fig. 1a). It is then possible to predict that a sample falls within the acceptance region (ellipse) of zero, one, or multiple known classes enabling new samples to be authenticated against a set of known materials. The case of zero is akin to anomaly detection in that the conclusion is the sample is a novel material. The case of multiple acceptances may also be insightful since a prediction that a sample may be coal, coke, and/or oil helps suggest which classes are easily distinguished from each other post hoc, and may help alleviate the training burden on models if it is deemed acceptable to combine such classes into a single one.

In this work, we evaluate machine learning models and other multivariate methods at discriminative and class modeling (authentication) tasks. We also consider two popular unsupervised dimensionality reduction methods to visually suggest rational class labelling schemes and infer which classes might be naturally separable by PGAA. We present a workflow to perform these tasks and illustrate how PGAA spectra may be used to build class models to authenticate materials.

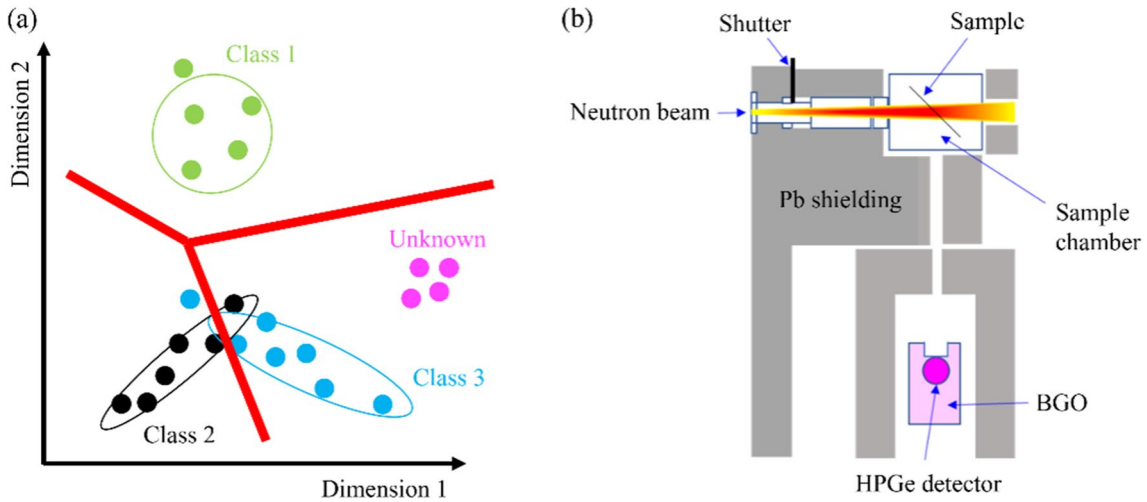


Fig. 1 **a** Example of decision boundaries for discriminative models (red lines) and class models (colored ellipses). The former models divide the latent space into disjoint regions which cover the entire space, while the latter may overlap and do not typically encompass the entire latent space. **b** Schematic of the PGAA instrument. A beam of neutrons enters the sample chamber irradiating the material, which then emits prompt gamma rays. This emission is detected by a high purity germanium (HPGe) detector (surrounded by a bismuth germanate (BGO) scintillator for Compton suppression). (Color figure online)

Methods

Prompt gamma ray activation analysis

The PGAA instrument is schematically shown in Fig. 1b. The sample under neutron beam irradiation emits prompt gamma rays that are collected by a high purity germanium detector. The acquisition lasts from minutes to hours, depending on the concentration of the elements of interest as well as the precision desired. The signal pulses are processed by digital signal processing electronics and sorted by a multi-channel analyzer into counts in each energy bin. The acquisition is controlled by computer via ethernet connection, and the spectra are recorded for off-line analysis.

Dataset summary

Our dataset consists of a variety of samples of different organic and inorganic materials. Figure 2a shows a summary of the different categories of materials used. Various SRMs and materials were selected as representative of each class, and complete descriptions of the selected materials in each category are available in the Supplemental Information (SI). For example, "steel" contains samples of various alloys, and "biomass" contains samples ranging from wood chips to plant leaves. We will partially evaluate the validity of these category choices later. In this work we only attempt to model materials for which we have at least 10 different samples. The remaining materials, those with less than 10 different samples, are kept in a held-out challenge set to test each model's novelty detection capabilities after they are trained. All data used in this work is available for download at Ref. [10].

Data collection and pre-processing

PGAA spectra were collected as histograms. The instrument used at NIST to obtain this data collects spectra in $2^{14} = 16{,}384$ energy bins spaced evenly to cover a range of up to approximately 12 MeV. The energy value for each bin is estimated by a calibration run which produces a linear fit of bin index to energy. This means the numerical energy value of a bin can vary slightly between measurements. All spectra were aligned to 2^{14} new bins evenly spaced between the global minimum and maximum energies in the dataset by linearly interpolating each spectrum at the fixed bin centers. Next, we coarsened the spectra by summing every 4 bins to produce aligned spectra with $2^{12} = 4{,}096$ total bins. Since very low energy portions of the spectra are considered unreliable, we removed the first 40 bins so that the spectra spanned from approximately 0.1 to 12 MeV using 4,056 bins. Finally, we normalized each spectrum so that the total number of counts summed to unity; while it is possible to normalize these measurements using the length of collection

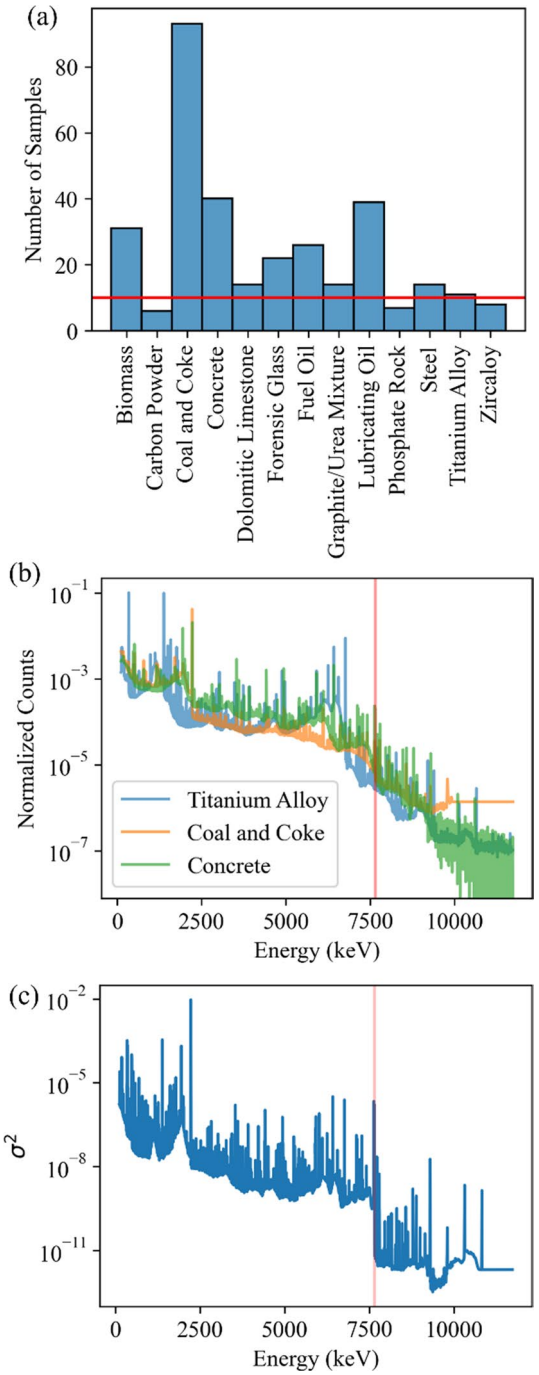


Fig. 2 Summary of PGAA spectral dataset used in this work. **a** Number of samples from each class. The horizontal red line indicates the minimum number of observations (10) needed to be considered for modeling. **b** A representative spectrum from several different classes. **c** Variance of the spectra at each energy bin across the entire dataset. The vertical red line in **b**, **c** at 7650 keV is a guide to the eye. (Color figure online)

time, calibrated neutron flux, and the mass of the sample [11] we found an empirical normalization to be simpler, and more consistent as it does not depend on the accurate measurement of other factors.

The energy range over which spectra are collected varies. Bins beyond an individual measurement's limit are fixed at the last measured value, creating an artifact; e.g., the Coal and Coke sample in Fig. 2b displays a flat line at high energy. Changing this value did not qualitatively affect the outcome of this study. Detector efficiency also decreases non-linearly at higher energies, leading to lower counts. As a result, both the global mean and variance over the dataset systematically decrease (cf. Fig. 2c) in higher energy regions of the spectra. These high-energy regions contain the artifacts, which should be regarded as spurious; Fig. 2b, c illustrate an estimated cutoff around 7650 keV after which artifacts from across the dataset create a notable decrease in bin variance. It is possible to simply truncate the spectra above this upper bound, however this eliminates all potential information that may be contained beyond this. Instead, we used a variance thresholding scheme whereby any energy bin which has a variance (measured over the available dataset) below some threshold value was removed. Bins with low variance contain essentially the same values and may be regarded as background noise. Furthermore, thresholding has the benefit of automatically trimming the spurious tail of the spectra, but the cutoff does not need to be determined a priori. Instead, it is a hyperparameter that can be optimized during model training.

Model training and comparison

Pipelines

All predictive models were built using pipelines in scikit-learn [12] and compatible python packages. A pipeline is a series of individual steps combined sequentially as shown in Fig. 3. The final step in a pipeline is the model which yields a prediction. Different pipelines may contain different preprocessing steps with various hyperparameters, which are optimized using cross-validation (CV). Testing and training data (or folds during CV) follow different paths. The performance of each pipeline is estimated using a nested scheme illustrated in the SI. The overall dataset is first broken into $R=5$ different data (sub)sets; K-fold CV is performed on each subset and the performances on the $K=2$ (validation) sets are recorded to determine the optimal pipeline. The $R\times K$ total scores may be averaged to estimate the performance and uncertainty that can be expected when this pipeline is optimized using CV, and is amenable to statistical testing, though it was not necessary in this work [13, 14]. Here, we report the mean and standard deviation of the $R\times K$ total scores as an estimate of the pipeline's performance and uncertainty after training on new data. Although these performance estimates may be biased, this generally does not affect the relative ranking of pipelines [15].

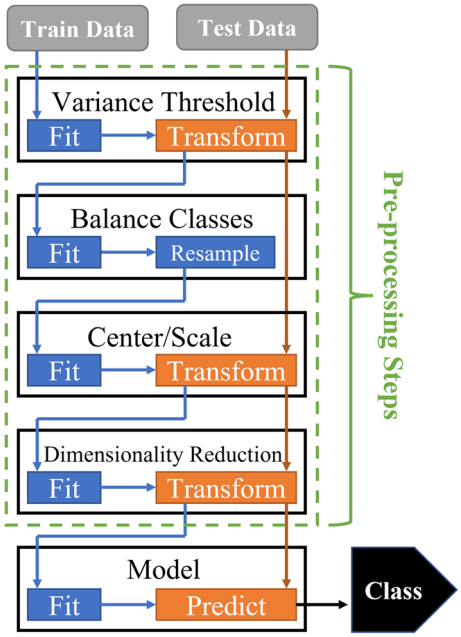


Fig. 3 Pipeline used for training models in this work. Training data (or folds) travel a different path (blue) through the steps which determine and store hyperparameters (e.g., mean or standard deviation) used during the testing phase (orange path). (Color figure online)

Once a final pipeline was selected, the overall data was split 80:20 in a stratified fashion into a single train and test set. A single fivefold CV loop was performed on the training set to determine the final hyperparameters and produce a final model. The final pipelines were evaluated on the held-out test set and in all cases were found to be consistent with the nested loop estimates used for ranking, suggesting minimal bias therein.

We used pipelines as implemented in the imbalanced-learn package [16]. Due to differences in the number of samples from each material class (cf. Fig. 2a), class balancing was performed either with SMOTEENN [16, 17] or with internal mechanisms such as class-based weighting when models allowed, but not both. In the former case, resampling was used to produce synthetic data used only during training (cf. Fig. 3); in the latter case, greater weight was assigned to incorrect predictions on minority classes during training and no additional data was generated. During the data centering and scaling step, the input data was always column-centered and divided by some scale. We allowed pipelines to select either standard or robust scaling procedures. Standard scaling (autoscaling) uses the mean and standard deviation for this transformation [18], while robust scaling uses the median and inter-quartile range instead. We also included a hyperparameter to enable Pareto scaling [19] for standard

and robust scaling, which instead divides by the square root of, respectively, the standard deviation or inter-quartile range. Dimensionality reduction was either performed in a linear fashion by principal components analysis (PCA), or in a non-linear fashion using the Pairwise Controlled Manifold Approximation Projection (PaCMAP) [20] technique. All other tools and models are from the scikit-learn python package [12]. For visualization and analysis we used Jupyter notebooks [21], seaborn [22], and pandas [23, 24]; notebooks producing these calculations are available online at Ref. [10]. All calculations were performed using the pychemauth package [25].

Discriminative models

In this work we considered several representative discriminative models for closed set classification including logistic regression and random forests. A discriminative model uses examples of a set of known classes to learn boundaries between them in some latent space [26]. This produces N disjoint regions each associated with one of the N total classes seen during training (cf. Fig. 1a). When classes are reasonably balanced and there is no reason to devote additional concern to any subset of them, accuracy is a reasonable metric that characterizes the performance of such models. If n_{ij} is the number of members of class i predicted to belong to class j, then the accuracy is:

$$acc = \frac{1}{M}\sum_{i=1}^{N} n_{ii} \tag{1}$$

where we have a total of M samples in the dataset and $M = \sum_{i=1}^{N}\sum_{j=1}^{N} n_{ij}$. The best model is the one with the best accuracy.

Class models

Class models, or one-class classifiers (OCC), are trained to determine if a new sample is consistent with a particular class; here, a separate model is trained for each class, though this may be done in several ways [8, 9, 26, 27]. For "rigorous" OCCs, only examples of the class itself are used during training [9, 28]. In contrast, "compliant" OCCs also use alternative classes during training to determine how well those alternates are rejected by the model; this can assist in the training, but it may introduce some bias based on which alternates are used.

In this work we used the popular OCC data-driven soft independent modeling by class analogy (DD-SIMCA) model [28–30]. In DD-SIMCA the input matrix, X, containing rows of spectra collected from only the target class are centered and possibly scaled, then PCA is used to perform dimensionality reduction. Regardless of the number of principal components, two distances determine class membership: the score distance, h, and the orthogonal distance, q (cf. SI for details and definitions). The former reflects a point's position within the class space, whereas the latter captures the distance away from the class space (the error introduced from the dimensionality reduction). The total distance for each point,

$$c = N_h\left(\frac{h}{h_0}\right) + N_q\left(\frac{q}{q_0}\right) \sim \chi^2_{N_h+N_q} \tag{2}$$

is assumed to be distributed according to a chi-squared distribution (h_0 and q_0 are scaling factors, cf. SI) such that class membership is determined by a critical chi-squared value with $N_h + N_q$ degrees of freedom and a significance level, α, which is the type I error rate selected (typically from 0.01 to 0.05).

This model can be assessed with several metrics: sensitivity, specificity, and efficiency [9, 31]. The model's total sensitivity (TSNS) is the rate of true positives: TSNS = TP/(TP + FN); a true positive, TP, occurs when the model correctly accepts a true class member, while a false negative, FN, occurs when a true member is rejected. The model's total specificity (TSPS) refers to the true negative rate; TSPS = TN/(TN + FP) = 1 − FP/(TN + FP), where TN is the number of true negatives and FP is the number of false positives. The geometric mean of these two is the total efficiency: $TEFF^2 = TSNS \times TSPS$.

For rigorous DD-SIMCA models, only TSNS can be computed since only members of the target class are available, so the optimal model is the one where $TSNS = 1 - \alpha$ [9]. In this case hyperparameters, such as the number of principal components, are adjusted to meet this target while α is fixed. For compliant DD-SIMCA models where alternative classes are available, TSPS can be computed and the model with the highest TEFF is selected; in this case, α is allowed to vary to increase TEFF.

Intermediate models

We consider partial least-squares discriminant analysis (PLS-DA) to be an intermediate approach between OCCs and discriminative models. PLS-DA performs partial least-squares regression (PLS2) against a one-hot encoded class target matrix for N classes to map the input to a latent space of N dimensions containing an N-simplex where each of the vertices corresponds to a different class; for example, a tetrahedron in 3D. The extra vertex located at the origin means the latent space effectively has one extra dimension; the approach we use here is to subsequently perform PCA using $N - 1$ components after the PLS mapping. This is more thoroughly discussed in Ref. [31].

Thus, PLS-DA yields a single model trained on a fixed set of known classes, akin to a discriminative model, but the boundary around each class center can be made to mimic either a discriminative model with "hard" boundaries or an OCC with "soft" boundaries. A hard PLS-DA model follows by constructing hyperplanes, e.g., by using a linear discriminant analysis approach where the (squared) distance from a given point projected into the latent space, t, to a projected class center, c_k, for a class k is given by [31]:

$$d_k^2 = \left(t - c_k\right)\Lambda^{-1}\left(t - c_k\right)^T \tag{3}$$

Here, Λ refers to the pooled sample covariance matrix which is a diagonal matrix composed of the (sorted) eigenvalues from the PCA. A sample point is assigned to the nearest class center, which is always one of the known classes trained on.

Alternatively, an elliptical boundary can be determined by the Mahalanobis distance to each class center to create a soft PLS-DA model [31]. The (squared) distance is now given by:

$$d_k^2 = \left(t - c_k\right)S_k^{-1}\left(t - c_k\right)^T \tag{4}$$

where S_k refers to the within-class sample covariance matrix of class k. Rather than simply predicting class membership based on the single nearest class, class membership is assumed for any, and all, classes such that:

$$d^2 < d_{crit}^2 \tag{5}$$

where d_{crit}^2 is the critical chi-squared value with N-1 degrees of freedom and a significance level of α. Equations for total and class-based specificity and sensitivity vary slightly because of this and are discussed in more detail in the SI.

Results and discussion

Unsupervised clustering

PGAA spectra have over 4,000 bins (features) in this work which is much larger than the number of classes we will attempt to distinguish. Appropriate pre-processing and dimensionality reduction (DR) are critical to developing good models. PCA is a common tool used for linear dimensionality reduction and focuses on preserving the global structure of the data. It is also a common component found within class models. A powerful non-linear alternative is the Pairwise Controlled Manifold Approximation Projection (PaCMAP) method [20]; this method produces a low dimensional embedding by looking at pairs of points at different distances (neighbors, mid-near, and further) and can preserve both the local and global structure of the data in the original space. Both DR methods can provide intuition as to which classes are naturally separable, and what sort of confusion we may expect to arise during modeling.

Both PCA and PaCMAP are unsupervised approaches and for comparison we elect to project the PGAA spectra into 2 dimensions to simplify visualization. Figure 4 illustrates the projection, colored by class, after variance thresholding and autoscaling. When all 4,056 features are included, the PCA loadings indicate that high energy bins (> 7650 keV) are contributing significantly to these principal components. As the minimum variance threshold, T, is increased, high energy features are systematically removed. When all features are included both DR methods show separation of classes into different clusters depending on the T value. Both methods fail to separate clusters when there are too few bins allowed, but PCA's failure is more pronounced as evidenced by more overlapping ellipses. However, there is an intermediate amount of thresholding ($T \approx 10^{-8}$) that produces very clean separation; importantly, this also corresponds to the point where most of the potentially spurious, high-energy bins have been eliminated. Note that regardless of thresholding,the Coal and Coke category tends to overlap with many similar organic materials. This suggests the category is very broad, which could lead to confusion in low dimensional models. Regardless, the categories it overlaps are chemically similar organic compounds.

In what follows we include variance thresholding in all pipelines unless otherwise stated. Although PaCMAP performs well, PCA is a simpler alternative which separates classes nearly as well for this dataset, so we elect to use this exclusively. Both the dimensionality of this space and the variance threshold are key hyperparameters that, when well-tuned, enable highly accurate predictive models to be obtained.

Supervised classification

Discriminative models

First, we developed pipelines which use discriminative models to distinguish the classes in Fig. 2a with a minimum of 10 observations from the other well-sampled classes. Inspection of successful models can provide insight into what characteristic differences exist between classes in this dataset, and how discriminating PGAA spectra are for this task. Table 1 summarizes the performance of different pipelines (cf. Fig. 3) we considered. All pipelines performed quite well. Regardless of pre-processing, pipelines using a random forest (RF) model [32] performed the best, while those using quadratic discriminant analysis (QDA) performed the worst. Regardless of the model, using only thresholding as a pre-processing step yielded the worst pipelines on average.

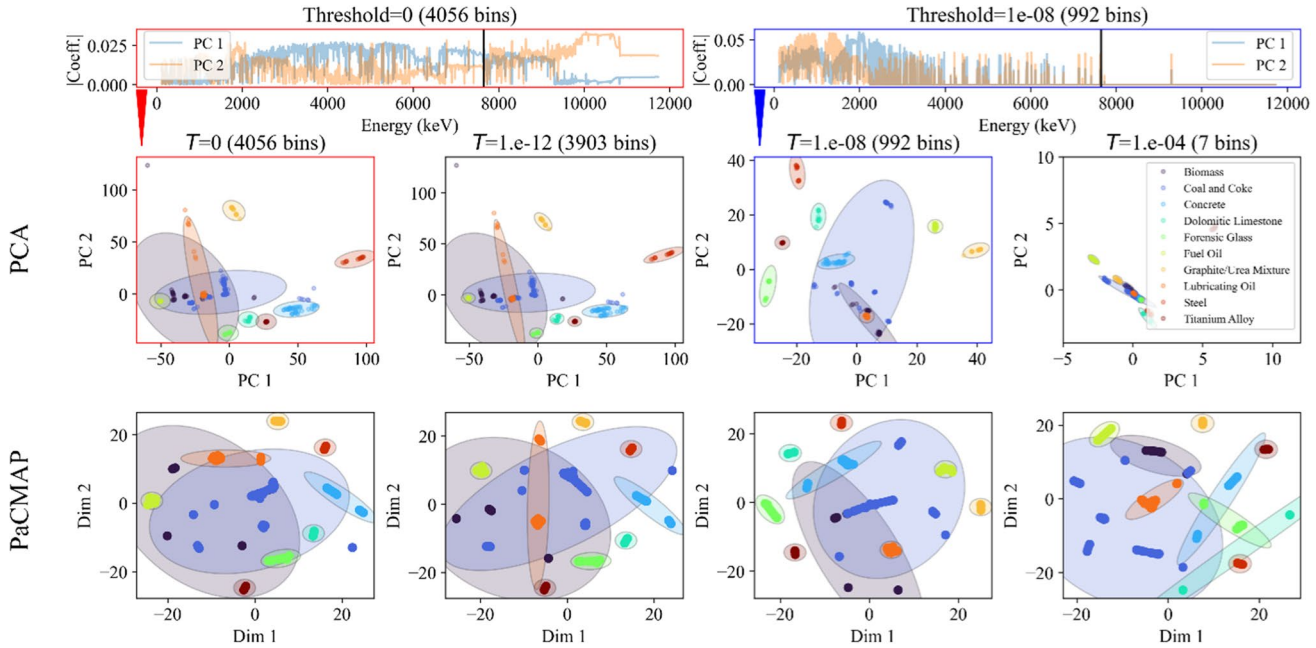


Fig. 4 Unsupervised dimensionality reduction of PGAA spectra. The absolute value of the coefficients for the 2 principal components used in PCA (first row) are shown at the top (loadings); the vertical black line denotes 7650 keV as a guide to the eye. PCA is performed after variance thresholding which removes low variance energy bins. The same variance threshold is employed before using PaCMAP to perform a non-linear dimensionality reduction into two dimensions (second row). Ellipses drawn around classes are a guide to the eye

Table 1 Accuracies and uncertainties (one standard deviation) of different pipelines with different pre-processing steps (columns) and models (rows); cf. Fig. 3

	Thresholding: standard scaling: PCA	Thresholding: robust scaling: PCA	Threshold-ing: standard scaling	Thresholding: robust scaling	Thresholding: PCA	Thresholding	Average
Random forest	1.0 (0)	1.0 (0)	1.0 (0)	1.0 (0)	0.997 (8)	1.0 (0)	0.999
Hard PLS-DA	0.997 (4)	0.993 (4)	0.999 (2)	0.999 (2)	0.986 (10)	0.999 (2)	0.995
Linear discriminant analysis	1.0 (0)	0.997 (4)	0.992 (5)	0.992 (5)	0.996 (4)	0.992 (5)	0.995
Logistic regression	0.997 (4)	0.998 (3)	0.998 (4)	0.998 (4)	0.984 (11)	0.970 (19)	0.991
Decision tree	0.991 (11)	0.997 (5)	0.988 (14)	0.988 (14)	0.988 (13)	0.988 (14)	0.990
Quadratic discriminant analysis	0.989 (9)	0.993 (5)	0.970 (19)	0.972 (24)	0.976 (12)	0.71 (7)	0.935
Average	0.996	0.996	0.991	0.991	0.988	0.943	

Steps in the pipelines are separated by semicolons. These performances were estimated as described in the SI. Pipelines using LDA, QDA, and the hard PLS-DA models used SMOTEENN to perform class balancing, while the rest used frequency-based weighting to balance the models

This was improved by adding data scaling or PCA steps, but the best pipelines employed both. This is unsurprising given the natural separation that can be found between the classes we used in this work even in only 2 dimensions (cf. Fig. 4). Robust scaling did not typically outperform standard scaling.

Although pipelines with RF models performed the best, RF models are typically difficult to interpret. Instead, here we examine pipelines using a logistic regression (LR) model, which is more naturally interpretable and did not substantially underperform RF-based pipelines. In these LR-based pipelines, multinomial LR was used to predict the un-normalized probability, p_k, that a sample, x, belongs to a class, k, with multilinear regression:

$$\ln(p_k) = a_{k,0} + \sum_{i=1}^{4,056} a_{k,i} s_i \left(\frac{x_i}{s_i}\right) \quad (6)$$

The softmax function was then used to compute a normalized probability for each class. The coefficients, $a_{k,i}$, on each histogram bin, x_i, give insight into the significance of each bin for each class. Variance thresholding effectively sets certain $a_{k,i} = 0$ reducing the number of terms which contribute to this sum. The scale of x_i factors into the interpretation of these coefficients; dividing an energy by its bin's sample standard deviation over the dataset implies the product, $a_{k,i}s_i$, can be taken as a dimensionless term indicating the significance of each bin.

A final LR-based pipeline with no other pre-processing besides variance thresholding was trained using an 80:20 train:test split of the data with fivefold CV used to optimize hyperparameters on the training set. The scaled coefficients are shown in Fig. 5a. Only results for 3 representative materials are given, since they all displayed prominent peaks at 3 main energy bins (roughly 2224, 1382, and 342 keV, cf. SI). The optimal variance threshold was found to be $T_{\text{opt}} = 1.0\text{e}{-10}$ which is slightly less than the optimal suggested by Fig. 4 if it is combined with dimensionality reduction. Regardless, this model, like others trained without variance thresholding, naturally found the primary differences originate in the lower energy portion of the spectra. This result is specifically premised on the 10 classes used here and may not hold true in general.

For comparison, we also optimized a pipeline that used a logistic regression model after pre-processing the spectra by centering, standard scaling, then using PCA. After CV, 10 latent variables in the PCA were found to be optimal, with an accuracy of 100% on both the test and training set (again, $T_{\text{opt}} = 1.0\text{e}{-10}$); however, the optimal pipeline when using only 2 dimensions yielded a 96.7% accuracy on the test set and is easier to visualize, so we report this model here ($T_{\text{opt}} = 1.0\text{e}{-8}$). Figure 5b shows the 2D latent space used by this pipeline, with the subsequently trained logistic regression model's decision boundaries. The training data depicted naturally separate in this space with the exception of the "Coal and Coke" class; PC 1 is essentially just the 2224 keV energy peak corresponding to hydrogen, while PC 2 is primarily composed of the titanium peaks at 342 and 1382 keV. Even after completely different pre-processing steps, these peaks are again the most important features to the LR model in these pipelines.

Collectively, this suggests that these 10 classes are characteristically different based on their hydrogen and inorganic/metallic content. Thus, we essentially have one pseudo-organic axis (PC 1) and one pseudo-inorganic axis (PC 2) along which these training materials separate very well. If our training set contained different materials, these dimensions could change. This is the central

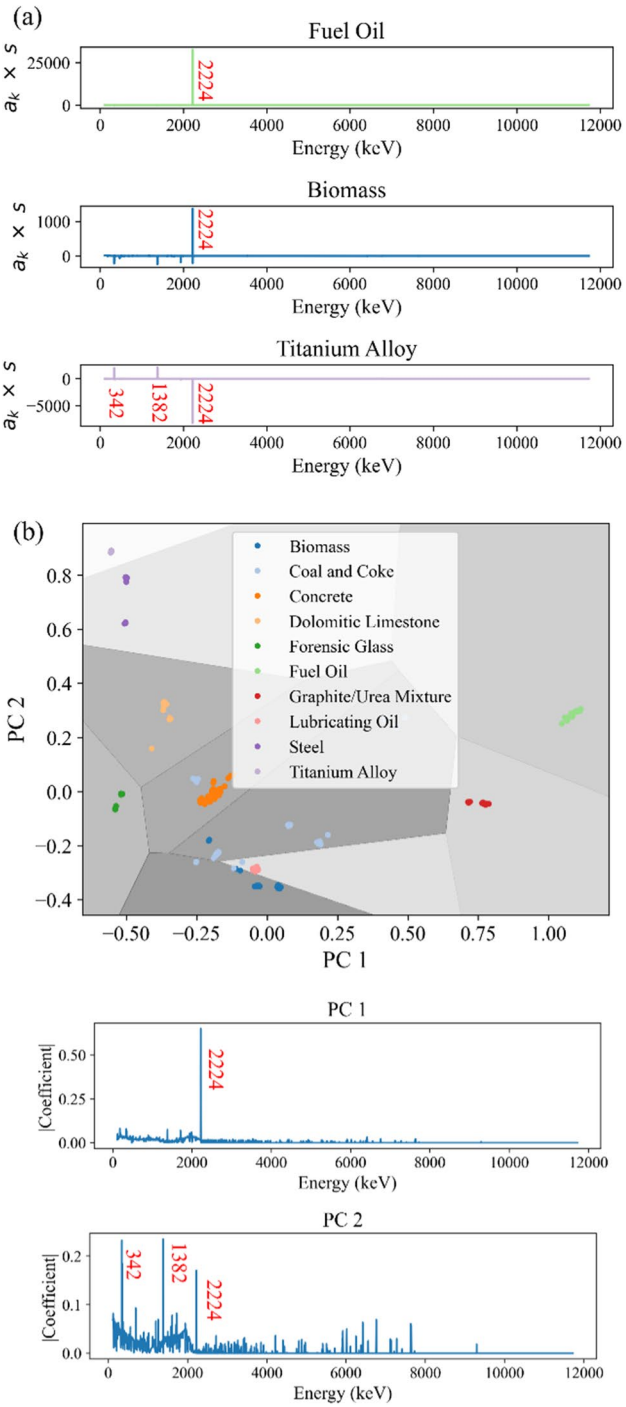


Fig. 5 Predictions using a logistic regression model. **a** Coefficients for 3 representative materials show that 3 energy bins (at roughly 2224, 1382, and 342 keV) are the most significant to the pipeline using only variance thresholding. **b** Logistic regression model decision boundaries after pre-processing by centering, (Pareto) standard scaling, then performing PCA. The loadings indicate that PC 1 and PC 2 reflect the same importance of these 3 peaks

problem with discriminative models since they focus on learning the differences between known training classes. In contrast, class models focus on determining the essence of a material.

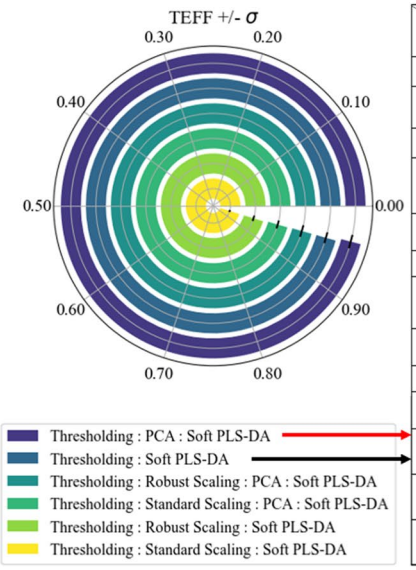
Soft PLS-DA model

Soft PLS-DA produces ellipsoidal acceptance regions around class centers in a latent space whose dimensionality is determined by the number of classes used during training (due to one-hot encoding). Thus, both the number and specific choice of classes to use during training can introduce some bias into the model. Regardless, the soft boundaries are amenable to authentication problems under open set conditions since they enable novelty detection. We trained various pipelines using soft PLS-DA as the model following the same set of different pre-processing steps reported in Table 1. The average performances and one standard deviation are shown in Fig. 6.

Here, pre-processing makes only very minor differences to the pipeline's final performance (TEFF ≈ 0.95 always). We compare the top two pipelines: one involving only variance thresholding, the other adding PCA to this step. Before any PCA is performed the data is always column-centered, but here it is not scaled. Still, including scaling yielded similar results. On the final test set the first pipeline yielded TEFF = 0.958, whereas the second yielded TEFF = 0.973.

The table in Fig. 6 details how materials from different classes (rows) are assigned (columns) for the final pipelines. Once again, $T_{opt} \geq 1.0e{-}8$ was found naturally by cross validation. As a baseline, the pipeline with only thresholding is reported in black, and changes which occur when including PCA are shown in red. For example, there were 6 Biomass samples in the test set. In the first pipeline all 6 were assigned to both the "Biomass" and "Coal and Coke" categories, whereas the second pipeline did not confuse any of these with "Coal and Coke." In general, the "Coal and Coke" column is much more populated by the first pipeline than the second. This is the primary source of error for this model, the cause of which is qualitatively reflected in Fig. 4; the "Coal and Coke" category is very broad and seems to encompass several other classes of organic materials. Including PCA in the pre-processing steps seems to help structure the latent space better, which allows the underlying PLS2 model to better exclude non-members. This pipeline has a substantially higher class specificity, CSPS = 0.952 vs. 0.548, as a result (cf. SI).

Despite its biases, soft PLS-DA performs very well on this dataset, especially with the appropriate pre-processing steps in place. If a reasonably complete set of classes can be enumerated and measured, PGAA spectra modeled by soft PLS-DA may enable high-quality authentication and classification of materials in practice. Note that hard PLS-DA was also the second best performing discriminative model, on average, in Table 1.



Assigned / Member	Biomass	Coal and Coke	Concrete	Dolomitic Limestone	Forensic Glass	Fuel Oil	Graphite / Urea Mixture	Lubricating Oil	Steel	Titanium Alloy	Unknown	M_{train}	M_{test}
Biomass	6	6 (0)	0	0	0	0	0	0	0	0	0	25	6
Coal and Coke	1 (0)	19 (18)	0	0	0	0	0	0	0	0	0 (1)	74	19
Concrete	0	0	8	0	0	0	0	0	0	0	0	32	8
Dolomitic Limestone	0	0	0	3	0	0	0	0	0	0	0	11	3
Forensic Glass	0	4 (0)	0	0	4	0	0	0	0	0	0	18	4
Fuel Oil	0	0	0	0	0	5	0	0	0	0	0	21	5
Graphite / Urea Mixture	0	1 (0)	0	0	0	0	2	0	0	0	0 (1)	11	3
Lubricating Oil	8 (0)	8 (2)	0	0	0	0	0	8	0	0	0	31	8
Steel	0	0	0	0	0	0	0	0	3	0	0	11	3
Titanium Alloy	0	0	0	0	0	0	0	0	0	1	1	9	2
Test CSPS	0.836 (1.000)	0.548 (0.952)	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0			
Test CSNS	1.0	1.000 (0.947)	1.0	1.0	1.0	1.0	0.667	1.0	1.0	0.500			
Test CEFF	0.915 (1.000)	0.740 (0.950)	1.0	1.0	1.0	1.0	0.816	1.0	1.0	0.707			

Fig. 6 Test set performances of pipelines using a soft PLS-DA model. The dial on the left indicates the mean performance (error bars are one standard deviation) over $R \times K$ different subsets of the data. The results for the second-best performer (only variance thresholding used to pre-process) are given in black; differences in the best model (which includes PCA in pre-processing) are shown in red. In the former, $T_{opt} = 1.0e{-}4$, $\alpha = 0.01$, and the number of latent variables in the PLS2 stage was 3. In the latter, $T_{opt} = 1.0e{-}8$, $\alpha = 0.05$, and the number of latent variables in the PLS2 stage was 2; the PCA compressed the data into 10 dimensions

Class models

Next, we developed both rigorous and compliant DD-SIMCA models for these materials. Compliant models make use of alternative classes to evaluate a model's TSPS (and TEFF) which affects hyperparameter selection during CV. This additional information may enable better performance against these known alternative classes, however, the impact of this bias is difficult to assess. Rigorous models do not need to be re-evaluated when new alternative classes become available, whereas compliant ones may benefit from this. Since DD-SIMCA performs PCA to create the class space in the first place, we did not bother with an additional PCA pre-processing step. Instead, we trained pipelines with only variance thresholding (where we enforced $T_{opt} \geq 1.0e{-}8$) and class balancing (via SMOTEENN). Non-robust estimates were used for the scaling parameters and degrees of freedom [29].

Table 2 summarizes the performance of these models on the final held-out test set, and the hyperparameters of the optimal models. Most models relied on only a single latent variable (LV) in the PCA and elected to use the lowest α value allowed (0.01). The values correspond to compliant models which use samples from the other 9 classes to compute TSPS used during training. The bold numbers correspond to rigorous models which do not compute TSPS during training since they see only examples of the class they are meant to model. In the latter, hyperparameters are tuned using CV to achieve TSNS = $1 - \alpha$; to make a fair comparison to compliant models, we selected $\alpha = 0.01$ as their target. Compliant models tended to use a higher variance threshold, focusing on lower energy portions of the spectra, and more latent variables than their rigorous counterparts. Representative models are summarized in Fig. 7.

Nearly all the models performed identically on the test set, with the notable exception of the model for Coal and Coke. TEFF is quite high for most models; TSNS is less than one for Graphite/Urea Mixtures and Titanium Alloy, though these have 3 or less samples in the test set to evaluate this on. More data would likely improve both the model itself and the accuracy of the test set performance estimate. The TSPS = 1 for all compliant models and only substantially different for the rigorous Coal and Coke model, for which TSPS drops to 0.619.

Again, the breadth of this category leads to issues making a model which is specific enough to exclude other similar organic materials. The acceptance plot for the compliant and rigorous DD-SIMCA models is given in Fig. 8. The training data is depicted to illustrate how these models have been tuned. The rigorous model, seeing only examples of Coal and Coke during training, correctly accepts all 74 training examples of this class, but also erroneously accepts all 25 Biomass samples and all 31 Lubricating Oil samples in the training set. Moreover, 3 Fuel Oil samples were erroneously accepted, and the remaining 18 (shown in yellow crosses in Fig. 8) were nearly accepted. This qualitatively agrees with Fig. 4 which shows how these intuitively similar categories can end up overlapping after PCA is applied. The advantage of using these alternative classes during training is clear, and the optimal compliant model uses 2 additional LVs to separate Coal and Coke from these other classes. Figure 8 shows that out of the training data only 1 of the 74 Coal and Coke samples ended up being erroneously rejected, and only 2 out of 25 Biomass samples were incorrectly accepted by the final compliant model; on the test set, TSPS = TSNS = 1 (cf. Table 2).

Overall, this suggests that accurate class models can be developed for PGAA spectra which perform as well as soft PLS-DA or discriminative models. While it is important to take care that classes are not so broad that they can be easily confused with similar materials, using compliant models can be a fruitful way to combat this effect, if necessary. For example, the compliant DD-SIMCA model's TEFF for Coal

Table 2 Test set performance of DD-SIMCA models

	Test TSNS	Test TSPS	Test CEFF	# LV	T_{opt}	α
Biomass	1	1 **(0.982)**	1 **(0.991)**	2 **(1)**	1e–4 **(1e–6)**	0.01
Coal and coke	1	1 **(0.619)**	1 **(0.787)**	3 **(1)**	1e–6 **(1e–8)**	0.01
Concrete	0.875	1	0.935	1	1e–8	0.01
Dolomitic limestone	1	1	1	1	1e–6	0.01
Forensic glass	1	1	1	1	1e–4 **(1e–8)**	0.01
Fuel oil	1	1	1	2	1e–8	0.01
Graphite/urea mixture	0.667	1	0.816	1	1e–4 **(1e–8)**	0.01
Lubricating oil	1	1	1	1	1e–8	0.01
Steel	1	1	1	1	1e–8	0.01
Titanium alloy	0.500	1	0.707	1	1e–4 **(1e–8)**	0.01

The results from the compliant model are shown in black; the rigorous model yielded identical results except where indicated in bold

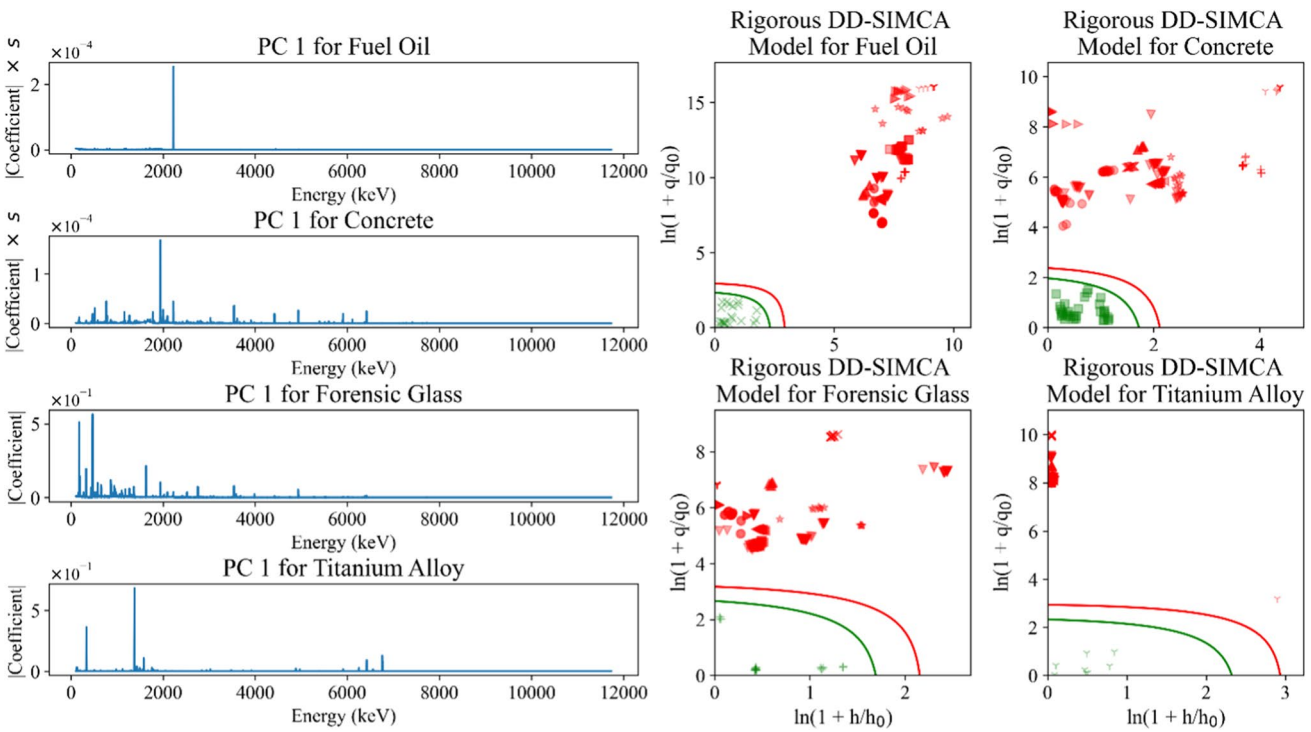


Fig. 7 Rigorous DD-SIMCA models for several materials. At the left, the absolute values of the coefficients (loadings) for the leading principal component is given. One hyperparameter in DD-SIMCA models is whether to column-wise divide the input by its sample standard deviation after centering (before PCA). To fairly compare models, if the data was scaled, then the loadings are multiplied by the scale(s) here (fuel oil and concrete); otherwise just the coefficients are reported. Acceptance plots are shown on the right. The green curve is the acceptance boundary for class membership ($\alpha = 0.01$), while the red line is an outlier threshold corresponding to the same significance level based on Ref. [29]. (Color figure online)

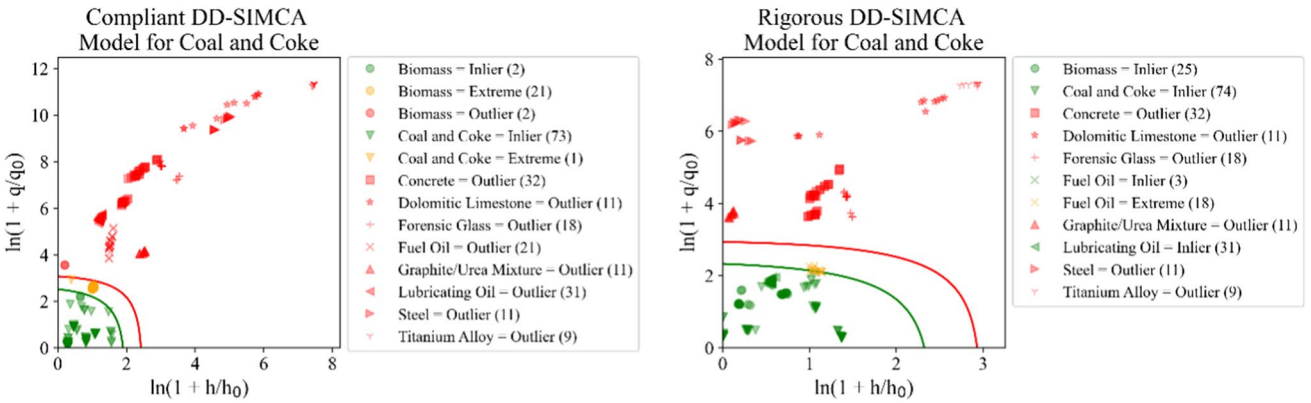


Fig. 8 Acceptance plots for DD-SIMCA models of Coal and Coke when trained to be compliant (left) vs. rigorous (right). Predictions made on the training set are shown here. Green symbols are accepted as a member of the class by the model, yellow and red are rejected, and different symbols are used to denote different classes. The green curve is the acceptance boundary for class membership ($\alpha = 0.01$), while the red line is an outlier threshold corresponding to the same significance level based on Ref. [29]. (Color figure online)

and Coke (TEFF = 1) exceeds the class efficiency for this category in the best PLS-DA model (0.950).

Authentication Tests

Class models are considered the most useful for authentication tasks, where an arbitrary sample is obtained and the goal is to ascertain if it is consistent with any known materials or

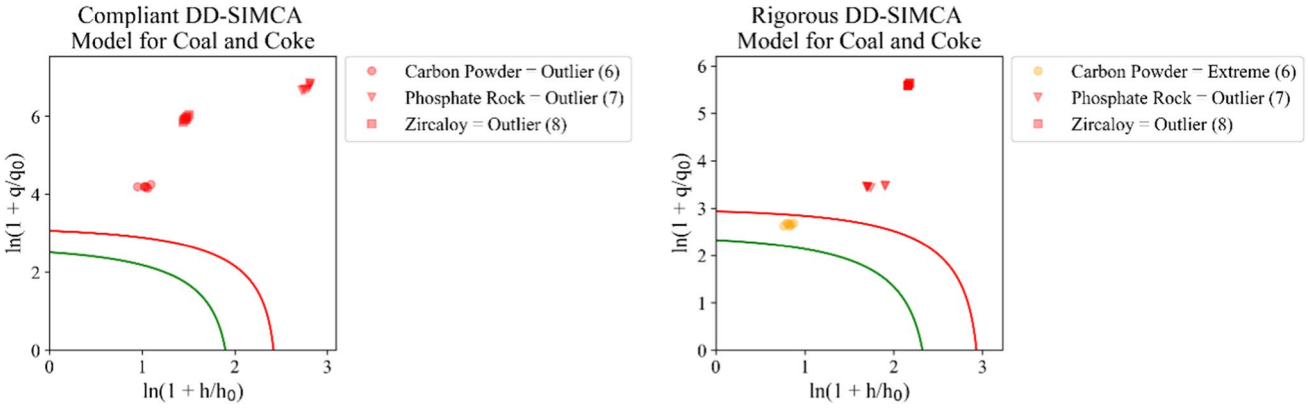


Fig. 9 Acceptance plots for DD-SIMCA models of Coal and Coke. These are the same models as in Fig. 8. Predictions are made for the classes which we did not use to train any models. The Carbon Powder samples are much closer to the acceptance boundary in the rigorous model than the compliant model. The green curve is the acceptance boundary for class membership ($\alpha = 0.01$), while the red line is an outlier threshold corresponding to the same significance level based on Ref. [29]. (Color figure online)

not. Here we test the DD-SIMCA models against the samples from the 3 categories in Fig. 2a which did not have enough samples to warrant the training of their own class model: Carbon Powder, Phosphate Rock, and Zircaloy. For such cases we can only evaluate the TSPS of the DD-SIMCA models to ensure all models reject these materials as members; in fact, both the rigorous and compliant models yielded a perfect TSPS = 1 against these materials for all classes. Figure 9 shows the acceptance plots for the Coal and Coke models, which are the most interesting. Both the rigorous and compliant models reject all alternate classes, but the Carbon Powder samples are much closer to the acceptance boundary (green curve) for the compliant model than the rigorous one. This suggests DD-SIMCA models based on PGAA spectra can be very accurate and robust for authentication purposes. These models did not require any preprocessing beyond simple thresholding making them easy to train and deploy.

Conclusions

In this work we developed a variety of models to differentiate materials using PGAA spectra. Key to these models' success is the appropriate use of pre-processing to ensure models do not overfit to unreliable regions of the spectra. Here, this was accomplished with variance thresholding and dimensionality reduction via PCA. We developed discriminative models to help explain the basic differences between materials being considered here; despite their excellent performance, these models are considered less applicable under most real-world conditions since they can only distinguish between the known set of materials they were trained on. To handle open set conditions more appropriately, we trained soft PLS-DA and DD-SIMCA models. Both performed very similarly, but each have individual caveats that may make one more appropriate than another in different scenarios. Soft PLS-DA carries an implicit bias owing to the number, and specific choice, of categories to use during training. However, if it is possible to sample most classes that are going to be encountered when deployed, this can be a powerful model for authenticating materials. If this is not possible, DD-SIMCA models were found to perform similarly on this dataset. Rigorous class models may be trained with sufficient data to represent intra-class variance but may struggle to distinguish very similar materials. Compliant models can circumvent this issue, often increasing the model complexity, but introduce bias which is difficult to fully quantify.

However, we caution that all conclusions reached here are based on models trained using laboratory prepared samples, many of which are standard reference materials. These materials are highly homogenous and designed to have very low variance between samples of the same material. Although we have defined many classes to be composed of multiple SRMs, it is important to note that this low variance represents an idealized version of a class. In practice, intra-class variance is expected to be higher in many real-world materials; this is likely to make it harder for models to achieve good specificity if a class is very broad, akin to issues seen here with Coal and Coke. Regardless, this study illustrates that PGAA spectra can be used to develop high-performance models for many classes of materials; moreover, when "broad" categories are encountered, adding dimensionality to models and using compliant approaches can circumvent issues with low specificity. It is primarily a matter of obtaining sufficient data to train such models, since the computational time required to authenticate a newly acquired material using any of these trained models is minimal (i.e., less than a second on any modern computer).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10967-023-09024-x.

Acknowledgements Certain equipment, instruments, software, or materials, commercial or non-commercial, are identified in this paper to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement of any product or service by NIST, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose. Contribution of the National Institute of Standards and Technology, not subject to US Copyright.

Data availability Data available from authors upon reasonable request.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

1. Eftekhari-Zadeh E, Feghhi SAH, Roshani GH (2017) Hybrid combination of multi-layer perceptron and neutron activation analysis in cement prediction. Pramana—J Phys. https://doi.org/10.1007/s12043-016-1327-2
2. Nunes WV, Da Silva AX, Crispim VR, Schirru R (2002) Explosives detection using prompt-gamma neutron activation and neural networks. Appl Radiat Isot 56:937–943
3. Hossny K, Hossny AH, Magdi S, et al (2020) Detecting shielded explosives by coupling prompt gamma neutron activation analysis and deep neural networks. Sci Rep. https://doi.org/10.1038/s41598-020-70537-6
4. Vaze S, Han K, Vedaldi A, Zisserman A (2021) Open-set recognition: a good closed-set classifier is all you need? arXiv:2110.06207
5. Geng C, Huang S, Chen S (2018). Recent advances in open set recognition: a survey. https://doi.org/10.1109/TPAMI.2020.2981604
6. Dietterich TG, Guyer A (2022) The familiarity hypothesis: explaining the behavior of deep open set methods. Pattern Recogn. https://doi.org/10.1016/j.patcog.2022.108931
7. Yang J, Zhou K, Li Y, Liu Z (2021) Generalized out-of-distribution detection: a survey. arXiv:2110.11334
8. Forina M, Oliveri P, Lanteri S, Casale M (2008) Class-modeling techniques, classic and new, for old and new problems. Chemom Intell Lab Syst 93:132–148. https://doi.org/10.1016/j.chemolab.2008.05.003
9. Oliveri P (2017) Class-modelling in food analytical chemistry: development, sampling, optimisation and validation issues—a tutorial. Anal Chim Acta 982:9–19. https://doi.org/10.1016/j.aca.2017.05.013
10. Mahynski NA, Monroe JI, Sheen DA, et al (2023) pgaa-material-authentication. https://github.com/mahynski/pgaa-material-authentication. Accessed 9 Feb 2023
11. Mackey EA, Paul RL, Lindstrom RM et al (2005) Sources of uncertainties in prompt gamma activation analysis. J Radioanal Nucl Chem 265:273–281
12. Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2858–2830
13. Nadeau C, Bengio Y (2003) Inference for the generalization error, pp 239–281. https://doi.org/10.1023/A:1024068626366
14. Bouckaert RR, Frank E (2004) Evaluating the replicability of significance tests for comparing learning algorithms. Lect Notes Comput Sci (Includ subser Lect Notes Artif Intell Lect Notes Bioinformat) 3056:3–12. https://doi.org/10.1007/978-3-540-24775-3_3
15. Wainer J, Cawley G (2021) Nested cross-validation when selecting classifiers is overzealous for most practical applications. Expert Syst Appl 182:115222
16. Lemaitre G, Nogueira F, Aridas CK (2017) Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res 18:1–5
17. Chawla N, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intel Res 16:321-357
18. Vandeginste BGM, Massart and DL, Buydens LMC, et al (1998) Analysis of measurement tables. In: Handbook of chemometrics and qualimetrics: part B. Elsevier, pp 87–160
19. van den Berg RA, Hoefsloot HCJ, Westerhuis JA et al (2006) Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genom 7:1–15
20. Wang Y, Huang H, Rudin C, Shaposhnik Y (2020) Understanding how dimension reduction tools work: an empirical approach to deciphering t-SNE, UMAP, TriMAP, and PaCMAP for data visualization, pp 1–63. arXiv:2012.04456
21. Kluyver T, Ragan-Kelley B, Pérez F et al (2016) Jupyter Notebooks - a publishing format for reproducible computational workflows. In: Loizides F, Schmidt B (eds) Positioning and power in Academic Publishing: players, agents and agendas. IOS Press, pp 87–90
22. Waskom M (2021) Seaborn: statistical data visualization. J Open Sour Softw 6:3021. https://doi.org/10.21105/joss.03021
23. Reback J, McKinney W, et al (2022) Pandas-dev/pandas: Pandas 1.4.3. 10.5281/zenodo.6702671
24. McKinney W (2010) Data structures for statistical computing in python. In: Proceedings of the 9th python in science conference, vol 1, pp 56–61. https://doi.org/10.25080/majora-92bf1922-00a
25. Mahynski NA (2022) pychemauth. https://github.com/mahynski/pychemauth. Accessed 9 Feb 2023
26. Kemsley EK, Defernez M, Marini F (2019) Multivariate statistics: considerations and confidences in food authenticity problems. Food Control 105:102–112. https://doi.org/10.1016/j.foodcont.2019.05.021
27. Rodionova OY, Titova AV, Pomerantsev AL (2016) Discriminant analysis is an inappropriate method of authentication. TrAC—Trends Analyt Chem 78:17–22. https://doi.org/10.1016/j.trac.2016.01.010
28. Rodionova OY, Oliveri P, Pomerantsev AL (2016) Rigorous and compliant approaches to one-class classification. Chemom Intell Lab Syst 159:89–96. https://doi.org/10.1016/j.chemolab.2016.10.002
29. Pomerantsev AL, Rodionova OY (2014) Concept and role of extreme objects in PCA/SIMCA. J Chemom 28:429–438. https://doi.org/10.1002/cem.2506
30. Pomerantsev AL (2008) Acceptance areas for multivariate classification derived by projection methods. J Chemom 22:601–609. https://doi.org/10.1002/cem.1147
31. Pomerantsev AL, Rodionova OY (2018) Multiclass partial least squares discriminant analysis: taking the right way—a critical tutorial. J Chemom 32:1–16. https://doi.org/10.1002/cem.3030
32. Breiman L (2001) Random forests. Mach Learn 45:5–32