SHAPCA: Consistent and Interpretable Explanations for Machine Learning Models on Spectroscopy Data

Mingxing Zhang[1,2,3[]], Nicola Rossberg[1,2[0009−0005−3883−5833]], Simone Innocente[1,4[0009−0002−2237−513X]], Katarzyna Komolibus[4[0000−0003−4545−6100]], Rekha Gautam[4[0000−0002−1176−8491]], Barry O'Sullivan[1,2,3[0000−0002−0090−2085]] Luca Longo[1,3[0000−0002−2718−5426]], and Andrea Visentin[1,2,3[0000−0003−3702−4826]]

[1] School of Computer Science and Information Technology, University College Cork, Ireland
[2] Centre for Research Training in Artificial Intelligence, University College Cork, Ireland
[3] Insight Centre for Data Analytics, University College Cork, Ireland
[4] Biophotonics@Tyndall, IPIC, Tyndall National Institute, Ireland

Abstract. In recent years, machine learning models have been increasingly applied to spectroscopic datasets for chemical and biomedical analysis. For their successful adoption, particularly in clinical and safety-critical settings, professionals and researchers must be able to understand and trust the reasoning behind model predictions. However, the inherently high dimensionality and strong collinearity of spectroscopy data pose a fundamental challenge to model explainability. These properties not only complicate model training but also undermine the stability and consistency of explanations, leading to fluctuations in feature importance across repeated training runs. Feature extraction techniques have been used to reduce the input dimensionality; these new features hinder the connection between the prediction and the original signal. This study proposes *SHAPCA*, an explainable machine learning pipeline that combines Principal Component Analysis (for dimensionality reduction) and Shapely Additive exPlanations (for post hoc explanation) to provide explanations in the original input space, which a practitioner can interpret and link back to the biological components. The proposed framework enables analysis from both global and local perspectives, revealing the spectral bands that drive overall model behaviour as well as the instance-specific features that influence individual predictions. Numerical analysis demonstrated the interpretability of the results and greater consistency across different runs.

Keywords: SHAP · Spectroscopy · Medical Diagnosis · PCA

1 Introduction

Spectroscopy techniques are broadly employed in the scientific and industrial domains, including biomedical research, chemistry, materials science, and envi-

ronmental monitoring. Techniques such as Raman, infrared, near-infrared, and nuclear magnetic resonance spectroscopy provide rich molecular fingerprint information, enabling non-destructive and label-free analysis of samples [8]. In recent years, the increased availability of high-resolution spectroscopy instruments has enabled the generation of large, complex datasets [24]. Machine learning techniques have been increasingly used to analyse and evaluate spectroscopic data for classification, regression, and anomaly detection [28]. In particular, supervised learning models have demonstrated outstanding performance in spectroscopic classification, effectively discriminating between different chemical compositions, material states, and disease conditions based on spectral features [16]. Looking forward, the deep integration of spectroscopy and machine learning is expected to play an increasingly important role in automated decision-making systems, diagnostics, and large-scale analytical pipelines [3].

Despite the strong predictive performance of machine learning models applied to spectroscopic data, their adoption in high-stakes domains remains limited, particularly in medical and clinical applications. In this high-risk field, the European Union Artificial Intelligence Act (EU AI Act) now imposes legally binding requirements for transparency, accountability, and meaningful human oversight in AI-driven decision-making systems [34]. At the same time, clinicians and healthcare practitioners must be able to understand and trust the reasoning behind algorithmic predictions in order to safely integrate such systems into clinical practice [38]. As a result, eXplainable Artificial Intelligence (XAI) has become a fundamental requirement for the responsible deployment of machine learning models, enabling predictions to be interpreted, validated, and justified in line with ethical and regulatory expectations.

However, achieving explainability in spectroscopic machine learning models is particularly challenging due to the intrinsic characteristics of spectroscopic data. Spectra are typically high-dimensional, comprising hundreds of features, and exhibit strong collinearity, as adjacent wavelengths originate from the same molecular vibrations and spectral peaks are inherently broad and overlapping [10]. As a result, information relevant to classification is often distributed across extended spectral regions rather than confined to isolated peaks [23]. These properties complicate both model training and post-hoc explanation: feature importance measures may become diluted across correlated variables and exhibit instability or inconsistency across repeated training runs [29]. Consequently, standard explainability techniques, when directly applied to raw spectroscopic features, may fail to yield reliable or physically meaningful interpretations. Feature extraction approaches, such as Principal Component Analysis (PCA), are widely used in spectroscopy to project the spectra, reducing correlation and noise. However, their components do not have a direct biological meaning, and they can limit the applicability of XAI approaches since they would explain the decisions in relation to the values in the reduced space.

To address the issues presented above, this paper introduces *SHAPCA*, a model-agnostic post-hoc explanation algorithm that captures spectral correlation structure by grouping highly correlated wavelengths into a low-dimensional

latent representation used by a classifier and reprojects their impact into the original input space, improving practitioners' understanding. Model training and classification are conducted in the latent space, reducing computational complexity while maintaining (or improving) predictive accuracy. We then compute Shapley Additive exPlanations (SHAP) values [17] at the component level to quantify the contribution of each latent component to the prediction. By performing explanation in the latent space, contributions are less susceptible to dilution across individual correlated features and exhibit improved stability across repeated runs. Finally, component-level contributions are back-projected to the original wavenumber axis to yield feature-space explanations that are directly interpretable and suitable for practical deployment. The explanation is superimposed over the original signal, highlighting which areas are relevant for the machine learning decision-making and if they are higher or lower than the other samples.

The structure of the paper is as follows: Section 2 positions this work in the relevant literature by providing an overview of the existing XAI approaches for spectroscopy data and of combinations of PCA and SHAP values for explainability. The datasets used, and the proposed approach are presented in Section 3. Section 4 presents and discusses the numerical analysis. Finally, Section 5 summarises the contributions and findings.

2 Literature Review

To contextualize this work, previous XAI applications to spectroscopic data are reviewed and evaluated. Additionally, the integration of PCA and SHAP in the domain of XAI is analysed.

2.1 XAI for spectroscopy

As the field of machine learning (ML) continues to grow, it is being increasingly integrated into the field of spectroscopy, permitting the automation of a range of predictive tasks. The applications of ML to spectroscopy have been especially impactful in fields such as chemistry [36,1,32], agriculture [31,9,21,25], and biomedicine [39,28,22,26]. These results demonstrate that machine learning models are capable of successfully classifying high-dimensional spectroscopy signals, demonstrating strong potential for the automation of a range of important tasks. However, one major concern with the implementation of ML models, especially in critical domains such as medicine, is their black-box nature and the associated inability to reliably and faithfully explain decision processes. In addition to being a policy requirement under the European Union AI act, the implementation of explainability is also vital to prevent problems such as model bias and shortcut learning and promote practitioner trust [33,14,20].

A wide range of XAI techniques have been proposed in the domain of ML, accounting for various data, model, and explanation types. However, in conjunction with spectroscopic signals, the implementation of XAI is challenging due

to the high-dimensional and intercorrelated nature of these signals. As a result, traditional explanatory methods frequently yield explanations that are inconsistent across model iterations and consequentially unreliable. A comprehensive review of existing explanatory approaches for spectroscopic signals is provided by Contreras et al. [5]. Their review indicates that relatively few XAI studies have been specifically developed for spectroscopy. Among the existing work, explanation methods can generally be categorized into four groups: activation-based, perturbation-based, gradient-based, and surrogate-based approaches, as well as various combinations of these techniques.

Rossberg et al. [27] proposed two model-specific XAI-based feature-selection methods for Raman spectroscopy, combining CNNs with Grad-CAM and Transformers with attention-derived importance scores. Their results showed that the CNN-Grad-CAM approach achieved the highest average accuracy, although no single method consistently outperformed all others across every scenario. Despite this strong predictive performance, both approaches are tied to specific model architectures, and the selected features remain mainly individual wavenumbers, which may limit physical interpretability in spectroscopy, where variables are highly correlated and meaningful information is often distributed across broader spectral regions rather than isolated features.

While Contreras et al. [6] proposed a spectral-zones-based SHAP and LIME framework for deep learning on spectral data, in which neighboring wavelengths are grouped into contiguous zones before perturbation so that the explanations emphasize interpretable spectral regions rather than single features. However, this strategy remains constrained by spatial adjacency: it can only group neighboring wavelengths and therefore cannot capture correlated but non-adjacent features that may originate from the same chemical component. As noted in the study, some biochemical signatures, such as fat-related signals, may appear in several separate spectral regions, meaning that chemically related but spatially distant variables are still treated independently.

In addition, [35] developed an ANN-based model to predict solution conductivity from plasma optical emission spectra and they extract the most dominent emission lines from each spectrum and employed LIME to identify influential ones for the model's predictions. Their results showed that OH, H_γ and H_β emission lines were critical to the model's decisions, and that these differed from the lines typically selected by human experts. This further highlights the value of grouping spectral features for explainability, but also suggests that more principled grouping strategies are needed.

Overall, these studies demonstrate that some form of feature grouping or segmentation is essential for explaining high-dimensional spectroscopy data. However, most of the existing approaches still rely mainly on architectural constraints or local spectral adjacency, and therefore may fail to capture the broader correlation structure of the data. In our study, we address this limitation by applying PCA to group correlated features into latent components, providing a more principled way to represent jointly varying spectral information and enabling ex-

planations at a level that is more robust and potentially more meaningful than isolated feature-wise attribution.

2.2 PCA for high-dimensional data

One approach to combat the high dimensionality and inter-correlation of spectroscopic data is the implementation of PCA. PCA generates new variables based on the directions of maximum variance within a dataset. As the generated components are perpendicular to each other within the feature space, they are uncorrelated and hence can be more reliably classified and explained through machine learning systems. In addition, the implementation of PCA for spectroscopic signals was found to reveal systematic variation and underlying patterns in complex spectroscopic datasets [37]. However, one drawback of the implementation of PCA is the associated loss of information linking the ML output to the original input variables. This, in turn, decreases transparency and hinders model implementation, as practitioners are unable to link explanations to spectral signatures. As such, the ability to decompose components back to the spectral inputs after ML classification would permit the successful integration of XAI techniques for spectroscopic data.

Some previous research has investigated the implementation of PCA in combination with XAI techniques in other fields that also have high-dimensional datasets, since there is not much similar work for spectroscopy data. Chen et al. [4] proposed an interpretable stacked ensemble framework for predicting SO_2 concentration in coal-fired boilers. After applying the Relief regression algorithm to select informative variables, they used Sparse PCA to reduce redundancy in the high-dimensional input features and constructed a base-model pool consisting of 17 machine learning models, which were then combined through stacking strategies. SHAP was subsequently employed to analyse the contribution of input variables to the model predictions. Although Sparse PCA and SHAP were both used within the framework, their integration was relatively loose, as Sparse PCA mainly served as a preprocessing step for dimensionality reduction rather than being directly coupled with the explanation stage. Nevertheless, the study still illustrates the respective advantages of Sparse PCA for handling high-dimensional data and SHAP for improving model interpretability. Other fields have successfully implemented PCA in conjunction with SHAP for applications such as the diagnosis of prostate cancer [2] and brain tumours [12].

In this study, for spectroscopy data, we extend these efforts by integrating Sparse PCA with SHAP and explicitly projecting all attribution scores back to the original wavenumber space. This design preserves the interpretability advantages of sparsity while enabling feature-level explanations that remain linked to physically meaningful spectral bands, allowing the importance of original spectral variables to be quantified in a stable and scientifically interpretable manner.

3 Methodology

This section describes the dataset used in this study, the proposed analysis pipeline, the back-projection steps applied, and the consistency-checking procedures used to evaluate the results. Together, these parts establish the study's logical and methodological foundation.

3.1 Data Preparation

We selected two datasets to evaluate the approach on both binary and multi-class classification tasks. To examine the behaviour of SHAP explanations in both settings, two datasets, collected using different spectroscopic modalities, are considered.

Raman Spectroscopy Dataset. This dataset was collected by [19] and consists of Raman spectroscopy measurements acquired from cheek mucosa tissue for cancer diagnosis. The data consists of 484 Raman spectra obtained from 77 patients, with each spectrum assigned a class label corresponding to the underlying tissue condition, resulting in three diagnostic classes: Potentially Malignant Healthy(PMH), Potentially Malignant Lesion (PML), and Healthy(H). Raman intensity values were recorded over the spectral range of 398.75–1998.75 cm^{-1}, forming high-dimensional feature vectors that capture the biochemical characteristics of the cheek mucosa tissue.

Prior to classification, the following preprocessing steps were taken:

- Samples labelled 'PMH' were excluded. Since this study focuses on distinguishing healthy from malignant cases, and PMH samples do not clearly fit into either category, this class was removed to maintain a clear binary classification.
- Cropped the spectral region below 411.25 cm^{-1}, as this range was dominated by noise.
- Applied Savitzky–Golay smoothing with a window length of 5 and a polynomial order of 2 along the spectral dimension.
- Performed baseline correction using the DRPLS algorithm with parameters $\lambda = 5 \times 10^5$ and $p = 0.003$.
- Applied maximum intensity normalization on each spectrum.
- The training and testing sets were split at the patient level to prevent data leakage.

DRS Dataset. This dataset was collected by [15] and comprises 5,215 spectra spanning six tissue classes commonly encountered in orthopaedic surgeries: Bone Cement (boneCement), Bone Marrow (boneMarrow), Cartilage (cartilage), Cortical Bone (cortBone), Muscle (muscle), and Trabecular Bone (traBone). Spectra were acquired over the range 355.016–1849.739 nm. Preprocessing was performed in accordance with the procedure described in the original study, prior to applying the proposed pipeline.

3.2 Pipeline

This study proposes an interpretable machine-learning pipeline for spectroscopic data that integrates PCA, supervised classification, and SHAP-based explainability. The procedure comprises the following steps:

1. PCA decomposition: the input spectrum is transformed into a low dimensional representation by projecting it through the PCA loadings W, yielding principal component (PC) values;
2. Supervised prediction: the PC values are used as inputs to a classifier to produce class predictions; in this work, we used Random Forest(RF) and Support Vector Classifiers (SVC);
3. SHAP explainability: Shapley values Φ are computed for the latent components to quantify their contributions to the model output;
4. Back-projection and visualisation: the Shapley values' contribution is reprojected into the original spectra as opacity to highlight the parts of the input that are more relevant to the decision. Their being higher or lower than the rest of the samples is driven by the PC values, which give the color red for higher and blue for lower. Combining and overlaying these two aspects onto the original spectral shape to yield physically interpretable, feature-level explanations.

Figure 1 shows a summary of the pipeline. The individual components and steps are described in the following Sections.

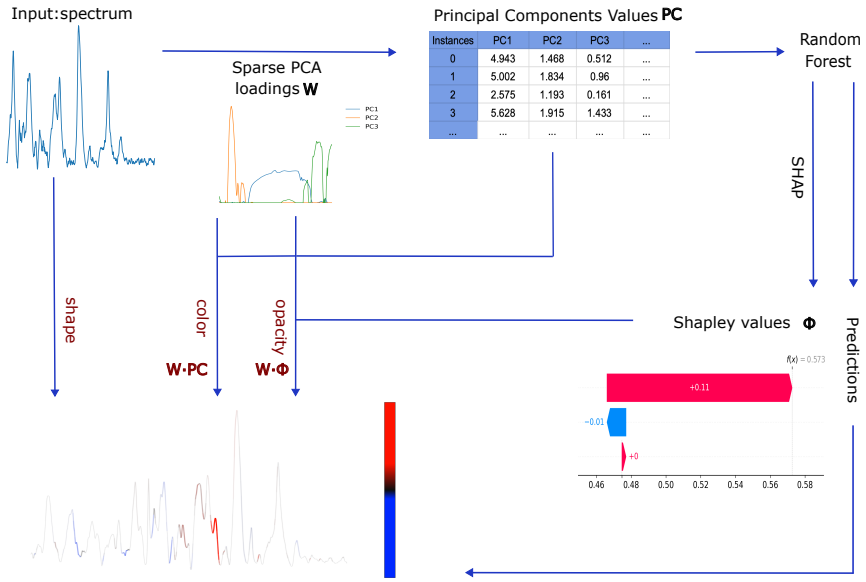

Fig. 1. Step-by-step workflow of the proposed interpretable spectroscopic pipeline.

Sparse PCA As discussed in Section 1, individual molecular vibrations influence a range of neighbouring wavenumbers, and multiple biochemical components contribute overlapping signals within the same spectral regions, resulting in strong collinearity among spectroscopic features. PCA is well-suited to capture this covariance structure because it identifies orthogonal directions in the data that explain the greatest variance. In spectroscopic data, groups of correlated variables often vary together and therefore contribute jointly to these dominant directions of variation. The resulting principal components offer an approximate representation of the main patterns present in the spectra, although they are derived by maximizing variance rather than explicitly modelling collinearity itself. Nevertheless, fully disentangling the latent structure underlying spectroscopic data is a highly challenging problem and is beyond the scope of the present study.

Given a dataset consisting of p original features (in this case, wavenumbers), PCA constructs a new set of orthogonal variables, referred to as principal components, which are ordered according to the amount of variance they explain. Each principal component is defined as a linear combination of the original features. Specifically, the k-th principal component can be written as

$$PC_k = \sum_{j=1}^{p} w_{jk} x_j, \tag{1}$$

where x_j denotes the j-th original feature and w_{jk} represents the corresponding loading. The vector of loadings

$$\mathbf{w}_k = (w_{1k}, w_{2k}, \ldots, w_{pk})^\top$$

defines the direction of the k-th principal component in the original feature space.

The loading vectors quantify the contribution of each original feature to a given component. Features with larger absolute loading values exert a stronger influence on the component, whereas features with near-zero loadings contribute minimally. As such, the loadings provide a direct link between the transformed representation and the original variables, enabling interpretation of principal components in terms of physically meaningful structures, such as spectral bands associated with underlying biochemical variations.

However, in standard PCA, each principal component is typically expressed as a linear combination of all original wavenumbers, even if many of the corresponding loadings have very small magnitudes. This results in dense components that are difficult to interpret, particularly for spectroscopic data, where diagnostically relevant information is often localised to specific biochemical bands rather than distributed across the entire spectrum.

To concentrate the explanations on the spectral regions most relevant to the model's decision-making process, we therefore employed Sparse Principal Component Analysis (Sparse PCA). Sparse PCA extends conventional PCA by imposing sparsity constraints on the component loadings, forcing many coefficients to be exactly zero while retaining non-zero contributions for only a limited subset

of wavenumbers. This sparsity property allows each component to be associated with a small number of informative spectral regions, rather than diffuse contributions across the full spectral range. Consequently, Sparse PCA yields more interpretable component representations in terms of localised spectral features or biochemical bands, while still preserving the ability to capture the dominant patterns of variation in the data.

Classification Sparse PCA component values are used as inputs to the downstream classifier. In our experiments, Random Forest and Support Vector Classifiers (SVC) are chosen.

Importantly, RF is particularly well-suited for SHAP-based interpretation. SHAP provides an exact and computationally efficient TreeExplainer for tree-based models, enabling stable and theoretically grounded Shapley value estimation. In contrast, SHAP explanations for non-tree-based classifiers, such as SVC, typically rely on model-agnostic explainers; we applied Kernel Explainer, which are computationally expensive and less stable in high-dimensional settings. Moreover, SVCs do not natively produce calibrated probability estimates, requiring additional probability calibration, which introduces further approximation and uncertainty. By contrast, Random Forests provide native probabilistic outputs through ensemble averaging, allowing SHAP values to be computed directly and consistently.

Train/test configuration and hyperparameters selection The Raman dataset was first split at the patient level into an 80/20 train–test partition to prevent information leakage across patients, whereas the DRS dataset was split at the sample level. Hyperparameter tuning of the pipeline was performed exclusively on the training portion using a two-stage strategy: an initial randomized search with cross-validation to identify stable regions of the hyperparameter space, followed by a grid search for fine-grained refinement. GroupKFold was used for the Raman dataset and stratified cross-validation was used for the DRS dataset. After tuning, the final model was retrained on the full training set and evaluated once on the held-out test split.

Because overly restrictive sparsity constraints in Sparse PCA can reduce classification performance, hyperparameters of both Sparse PCA and the downstream classifier were jointly optimized. Candidate models were evaluated based on predictive accuracy and the sparsity of the PCA components. Among configurations with comparable accuracy, preference was given to models exhibiting higher sparsity to enhance interpretability, thereby ensuring a balanced trade-off between classification performance and clarity of feature-level explanations. For both the Raman and DRS datasets, the selected Sparse PCA configuration retained 10 components, with the sparsity parameter $\alpha = 1$ for both the RF and SVC based pipelines.

SHAP After classification, all model predictions were provided to SHAP for explanation. SHAP is a post-hoc explainability method grounded in cooperative

game theory. It quantifies the contribution of each feature (in this study, each latent component) to a model's prediction by computing its average marginal contribution across all possible feature subsets [17]. Under the additive explanation framework, the prediction for an instance x is decomposed as

$$f(x) = \phi_0 + \sum_{k=1}^{K} \phi_k(PC_k), \tag{2}$$

where K is the number of components, $\phi_k(PC_k)$ denotes the contribution of component k, and ϕ_0 is a baseline term, commonly defined as the expected model output over a reference (background) distribution:

$$\phi_0 = \mathbb{E}[f(X)]. \tag{3}$$

Shapley values represent additive, sample-wise contributions defined with respect to a common baseline, which makes them comparable across samples and amenable to meaningful aggregation. This property distinguishes SHAP from other attribution methods, such as gradient-based saliency maps or attention weights, whose values are often sensitive to input scaling, lack a consistent reference across samples, and therefore do not support semantically reliable aggregation.[18].

Importantly, for the DRS dataset, which involves a multi-class classification task, component attributions are inherently class-conditional, as the model learns distinct decision functions for each output class. Accordingly, SHAP computes component contributions separately for each class, yielding class-specific Shapley values that quantify how each component influences different class predictions. As a result, global explanations must be derived in a class-wise manner to preserve semantic meaning. Aggregating attributions across classes would mix contributions that correspond to different class outputs, thereby blurring or even cancelling class-specific evidence and obscuring the distinct decision logic learned for each class. By contrast, aggregating SHAP values within each class preserves the class-conditional semantics of the attributions and yields global patterns that remain interpretable and directly aligned with the multi-class classification objective.

3.3 Visualization

The goal of the visualisation is to enable practitioners to interpret the results in terms of biologically meaningful spectral regions. In particular, the visualisation aims to answer two key questions of practical relevance: which regions of the spectrum contribute to the model's decision, and whether these regions exhibit higher or lower intensity than expected.

To this end, two component-level quantities are back-projected to the original feature space: (i) component-wise Shapley values, which are mapped to opacity (alpha) to indicate the strength of contribution to the prediction, and (ii) component values, which are encoded by colour to represent their relative magnitude

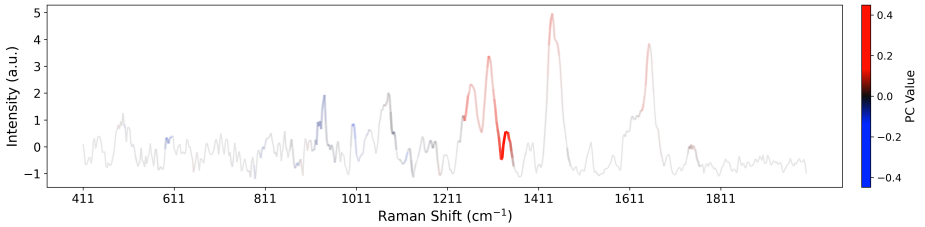

Fig. 2. A Raman spectroscopy sample. Opacity indicates the most decisive spectral regions for the prediction, while colour encodes intensity relative to the dataset average (red = higher, blue = lower).

across classes. This joint encoding enables simultaneous assessment of where the model draws evidence from and whether the corresponding spectral features are higher or lower than expected. Figure 2 shows an example of a spectra explanation. In this example, the peak around 1350 Raman shift is the most impactful for the classification, and it is higher than normal.

Global Explanation Global explanation aims to characterize the overall decision behaviour of the trained model by identifying spectral regions that consistently contribute to class discrimination across the dataset. Rather than focusing on individual samples, this analysis aggregates model explanations at the class level, thereby reducing the influence of sample-specific noise and highlighting stable, representative patterns. As shown in Algorithm 1, for each class, SHAP values are extracted and averaged across samples predicted as that class to obtain a mean component-level contribution vector $\bar{\boldsymbol{\phi}}^{(c)}$. These component-wise importance scores are then back-projected to the original wavenumber space using the absolute Sparse PCA loadings $\mathbf{W}$, yielding a class-specific feature importance vector $\boldsymbol{\Psi}^{(c)}$:

$$\boldsymbol{\Psi}^{(c)} \;=\; \left(\bar{\boldsymbol{\phi}}^{(c)}\right)^{\top} W. \tag{4}$$

In Figure 1, this encodes the opacity (transparency) of the explanation superimposed on the original spectra.

As for component values, they are first normalized across samples and then averaged within each predicted class to obtain class-wise normalized component profiles $\bar{\boldsymbol{pc}}^{(c)}$. These profiles are subsequently projected back to the original feature space using the Sparse PCA loadings $\mathbf{W}$, and the contributions from all normalised components are summed to yield a class-specific projection vector $\boldsymbol{PC}^{(c)}$:

$$\boldsymbol{PC}^{(c)} = \sum_{k=1}^{K} \left(\bar{\boldsymbol{pc}}^{(c)}\right)_k \mathbf{W}_k \tag{5}$$

Algorithm 1 Class-wise SHAP Aggregation and Back-Projection to Feature Space

Input:
$\boldsymbol{\Phi} \in \mathbb{R}^{N \times K \times C}$: SHAP values
(N: number of samples, K: number of components, C: number of classes)
$\hat{\mathbf{y}} \in \{1, \ldots, C\}^N$: predicted class label for each sample
$\mathbf{W} \in \mathbb{R}^{K \times P}$: Sparse PCA loadings
(P: number of original features / wavenumbers)
Output:
$\boldsymbol{\Psi}^{(c)} \in \mathbb{R}^P$: feature-level importance vector for each class c

for each class $c \in \{1, \ldots, C\}$ **do**
  Extract SHAP values for class c: $\boldsymbol{\Phi}^{(c)} \in \mathbb{R}^{N \times K}$
  Select samples predicted as class c
  Compute the mean SHAP value of each component over these samples: $\bar{\boldsymbol{\phi}}^{(c)} \in \mathbb{R}^K$
  Initialise feature-level importance vector: $\boldsymbol{\Psi}^{(c)} \leftarrow \mathbf{0} \in \mathbb{R}^P$
  for each component k in $\{1, \ldots, K\}$ **do**
    Distribute component importance to original features: $\boldsymbol{\Psi}^{(c)} \leftarrow \boldsymbol{\Psi}^{(c)} + \bar{\phi}_k^{(c)} \cdot |\mathbf{W}_k|$
  end for
end for
return $\{\boldsymbol{\Psi}^{(c)}\}_{c=1}^{C}$

This highlights whether the specific part of the spectra encoded by the component is higher or lower compared to the other datapoints. The information is encoded through color (Figure 1) with a gradient similar to SHAP beeswarm plots.

Local Explanation To explain a given spectrum for a target class, the component-level SHAP values were separated into positive and negative parts to preserve their signs. The approach is summarised in Algorithm 2. Positive contributions correspond to components that support the class prediction, whereas negative contributions indicate suppressive effects. These two parts were then independently back-projected to the original feature space using the absolute Sparse PCA loadings $\mathbf{W}$, producing separate spectral contribution maps, $\boldsymbol{\Psi}_i^+$ and $\boldsymbol{\Psi}_i^-$, that highlight class-supporting and class-opposing evidence, respectively.

$$\boldsymbol{\Psi}_i^+ = \left(\boldsymbol{\phi}_i^+\right)^\top \mathbf{W}, \qquad \boldsymbol{\Psi}_i^- = \left(\boldsymbol{\phi}_i^-\right)^\top \mathbf{W}. \tag{6}$$

where i denotes the i-th instance. $\boldsymbol{\Psi}_i^+$ represents the features that are supporting the class prediction of instance i, while $\boldsymbol{\Psi}_i^-$ the features of the negative contribution.

For the component values, a similar back-projection procedure to that used in the global explanation was applied. Specifically, the normalised component

Algorithm 2 Sample-wise SHAP and Back-projection to Feature Space

Input:
SHAP values $\boldsymbol{\Phi}$
spectrum index i
spectrum predicted class c
Sparse PCA loadings $\mathbf{W}$
Output:
Feature-level contribution maps $\boldsymbol{\Psi}^+, \boldsymbol{\Psi}^-$

Extract component-level SHAP values for spectrum i and class c:
$\boldsymbol{\phi}_i^{(c)} \in \mathbb{R}^K$
Separate positive and negative SHAP contributions:

$$\phi_{i,k}^+ = \begin{cases} \phi_{i,k}^{(c)}, & \phi_{i,k}^{(c)} > 0 \\ 0, & \text{otherwise} \end{cases}$$

$$\phi_{i,k}^- = \begin{cases} \phi_{i,k}^{(c)}, & \phi_{i,k}^{(c)} < 0 \\ 0, & \text{otherwise} \end{cases}$$

Initialise feature-level intensity vectors:
$\boldsymbol{\Psi}^+, \boldsymbol{\Psi}^- \leftarrow \mathbf{0}$
for each component k in $\{1, \ldots, K\}$ **do**
    $\boldsymbol{\Psi}^+ \leftarrow \boldsymbol{\Psi}^+ + \phi_{i,k}^+ \cdot |\mathbf{W}_k|$
    $\boldsymbol{\Psi}^- \leftarrow \boldsymbol{\Psi}^- + \phi_{i,k}^- \cdot |\mathbf{W}_k|$
end for
return feature-level positive and negative contribution maps $\boldsymbol{\Psi}^+$ and $\boldsymbol{\Psi}^-$

values of the target instance $\boldsymbol{pc}_k$ were multiplied by the corresponding Sparse PCA loadings $\mathbf{W}_k$ and summed, thereby projecting the instance-level component representation back to the original feature space:

$$\mathbf{PC} = \sum_{k=1}^{K} \boldsymbol{pc}_k \, \mathbf{W}_k \tag{7}$$

where $\boldsymbol{PC}$ encodes the color that relate spectrum i to the other spectra. The color component is used both for the positive and negative explanations.

3.4 Consistency Check

To assess the stability and consistency of the explanation methods, both the baseline pipeline (SHAP) and the proposed pipeline (SHAPCA) were trained under a five-fold cross-validation protocol, yielding five independently trained models per method. For each trained model, global and local explanations were obtained by back-projecting the corresponding Shapley values to the original spectral domain and evaluating these projections on an identical held-out test set. Explanation stability was measured using pairwise similarity and correlation between the five SHAP models and between the five SHAPCA models, for both global and local projections. The resulting consistency scores were then averaged within each method and compared between SHAP and SHAPCA to determine

whether the proposed pipeline produces more reproducible explanations across folds. This approach has been inspired by [11].

Cosine similarity and Pearson correlation are adopted in this study. Cosine similarity quantifies the directional alignment between two vectors and is therefore primarily sensitive to the overall pattern (i.e., the relative distribution of peaks and valleys) of the back-projected Shapley vectors rather than their absolute magnitudes. For two back-projected Shapley vectors $\boldsymbol{\Psi}_m$ and $\boldsymbol{\Psi}_n$, cosine similarity is defined as

$$\cos\_\mathrm{sim}(\boldsymbol{\Psi}_m, \boldsymbol{\Psi}_n) = \frac{\boldsymbol{\Psi}_m^\top \boldsymbol{\Psi}_n}{\|\boldsymbol{\Psi}_m\|_2 \, \|\boldsymbol{\Psi}_n\|_2}, \tag{8}$$

where $\boldsymbol{\Psi}_m^\top \boldsymbol{\Psi}_n$ denotes the dot product and $\|\cdot\|_2$ denotes the Euclidean norm. By construction, $\cos\_\mathrm{sim}(\boldsymbol{\Psi}_m, \boldsymbol{\Psi}_n) \in [-1, 1]$, with values close to 1 indicating highly aligned explanation patterns, values near 0 indicating weak alignment, and values close to -1 indicating opposing patterns.

Pearson correlation complements cosine similarity by measuring how strongly two explanation vectors co-vary after removing their mean offsets. In practice, this indicates whether two back-projected attribution maps exhibit peaks and valleys at consistent spectral locations, independent of any constant baseline shift. This is particularly relevant for spectroscopy explanations because back-projected Shapley vectors obtained from different folds may differ by an additive offset (or overall centering) even when the underlying discriminative pattern is preserved. For two back-projected Shapley vectors $\boldsymbol{\Psi}_m$ and $\boldsymbol{\Psi}_n$ with p spectral features, Pearson correlation is defined as

$$\rho(\boldsymbol{\Psi}_m, \boldsymbol{\Psi}_n) = \frac{\sum_{i=1}^{p} \left(\Psi_{m,i} - \bar{\Psi}_m\right)\left(\Psi_{n,i} - \bar{\Psi}_n\right)}{\sqrt{\sum_{i=1}^{p} \left(\Psi_{m,i} - \bar{\Psi}_m\right)^2} \sqrt{\sum_{i=1}^{p} \left(\Psi_{n,i} - \bar{\Psi}_n\right)^2}}, \tag{9}$$

where $\Psi_{m,i}$ denotes the i-th element of $\boldsymbol{\Psi}_m$, and $\bar{\Psi}_m = \frac{1}{p}\sum_{i=1}^{p} \Psi_{m,i}$ (analogously, $\bar{\Psi}_n$) is the mean of the corresponding explanation vector. The coefficient satisfies $\rho \in [-1, 1]$, where values close to 1 indicate strong agreement in the mean-centred spectral pattern, values near 0 indicate weak linear association, and negative values indicate opposing patterns. Reporting both cosine similarity and Pearson correlation, therefore, provides a more complete assessment of explanation stability.

Distance-based metrics such as Euclidean distance have been used in related work [13]; however, they are not well suited to the present setting, since the SHAP provides $\boldsymbol{\Psi}$ values for a limited number of features, their average intensity is low and many values are 0, leading to smaller errors overall.

4 Results and Discussion

This section presents the pipeline performance across the two datasets. In each application, we computed the global and local explanations and investigated

Table 1. Mean test accuracy and F1 score for Raman spectroscopy and DRS data, reported as mean ± 95% confidence interval.

Pipeline	Raman spectroscopy		DRS	
	Accuracy	**F1 Score**	**Accuracy**	**F1 Score**
Sparse PCA + SVC	0.681 ± 0.023	0.663 ± 0.023	0.972 ± 0.002	0.976 ± 0.002
SVC	0.757 ± 0.025	0.735 ± 0.028	0.993 ± 0.001	0.994 ± 0.001
Sparse PCA + RF	0.723 ± 0.024	0.705 ± 0.026	0.936 ± 0.003	0.944 ± 0.002
RF	0.740 ± 0.026	0.725 ± 0.028	0.945 ± 0.003	0.952 ± 0.003

the biological components associated with them. Then, the consistency of the explanation across different iterations for SHAPCA and the baseline SHAP is analysed. For the sake of reproducibility, the code used in this study is available online [5].

Table 1 summarises the average predictive performance on both datasets. Each pipeline was evaluated over 100 independent runs, and the results are reported in terms of mean test accuracy and F1 score, together with their 95% confidence intervals. Overall, across both datasets, introducing Sparse PCA led to a decrease in both accuracy and F1 score compared with using the original feature space. Although the two classifiers performed at a broadly similar level on both datasets, RF was selected as the reference model for the global and local explanation visualisations because it enables the use of TreeExplainer and yielded more stable explanations across repeated runs, as shown later in the consistency analysis.

4.1 Raman Spectroscopy Dataset

Global Explanation Figure 3 shows the global explanations for the binary Raman spectroscopy classification task. The grey curve corresponds to the mean Raman spectrum of each class, while the colour encodes the average normalised principal component PC values of the corresponding class projected back to the original wavenumber space. In addition, the opacity represents the averaged Shapley values, also back-projected to the original wavenumbers.

From the global explanation, several spectral regions consistently exhibit strong contributions to the model's classification decisions, notably around 600–670 cm^{-1}, 860–900 cm^{-1}, 1010–1030 cm^{-1}, 1110–1130 cm^{-1}, 1310–1360 cm^{-1}. These regions therefore constitute the most decisive spectral features at the class level.

For the H class, the peaks in the 1310–1360 cm^{-1} region, which are commonly associated with lipids and nucleic acids, are notably higher than the overall average spectrum, whereas the spectral intensities in the remaining regions 600–670 cm^{-1}, 860–900 cm^{-1}, 1010–1030 cm^{-1}, and 1110–1130 cm^{-1} are consistently lower than the average. In other words, these characteristic patterns serve as a

[5] https://github.com/appleeye007/SHAPCA

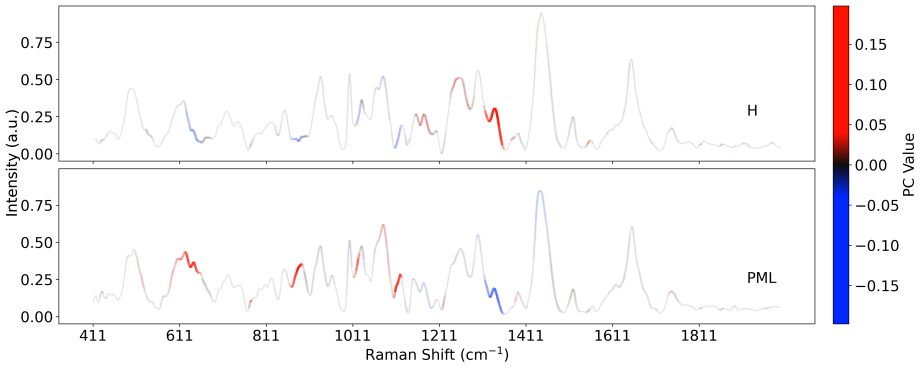

Fig. 3. Global explanations for the binary Raman spectroscopy task (classes H and PML). The grey curves show the class-mean spectra.

key discriminative signal that the model associates with the Healthy class. In contrast, the class PML exhibits an essentially reversed trend. Consequently, when the model observes this opposite spectral pattern, it is more likely to classify the sample as PML.

Local Explanation Figure 4 illustrates the local explanation for a single instance in the binary Raman spectroscopy classification task. The instance is labeled as PML and is correctly predicted as PML by the model. Color encodes the normalized PC value of this instance, while opacity encodes the back-projected Shapley value, which is separated into positive and negative.

In this example, elevated intensities around 620–640 cm^{-1}, 860–880 cm^{-1}, and 1110–1130 cm^{-1} provide positive evidence for the PML prediction, while reduced intensity in 1310–1360 cm^{-1} and 1450–1500 cm^{-1} further supports this classification. In contrast, lower-than-average intensities in 640–670 cm^{-1}, 880–910 cm^{-1} and 1020–1030 cm^{-1} act against the PML label, as these patterns are more consistent with an H-like spectral profile. Overall, the model integrates both supporting and opposing evidence; here, the PML-supporting regions dominate, leading to a PML classification. Notably, the directions of these local deviations align with the class-level patterns identified in the global explanation.

Consistency Tables 2 and 3 summarise explanation reproducibility for SHAPCA and baseline SHAP using two complementary measures: cosine-based consistency and Pearson correlation. Table 2 reports the global, class-level results and shows that SHAPCA achieves higher agreement than baseline SHAP across classes under both metrics. The SVC-based results are also markedly lower than the RF-based results, which is likely attributable to the explainer used in each case: TreeExplainer leverages the tree structure to compute SHAP values efficiently and more stably for RF, whereas KernelExplainer for SVC relies on a sampling-based approximation with limited background data as the reference distribution,

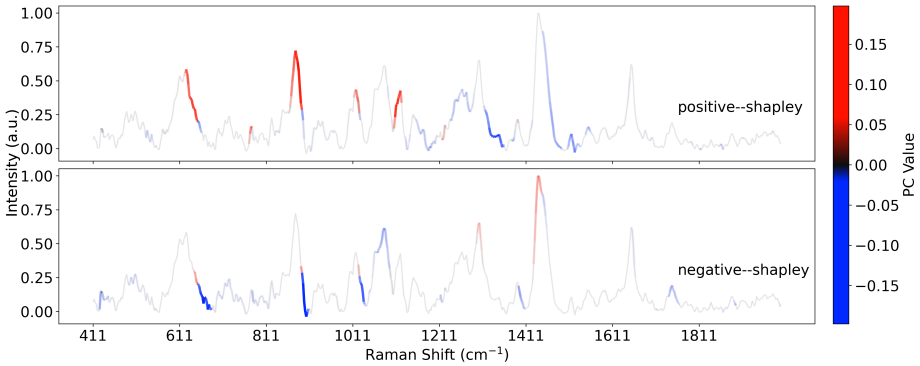

Fig. 4. Local explanation for a binary classification instance correctly predicted as PML. The grey curve represents the Raman spectrum of the instance.

making the resulting attributions more sensitive to sampling variation and considerably more expensive to compute. Table 3 reports the local, instance-level results; as expected, local explanations are less reproducible than global explanations because single-sample attributions are more sensitive to noise and local decision-boundary geometry. Even so, SHAPCA continues to outperform baseline SHAP under both metrics. In particular, the local explanations produced by SVC show extremely low consistency, in some cases even yielding negative values. One possible reason is that the selected regularization setting is not well suited for explanation stability. Although the SVC hyperparameters were tuned to achieve the best predictive accuracy, the resulting model may be suboptimal from an interpretability perspective, leading to unstable local explanations across runs. Overall, these results indicate that Sparse PCA improves the stability of the resulting explanations, and that RF with TreeExplainer provides a more computationally efficient and reproducible explanation framework than SVC with KernelExplainer.

Table 2. Comparison of global explanation consistency and correlation across classes for Raman spectroscopy data.

Class	SHAPCA				SHAP			
	RF+Sparse PCA		SVC+Sparse PCA		RF		SVC	
	Cosine	Correlation	Cosine	Correlation	Cosine	Correlation	Cosine	Correlation
H	0.820	0.811	0.458	0.406	0.715	0.698	0.370	0.333
PML	0.842	0.835	0.567	0.517	0.759	0.747	0.437	0.403

Table 3. Comparison of local explanation consistency and correlation for Raman spectroscopy data.

	SHAPCA				SHAP			
	RF+Sparse PCA		SVC+Sparse PCA		RF		SVC	
	Cosine	Correlation	Cosine	Correlation	Cosine	Correlation	Cosine	Correlation
Local	0.622	0.615	0.242	0.239	0.573	0.568	-0.068	-0.060

4.2 DRS Dataset

Global Explanation Figure 5 presents global explanations for the six classes. The following references were used as guidelines. For DRS, [40,7,30] were consulted. Notably, wavelength bands that contribute most to classification performance largely coincide with known absorption features of biological tissue chromophores, including hemoglobin, water, lipids, and collagen.

For Bone Cement, high feature importance is observed around 400 nm and 650 nm. The band at 400 nm is likely informative because the reflectance profile in this region highlights the absence of hemoglobin, which is otherwise present in all investigated biological tissues. In contrast, the band near 650 nm is likely due to artificial pigments added to improve its visibility during surgery. This valley is clearly not present in all the biological components.

In the case of Bone Marrow, the model primarily relies on wavelength bands around 1200 nm and 1700 nm, consistent with reflectance characteristics associated with lipid absorption, reflecting the high lipid content of this tissue. Cartilage classification is dominated by the shorter wavelenghts, where its structure shows lower absorption.

For Cortical Bone, several wavelength bands contribute to classification; however, the region around 450 nm is of particular interest, as it aligns with a reflectance dip linked to the presence of oxygenated hemoglobin. Muscle tissue shows high feature importance around 420 nm relative to the shoulder of the haemoglobin peak, and the oxygenation status seems important, as the slope of the oxy and deoxy hemoglobin absorption bands in this spectral region is different, so it might differentiate more based on the shoulder. A band in 1700 nm highlighting the absence of a lipid relative to other similar tissue types

Finally, Trabecular Bone classification predominantly relies on features between 1400 nm, 1500 nm, and 1700 nm, coinciding with a reflectance minimum primarily attributable to elevated concentrations of water and lipids.

Overall, these results indicate that while many classification-relevant features correspond to absorption characteristics of key biological components and tissue microstructure, others are selected because of their distinctive spectral contrast relative to the remaining classes.

Local Explanation Figure 6 presents a local explanation for a single instance that is correctly classified as cortBone. The grey curve represents the DRS spectrum of the instance.

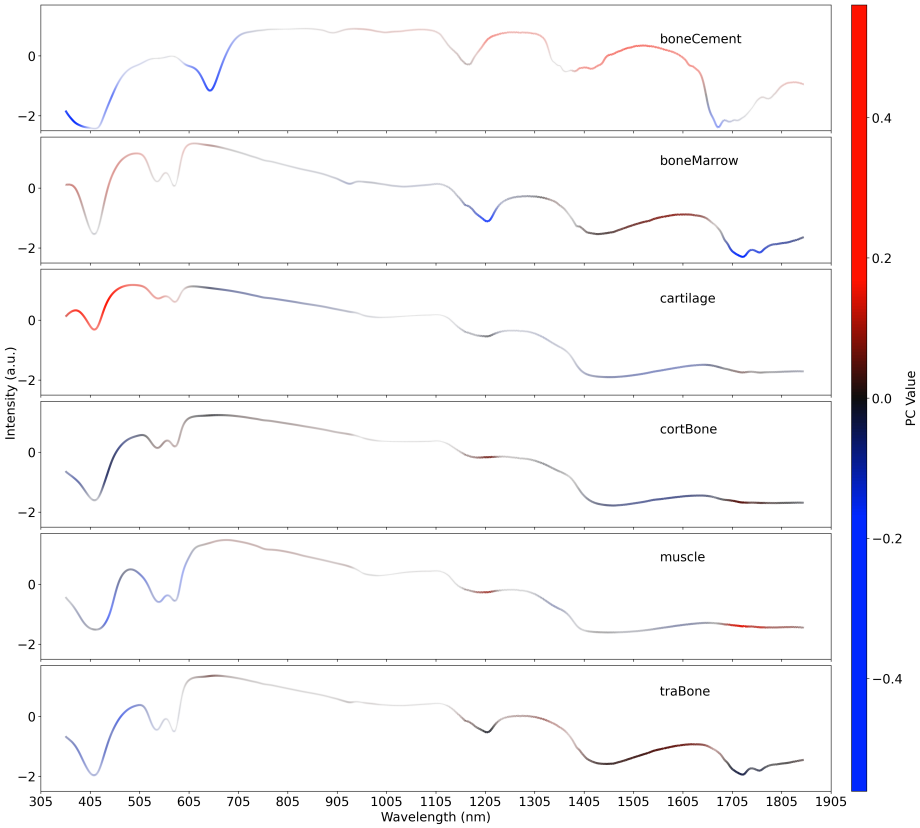

Fig. 5. Global explanation of DRS multiclass classification across all classes. The grey curve shows the mean spectrum of each class.

Evidence supporting the cortBone prediction includes reduced intensity in the 350–500 nm region and elevated intensity in the mid-range (705–905 nm). In contrast, the dominant negative evidence is concentrated around 505–650 nm (blue with high opacity), indicating bands where the instance deviates toward a more muscle-like signature and thus lowers the cortBone score. Overall, the decision is governed by the strongly supportive cortBone-aligned regions; the competing short-wavelength evidence is insufficient to offset them, and the instance is therefore classified as cortBone.

Consistency Across repeated splits, SHAPCA yields substantially more reproducible explanations than baseline SHAP under both cosine similarity and Pearson correlation. For global explanations, the RF + Sparse PCA pipeline achieves consistently very high agreement across all classes, with both metrics remaining close to 1, indicating near-perfect stability. SVC + Sparse PCA is generally less stable than RF + Sparse PCA, but it still outperforms baseline SHAP in most

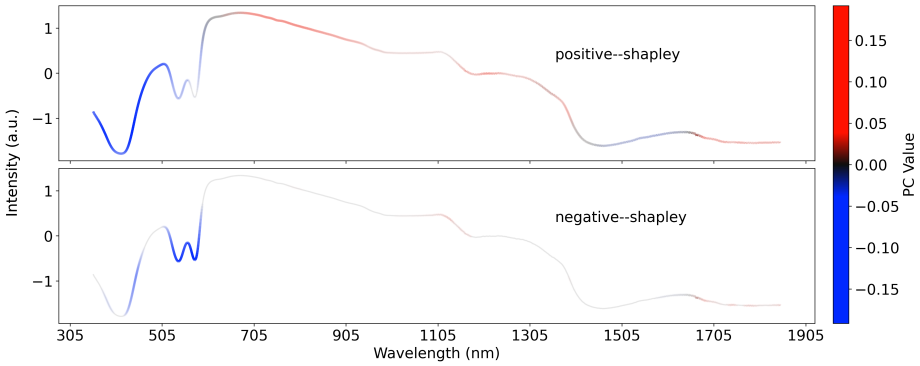

Fig. 6. Local explanation of multi-classification for an instance correctly predicted as cortBone

classes. By contrast, the baseline SHAP results for RF are markedly lower and more variable, while SHAP with SVC shows the weakest overall reproducibility, despite moderate performance for a few classes such as CortBone. A similar pattern is observed for local explanations, where SHAPCA again produces substantially higher consistency than SHAP, especially for RF + Sparse PCA. Taken together, the improvements in both cosine similarity and correlation suggest that combining Sparse PCA with SHAP leads to more stable explanations across repeated splits, likely because grouping correlated spectral variables into components reduces sensitivity to sampling variation. In addition, the tree-based RF setting appears more robust than SVC in both the SHAPCA and baseline SHAP frameworks.

The consistency checks for the DRS dataset yield stronger results than those for the Raman dataset. DRS spectra have wider peaks, leading to higher collinearity between the features. In addition, the DRS dataset is considerably larger, which stabilizes model training and reduces sampling noise. As a result, when Sparse PCA estimates component directions, the extracted components are more reproducible across different random seeds and resampling runs. This improved stability propagates through the pipeline, leading to more consistent SHAP attributions and more reliable back-projected explanations. This is particularly clear in the BoneCement, where the sample does not have the molecular variability of live tissue. In this case, the sample is considerably more reflective, providing a signature that is very different than the other classes. Consequently, many parts of the spectra can be used to discriminate it, making the SHAP explanations of the individual wavelengths quite aleatory, leading to poor consistency.

Table 4. Comparison of global explanation consistency and correlation across classes for DRS data.

Class	SHAPCA				SHAP			
	RF+Sparse PCA		SVC+Sparse PCA		RF		SVC	
	Cosine	Corr.	Cosine	Corr.	Cosine	Corr.	Cosine	Corr.
Muscle	0.992	0.984	0.784	0.787	0.694	0.679	0.340	0.280
CortBone	0.988	0.948	0.792	0.706	0.577	0.576	0.751	0.748
TraBone	0.990	0.976	0.842	0.735	0.729	0.694	0.620	0.607
Cartilage	0.981	0.963	0.942	0.950	0.648	0.647	0.387	0.370
BoneMarrow	0.993	0.988	0.770	0.647	0.640	0.636	0.422	0.409
BoneCement	0.971	0.955	0.978	0.957	0.334	0.300	0.150	0.045

Table 5. Comparison of local explanation consistency and correlation for DRS data.

	SHAPCA				SHAP			
	RF+Sparse PCA		SVC+Sparse PCA		RF		SVC	
	Cosine	Corr.	Cosine	Corr.	Cosine	Corr.	Cosine	Corr.
Local	0.982	0.977	0.459	0.631	0.691	0.685	0.371	0.367

5 Conclusion

This work addressed the challenge of producing explanations that are stable and understandable by practitioners for machine learning models applied to spectroscopic data, where high dimensionality and strong feature collinearity often undermine the reliability of standard post-hoc explanation methods. To this end, we proposed SHAPCA, an interpretable pipeline that combines PCA with SHAP to generate feature-level explanations grounded in the underlying spectral structure.

By grouping correlated wavenumbers into sparse components, SHAPCA reduces redundancy and stabilises both model training and explanation. Performing SHAP attribution in the latent space and subsequently back-projecting contributions to the original spectral axis preserves interpretability while improving consistency across repeated runs. Experiments on Raman and DRS datasets showed that the proposed approach yields explanations that align with known biochemical and tissue-specific spectral features, while consistently outperforming baseline SHAP in terms of explanation stability, particularly at the global level.

The proposed visualisation, which jointly encodes attribution strength and relative spectral intensity, enables intuitive interpretation in terms of contiguous spectral bands rather than isolated variables, facilitating practical use by domain experts. Overall, SHAPCA provides a simple strategy for improving the robustness and interpretability of explanations in spectroscopic machine learning, supporting more trustworthy deployment in biomedical and other safety-critical applications. Future work will explore the extension of SHAPCA to other dimensionality-reduction techniques, as well as its application to longitudinal and multimodal spectroscopic datasets and other 1-D classification tasks.

Acknowledgments. This work was conducted with the financial support of Research Ireland - Taighde Éireann, under Grant Nos. 18/CRT/6223 and 12/RC/2289-P2, the latter being co-funded under the European Regional Development Fund. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Anker, A.S., Butler, K.T., Selvan, R., Jensen, K.M.: Machine learning for analysis of experimental scattering and spectroscopy data in materials chemistry. Chemical Science **14**(48), 14003–14019 (2023)
2. B. Gavade, A., B. Nerli, R., A. Gavade, P., Kumar, M., Mehta, U.: Innovative prostate cancer classification: Merging autoencoders, pca, shap, and machine learning techniques. In: International Conference on Advanced Robotics, Control, and Artificial Intelligence. pp. 375–390. Springer (2024)
3. Bertazioli, D., Piazza, M., Carlomagno, C., Gualerzi, A., Bedoni, M., Messina, E.: An integrated computational pipeline for machine learning-driven diagnosis based on raman spectra of saliva samples. Computers in Biology and Medicine **171**, 108028 (2024)
4. Chen, Y., Wang, Q., Xie, G., Tian, Z., Zhang, B., Yue, J.: Sparse principal component analysis and shap-based explainable framework for so2 concentration prediction: A multi-method stacked ensemble model. Journal of Environmental Chemical Engineering p. 118330 (2025)
5. Contreras, J., Bocklitz, T.: Explainable artificial intelligence for spectroscopy data: a review. Pflügers Archiv-European Journal of Physiology **477**(4), 603–615 (2025)
6. Contreras, J., Winterfeld, A., Popp, J., Bocklitz, T.: Spectral zones-based shap/lime: Enhancing interpretability in spectral deep learning models through grouped feature analysis. Analytical Chemistry **96**(39), 15588–15597 (2024)
7. Dasa, M.K., Markos, C., Maria, M., Petersen, C.R., Moselund, P.M., Bang, O.: High-pulse energy supercontinuum laser for high-resolution spectroscopic photoacoustic imaging of lipids in the 1650-1850 nm region. Biomedical optics express **9**(4), 1762–1770 (2018)
8. Esmonde-White, K.A., Cuellar, M., Lewis, I.R.: The role of raman spectroscopy in biopharmaceuticals from development to manufacturing. Analytical and Bioanalytical Chemistry **414**(2), 969–991 (2022)

9. Farber, C., Kurouski, D.: Raman spectroscopy and machine learning for agricultural applications: chemometric assessment of spectroscopic signatures of plants as the essential step toward digital farming. Frontiers in Plant Science **13**, 887511 (2022)
10. Guo, K., Shen, Y., Gonzalez-Montiel, G.A., Huang, Y., Zhou, Y., Surve, M., Guo, Z., Das, P., Chawla, N.V., Wiest, O., Zhang, X.: Artificial intelligence in spectroscopy: Advancing chemistry from prediction to generation and beyond (2025), https://arxiv.org/abs/2502.09897
11. Gwinner, F., Tomitza, C., Winkelmann, A.: Comparing expert systems and their explainability through similarity. Decision Support Systems **182**, 114248 (2024)
12. Happila, T., Simbu, M., Rajendran, A., Ranjith Kumar, P., Rajakumar, S., Hariprakash, P.: Enhancing brain tumor diagnosis: A hybrid machine learning model with pca and shap for interpretability. In: 2025 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI). pp. 1–7. IEEE (2025)
13. Karagoz, G., Özçelebi, T., Meratnia, N.: Systematic benchmarking of local and global explainable ai methods for tabular healthcare data. In: Guidotti, R., Schmid, U., Longo, L. (eds.) Explainable Artificial Intelligence: Third World Conference, xAI 2025, Istanbul, Turkey, July 9–11, 2025, Proceedings, Part I. Communications in Computer and Information Science, vol. 2576, pp. 337–358. Springer, Cham (2025). https://doi.org/10.1007/978-3-032-08317-3_16
14. Kästner, L., Langer, M., Lazar, V., Schomäcker, A., Speith, T., Sterz, S.: On the relation of trust and explainability: Why to engineer for trustworthiness. In: 2021 IEEE 29th international requirements engineering conference workshops (REW). pp. 169–175. IEEE (2021)
15. Li, C.L., Fisher, C.J., Komolibus, K., Lu, H., Burke, R., Visentin, A., Andersson-Engels, S.: Extended-wavelength diffuse reflectance spectroscopy dataset of animal tissues for bone-related biomedical applications. Scientific Data **11**(1), 136 (2024)
16. Liu, Y., Wu, Y., Wang, J., Qi, J., Zhou, C., Xue, Y.: Recent advances in raman spectral classification with machine learning. Sensors **26**(1), 341 (2026)
17. Lundberg, S., Lee, S.I.: A unified approach to interpreting model predictions (2017), https://arxiv.org/abs/1705.07874
18. Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.I.: From local explanations to global understanding with explainable ai for trees. Nature machine intelligence **2**(1), 56–67 (2020)
19. Maryam, S., Benazza, A., Fahy, E., Sekar, S.K.V., US, D., Olivo, M., Riordain, R.N., Andersson-Engels, S., Humbert, G., Komolibus, K., et al.: Liquid saliva analysis using optofluidic photonic crystal fiber for detection of oral potentially malignant disorders. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy **332**, 125788 (2025)
20. Mensah, G.B.: Artificial intelligence and ethics: a comprehensive review of bias mitigation, transparency, and accountability in ai systems. Preprint, November **10**(1), 1 (2023)
21. Mokere, R., Ghassan, M., Barra, I.: Soil spectroscopy evolution: A review of homemade sensors, benchtop systems, and mobile instruments coupled with machine learning algorithms in soil diagnosis for precision agriculture. Critical reviews in analytical chemistry **55**(7), 1304–1323 (2025)
22. Nogueira, M.S., Maryam, S., Amissah, M., Killeen, S., O'Riordain, M., Andersson-Engels, S.: Diffuse reflectance spectroscopy for colorectal cancer surgical guidance: towards real-time tissue characterization and new biomarkers. Analyst **149**(1), 88–99 (2024)

23. Pandiselvam, R., Prithviraj, V., Manikantan, M., Kothakota, A., Rusu, A.V., Trif, M., Mousavi Khaneghah, A.: Recent advancements in nir spectroscopy for assessing the quality and safety of horticultural products: A comprehensive review. Frontiers in Nutrition **9**, 973457 (2022)
24. Parker, S.F.: The analysis of vibrational spectra: Past, present and future. ChemPlusChem **90**(1), e202400461 (2025)
25. Peng, Y., Wang, T., Xie, S., Liu, Z., Lin, C., Hu, Y., Wang, J., Mao, X.: Estimation of soil cations based on visible and near-infrared spectroscopy and machine learning. Agriculture **13**(6), 1237 (2023)
26. Rodriguez Troncoso, J., Marium Mim, U., Ivers, J.D., Paidi, S.K., Harper, M.G., Nguyen, K.G., Ravindranathan, S., Rebello, L., Lee, D.E., Zaharoff, D.A., et al.: Evaluating differences in optical properties of indolent and aggressive murine breast tumors using quantitative diffuse reflectance spectroscopy. Biomedical Optics Express **14**(12), 6114–6126 (2023)
27. Rossberg, N., Gautam, R., Komolibus, K., O'Sullivan, B., Visentin, A.: Explainable ai-based feature selection approaches for raman spectroscopy. Diagnostics **15**(16), 2063 (2025)
28. Rossberg, N., Li, C.L., Innocente, S., Andersson-Engels, S., Komolibus, K., O'Sullivan, B., Visentin, A.: Machine learning applications to diffuse reflectance spectroscopy in optical diagnosis: a systematic review. Applied Spectroscopy Reviews pp. 1–52 (2025)
29. Salih, A.M.: Explainable artificial intelligence for dependent features: Additive effects of collinearity (2024), `https://arxiv.org/abs/2411.00846`
30. Sekar, S.K.V., Bargigia, I., Mora, A.D., Taroni, P., Ruggeri, A., Tosi, A., Pifferi, A., Farina, A.: Diffuse optical characterization of collagen absorption from 500 to 1700 nm. Journal of biomedical optics **22**(1), 015006–015006 (2017)
31. Su, W.H.: Advanced machine learning in point spectroscopy, rgb-and hyperspectral-imaging for automatic discriminations of crops and weeds: A review. Smart Cities **3**(3), 767–792 (2020)
32. Tetef, S., Govind, N., Seidler, G.T.: Unsupervised machine learning for unbiased chemical classification in x-ray absorption spectroscopy and x-ray emission spectroscopy. Physical Chemistry Chemical Physics **23**(41), 23586–23601 (2021)
33. Thalpage, N.: Unlocking the black box: Explainable artificial intelligence (xai) for trust and transparency in ai systems. J. Digit. Art Humanit **4**(1), 31–36 (2023)
34. Union, E.: Regulation (eu) 2024/1689 of the european parliament and of the council. Official Journal of the European Union (2024)
35. Wang, C.Y., Ko, T.S., Hsu, C.C.: Machine learning with explainable artificial intelligence vision for characterization of solution conductivity using optical emission spectroscopy of plasma in aqueous solution. Plasma Processes and Polymers **18**(12), 2100096 (2021)
36. Westermayr, J., Marquetand, P.: Machine learning spectroscopy to advance computation and analysis. Chemical Science **16**(46), 21660–21676 (2025)
37. Wiklund, S.: Spectroscopic data and multivariate analysis: tools to study genetic perturbations in poplar trees. Ph.D. thesis, Kemi (2007)
38. Yang, C.C.: Explainable artificial intelligence for predictive modeling in healthcare. Journal of healthcare informatics research **6**(2), 228–239 (2022)
39. Zhao, B., Zhai, H., Shao, H., Bi, K., Zhu, L.: Potential of vibrational spectroscopy coupled with machine learning as a non-invasive diagnostic method for covid-19. Computer Methods and Programs in Biomedicine **229**, 107295 (2023)

40. Zhao, T., Desjardins, A.E., Ourselin, S., Vercauteren, T., Xia, W.: Minimally invasive photoacoustic imaging: Current status and future perspectives. Photoacoustics **16**, 100146 (2019)