



# Enhancing the classification metrics of spectroscopy spectrums using neural network based low dimensional space

Mohamed Yousuff<sup>1</sup> · Rajasekhara Babu<sup>1</sup>

Received: 27 September 2022 / Accepted: 4 December 2022 / Published online: 23 December 2022  
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

## Abstract

Spectroscopy is a methodology for gaining knowledge of particles, especially biomolecules, by quantifying the interactions between matter and light. By examining the level of light absorbed, reflected or released by a specimen, its constituents, properties, and volume can be determined. Spectra obtained through spectroscopy procedures are quick, harmless and contactless; hence nowadays preferred in chemometrics. Due to the high dimensional nature of the spectra, it is challenging to build a robust classifier with good performance metrics. Many linear and nonlinear dimensionality reduction-based classification models have been previously implemented to overcome this issue. However, they lack in capturing the subtle details of the spectra into the low dimension space or cannot efficiently handle the nonlinearity present in the spectral data. We propose a graph-based neural network embedding approach to extract appropriate features into latent space and circumvent the spectrums' nonlinearity problem. Our approach performs dimensionality reduction into two phases: constructing a nearest neighbor graph and producing almost linear embedding using a fully connected neural network. Further, the low dimensional embedding is subjected to classification using the Random Forest algorithm. In this paper, we have implemented and compared our technique with four nonlinear dimensionality techniques widely used for spectral data analysis. In this study, we have considered five different spectral datasets belonging to specific applications. The various classification performance metrics of all the techniques are evaluated. The proposed approach is able to perform competitively well on six different low-dimensional spaces for each dataset with an accuracy score above 95% and Matthew's correlation coefficient value close to 1. The trustworthiness score of almost 1 show that the presented dimensionality reduction approach preserves the closest neighbor structure of high dimensional spectral inputs into latent space.

**Keywords** Dimensionality reduction · Machine learning · Spectroscopy · Random Forest · COVID-19 · Chemometrics

## Introduction

Spectroscopy allows an investigation of the interplay between matter and radiation as a relation and dependence of wavelength, which is a not annihilative, harmless, non-contactable and quick methodology compared to the conventional approaches in chemometrics (Fu and Ying 2016; Zheng et al. 2017). Raman spectroscopy (RS) and Infrared spectroscopy (IS) are two more prominent techniques with

a plethora of applications from science and agriculture to engineering. Raman is a light dispersion methodology in which a molecule disperses the ray of incident light from an intensified laser light emitter. Maximum dispersed light holds the same wavelength as the light source; hence, it fails to provide valuable and beneficial information. Meaningful information is hidden in a little ray of light scattered at differing wavelengths called Raman scattering (Chen et al. 2022b; Araújo et al. 2021).

IS deals with observing molecules stimulated by an infrared light beam resulting in an infrared absorbance spectrum. IS absorbance spectrum being a "fingerprint" of any (bio)chemical component, provides intrinsic information about the substances, which is necessary for many investigations. Recently, RS and IS combined with Machine Learning (ML) and Dimensionality Reduction (DR) approaches have numerous realistic and pragmatic

---

Communicated by: H. Babaie

✉ Rajasekhara Babu  
mrajasekharababu@vit.ac.in

<sup>1</sup> School of Computer Science and Engineering, Vellore Institute of Technology, Vellore Campus, Vellore 632014, Tamilnadu, India

applications like quantitative and qualitative assessments of soil attributes to ensure fertility and productivity via the formulation and recommendation of modified fertilizer compositions (Barra et al. 2021). Effectual guidance is provided in diagnosing different types of cancer based on the spectrums generated on the blood plasma and skin samples (Chen et al. 2022b; Araújo et al. 2021; Mohamed Yousuff and Rajasekhara Babu 2022). Detection of food adulteration, especially in seafood, honey, and edible oils, helps differentiate the quality and debasement of fuels such as diesel and gasoline (Dumancas and Ellis 2022; Owen et al. 2021; Zhao et al. 2022; Li et al. 2020; Wang et al. 2000a, 2000b). Evaluation of antibiotic susceptibility of certain microorganisms such as bacteria and finding the types of diseases that affected the crop leaves (Suleiman et al. 2022; Mohamed Yousuff et al. 2020). Microbial spoilage detection and classification of muscle foods (Ellis et al. 2004; Yan et al. 2021). Investigating the quality of products, specifically tea, coffee and fruit-based beverages (Mishra et al. 2018; Hu et al. 2022; Bizzani et al. 2020). Chemical component analysis and many more medical and health-related applications are apparently implemented using these technologies (Gao et al. 2022; Chen et al. 2022a; Ralbovsky et al. 2021; Liu et al. 2021).

Spectra usually have many wavelength features ranging from a hundred to a thousand dimensions. Classification approaches utilizing the entire spectrum of features become a time-consuming process, and extraneous information within the spectrum will compromise the model's precision and stability. Especially if the dimensionality of the input spectral features is very high, the modeling time and cost will be extraordinarily expensive. Regarding sampling cost in industrial and medical applications, the number of spectral observations or samples ( $s$ ) collected for any scenario is usually lower than the number of features ( $f$ ). Without well-formed and conditioned observation matrix, such as  $s \gg f$ , ML models cannot deliver accurate predictions. In other words, RS and IS spectrums are high-dimensional data which require an effective DR approach to achieve better classification metrics (Mishra et al. 2018; Zheng et al. 2019).

Principal Component Analysis (PCA) is among the most widely implemented linear DR algorithm for several High Dimensional Space (HDS) data. PCA is a projection-based approach that transforms the data points by mapping them onto orthogonal axes. PCA identifies the optimal linear combinations of the actual spectral features so that the variance or dispersion along the transformed feature is maximized (Zhang et al. 2022). In the context of spectral data, an eigenvector indicates a direction or axis of the data, and the associated eigenvalue reflects the dispersion along that eigenvector. The greater the eigenvalue, the greater the variance, especially along the corresponding eigenvector. The outcome of the PCA algorithm is the principal components

which are linear combinations of actual wavelengths of the spectrum (Zhao et al. 2022).

Therefore, PCA cannot comprehend intricate polynomial correlations between features. Thus, a major issue with PCA is that it fails to produce efficient Low Dimensional Space (LDS) if there exists more nonlinearity in the spectrums (Liu et al. 2017). To overcome this problem, Kernel PCA (KPCA) is introduced to manage the nonlinear aspects of the spectrums. Kernel PCA is implemented to aid in the determination of data points whose decision boundaries are characterized by a nonlinearly separable function. The notion behind the concept of the kernel is to move to an HDS in which the decision boundary of the spectral features turns linear. A general nonlinear integration of the original features will generate a large number of new components or features after the implementation of the kernel function, which exponentially increases the problem's computational complexity compared to PCA (Sun et al. 2019).

KPCA cannot outperform PCA if many data points present in the spectral data are linearly separable; hence, using a nonlinear kernel may result in performance degradation due to overfitting (Li et al. 2020). Multi-Dimensional Scaling (MDS) is a method for learning manifolds that preserves distance. Methods that preserve distance presuppose that a manifold is given by the pairwise distances between its data points. In distance-preserving approaches, an LDS is created from an HDS such that pairwise distances between data points stay unchanged. MDS maintains spatial distances, whereas other approaches maintain graph distances. The dissimilarity matrix is computed from the input spectrums. MDS considers the dissimilarity matrix and creates a corresponding mapping on an LDS, retaining the dissimilarities of the data points as precisely as possible. Generating the dissimilarity matrix at each step of MDS needs a significant amount of processing resources. It is not easy to incorporate new data into MDS (Mishra et al. 2018).

Isometric mapping (ISOMAP) is a nonlinear DR technique based on spectral theory that attempts to retain geodesic distances in the LDS. ISOMAP begins by constructing a neighborhood network. The graph distance is used to estimate the geodesic distance between each pair of data points, then uses eigenvalue disintegration of the geodesic distance matrix, and the LDS of the dataset is subsequently determined. The geodesic distance is computed as the summation of path weights on the shortest path connecting two data points. When the manifolds are not adequately sampled and have gaps, ISOMAP fails miserably. Creating a neighborhood graph is challenging, and a small error in the parameters can have negative implications (Wang et al. 2020b; Mishra et al. 2018).

t-distributed Stochastic Neighbor Embedding (t-SNE) is a revolutionary DR and data visualization technique. t-SNE incorporates not only the local patterns of the HDS but also tries to maintain the global features of the data. It has a

remarkable capacity to form well-defined, distinct clusters. Student-t distribution is implemented to quantify the similarity between the data points in the LDS, and t-SNE uses a symmetric probability distribution for the HDS (Wang et al. 2020a; Luo et al. 2021). When the LDS dimension exceeds 3, t-SNE has execution issues. Similar to other gradient descent-based algorithms, t-SNE tends to become trapped in local optima. t-SNE is quite responsive to the perplexity value, producing misleading clusters. The fundamental t-SNE implementation is slow due to search requests for nearest neighbors (Wang et al. 2021b).

To overcome the above issues, we propose a Graph-based Neural Network Embedding (GNNE) approach to produce an appropriate and reliable LDS representation of HDS spectral inputs. GNNE starts with building a  $k$  closest neighbors' graph of input data points and computing edge probabilities of HDS input observations. Then a Fully Connected Neural Network (FCNN) with a nonlinear activation function is employed to obtain the embedding of desired low dimensions. The probability values of LDS or latent space or embedding are computed, and finally, the difference between both HDS and LDS probability distributions is optimized using the cross-entropy cost function to extract efficient embeddings. The content of our work is organized as follows, various spectral datasets description, spectra preprocessing strategies, visualization of preprocessed spectrums and proposed methodology is explained in **Method and Implementation** section. The procedures executed to extract the LDS using the proposed technique and comparative visualizations of 2-Dimensional embeddings obtained from various DR techniques and GNNE for all the spectral datasets are discussed and depicted in **Experiments** section. In **Results and Discussions** section, the spectra classification model and its performance metrics are presented along with the DR evaluation metric. Finally, the conclusion part is mentioned in **Conclusion** section.

## Method and implementation

### Spectral datasets description

The absorbance spectra from five different spectral datasets are considered in our study to examine the performance of the proposed DR approach. The datasets such as coffee, fresh meat, olive oil, and fruit purees are available online at <https://data.mendeley.com/datasets/frrv2yd9rg/1>. The COVID-19 Raman spectroscopy dataset is available on <https://doi.org/10.6084/m9.figshare.12159924.v1>. All the five datasets are described as follows:

- (i) Coffee: Beans of coffee collected from different world regions are roasted to too many degrees of temperature; finally, it is well processed to form a lyophilized powder. The coffee powder is stored in an air-tight plastic container at  $-20^{\circ}\text{C}$  before subjecting it to spectroscopy. After a while, Fourier Transformed Infrared Spectroscopy (FTIS) is used to generate the spectra of coffee for 56 samples belonging to two classes, namely arabica (29 samples) and robusta (27 samples). Each of these coffee spectra consists of 286-dimension features in a wavelength range between  $811$  to  $1910\text{ cm}^{-1}$  (Downey et al. 1997).
- (ii) Fresh Meat: The meat belonging to three categories, namely chicken, turkey, and pork, of approximately 100 g, is collected over a period of 14 days. Pork chops are taken; similarly, breast pieces of chicken and turkey are preferred for the spectroscopy. After removing skin and bones, all the meat is softened using a blender and cleansed using 2% Triton-X solution and distilled water. The Mid-IS spectrums are taken for 20 samples belonging to each category under frozen and thawed conditions. Each of these fresh meat spectra consists of 448-dimension features in a wavelength range between  $1006$  to  $1867\text{ cm}^{-1}$  (Al-Jowder et al. 1997).
- (iii) Olive Oil: Sixty specimens of genuine virgin olive oils are obtained from four reputed European nations well known for their oil production. The samples gathered are 10, 17, 8, and 25 from Greece, Italy, Portugal and Spain, respectively. Two distinct durations of around 14 days are spent to collect the data. Before and between spectral observations, specimens were kept in the dark at room temperature. Each of these virgin olive oil spectra consists of 570-dimension features in a wavelength range between  $799$  to  $1896\text{ cm}^{-1}$  (Tapp et al. 2003).
- (iv) Fruit Purees: Mid-IS spectrums are measured on two different types of verified fruit purees. The first type contains 351 spectra belonging to the 'Strawberry' category. Fresh strawberry fruits are collected, and purees are prepared by the researchers, which are then subjected to spectroscopy. The second category contains 632 spectra belonging to the 'Non-Strawberry' class. It is a kind of adulterated strawberry purees mixed with several other fruit juices and sugar solutions. Each of these fruit purees spectra consists of 235-dimension features in a wavelength range between  $900$  to  $1802\text{ cm}^{-1}$  (Holland et al. 1998).
- (v) COVID-19: The blood serum of confirmed COVID-19 patients, healthy persons, and suspected individuals are obtained. The RS analysis process is performed, and spectra are measured on all types of serum specimens. 465 spectra are totally acquired, out of which 159 spectra come under the COVID-19 category, 156 spectra belong to the suspected class, and 150 fit in the healthy class. Each of these COVID-19 spectra con-

sists of 900-dimension features in a wavelength range between 400 to 2112  $\text{cm}^{-1}$  (Yin et al. 2019).

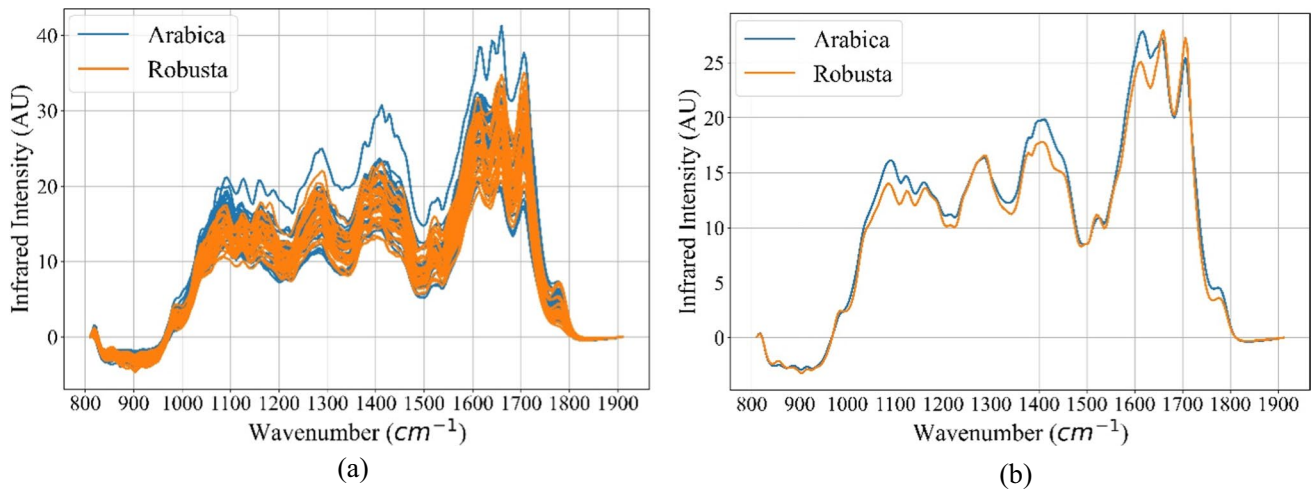
## Spectra preprocessing

The spectral data analysis procedure can assure better results with effective preprocessing steps implemented prior to the analysis phase. A sequence of preprocessing steps like background noise elimination, normalization, smoothing, and baseline emendation is carried out in order to ensure finer and enhanced classification metrics (Khan et al. 2018). Applying a Savitzky-Golay filter, all spectrums are smoothed. A digital filter such as the Savitzky–Golay can be used to smooth the data, enhancing its lucidity, sharpness, and resolution without altering the spectrum's rudimentary pattern. In a convolution approach, a low-degree polynomial

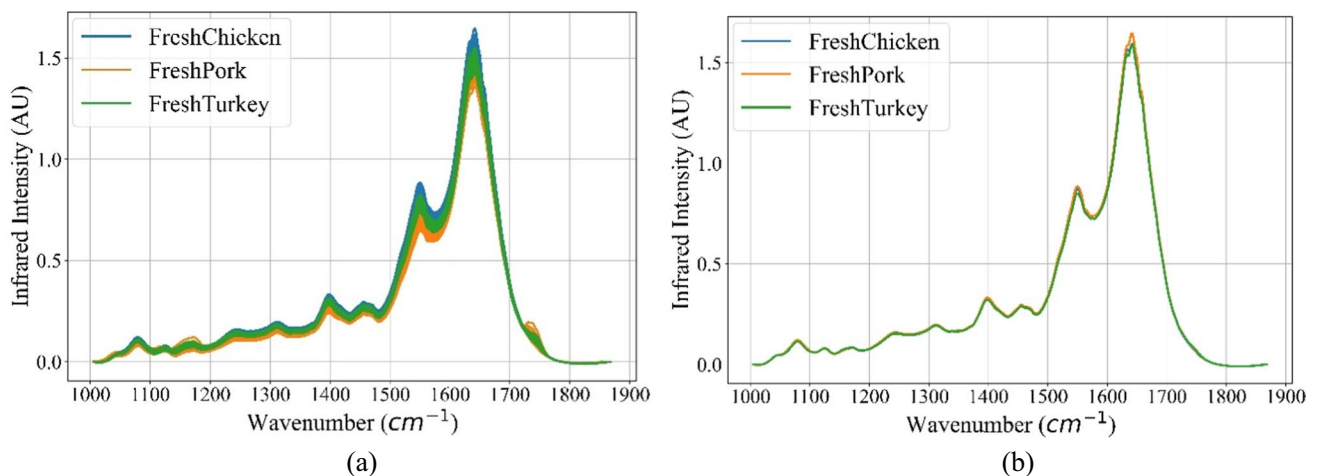
is adapted to successive subsets of adjacent data points using the linear least-squares technique to obtain the smoothing outcome (Mohamed Yousuff and Rajasekhara Babu 2022; Schafer 2011). The preprocessed spectrums and their corresponding average spectrums of all the datasets are depicted in Figs. 1, 2, 3, 4, and 5, respectively.

## Methodology

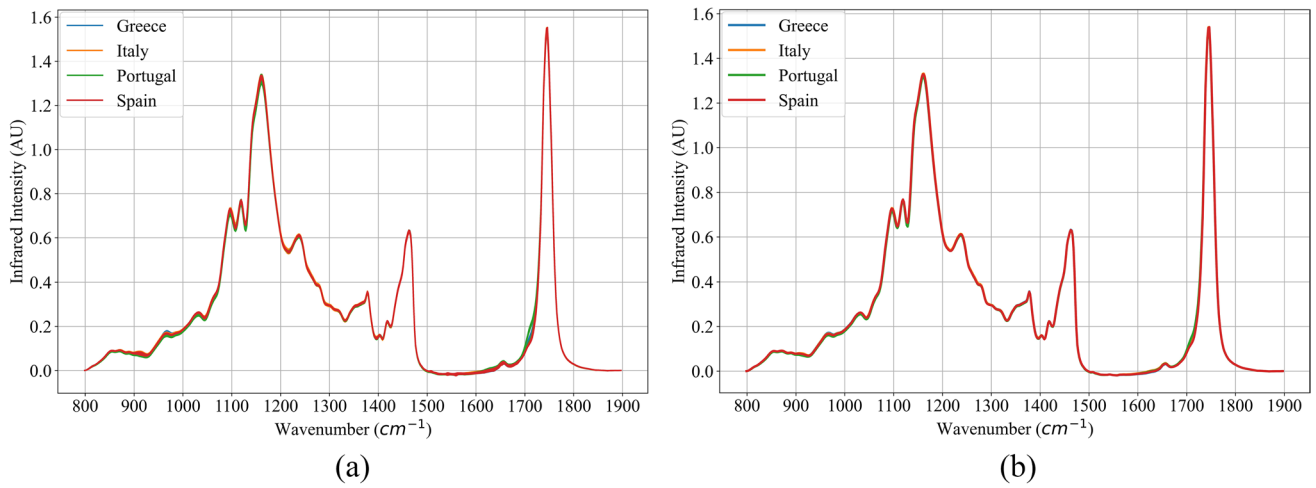
Similar to t-SNE and associated techniques, we presume that the data points which are in close proximity to each other in the HDS as per a pertinent metric should also be closer to one another in the embedding space. Consequently, we also assume that the data points which are far away in proximity in the HDS should also be isolated accordingly in the LDS. We suppose a metric like the Euclidean distance



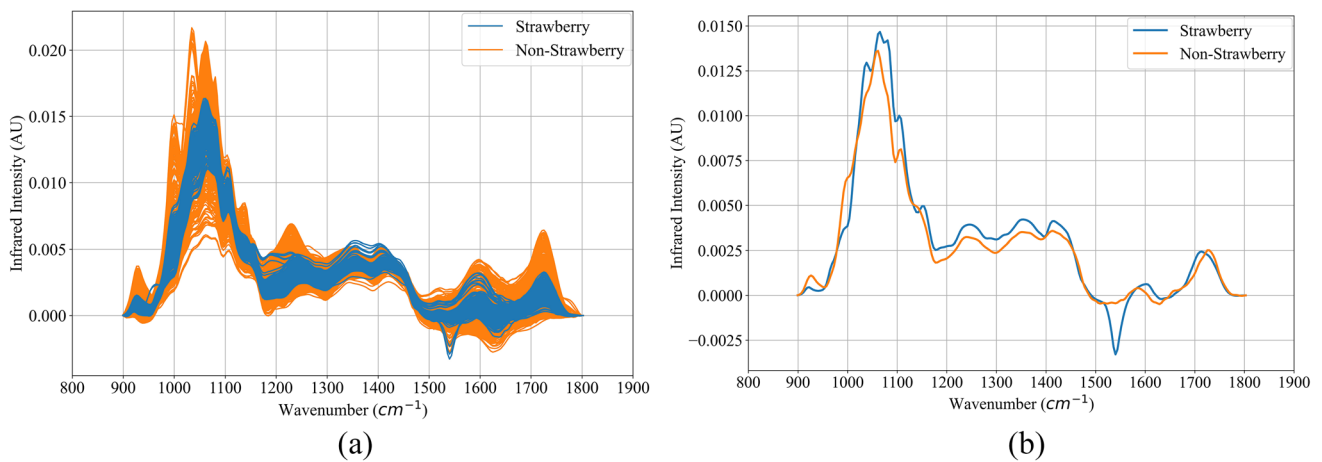
**Fig. 1** Preprocessed FTIS spectrums of arabica and robusta coffee variety (a) Set of all preprocessed coffee spectrums (b) Preprocessed average coffee spectrums



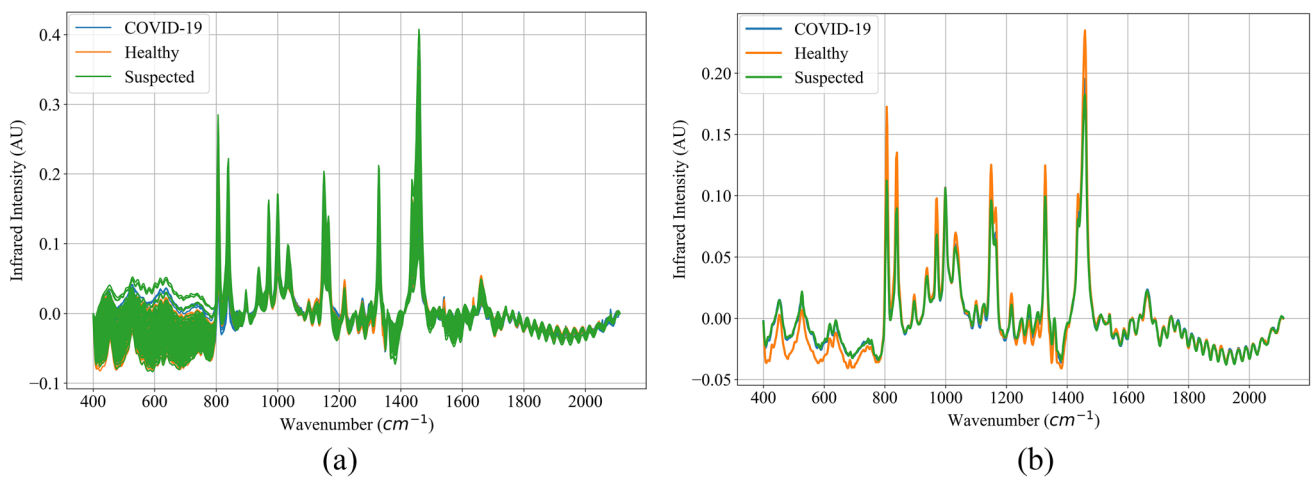
**Fig. 2** Preprocessed Mid-IS spectrums of chicken, pork and turkey fresh meat (a) Set of all preprocessed fresh meat spectrums (b) Preprocessed average fresh meat spectrums



**Fig. 3** Preprocessed FTIS spectra of Greece, Italy, Portugal and Spain olive oils (a) Set of all preprocessed olive oil spectra (b) Preprocessed average olive oil spectra



**Fig. 4** Preprocessed Mid-IS spectra of strawberry and non-strawberry fruit purees (a) Set of all preprocessed fruit purees spectra (b) Preprocessed average fruit purees spectra



**Fig. 5** Preprocessed RS spectra of COVID-19, healthy and suspected blood serum (a) Set of all preprocessed blood serum spectra (b) Preprocessed average blood serum spectra

in the HDS is adequate to depict a manifold on which the input observations lie (van der Maaten and Hinton 2008). The objective of the proposed GNNE approach: Given  $D$ -dimensional spectral data points  $S \in \mathbb{R}^D$ , create a  $d$ -dimensional LDS or embedding  $E \in \mathbb{R}^d (d \ll D)$  such that the data points nearer in proximity in  $S$  (for example  $S_i$  and  $S_j$ ) should also be closer to one another in  $E$  ( $E_i$  and  $E_j$ ). GNNE computes a nearest-neighbors graph and edge probability values of each input spectral datapoint followed by extraction of needed latent space using FCNN with a nonlinear activation function. The cross-entropy cost function is implemented to reduce the difference between high dimensional and low dimensional probability distributions, resulting in an effective embedding that can be further utilized for visualization and classification of the spectra.

### Data graph in the HDS (Spectral Input Space)

Considering the spectral dataset  $S = [s_1, s_2, \dots, s_n] \in \mathbb{R}^{D \times N}$  where  $N$  is the number of spectral observations and  $D$  is their corresponding dimensionality. We build a  $k$ -Nearest Neighbors ( $k$  NN) graph ( $k$  is considered as a hyper parameter) for the given spectral input space (Dong et al. 2011). The  $j$ -th neighbor of  $s_i$  is indicated as  $s_{i,j}$  then  $\eta_i := \{s_{i,1}, s_{i,2}, \dots, s_{i,k}\}$  represents the set of neighbor data points for the observations  $s_i$ . We considered the neighbor affinity relationship among the data points randomly. Radial Basis Function (RBF) or Gaussian kernel is used to compute the similitude between the input data points in the HDS (van der Maaten and Hinton 2008; Hinton and Roweis 2002; Ghogh et al. 2020). The probability that a spectral data point  $s_i$  has  $s_j$  as its neighbor can be calculated using the similarity of these data points as given in Eq. 1.

$$\mathbb{P}_{j|i} = \begin{cases} \exp\left(-\frac{\|s_i - s_j\|_2 - \zeta_i}{\psi_i}\right), & \text{if } s_j \in \eta_i \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where  $\|\cdot\|_2$  indicates the  $L_2$  norm,  $\zeta_i$  is the measure of distance between  $s_i$  and its nearest neighbor data point given by Eq. 2.

$$\zeta_i = \min\{\|s_i - s_{i,j}\|_2 \mid 1 \leq j \leq k\} \quad (2)$$

The  $\psi_i$  is the scaling variable, calculated so as to normalize the total similarity of the data point  $s_i$  to its  $k$  NNs. Using binary search  $\psi_i$  is determined to gratify Eq. 3.

$$\sum_{j=1}^k \exp\left(\frac{\|s_i - s_{i,j}\|_2 - \zeta_i}{\psi_i}\right) = \log_2(k) \quad (3)$$

t-SNE uses entropy as perplexity for a similar scale search. Since the scaling for a data point in a crowded section of the dataset turns small, the scaling for a data point

in a sparsely dispersed area of the dataset becomes vast; as a result, these searches cause the neighborhoods of diverse data points to act similarly. In other terms, we presume that the observations are distributed uniformly on an LDS manifold. Directional similitude measure is given in Eq. 1 whereas Eq. 4 gives the symmetric measure of similitude between data points  $s_i$  and  $s_j$  in the high dimensional spectral input space.

$$\mathbb{R} \ni \mathbb{P}_{ij} = \mathbb{P}_{ij} + \mathbb{P}_{ji} - \mathbb{P}_{ij}\mathbb{P}_{ji} \quad (4)$$

### Data graph in the LDS (Embedding)

Let the embeddings of the spectral data points be  $E = [e_1, e_2, \dots, e_n] \in \mathbb{R}^{d \times n}$  where  $d$  is the dimensionality of the LDS, which is always considerably smaller than the HDS or spectral input space ( $d \ll D$ ) and  $n$  is the number of data points ( $N := n$ ). Notice that  $e_i$  is the LDS commensurate to  $s_i$ . In the LDS, the probability that a data point  $e_j$  is the neighbor of  $e_i$  can be calculated by the similitude of these data points given in Eq. 5.

$$\mathbb{R} \ni \mathbb{Q}_{ij} = (1 + u \|e_i - e_j\|_2^{2v})^{-1} \quad (5)$$

The variables  $u > 0$  and  $v > 0$  are the two hyperparameters influenced by the user. We have considered the value of  $u \approx 1.9289$  and  $v \approx 0.7914$ , since it has been empirically demonstrated that choosing  $u = v = 1$  has no qualitative effect on the outcomes (Böhm et al. 2022).

### Neural network-based embedding

The proposed technique tries to create a similarity between the data graph in the HDS and the data points in the LDS. In other words, we interpret Eq. 4 and 5 as probability distributions and decrease the disparity between them so that the similitude of data points in the LDS resembles the similarities of data points in the HDS. The modified cross-entropy cost function (C) given in Eq. 6 (Tang et al. 2016) is used as a measurement for the variation between these two probability distributions.

$$\mathbb{C} = \sum_{i=1}^n \sum_{j=1, j \neq i}^n \left( \mathbb{P}_{ij} \ln\left(\frac{\mathbb{P}_{ij}}{\mathbb{Q}_{ij}}\right) + (1 - \mathbb{P}_{ij}) \ln\left(\frac{1 - \mathbb{P}_{ij}}{1 - \mathbb{Q}_{ij}}\right) \right) \quad (6)$$

where  $\ln(\cdot)$  denotes the natural logarithm. The first part of Eq. 6 specifies the pulling force that pulls neighboring data point embeddings towards one another. This component of the equation can only exist when  $\mathbb{P}_{ij} \neq 0$ , indicating that there are three possibilities such as  $s_i$  is closer in proximity to  $s_j$  or  $s_j$  is a neighbor of  $s_i$ , or both. The second element in Eq. 6 is the repelling force that separates the embeddings of data points that are far in proximity or non-neighbors. The loss

function of the neural network ( $f_{\vartheta}(\cdot)$ ) is the cost function of the algorithm, and the weights  $\vartheta$  (parameters) of the network are updated while training using backpropagation of loss function errors. Instead of optimizing the cost for the entire dataset, optimization is done in mini-batches so that the model is capable of taking many more observations. The nonlinearity in the layers of the neural network is achieved by implementing an effective nonlinear activation function; hence the model can easily manage very high nonlinear inputs. The steps and procedures involved in the proposed GNNE approach are given in Algorithm 1.

### Hyperparameters

Because networks include many parameters, some important hyperparameters are fixed before training the model. Bayesian optimization technique is used for the selection of hyperparameters used in the work. Some of the hyperparameters to be tuned the neural network are number of neurons, activation function, learning rate, batch size, number of epochs etc. The proposed algorithm has four hyperparameters as follows:

1.  $k$ , the number of nearby neighbors to take into account while estimating the local (Euclidean) distant metric. The  $k$  parameter enables a quantifiable indication of how effectively the embedding has kept the crucial local

structure of the input HDS data. By altering the values of  $k$ , we can further examine how structure preservation changes during the shift from exclusively local to non-local to global structure.  $k$  exemplifies a level of trade-off between granular and massive scale manifold or diverse nonlinear features. Smaller values assure meticulous nonlinear structure is precisely apprehended, while larger values grab massive scale nonlinear structures. With smaller  $k$  values, the manifold begins to fragment into a multitude of small, interconnected components. It is intuitive and empirically evident to choose a range of smaller values ranging between 5 to 50 for datasets with fewer observations. On the other hand,  $k$  value can be more than hundred in case of large datasets.

2.  $d$ , the dimension of the expected embeddings or LDS. 2 or 3-dimensional embeddings are very much essential to visualize the HDS in LDS and derive significant insights. In contrast, more than 3-dimensional LDS can be utilized for better and more efficient classification tasks achieving high-performance metric values.
3.  $e - dist$ , an intended distance of separation between adjacent points in the embedding space.  $e - dist$  regulates how tightly data points can be compacted to one another in the LDS representation. Smaller values of  $e - dist$  will likely result in crowded sections but will more accurately depict the manifold structure. Larger value of  $e - dist$  will drive the embedding to disperse

#### Algorithm 1 Graph based Neural Network Embeddings

---

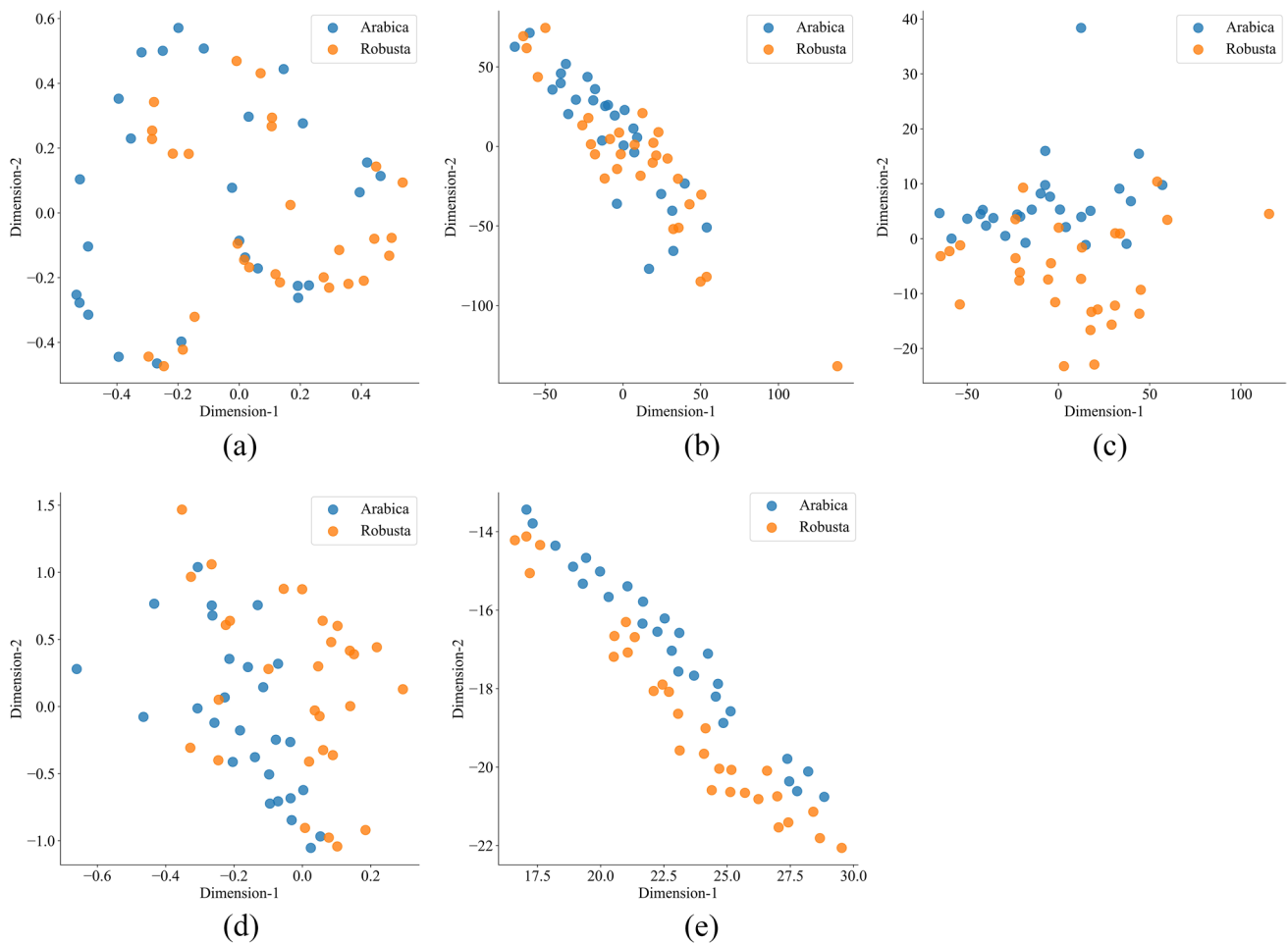
**Input:** High dimensional spectral observations:  $\{S_i \in \mathbb{R}^D\}, i = 1, 2, 3, \dots, N$

**Output:** Low dimensional embeddings of input spectral observations:  $\{E_i \in \mathbb{R}^d, d \ll D\}, i = 1, 2, 3, \dots, N$

**Initialization:** Stochastic initialization of network weights ' $\vartheta$ ' and  $\{E_i\}_{i=1}^n \leftarrow \{f_{\vartheta}(S_i)\}_{i=1}^n$

---

1. Build a  $k$ NN graph from the given HDS spectral inputs
  2. Calculate HDS edge probabilities  $\mathbb{p}_{ij}$  for every observation using Eqn. 4
  3. Create a Tensorflow dataset that iterates throughout the batches of  $k$ NNs depending on the probabilities indicated by the created graph.
  4. Train an FCNN to minimize the loss for the dataset obtained in step 3.
  5. Extract the embeddings corresponding to each batch and compute the probabilities  $\mathbb{q}_{ij}$  for the LDS using Eqn. 5.
  6. Reduce the difference between HDS and LDS probability distributions using the cost function given in Eqn. 6.
  7. Obtain the final LDS values from the last layer of the model, i.e.,  $E_i = f_{\vartheta}(S_i), \forall i \in \{1, 2, 3, \dots, N\}$
-



**Fig. 6** 2D visualization of coffee spectra using different DR techniques (a) KPCA (b) MDS (c) ISOMAP (d) t-SNE (e) Proposed Approach

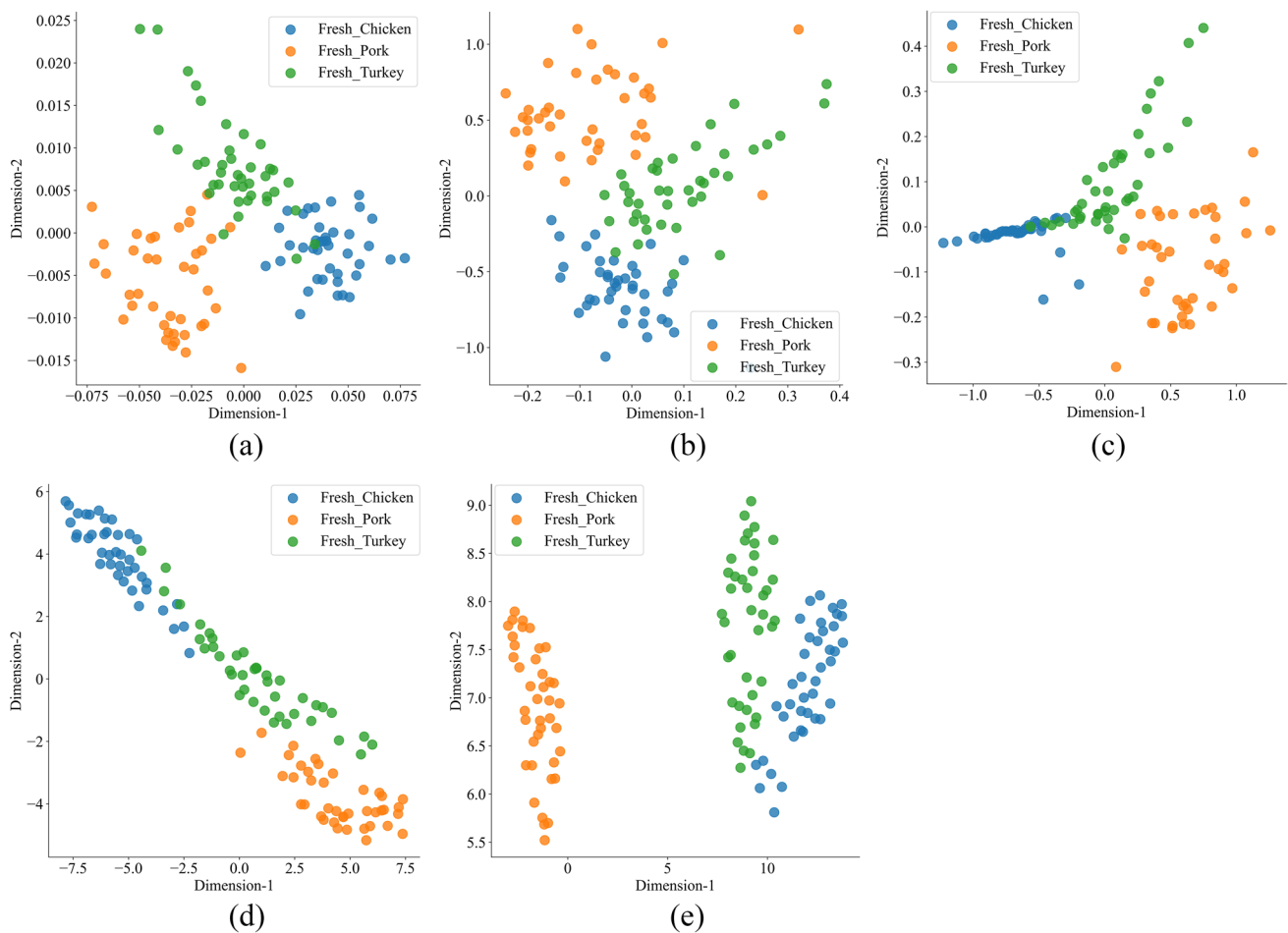
the data points further, so facilitating visualization by preventing any overlapping and overplotting concerns. We consider  $e - dist$  to be a purely aesthetic factor that influences the visual aspect of the LDS, and therefore it is more relevant while performing visualization task. Higher values of  $e - dist$  causes the distinguishable clusters to get squeezed together, diminishing the boundaries between the clusters.

4.  $n - epochs$ , the number of epochs required during the training phase to optimize the LDS representation. In every epoch, each input observation in the batches of the training dataset gets a chance to internally fine-tune the model weights resulting in minimization of error and obtaining better embedding as output.

## Experiments

The GNNE algorithm takes five parameters as input, such as the HDS dataset,  $k$ ,  $d$ ,  $e - dist$  and  $n - epochs$ . For all the experiments on the five aforementioned datasets we have considered  $k = 15$ ,  $d = \{2, 3, 5, 10, 15, 20\}$ ,  $e - dist = 0.1$  and  $n - epochs = 10$ . The Nearest-Neighbor-Descent (NND) technique offers a cost-effective approximation of the  $k$ -nearest-neighbor calculation. Due to the error inherent in DR techniques, such an estimation is far beyond sufficient for these tasks. Even though no conceptual complexity limitations have been determined for NND, the authors mention an experimental complexity of  $O(N^{1.14})$  (Dong et al. 2011). The main advantage of NND is its versatility; it can be applied





**Fig. 7** 2D visualization of fresh meat spectra using different DR techniques (a) KPCA (b) MDS (c) ISOMAP (d) t-SNE (e) Proposed Approach

to any reasonable dissimilarity measure and is efficacious even for HDS. Hence NND with Euclidean distant metric is used in constructing the  $k$  NN graph. Based on the probabilities reported by the  $k$  NN graph a Tensorflow dataset is built through an iteration process on the batches of  $k$  NNs. These input data are then subjected to an FCNN for optimization and efficient generation of embeddings.

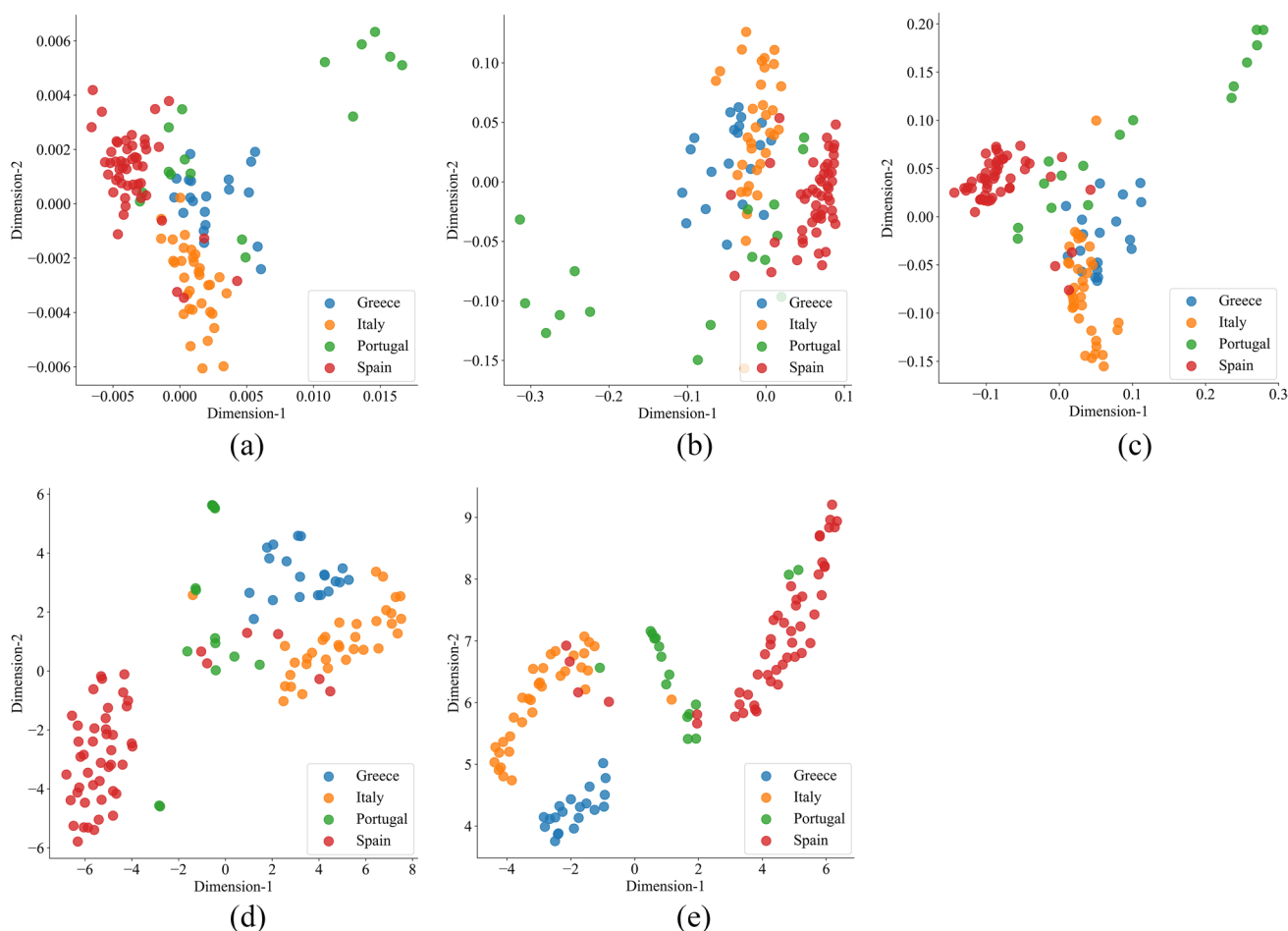
The FCNN with four layers is used to obtain the LDS. The input layer is meant to acquire the spectral observations. The two hidden layers, followed by the input layer, have 128 and 64 neurons along with Scaled Exponential Linear Unit (SELU) nonlinear activation function. Implementing the SELU activation function gives the layers of neural network self-normalizing characteristics. The SELU

activation function and its derivate is depicted in Eqs. 7 and 8 (Klambauer et al. 2017).

$$SELU(u) = \lambda \begin{cases} u, & \text{if } u > 0 \\ \alpha e^u - \alpha, & \text{if } u \leq 0 \end{cases} \quad (7)$$

$$SELU'(u) = \lambda \begin{cases} 1, & \text{if } u > 0 \\ \alpha e^u, & \text{if } u \leq 0 \end{cases} \quad (8)$$

where both  $\lambda$  and  $\alpha$  parameters are set to a value of 1.0507 and 1.6733 empirically to infer better performance from FCNN (Klambauer et al. 2017). Each dense layer of the FCNN is incorporated with alpha dropout layers in between with a 0.1 dropout value to promote regularization of FCNN

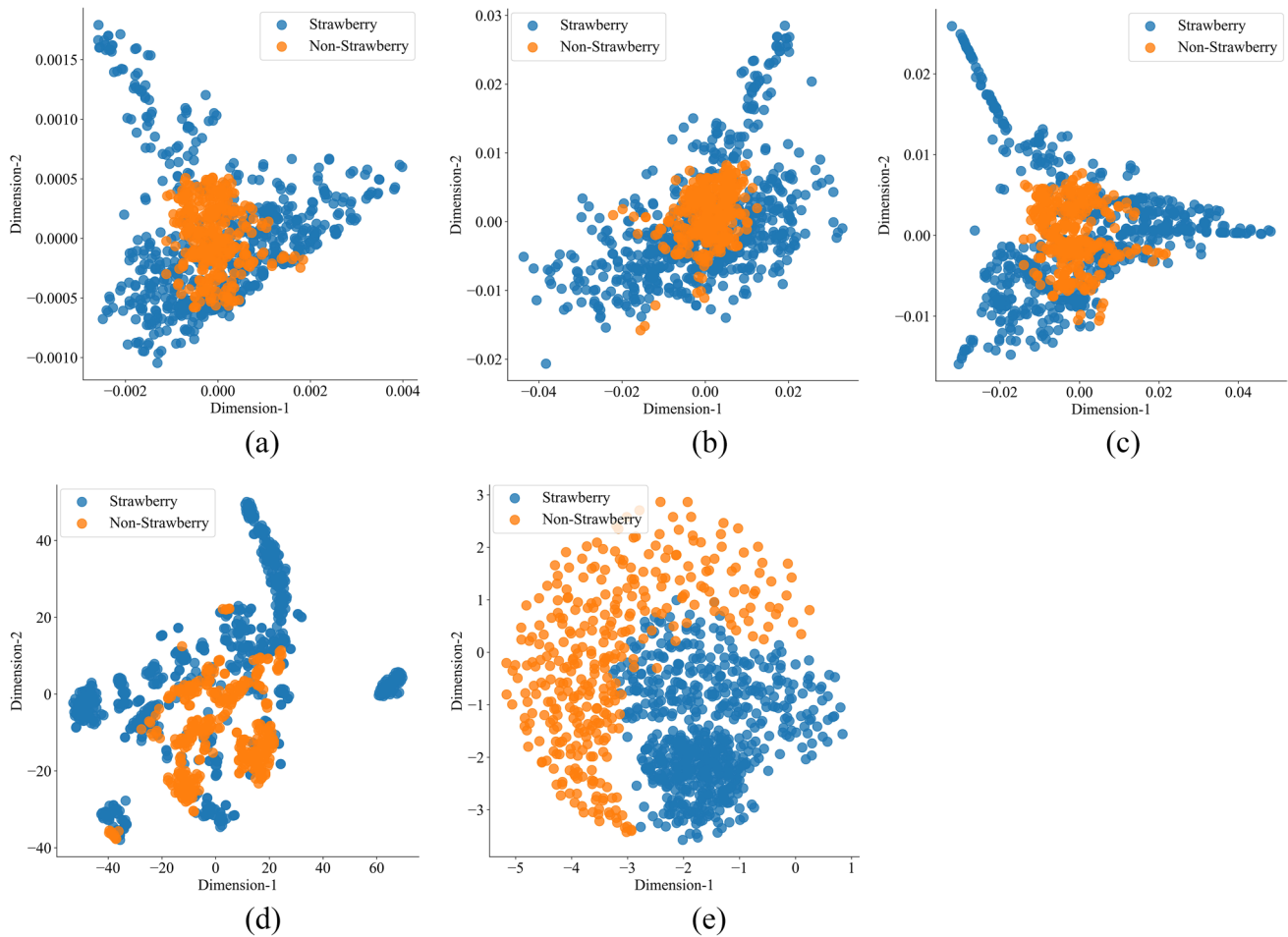


**Fig. 8** 2D visualization of olive oil spectra using different DR techniques (a) KPCA (b) MDS (c) ISOMAP (d) t-SNE (e) Proposed Approach

and to avoid over-fitting. These FCNN layers set 10% of input units to zero at each iteration but are meant to function with SELU to retain the self-normalization characteristic by preserving the variance and mean of inputs. SELU activation function is preferred for the following reasons: (i) It can have both negative and positive values, allowing the mean to be controlled. (ii) It facilitates saturation zones, allowing to soften large variances from preceding layers. (iii) It has a slope value greater than one, which enables it to raise low variances from the preceding layers. (iv) They form a seamless curve that guarantees a stable point between variance dampening and rising.

The backpropagation approach is utilized as a primary technique to train the FCNN. Weights are consistently updated after each iteration; as a result, the error rate is decreased, assuring the model's stability. Adamax

optimizer with hyperparameters value such as learning rate = 0.001,  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$  is chosen for faster convergence (Kingma and Ba 2015). The last layer of FCNN is the nodal point to extract the embeddings. Hence the number of units in this layer is equal to the dimension of LDS. We have considered six different LDS dimensions namely 2,3,5,10,15,20. LDS with 2-dimension (2D) and 3-dimension (3D) are helpful for visualization tasks, whereas higher dimensions are intended to perform the classification task. 2D graphs of various DR techniques and the GNNE approach on various datasets are illustrated in Figs. 6, 7, 8, 9, and 10. All DR techniques used in our work are implemented in python using Scikit-learn (Pedregosa et al. 2011), and all the visualizations are aid by NumPy (Walt et al. 2011), Pandas (McKinney 2010), Matplotlib (Hunter 2007), and Seaborn (Waskom et al. 2014) libraries.



**Fig. 9** 2D visualization of fruit purees spectra using different DR techniques (a) KPCA (b) MDS (c) ISOMAP (d) t-SNE (e) Proposed Approach

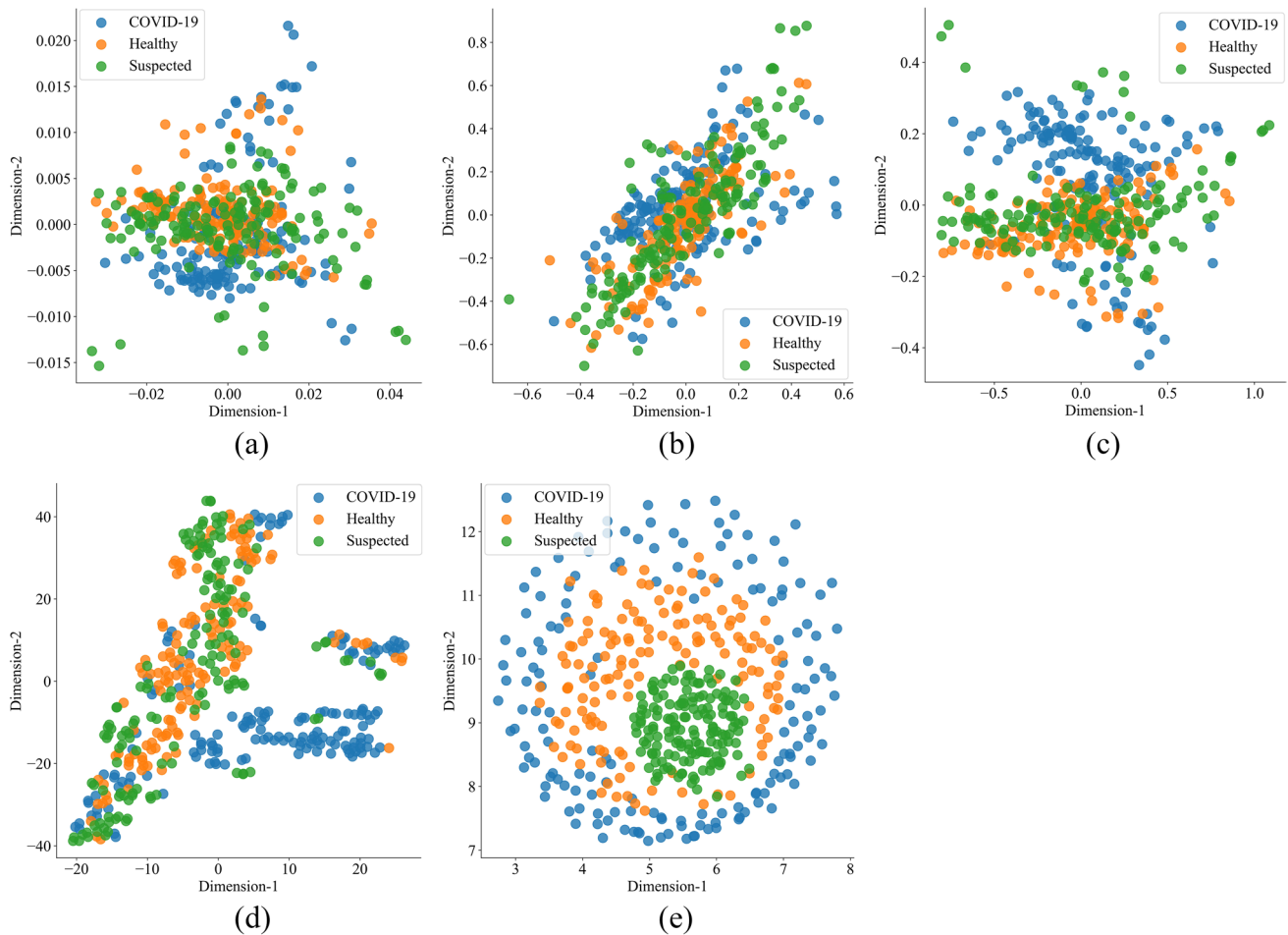
The coffee and fruit purees spectra have binary labels, whereas fresh meat, olive oil, and COVID-19 spectra are categorized into multiple labels. We can infer from the preprocessed spectral graphs (i.e., from Figs. 1, 2, 3, 4 and 5) that all datasets have a high degree of nonlinearity. The GNNE technique can handle the nonlinear data points considerably well compared to the existing methods, which is apparent in the 2D visualization of Figs. 6, 9, and 10. All the techniques did provide an interpretable 2D visualization for fresh meat spectra, as shown in Fig. 7. Additionally, the proposed approach is able to maintain the compactness between the spectra of similar category. Due to numerous nonlinear spectra in olive oil dataset, all the techniques have overlaps in projecting the HDS in 2D, as illustrated in Fig. 8.

### Results and discussions

The DR techniques are further evaluated using classification task performance metrics and trustworthiness. The performance metrics such as accuracy, precision, recall, F1-score, and Matthew's Correlation Coefficient (MCC) are evaluated on 2D, 3D, 5D, 10D, 15D, and 20D LDS of all the spectral datasets. Coffee and fruit purees spectra are meant for binary classification, whereas the other three datasets are subjected to multi-class classification.

### Spectra classification model and evaluation

Random Forest Classifier (RFC) is a predominant and accustomed model implemented for spectral features classification



**Fig. 10** 2D visualization of COVID-19 spectra using different DR techniques (a) KPCA (b) MDS (c) ISOMAP (d) t-SNE (e) Proposed Approach

(Breiman 2001; Gomes Marques et al. 2021; Wang et al. 2021a; Zhou et al. 2020). RFC is a collection of Decision Tree (DT) classifiers where each tree takes on the inputs of an independent random vector selected with a similar distribution for all DT in the forest. The fundamental principle of RFC is that a number of weak learners (DT) can be combined to create a strong learner. By introducing randomization into the sample choosing process via Bootstrap resample, many distinct trees are created, making the RFC less susceptible to over-fitting. Feed the input spectral features to every DT in the forest for the classification task. Each DT provides a classification or votes for a particular category or class. The forest selects the classification with the highest votes (Breiman 2001). Even in noisy situations, RFC detects significant features adequately, and it is adept at

handling HDS of the spectrum features (Ghebleh Goydaragh et al. 2021; Wójtowicz et al. 2021).

$K$ -fold cross-validation is a statistical method used to assess the model's performance. First,  $K$  segments are created from the overall spectral observations. The data points segments are then divided into training and testing batches randomly. The model picks one data segment for testing and the leftover  $K - 1$  data points segments for training. Therefore, in this manner, the model predicts the observations consecutively by repeating running the exact procedure  $k$  times. We have considered  $K = 10$ ; hence the whole spectral data is split up into 10 subgroups resulting in tenfold cross-validation. During the training process, the model extracts the perceived features of the LDS spectra obtained after DR and then predicts the labels of the

**Table 1** Coffee spectra classification metrics

Algorithm	Dimensionality	Precision	Recall	F1-Score
Without DR	286D	0.73	0.73	0.73
KPCA	2D	0.50	0.50	0.50
	3D	0.54	0.54	0.54
	5D	0.77	0.77	0.77
	10D	0.80	0.80	0.80
	15D	0.79	0.79	0.79
	20D	0.89	0.89	0.89
ISOMAP	2D	0.73	0.73	0.73
	3D	0.80	0.80	0.80
	5D	0.89	0.89	0.89
	10D	0.91	0.91	0.91
	15D	0.88	0.88	0.88
	20D	0.73	0.73	0.73
MDS	2D	0.52	0.52	0.52
	3D	0.89	0.89	0.89
	5D	0.77	0.77	0.77
	10D	0.82	0.82	0.82
	15D	0.89	0.89	0.89
	20D	0.93	0.93	0.93
t-SNE	2D	0.64	0.64	0.64
	3D	0.79	0.79	0.79
	5D	0.79	0.79	0.79
	10D	0.73	0.73	0.73
	15D	0.75	0.75	0.75
	20D	0.75	0.75	0.75
Proposed approach	2D	0.80	0.80	0.80
	3D	0.86	0.86	0.86
	5D	0.95	0.95	0.95
	10D	0.88	0.88	0.88
	15D	0.89	0.89	0.89
	20D	0.95	0.95	0.95

**Table 2** Fresh meat spectra classification metrics

Algorithm	Dimensionality	Precision	Recall	F1-Score
Without DR	448D	0.81	0.81	0.81
KPCA	2D	0.92	0.92	0.92
	3D	0.93	0.93	0.93
	5D	0.93	0.93	0.93
	10D	0.95	0.95	0.95
	15D	0.93	0.93	0.93
	20D	0.92	0.92	0.92
ISOMAP	2D	0.87	0.87	0.87
	3D	0.90	0.90	0.90
	5D	0.92	0.92	0.92
	10D	0.89	0.89	0.89
	15D	0.88	0.88	0.88
	20D	0.87	0.87	0.87
MDS	2D	0.91	0.91	0.91
	3D	0.91	0.91	0.91
	5D	0.90	0.90	0.90
	10D	0.94	0.94	0.94
	15D	0.90	0.90	0.90
	20D	0.91	0.91	0.91
t-SNE	2D	0.90	0.90	0.90
	3D	0.92	0.92	0.92
	5D	0.92	0.92	0.92
	10D	0.89	0.89	0.89
	15D	0.81	0.81	0.81
	20D	0.82	0.82	0.82
Proposed approach	2D	0.96	0.96	0.96
	3D	0.98	0.98	0.98
	5D	0.99	0.99	0.99
	10D	0.98	0.98	0.98
	15D	0.98	0.98	0.98
	20D	0.98	0.98	0.98

data points in the testing phase, ultimately resulting in the classification.

The confusion matrix consists of four terms such as (i) True Positive (TP), (ii) False Positive (FP), (iii) False Negative (FN), and (iv) True Negative (TN). Consequently, a superior classification model contains a greater number of TP and TN. Moreover, FP and FN values represent the model's errors and misclassifications. These four confusion matrix values are the essential factors to compute the metrics like accuracy, precision, recall, F1-score, and MCC using Eqs. 9 to 13. MCC computes the correlation between the

actual and predicted categories, yielding a value between -1 and 1. Hence, a value close to 1 is meant to be a good score which is only be achieved if the model is precise and reliable in all confusion matrix terms (Chicco et al. 2021). We have computed Micro Average Precision (MAP) and Micro Average Recall (MAR) for multi-class classification spectra as given in Eqs. 14 and 15.

Various evaluation metrics of classification tasks performed on the 2D, 3D, 5D, 10D, 15D, and 20D LDS spectra are given in Tables 1, 2, 3, 4, and 5, whereas accuracy and MCC metrics are depicted in Figs. 11 and 12. All the DR

**Table 3** Olive Oil spectra classification metrics

Algorithm	Dimensionality	Precision	Recall	F1-Score
Without DR	570D	0.52	0.52	0.52
KPCA	2D	0.66	0.66	0.66
	3D	0.78	0.78	0.78
	5D	0.78	0.78	0.78
	10D	0.78	0.78	0.78
	15D	0.78	0.78	0.78
	20D	0.78	0.78	0.78
ISOMAP	2D	0.62	0.62	0.62
	3D	0.78	0.78	0.78
	5D	0.77	0.77	0.77
	10D	0.74	0.74	0.74
	15D	0.75	0.75	0.75
	20D	0.70	0.70	0.70
MDS	2D	0.59	0.59	0.59
	3D	0.78	0.78	0.78
	5D	0.78	0.78	0.78
	10D	0.79	0.79	0.79
	15D	0.72	0.72	0.73
	20D	0.78	0.78	0.78
t-SNE	2D	0.77	0.77	0.77
	3D	0.80	0.80	0.80
	5D	0.73	0.73	0.73
	10D	0.56	0.56	0.56
	15D	0.53	0.53	0.53
	20D	0.60	0.60	0.60
Proposed approach	2D	0.82	0.82	0.82
	3D	0.81	0.81	0.81
	5D	0.81	0.81	0.81
	10D	0.81	0.81	0.81
	15D	0.82	0.82	0.82
	20D	0.82	0.82	0.82

**Table 4** Fruit purees spectra classification metrics

Algorithm	Dimensionality	Precision	Recall	F1-Score
Without DR	235D	0.74	0.74	0.74
KPCA	2D	0.74	0.74	0.74
	3D	0.86	0.86	0.86
	5D	0.90	0.90	0.90
	10D	0.94	0.94	0.94
	15D	0.95	0.95	0.95
	20D	0.95	0.95	0.95
ISOMAP	2D	0.81	0.81	0.81
	3D	0.89	0.89	0.89
	5D	0.92	0.92	0.92
	10D	0.92	0.92	0.92
	15D	0.92	0.92	0.92
	20D	0.92	0.92	0.92
MDS	2D	0.81	0.81	0.81
	3D	0.84	0.84	0.84
	5D	0.86	0.86	0.86
	10D	0.85	0.85	0.85
	15D	0.86	0.86	0.86
	20D	0.85	0.85	0.85
t-SNE	2D	0.93	0.93	0.93
	3D	0.93	0.93	0.93
	5D	0.93	0.93	0.93
	10D	0.94	0.94	0.94
	15D	0.94	0.94	0.94
	20D	0.86	0.86	0.86
Proposed approach	2D	0.96	0.96	0.96
	3D	0.94	0.94	0.94
	5D	0.93	0.93	0.93
	10D	0.97	0.97	0.97
	15D	0.94	0.94	0.94
	20D	0.95	0.95	0.95

techniques have improved the performance metrics, which is apparently visible in the results. The proposed technique has further enhanced the performance metrics significantly than the other DR techniques, especially in the case of multi-class spectra classification.

$$Accuracy = \frac{TN + TP}{TP + TN + FP + FN} \quad (9)$$

$$Precision = \frac{TP}{FP + TP} \quad (10)$$

$$Recall = \frac{TP}{FN + TP} \quad (11)$$

$$F1 - Score = 2 \left( \frac{Precision * Recall}{Precision + Recall} \right) \quad (12)$$

$$MCC = \frac{(TN * TP) - (FN * FP)}{\sqrt{(FP + TP) * (FN + TP) * (FP + TN) * (FN + TN)}} \quad (13)$$

$$MAP = \frac{(TP_{class_1} + TP_{class_2} \dots + TP_{class_n})}{(FP_{class_1} + FP_{class_2} \dots + FP_{class_n}) + (TP_{class_1} + TP_{class_2} \dots + TP_{class_n})} \quad (14)$$

$$MAR = \frac{(TP_{class_1} + TP_{class_2} \dots + TP_{class_n})}{(FN_{class_1} + FN_{class_2} \dots + FN_{class_n}) + (TP_{class_1} + TP_{class_2} \dots + TP_{class_n})} \quad (15)$$

**Table 5** COVID-19 spectra classification metrics

Algorithm	Dimensionality	Precision	Recall	F1-Score
Without DR	900D	0.46	0.46	0.46
KPCA	2D	0.33	0.33	0.33
	3D	0.37	0.37	0.37
	5D	0.62	0.62	0.62
	10D	0.69	0.69	0.69
	15D	0.64	0.64	0.64
	20D	0.66	0.66	0.66
ISOMAP	2D	0.46	0.46	0.46
	3D	0.52	0.52	0.52
	5D	0.58	0.58	0.58
	10D	0.60	0.60	0.60
	15D	0.56	0.56	0.56
	20D	0.57	0.57	0.57
MDS	2D	0.35	0.35	0.35
	3D	0.45	0.45	0.45
	5D	0.48	0.48	0.48
	10D	0.52	0.52	0.52
	15D	0.53	0.53	0.53
	20D	0.55	0.55	0.55
t-SNE	2D	0.53	0.53	0.53
	3D	0.55	0.55	0.55
	5D	0.57	0.57	0.57
	10D	0.60	0.60	0.60
	15D	0.61	0.61	0.61
	20D	0.60	0.60	0.60
Proposed approach	2D	0.83	0.83	0.83
	3D	0.82	0.82	0.82
	5D	0.83	0.83	0.83
	10D	0.89	0.89	0.89
	15D	0.92	0.92	0.92
	20D	0.92	0.92	0.92

### Trustworthiness $T(k)$

The degree to which a set of LDS retains the local structure of collection of the features is a measure of its trustworthiness. It is quantified by an act of examining resemblances of the nearest neighbors of each datapoint in LDS with actual input HDS datapoint. Let  $\mathbb{N}$  be the size of the total observations present in the input dataset and  $r(x, y)$  be the rank of the data point  $y$  in the arrangement based on the distance from  $x$  in the actual input HDS.  $Z_k(x)$  is a collection of those data observations of size  $k$  that are closest to the data point  $x$ . The measure of trustworthiness is given in Eq. 16 (Venna et al. 2001; Venna and Kaski 2006).

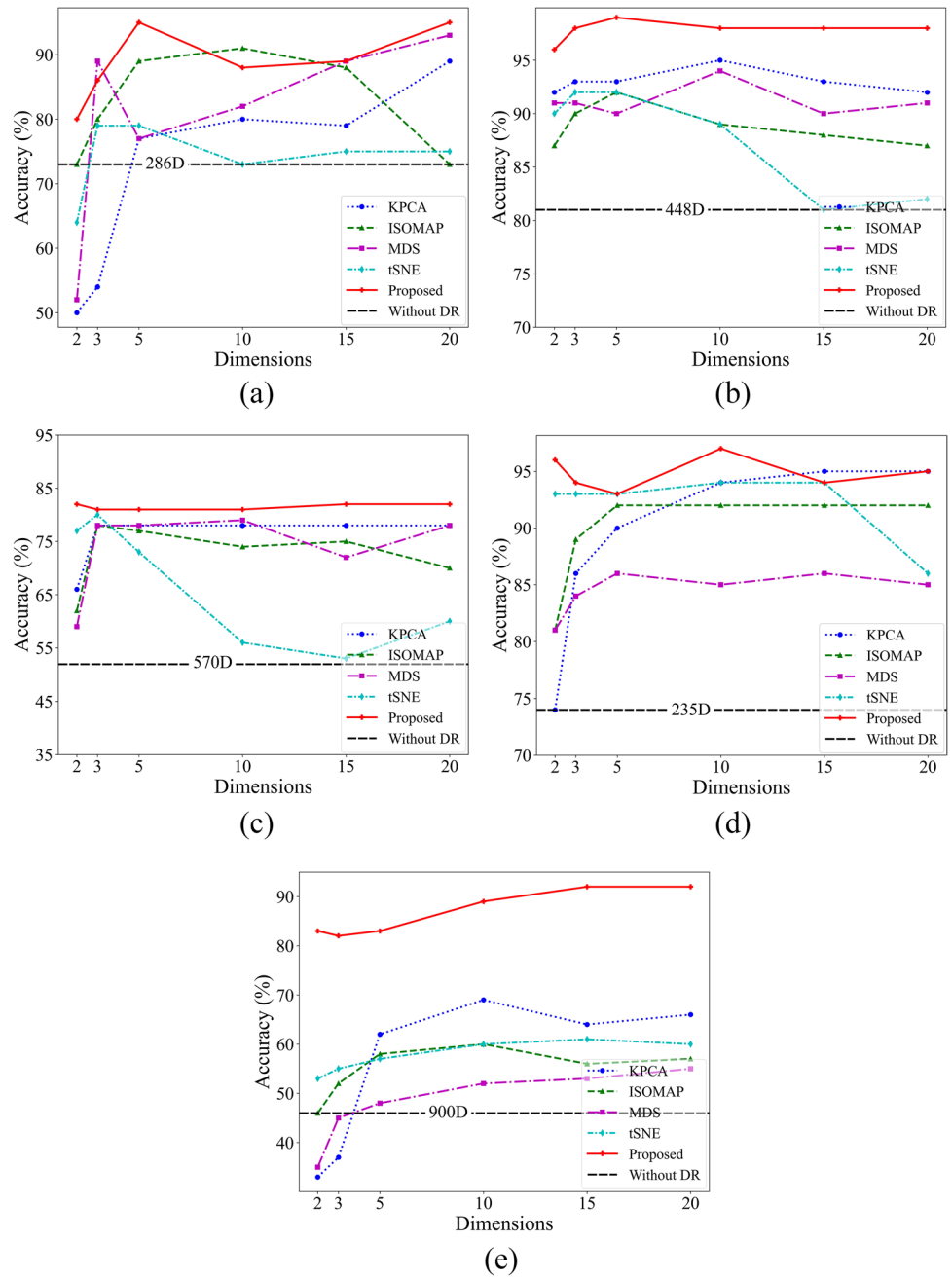
$$T(k) = 1 - \frac{2}{\mathbb{N}k(2\mathbb{N} - 3k - 1)} \sum_{x=1}^{\mathbb{N}} \sum_{y \in Z_k(x)} r(x, y) - k \quad (16)$$

The trustworthiness scale ranges from 0 to 1, with 1 being the most trustworthy.  $T(k)$  values for 2D, 3D, 5D, 10D, 15D, and 20D embeddings of various DR techniques are computed and depicted in Fig. 13. The proposed method is equally trustworthy as other DR techniques because it is able to achieve  $T(k)$  values almost equal to 1.

### Conclusion

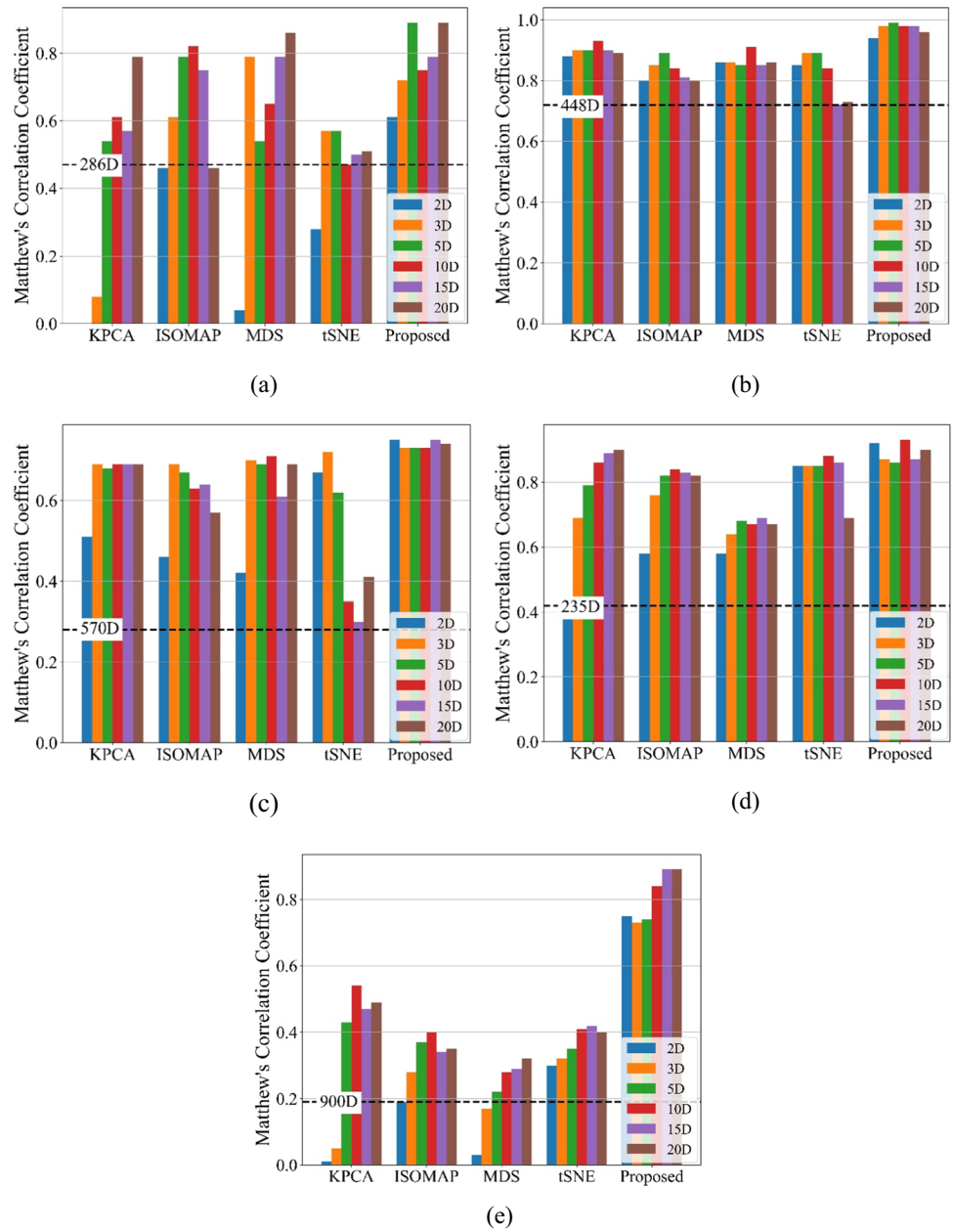
In this study, we presented a dimensionality reduction approach for spectroscopy spectra achieved using graph-based neural network embeddings. The spectral data collected from various sources and applications are of high dimensional nature. The classification performance of such spectra can be enhanced by effectively reducing the dimensions. The proposed technique is implemented on

**Fig. 11** Accuracy metric of various spectral datasets embeddings (a) Coffee spectra (b) Fresh meat spectra (c) Olive oil spectra (d) Fresh purees spectra (e) COVID-19 spectra

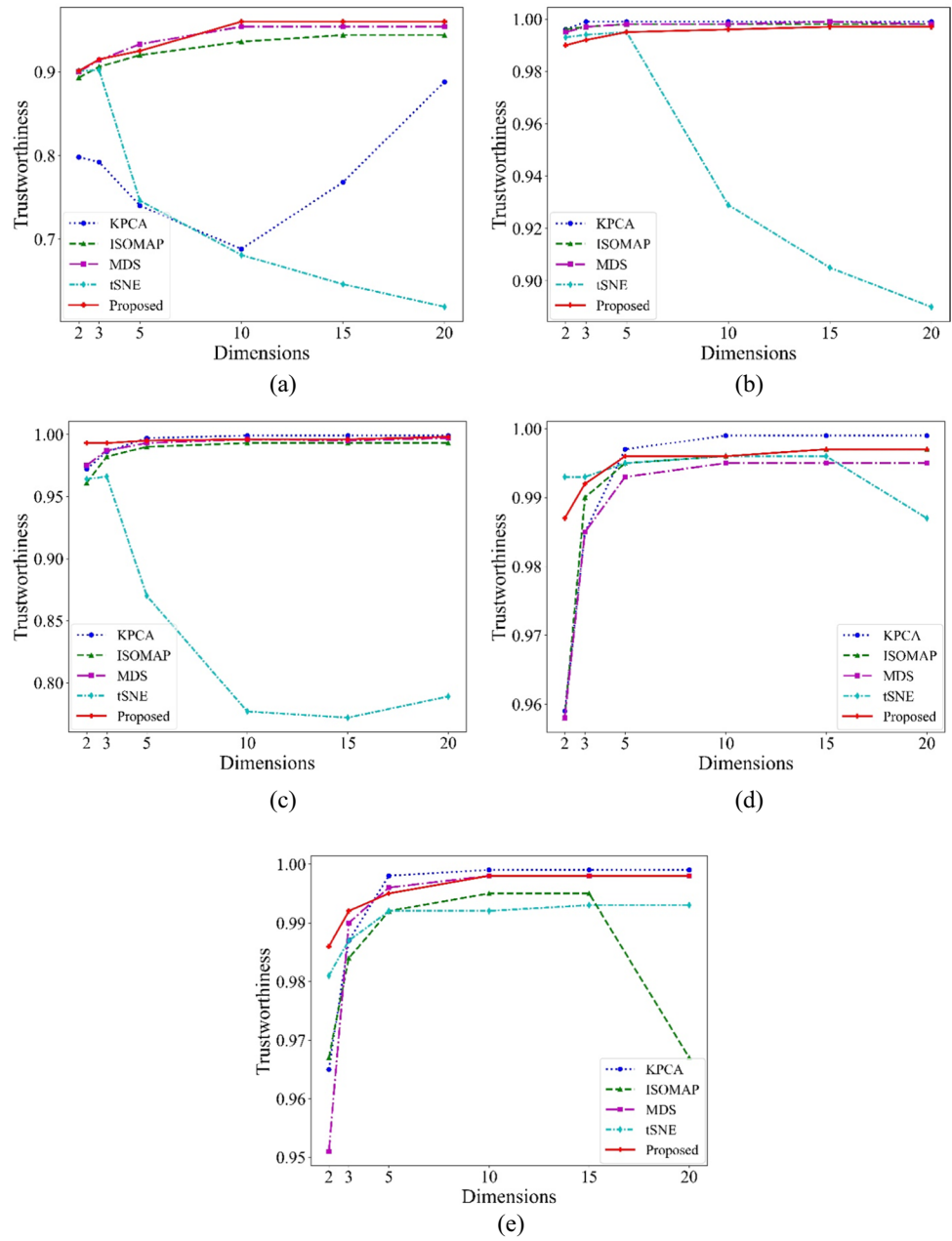




**Fig. 12** MCC metric of various spectral datasets embeddings (a) Coffee spectra (b) Fresh meat spectra (c) Olive oil spectra (d) Fresh purees spectra (e) COVID-19 spectra



**Fig. 13** Trustworthiness metric of various spectral datasets embeddings (a) Coffee spectra (b) Fresh meat spectra (c) Olive oil spectra (d) Fresh purees spectra (e) COVID-19 spectra



five different types of spectroscopy data, and its results are compared with existing prominent DR techniques. The 2D visualizations of the spectral datasets using our approach have shown a competitive and better LDS visualization of HDS. Nonlinearity present in the data is handled efficaciously using a nonlinear activation function; as a result, all the performance metrics of the classification task, including accuracy and MCC, have been remarkably improved. The multi-classification task of spectra has shown slightly better outcomes in comparison with the binary classification. A trustworthiness metric value of almost 1 proves that the HDS features of spectral observations are finely preserved in the latent space. Further access and availability to more spectral data points, especially in medical subdomains, can be an advantage in training a reliable model. Novel nonlinear activation functions can be explored in future to manage the high dimensional and nonlinear spectra more efficiently.

**Acknowledgements** We gratefully acknowledge VIT University for providing the resources and support

**Authors contributions** All authors contributed equally in formulation and execution of this work. Rajasekhara Babu was involved in planning and supervising the work and wrote the first draft of the manuscript. Mohamed Yousuff performed data collection, processed the experimental data, data analysis, and designed the figures. All authors read and approved the final manuscript.

**Data availability** The spectral datasets such as coffee, fresh meat, olive oil, and fruit purees are available online at <https://data.mendeley.com/datasets/frfv2yd9rg/1>. The COVID-19 Raman spectroscopy dataset is available on <https://doi.org/10.6084/m9.figshare.12159924.v1>.

## Declarations

**Competing interests** Authors declare no conflict of interest.

## References

- Al-Jowder O, Kemsley EK, Wilson RH (1997) Mid-infrared spectroscopy and authenticity problems in selected meats: a feasibility study. *Food Chem* 59(2):195–201
- Araújo DC, Veloso AA, de Oliveira Filho RS, Giraud M-N, Raniero LJ, Ferreira LM et al (2021) Finding reduced Raman spectroscopy fingerprint of skin samples for melanoma diagnosis through machine learning. *Artif Intell Med* 120:102161
- Barra I, Haefele SM, Sakrabani R, Kebede F (2021) Soil spectroscopy with the use of chemometrics, machine learning and pre-processing techniques in soil diagnosis: recent advances—a review. *TrAC Trends Anal Chem* 135:116166
- Bizzani M, William Menezes Flores D, Alberto Colnago L, David FM (2020) Monitoring of soluble pectin content in orange juice by means of MIR and TD-NMR spectroscopy combined with machine learning. *Food Chem* 332:127383
- Breiman L (2001) Random Forests. *Mach Learn* 45(1):5–32
- Böhm JN, Berens P, Kobak D (2022) Attraction–repulsion spectrum in neighbor embeddings. *J Mach Learn Res [Internet]* 23(95):1–32. Available from: <http://jmlr.org/papers/v23/21-0055.html>
- Chen H, Huang Q, Lin Z, Tan C (2022a) Detection of adulterants in medicinal products by infrared spectroscopy and ensemble of window extreme learning machine. *Microchem J* 173:107009
- Chen F, Sun C, Yue Z, Zhang Y, Xu W, Shabbir S et al (2022b) Screening ovarian cancers with Raman spectroscopy of blood plasma coupled with machine learning data processing. *Spectrochim Acta Part A Mol Biomol Spectrosc* 265:120355
- Chicco D, Tötsch N, Jurman G (2021) The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Min* 14(1):13
- Dong W, Moses C, Li K (2011) Efficient K-Nearest neighbor graph construction for generic similarity measures. In: Proceedings of the 20th international conference on world wide web. Association for Computing Machinery, pp 577–586
- Downey G, Briandet R, Wilson RH, Kemsley EK (1997) Near- and mid-infrared spectroscopies in food authentication: coffee varietal identification. *J Agric Food Chem* 45(11):4357–4361
- Dumancas GG, Ellis H (2022) Comprehensive examination and comparison of machine learning techniques for the quantitative determination of adulterants in honey using Fourier infrared spectroscopy with attenuated total reflectance accessory. *Spectrochim Acta Part A Mol Biomol Spectrosc* 276:121186
- Ellis DI, Broadhurst D, Goodacre R (2004) Rapid and quantitative detection of the microbial spoilage of beef by Fourier transform infrared spectroscopy and machine learning. *Anal Chim Acta* 514(2):193–201
- Fu X, Ying Y (2016) Food safety evaluation based on near infrared spectroscopy and imaging: a review. *Crit Rev Food Sci Nutr* 56(11):1913–1924
- Gao W, Zhou L, Liu S, Guan Y, Gao H, Hui B (2022) Machine learning prediction of lignin content in poplar with Raman spectroscopy. *Bioresour Technol* 348:126812
- Ghebleh Goydaragh M, Taghizadeh-Mehrjardi R, Golchin A, Asghar Jafarzadeh A, Lado M (2021) Predicting weathering indices in soils using FTIR spectra and random forest models. *Catena* 204:105437
- Ghojogh B, Ghodsi A, Karray F, Crowley M (2020) Stochastic neighbor embedding with Gaussian and Student-t distributions: tutorial and survey
- Gomes Marques de Freitas A, AlmirCavalcante Minho L, Elizabeth Alves de Magalhães B, Nei Lopes dos Santos W, Soares Santos L, Augusto de Albuquerque Fernandes S (2021) Infrared spectroscopy combined with random forest to determine tylosin residues in powdered milk. *Food Chem* 365:130477
- Hinton G, Roweis S (2002) Stochastic neighbor embedding. In: Proceedings of the 15th international conference on neural information processing systems. MIT Press, pp 857–864. (NIPS’02)
- Holland JK, Kemsley EK, Wilson RH (1998) Use of Fourier transform infrared spectroscopy and partial least squares regression for the detection of adulteration of strawberry purées. *J Sci Food Agric* 76(2):263–269
- Hu Q, Sellers C, Kwon JS-I, Wu H-J (2022) Integration of surface-enhanced Raman spectroscopy (SERS) and machine learning tools for coffee beverage classification. *Digit Chem Eng* 3:100020
- Hunter JD (2007) Matplotlib: a 2D graphics environment. *Comput Sci Eng* 9(3):90–95
- Khan S, Ullah R, Shahzad S, Javaid S, Khan A (2018) Optical screening of nasopharyngeal cancer using Raman spectroscopy and support vector machine. *Optik (Stuttg) [Internet]* 157:565–70. Available from: <https://www.sciencedirect.com/science/article/pii/S0030402617315176>
- Klambauer G, Unterthiner T, Mayr A, Hochreiter S (2017) Self-normalizing neural networks
- Kingma DP, Ba J (2015) Adam: a method for stochastic optimization. In: 3rd international conference for learning representations, San Diego

- Li Y, Chen S, Chen H, Guo P, Li T, Xu Q (2020) Effect of thermal oxidation on detection of adulteration at low concentrations in extra virgin olive oil: study based on laser-induced fluorescence spectroscopy combined with KPCA–LDA. *Food Chem* 309:125669
- Liu T, Li Z, Yu C, Qin Y (2017) NIRS feature extraction based on deep auto-encoder neural network. *Infrared Phys Technol* 87:124–128
- Liu D, Caliskan S, Rashidfarokhi B, Oldenhof H, Jung K, Sieme H et al (2021) Use of Fourier transform infrared spectroscopy combined with machine learning to detect oxidative damage in freeze-dried heart valve scaffolds. *Cryobiology* 103:160
- Luo N, Yang X, Sun C, Xing B, Han J, Zhao C (2021) Visualization of vibrational spectroscopy for agro-food samples using t-Distributed Stochastic neighbor embedding. *Food Control* 126:107812
- McKinney W, others (2010) Data structures for statistical computing in python. In: Proceedings of the 9th python in science conference, pp 51–56
- Mishra P, Nordon A, Tschannerl J, Lian G, Redfern S, Marshall S (2018) Near-infrared hyperspectral imaging for non-destructive classification of commercial tea products. *J Food Eng [Internet]* 238(January):70–7. Available from: <https://doi.org/10.1016/j.jfoodeng.2018.06.015>
- Mohamed Yousuff AR, RajasekharaBabu M (2020) Improving the accuracy of prediction of plant diseases using dimensionality reduction-based ensemble models. In: Venkata Krishna P, Mohammad Obaidat S (eds) *Emerging research in data engineering systems and computer communications*. Springer Singapore, pp 121–129
- Mohamed Yousuff AR, Rajasekhara Babu M (2022) Deep autoencoder based hybrid dimensionality reduction approach for classification of SERS for melanoma cancer diagnostics. *J Intell Fuzzy Syst*. Pre-Press:1–15.
- Owen S, Cureton S, Szuhan M, McCarten J, Arvanitis P, Ascione M et al (2021) Microplastic adulteration in homogenized fish and seafood - a mid-infrared and machine learning proof of concept. *Spectrochim Acta Part A Mol Biomol Spectrosc* 260:119985
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O et al (2011) Scikit-learn: machine learning in {P}ython. *J Mach Learn Res* 12:2825–2830
- Ralbovsky NM, Fitzgerald GS, McNay EC, Lednev IK (2021) Towards development of a novel screening method for identifying Alzheimer's disease risk: Raman spectroscopy of blood serum and machine learning. *Spectrochim Acta Part A Mol Biomol Spectrosc* 254:119603
- Schafer RW (2011) What is a Savitzky-Golay filter? [Lecture Notes]. *IEEE Signal Process Mag* 28(4):111–117
- Suleiman M, Abu-Aqil G, Sharaha U, Riesenber K, Lapidot I, Salman A et al (2022) Infra-red spectroscopy combined with machine learning algorithms enables early determination of *Pseudomonas aeruginosa*'s susceptibility to antibiotics. *Spectrochim Acta Part A Mol Biomol Spectrosc* 274:121080
- Sun H, Lv G, Mo J, Lv X, Du G, Liu Y (2019) Application of KPCA combined with SVM in Raman spectral discrimination. *Optik (Stuttg)* 184:214–219
- Tang J, Liu J, Zhang M, Mei Q (2016) Visualizing large-scale and high-dimensional data. In: Proceedings of the 25th international conference on world wide web. international world wide web conferences steering committee, pp 287–297
- Tapp HS, Defernez M, Kemsley EK (2003) FTIR spectroscopy and multivariate analysis can distinguish the geographic origin of extra virgin olive oils. *J Agric Food Chem* 51(21):6110–6115
- van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. *J Mach Learn Res [Internet]* 9:2579–605. Available from: <http://www.jmlr.org/papers/v9/vandermaaten08a.html>
- van der Walt S, Colbert SC, Varoquaux G (2011) The NumPy array: a structure for efficient numerical computation. *Comput Sci Eng* 13(2):22–30
- Venna J, Kaski S (2001) Neighborhood preservation in nonlinear projection methods: an experimental study. In: Dorffner G, Bischof H, Hornik K (eds) *Artificial neural networks -- ICANN 2001*. Springer Berlin Heidelberg, pp 485–491
- Venna J, Kaski S (2006) Local multidimensional scaling. *Neural Netw* 19(6):889–899
- Wang S, Liu S, Yuan Y, Zhang J, Wang Z, Che X (2020a) A novel CC-tSNE-SVR model for rapid determination of diesel fuel quality by near infrared spectroscopy. *Infrared Phys Technol* 106:103276
- Wang S, Liu S, Zhang J, Che X, Wang Z, Kong D (2020b) Feasibility study on prediction of gasoline octane number using NIR spectroscopy combined with manifold learning and neural network. *Spectrochim Acta Part A Mol Biomol Spectrosc* 228:117836
- Wang L, Huang Z, Wang R (2021a) Discrimination of cracked soybean seeds by near-infrared spectroscopy and random forest variable selection. *Infrared Phys Technol* 115:103731
- Wang Y, Huang H, Rudin C, Shaposhnik Y (2021b) Understanding how dimension reduction tools work: an empirical approach to deciphering t-SNE, UMAP, TriMap, and PaCMAP for Data Visualization. *J Mach Learn Res [Internet]* 22(201):1–73. Available from: <http://jmlr.org/papers/v22/20-1061.html>
- Waskom M, Botvinnik O, Hobson P, Cole JB, Halchenko Y, Hoyer S et al (2014) seaborn: v0.5.0 (November 2014) [Internet]. Zenodo. Available from: <https://doi.org/10.5281/zenodo.12710>
- Wójtowicz A, Piekarczyk J, Czernecki B, Ratajkiewicz H (2021) A random forest model for the classification of wheat and rye leaf rust symptoms based on pure spectra at leaf scale. *J Photochem Photobiol B Biol* 223:112278
- Yan S, Wang S, Qiu J, Li M, Li D, Xu D et al (2021) Raman spectroscopy combined with machine learning for rapid detection of food-borne pathogens at the single-cell level. *Talanta* 226:122195
- Yin G, Li L, Lu S, Yin Y, Su Y, Zeng Y et al (2020) Data and code on serum Raman spectroscopy as an efficient primary screening of coronavirus disease in 2019 (COVID-19). Available from: [https://figshare.com/articles/dataset/Data\\_and\\_code\\_on\\_serum\\_Raman\\_spectroscopy\\_as\\_an\\_efficient\\_primary\\_screening\\_of\\_coronavirus\\_disease\\_in\\_2019\\_COVID-19/12159924](https://figshare.com/articles/dataset/Data_and_code_on_serum_Raman_spectroscopy_as_an_efficient_primary_screening_of_coronavirus_disease_in_2019_COVID-19/12159924)
- Zhang L, Li C, Peng D, Yi X, He S, Liu F et al (2022) Raman spectroscopy and machine learning for the classification of breast cancers. *Spectrochim Acta Part A Mol Biomol Spectrosc* 264:120300
- Zhao H, Zhan Y, Xu Z, John Nduwamungu J, Zhou Y, Powers R et al (2022) The application of machine-learning and Raman spectroscopy for the rapid detection of edible oils type and adulteration. *Food Chem* 373:131471
- Zheng W, Fu X, Ying Y (2017) Similar offspring voting genetic algorithm for spectral variable selection. *J Chemom* 31(7):e2893
- Zheng W, Shu H, Tang H, Zhang H (2019) Spectra data classification with kernel extreme learning machine. *Chemom Intell Lab Syst* 192:103815
- Zhou Y, Zuo Z, Xu F, Wang Y (2020) Origin identification of Panax notoginseng by multi-sensor information fusion strategy of infrared spectra combined with random forest. *Spectrochim Acta Part A Mol Biomol Spectrosc* 226:117619

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.