Article:
Khadem, H. orcid.org/0000-0002-6878-875X, Nemat, H., Elliott, J. orcid.org/0000-0002-7867-9987 et al. (1 more author) (2022) Signal fragmentation based feature vector generation in a model agnostic framework with application to glucose quantification using absorption spectroscopy. Talanta, 243. 123379. ISSN 0039-9140

https://doi.org/10.1016/j.talanta.2022.123379

Signal fragmentation based feature vector generation in a model agnostic framework with application to glucose quantification using absorption spectroscopy

Heydar Khadem [a,*], Hoda Nemat [a], Jackie Elliott [b], Mohammed Benaissa [a]

[a] Department of Electronic and Electrical Engineering, University of Sheffield, UK, [b] Department of Oncology and Metabolism, University of Sheffield, UK, * Corresponding author

E-mail addresses: h.khadem@sheffield.ac.uk, hoda.nemat@sheffield.ac.uk, j.elliott@sheffield.ac.uk, m.benaissa@sheffield.ac.uk

ABSTRACT

This paper proposes feature vector generation based on signal fragmentation equipped with a model interpretation module to enhance glucose quantification from absorption spectroscopy signals. For this purpose, near-infrared (NIR) and mid-infrared (MIR) spectra collected from experimental samples of varying glucose concentrations are scrutinised. Initially, a given spectrum is optimally dissected into several fragments. A base-learner then studies the obtained fragments individually to estimate the reference glucose concentration from each fragment. Subsequently, the resultant estimates from all fragments are stacked, forming a feature vector for the original spectrum. Afterwards, a meta-learner studies the generated feature vector to yield a final estimation of the reference glucose concentration pertaining to the entire original spectrum. The reliability of the proposed approach is reviewed under a set of circumstances encompassing modelling upon NIR or MIR signals alone and combinations of NIR and MIR signals at different fusion levels. In addition, the compatibility of the proposed approach with an underlying preprocessing technique in spectroscopy is assessed. The results substantiate the utility of incorporating the designed feature vector generator into standard benchmarked modelling procedures under all considered scenarios. Finally, to

promote the transparency and adoption of the propositions, SHapley additive exPlanations (SHAP) is leveraged to interpret the quantification outcomes.

Keywords: Glucose quantification; Near-infrared spectroscopy; Mid-infrared spectroscopy; Machine learning, SHAP

1. Introduction

In vitro glucose quantification has practical applications in a variety of areas, e.g., food science, biology, and botany [1–4] Consequently, continued research is underway to expand this area of knowledge [5].

In this context, two optical modalities of near-infrared (NIR) and mid-infrared (MIR) have been broadly pursued in glucose quantification studies [6,7]. NIR and MIR signals are within the wavelength range of 750–2500 nm and 2500–10000 nm, respectively [8]. One advantage of using these technologies for glucose sensing is that the absence of reagents makes them economically appropriate for regular measurements [9].

NIR light possesses a high penetration rate enabling it to enter deeper parts of opaque compounds to seek the glucose trace [10–12]. On the other hand, the MIR region includes sharp peaks of glucose [13]. Of other merits of MIR spectroscopy for glucose sensing are the attenuated scattering phenomena and intensified absorption due to longer wavelengths [14]. Hence, there are stimuli to investigate glucose quantification from the combination of NIR and MIR spectra, as well.

As NIR/MIR light is traversing through an object and as a result of the interaction with physiological compounds of the object, some beams frequencies get scattered and absorbed [15,16]. These absorption and scattering patterns could be scrutinised using appropriate tools to derive information concerning the analyte(s) of interest [8]. Specifically, machine learning (ML) multivariate calibration algorithms, in particular, partial least squares regression (PLSR), are typically suggested for quantifying glucose from recorded NIR/MIR spectra [17,18].

Notwithstanding the general suitability of such algorithms, further advancements in the analysis are necessary towards achieving decisive glucose quantifications from NIR/MIR spectra [19,20]. In this regard, scopes exist to enhance the accuracy of the analysis by exploiting state-of-the-art ML techniques such as stack learning. Stack learning is an ensemble method for improving the competence of ML models in which a meta-learner integrates outputs of multiple base-learner to produce a final output [21].

In conjunction with algorithms like stack learning, model interpretation frameworks could also be incorporated to expand the clarity of the analysis and further support the findings [22,23]. In this respect, SHapley additive exPlanations (SHAP) is an elaborate game-theoretic model agnostic approach to explain ML models [24]. SHAP joins optimal credit allocation with local explanations via the concept of Shapley values from cooperative game theory [25]. Resultant SHAP values designate the contribution of attributes to deviations from average estimations, a measurement to elucidate the effect of individual features on models' outputs [24].

This article suggests signal fragmentation based feature vector generation (SFFVG) dressed with model interpretations for in vitro glucose estimation upon absorbance spectroscopy data. First, a given signal was efficiently segmented into a number of sub-signals. The sub-signals were then autonomously investigated using a base-learner to estimate the reference glucose concentration. These fragmentary estimations were thereafter concatenated, forming a feature vector for the given signal. A meta-learner, utilising the concept of stack learning, later aggregated the generated feature vector's elements, creating an estimation related to the entire signal. The flexibility of the proposed approach was monitored by implementing it on NIR signals, MIR signals, and the fusion of NIR and MIR signals. Furthermore, the compatibility of the method with a conventional preprocessing technique in spectroscopy was examined. Finally, to spur the adoption of the propositions by increasing the clarity of the analysis, SHAP was carried out to delineate the influence of constructed features on the formation of final estimations.

2. Material and methods

For glucose quantification, this research used a dataset consisting of NIR and MIR spectra related to100 mixture solutions of various glucose concentrations of 5–500 mg dL^{-1}, at 5 mg dL^{-1} intervals [26]. SFFVG was proposed to advance glucose quantification from these absorption spectroscopic data. The effectiveness of the proposed method was examined within six different modelling strategies and with and without including a classical preprocessing technique. Finally, to extend the transparency of the proposed method, SHAP was deployed to interpret the created models. The dataset and details of implementation steps are described in this section.

2.1. Experimental data

The experimental samples were prepared at the laboratories of the Department of Chemistry, University of Sheffield, Sheffield, UK. Two aqueous solutions were prepared with the same volume (0.5 l), pH (7.4), phosphate (0.01 M/dl), and human serum albumin (5 g/dl), where the first solution contained glucose as well (500 mg dL^{-1}), but the second solution did not contain glucose. 5 ml of the first solution (with glucose) was extracted and preserved in a sealed tube, forming the first sample (glucose 500 mg dL^{-1}). Then, 5 ml of the second solution (without glucose) was added to the first solution, reducing its glucose concentration to 495 mg dL-1. Similarly, storing 5 ml of the first solution in another sealed tube, the second sample (glucose 495 mg dL^{-1}) was acquired. The removed amount from the first solution was again replaced with 5 ml of the second solution, decreasing the glucose concentration of the first solution to 490 mg dL^{-1}. The same stages were recured to obtain 100 samples with 5–500 mg dL^{-1} glucose concentrations in 5 mg dL^{-1} increments.

Spectra were collected using a Fourier transform infrared spectrometer (PerkinElmer Inc., USA) in uncontrolled laboratory conditions at the Department of Materials Science and Engineering, University of Sheffield, Sheffield, UK. The sensing lens of the device was cleaned utilising ethanol wipe prior to placing each sample for recording. After that, the entire surface of the lens was overlaid with a layer of the sample.

The spectrometer then recorded the absorption signals with the attenuated total reflection technique. The recorded spectra laid in the wavelength range of 2100–8000 nm (1.7 nm resolution). The wavelengths within 2100–2500 nm and 2500-8000 nm were part of the NIR and MIR region, respectively. To achieve authentic spectra, the spectrometer was configured to take four readings for each sample and return the average as the output [18]. Some of the collected raw spectra are displayed in Figure 1.

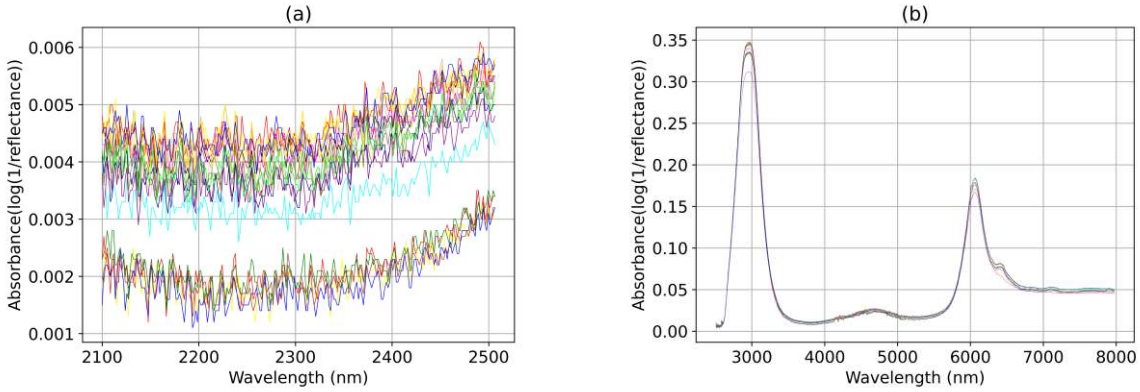


Figure 1. 20 randomly selected raw spectra collected from chemical samples (a) NIR signals, (b) MIR signals.

2.2. Calibration-validation split

For creating quantitative models, 80% of the data points were randomly selected and allocated as the calibration set, and the remaining 20% were considered as the validation set. Table 1 summarises some statistical characteristics of the calibration and validation set. All subsequent model training and hyperparameter tuning operations were carried out using only the calibration set, whereas the validation set remained unseen for evaluation and model interpretation analysis.

Table 1. Characteristics of the calibration and validation set.

	Samples	Mean (mg dL^{-1})	Standard Deviation (mg dL^{-1})
Calibration set	80	250.3	146.4
Validation set	20	261.2	135.2

2.3. Feature vector generation

Figure 2 depicts the block diagram of SFFVG consisting of a signal fragmentation, regression, and concatenation unit. In the first step, the fragmentation unit efficiently breaks signals into several intervals. After that, the regression unit studies the obtained fragments independently to produce a corresponding fragmentary estimation of the reference glucose concentration. It should be noted that this regression block is trained separately for each interval using the corresponding fragments from the calibration set. Finally, the concatenation unit stack the outputs of the regression unit, forming a feature vector for the original input signal.

The fragmentation unit was optimised for three separate scenarios depending on input data: NIR signals, MIR signals, or concatenation of NIR and MIR signals (hereafter referred to as NIR-MIR signals). For simplicity, equidistant fragmentation was considered, and signals were inputted in raw form. Values of 1 to 20 were explored as the number of intervals, and the one resulting in estimations (by the regression unit) with the lowest root mean square error (*RMSE*) of five-fold cross-validation on the calibration set was selected.

For the regressors block, PLSR was assigned, which previously has demonstrated to be an excellent method in spectroscopic data analysis [27,28]. For tunning the number of PLSR components, values of 1 to max (10, the length of the input variable) were sought, and the one delivering the minimum *RMSE* of glucose quantification based on five-fold cross-validation on the calibration set was decided.

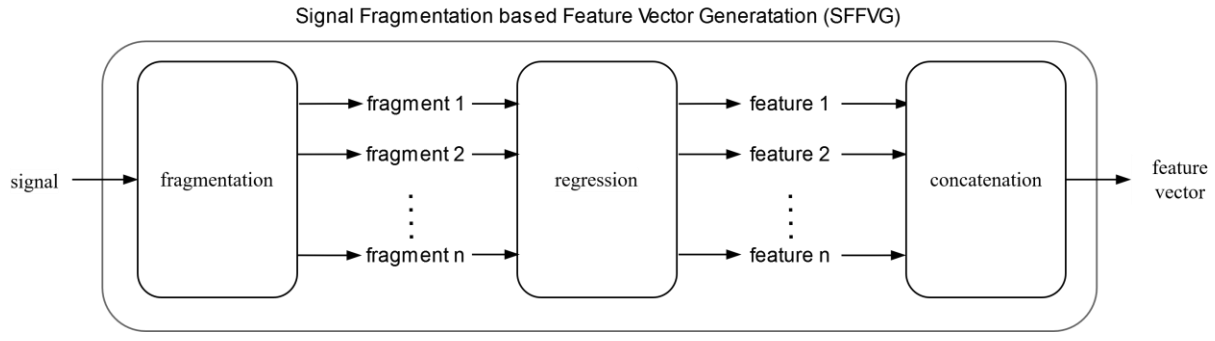


Figure 2. The general scheme of the proposed signal fragmentation based feature vector generation (SFFVG) method consists of signal fragmentation, regression, and concatenation. The input spectrum is optimally divided into a number of fragments. Next,

each fragment is used as the input of a regressor (partial least square regression) to estimate the glucose concentration. Outputs of regressors were then stacked according to the order of the relevant fragments to form a generated feature vector.

2.4. Chemometric

In this work, for creating glucose quantification models, we assigned six different modelling strategies with the general block diagrams exhibited in Figure 3 [29–32]. As can be observed, SFFVG was a building block of all considered strategies.

For preprocessing units in the structure of the strategies shown in the figure, Savitzky-Golay (SG) smoothing filter was considered with a second-order polynomial and a five-point window [33,34]. Including preprocessing was to examine the compatibility of SFFVG with this prominent stage in spectroscopy. For regression units, PLSR was appointed with the same tuning process described in subsection 2.2.

It is distinguishable from the block diagram that the first two modelling strategies were unimodal, where only NIR or MIR spectra took part in the modelling process. In contrast, the other four strategies were bimodal, utilising both NIR and MIR signals. Also, each dashed block in Figure 3 signifies two possible model creation scenarios for the associated modelling strategy by incorporating or not incorporating that particular block. Therefore, the working mechanism of modelling strategies was as follows.

a) NIR Modelling (Figure 3a): the raw or preprocessed form of a given NIR signal or their feature vector were input to a regression unit for making a final glucose estimation.
b) MIR Modelling (Figure 3b): this strategy was akin to NIR Modelling, except MIR signals were studied instead of NIR.
c) Raw Spectra Fusion Modelling (Figure 3c): a given NIR and MIR signal was first concatenated, forming an NIR-MIR signal. Raw or preprocessed form of the NIR-MIR signal or their feature vector was then fed to a regressor, creating a final estimation.

d) Preprocessed Spectra Fusion Modelling (Figure 3d): initially, a given NIR and MIR signal were separately preprocessed and then mixed. The resultant NIR-MIR signal or its feature vector were then given to a regressor to accomplish a final estimation.

e) Feature Fusion Modelling (Figure 3e): first, feature vectors were generated distinctly from the raw or preprocessed form of a given NIR and MIR signal and thereafter coupled. The obtained combined feature vector was then input to a regressor, making a final quantification.

f) Decision Fusion Modelling (Figure 3f): estimations created individually using a given NIR and MIR signal were ensembled by a regressor to generate a final estimation.

The goal of including different strategies was to comprehensively investigate the robustness of SFFVG under diverse circumstances. The idea was to generate quantitative models according to all possible permutations for each strategy and later perform intra-strategy comparisons between models with SFFVG and those without SFFVG as benchmarks. Thus, by incorporating or skipping preprocessing and SFFVG blocks each, four models were constructed using each of NIR Modelling, MIR Modelling, Raw Spectra Fusion Modelling, and Decision Fusion Modelling. On the other hand, two models were created through each of Preprocessed Spectra Fusion Modelling and Feature Fusion Modelling by incorporating or not incorporating their sole dashed unit. It is worth clarifying that the preprocessing unit in Preprocessed Spectra Fusion Modelling and SFFVG unit in Feature Fusion Modelling was not skippable due to the essence of these strategies.

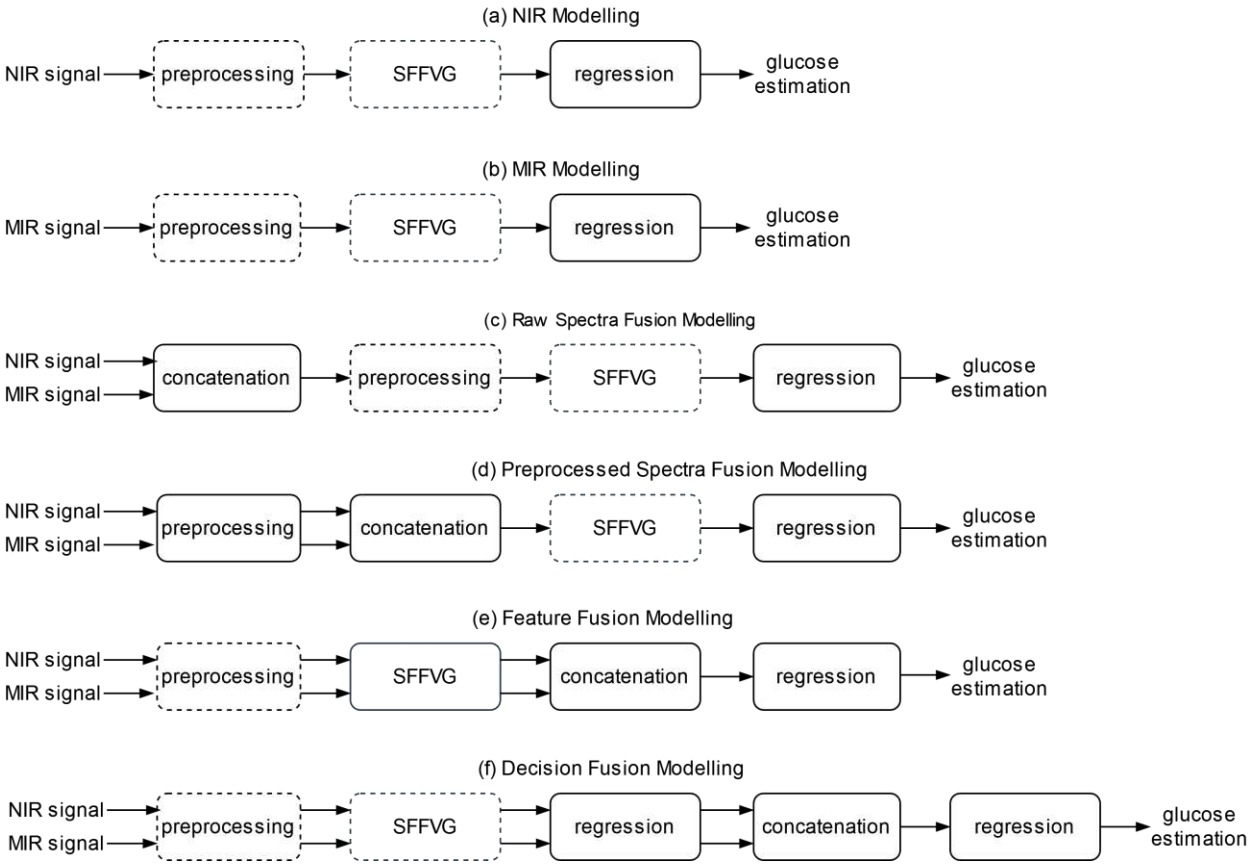


Figure 3. The general block diagram of the six considered strategies for creating glucose estimation models.
Note. For the preprocessing and regression blocks, the Savitzky-Golay filter and partial least square regression were used, respectively. SFFVG (signal fragmentation based feature vector generation) block's internal architecture is shown in Figure 2. The dashed blocks indicate that both conditions with or without including the block were investigated separately. (a, b) NIR Modelling and MIR Modelling, two unimodal strategies where glucose concentrations were estimated from NIR or MIR signals alone. (c) Raw Spectra Fusion Modelling where NIR and MIR data were fused in their raw format and then used to create quantitative models, (d) Preprocessed Spectra Fusion Modelling where NIR and MIR signals were fused after the preprocessing and then used for constructing quantitative models, (e) Feature Fusion Modelling where features generated from NIR and MIR signal were fused and used to create quantitative models, and (f) Decision Fusion Modelling where quantitative models created from NIR and MIR signals were ensembled to form a combined model.

2.5. Model evaluation

The developed models were evaluated considering three frequently used regression metrics for estimations one the evaluation set; *RMSE* as Eq. (1) and mean absolute percentage deviation (*MAPD*) as Eq. (2) to reflect the error of quantifications [35], [36], and coefficient of determination (r^2) as Eq. (3) as a statistical measure to indicate correlations between the reference and estimated values [37].

$$RMSE = \sqrt{(\sum_{i=1}^{N}(y_i - f_i)^2)/N} \quad (1)$$

$$MAPD = ((\sum_{i=1}^{N}|(y_i - f_i)/y_i|)/N) \times 100 \quad (2)$$

$$r^2 = 1 - (RSS/TSS) \quad (3)$$

where, in Eqs. (1) and (2), N, y_i, and f_i are respectively the size of the evaluation set, actual value, and estimated value; and in Eq. (3), *RSS* and *TSS* respectively represent the residual sum of squares and the total sum of squares.

2.6. Model interpretation

SHAP is a game-theoric ML explainability technique. It stimulates how an ML model produces an estimation for a data instance as a game between input variables. Then, using the Shapley value concept from game theory [25], each input variable's contribution to generated estimation for the data instance is quantified as Eq. (4) [24].

$$SHAP_x(f) = \sum_{F:f\in F}\left(|F| \times \binom{f}{|F|}\right)^{-1} \times \left(\hat{x}_F - \hat{x}_{F\setminus f}\right) \quad (4)$$

where f is a given input variable; *x* is a given instance of data; $SHAP_x(f)$ represents the quantified contribution level of variable *f* in the generated estimation for *x (*SHAP value of variable *f* for *x); F* represents all possible subsets of variables with *f* included; $|F|$ is the size of *F* (number of variables in *F*); $\hat{x}_F$ represents the model's estimation for *x* from *F;* $\hat{x}_{F\setminus f}$ is the model's estimation for *x* from *F* excluding *f*

Following the evaluation analysis, SHAP was deployed to globally interpret models assimilating SFFVG, i.e., explaining the impact of the features generated from sub-signals in producing the estimations across the entire validation set. For this purpose, the mean absolute of features' SHAP values presented in Eq. (4) was used.

This analysis allows analogies to be drawn between the importance of the segregated intervals, thereby increasing the transparency of investigations utilising SFFVG. For conciseness, interpretation was

undertaken only for the best model generated by each modelling strategy according to their evaluation results presented later in subsection 4.1.

3. Results and discussion

This section reports the evaluation and model interpretation results alongside the corresponding discussion.

3.1. Signal fragmentation

Based on optimisation for the number of intervals, NIR signals were divided into four equal fragments (100 nm wide apiece), MIR signals into six equal fragments (≈ 916 nm wide apiece), and NIR-MIR signals into ten equal fragments (590 nm wide apiece). The results of spectra fragmentation are summarised in Table 2.

Consequently, for NIR Modelling, the fragmentation module divided signals into the four NIR intervals represented in the table and then the corresponding NIR features were extracted from these intervals. Similarly, for MIR Modelling, the signals were divided into the six MIR intervals shown in the table, and then the corresponding MIR features were generated. For Raw Spectra Fusion Modelling and Preprocessed Spectra Fusion Modelling, signals were divided into the ten NIR-MIR intervals shown in the table and then the associated NIR-MIR features were constructed. Finally, for Feature Fusion Modelling and Decision Fusion Modelling, NIR and MIR signals were separately fragmented into respectively the four NIR intervals and the six MIR intervals presented in the table, and then the relevant features were created.

Table 2. Signal fragmentation process outcomes including the generated intervals their associated name and feature.

Region	Interval	Interval name	Generated feature
NIR	2100–2200 nm	NIR interval 1	NIR feature 1
	2200–2300 nm	NIR interval 2	NIR feature 2
	2300–2400 nm	NIR interval 3	NIR feature 3
	2400–2500 nm	NIR interval 4	NIR feature 4
MIR	2500–3416 nm	MIR interval 1	MIR feature 1
	3416–4333 nm	MIR interval 2	MIR feature 2
	4334–5250 nm	MIR interval 3	MIR feature 3
	5250–6166 nm	MIR interval 4	MIR feature 4
	6166–7084 nm	MIR interval 5	MIR feature 5
	7084–8000 nm	MIR interval 6	MIR feature 6
NIR-MIR	2100–2690 nm	NIR-MIR interval 1	NIR-MIR feature 1
	2690–3280 nm	NIR-MIR interval 2	NIR-MIR feature 2

	3280–3870 nm	NIR-MIR interval 3	NIR-MIR feature 3
	3870–4460 nm	NIR-MIR interval 4	NIR-MIR feature 4
	4460–5050 nm	NIR-MIR interval 5	NIR-MIR feature 5
	5050–5640 nm	NIR-MIR interval 6	NIR-MIR feature 6
	5640–6230 nm	NIR-MIR interval 7	NIR-MIR feature 7
	6230–6820 nm	NIR-MIR interval 8	NIR-MIR feature 8
	6820–7410 nm	NIR-MIR interval 9	NIR-MIR feature 9
	7410–8000 nm	NIR-MIR interval 10	NIR-MIR feature 10

Note. NIR: near-infrared; MIR: mid-infrared; NIR-MIR: the combination of near- and mid-infrared.

3.2. Model evaluation

Table 3 lists the results of *RMSE*, *MAPD* and r^2 for all created models. The table is compartmentalised with the results of the modelling strategies to facilitate intra-strategy comparisons of SFFVG-included models versus non-SFFVG models.

Each improvement ratio in Table 3 compares the result of an evaluation metric achieved by an SFFVG-included model versus the model with the same strategy and preprocessing but without SFVG (benchmarked model), reported in the raw above. According to the table, in all pair-wise comparisons, the majority of improvement ratios convey the efficacy of SFFVG-included models over non-SFFVG counterparts.

The values in bold in Table 3 are the best result(s) obtained for evaluation metrics through each modelling strategy. Grey cells in the table highlight the model(s) with the highest number of best values for the evaluation metrics amongst the models created using the same modelling strategy. These highlights indicate that the best model of all strategies was SFFVG-included. Explicitly, applying SFFVG without preprocessing granted the highest overall performance with the best *MAPD* and r^2 values for the NIR Modelling, whilst applying SFFVG with preprocessing gave the lowest *RMSE* in this case. In MIR Modelling, SFFVG without preprocessing yielded the best overall results and for all evaluation metrics. Moreover, for Raw Spectra Fusion Modelling, Preprocessed Data Fusion Modelling, and Feature Fusion Modelling, SFFVG joined with preprocessing conferred the best performance overall and according to each evaluation criterion. Finally, the best performance for Decision Fusion Modelling was achieved by SFFVG

without preprocessing, the best *RMSE* and r^2, whilst the best *MAPD* was for non-SFFVG with preprocessing.

Overall, incorporating SFFVG in the six modelling strategies enhanced the accuracy of glucose estimation. Moreover, SFFVG maintained its effectiveness when a preprocessing step was also present in the modelling process. Such attainments underpin the functionality and flexibility of the proposed SFFVG approach. The coordinating power of stack learning could justify such fulfilments; deriving glucose information from fragments of a signal and then aggregating the outcomes dominated studying the whole signal at once. Finally, it is noteworthy that pre-partitioning procedures have recently found successful applications in image processing tasks, supporting the relevance of the core idea involved in this work to other areas where further exploration would be desirable [38,39].

Comparing bimodal strategies with MIR Modelling reveals that Raw Spectra Fusion Modelling, Preprocessed Data Fusion Modelling, and Feature Fusion Modelling produced results on par with MIR modelling whilst not conclusively outperforming it. Nevertheless, rather than taking advantage of synergistic effects, the object of including bimodal strategies in this work was to test SFFVG's capability under a broader range of spectra with different data fusion strategies.

Table 3. Evaluation results for all generated quantitative models.

Strategy	Preprocessing	SFFVG	RMSE (mg dL^{-1})	RMSE IR (%)	MAPD (%)	MAPD IR (%)	r^2	r^2 IR (%)
NIR Modelling	No	No	98.1	—	66.6	—	0.47	—
		Yes	91.0	+7.2	**46.5**	+30.1	**0.58**	+23.4
	Yes	No	98.7	—	67.3	—	0.46	—
		Yes	**89.7**	+9.1	48.0	+28.6	0.55	+19.5
MIR Modelling	No	No	36.3	—	28.0	—	0.92	—
		Yes	**24.5**	+32.5	**24.4**	+12.8	**0.96**	+4.3
	Yes	No	36.2	—	27.8	—	0.92	—
		Yes	24.6	+32.0	24.7	+11.1	**0.96**	+4.3
Raw Spectra Fusion Modelling	No	No	34.4	—	28.8	—	0.93	—
		Yes	32.3	+6.1	26.8	+6.9	**0.94**	+1.0
	Yes	No	34.5	—	29.0	—	0.93	—
		Yes	**32.1**	+6.6	**25.6**	+11.7	**0.94**	+1.0
Preprocessed Spectra Fusion Modelling	Yes	No	34.2	—	28.7	—	0.93	—
		Yes	**32.2**	+5.8	**25.7**	+10.4	**0.94**	+1.0
Feature Fusion Modelling	No	Yes	27.7	—	25.1	—	0.95	—
	Yes	Yes	**26.6**	+3.9	**24.1**	+3.9	**0.96**	+1.0
Decision Fusion Modelling	No	No	60.1	—	35.2	—	0.80	—
		Yes	**47.0**	+21.7	37.0	-5.1	**0.87**	+8.7
	Yes	No	60.0	—	**35.0**	—	0.81	—
		Yes	49.6	+17.3	38.3	-8.6	0.86	+5.8

Note. SFFVG: signal fragmentation based feature vector generation; RMSE: root mean square error; IR: improvement ratio (comparing the results of an SFFVG-included model versus the benchmarked non-SFFVG model reported in the raw above.); MAPD: mean absolute percentage deviation; r2: coefficient of determination; NIR: near-infrared; MIR: mid-infrared. The values in bold font indicate the best result for each evaluation metric in each strategy. The grey cells indicate the model(s) with the highest number of best-obtained evaluation metrics amongst models developed using the same modelling strategy.

3.3. Model interpretation

Figure 4 represents the variable importance graphs for the best model of each strategy (marked with grey cells in Table 3). The length of each bar in the graphs expresses the importance rate of the corresponding feature according to mean absolute SHAP values over the entire validation set.

As presented in Figure 4a, NIR feature 4 (associated with interval 2400–2500 nm) was the most informative variable for the best model from NIR Modelling. NIR features 1, 3, and 2 (intervals 2100–2200 nm, 2300–2400 nm, and 2200–2300 nm, respectively) in order placed in ranks 2 to 4.

According to Figure 4b, MIR feature 1 (interval 2500–3416 nm) had the dominant influence on the best model of MIR Modelling with a mean absolute SHAP value remarkably superior to others. In contrast, MIR feature 2 (interval 3416–4333 nm) carried the most inferior influence with a mean absolute SHAP value considerably lower than others. In comparison, MIR features 3, 4, 5, and 6 (intervals 4333–5250 nm, 5250–6166 nm, 6166–7084 nm, and 7084–8000 nm, respectively) induced comparable and medium impacts on the model.

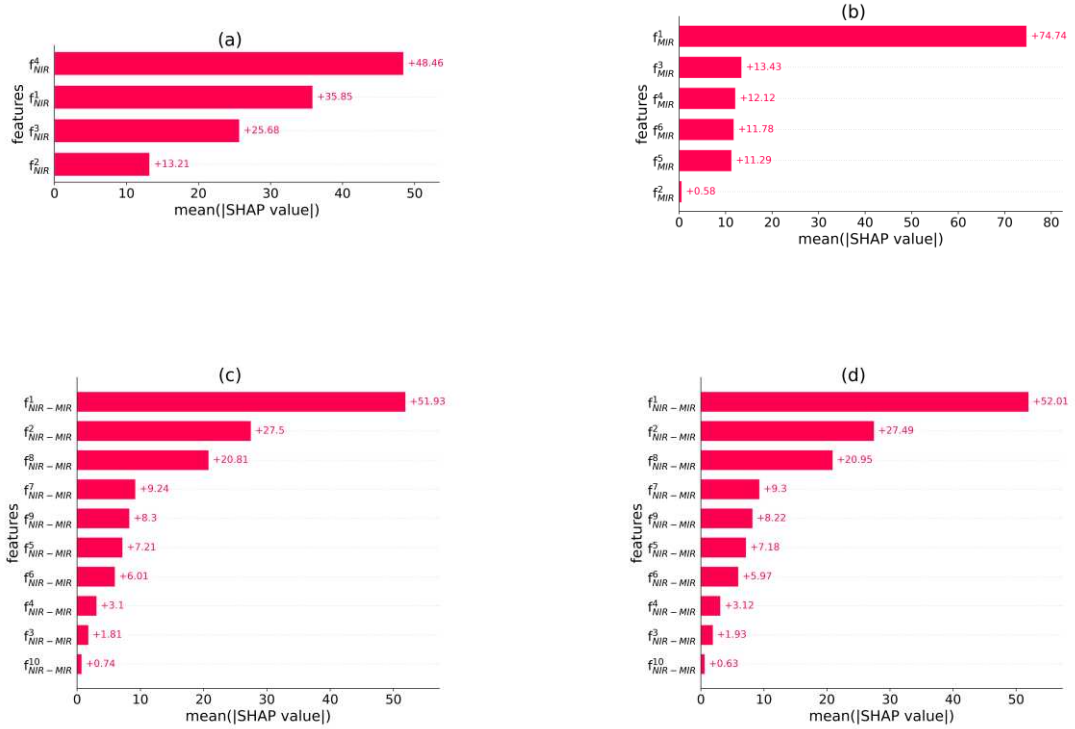
For the best model of Raw Spectra Fusion Modelling (Figure 4c) and Preprocessed Spectra Fusion Modelling (Figure 4d), NIR-MIR feature 1 (interval 2100–2690 nm) supplied the maximum impact on the model with a mean absolute SHAP value appreciably higher than others. NIR-MIR features 2 and 8 (intervals 2690–3280 nm and 6230–6820 nm, respectively) placed the second and third rank. Other features had relatively subordinate effects.

Based on Figure 4e, for the best model of Feature Fusion Modelling, the impact of MIR feature 1 (interval 2500–3416 nm) outweighed that of other MIR and NIR features with a mean absolute SHAP value markedly higher than others. MIR feature 3 (interval 4333–5250 nm) and NIR feature 4 (interval

2400–2500 nm) placed in the second and third rank. The other three NIR features (1, 2, and 3) had the weakest impact on the model.

Figure 4f displays the variable importance plot for the best model of Decision Fusion Modelling. In this case, since decisions of NIR and MIR models were combined at the final stage, the effect of NIR and MIR decisions on the final estimations were compared. The results illustrate that the influence of MIR decisions on the models' outcomes surpassed NIR decisions.

According to the interpretation analysis, potential associations between the most informative intervals detected and the nearest glucose-informative bands according to the ordinary glucose signature in NIR and MIR regions could be inferred [40]. For instance, information possessed by the most influential features in different regions were potentially connected to the following vibrations in glucose molecules bonds: a combination of vibrations in CH and CH2 bonds for NIR feature 4, stretching vibrations in OH and CH bonds for MIR feature 1, a combination of vibrations in OH, CH, and CH2 bonds for NIR-MIR feature 1.



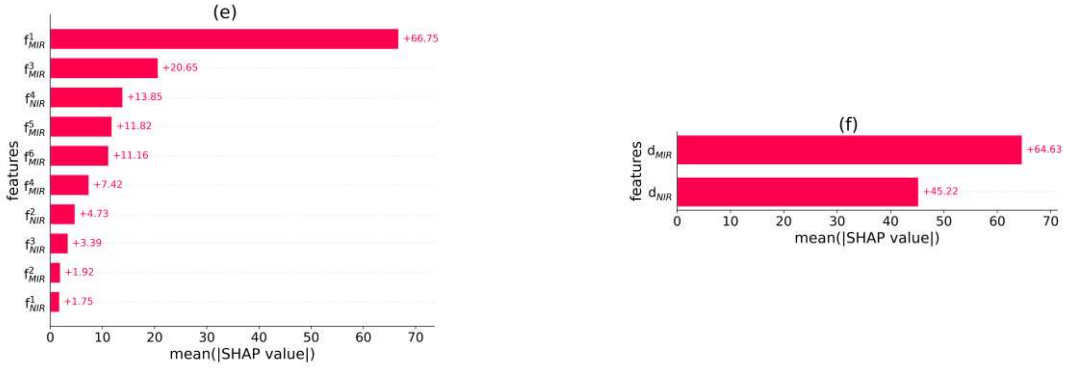


Figure 4. Feature importance plots which indicate the influence of variables upon collective absolute (SHapley Additive exPlanations) SHAP values for the best model from each modelling strategy (a) NIR modelling, (b) MIR modelling, (c) raw spectra fusion modelling, (d) preprocessed spectra fusion modelling, (e) feature fusion modelling, and (f) decision fusion modelling.
Note. f_s^i: feature generated from the ith fragment of s signals. d_s: decision from s signals.

3.4. Complementary analysis

3.4.1. Comparative analysis

Interval partial least squares (iPLS) is a well-known variable selection technique in spectroscopy data analysis [40]. The technique starts with breaking signals into several intervals. Then, some of these intervals are selected for subsequent modelling analysis, where the selected intervals are stacked and inputted into a prediction model.

We conducted a basic comparative analysis between iPLS and the proposed SFFVG method. To this end, using the same block diagram shown in Figure 3, for each model with SFFVG, a comparator model with iPLS was constructed. The building blocks of each comparator model was similar to its reference model, except the SFFVG unit was replaced with an iPLS unit. For the sake of fair comparisons, the units of each comparator model underwent the same optimisation process performed for its reference model's units. As a result of identical fragmentation optimisation, the same intervals represented in Table 2 were utilised for comparator models with iPLS.

In iPLS analysis, first, an autonomous glucose quantification model was created for each interval. The interval providing the lowest RMSE of five-fold cross-validation on the calibration set was selected. Then,

the selected interval combined with the remaining intervals, one at a time, were used to build quantitative models. The combination of intervals that produced the model with the lowest RMSE of five-fold cross-validation on the calibration set was selected. This successive interval selection cycle was repeated until adding a new interval could not lower the RMSE of five-fold cross-validation on the calibration set.

Table 4 presents the results of the comparative analysis of SFFVG and the iPLS. Values in bold and grey cells indicate the same information as explained in subsection 3.2 for Table 3. According to the grey cells in the table, four of the models with the dominant number of best-obtained evaluation metrics (amongst the models generated using each modelling strategy) were with SFFVG. These outcomes further support the capability of SFFVG.

Table 4. Evaluation results for the comparision analysis between SFFVG and iPLS.

Strategy	Preprocessing	Feature enineeing	RMSE (mg dL^{-1})	MAPD (%)	r^2
NIR Modelling	No	SFFVG	91.0	**46.5**	**0.58**
		iPLS	99.9	69.7	0.45
	Yes	SFFVG	**89.7**	48.0	0.55
		iPLS	99.9	71.3	0.44
MIR Modelling	No	SFFVG	**24.5**	24.4	**0.96**
		iPLS	29.3	18.5	0.95
	Yes	SFFVG	24.6	24.7	**0.96**
		iPLS	28.7	**18.4**	**0.96**
Raw Spectra Fusion Modelling	No	SFFVG	32.3	26.8	**0.94**
		iPLS	33.7	19.2	0.93
	Yes	SFFVG	32.1	25.6	**0.94**
		iPLS	**31.8**	**17.7**	**0.94**
Preprocessed Spectra Fusion Modelling	Yes	SFFVG	32.2	25.7	**0.94**
		iPLS	**31.8**	**17.7**	**0.94**
Feature Fusion Modelling	No	SFFVG	27.7	25.1	0.95
		iPLS	29.5	16.8	0.95
	Yes	SFFVG	**26.6**	24.1	**0.96**
		iPLS	29.1	**16.1**	0.95
Decision Fusion Modelling	No	SFFVG	**47.0**	**37.0**	**0.87**
		iPLS	49.6	38.4	0.86
	Yes	SFFVG	49.6	38.3	0.86
		iPLS	49.5	38.5	0.86

Note. iPLS: interval partial least squares; SFFVG: signal fragmentation based feature vector generation; RMSE: root mean square error; IR: improvement ratio (comparing the results of an SFFVG-included model versus the benchmarked non-SFFVG model reported in the raw above.); MAPD: mean absolute percentage deviation; r2: coefficient of determination; NIR: near-infrared; MIR: mid-infrared. The values in bold font indicate the best result for each evaluation metric in each strategy. The grey cells indicate the model(s) with the highest number of best-obtained evaluation metrics amongst models developed using the same modelling strategy.

3.4.2. Reevaluation analysis

To further examine the functionality of SFFVG, after reshuffling the data and performing another 80-20 calibration and validation split, we reconducted the model generation and evaluation analysis. The

results of this extra analysis are summarised in Table 5. Values in bold and grey cells in the table denote the same information as explained in subsection 3.2 for Table 3. Overall, according to Table 5, intra-strategy analogies reaffirmed the principal outcomes reported in subsection 3.2. Explicitly, again, models with SFFVG outperformed their counterparts without SFFVG in most scenarios. Also, in all strategies, the model(s) with the highest number of best-obtained evaluation metrics included SFFVG.

Table 5. Results of reevaluation analysis for all investigated scenarios.

Strategy	Preprocessing	SFFVG	RMSE (mg dL^{-1})	MAPD (%)	r^2
NIR Modelling	No	No	101.4	83.3	0.48
		Yes	97.4	53.7	0.52
	Yes	No	103.4	77.6	0.46
		Yes	**95.4**	**52.6**	**0.54**
MIR Modelling	No	No	37.4	31.7	0.93
		Yes	35.6	**21.3**	**0.94**
	Yes	No	37.3	30.9	0.93
		Yes	**35.0**	21.4	**0.94**
Raw Spectra Fusion Modelling	No	No	33.0	25.0	0.94
		Yes	32.8	**17.3**	**0.95**
	Yes	No	33.1	22.1	0.94
		Yes	**31.1**	17.4	**0.95**
Preprocessed Spectra Fusion Modelling	Yes	No	32.8	22.0	0.94
		Yes	**31.0**	**17.3**	**0.95**
Feature Fusion Modelling	No	Yes	**36.7**	36.9	**0.94**
		Yes	37.1	**35.7**	0.93
Decision Fusion Modelling	No	No	57.6	49.3	0.83
		Yes	**54.2**	**30.3**	**0.85**
	Yes	No	101.4	83.3	0.48
		Yes	97.4	53.7	0.52

Note. SFFVG: signal fragmentation based feature vector generation; RMSE: root mean square error; IR: improvement ratio (comparing the results of an SFFVG-included model versus the benchmarked non-SFFVG model reported in the raw above.); MAPD: mean absolute percentage deviation; r2: coefficient of determination; NIR: near-infrared; MIR: mid-infrared. The values in bold font indicate the best result for each evaluation metric in each strategy. The grey cells indicate the model(s) with the highest number of best-obtained evaluation metrics amongst models developed using the same modelling strategy.

4. Summary and conclusion

Feature vector generation based on signal partitioning and framed with model interpretation analysis enhanced in vitro glucose quantification from absorption spectroscopy. First, a given spectrum was sliced into some fragments. A base-regressor then analysed these fragments individually, forming preliminary glucose concentration estimations. These estimations were then stacked, generating a feature vector for the original spectrum. Later, leveraging the concept of stack learning, a meta-regressor investigates this feature vector to produce a final estimation of the reference glucose concentration. The versatility of the proposed

method was tested under an array of modelling strategies. Moreover, the compatibility of the proposed method with a standard preprocessing technique was investigated. Overall, the results obtained accentuated the efficacy of the proposed method in improving glucose quantifications for all modelling strategies. The method maintained its functionality when a preprocessing step was also incorporated in the modelling process. Finally, SHAP was employed to interpret the outcomes of the quantitative analysis. Such interpretation encourages the adoption of the proposed method by extending the transparency of the analysis. For future work, applying the proposed methodology with ununiformed spectra fragmentation is recommended.

Acknowledgement

We thank Dr Osamah Alrezj for his efforts in preparing the experimental data used in this work.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Code availability

We coded in Python (3.6.7); the packages Pandas, NumPy and Sklearn were used for the analysis. The source code of implementations is publicly available in this repository.

References

[1] A.L. Galant, R.C. Kaufman, J.D. Wilson, Glucose: Detection and analysis, Food Chem. 188 (2015) 149–160. https://doi.org/http://dx.doi.org/10.1016/j.foodchem.2015.04.071.

[2] R. Ahmad, M. Khan, N. Tripathy, M.I.R. Khan, A. Khosla, Hydrothermally Synthesized Nickel Oxide Nanosheets for Non-Enzymatic Electrochemical Glucose Detection, J. Electrochem. Soc. 167 (2020) 107504. https://doi.org/10.1149/1945-7111/ab9757.

[3] J. Boudrant, L.P. Fonseca, A.N. Reshetilov, K. Pontius, D. Semenova, Y.E. Silina, K. V Gernaey, H. Junicke, Automated Electrochemical Glucose Biosensor Platform as an Efficient Tool Toward On-Line Fermentation Monitoring: Novel Application Approaches and Insights, Front. Bioeng. Biotechnol. 8 (2020) 1–15. https://doi.org/10.3389/fbioe.2020.00436.

[4] I. Delfino, C. Camerlingo, M. Portaccio, B. Della Ventura, L. Mita, D.G. Mita, M. Lepore, Visible micro-Raman

spectroscopy for determining glucose content in beverage industry, Food Chem. 127 (2011) 735–742. https://doi.org/http://dx.doi.org/10.1016/j.foodchem.2011.01.007.

[5] M. Shokrekhodaei, Non-Invasive In-Vitro Glucose Monitoring Using Optical Sensor and Machine Learning Techniques for Diabetes Applications, The University of Texas at El Paso, 2021.

[6] A. Al-Mbaideen, M. Benaissa, Coupling subband decomposition and independent component regression for quantitative NIR spectroscopy, Chemom. Intell. Lab. Syst. 108 (2011) 112–122. https://doi.org/10.1016/j.chemolab.2011.05.012.

[7] J. Haas, B. Mizaikoff, Advances in Mid-Infrared Spectroscopy for Chemical Analysis, Annu. Rev. Anal. Chem. 9 (2016) 45–68. https://doi.org/10.1146/annurev-anchem-071015-041507.

[8] S.K. Vashist, Non-invasive glucose monitoring technology in diabetes management: A review, Anal. Chim. Acta. 750 (2012) 16–27. https://doi.org/10.1016/j.aca.2012.03.043.

[9] D.A. Burns, E.W. Ciurczak, Handbook of near-infrared analysis, CRC press, 2007.

[10] S. Delbeck, H.M. Heise, Evaluation of opportunities and limitations of mid-infrared skin spectroscopy for noninvasive blood glucose monitoring, J. Diabetes Sci. Technol. 15 (2021) 19–27. https://doi.org/https://doi.org/10.1177/1932296820936224.

[11] B. Rabinovitch, W.F. March, R.L. Adams, Noninvasive glucose monitoring of the aqueous humor of the eye: Part I. Measurement of very small optical rotations, Diabetes Care. 5 (1982) 254–258. https://doi.org/https://doi.org/10.2337/diacare.5.3.254.

[12] J. Yadav, A. Rani, V. Singh, B.M. Murari, Prospects and limitations of non-invasive blood glucose monitoring using near-infrared spectroscopy, Biomed. Signal Process. Control. 18 (2015) 214–227. https://doi.org/10.1016/j.bspc.2015.01.005.

[13] C.-F. So, K.-S. Choi, T.K.S. Wong, J.W.Y. Chung, Recent advances in noninvasive glucose monitoring, Med. Devices Evid. Res. 5 (2012) 45–52. https://doi.org/https://dx.doi.org/10.2147%2FMDER.S28134.

[14] H. von Lilienfeld-Toal, M. Weidenmüller, A. Xhelaj, W. Mäntele, A novel approach to non-invasive glucose measurement by mid-infrared spectroscopy: The combination of quantum cascade lasers (QCL) and photoacoustic detection, Vib. Spectrosc. 38 (2005) 209–215. https://doi.org/https://doi.org/10.1016/j.vibspec.2005.02.025.

[15] B.K. Mekonnen, W. Yang, T.-H. Hsieh, S.-K. Liaw, F.-L. Yang, Accurate prediction of glucose concentration and identification of major contributing features from hardly distinguishable near-infrared spectroscopy, Biomed. Signal Process. Control. 59 (2020) 1–15. https://doi.org/https://doi.org/10.1016/j.bspc.2020.101923.

[16] A. Tura, A. Maran, G. Pacini, Non-invasive glucose monitoring: Assessment of technologies and devices according to quantitative criteria, Diabetes Res. Clin. Pract. 77 (2007) 16–40. https://doi.org/10.1016/j.diabres.2006.10.027.

[17] G. Han, S. Chen, X. Wang, J. Wang, H. Wang, Z. Zhao, Noninvasive blood glucose sensing by near-infrared spectroscopy based on PLSR combines SAE deep neural network approach, Infrared Phys. Technol. 113 (2021) 1–10. https://doi.org/https://doi.org/10.1016/j.infrared.2020.103620.

[18] J. Tenhunen, H. Kopola, R. Myllylä, Non-invasive glucose measurement based on selective near infrared absorption; requirements on instrumentation and spectral range, Meas. J. Int. Meas. Confed. 24 (1998) 173–177. https://doi.org/10.1016/S0263-2241(98)00054-2.

[19] Å. Rinnan, F. van den Berg, S.B. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, TrAC - Trends Anal. Chem. 28 (2009) 1201–1222. https://doi.org/10.1016/j.trac.2009.07.007.

[20] H. Khadem, M.R. Eissa, H. Nemat, O. Alrezj, M. Benaissa, Classification before regression for improving the accuracy of glucose quantification using absorption spectroscopy, Talanta. 211 (2020) 1–10. https://doi.org/https://doi.org/10.1016/j.talanta.2020.120740.

[21] Z.-H. Zhou, Ensemble methods: foundations and algorithms, Chapman and Hall/CRC, 2019.

[22] E. Mauer, J. Lee, J. Choi, H. Zhang, K.L. Hoffman, I.J. Easthausen, M. Rajan, M.G. Weiner, R. Kaushal, M.M. Safford, others, A predictive model of clinical deterioration among hospitalized COVID-19 patients by harnessing hospital course trajectories, J. Biomed. Inform. 118 (2021) 1–12. https://doi.org/10.1016/j.jbi.2021.103794.

[23] S. Bhatt, A. Cohon, J. Rose, N. Majerczyk, B. Cozzi, D. Crenshaw, G. Myers, Interpretable machine learning models for clinical decision-making in a high-need, value-based primary care setting, NEJM Catal. Innov. Care Deliv. 2 (2021). https://doi.org/https://doi.org/10.1056/CAT.21.0008.

[24] S. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: 31th Conf. Neural Inf. Process. Syst., 2017: pp. 4765–4774.

[25] L.S. Shapley, A value for n-person games, Contrib. to Theory Games. 2 (1953) 307–317.

[26] O. Alrezj, M. Benaissa, S.A. Alshebeili, Digital bandstop filtering in the quantitative analysis of glucose from near-infrared and midinfrared spectra, J. Chemom. 34 (2020) e3206. https://doi.org/https://doi.org/10.1002/cem.3206.

[27] Ian T. Jolliffe, A Note on the Use of Principal Components in Regression, J. R. Stat. Soc. 31 (1982) 300–303.

[28] G.M. Escandar, P.C. Damiani, H.C. Goicoechea, A.C. Olivieri, A review of multivariate calibration methods applied to biomedical analysis, Microchem. J. 82 (2006) 29–42. https://doi.org/10.1016/j.microc.2005.07.001.

[29] Y. Li, Y. Xiong, S. Min, Data fusion strategy in quantitative analysis of spectroscopy relevant to olive oil adulteration, Vib. Spectrosc. 101 (2019) 20–27. https://doi.org/10.1016/j.vibspec.2018.12.009.

[30] Y. Li, J.Y. Zhang, Y.Z. Wang, FT-MIR and NIR spectral data fusion: a synergetic strategy for the geographical traceability of Panax notoginseng, Anal. Bioanal. Chem. 410 (2018) 91–103. https://doi.org/10.1007/s00216-017-0692-0.

[31] W. Sun, X. Zhang, Z. Zhang, R. Zhu, Data fusion of near-infrared and mid-infrared spectra for identification of rhubarb, Spectrochim. Acta - Part A Mol. Biomol. Spectrosc. 171 (2017) 72–79. https://doi.org/10.1016/j.saa.2016.07.039.
[32] L. Tao, B. Via, Y. Wu, W. Xiao, X. Liu, NIR and MIR spectral data fusion for rapid detection of Lonicera japonica and Artemisia annua by liquid extraction process, Vib. Spectrosc. 102 (2019) 31–38. https://doi.org/10.1016/j.vibspec.2019.03.005.
[33] Y. Wang, M. Yang, G. Wei, R. Hu, Z. Luo, G. Li, Improved PLS regression based on SVM classification for rapid analysis of coal properties by near-infrared reflectance spectroscopy, Sensors Actuators, B Chem. 193 (2014) 723–729. https://doi.org/10.1016/j.snb.2013.12.028.
[34] A. Savitzky, M.J.E. Golay, Smoothing and Differentiation of Data by Simplified Least Squares Procedures, Anal. Chem. 36 (1964) 1627–1639. https://doi.org/10.1021/ac60214a047.
[35] M.L.F. Simeone, R.A.C. Parrella, R.E. Schaffert, C.M.B. Damasceno, M.C.B. Leal, C. Pasquini, Near infrared spectroscopy determination of sucrose, glucose and fructose in sweet sorghum juice, Microchem. J. 134 (2017) 125–130. https://doi.org/10.1016/j.microc.2017.05.020.
[36] A. De Myttenaere, B. Golden, B. Le Grand, F. Rossi, Mean absolute percentage error for regression models, Neurocomputing. 192 (2016) 38–48. https://doi.org/https://doi.org/10.1016/j.neucom.2015.12.114.
[37] and W.A.N. Lee Rodgers, Joseph, Thirteen ways to look at the correlation coefficient., Am. Stat. 42 (1988) 59–66. https://doi.org/https://doi.org/10.1080/00031305.1988.10475524.
[38] P. Mlynarski, H. Delingette, A. Criminisi, N. Ayache, 3D convolutional neural networks for tumor segmentation using long-range 2D context, Comput. Med. Imaging Graph. 73 (2019) 60–72. https://doi.org/https://doi.org/10.1016/j.compmedimag.2019.02.001.
[39] H. Hui, X. Zhang, F. Li, X. Mei, Y. Guo, A Partitioning-Stacking Prediction Fusion Network Based on an Improved Attention U-Net for Stroke Lesion Segmentation, IEEE Access. 8 (2020) 47419–47432. https://doi.org/10.1109/ACCESS.2020.2977946.
[40] M. Golic, K. Walsh, P. Lawson, Short-wavelength near-infrared spectra of sucrose, glucose, and fructose with respect to sugar concentration and temperature, Appl. Spectrosc. 57 (2003) 139–145. https://doi.org/10.1366/000370203321535033.