Raman Biosensing of Sweat Metabolites: Univariate vs. Multivariate Algorithms

Raman Biosensing of Sweat Metabolites: Univariate vs. Multivariate Algorithms

Leonardo Iannucci
Department of Applied Science and Technology
Politecnico di Torino
Turin, Italy
leonardo.iannucci@polito.it

Ata Golparvar
Bio/CMOS Interfaces Laboratory
EPFL
Neuchâtel, Switzerland
ata.golparvar@epfl.ch

Francesco Giraudo
Department of Applied Science and Technology
Politecnico di Torino
Turin, Italy
francesco.giraudo@studenti.polito.it

Sandro Carrara
Bio/CMOS Interfaces Laboratory
EPFL
Neuchâtel, Switzerland
sandro.carrara@epfl.ch

Sabrina Grassini
Department of Applied Science and Technology
Politecnico di Torino
Turin, Italy
sabrina.grassini@polito.it

***Abstract*—Over the past eight years, there has been a remarkable surge in the field of sweat analysis, as a non-invasive, continuous, and multi-metabolite monitoring solution tailored for wearable devices. However, its full potential has yet to be fully realized due to the limitations of existing biosensing transducers. Despite years of research, wearable devices still fall short of providing biochemical insights into human functions, largely due to the longevity issues associated with colorimetric and electrochemical biosensing methods stemming from their biorecognition elements. However, optical methods such as Raman scattering measurements offer an alternative, inherently selective biosensing mechanism without the longevity issues seen in other methods. While the main hurdle in the past was the bulky instrumentation required, advancements in microengineering and laser technology have paved the way for the development of compact Raman systems. Nevertheless, research at the intersection of Raman systems and sweat analysis (or other alternative biofluids to blood) is still in its infancy, with no comparative studies to assess the efficiency of multivariate versus univariate data analysis techniques in biosensing. To address this, the present work analyzes two of these widely used data processing methods in multiplexed human sweat glucose, urea, and lactate biosensing. Experimental findings suggest that multivariate analysis, particularly Principal Components Regression (PCR), demonstrates better performance especially in datasets containing interferents, outperforming univariate analysis. This paper also delves into the potential advantages and limitations associated with the two investigated algorithms, shedding light on their applicability in sweat analysis for future wearables Raman systems.**

***Index Terms*—Sweat analysis, Raman spectroscopy, Multivariate analysis, Principal components regression.**

I. Introduction

Current wearable devices excel in assessing continuous physical biomarkers monitoring but fall short in chemical biomarkers monitoring [1]. This limitation stems from the reliance of current biosensing transducers on external interface biorecognition elements, which possess a finite lifespan and degrade over time or with prolonged use. Consequently, cutting-edge biosensing devices, such as microneedle-based continuous glucose monitoring devices, typically operate for up to 2 weeks before necessitating replacement [2]. In contrast, optical transducing mechanisms offer extended operational lifetimes. For instance, Eversense's continuous glucose monitoring implant, leveraging fluorescence detection, boasts an impressive 3-month lifespan [3]. Label-free methods like vibrational spectroscopy techniques, such as absorption spectroscopy, for continuous metabolite monitoring also exist [4].

Historically, the primary limitation of optical biosensing was the bulky size of instrumentation. However, with advancements in microengineering and laser technology, compact point-of-care Raman biosensing systems have already been developed [5]. Furthermore, recent breakthroughs in wavelength-specific Raman biosensing and Raman-on-chip systems are expected to further drive the development of even smaller Raman systems, facilitating their seamless integration into wearable technologies [6], [7].

Similarly, the field of sweat analysis has seen a rise in interest since the original paper from the Javey's Lab [8]. Sweat provides an attractive medium for continuous multi-metabolite monitoring, especially with the routine use of on-demand sweat extraction systems such as iontophoresis and epidermal microfluidics. While the combination of sweat and Raman biosensing is a less explored avenue, it is gaining momentum [9]–[13].

In literature, two main approaches can be identified to process spectroscopic data: multivariate (or univariate) analysis [14], [15] and machine learning techniques [16]. The former approach is generally preferred when the available dataset

is not very large. Actually, a lower number of spectra is required to create a reliable model and the relatively simple mathematical operations generally preserve a clear correlation between the regression analysis and spectrum interpretation from chemical point of view. On the other hand, machine learning algorithms may reach interesting performance in terms of concentration prediction, but require large datasets to train and optimize the model [17]. Moreover, as they work as a 'black box', it is not possible to derive a rationale explaining their functioning. Despite many studies have analyzed the performance of different machine learning algorithms [16], [18], there is currently no comparative research which clearly addresses univariate and multivariate methods.

For this reason, this paper specifically compares the performance of univariate and multivariate analysis, to estimate the concentration of different analytes from Raman spectra. This study specifically focuses on the three main components of human sweat, namely glucose, lactate, and urea, predicting their concentration in different matrices deionized water and artificial sweat. Possible advantages and limitations are discussed for the two investigated algorithms.

II. Materials and Methods

Raman measurements were carried out with a backscattered confocal micro-Raman microscope (LabRAM HR, HORIBA, Japan) in the spectral region from 300 cm^{-1} to 1500 cm^{-1}. The source was a 532-nm green laser, set to 200 mW of power through the built-in neutral density filters. The filtered beam was focused using a ×50 objective lens, and the confocal hole size was adjusted to 400 μm. Calibration of the spectrometer was carried out before the measurement sessions using the characteristic peak of silicon at 520 cm^{-1}. The acquisition time for each scan was 120 s.

Analytical grade powder reagents, D-(+)-glucose ($C_6H_{12}O_6$, $\geq$ 99.5%), sodium L-lactate ($C_3H_5NaO_3$, 98%), urea (CH_4N_2O, 98%), were purchased from Sigma-Aldrich (MilliporeSigma, USA) and used as received without further purification. During solutions preparation, the powders were weighed with an analytical balance, dissolved in ultra-pure deionized (DI) water, then stirred with an orbital shaker, and refrigerated overnight to reach equilibrium. The following solutions were prepared, analyzing 3 samples per each concentration: glucose in (DI) water (from 0 mM to 100 mM), lactate in DI water (from 0 mM to 100 mM), urea in DI water (from 0 mM to 100 mM). Then, the following solutions were analysed, to mimic human sweat and assess the effect of possible interferents:

- 4 mM of lactate, 4 mM of urea and glucose concentration ranging from 0 mM to 10 mM;
- 8 mM of glucose, 4 mM of urea and lactate concentration ranging from 0 mM to 40 mM;
- 8 mM of glucose, 4 mM of lactate and urea ranging from 0 mM to 40 mM.

For multivariate analysis, the principal components were computed using a Principal component analysis (PCA) script written in Python, as described in [19], [20]. After importing the spectra, the following pre-processing steps were performed:

- Interval selection: depending on the analyzed compound, a specific Raman shift range was chosen. This was from 950 cm^{-1} to 1450 cm^{-1} for glucose, from 750 cm^{-1} to 1150 cm^{-1} for lactate, and from 880 cm^{-1} to 1450 cm^{-1} for urea;
- Baseline removal: symmetric least square smoothing, as described in [21];
- Smoothing: with the Savitzky-Golay filter [22], using a second-order polynomial and window length of 15 cm^{-1}.

Spectra normalization was not performed, in order to preserve the correlation between the peaks height and the analytes concentration. Both the computation of the principal components and the PCR were performed using the Scikit-Learn library [23].

The univariate analysis was carried out by performing a linear regression of the intensity of the main peak present in the spectrum for each of the investigated analytes. The selected peak was: at (1125 ± 10) cm^{-1} for glucose, at (861 ± 10) cm^{-1} for lactate, and at (1005 ± 10) cm^{-1} for urea.

III. Results and Discussion

Raman spectroscopy stands out as a distinguished analytical tool, seamlessly offering a blend of qualitative richness and quantitative precision in the exploration of diverse samples. By examining the position and shape of distinct peaks, the compounds under investigation can be identified, while the intensity of these peaks can be utilized to establish a quantitative model for assessing the concentration of the analyte in a liquid sample.

Indeed, a Raman spectrum typically comprises two signals: a baseline attributed to fluorescence (elastic scattering) and peaks corresponding to specific molecular vibrations (inelastic scattering). Therefore, prior to developing a quantitative model for analyte concentration, it is imperative to conduct preprocessing to eliminate the contribution of fluorescence, as it is unrelated to the concentration of the compound under investigation. In this study, symmetric least square smoothing was employed to fit the baseline, which was subsequently subtracted from the raw spectrum. Furthermore, to enhance the signal-to-noise ratio (SNR), the Savitzky-Golay filter was applied, and a specific Raman shift range was selected. The result of these preprocessing steps (interval selection, baseline removal, and smoothing) is illustrated in Figure 1, showcasing the spectrum obtained from a 50 mM solution of lactate in DI water. As observed, the processed spectrum exhibits reduced absolute intensity values; however, the peaks are more discernible and no longer overshadowed by baseline intensity.

The datasets acquired on the solutions of glucose, lactate, and urea in DI water are reported in Figure 2, superimposing the spectra related to concentrations from 1 mM to 100 mM. As can be seen, for each analyte some specific peaks are present, associated to the different molecular vibrations for each compound. The spectra acquired on glucose solutions (Figure 2a) are characterized by the main peak at 1125 cm^{-1}

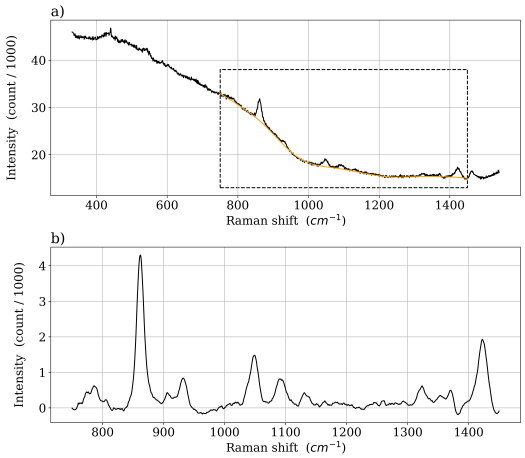

Fig. 1. (a) Raman spectrum of a 50 mM aqueous lactate solution. The dotted box delimits the spectrum interval selected for the subsequent processing, while the yellow line represents the computed spectrum baseline. (b) The same spectrum after data pre-processing.

related to the bending motion of COH bonds and then a secondary peak at 1060 cm^{-1} due to the CO stretching [24]. From 800 cm^{-1} to 1000 cm^{-1} the anomeric region is present, where vibrations related to the anomeric carbon are observed. The part of the spectrum above 1300 cm^{-1} shows the peaks associated with the scissoring and bending modes of CH_2 groups [24]. In the spectra acquired on lactate solutions (Figure 2b), the peak at 861 cm^{-1} is the most intense and it is attributed to CH_3 group torsion. Additional peaks are found at 1020 cm^{-1}, 1080 cm^{-1}, and 1120 cm^{-1}, related to CO stretching in the fingerprint region [24]. Finally, in the spectra acquired on urea solutions (Figure 2c), the peak at 1005 cm^{-1} is linked to the $N-C-N$ stretching vibration and the one at 1160 cm^{-1} to the rocking motion of NH_2 bonds [25]. As can be seen even from a qualitative point of view, each of the three compounds is characterized by some peculiar peaks (generally the most intense ones) and then additional minor peaks that can be in common with other molecules, because they are constituted by the same chemical bonding.

When employing PCR to conduct a linear regression of the spectra, the first step involves computing the principal components, commonly referred to as 'loadings'. These loadings take the form of vectors with dimensions equal to the number of points in each spectrum and represent the directions of maximum variance for the input datasets. For each analyte, a model is constructed using spectra acquired at increasing concentrations (Figure 3). As expected, given that variations in the spectra correspond to changes in analyte concentration, the loadings identify both major and minor peaks in the spectrum. Consequently, regression is performed subsequent to computing the eigenvalues, known as 'scores', obtained from the following mathematical operation:

$$S = R \cdot L$$

where S is the scores vector, R signifies the pre-processed

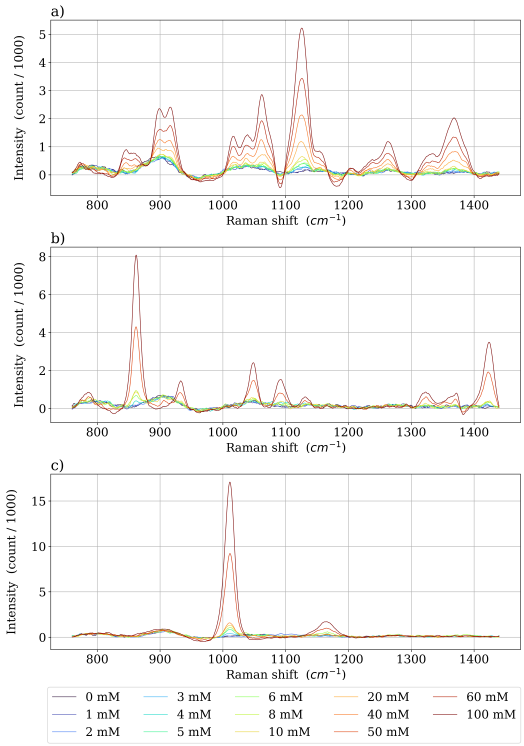

Fig. 2. Processed Raman spectra of aqueous solutions of (a) glucose, (b) lactate, and (c) urea. The concentration of each analyte in the solution ranges from 0 mM to 100 mM.

Raman spectrum, and L is the matrix containing the loadings. The computed principal components can effectively capture a substantial percentage of the total variance, even with a single component. Specifically, for glucose, the first principal component (i.e., PC1) accounts for 99.6%, (with PC2 accounting for 0.1%); for lactate, PC1 represents 98.9% (PC2 0.3%); for urea, PC1 encompasses 99.7% (PC2 covering 0.2%). This indicates that a model can be constructed for each analyte using just the first loading PC1, as it sufficiently captures the experimental data alone.

The PCR model is initially trained by regressing the score values computed from the known dataset. Subsequently, the model can be utilized to predict the concentration of the analyte in an unknown sample. In this process, the algorithm leverages the information encapsulated in the loadings, thereby considering all peaks present in PC1. The outcomes are depicted in Figure 4 for the three analytes examined. For all compounds, the interpolating curve exhibits a slope close to 1 (above 0.99 in all three cases) and an intercept with the y-axis approaching 0 (below 0.05 mM). The high linearity of the fitting is further underscored by R^2 values (min 0.9976) and a low mean squared error (MSE) with a maximum of 1.45 mM.

When analyzing a solution containing a single analyte in DI water, comparable performance is achieved by univariate algorithms. Actually, performing a regression with the intensity of the highest peak for each of the compounds, R^2 values

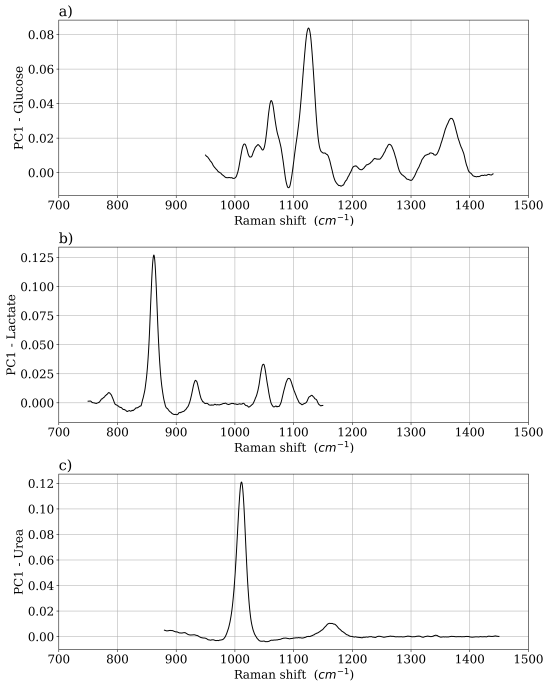

Fig. 3. The first principal component (PC1) for each of the three analytes: (a) Glucose, (b) Lactate, and (c) Urea.

above 0.99 are obtained in all three cases. This result confirms that, in the absence of interferents, univariate analysis serves as a straightforward yet effective approach, significantly streamlining the computational complexity required for predicting analyte concentrations in the sample [26].

When the multivariate algorithm is applied to a dataset where both the analyte of interest and other interferents are present (i.e., artificial sweat solutions), the performance is slightly compromised. Indeed, the presence of additional peaks, not attributable to the main compound, may lead to an overestimation of the predicted concentration. This is because if the molecules share the same chemical bonds or similar ones, their peaks will be positioned at the same Raman shift values. Figure 5 illustrates the results from PCR conducted on artificial sweat solutions. The concentration of the analyte of interest in artificial sweat extended up to 10 mM for glucose (given its concentration in sweat typically falls way below this threshold), whereas it reached up to 40 mM for lactate and urea. Despite the presence of interferents, linearity remains robust across all three datasets (i.e., R^2 above 0.98) and the intercept of the interpolating line with the y-axis is close to zero.

Using univariate analysis, a similar performance was achieved by the models for lactate and urea in artificial sweat solutions. In these cases, the obtained R^2 values were equal to 0.9809 and 0.9876, respectively. On the other hand, R^2 was equal to 0.9322 for the dataset of glucose with interferents. The reason should be ascribed to the low concentration of glucose in these solutions (i.e. up to 10 mM): in this condition, the performance of univariate analysis is affected by the presence

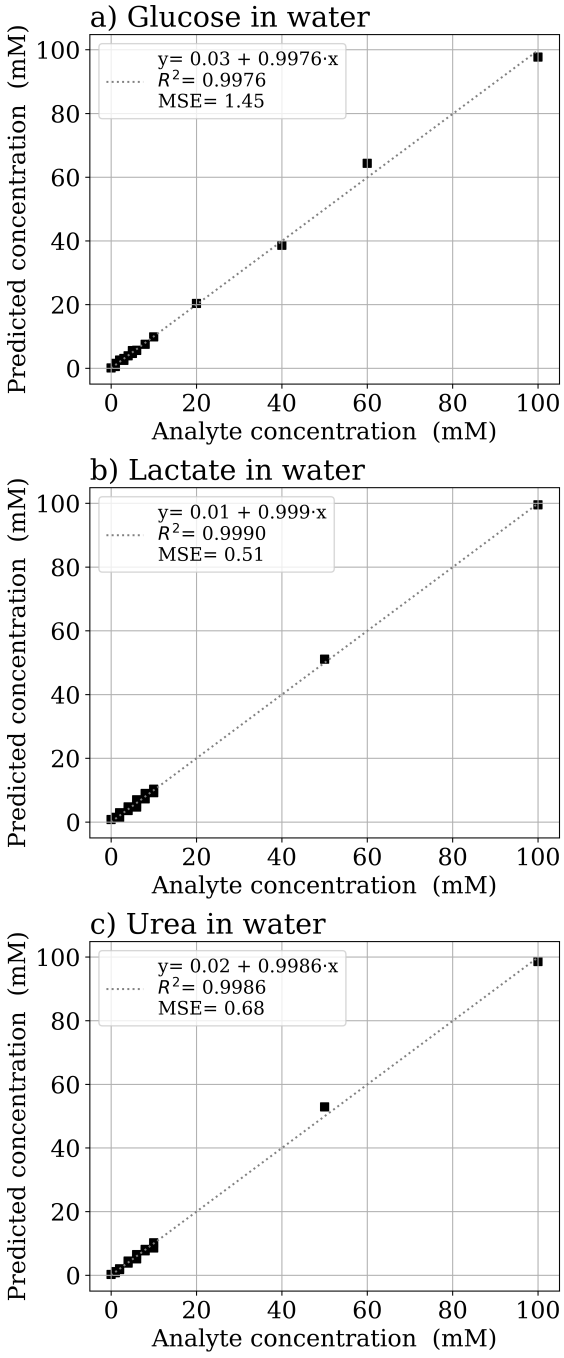

Fig. 4. Results of PCR for the three analytes in DI water. The dotted line represents the linear regression, whose equation is reported in the inset.

of other compounds such as lactate and urea. On the other hand, as multivariate analysis exploits a larger portion of the Raman spectrum to perform the prediction, it is less biased by interferents.

Indeed, considering only the low-concentration samples (i.e. up to 10 mM) also for lactate and urea, and performing a regression only in this reduced range, a similar behaviour was found also for lactate and urea. With multivariate analysis, R^2 is equal to 0.9565 for lactate and to 0.9742 for urea, but it falls to 0.7519 and 0.9348 for the same compounds when univariate analysis is used.

Thus, while the performance is similar for the two algorithms if a large concentration range is considered, PCR outperforms univariate analysis when the low-concentration range is investigated. Actually, using the principal components of the spectrum instead of a single peak allows the regressing algorithm to identify more precisely the variations related to the analyte of interest, without a relevant bias from the other interferents present in the analyzed solution.

IV. Conclusion

The paper presented a comparison of the performance of univariate and multivariate algorithms to estimate the concentration of glucose, lactate and urea in different matrices. As these three compounds are the main constituents of sweat, they give the most intense signal in the Raman spectrum. Moreover, as they have similar chemical bonds, they are expected to interfere with each other in the analysis.

Results showed that, in simple aqueous solutions containing the analyte of interest dissolved in DI water, the two types of algorithms have similar performance and both are able to effectively predict the analyte concentration. When solutions mimicking artificial sweat are taken into consideration, good linearity is obtained again by both algorithms if a large concentration range is considered. When focusing only on the low-concentration range, PCR is able to outperform univariate analysis, as it is less affected by the presence of interferents.

Future work will test the performance of multivariate algorithms in even more complex matrices, increasing the concentration of the interferents to further check the robustness of PCR in sweat analysis. Moreover, different laser source wavelengths will be used, to try to reduce the effect of fluorescence in the acquired spectra.

References

[1] R. Ghaffari, J. A. Rogers, and T. R. Ray, "Recent progress, challenges, and opportunities for wearable biochemical sensors for sweat analysis," *Sensors and Actuators, B: Chemical*, vol. 332, 2021.

[2] O. Didyuk, N. Econom, A. Guardia, K. Livingston, and U. Klueh, "Continuous glucose monitoring devices: Past, present, and future focus on the history and evolution of technological innovation," *Journal of Diabetes Science and Technology*, vol. 15, no. 3, p. 676 – 683, 2020.

[3] J. Kropff, P. Choudhary, S. Neupane, K. Barnard, S. C. Bain, C. Kapitza, T. Forst, M. Link, A. Dehennis, and J. H. De Vries, "Accuracy and longevity of an implantable continuous glucose sensor in the precise study: A 180-day, prospective, multicenter, pivotal trial," *Diabetes Care*, vol. 40, no. 1, p. 63 – 68, 2017.

[4] https://indigomed.com/, Indigo -Revolutionizing the management of chronic disease, last checked on 13th of March 2024.

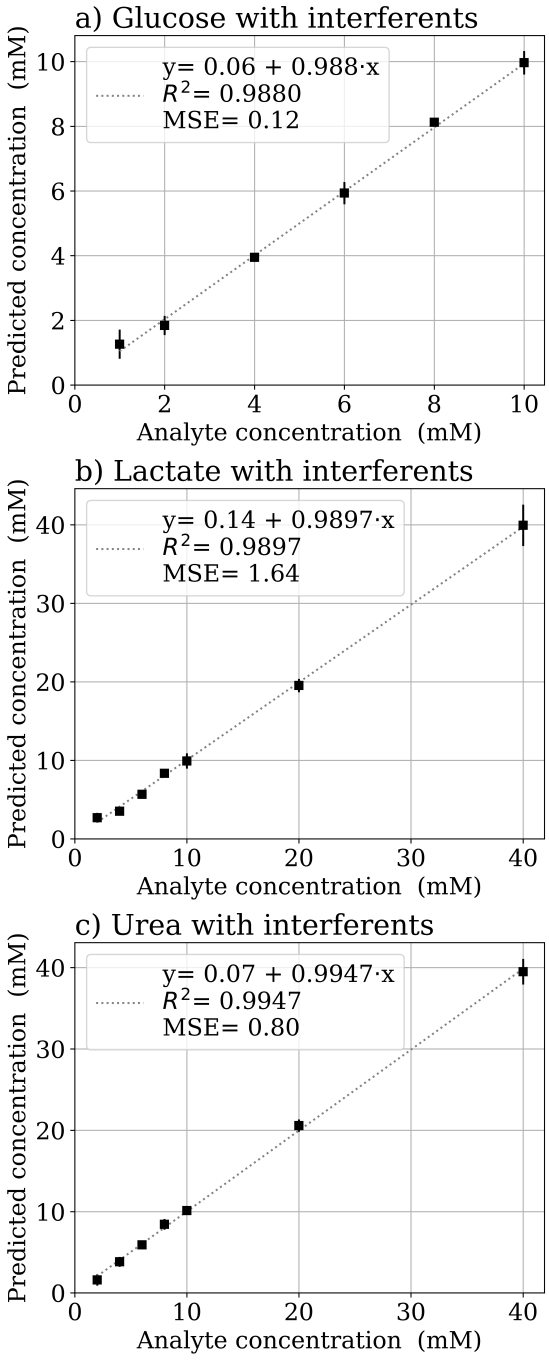

Fig. 5. Results of PCR for the three analytes in artificial sweat solution. The dotted line represents the linear regression, whose equation is reported in the inset.

[5] A. Pors, K. G. Rasmussen, R. Inglev, N. Jendrike, A. Philipps, A. G. Ranjan, V. Vestergaard, J. E. Henriksen, K. Nørgaard, G. Freckmann, K. D. Hepp, M. C. Gerstenberg, and A. Weber, "Accurate post-calibration predictions for noninvasive glucose measurements in people using confocal raman spectroscopy," *ACS Sensors*, vol. 8, no. 3, p. 1272 – 1279, 2023.

[6] Y. Park, U. J. Kim, S. Lee, H. Kim, J. Kim, H. Ma, H. Son, Y. Z. Yoon, J.-s. Lee, M. Park, H. Choo, Q.-H. Park, and Y.-G. Roh, "On-chip raman spectrometers using narrow band filter array combined with cmos image sensors," *Sensors and Actuators B: Chemical*, vol. 381, 2023.

[7] A. Golparvar, J. Kim, A. Boukhayma, D. Briand, and S. Carrara, "Highly accurate multimodal monitoring of lactate and urea in sweat by soft epidermal optofluidics with single-band raman scattering," *Sensors and Actuators B: Chemical*, vol. 387, 2023.

[8] W. Gao, S. Emaminejad, H. Y. Y. Nyein, S. Challa, K. Chen, A. Peck, H. M. Fahad, H. Ota, H. Shiraki, D. Kiriya, D.-H. Lien, G. A. Brooks, R. W. Davis, and A. Javey, "Fully integrated wearable sensor arrays for multiplexed in situ perspiration analysis," *Nature*, vol. 529, no. 7587, p. 509 – 514, 2016.

[9] A. Golparvar, L. Thenot, A. Boukhayma, and S. Carrara, "Soft epidermal paperfluidics for sweat analysis by ratiometric raman spectroscopy," *Biosensors*, vol. 14, no. 1, 2024.

[10] U. Mogera, H. Guo, M. Namkoong, M. S. Rahman, T. Nguyen, and L. Tian, "Wearable plasmonic paper-based microfluidics for continuous sweat analysis," *Science Advances*, vol. 8, no. 12, 2022.

[11] X. He, C. Fan, Y. Luo, T. Xu, and X. Zhang, "Flexible microfluidic nanoplasmonic sensors for refreshable and portable recognition of sweat biochemical fingerprint," *npj Flexible Electronics*, vol. 6, no. 1, 2022.

[12] E. H. Koh, W.-C. Lee, Y.-J. Choi, J.-I. Moon, J. Jang, S.-G. Park, J. Choo, D.-H. Kim, and H. S. Jung, "A wearable surface-enhanced raman scattering sensor for label-free molecular detection," *ACS Applied Materials and Interfaces*, vol. 13, no. 2, p. 3024 – 3032, 2021.

[13] Y. Wang, C. Zhao, J. Wang, X. Luo, L. Xie, S. Zhan, J. Kim, X. Wang, X. Liu, and Y. Ying, "Wearable plasmonic-metasurface sensor for non-invasive and universal molecular fingerprint detection on biointerfaces," *Science Advances*, vol. 7, no. 4, 2021.

[14] Y. Han, X. Fang, H. Li, L. Zha, J. Guo, and X. Zhang, "Sweat sensor based on wearable janus textiles for sweat collection and microstructured optical fiber for surface-enhanced raman scattering analysis," *ACS Sensors*, vol. 8, no. 12, p. 4774 – 4781, 2023.

[15] W. Sun, S. Song, B. Qian, D. Wen, D. Jiang, Y. Fu, and Q. Wang, "Quantitative analysis of blood glucose by ft-raman spectroscopy and multivariate statistical analysis," *Microwave and Optical Technology Letters*, vol. 66, no. 1, 2024.

[16] N. González-Viveros, P. Gómez-Gil, J. Castro-Ramos, and H. Cerecedo-Núñez, "On the estimation of sugars concentrations using raman spectroscopy and artificial neural networks," *Food Chemistry*, vol. 352, 2021.

[17] S. Song, Q. Wang, X. Zou, Z. Li, Z. Ma, D. Jiang, Y. Fu, and Q. Liu, "High-precision prediction of blood glucose concentration utilizing fourier transform raman spectroscopy and an ensemble machine learning algorithm," *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, vol. 303, 2023.

[18] F. Chen, C. Chen, C. Chen, Z. Yan, R. Gao, H. Han, W. Li, and X. Lv, "Application of plsr in rapid detection of glucose in sheep serum," *Optik*, vol. 224, 2020.

[19] L. Iannucci, "Chemometrics for data interpretation: Application of principal components analysis (pca) to multivariate spectroscopic measurements," *IEEE Instrumentation and Measurement Magazine*, vol. 24, no. 4, p. 42 – 48, 2021.

[20] L. E. Sebar, L. Iannucci, Y. Goren, P. Fabian, E. Angelini, and S. Grassini, "Raman investigation of corrosion products on roman copper-based artefacts," *Acta IMEKO*, vol. 10, no. 1, p. 129 – 135, 2020.

[21] P. H. C. Eilers, "A perfect smoother," *Analytical Chemistry*, vol. 75, no. 14, p. 3631 – 3636, 2003.

[22] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Analytical Chemistry*, vol. 36, no. 8, p. 1627 – 1639, 1964.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, p. 2825 – 2830, 2011.

[24] M. Dudek, G. Zajac, E. Szafraniec, E. Wiercigroch, S. Tott, K. Malek, A. Kaczor, and M. Baranska, "Raman optical activity and raman spectroscopy of carbohydrates in solution," *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, vol. 206, p. 597 – 612, 2019.

[25] K. M. Khan, H. Krishna, S. K. Majumder, and P. K. Gupta, "Detection of urea adulteration in milk using near-infrared raman spectroscopy," *Food Analytical Methods*, vol. 8, no. 1, p. 93 – 102, 2015.

[26] A. Golparvar, A. Boukhayma, T. Loayza, A. Caizzone, C. Enz, and S. Carrara, "Very selective detection of low physiopathological glucose levels by spontaneous raman spectroscopy with univariate data analysis," *BioNanoScience*, vol. 11, no. 3, p. 871 – 877, 2021.