Towards a Global Model for Diabetic Kidney Disease Screening using ATR-FTIR

Víctor Navarro-Esteve[1], Ángel Sánchez-Illana[1], Jose Pórtoles[2,3], Maria Marques-Vidas[2], Josep Ventura-Gayete[4], Nuria Estañ Capell[4], Iris Viejo-Boyano[5], Francisco Valero-Mena[5], Antonio J. Sanchez-Lopez[6,7], Bayden R. Wood*[8], David Pérez-Guaita**[1]

[1] Department of Analytical Chemistry, University of Valencia, Burjassot, Spain
[2] University Hospital Puerta de Hierro de Majadahonda, Majadahonda, Spain. RISCORS2040 RD24/0004/0028
[3] Department of Medicine. Universidad Autónoma de Madrid, Spain
[4] University Hospital Doctor Peset Aleixandre, Valencia, Spain
[5] University Hospital La Fe, Valencia, Spain
[6] Neuroimmunology Unit, Instituto de Investigación Sanitaria Puerta de Hierro-Segovia de Arana, Madrid, Spain.
[7] Biobank, Instituto de Investigación Sanitaria Puerta de Hierro-Segovia de Arana, Madrid, Spain.
[8] Centre for Biospectroscopy, Monash University, Clayton, Victoria, Australia
*bayden.wood@monash.edu, **david.perez-guaita@uv.es

Abstract

Infrared (IR) spectroscopy of urine extracts coupled with Machine Learning (ML) methods has been proposed as a promising technique for Diabetic Kidney Disease (DKD) screening. However, for clinical translation, predictive models based on machine learning techniques necessitate substantial sets of samples for calibration and testing under various experimental parameters and populations, posing a critical barrier to the development of globally deployable tools. Here, we aim to assess the methodology's ability to establish DKD diagnostic models applicable across diverse populations, instruments and experimental conditions worldwide. Two datasets were compared. The Australian set included 155 DKD and 22 control samples from a 24-hour urine collection, with preconcentrated proteins measured on a Bruker spectrometer. The Spanish set, comprising 35 DKD and 26 control spot urine samples, was analyzed using a Perkin-Elmer spectrometer. Different ML methods were developed to identify DKD and microalbuminuria, aiming to compare their performance in terms of generalization and adaptation to different datasets. Models developed using Australian spectra successfully predicted Spanish samples, achieving AUROC values of 0.87 and 0.98 for DKD and microalbuminuria identification, respectively. Both values improved to 0.99 when a global model was calibrated and independently tested with a combined set integrating samples from both countries. Results evidence that the spectral markers found in the IR spectra, based on signals arising from albumin and other glycoproteins, have proven to be robust, minimizing the effects of population and instrument variability. Results exemplify the potential of developing global big-data spectroscopic datasets to facilitate the deployment of IR-based diagnostic methods in real-world settings.

Keywords: Infrared; Diabetic kidney disease (DKD); Machine Learning; Model Generalization; Domain Adaptation; Microalbuminuria; ATR-FTIR

1. Introduction

Diabetic kidney disease (DKD) is caused by microvascular damage sustained as a result of diabetes and is the leading cause of chronic kidney disease (CKD) [1]. CKD affects 12% of people and will be the fifth cause of mortality worldwide in 2040 [2]. In 2021, an estimated 537 million people were estimated living with diabetes, 40% of whom are expected to develop DKD [3]. The main early marker of this disease is albuminuria (i.e., presence of albumin in the urine), which identifies the risk of progression of DKD and it is a therapeutic target and surrogate marker of clinical events [4]. Given the predictive value of albuminuria for the outcome of the disease [5], there is a need for an affordable and portable screening tool that can detect albumin, unlike immunoassay-based autoanalyzers, which provide optimal analytical and clinical performance but are best suited for batch analysis in hospital clinical laboratories. Urine dipsticks are the main screening tool, but have low sensitivity and specificity in the microalbuminuria range (30-300 mg/L) [6,7]. Considering the severity and high prevalence of DKD, the importance of accurately detecting and monitoring albuminuria should be accessible at any point of care within the health system such as pharmacies and primary care units.

Recently, we have developed a novel methodology integrating centrifugal ultrafiltration and Attenuated Total Reflectance-Fourier transform infrared (ATR-FTIR) [8,9]. The technique relies on machine learning (ML) methodologies to extract the information about the presence and abundance of specific proteins in the sample associated to DKD. Indeed, the use of sophisticated ML techniques to create predictive models is a common approach in clinical vibrational spectroscopy, where infrared (IR) and Raman spectra are treated with classification and regression approaches to offer label-free, fast, and cost-effective alternatives to label-based techniques. As an example, diseases like cancer [10], malaria [11], and COVID-19 [12] have been targeted for diagnosis using vibrational spectroscopy techniques, and these methods have demonstrated potential in identifying spectral markers, facilitating early detection and diagnosis.

However, the use of these powerful data analysis methodologies for the elucidation of spectral markers with discriminating power, while essential for exploiting the large amount of information contained on spectra of biological samples, comes with a cost. Due to the black-box nature of the 'vibrational signature' obtained by those highly sophisticated ML procedures, it is hard to establish how variations in the experimental process can affect the results of the methodology beyond the ones used in the calibration and test steps. This includes types and batches of instruments, personnel as well as demographics. In turn, this can be an impediment when scaling the methodology from proof-of-concept research to commercial solutions in the real world.

For example, the proposed methodology for DKD diagnosis was only calibrated and validated using male samples from the same region, Australia. As many diseases, DKD presents gender and ethnic differences [10,11], so different global regions present not only distinct incidence but also varied presentations of the disease [12,13]. Different health systems may have slightly different criteria for the diagnosis of the patients and their inclusion in the clinical study. Therefore, a new set of samples with different demographic characteristics may not generalize well, leading to errors that differ significantly from those observed in the original population.

Furthermore, apart from biological variance, the use of different equipment is a key limitation in the transference of a method to a new laboratory, as changes introduced by manufacturers introduce unknown variables that may have significant impacts on method performance [13]. In the case of FTIR, these non-biological variations could affect significantly the result, as usually spectral artifacts can appear under differing conditions [14]. Furthermore, in our case study we have to add the variability of sample drying, as differences in air flow can strongly influence such a complex process [8,15]. Beyond, as different filter materials may interact differently with the interest analyte and/or contaminants [16,17], it is a mandatory point to explore if this is also a factor affecting albumin preconcentration and hence, predictive capacities for DKD.

It is common to transfer analytical methods from the original location to other laboratories when scaling up clinical processes. However, it must be ensured that the receiving laboratories can obtain results with the same accuracy, precision and sensitivity as the source laboratory. In the field of ML based models, this step is critical, as differences in data collection, preprocessing, and confounding variables can significantly impact model performance, particularly in regression and classification tasks. The domain adaptation discipline aims to develop models capable of predicting samples outside the calibration space [18]. This is crucial for translating clinical spectroscopic models into real-world applications, as models should generalize effectively without requiring re-calibration for different clinics, instruments, or patient populations, which would be impractical and extremely costly.

In this research, we evaluated the IR based methodology for diagnosing microalbuminuria in a global setting, by employing new sample data obtained under conditions markedly distinct from those of the calibration dataset. Our aim was to evaluate the robustness of its diagnostic capabilities using different machine learning algorithms for practical application across various laboratories worldwide. To overcome challenges in measuring certain samples with both types of instruments for calibration transfer, we opted for a strategy of interpolating wavenumber values and normalizing spectra. Furthermore, we investigated the enhancement in overall performance of the classification models when the model was recalibrated using samples from both instruments. In total, we compared the adaptability capabilities of three different ML algorithms: Partial

Least Squares (PLS), Support Vector Machines (SVM) and Random Forest (RF). We also extended the reach of the method by evaluating the methodology for the identification of patients diagnosed with DKD, regardless of their levels of albuminuria.

2. Materials and Methods

2.1 Sample collection

Table 1 summarizes the differences among the two datasets compared in this study. Australian (N=155) samples were collected as described elsewhere [9]. Briefly, urine samples from male patients diagnosed with diabetes were collected during 24-hours. Urine control samples from male healthy volunteers with no history of diabetes, pre-diabetes, kidney disease, kidney stones or hypertension and who were not currently taking any medications or drugs were collected as a spot urine. This study was approved by the Austin Health Human Research Ethics Committee (Ethics ID: H2013/04937) and the Monash University Human Research Ethics Committee (Ethics ID:21405). All individuals provided written consent before analysis commenced.

Samples origin	DKD samples (n(%Males))	Control samples (n(%Males))	Sample collection	0.22 μM filter	10kDa filter	Spectrometer	Pipetted volume (μL)
Australia	155 (100)	22 (100)	24 hours	Corning cellulose acetate membrane	Amicon© Regenerated Cellulose	Bruker Alpha	0.8
Valencia (Spain)	20 (70)	26 (62)	Spot urine	Scharlau nylon membrane	Vivaspin® polyethersulfone (PES)	Perkin-Elmer Spectrum Two	1.0
Madrid (Spain)	15 (80)	0 (0)					

Table 1. Characteristics of the samples of this study, including the filters used for the preconcentration and how the spectra were acquired

Valencian spot urine samples (N=46) were collected at the Hospital Universitari i Politècnic La Fe. DKD patients did not exhibit any additional pathologies associated with increased albuminuria. Control urine samples were collected from healthy volunteers with no declared history of DKD and stored at -20ºC until use (Ethical approval by The Ethics Committee of Research in Humans of the Ethics Commission in Experimental Research of the University of Valencia, register code 2425449).

Madrid spot urine samples (N=15) were provided by the Hospital Universitario Puerta de Hierro de Majadahonda (HUPHM)/Instituto de Investigación Sanitaria Puerta de Hierro-Segovia de Arana (IDIPHISA) Biobank (Carlos III Health Institute Biomodels and Biobanks Platform – PT23/00015). They were processed following standard operating procedures with the appropriate approval of the Ethics and Scientific Committees (Nº 0128).

Additionally, the albumin content of each sample was obtained from the corresponding clinical laboratory, which served as reference values. Specifically, ELISA was used for the Australian samples [9] while immunoturbidimetry was used for the Spanish samples.

2.2 Sample processing

For the Australian dataset, samples were processed as described elsewhere [9]. In short, 700 μL of urine were drawn into a needleless syringe and dispensed through a 0.2 μM filter. 500 μL of filtrate were transferred into a centrifuge filter with a pore size of 10kDa, which had been washed as per the manufacturer's instructions. Samples were centrifuged at 14000 g for 15 minutes and the filtrate discarded. 450 μL of ultrapure water was added to the filter and the sample was spun at 14000 g for 15 minutes and the filtrate discarded. The washing process was repeated three times.

For the Spanish dataset (Madrid and Valencia), the procedure was the same but using 0.2 μM Scharlau syringe nylon filter (Scharlab, S.L., Barcelona, Spain) and Vivaspin® 500 10kDa filter (Sartorius, Göttingen, Germany).

2.3 ATR FTIR measurements

The Australian dataset was recorded using an Alpha FTIR spectrometer from Bruker (Ettlingen, Germany) as described elsewhere [9]. 0.8 μL of protein extract were pipetted onto the 1 reflexion ATR diamond cell and dried under gentle air flow for 2 minutes. 40 scans were co-averaged (measurement time ≈ 1 min) from 4000 cm^{1} to 800 cm^{1}. A background measurement of the clean diamond crystal with 128 scans was recorded between each biological replicate. The spectral resolution was 4 cm^{-1}. The instrument was operated with the OPUS software from Bruker, and the CO_2/H_2O correction integrated in this software was performed over the spectra prior to analysis.

The spectra of the Spanish samples were measured on a Spectrum Two FT-IR spectrometer from Perkin-Elmer (Waltham, MA, USA) equipped with the UATR accessory. 1 μL of protein extract sample was pipetted onto the 1 reflexion ATR diamond cell and dried under gentle airflow for 2 minutes. The spectrometer was controlled using the Spectrum (v10.03.06) software (Perkin-Elmer) and the data was corrected using the in-built CO_2/H_2O correction available in it. Spectra were

recorded from 450 cm^{-1} to 4000 cm^{-1} averaging 10 scans (measurement time $\approx$ 1 min) at a spectral resolution of 4 cm^{-1}. A background measurement of the clean diamond crystal was recorded averaging 10 scans.

2.4 Data pre-processing and analysis

Data visualization and analysis was carried out in MATLAB 2023b (MathWorks Inc., Natic, USA) using in-house written scripts, the MATLAB statistical and machine learning toolbox, the PLS Toolbox (Eigenvector Research Inc., Manson, USA) and the LIBSVM library [19]. All spectra were trimmed to the 3994-800 cm^{-1} region and interpolated to ensure they contained the same number of points at the same wavenumbers, using the interp1 function. To correct for differences in deposited volume, the Spanish spectra were scaled by a factor of 0.8.

First, principal component analysis (PCA) was performed for data exploration purposes. Then different ML models were constructed for both classification and regression tasks. All spectra underwent standard normal variate (SNV) scaling, Savitzky–Golay first derivative [20], and mean centering. The classification models aimed to distinguish between diabetic kidney disease (DKD) and control groups, including partial least squares discriminant analysis (PLS-DA), support vector machines for classification (SVM-DA), and random forest (RF) for classification. Additionally, corresponding regression models were built using albumin concentration as the target variable, specifically partial least squares (PLSR), support vector regression (SVR), and RF. To assess domain transfer classification and the feasibility of constructing a global model, different combinations of training and test sets were evaluated, considering the sample origin.

PLSR, PLS-DA, SVMR and SVM-DA were developed using the PLS toolbox libraries. The hyperparameters' optimization was carried out using a 10 splits Venetian Blinds cross-validation (CV) (i.e., 10-fold CV with systematic interleaving). The optimal number of latent variables (LVs) of the PLS and PLS-DA models was optimized by minimizing the RMSECV (root mean square error of CV) and the CV classification error for regression and discriminant analysis, respectively. Regarding SVM, the radial basis function (RBF) was chosen as the kernel function to effectively capture non-linear relationships within the data. The optimal parameters, (C and γ), were determined using a grid search approach, aiming to minimize cross-validation (CV) error. The range explored for C and γ values were 10^{-3} to 100 and 10^{-6} to 10, respectively. RF models were developed using the statistical and machine learning toolbox from Matlab. The number of learning cycles, the learn rate and the maximum number of splits were optimized using bayesian optimization.

All scripts used for model implementation, along with the raw data and detailed parameters, are available in the public dataset hosted on the Zenodo repository (https://doi.org/10.5281/zenodo.14762603).

3. Results and Discussion

3.1 Data overview

First, we explored visually how the differences between the urine datasets could affect the spectra. Figure 1A depicts the average DKD and control raw spectra. It can be seen several typical bands of proteins such as the Amide A N-H stretching at 3286 cm^{-1}, Amide I at 1644 cm^{-1} and Amide II at 1538 cm^{-1} [21]. In addition, the set of overlapped bands found in the 1200-1000 cm^{-1} region can be assigned to glycoproteins [22], being the glycoprotein uromodulin the most abundant protein excreted in urine of healthy individuals [23]. In general, the DKD patients show larger intensities of protein bands than controls for both the Spanish and Australian set of data, indicative of an increase of urinary albumin excretion. Standard Normal Variate (SNV) normalization of the spectra (See Figure 1B) allows for a qualitative comparison of the protein composition without considering total protein concentration. In general, the controls from Australia and Spain showed higher intensities of the bands associated with glycoproteins in the 1200–1000 cm^{-1} region. This indicates that the proportion of albumin relative to the total protein content is higher in the urine of DKD patients, while the control samples exhibit more abundance in glycoproteins (uromodulin).
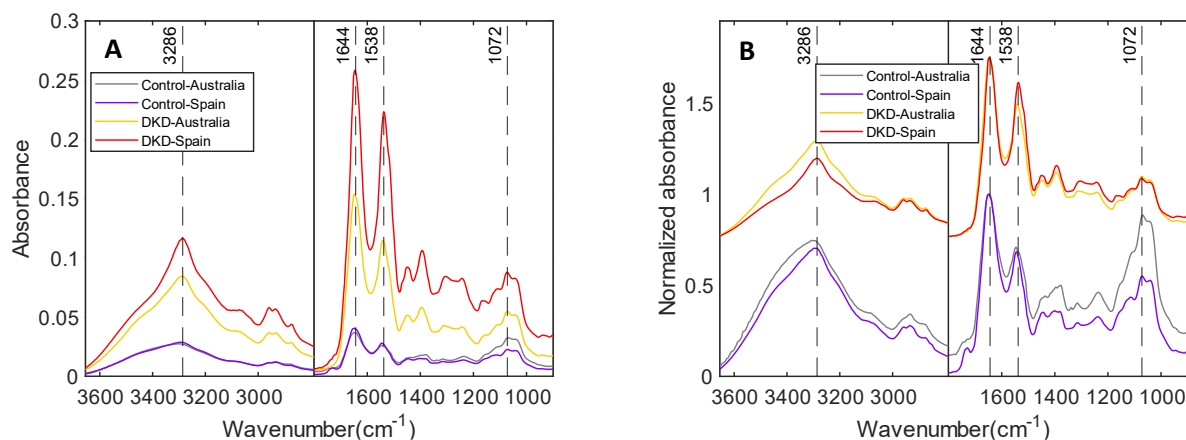
Figure 1: Mean spectra of Control vs. DKD (offset) for Spanish and Australian samples: Raw (A) and Standard Normal Variate (SNV) normalized (B).

Principal Component Analysis (PCA) was performed on the merged dataset including samples from both countries. The spectra were pre-processed by applying a first derivative followed by mean centering. Only the 3654-2866 cm^{-1} and the 1786-1012 cm^{-1} regions were included in the modelling. The PC1 loadings, which explains the 94% of the spectral variance, are depicted in Figure 2C (blue) and correspond with a typical spectrum of albumin, highlighting the Amide I, II and A bands. Control samples have, as expected, lower score values of PC1 than DKD samples (Fig. 2A). PC2 unfortunately, differentiates DKD samples according to their origin. It can be seen that PC2 decreases concurrently with an increase in PC1 (albumin content) between -0.5 to 0 until a critical point is reached (See Arrows in the x-axis). Notably, Spanish samples reach this inflection point at a lower PC1 value compared to Australian samples (see arrows in Fig. 2A).
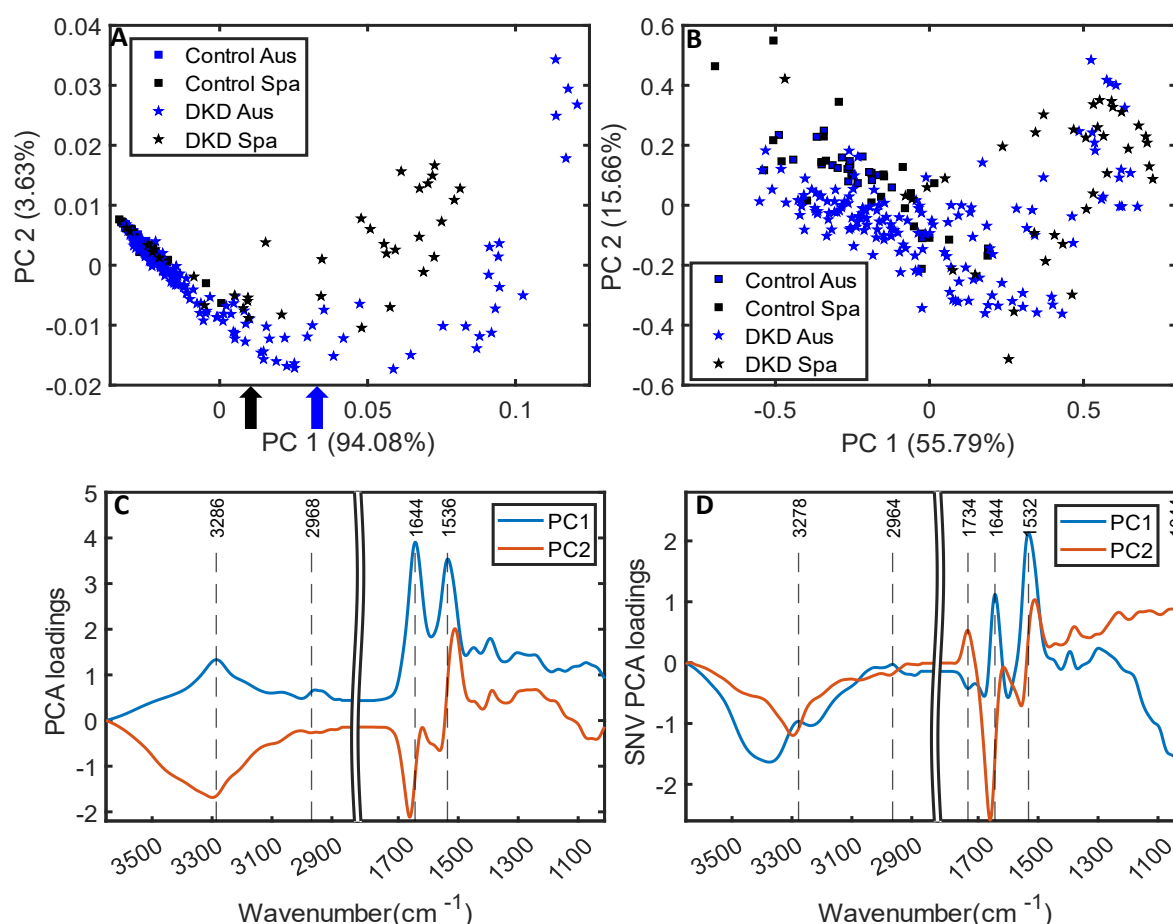
Figure 2: PC1 and PC2 scores for the whole dataset and their corresponding loadings without (A, C) and with spectral normalization (B, D). Squares and stars represent controls and DKD, respectively. Blue and black colors represent Australian and Spanish samples, respectively. Arrows (A) indicate the inflection point for both countries. Since the spectra were derivatized, the depicted loadings represent the cumulative sum of the obtained loadings, computed using MATLAB cumsum function to aid interpretation.

To minimize differences related to sample origin, an additional PCA was conducted, this time incorporating SNV normalization as a preprocessing step. In this analysis, PC1 and PC2 accounted for 56% and 16% of the spectral variance, respectively. The score values in Figure 2C indicate that the samples do not cluster according to their origin, demonstrating

that normalization effectively integrated both datasets. The clustering associated with the DKD class appeared less distinct compared to the non-normalized data, although the control samples were generally located in the upper left corner of the score plot. Notably, the loading in Figure 2D (blue) shows positive bands associated with albumin alongside negative bands linked to glycoproteins and other metabolites naturally present in healthy samples. Furthermore, PC2 exhibited considerable similarities with the corresponding loading from the previous PCA analysis.

3.2 Domain Transfer Classification

Next, we assessed the adaptation capability of classification models created with the Australian dataset to predict samples from the Spanish dataset. To that end, PLSDA, SVMDA and RF models were calibrated and optimized using the Australian samples and then models were independently tested using the Spanish set.

Table 2. Analytical figures of merit for the independent validation of the different machine learning models for the domain transfer classification. Note: AUROC stands for area under the receiver operating characteristic

	CAL Samples	TEST Samples	AUROC			Sensitivity (%)			Specificity(%)		
			PLSDA	SVMDA	RF	PLSDA	SVMDA	RF	PLSDA	SVMDA	RF
DKD	155 + (AUS) 22 – (AUS)	35 + (SPA) 26 – (SPA)	0.870	0.787	0.697	82.9	100	97.1	88.5	27.0	42.3
Albuminuria	64 + (AUS) 113 – (AUS)	34 + (SPA) 27 – (SPA)	0.950	0.980	0.934	94.1	94.1	94.1	66.7	85.2	92.6

The technique was previously evaluated for the detection of microalbuminuria levels [24], but patients diagnosed with DKD may did not exhibit microalbuminuria levels in their urine samples. Therefore, here we studied the differentiation of DKD diagnosed patients' samples (including those with low levels of albuminuria) from control group samples. Classification performance indicators can be found in the first row of Table 2 and Figure 3. PLSDA yielded the best prediction error, obtaining sensitivity, specificity and AUROC values of 82.9, 88.5 and 0.87, respectively. Figure 3A depicts the probabilities predicted by the PLSDA model, showing that the separation among classes enabled the independent prediction of 52 of the 61 Spanish samples correctly (85%). This value clearly indicates that the spectral markers found in the Australian model were robust and consistent enough to be transferred to the Spanish dataset, even if the model had never been trained with any Spanish data and even both datasets contained inconsistencies in sample population and experimental measurement conditions.

On the other hand, interestingly, SVMDA and RF exhibited very poor specificities of 27.0% and 42.3%, respectively. This may be explained by the imbalance of the datasets. Australian number of control samples, 22, was really small compared to the 155 DKD Australian samples, making models prone to predict spectra as positive. Figure 3B and 3C shows that the predictions tended to favour the classification of almost all the samples as DKD, leading to the poor performance of these non-linear machine learning methods.
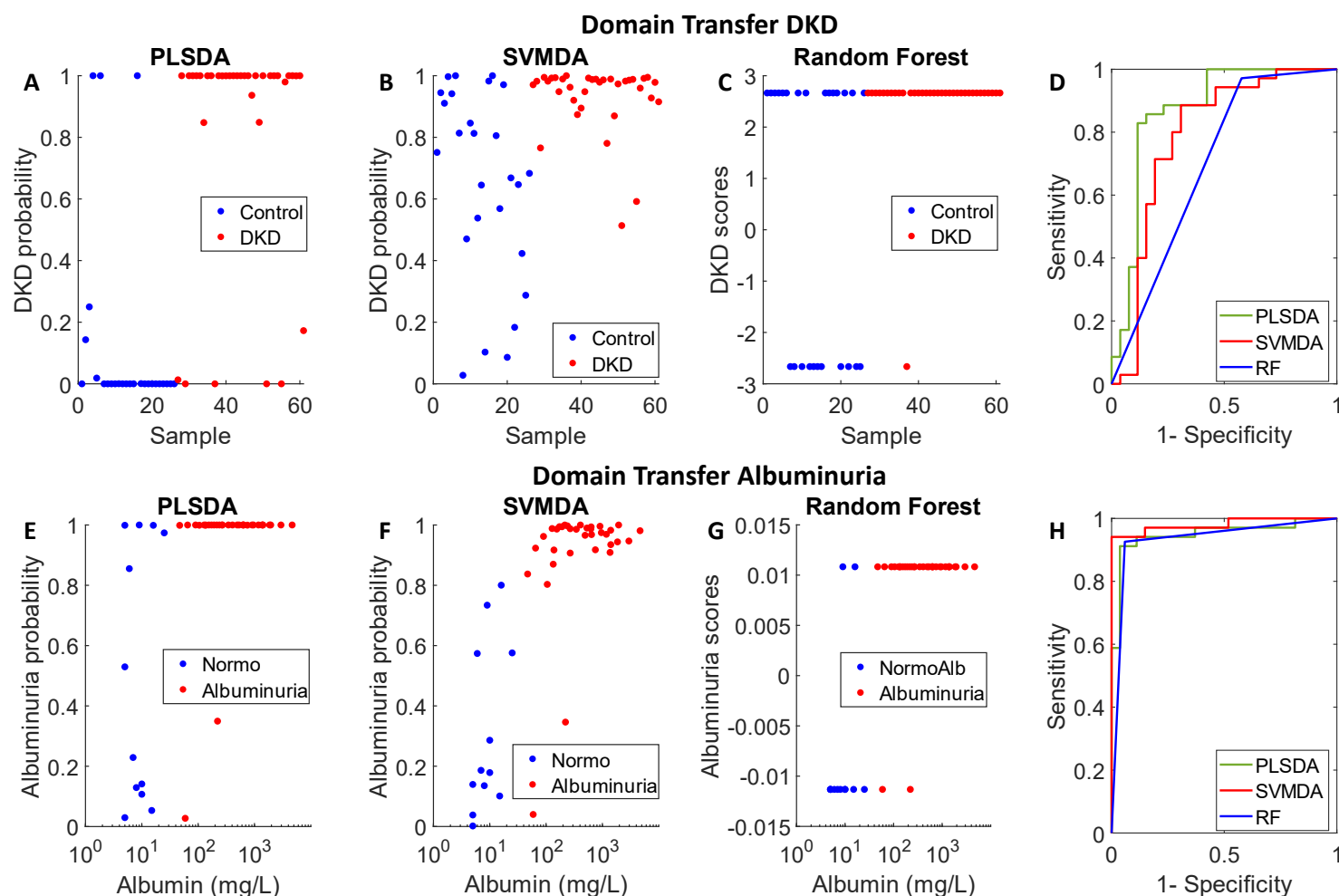
Figure 3: Independent testing results obtained for the prediction of the Spanish dataset with models calibrated solely with the Australian samples. Predicted Probability of being DKD for PLSDA (A), SVMDA (B) and RF (C) models and their respective ROC curves (D). Probability of being an albuminuric sample represented vs. albumin concentrations in logarithmic scale for PLSDA (E), SVMDA (F) and RF (G) and their respective ROC curves (H). For RF, the scores are displayed.

Secondly, we investigated the capability of machine learning models to discriminate samples according to its albumin content, defining two ranges: normoalbuminuria (albumin concentration <30 mg/L) or micro/macroalbuminuria (albumin concentration ≥ 30 mg/L) [3]. Overall, this represented a more balanced (64 & 113) sample set for the calibration. Statistical values can be found in the second row of Table 2 and Figure 3. This time the prediction capabilities were improved for both SVMDA and RF, achieving to classify urine samples based on their albumin excretion level with AUROC values above 0.930 for all three machine learning models (See Figure 3H), with RF reaching the best sensitivity and specificity values of

94.1 and 92.6, respectively. Figure 3E-G also shows the prediction probabilities as a function of the albumin concentration, evidencing a sigmoidal distribution of the probabilities with a grey zone of diagnosis around the cut-off value.

In summary, considering all the technical and biological variations between samples of Australia and Spain, classification models retained relatively strong predictive capabilities. Results indicate that the predictive spectral markers were robust enough to overcome most of the potential batch effects even if only Australian samples were utilized for model calibration, and Spanish samples were not included in the construction of the model.
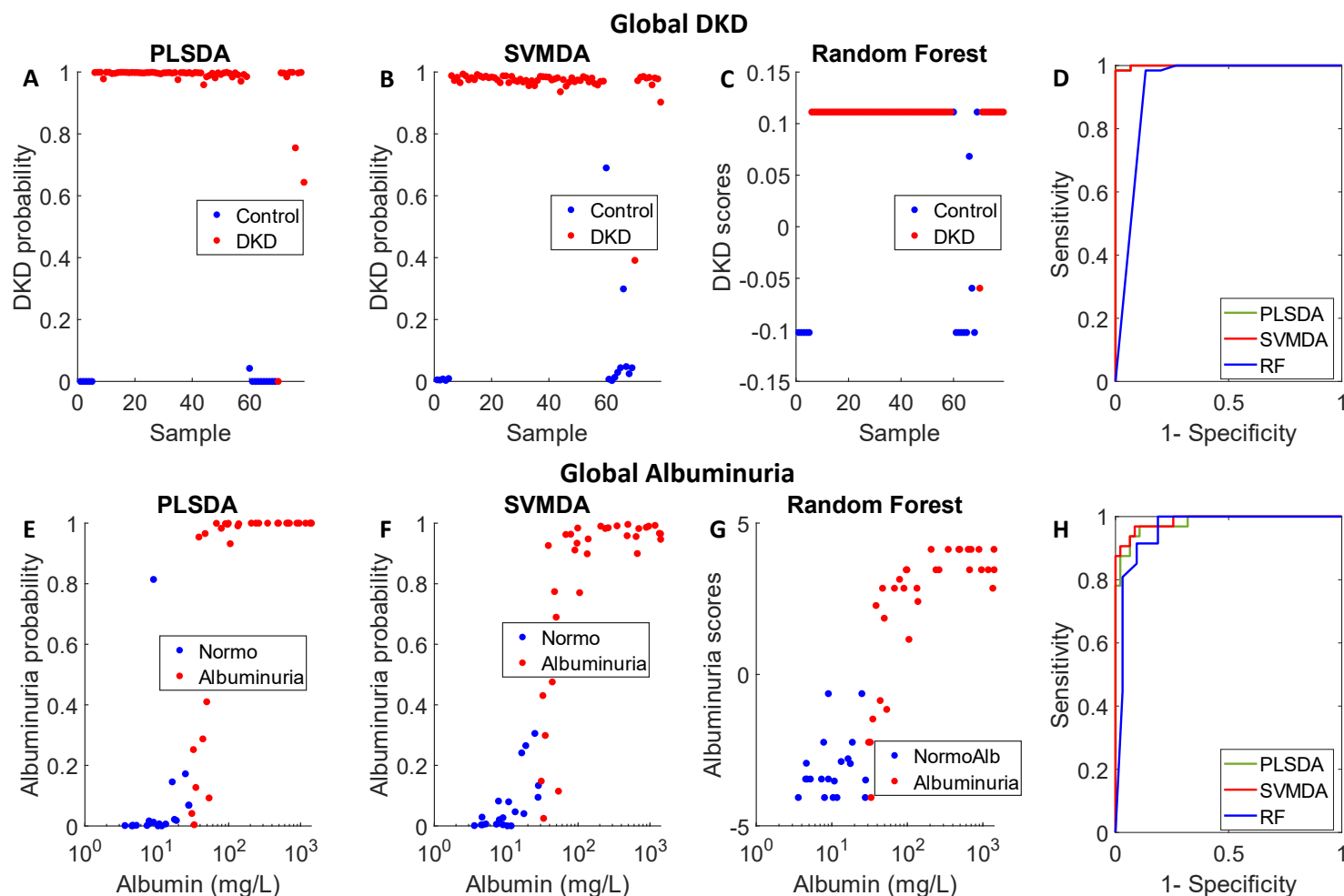
3.3 Global Model

Seeking to enhance the prediction robustness, we also followed a global approach, where both, the Spanish and the Australian spectra were integrated as a unique dataset. For this, the whole dataset was divided into a calibration (2/3) and test dataset (1/3) using the Kennard-Stone algorithm, which selects samples providing uniform coverage of the dataset [25]. Thus, both the calibration and validation datasets included samples from both countries.

Independent testing statistics available in Table 3 indicated that the incorporation of the integrated dataset improved the classification performance for both DKD vs. control and normo vs. micro/macroalbuminuria. Notably, the choice of the ML method became less crucial as in this case all perform more similarly. For the DKD (See Figure 4 A-C) classification, PLSDA and SVMDA obtained AUROC, sensitivity and specificity values were found to be above 0.99, 98.4% and 93.3%, respectively.

Table 3. Analytical figures of merit for the independent validation of the different machine learning models for the global classification. Note: AUROC stands for area under the receiver operating characteristic.

	CAL Samples	TEST Samples	AUROC			Sensitivity (%)			Specificity(%)		
			PLSDA	SVMDA	RF	PLSDA	SVMDA	RF	PLSDA	SVMDA	RF
DKD	101 + (AUS) 25 + (ESP) 17 – (AUS) 16 – (ESP)	54 + (AUS) 10 + (ESP) 5 – (AUS) 10 – (ESP)	0.999	0.999	0.931	98.4	98.4	98.4	100	93.3	80.0
Albuminuria	43 + (AUS) 23 + (ESP) 75 – (AUS) 18 – (ESP)	21 + (AUS) 11 + (ESP) 38 – (AUS) 9 – (ESP)	0.981	0.987	0.957	78.1	81.3	81.3	97.9	100	100

The albuminuria classification achieved AUROC values above 0.96 and specificity values exceeding 97.9%. In this case, sensitivity values were more modest (around 80%), but Figure 4E–G shows that most misclassifications occurred around the 30 mg/L cut-off for the microalbuminuria region. Since sample classification criteria are based on a continuous scale, some classification uncertainty within this range is expected.


Figure 4: Independent testing results obtained for the prediction of the independent test set in the "global" model, which integrated Spanish and Australian spectra. Predicted Probability of being DKD for PLSDA (A), SVMDA (B) and RF (C) models and their respective ROC curves (D). Probability of being an albuminuric sample represented vs. albumin concentrations in logarithmic scale for PLSDA (E), SVMDA (F) and RF (G) global models. (H) Their respective ROC curves. For RF, the scores are displayed.

In general, global models proved to outperform those calibrated using only samples from one dataset, showing the compatibility of the spectral markers obtained in the different regions. Overall, PLS was outperformed by non-linear machine learning algorithms, although in the global models the differences between models were reduced. The only condition under which SVM and RF struggled was the domain transfer DKD classification, where the non-linear models lose performance due to unbalanced classes in the training dataset [26,27]. Thus, strategies to compensate for the unbalanced set of samples should be considered before choosing a ML algorithm.

3.3 Study of discriminative features.

Classification results indicated that spectral markers were robust and comparable across datasets, and inter-batch differences were not considerably enough to profoundly compromise their use. We aimed to study the aetiology of these spectral markers, which represented quantitative and qualitative changes in protein composition associated to DKD and microalbuminuria. In general, both classifications were expected to be correlated, as patients diagnosed with DKD are expected to have high levels of albuminuria. However, they should also show differences, because DKD diagnostic criteria also consider the estimated glomerular filtration rate (eGFR). In our datasets there were several normoalbuminuric samples from DKD patients, as there is a substantial number of cases diagnosed with this renal disease that present normal albuminuria levels [28]. Results described in sections 3.1 and 3.2 showed that the DKD classification models were able to discriminate between normoalbuminuric samples from DKD patients and those from controls.

PLSDA was used to study the spectral markers associated to qualitative and quantitative changes in proteins, as the black-box nature of SVMDA and RF poses several challenges in the interpretation of the models. PLSDA regression vectors for global classification models are shown in Figure 5, with points coloured according to the variable importance projection (VIP) score. While both regression vectors show similarities in the higher region (3200-2800 cm^{-1}), they also display notable differences.
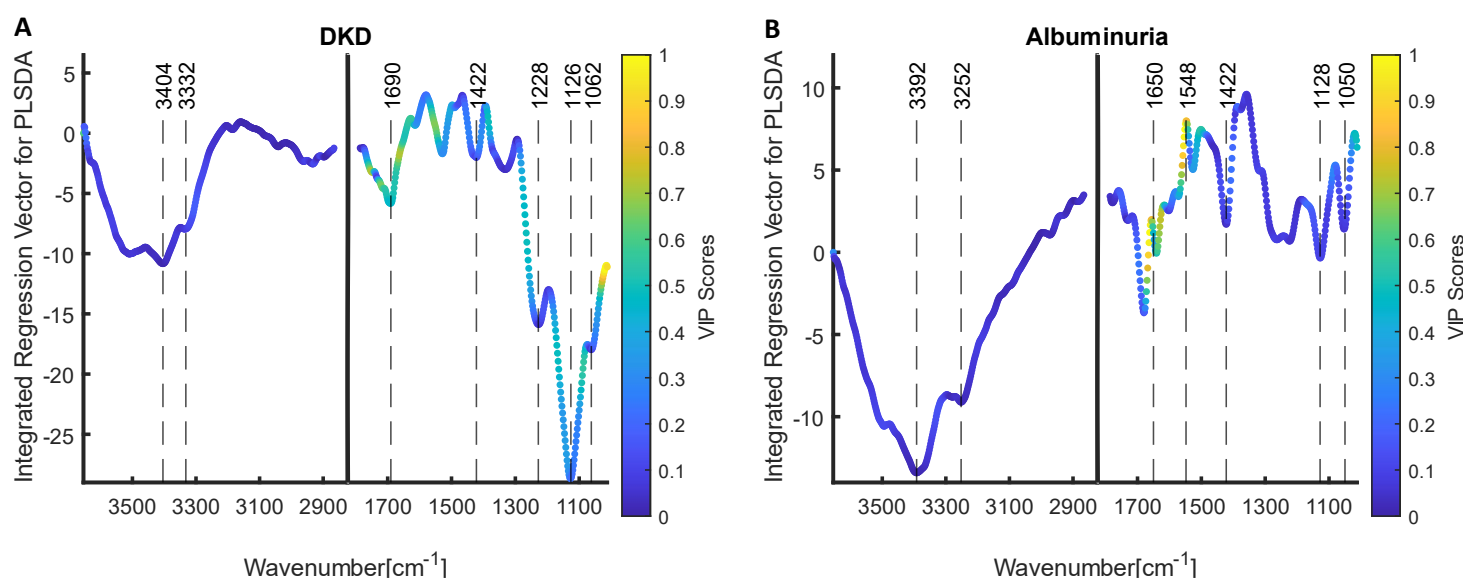
Figure 5: Integrated regression vector of the PLSDA classification global model for DKD vs. control (A) and normoalbuminuria vs. micro/macroalbuminuria (B). These are colored according to their variable importance in projection (VIP) scores. Since the spectra were derivatized, the depicted regression vectors represent the cumulative sum of the obtained regression vectors, computed using MATLAB cumsum function to aid interpretation.

In the DKD classification (see Figure 5A), the C-O and C-C stretching vibrations found at 1200-1000 cm^{-1} associated with glycoproteins are extremely negative [29]. Furthermore, the VIP scores highlight the importance of these variables along with the Amide I and Amide II region. All these differences are in accordance with the normalized mean spectra of Control and DKD samples from both datasets (Fig. 1), where the 3600-3200 cm^{-1} and 1200-1000 cm^{-1} regions show a higher absorbance in controls. In summary, the analysis of the discriminant features following spectral normalization for DKD indicates that patients are classified based on their specific protein profiles. In healthy individuals, low levels of proteins, primarily uromodulin, are typically excreted, and a reduced urinary proportion of this glycoprotein has been linked to kidney damage [30]. In contrast, the presence of DKD increases considerably the proportion of albumin.
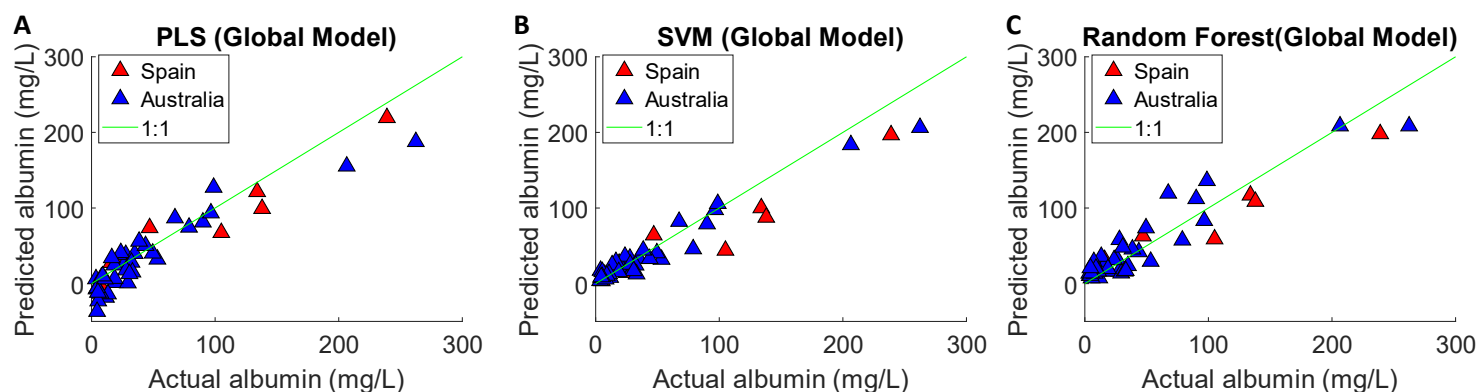
In the case of the regression vector of the albuminuria model (See Figure 5B), discriminative features are similar in the amide bands and to a lesser extent in the 1200-1000 cm^{-1} region. In this case, VIP scores emphasize the significance of the positive bands assigned to Amide I (1650 cm^{-1}) and Amide II (1548 cm^{-1}), which show positive values directions in the regression vector, indicating that the model was driven by the presence and concentration of typical protein bands.

3.4 Regression

Finally, regression models to quantify the concentration of albumin in urine using the global dataset were developed. Again, the global dataset containing all Australian and Spanish dataset was split between calibration (2/3) and test (1/3) sets using the Kennard-Stone algorithm [25]. The calibration set was used to develop and optimize PLSR, SVR and RF models that were independently tested with the test set. Since samples with high values of macroalbuminuria skewed the model and provoked high errors at lower concentrations around the low levels, which are the clinically significant ones, we opted to exclude macroalbuminuric samples directly, aiming to enhance the diagnostic accuracy within the normo/microalbuminuria range (up to 300 mg/L). Table 4 depicts the R^2 values and root mean square errors of prediction (RMSEP) for the independent testing, showing values normalized to the mean value of normoalbuminuria and microalbuminuria samples. Overall, the SVR model outperformed PLSR and RF in the normoalbuminuria range, with errors below 10 mg/L, while for the microalbuminuria range, errors where comparable (28 mg/L) for the three methodologies. Figure 6 depicts the actual vs. predicted plots, demonstrating that the model exhibited strong predictive performance for albuminuria concentrations across the entire range, with R^2 values exceeding 0.89. Figure 6A highlights the limitations of the PLSR algorithm in predicting low-concentration samples, with several predictions yielding negative values. Most importantly, samples from Australia and Spain were predicted similarly, and do not show major clustering or bias in their predictions.

Table 4. Characteristics of different global regression models for the quantification of albumin in urine samples. Note: RMSEP stans for Root Mean Squared Error of Prediction.

	R²(Pred)	RMSEP (mg/L) (NormoAlb)	RMSEP (mg/L) Normalized (NormoAlb)	RMSEP (mg/L) (MicroAlb)	RMSEP (mg/L) Normalized (MicroAlb)
PLSR	0.89	18	1.26	28	0.29
SVR	0.93	7	0.52	28	0.30.
RF	0.89	12	0.87	28	0.30

Figure 6: Actual vs. Predicted concentrations of albumin for the PLS (A), SVM (B) and Random Forest (C) regression models. Spanish and Australian samples are colored in red and blue, respectively.

4. Conclusions and outlook

In this study we have evaluated the effectiveness and adaptability of a methodology based on IR spectroscopy and machine learning for the identification of microalbuminuria and DKD in a global context at any point of care. First, we validated models created using the Australian dataset by predicting Spanish data collected under significantly different conditions. For example, the Australian model was constructed using only male samples, whereas the new Spanish dataset included some female samples, which may have impaired predictions due to documented gender-related differences in the urine proteome [31]. Moreover, the Spanish samples were derived from spot urine, which may not fully reflect the daily albumin excretion, potentially causing the model to rely on the qualitative composition of proteins.

Despite this, PLS models were able to adapt to the new dataset and provide consistent identifications of DKD, while PLS, RF and SVM models showed strong prediction capabilities for the diagnosis of albuminuria. Furthermore, incorporating a representative subset from both datasets into the calibration set and updating the model substantially increased considerably the prediction capability of the model, providing sensitivity and specificity values above 90%. Regression vectors showed that the spectral signatures captured by the models featured bands of albumin for the albuminuria prediction and bands of albumin and glycoprotein for the diagnosis of DKD.

Results indicated the robustness and adaptability of this methodology, enhancing its potential for use in clinical laboratories with different equipment and sample types. The models could adapt to deal with variations among samples and non-biological factors, minimizing their impact on performance and reducing the need to process and measure a subset of samples in both laboratories. Results also encourage the compilation of samples from different setups into global datasets and repositories, which can help develop big data-driven diagnostic models for routine diagnosis and monitoring of albuminuria. These models could be applied at any point of care within the health system. To support this, we have provided the full dataset used, which can be leveraged to expand the dataset size and further enhance the performance of future models based on this approach.

CRediT authorship contribution statement

Conceptualization, Data curation, Software, Experimental Design, Formal analysis and Methodology: VNE and DPG; Supervision: DPG; Funding acquisition: DPG, ASI, MMV, JP; Resources: all authors; Validation and Writing - Original Draft: VNE; Investigation: VNE, DPG, ASI; Writing - Review and editing: VNE, DPG, ASI, JP, MMV, IVB, AJSL, BRW.

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.

Data availability

Datasets and details of the data processing can be found at the Zenodo repository (https://doi.org/10.5281/zenodo.14762603).

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT-4o to refine the text. After using this service, authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

References

[1] M.C. Thomas, M. Brownlee, K. Susztak, K. Sharma, K. Jandeleit-Dahm, S. Zoungas, P. Rossing, P.-H. Groop, M.E. Cooper, Diabetic kidney disease, Nat Rev Dis Primers 1 (2015) 15018. https://doi.org/10.1038/nrdp.2015.18.

[2] A. Ortiz, E.K.P.F. (EKPF) Asociación Información Enfermedades Renales Genéticas (AIRG-E) Federación Nacional de Asociaciones para la Lucha Contra las Enfermedades del Riñón (ALCER), Fundación Renal Íñigo Álvarez de Toledo (FRIAT), Red de Investigación Renal (REDINREN), Resultados en Salud 2040 (RICORS2040), Sociedad Española de Nefrología (SENEFRO) Council, Sociedad Española de Trasplante (SET) Council, Organización Nacional de Trasplantes (ONT), RICORS2040: the need for collaborative research in chronic kidney disease, Clinical Kidney Journal 15 (2022) 372–387. https://doi.org/10.1093/ckj/sfab170.

[3] P. Rossing, M.L. Caramori, J.C.N. Chan, H.J.L. Heerspink, C. Hurst, K. Khunti, A. Liew, E.D. Michos, S.D. Navaneethan, W.A. Olowu, T. Sadusky, N. Tandon, K.R. Tuttle, C. Wanner, K.G. Wilkens, S. Zoungas, I.H. de Boer, KDIGO 2022 Clinical Practice Guideline for Diabetes Management in Chronic Kidney Disease, Kidney International 102 (2022) S1–S127. https://doi.org/10.1016/j.kint.2022.06.008.

[4] CKD Evaluation and Management – KDIGO, (n.d.). https://kdigo.org/guidelines/ckd-evaluation-and-management/ (accessed February 21, 2024).

[5] V. Lezaic, Albuminuria as a Biomarker of the Renal Disease, in: V.B. Patel (Ed.), Biomarkers in Kidney Disease, Springer Netherlands, Dordrecht, 2015: pp. 1–18. https://doi.org/10.1007/978-94-007-7743-9_31-1.

[6] D. Vuljanić, A. Dojder, V. Špoljarić, A. Saračević, L. Dukić, J. Leniček-Krleža, J. Vlašić-Tanasković, I. Maradin, A. Grzunov, Ž. Vogrinc, A.-M. Šimundić, Analytical verification of 12 most commonly used urine dipsticks in Croatia: comparability, repeatability and accuracy, Biochem Med (Zagreb) 29 (2019) 010708. https://doi.org/10.11613/BM.2019.010708.

[7] J.R. Mejia, J.E. Fernandez-Chinguel, G. Dolores-Maldonado, N. Becerra-Chauca, S. Goicochea-Lugo, P. Herrera-Añazco, J.H. Zafra-Tanaka, A. Taype-Rondan, Diagnostic accuracy of urine dipstick testing for albumin-to-creatinine ratio and albuminuria: A systematic review and meta-analysis, Heliyon 7 (2021) e08253. https://doi.org/10.1016/j.heliyon.2021.e08253.

[8] D. Perez-Guaita, Z. Richardson, P. Heraud, B. Wood, Quantification and Identification of Microproteinuria Using Ultrafiltration and ATR-FTIR Spectroscopy, Anal. Chem. 92 (2020) 2409–2416. https://doi.org/10.1021/acs.analchem.9b03081.

[9] Z. Richardson, A. Kincses, E. Ekinci, D. Perez-Guaita, K. Jandeleit-Dahm, B.R. Wood, ATR-FTIR Spectroscopy for Early Detection of Diabetic Kidney Disease, Analysis & Sensing n/a (n.d.) e202200094. https://doi.org/10.1002/anse.202200094.

[10] M. Kujdowicz, D. Perez-Guaita, P. Chłosta, K. Okon, K. Malek, Towards the Point of Care and noninvasive classification of bladder cancer from urine sediment infrared spectroscopy. Spectral differentiation of normal, abnormal and cancer patients, Microchemical Journal 168 (2021) 106460. https://doi.org/10.1016/j.microc.2021.106460.

[11] D. Perez-Guaita, K.M. Marzec, A. Hudson, C. Evans, T. Chernenko, C. Matthäus, M. Miljkovic, M. Diem, P. Heraud, J.S. Richards, D. Andrew, D.A. Anderson, C. Doerig, J. Garcia-Bustos, D. McNaughton, B.R. Wood, Parasites under the Spotlight: Applications of Vibrational Spectroscopy to Malaria Research, Chem. Rev. 118 (2018) 5330–5358. https://doi.org/10.1021/acs.chemrev.7b00661.

[12] B.R. Wood, K. Kochan, D.E. Bedolla, N. Salazar-Quiroz, S.L. Grimley, D. Perez-Guaita, M.J. Baker, J. Vongsvivut, M.J. Tobin, K.R. Bambery, D. Christensen, S. Pasricha, A.K. Eden, A. Mclean, S. Roy, J.A. Roberts, J. Druce, D.A. Williamson, J. McAuley, M. Catton, D.F.J. Purcell, D.I. Godfrey, P. Heraud, Infrared Based Saliva Screening Test for COVID-19, Angew Chem Int Ed Engl 60 (2021) 17102–17107. https://doi.org/10.1002/anie.202104453.

[13] G.L. Reid, Chapter 6 - Analytical method transfer, in: C.M. Riley, T.W. Rosanske, G. Reid (Eds.), Specification of Drug Substances and Products (Second Edition), Elsevier, 2020: pp. 125–148. https://doi.org/10.1016/B978-0-08-102824-7.00006-3.

[14] C.L.M. Morais, M. Paraskevaidi, L. Cui, N.J. Fullwood, M. Isabelle, K.M.G. Lima, P.L. Martin-Hirsch, H. Sreedhar, J. Trevisan, M.J. Walsh, D. Zhang, Y.-G. Zhu, F.L. Martin, Standardization of complex biologically derived spectrochemical datasets, Nat Protoc 14 (2019) 1546–1577. https://doi.org/10.1038/s41596-019-0150-x.

[15] J.M. Cameron, H.J. Butler, D.S. Palmer, M.J. Baker, Biofluid spectroscopic disease diagnostics: A review on the processes and spectral impact of drying, Journal of Biophotonics 11 (2018) e201700299. https://doi.org/10.1002/jbio.201700299.

[16] G. Vergauwen, B. Dhondt, J. Van Deun, E. De Smedt, G. Berx, E. Timmerman, K. Gevaert, I. Miinalainen, V. Cocquyt, G. Braems, R. Van den Broecke, H. Denys, O. De Wever, A. Hendrix, Confounding factors of ultrafiltration and protein analysis in extracellular vesicle research, Sci Rep 7 (2017) 2704. https://doi.org/10.1038/s41598-017-02599-y.

[17] A. Kratzer, F. Kees, C. Dorn, Unbound fraction of fluconazole and linezolid in human plasma as determined by ultrafiltration: Impact of membrane type, Journal of Chromatography B 1039 (2016) 74–78. https://doi.org/10.1016/j.jchromb.2016.10.040.

[18] J. Li, Z. Yu, Z. Du, L. Zhu, H.T. Shen, A Comprehensive Survey on Source-Free Domain Adaptation, IEEE Transactions on Pattern Analysis and Machine Intelligence 46 (2024) 5743–5762. https://doi.org/10.1109/TPAMI.2024.3370978.

[19] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011) 27:1-27:27. https://doi.org/10.1145/1961189.1961199.

[20] Abraham. 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.

[21] A. Barth, Infrared spectroscopy of proteins, Biochimica et Biophysica Acta (BBA) - Bioenergetics 1767 (2007) 1073–1101. https://doi.org/10.1016/j.bbabio.2007.06.004.

[22] S.-M. Kim, Y.-R. Chang, J. Melby, Y.J. Kim, D. Davis, Y.J. Lee, Quantum Cascade Laser Infrared Spectroscopy for Glycan Analysis of Glycoprotein Solutions, Anal. Chem. 96 (2024) 13120–13130. https://doi.org/10.1021/acs.analchem.4c01772.

[23] S. Youhanna, J. Weber, V. Beaujean, B. Glaudemans, J. Sobek, O. Devuyst, Determination of uromodulin in human urine: influence of storage and processing, Nephrol Dial Transplant 29 (2014) 136–145. https://doi.org/10.1093/ndt/gft345.

[24] Z. Richardson, A. Kincses, E. Ekinci, D. Perez-Guaita, K. Jandeleit-Dahm, B.R. Wood, ATR-FTIR Spectroscopy for Early Detection of Diabetic Kidney Disease, Analysis & Sensing 3 (2023) e202200094. https://doi.org/10.1002/anse.202200094.

[25] R.W. Kennard, L.A. Stone, Computer Aided Design of Experiments, Technometrics 11 (1969) 137–148. https://doi.org/10.2307/1266770.

[26] J. Tang, A. Henderson, P. Gardner, Exploring AdaBoost and Random Forests machine learning approaches for infrared pathology on unbalanced data sets, Analyst 146 (2021) 5880–5891. https://doi.org/10.1039/D0AN02155E.

[27] R. Akbani, S. Kwek, N. Japkowicz, Applying Support Vector Machines to Imbalanced Datasets, in: J.-F. Boulicaut, F. Esposito, F. Giannotti, D. Pedreschi (Eds.), Machine Learning: ECML 2004, Springer Berlin Heidelberg, Berlin, Heidelberg, 2004: pp. 39–50. https://doi.org/10.1007/978-3-540-30115-8_7.

[28] J.P. Dwyer, H.-H. Parving, L.G. Hunsicker, M. Ravid, G. Remuzzi, J.B. Lewis, Renal Dysfunction in the Presence of Normoalbuminuria in Type 2 Diabetes: Results from the DEMAND Study, Cardiorenal Med 2 (2012) 1–10. https://doi.org/10.1159/000333249.

[29] A. Derenne, K.-M. Derfoufi, B. Cowper, C. Delporte, C.I. Butré, E. Goormaghtigh, Analysis of Glycoproteins by ATR-FTIR Spectroscopy: Comparative Assessment, in: A. Delobel (Ed.), Mass Spectrometry of Glycoproteins: Methods and Protocols, Springer US, New York, NY, 2021: pp. 361–374. https://doi.org/10.1007/978-1-0716-1241-5_25.

[30] R. Thielemans, R. Speeckaert, C. Delrue, S. De Bruyne, M. Oyaert, M.M. Speeckaert, Unveiling the Hidden Power of Uromodulin: A Promising Potential Biomarker for Kidney Diseases, Diagnostics 13 (2023) 3077. https://doi.org/10.3390/diagnostics13193077.

[31] C. Shao, M. Zhao, X. Chen, H. Sun, Y. Yang, X. Xiao, Z. Guo, X. Liu, Y. Lv, X. Chen, W. Sun, D. Wu, Y. Gao, Comprehensive Analysis of Individual Variation in the Urinary Proteome Revealed Significant Gender Differences, Mol Cell Proteomics 18 (2019) 1110–1122. https://doi.org/10.1074/mcp.RA119.001343.