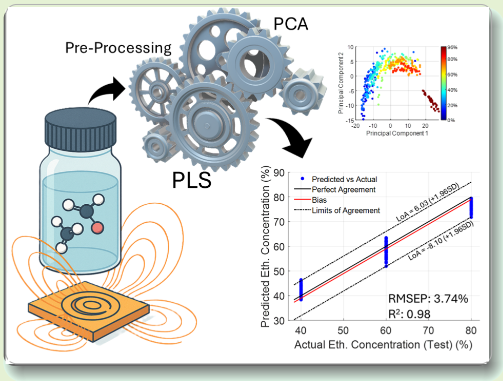
CSRR chemical sensing in uncontrolled environments by PLS regression

Javier Alonso-Valdesueiro, Luis Fernández, Agustín Gutiérrez-Gálvez, and Santiago Marco

***Abstract*—Complementary Split Ring Resonators (CSRRs) have been extensively studied as planar sensors in the last two decades. However, their practical use remains limited to controlled environments and classification problems. Their performance reliance on high-end Vector Network Analyzers (VNAs), highly repeatable laboratory conditions, and special sample holders or microfluidic circuits hinders its regular use in chemistry laboratories as analytical tool. Temperature drifts and humidity variations during measuring, uncertainties in the electromagnetic properties of the sample containers, and careles sample handling introduce significant uncertainties in measurements, leading to unreliable results. Therefore, the prediction of target compounds concentration in samples have been out of the research focus up to now. Machine Learning algorithms can help to mitigate these uncertainties and open the applicability of CSRR sensors to quantification problems, where is necessary to determine the amount of a substance in a liquid (or solid) sample.**



This work presents a novel approach that takles this issue, combining a CSRR sensor with well stablised ML algorithms that enhances its quantification performance. For illustration purposes, a low-cost, benchtop CSRR-based system is proposed to predict ethanol concentration in water solutions. Ethanol samples from 10% to 96% concentration were prepared in commercial vials, generating 450 randomized measurements. Principal Component Analysis (PCA) was employed for data exploration, while a Partial Least Squares regression model (PLS), tuned with Leave-One-Group-Out Cross-Validation, was trained for ethanol concentration prediction. No feature extraction technique or noise reduction strategy was applied. Although this straightforward workflow is well known in the chemical sensing field, it has not been applied to data acquired with CSRR sensors.

The trained model achieved a Root Mean Square Error in Prediction (RMSEP) of 3.7%. Compared with 23.4% RMSEP when using univariate calibration at optimized frequencies, it presentes a prediction performance reduced by a factor of 6. No evidence of underfitting or overfitting was observed during test of the trained model. The low RMSEP achieved by the presented setup demonstrates the potential of CSRR-based sensors when combined with ML techniques for concentration prediction working in realistic, uncontrolled conditions. This pushes forward the applicability of CSRR sensors in the chemical analysis field, which might lead to benchtop, lowcost and relaible analysis devices for many laboratories.

***Index Terms*—CSRR Sensors, Machine Learning, Concentration Prediction, Chemical Analysis, RF sensors, Resonators.**

I. Introduction

COMPLEMENTARY Split Ring Resonators (CSRR) were introduced as metasurfaces and metamaterials [1], [2], and originally designed for microwave stopband filters, transmission lines, and antennas [3]–[8]. Their sharp filtering behavior and sensitivity to surface surroundings [4], [9], [10] soon positioned them as sensors [11], particularly for measuring complex permittivity of materials [12]–[17].

In recent years, CSRR sensors have gained prominence in chemical analysis, particularly for quantifying solute concentrations in solvents [18]–[20]. Notable applications include glucose detection in aqueous solutions [19], [21], alongside microfluidic device integration [22]–[25]. However, despite promising laboratory results, only a few these studies have demonstrated predicting capabilities of the studied substance concentration levels in the Sample under Test (SUT) when the study is performed under real laboratory conditions. This pushes away the usability of CSRR sensors in real laboratory environments, or when the samples are in commercial vials.

Accordingly, the objective of this contribution is threefold: First, to enhance the performance of CSRRs in predicting the concentration of substances in aqueous solutions. Sec-

J. Alonso-Valdesueiro is with University of Barcelona, Carrer Martí i Franquès,1. 08028, Barcelona, Spain (e-mail: javier.alonsov@ub.edu).

L. Fernández, is with University of Barcelona, Carrer Martí i Franquès,1. 08028, Barcelona, Spain (e-mail: lfernandez@ub.edu) and Institute for Bioengineering of Catalonia (IBEC), Barcelona Institute of Science and Technology, Baldiri Reixac 10-12 ,08028 Barcelona, Spain (email: lfernandez@ibecbarcelona.eu).

A. Gutiérrez-Gálvez, is with University of Barcelona, Carrer Martí i Franquès,1. 08028, Barcelona, Spain (e-mail: agutierrez@ub.edu).

S. Marco, is with University of Barcelona, Carrer Martí i Franquès,1. 08028, Barcelona, Spain (e-mail: santiago.marco@ub.edu) and Institute for Bioengineering of Catalonia (IBEC), Barcelona Institute of Science and Technology, Baldiri Reixac 10-12 ,08028 Barcelona, Spain (email: smarco@ibecbarcelona.eu).

ond, to develop a system that operates directly with standard vials—eliminating the need for dedicated microfluidic circuits—under uncontrolled environmental conditions (temperature and humidity), and using low-cost, commercially available bench top equipment. Third, to demonstrate that a simple and well-established workflow—namely, Principal Component Analysis (PCA) for data visualization and Partial Least Squares (PLS) regression for quantification—can be effectively applied to CSRR sensor data without requiring feature extraction or noise reduction techniques, while still achieving outstanding performance.

In a bench top, low-cost CSRR based system, the sources of uncertainty can be summarized as follows: (i) the CSRR sensor, which is sensitive to the position of the Sample Under Test (SUT), (ii) the SUT itself, which is sensitive to temperature and humidity variations, (iii) the Vector Network Analyzer (VNA), which, in the case of a low-cost device, introduces a considerable amount of uncertainty, and (iv) the measurement routine, which is sensitive to SUT preparation and the handling of vials. Furthermore, the presence of this uncertainty is not visible in small well selected datasets [26], [27]. Therefore, these factors can lead to significant variability in measurements, making it challenging to obtain reliable results using conventional regression techniques based on univariate calibration at optimized frequencies [11], [28], [29].

Machine Learning (ML) algorithms might solve this situation. They have been increasingly applied to CSRR sensors in the last five years [29]–[32], primarily to fit models linking Scattering Parameter (S-Parameter) features to concentration of solute, classifying different samples [21]. Recently, Multivariate Analysis (MVA) has been briefly introduced to this purpose [27]. Despite these advances, the full potential of ML algorithms to predict concentration of solutes in samples and to mitigate noise and variability in CSRR-based measurements under realistic conditions has not yet been fully explored. Furthermore, the literature lacks studies involving large datasets and appropriate sampling techniques, which are essential to build robust and generalizable models for prediction.

As a benchmark experiment for sensitivity characterization, the presence of ethanol in water solutions using CSRR sensors has been widely studied in the literature [32]–[35]. Ethanol is a common solvent in laboratories, and its quantification is relevant in various fields, including food safety, environmental monitoring, and clinical diagnostics [36]–[38]. Therefore, the prediction of ethanol concentration will be evaluated with the proposed CSRR bench top system as benchmark test of its performance.

Within this framework, this work proposes a portable, low-cost bench top CSRR-based system prone to uncertainties from the sensor, the Vector Network Analyzer (VNA), and measurement routine. Partial Least Squares regression (PLS), is employed to enhance system performance. This approach demonstrates that ML algorithms can enable CSRR-based sensors to be deployed as concentration prediction tools in chemical analysis outside strictly controlled environments, increasing the versatility and usability of the CSRR sensor concept.

Consequently, the text is organized as follows: Section II describes the measurement system, including the CSRR sensor, VNA, sampling methodology, and ML workflow. Section III presents data visualization, traditional CSRR performance modeling, PCA exploration, and PLS training results. Section IV compares the system's performance with conventional curve fitting methods. Finally, Section V summarizes the findings and the potential of CSRR-based systems in chemical analysis.

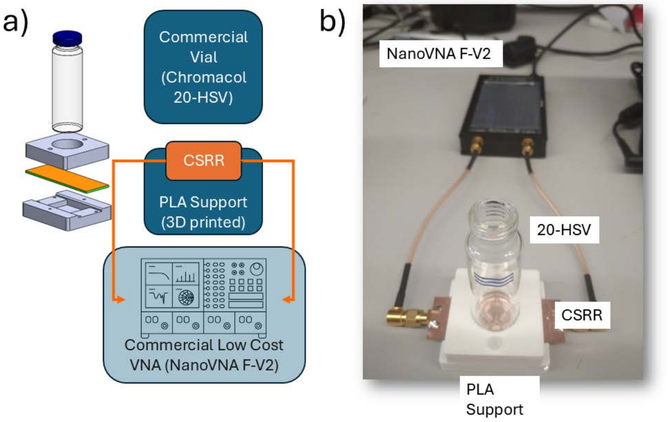

a)
Commercial Vial (Chromacol 20-HSV)
CSRR
PLA Support (3D printed)
Commercial Low Cost VNA (NanoVNA F-V2)
b)
NanoVNA F-V2
20-HSV
CSRR
PLA Support

Fig. 1. Proposed bench top CSRR system. (a) The commercial vial (Chromacol 20-HSV) containing the SUT is placed on the CSRR using a 3D-printed PLA support. The CSRR is connected to the low-cost VNA (NanoVNA F-V2) via two SMA cables. The NanoVNA operates either through its built-in graphical interface or via serial commands from a Python application running on a standard laptop. (b) Photograph of the real setup with a vial on top of the CSRR.

II. CSRR Benchtop System

A. System Block Diagram

A complete setup has been developed for the purposes of this work. Figure 1(a) presents the block diagram of the CSRR-based system. Describing the measurement system is essential because, as discussed in Section I, most studies optimize measurement setups to maximize sensor performance under controlled conditions.

In this contribution, the CSRR system is designed as a bench top tool for chemical analysis. The Sample Under Test (SUT) is placed in a commercial vial (Chromacol 20-HSV, Thermo Fisher Scientific, Inc. Waltham, MA, USA) and positioned on the sensor using a 3D-printed PLA holder. This holder ensures the vial is consistently centered over the CSRR resonant structure, minimizing positional variability during measurements.

Bench top equipment must be compact, portable, and affordable to fit in standard laboratory environments. The system measures the S_{21} parameter, which some studies achieve through custom electronics [19], [26], while others rely on expensive commercial VNAs [22]–[24]. Here, a low-cost commercial VNA (NanoVNA F-V2, Sysjoint Co., Ltd., Hangzhou, China) was selected due to its affordability, performance in the 1–3 GHz band, and compatibility with external computers via serial commands.

The VNA is controlled by a custom GUI developed in Python, running on a commercial laptop. The GUI facilitates

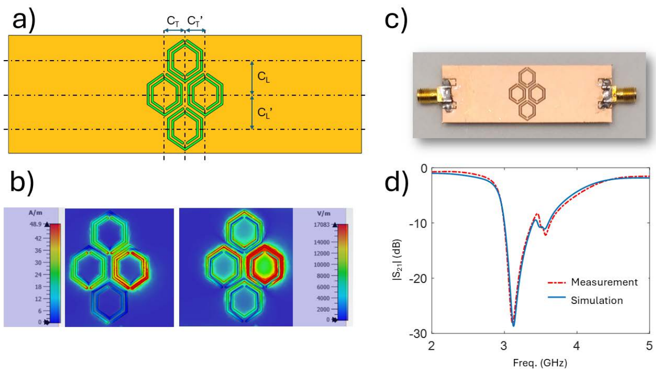


Fig. 2. Modified CSRR. (a) Structure of the measured CSRR with asymmetrical honeycomb placement on a $\mathbf{22 \times 66}$ mm PCB. The longitudinal honeycombs are positioned at $C_L = \mathbf{6.3}$ mm and $C_L' = \mathbf{6.4}$ mm from the PCB center, while the transverse honeycombs are placed at $C_T = \mathbf{3.7}$mm and $C_T' = \mathbf{3.8}$mm. (b) Simulated electric ($\vec{E}$V/m) and magnetic field ($\vec{H}$A/m) distributions on the honeycomb structure surface. (c) CSRR mounted with SMA connectors. (d) S_{21} comparison between simulations and measurements with the CSRR unloaded (Keysight E5071C-240).

data acquisition, organizes data into structured databases, and enables continuous measurement. It also includes a prediction module to integrate trained ML models for real-time concentration predictions from S-parameter measurements.

B. CSRR Sensor

One of the simplest and more affordable CSRR sensors was introduced by Omer et al. [26].

Figure 2(a) shows the structure of the CSRR used in this contribution, manufactured by Eurocircuits NV (Mechelen, Belgium). The design is based on the honeycomb CSRR, with modifications to change its spectral response. Specifically, the vertical and horizontal symmetry were modified by 100 μm with respect to the original design, introducing an additional resonance around 3.5 GHz close to the main resonance at 3.2 GHz. This asymmetry intentionally increases the system's sensitivity to positional variations of the SUT. Therefore, it enhances the sensitivity of the sensor to SUT electromagnetic anisotropies, which enriches the variability observed by using commercial vials. The coupling transmission line and the dimensions of each honeycomb are as defined in [26].

C. Sampling Methodology

The primary goal in any sampling methodology is to block potential unwanted influences on the measurements. To achieve this goal a heavily randomize strategy was followed in this contribution. This randomization mitigates undesired effects on the dataset, such as batch effects, repetition correlation, and instrumental error propagation [39].

In this study, samples were organized according to ethanol concentration in solution. The concentrations ranged from almost diluted (10% ethanol) to ethanol (96% purity),using 10% steps (10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, and 96%). Badge solutions were prepared with the specified concentrations and poured into 10 commercial vials, randomly selected from a pool of 100 vials (vials are 22.5 × 75 mm with 2 mm wall). In the vials, 1.2 mL of each badge solution were poured by micropipeting. Uncertainties in ethanol concentration at each vial ranged from ±0.45% for the most diluted preparation to ±3.6% for the samples prepared with 80% badge solution.

The 90 vials were labeled for randomization and identification, then stored together in a fridge at 5 °C for one day. The following day, the vials were removed from the fridge and left at 23 °C (room temperature) for an hour. Five Measurement Rounds (MR) were performed, with the order of vials randomized in each round. Each MR consisted of ten repetitions for each concentration, using different vials for each repetition, resulting in a total of 500 measurements (50 measurements for each concentration). Before each round, the unloaded response of the CSRR sensor (S21) was measured and used for baseline correction. Each measurement consisted of 201 data points recording $20\log\left(|S21|\right)$ from 1.6 GHz to 3 GHz using the NanoVNA F-V2. A large and comprehensive database with 500 repetitions of 10 samples of 9 different concentrations of ethanol (10% to 96%) resulted from this sampling methodology. The entire procedure is depicted in Figure 3. In this figure, the vial labelling and the randomization of the order on which the vials were measured is shown. The unloaded response of the CSRR measured before each MR also appears in the figure marked as baseline.

D. Machine Learning Workflow

As shown in figure 4, the workflow for training and test the machine learning (ML) models consists in three main steps: data preprocessing, data exploration, and model training/test. Initially, data is prepared and preprocessed to ensure consistency, quality and complexity control. Next, Principal Component Analysis (PCA) is applied to extract meaningful insights from the dataset and improve the visualization of the data. Finally, Partial Least Squares regression (PLS) is employed for model training and prediction. No feature extraction or noise reduction techniques were applied, as the aim of this

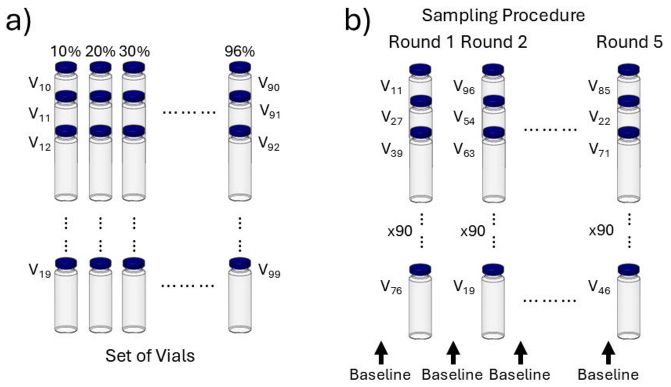


Fig. 3. Sampling Method. (a) 10 vials are assigned to each concentration randomly from the vial batch. The vials are poured with the corresponding ethanol concentration and labelled for identification. (b) Each Measurement Round (MR) consists of measures with the system described in Figure 1 the $\mathbf{20\log\left(|S_{21}|\right)}$ of the CSRR with a vial on top. The order of vials is randomized in each MR, and the unloaded response of the CSRR (baseline) is measured before each MR for baseline correction. In total $\mathbf{5}$ MR were carried out (450 measurements) with one baseline for each MR.

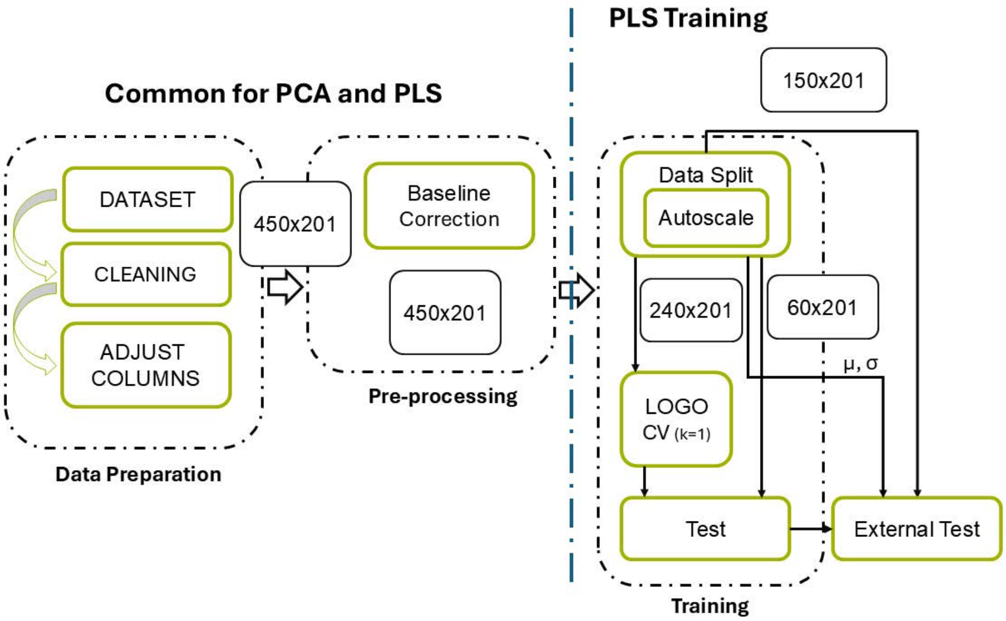


Fig. 4. Block diagram of the developed workflow for training the bench-top CSRR system. In the Data Preparation step, the database is cleaned by extracting the actual data (features) from the metadata, resulting in a $\mathbf{450 \times 201}$ feature matrix and a $\mathbf{5 \times 201}$ baseline matrix. In the Pre-Processing step, the feature matrix baseline is corrected using the baseline matrix, applied to each round separately. These two steps are common to both PCA and PLS training algorithms. In the Training step, the feature matrix is split into the training dataset ($\mathbf{300 \times 201}$ matrix) and the test dataset ($\mathbf{150 \times 201}$ matrix). In this step the train dataset is autoscaled and the mean and the standard deviation ($\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$) are stored for autoscaling the test dataset afterward. The train dataset consisted in the repetitions of $\mathbf{6}$ concentrations ($\mathbf{10\%}$, $\mathbf{20\%}$, $\mathbf{30\%}$, $\mathbf{50\%}$, $\mathbf{70\%}$, and $\mathbf{96\%}$) and the test dataset consisted in the repetitions of $\mathbf{3}$ concentrations ($\mathbf{40\%}$, $\mathbf{60\%}$, and $\mathbf{80\%}$). The training dataset is then introduced into a leave-one-block-out cross-validation (LOGO-CV) scheme, which produces an optimized PLS model. In the External Test step, the performance of the optimized model is evaluated by obtaining predictions from the test dataset.

work is not to implement the most advanced approaches in the machine learning field

1) Principal Component Analysis: Principal Component Analysis (PCA) was performed on the database generated as described in section II-C. The dataset was prepared and pre-processed as shown in Figure 4, and PCA was carried out using the Statistics and Machine Learning Toolbox (v12.5) in MATLAB (R2023a).

As shown in Figure 4, the dataset is cleaned by extracting metadata (labels, date of acquisition, etc.) and adjusting the number of features for each repetition, if necessary. The data matrix (450×201) is then pre-processed in three steps. First, A non-linear transformation of the measured S_{21} to dB is performed over the 201 point of each measurement $(20\log(|S_{21}|))$. Then the baseline of each measurement round is corrected using the measurement of the unloaded CSRR by simply subtraction. Second, each feature is auto-scaled by calculating the mean, μ_C, and standard deviation, σ_C, for each concentration in each MR.

The explained variance (EV) of the dataset was calculated for each PC, and the cumulative EV was plotted (see figure 7 (c)). The distribution of the repetitions in the reduced vector space was also plotted for the first two PCs (see Figure 7 (b)).

2) Partial Least Squares regression: As it shown in Figure 5, the complexity of the dataset is high and the number of repetitions per concentration of is lower than the features of each repetition. This makes the ${}^{sample}/_{dimensionality}$ ratio unfavorable for regular regression models. In these situations, Partial Least Squares (PLS) presents the most appropriate approach in order to avoid overfitting and deal with collinearity. The PLS algorithm compresses the data before regression by projecting the data into a lower-dimensional space. This helps to mitigate overfitting and improve model generalization [40]. Although PLS is a well-known technique in the chemical sensing field and the proposed workflow does not incorporate the latest advances in machine learning, its simplicity and effectiveness in this context make it the ideal candidate to demonstrate how ML algorithms can enhance the applicability of CSRR-based systems.

Before starting the PLS training, the dataset was pre-processed in the same way as for the PCA analysis (see Figure 4). The pre-processed data was then split into two subsets. Data from six concentrations was used for training and internal validation (training dataset, repetitions of 6 concentrations, 10%, 20%, 30%, 50%, 70%, and 96%), while the remaining data was set aside for external test (test dataset, repetitions of 3 concentrations, 40%, 60%, and 80%). Therefore, the training dataset consisted of a 300×201 feature matrix, and the test dataset of a 150×201 matrix. The training dataset was then autoscaled and the mean and standard deviation (μ and σ in Figure 4) were stored for later autoscaling of the test dataset.

A Leave-One-Group-Out cross-validation (LOGO-CV) scheme was used for model training [33]. This approach uses data from five concentrations for training and reserves data from one of the concentrations for internal validation. For each iteration, a set of predictions for each concentration is obtained and stored. At the end, for each level of complexity (number of Latent Variables, or LVs), the Root Mean Square Error in Cross Validation (RMSECV) is calculated. The optimal

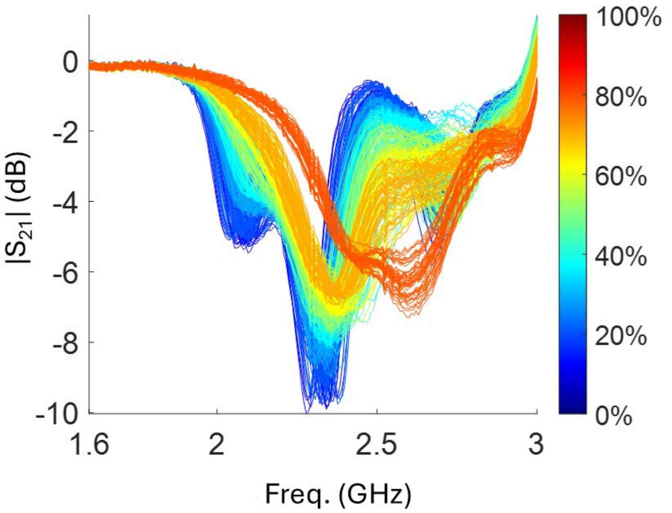


Fig. 5. $\mathbf{20log}\,(|\dot{S}_{21}|)$ of each measurement in the generated dataset. Ethanol concentrations in clean water range from $\mathbf{10\%}$ to $\mathbf{96\%}$. Ten random commercial vials were selected for each concentration from a pool of $\mathbf{100}$. Five rounds of measurements were performed, where in each round, the $\mathbf{20log}\,(|\dot{S}_{21}|)$ of the CSRR was recorded by placing a randomly selected vial on top of the sensor and continuing through the set of $\mathbf{90}$ vials. Before each round, the $\mathbf{20log}\,(|\dot{S}_{21}|)$ of the unloaded CSRR was measured for baseline correction of the data.

number of LVs is selected based on the minimum RMSECV value. The PLS model is then trained using the entire training dataset with the selected number of LVs. Once the PLS model is trained, external validation is performed by introducing the test dataset in the model and making predictions for the concentrations. In this case, predictions for 40%, 60%, and 80% concentrations were estimated using the resulting model.

III. CSRR System Performance

In this section, we present the results from the measurements, characterization of the frequency shift, and machine learning analysis of the CSRR system. The performance of the system is evaluated in terms of its ability to quantify the concentration of ethanol in water. We begin by discussing the traditional characterization of the system, followed by the PCA analysis of the dataset, and conclude with the evaluation of the PLS model's performance.

A. Acquired Database

Figure 5 shows the $20\log(|\dot{S}_{21}|)$ of each measurement in the dataset used for ML model training and testing. Several features are evident upon visual inspection. First, a wide dispersion of the curves is observed for each concentration. Notably, the dispersion of the deepest resonance seems to decrease as the ethanol concentration increases.

Features around 2.1 GHz and 2.5 GHz show better discrimination between samples. Additionally, for 20% ethanol, two repetitions exhibit an outlier behavior, with $20\log(|\dot{S}_{21}|)$ values deviating from the trend observed in the rest of the dataset.

Furthermore, outliers were identified for certain concentrations (10%, 20%, and 40%) in rounds 1, 2, 3, and 5. The repetitions corresponding to 40% ethanol showed one outlier in each round, likely due to a vial characteristic that differed significantly from the rest of the pool. However, the outliers for 10% and 20% ethanol appeared in different rounds, suggesting inconsistent handling of the vials during measurements.

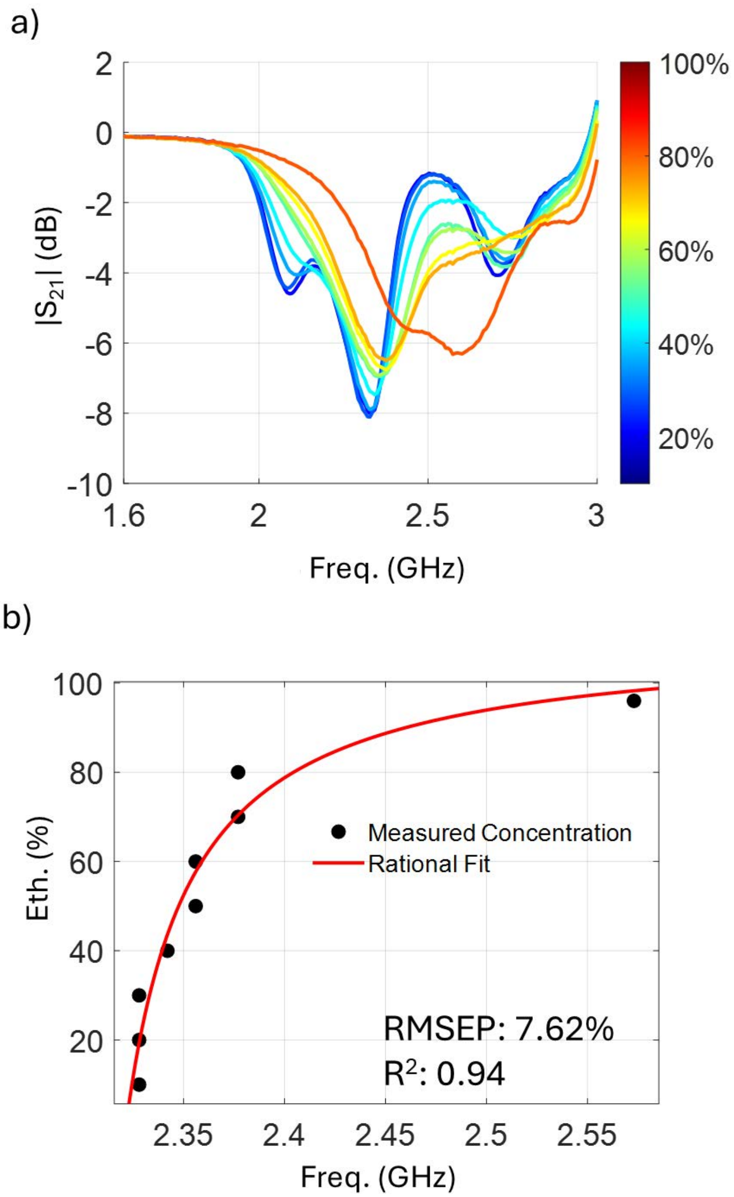


Fig. 6. Traditional characterization of the CSRR system when measuring concentrations of ethanol diluted in clean water and poured in commercial vials. a) Averaged $\mathbf{20log}\,(|\dot{S}_{21}|)$ response of the generated database. b) Rational model fitting for the frequency of the minimum observed in the averaged $\mathbf{20log}\,(|\dot{S}_{21}|)$ when concentrations of the training test are considered ($\mathbf{10}$, $\mathbf{20}$, $\mathbf{30}$, $\mathbf{50}$, $\mathbf{70}$, and $\mathbf{96}$ $\mathbf{\%}$)

B. Frequency Shift Characterization

Following studies presented in the literature, Figure 6(a) shows the averaged $20\log(|\dot{S}_{21}|)$ for each concentration. This plot illustrates how the shape of the S_{21} parameter evolves with concentration, indicating a non-linear response of the system to ethanol. In order to compare fair comparisson, dataset spllit presented in II-C (data from concentrations 10%, 20%, 30%, 50%, 70%, and 96%) was applied to the dataset. By evaluating the frequency shift of the most prominent resonance in the training dataset [32], a rational fit of the deepest resonance yields the following expression:

$$C_{\%}(\nu) = \frac{p1 \cdot \nu + p2}{\nu + q} \qquad (1)$$

where the frequency, ν, is in GHz, $p1 = 111.8\ {}^{\%}/_{GHz^2}$, $p2 = -259.2\ {}^{\%}/_{GHz}$, and $q = -2.3$ GHz. The R^2 of the fit is 0.96 and the RMSECV is $8.72\ \%$. The fitting was performed using the Curve Fitting Toolbox (v3.9) from MATLAB (R2023a). This model is shown in Figure 6 (b).

A set of predictions were obtained using the rational model of Equation 1. As test dataset, the data of the 3 concentrations not used for fitting (40%, 60%, and 80%) were selected in order to fairly compare this methodology with the ML workflow proposed in section II-D. For predictions, the frequency in GHz of the minimum observed in each $20\log\left(|\dot{S}_{21}|\right)$ for each repetition was introduced into the rational model. The predicted concentration was compared with the actual concentration of the sample. The RMSEP calculated in this way was about 23.4%.

C. Principal Component Analysis

As shown in Figure 7 (a), 95% of the Explained Variance (EV) of the dataset is contained within the first three Principal Components (PCs). Most of this variance is concentrated in the first PC (73%), while the remaining components account for less than 15% of the EV. While the first PC might be associated with the ethanol concentration in the solution, the second and third PCs are likely related to the variability introduced by the commercial vials and the measurement process.

Figures 7 (b), shows the distribution of the repetitions in the reduced vector space of the extracted for the two most relevant PCs. The complexity of the dataset is evident in the distribution of the repetitions in the reduced vector space. Each repetition is represented by a different marker, with the color indicating the ethanol concentration.

D. Partial Least Squares Predictive Model

Once the workflow presented in Figure 4 is completed, several intermediate results provide insights into the performance of the obtained model. Figure 8 (a) shows the evolution of the Root Mean Square Error in Cross-Validation (RMSECV) with respect to the number of components used in the PLS during the LOGO-CV scheme iterations. At the end, the optimal model was obtained with 7 LVs.

Additionally, the most important features in the training dataset were identified using the Variable Importance in Projection (VIP) scores for each feature [41]. The calculated scores provide information on the average importance of each feature in the projection of the repetitions onto the reduced vector space handled by the PLS.

This information can be used to reduce the feature matrix and focus on the most relevant parts of the $20\log\left(|\dot{S}_{21}|\right)$ response. There are several ways to define thresholds for feature importance [15]; a common approach is to select those with a mean VIP score greater than 1. However, since the objective of this study is not to assess the impact of feature selection on model performance, no such selection was performed. This aspect is reserved for future work.

1) Performance in Cross Validation: As mentioned in section III-D, after the LOGO-CV, the PLS model with complexity of 7 LVs provided the best performance. A model with this level of complexity produced predictions for the 20%, 30%, 50%, 70%, and 96%. The RMSE in cross-validation (RMSECV) was around 2.63%, with an R^2 of 0.98. The model introduced a linear bias of -0.9% on the predicted concentration, and the calculated Limits of Agreement (LoA), using the Bland-Altman method [42] for the 95% Confidence Interval (CI), resulted in a $\sim \pm 6\%$ maximum deviation from linear prediction.

Figure 8 (b) shows the VIP scores for the training dataset

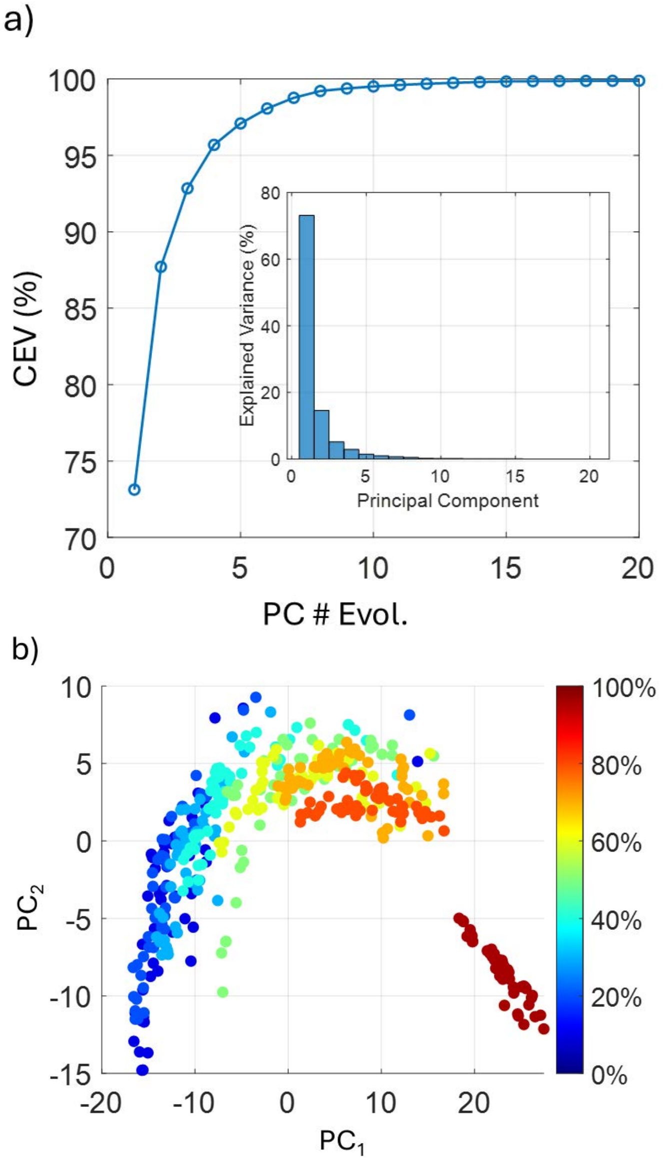


Fig. 7. Principal Component Analysis (PCA) of the dataset generated as described in section II-C. a) Cumulative Explained Variance (CEV) of the dataset (in %) with respect to the considered PCs and Explained Variance distribution over the set of PCs considered. b) Distribution of each repetition in the reduced vectorial space when representing the second Principal Component (PC$_2$) with respect to the first PC (PC$_1$).

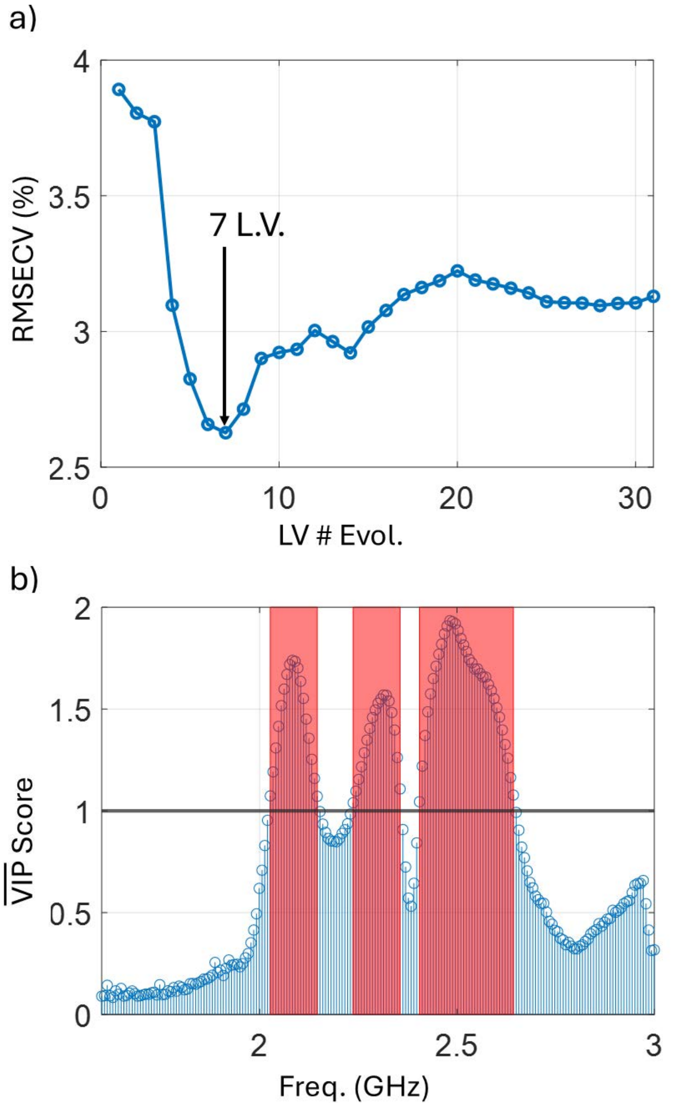


Fig. 8. Statistics of the Partial Least Squares Regressor (PLS) trained with the generated dataset: a) Evolution of the Mean Squared Error (MSE) with the number of Latent Variables (LVs) during the Leave-One-Block-Out Cross-Validation (LOGO-CV). b) Mean Variable Importance in Projection (VIP) scores for each feature from 1.6GHz to 3GHz, with scores greater than 1 highlighted in pink (🟥).

after optimizing the PLS to 7 LVs. In this figure, pink areas (🟥) highlight scores greater than 1. It can be observed that the most important parts of the $20\log\left(|S_{21}|\right)$ correspond to data around 2.1 GHz, 2.45 GHz, and 2.5 GHz. These frequency ranges align with the most significant differences between concentrations observed in Figure 5.

2) Performance in Test: After internal validation of the model, a external validation was performed. The purpose of this external validation was to evaluate the performance of the PLS model when facing repetitions of samples which the model had never seen before. For this purpose, the total database was divided into training dataset and test dataset as explained in section II-D. The test dataset contained repetitions of 40%, 60% and 80% samples.

Figure 9 shows the predictions obtained for the test dataset. These predictions produced a RMSE in Prediction (RMSEP) about 3.7 %, close to the RMSECV, with a R^2 coefficient of 0.98. The LoAs were calculated in the same way presented in previous section, resulting in a maximum deviation of $\sim \pm 7\%$ with a 95% of Confident Interval. The PLS models also showed a -1.5% bias when predicting the concentrations of the test dataset.

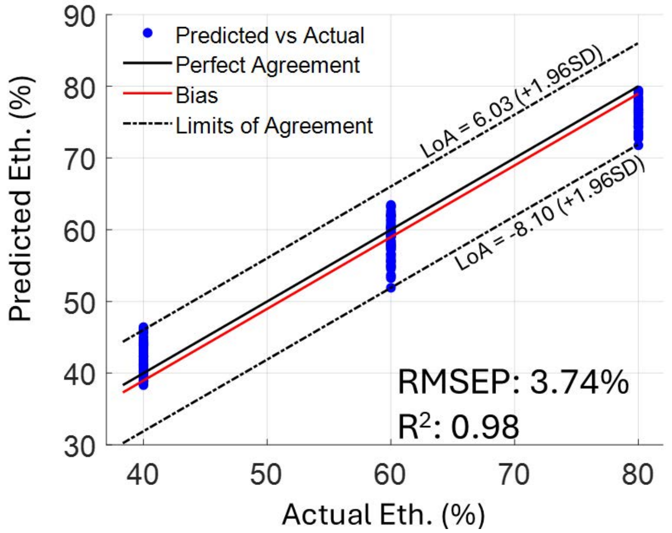


Fig. 9. Performance of the PLS at test. The plot includes Predicted versus Actual ethanol concentration in $\%$, the ideal perfect concentration, the bias introduced by the model and the upper and lower LoAs at $\mathbf{95\%}$ CI.

IV. CSRR System Evaluation and Discussion

In this study, we demonstrate the impact of applying a well established set of ML algorithms in measuring liquid concentrations with a CSRR-based bench-top system. When properly applied, even the most simply combination of data visualization (through PCA) and multivariate linear regression (through PLS modelling), not only reveal the properties of the generated datasets but also enhance the prediction performance of low-cost bench-top systems for predicting concentrations of substances under test (SUTs) when measurement conditions are very close to a real application.

For a bench-top system like the presented in section II the measurement procedure and the commercial vials used regularly in laboratories, introducing a considerable the amount of the observable variability across repetitions. In the literature, this variability is managed, either enhancing the hardware of the system, which might be impossible in many of the cases due to the increase in hardware cost, or designing ad-hoc solutions that reduce variability, as it happens with the vials. Table I shows a comparison of the presented system with other CSRR-based systems used for ethanol concentration determination. The table highlights that most of the systems do not report any error in the test, which is likely due to the fact that they are not tested with a large dataset, or they are not tested at all. With one recent exception where a convolutional neural network was trained with a large dataset, the rest of the systems are based on the frequency shift of the S$_{21}$ parameter.

In that sense, the presented system is the first one to report a large dataset of measurements and to apply a multivariate analysis to the data. Also, the novelty of the presented system

TABLE I
COMPARISON WITH DIFFERENT CSRR-BASED SYSTEM USED FOR ETHANOL CONCENTRATION DETERMINATION.

Ref	Sample Container	Model Parameters	Reported Error in Test
[20]	Plexiglas Sheets	Δf_0 $\Delta\|S_{21}\|$	No
[23]	ABS 3D printed	Δf_0 $\Delta\|S_{21}\|$	No
[24]	Microfluidic Channels	Convolutional Neural Network	Yes
[28]	PDMS microfluidic Channels	Δf_0 $\Delta\|S_{21}\|$	No
[6]	PDMS microfluidic Channels	No Reported	No
[43]	Capillary Tube	Δf_0 $\Delta\|S_{21}\|$	No

is based on the fact that it does not require any ad-hoc solution to reduce variability, such as microfluidic [24].

Furthermore, the inefficiency of using one-dimensional analysis, such as frequency shift, has been demonstrated in this study when a more realistic situation than the presented in the literature is considered. Section III-B highlights how a large and robust dataset (acquired as explained in section III-A) reveals significant variability and large standard deviations across repetitions, making it clear that simple curve fitting is unlikely to accurately predict the concentration of the SUT. In particular, with the dataset generated in section III-A, the combination of the presented system, measurement uncertainties, and commercial vials resulted in a curve-fitting model based on the frequency shift of the S_{21} parameter with ethanol concentration, yielding an RMSE of 23.4%. This low performance shows that multivariate analysis is essential to improve both accuracy and robustness.

In particular, the proposed ML workflow, with a preliminary exploration of the dataset and a full training of a PLS model have demonstrated their boosting capabilities reducing 6 times the error when a singular prediction is performed. The price to pay is the complexity of the approach at its beginning, which is compensated by the robustness of the model and the reduction of the error in the prediction.

For the benchmark problem of ethanol concentration diluted in clean water, the PCA shows the variability of the dataset concentrated around 4 Principal Components (PCs), as it is shown in Figure 7. The first PC is clearly related to the ethanol concentration in the solution. This is confirmed in Figure 7 (b). The figure also informs about the spread in the second PC. This spread can be attributed to variability introduced by the commercial vials and the measurement process itself.

This variability suggests the complex evolution of the $20\log\left(|\dot{S}_{21}|\right)$ of the system with the concentration of ethanol in solutions under the measurement conditions proposed here. This complexity is not captured by the traditional curve fitting model, as it is shown in section III-B and discussed above.

In accordance with this, the golden standard in multivariate analysis, the PLS model, was trained with the dataset generated in this study. The training evolution of the PLS model showed that 7 LVs minimized the RMSECV (see figure 8 (a)) which is not far away from the 4 PCs obtained in the PCA of section II-D.1. Moreover, the VIP scores analysis confirmed the multivariate nature of the problem (see figure 8 (b)) showing that different features of the spectrum have relevance during the training process.

In fact, the trained PLS model using the workflow presented in section II-D, shows a better performance than the univariate model. In particular, a reduction by a factor of ~ 6 in the RMSE at test (RMSEP $= 3.74$ %). Also, the LoAs are stable over the range of concentrations considered in the test dataset and similar at training and test. This behavior in training and test shows the robustness of the PLS model compared with the quantification models presented in the literature.

V. CONCLUSIONS

As a result of the presented study, it has been demonstrated that the performance of low-cost, bench top devices, ready to be used for predicting concentrations in solutions for laboratory settings, can be significantly enhanced by applying a very simple Machine Learning Workflow. Due to the popularity of commercial vials for sample handling and storage this largely improves the usability of the device for quantitative purposes. This demonstration was carried out using a representative example of the benchmark problem of quantifying ethanol concentration diluted in clean water. The results show that an affordable sampling campaign, coupled with the use of ML algorithms (specifically PCA and PLS) can improve the performance of traditional curve fitting by a factor of 6. Additionally, no feature extraction methods were applied in order to maintain a simple and straightforward workflow. As discussed in Section III-D.1, the impact of feature extraction may be explored in future studies to further improve model performance.

PCA analysis was used to explore the dataset and visualize the complexity of the problem. This approach enables linking the dataset's complexity to the physical properties and design of the CSRR sensor, as illustrated in Figure 8 (b). The VIP scores obtained during the training of the PLS model identify the most relevant features in the $20\log\left(|\dot{S}_{21}|\right)$ spectra, which correspond to the sensor's physical characteristics and design.

To maintain simplicity in the workflow, PLS regression was selected as the quantification algorithm. It is well-suited for scenarios with a low sample-to-feature ratio and is widely used in the chemical sensing field. Although other ML algorithms, such as Lasso or Random Forest, may achieve better performance in capturing dataset non-linearity [44]. PLS regression offers superior stability, interpretability, and robustness against overfitting [45], [46].

The presented results overcome the limitations associated with univariate analysis of the S_{21} parameter in CSRR-based systems in predicting problems, particularly when applied to real-world scenarios such as chemistry laboratories. They also open up the opportunity to utilize CSRR-based systems as bench top devices in chemical analysis and biomedical applications, advancing their development as valuable tools for analysis, monitoring, screening, and diagnosis.

At the same time, the study of the VIP scores during the training of the PLS model, identifies the most relevant features in the $20\log\left(\dot{|}S_{21}|\right)$ spectra. This information can be linked with the physical properties of the CSRR sensor and its design which could lead to improvements in the sensor design. Therefore, the link of between microwave sensor design and ML algorithms can also open new ways to improve the performance of CSRR-based sensors in the future.

It is important to note that there is still room for improvement in the portability and robustness analysis of the CSRR-based system presented in this work. Future studies could focus on evaluating the system's performance under varying environmental conditions, such as temperature and humidity changes, to ensure its reliability in real-world applications. These investigations are facilitated by the low-cost and portability of the proposed system. Its compact size and affordability make it suitable for deployment in diverse settings, including field applications and resource-limited environments.

In this work, the focus has been on predicting ethanol concentration in clean water, a benchmark problem in the field. However, the proposed workflow and system can be readily adapted to other substances and mixtures, making it a versatile tool for various applications in chemical analysis and biomedical fields. Future investigations might also address questions related to the vial–sensor interface and the influence of vial anisotropy on system sensitivity.

Acknowledgment

This work was financially supported by the Institut de Bioenginyeria de Catalunya (IBEC). IBEC is a member of the CERCA Programme from Generalitat de Catalunya. The authors would also like to acknowledge the Departament d'Universitats, recerca i Societat de la Informació of Generalitat de Catalunya (expedient 2021 SGR 01393) and the he Comissionat per a Universitats i Recerca del DIUE de la Generalitat de Catalunya; and the European Social Fund (ESF) for its financial support. Authors want also to thank Professor Antonio Pardo and Gema Guedes, Ph.D. for their collaboration during experiments in the laboratory and their understanding on the increasing entropy inside it.

References

[1] F. Falcone, T. Lopetegi, J. Baena, R. Marques, F. Martin, and M. Sorolla, "Effective negative-/spl epsiv/ stopband microstrip lines based on complementary split ring resonators," *IEEE Microwave and Wireless Components Letters*, vol. 14, no. 6, pp. 280–282, 2004.

[2] J. Baena, J. Bonache, F. Martin, R. Sillero, F. Falcone, T. Lopetegi, M. Laso, J. Garcia-Garcia, I. Gil, M. Portillo, and M. Sorolla, "Equivalent-circuit models for split-ring resonators and complementary split-ring resonators coupled to planar transmission lines," *IEEE Transactions on Microwave Theory and Techniques*, vol. 53, no. 4, pp. 1451–1461, 2005.

[3] J. Garcia-Garcia, F. Martin, F. Falcone, J. Bonache, J. Baena, I. Gil, E. Amat, T. Lopetegi, M. Laso, J. Iturmendi, M. Sorolla, and R. Marques, "Microwave filters with improved stopband based on sub-wavelength resonators," *IEEE Transactions on Microwave Theory and Techniques*, vol. 53, no. 6, pp. 1997–2006, 2005.

[4] J. Bonache, I. Gil, J. Garcia-Garcia, and F. Martin, "Novel microstrip bandpass filters based on complementary split-ring resonators," *IEEE Transactions on Microwave Theory and Techniques*, vol. 54, no. 1, pp. 265–271, 2006.

[5] M. Mandal, P. Mondal, S. Sanyal, and A. Chakrabarty, "Low insertion-loss, sharp-rejection and compact microstrip low-pass filters," *IEEE Microwave and Wireless Components Letters*, vol. 16, no. 11, pp. 600–602, 2006.

[6] M. Gil, J. Bonache, J. Selga, J. Garcia-Garcia, and F. Martin, "Broad-band resonant-type metamaterial transmission lines," *IEEE Microwave and Wireless Components Letters*, vol. 17, no. 2, pp. 97–99, 2007.

[7] A. Velez, J. Bonache, and F. Martin, "Varactor-loaded complementary split ring resonators (vlcsrr) and their application to tunable metamaterial transmission lines," *IEEE Microwave and Wireless Components Letters*, vol. 18, no. 1, pp. 28–30, 2008.

[8] H. Zhang, Y.-Q. Li, X. Chen, Y.-Q. Fu, and N.-C. Yuan, "Design of circular/dual-frequency linear polarization antennas based on the anisotropic complementary split ring resonator," *IEEE Transactions on Antennas and Propagation*, vol. 57, no. 10, pp. 3352–3355, 2009.

[9] T. Grzegorczyk, C. Moss, J. Lu, X. Chen, J. Pacheco, and J. A. Kong, "Properties of left-handed metamaterials: transmission, backward phase, negative refraction, and focusing," *IEEE Transactions on Microwave Theory and Techniques*, vol. 53, no. 9, pp. 2956–2967, 2005.

[10] I. Stevanovic, P. Crespo-Valero, K. Blagovic, F. Bongard, and J. Mosig, "Integral-equation analysis of 3-d metallic objects arranged in 2-d lattices using the ewald transformation," *IEEE Transactions on Microwave Theory and Techniques*, vol. 54, no. 10, pp. 3688–3697, 2006.

[11] M. S. Boybay and O. M. Ramahi, "Material characterization using complementary split-ring resonators," *IEEE Transactions on Instrumentation and Measurement*, vol. 61, no. 11, pp. 3039–3046, 2012.

[12] K. Song and P. Mazumder, "Design of highly selective metamaterials for sensing platforms," *IEEE Sensors Journal*, vol. 13, no. 9, pp. 3377–3385, 2013.

[13] C. S. Lee and C. L. Yang, "Thickness and permittivity measurement in multi-layered dielectric structures using complementary split-ring resonators," *IEEE Sensors Journal*, vol. 14, no. 3, pp. 695–700, 2014.

[14] C. S. Lee and C. L. Yang, "Complementary split-ring resonators for measuring dielectric constants and loss tangents," *IEEE Microwave and Wireless Components Letters*, vol. 24, no. 8, pp. 563–565, 2014.

[15] M. A. H. Ansari, A. K. Jha, and M. J. Akhtar, "Design and application of the csrr-based planar sensor for noninvasive measurement of complex permittivity," *IEEE Sensors Journal*, vol. 15, no. 12, pp. 7181–7189, 2015.

[16] A. Standaert, M. Rousstia, S. Sinaga, and P. Reynaert, "Permittivity measurements in millimeter range of ptfe foams," *IEEE Microwave and Wireless Components Letters*, vol. 27, no. 8, pp. 766–768, 2017.

[17] L. Su, J. Mata-Contreras, P. Vélez, and F. Martín, "Splitter/combiner microstrip sections loaded with pairs of complementary split ring resonators (csrrs): Modeling and optimization for differential sensing applications," *IEEE Transactions on Microwave Theory and Techniques*, vol. 64, no. 12, pp. 4362–4370, 2016.

[18] P. Vélez, K. Grenier, J. Mata-Contreras, D. Dubuc, and F. Martín, "Highly-sensitive microwave sensors based on open complementary split ring resonators (ocsrrs) for dielectric characterization and solute concentration measurement in liquids," *IEEE Access*, vol. 6, pp. 48 324–48 338, 2018.

[19] A. E. Omer, G. Shaker, S. Safavi-Naeini, K. Ngo, R. M. Shubair, G. Alquié, F. Deshours, and H. Kokabi, "Multiple-cell microfluidic dielectric resonator for liquid sensing applications," *IEEE Sensors Journal*, vol. 21, no. 5, pp. 6094–6104, 2021.

[20] K. Zhang, R. K. Amineh, Z. Dong, and D. Nadler, "Microwave sensing of water quality," *IEEE Access*, vol. 7, pp. 69 481–69 493, 2019.

[21] M. Martinic, M. Mertens, D. Schreurs, B. Nauwelaers, G. A. E. Vandenbosch, and T. Markovic, "Highly sensitive impedance-matched microwave dielectric sensor for glucose concentration measurements," *IEEE Sensors Journal*, pp. 1–1, 2025.

[22] S. Patel, S. Das, D. Mitra, S. Sarkar, and C. Koley, "Biological liquid monitoring using microwave resonator," in *2022 URSI Regional Conference on Radio Science (USRI-RCRS)*, 2022, pp. 1–7.

[23] S. Jiang, G. Liu, M. Wang, Y. Wu, and J. Zhou, "Design of high-sensitivity microfluidic sensor based on csrr with interdigital structure," *IEEE Sensors Journal*, vol. 23, no. 16, pp. 17 901–17 909, 2023.

[24] S. Liu, P. Cheong, C. Yang, Y. Ye, and W.-W. Choi, "Microfluidic sensor for simultaneous liquid classification and concentration detection," *IEEE Microwave and Wireless Technology Letters*, vol. 34, no. 3, pp. 358–361, 2024.

[25] Q. Zhang, Z. Kou, J. Yang, Y. Li, Z. Yin, and G. Deng, "Highly sensitive and robust metasurface-inspired microfluidic sensor for oil detection," *IEEE Sensors Journal*, vol. 24, no. 24, pp. 40 847–40 854, 2024.

[26] A. E. Omer, G. Shaker, and S. Safavi-Naeini, "Portable radar-driven microwave sensor for intermittent glucose levels monitoring," *IEEE Sensors Letters*, vol. 4, no. 5, pp. 1–4, 2020.

[27] S. Trovarello, O. Afif, A. Di Florio Di Renzo, D. Masotti, M. Tartagni, and A. Costanzo, "A non-invasive, machine learning assisted skin-

hydration microwave sensor," in *2024 54th European Microwave Conference (EuMC)*, 2024, pp. 932–935.

[28] A. Ebrahimi, W. Withayachumnankul, S. Al-Sarawi, and D. Abbott, "High-sensitivity metamaterial-inspired sensor for microfluidic dielectric characterization," *IEEE Sensors Journal*, vol. 14, no. 5, pp. 1345–1351, 2014.

[29] L. Harrsion, M. Ravan, D. Tandel, K. Zhang, T. Patel, and R. K. Amineh, "Material identification using a microwave sensor array and machine learning," *Electronics*, vol. 9, no. 2, 2020.

[30] D. Prakash and N. Gupta, "High-sensitivity grooved csrr-based sensor for liquid chemical characterization," *IEEE Sensors Journal*, vol. 22, no. 19, pp. 18 463–18 470, 2022.

[31] N. Kazemi and P. Musilek, "Enhancing microwave sensor performance with ultrahigh q features using cyclegan," *IEEE Transactions on Microwave Theory and Techniques*, vol. 70, no. 12, pp. 5369–5382, 2022.

[32] M. Abdolrazzaghi, N. Kazemi, V. Nayyeri, and F. Martin, "Ai-assisted ultra-high-sensitivity/resolution active-coupled csrr-based sensor with embedded selectivity," *Sensors*, vol. 23, no. 13, 2023.

[33] P. Filzmoser, B. Liebmann, and K. Varmuza, "Repeated double cross validation," *Journal of Chemometrics*, vol. 23, no. 4, pp. 160–171, 2009.

[34] A. Salim and S. Lim, "Complementary split-ring resonator-loaded microfluidic ethanol chemical sensor," *Sensors*, vol. 16, no. 11, 2016.

[35] Y. Beria, G. S. Das, A. Buragohain, and B. B. Chamuah, "Highly sensitive miniaturized octagonal ds-csrr sensor for permittivity measurement of liquid samples," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–9, 2023.

[36] N. H. Holford, "Clinical pharmacokinetics of ethanol," *Clin. Pharmacokinet.*, vol. 13, no. 5, pp. 273–292, nov 1987.

[37] D. Zuba, W. Piekoszewski, J. Pach, L. Winnik, and A. Parczewski, "Concentration of ethanol and other volatile compounds in the blood of acutely poisoned alcoholics," *Alcohol*, vol. 26, no. 1, pp. 17–22, jan 2002.

[38] Z. Zhang, Z. Li, K. Wei, Z. Cao, Z. Zhu, and R. Chen, "Sweat as a source of non-invasive biomarkers for clinical diagnosis: An overview," *Talanta*, vol. 273, p. 125865, 2024.

[39] C. Wu and M. E. Thompson, *Sampling theory and practice*, 2020th ed., ser. ICSA Book Series in Statistics. Cham, Switzerland: Springer Nature, may 2020.

[40] S. Wold, M. Sjöström, and L. Eriksson, "Pls-regression: a basic tool of chemometrics," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 109–130, 2001, pLS Methods.

[41] I. Chong and J. C.H., "Performance of some variable selection methods when multicollinearity is present," *Chemometrics and Intelligent Laboratory Systems*, vol. 78, no. 1, pp. 103–112, 2005.

[42] J. Martin Bland and D. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *The Lancet*, vol. 327, no. 8476, pp. 307–310, 1986, originally published as Volume 1, Issue 8476.

[43] E. L. Chuma, Y. Iano, G. Fontgalland, and L. L. Bravo Roger, "Microwave sensor for liquid dielectric characterization based on metamaterial complementary split ring resonator," *IEEE Sensors Journal*, vol. 18, no. 24, pp. 9978–9983, 2018.

[44] S. Tateishi, H. Matsui, and S. Konishi, "Nonlinear regression modeling via the lasso-type regularization," *Journal of Statistical Planning and Inference*, vol. 140, no. 5, pp. 1125–1134, 2010.

[45] B. H. Menze, B. M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, and F. A. Hamprecht, "A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data," *BMC Bioinformatics*, vol. 10, no. 1, p. 213, Jul 2009.

[46] W. Song, Z. Hou, M. S. Afgan, W. Gu, H. Wang, J. Cui, Z. Wang, and Y. Wang, "Validated ensemble variable selection of laser-induced breakdown spectroscopy data for coal property analysis," *J. Anal. At. Spectrom.*, vol. 36, pp. 111–119, 2021.

Dr. Javier Alonso-Valdesueiro Was born in Madrid, Spain, in 1980. He earned a Telecommunications Engineering degree from the University of Alcalá, Madrid, in 2006, and a Ph.D. from the Polytechnic University of Catalonia, Barcelona, in 2011. He was awarded with Marie Skłodowska-Curie Individual Fellowship in 2017. Over the past years, he has advanced his career as an RF and instrumentation engineer in both private companies and public research centres and recently the University of Barcelona. His current research focuses on radio-frequency (RF) devices for MRI, instrumentation electronics for gas sensing applications, and advancements in RF sensor technologies.

Dr. Luis Fernández is a physicist (BSc in Physics, 2006) and electronic engineer (BSc in Electronic Engineering, 2011), both degrees obtained from the University of Barcelona, where he also earned his PhD in Physics in 2016. His research focuses on data analysis and signal processing for chemical sensing systems, particularly artificial olfaction technologies such as electronic noses and analytical instruments like gas chromatography–ion mobility spectrometry (GC-IMS). Dr. Fernández is currently an associate professor at the University of Barcelona, where he continues his work on computational tools for sensor systems and biomedical applications.

Dr. Agustín Gutiérrez-Gálvez Agustin Gutierrez-Galvez received the B.S. degree in physics and in electrical engineering from the Universitat de Barcelona, Spain, in 1995 and 2000, respectively. He received the Ph.D. degree in computer science from Texas A&M University, College Station, in 2005. His research interests are computational models of the olfactory system and pattern recognition applied to odour detection with sensor arrays. He is currently an Associate Professor in the Department of Electronics and Biomedical Engineering, Universitat de Barcelona.

Dr. Santiago Marco-Colás Dr. Santiago Marco is Full Professor at the department of Electronics and Biomedical Engineering in the University of Barcelona (UB). He obtained his degree in and his PhD in Physics from the UB in 1988 and 1993 respectively. In 1994 he was a post-doc at the University of Rome "Tor Vergata", working on Data Processing for Artificial Olfaction. In 2004 he was in a sabbatical leave at the EADS-Corporate Research in Munich working in Ion Mobility Spectrometry. His current research interests are focused on data analysis and signal processing for chemical sensing systems, particularly electronic noses and analytical instruments such as gas chromatography–ion mobility spectrometry (GC-IMS).