Tree-driven optimization inference strategy for highly resource-demanding data acquisition: Application to chemical sensors

Marino Pavone[1], Roberto Alesii[1], Stefano Tennina[2], Giordano Pola[1]

***Abstract*— Radio-frequency-based chemical sensors are a promising approach for noninvasive, low-cost identification of substances across a wide range of applications, including medical and biological contexts. However, their practical deployment is often limited by high costs or long acquisition times required to collect high-resolution spectral features, which adversely affect energy consumption, responsiveness, and the ability to capture fast-evolving chemical phenomena in healthcare applications. In this work, we address this limitation by proposing an acquisition-aware classification framework that exploits the intrinsic sequential structure of tree-based learning algorithms.**

Building on a previously introduced wireless chemical-sensing platform based on spectral analysis, we investigate the combined use of feature selection and decision-tree-based inference to reduce both data acquisition and inference costs. The proposed approach models the sensing process as a query-driven system in which radio-frequency features are acquired conditionally as a classification tree is traversed. As a result, the number of acquired features required for inference is bounded by the depth of the tree rather than by the dimensionality of the full feature space.

On a preliminary dataset spanning 3–6 GHz (1000 candidate features) and multiple substances, our results indicate that accurate classification can be achieved with a small subset of features, yielding a substantial reduction of expected acquisition/inference time in balanced operating conditions ($> 30\%$), with further gains when class priors are available. Furthermore, the proposed inference strategy allows additional savings when prior knowledge about class occurrence probabilities is available. These findings highlight the potential of acquisition-aware, tree-based models as an effective solution for real-world radio-frequency chemical sensing systems operating under strict time and energy constraints.

I. Introduction

Smart sensing systems empowered by machine learning are increasingly adopted in domains requiring adaptability and robustness to variability, including healthcare and biology [1]. In this context, chemical sensors are essential tools for substance identification across applications ranging from environmental monitoring to medical and biological analysis [2]. Nevertheless, many current solutions remain constrained by high costs, limited flexibility, poor scalability across multiple substances, and the necessity for complex or invasive setups. These limitations drive ongoing research toward precise, trainable, low-cost, and noninvasive detection platforms.

[1]M. Pavone, R. Alesii, and G. Pola are with the Department of Information Engineering, Computer Science, and Mathematics, Center of Excellence for Research DEWS, University of L'Aquila, L'Aquila, Italy. (e-mail: marino.pavone,roberto.alesii,giordano.pola@univaq.it).

[2]S. Tennina is with WEST Aquila Srl, L'Aquila AQ, 67100 Italy. (stefano.tennina@westaquila.com).

Healthcare is shifting toward decentralized, patient-centric paradigms, where screening and monitoring increasingly occur outside centralized laboratories. This trend necessitates rapid, low-power tools for analyzing biological fluids (e.g., saliva, sweat, urine) in point-of-care and home-care settings. Saliva is particularly advantageous due to its accessibility and its ability to reflect both physiological and pathological conditions [3]. The widespread adoption of rapid tests during the COVID-19 pandemic further demonstrated the value of convenient, low-cost, and accurate screening solutions [4]. In these contexts, system performance is constrained not only by classification accuracy but also by latency, energy consumption, and usability.

A wireless Radio-Frequency (RF) chemical-sensing platform utilizing a microstrip sensor coupled with a Software-Defined Radio (SDR) was recently introduced [5], demonstrating promising multi-class discrimination by correlating frequency-domain responses with the properties of tested substances. However, several challenges related to real-world deployment persist, with the feature-acquisition phase as the primary bottleneck. High-resolution scans over wide frequency ranges result in longer measurement times, which negatively affect system responsiveness and energy consumption and limit the ability to capture rapidly evolving chemical phenomena [6]. This limitation is especially pronounced for unstable substances and biological samples, whose properties may change during acquisition.

In this paper, we address these limitations via an acquisition-aware analytical and algorithmic model that reduces both acquisition time and inference complexity. We combine embedded feature selection with decision tree classification to enable accurate identification without acquiring or evaluating the full RF feature set. By exploiting the structure of decision trees, inference can be performed using only a small subset of features, with acquisition cost bounded by tree depth, thereby permitting efficient and adaptive sensing strategies. Beyond reducing identification time, the proposed methodology provides practical indications for dataset expansion and, more broadly, for guiding sensor design.

Our main contributions are summarized as follows:

- **Acquisition-aware RF/SDR formulation.** We model feature acquisition as a cost-dominant component and cast decision-tree inference as a sequential process whose reached tree depth bounds required measurements.
- **Query-driven sensing pipeline.** We implement a tree-guided strategy where only the features along the visited decision path are acquired, enabling early decisions

when sufficient information is available.

- **Feature selection and validation.** On an SDR dataset with 14 substances in the 3–6 GHz band (1000 candidate features, 80–20 split), we show that accurate classification can be achieved using a small feature subset (on the order of 10), with a substantial reduction in expected acquisition/inference time under balanced conditions and further savings when class priors are available.

II. Related Work

A. Feature Selection in Spectral and RF-Based Sensing

Feature selection is a fundamental step in designing machine learning pipelines for chemical sensing systems, particularly when dealing with high-dimensional spectral or radio-frequency measurements [7]. In many sensing applications, a large number of features are acquired to capture subtle variations in the physical response of the system under test. However, the relevance of these features is often highly heterogeneous, and a significant portion may be redundant or weakly informative for classification.

This observation motivates a multi-level analysis in which feature selection is performed at different downsampling resolutions of the same dataset. At each resolution level, the relevance of individual features may change, leading to different optimal subsets. Classical dimensionality reduction techniques, such as Principal Component Analysis (PCA), Autoencoders, or alternative feature extraction techniques specific to frequency-domain data [8], cannot be employed, as they usually require the full feature vector.

Other feature selection strategies can be broadly categorized into filter, wrapper, and embedded methods [9]. Filter methods rely on statistical criteria such as variance, entropy, mutual information, or correlation measures, offering low computational cost but limited interaction with the final classifier. Wrapper methods evaluate feature subsets by repeatedly training and testing a classifier, often achieving better performance at the cost of high computational complexity and poor scalability.

In acquisition-constrained sensing systems, however, the effectiveness of feature selection is inseparable from the inference mechanism itself. This consideration naturally shifts attention to embedded feature selection approaches, in which the selection process is intrinsically tied to the classification algorithm.

B. Embedded Feature Selection and Acquisition-Aware Inference

Embedded feature selection methods integrate feature-relevance assessment directly into model training, optimizing feature use alongside classification performance. Tree-based learning algorithms are among the most prominent examples of this paradigm. In decision trees, features are selected at each split based on criteria such as information gain or impurity reduction, resulting in a hierarchical structure where only a subset of features is evaluated along each inference path. This property is particularly advantageous in sensing systems where feature acquisition is costly [10]. Unlike conventional classifiers that require the complete feature vector at inference time, decision trees enable conditional and sequential feature acquisition. The number of features required to reach a decision is bounded by the tree's depth, which directly translates into a bound on acquisition time. Moreover, the selected features retain their physical meaning, allowing a direct association between decision logic and specific frequency regions.

Even if ensemble methods based on decision trees, such as Random Forests, Extremely Randomized Trees, and Gradient boosting, further improve robustness and generalization (by combining multiple trees' inference), at inference time, different trees may require different features, potentially increasing the total number of acquired features [10], [11].

A more explicit formulation of the acquisition problem is provided by cost-sensitive decision trees [12], where each feature is associated with an acquisition cost. This approach is suitable for applications in which each feature has a different acquisition cost.

Rule-based learners share conceptual similarities with decision trees, as each rule typically involves a small number of features and can be evaluated sequentially, with early-exit conditions. While offering high interpretability and acquisition efficiency, these methods often struggle to match the performance of tree-based algorithms on complex, high-dimensional datasets [13].

Finally, approaches such as classifier cascades [14] and reinforcement-learning-based acquisition policies [15] can combine sequential inference with explicit class-priority assignment. However, they introduce substantial additional complexity: cascade architectures require independent training and threshold calibration for each binary stage, while reinforcement-learning formulations demand significantly larger datasets to reliably estimate an optimal acquisition policy — a condition that is not met by the preliminary dataset considered in this work.

In summary, while numerous machine learning approaches have been proposed for chemical sensing, only a limited subset explicitly supports partial, conditional, and acquisition-aware feature usage at inference time. Given that feature acquisition time is the dominant operational cost in the radio-frequency chemical sensor considered, embedded feature selection methods based on tree-based models provide a particularly suitable foundation for efficient and scalable real-world deployments.

III. Methodology

The objective of the proposed methodology is to reduce both data acquisition time and inference complexity by exploiting the intrinsic sequential structure of tree-based classifiers. In this section, we first formalize the relationship between decision-tree inference and conditional feature acquisition, and then describe the sensing system as a query-driven process controlled by the classifier.

A. Sequential Feature Access in Decision Trees

A decision tree classifier performs inference by traversing a path from the root node to a leaf node, where each internal node evaluates a decision rule based on a single feature. Let $d \in \mathbb{N}^+$ denote the depth of a given leaf, defined as the number of decision nodes encountered along the path from the root to that leaf. Since each decision node requires the evaluation of exactly one feature, the number of features required to classify a sample associated with that leaf is upper bounded by d.

Formally, let $\mathcal{F} = \{f_1, f_2, \ldots, f_M\}$ be the full set of available features, with $M \gg 1$, and let $\mathcal{P}_\ell \subseteq \mathcal{F}$ be the subset of features evaluated along the path leading to leaf $\ell \in \mathbb{N}^+$. Then:

$$|\mathcal{P}_\ell| \leq d_\ell$$

where d_ℓ is the depth of leaf ℓ. Since d_ℓ is typically much smaller than M, especially when shallow or regularized trees are employed, this property enables classification without requiring the acquisition of the full feature vector.

This behavior contrasts sharply with conventional classifiers, such as distance- or kernel-based models, which require all features to be available before inference. In the considered sensing scenario, where acquiring each feature incurs a non-negligible time cost, this distinction becomes critical.

B. Sensor as a Query-Driven Acquisition System

Based on the above observation, the sensing system can be interpreted as a query-driven process in which the decision tree dynamically controls feature acquisition during inference. Rather than acquiring all radio-frequency features upfront, the sensor operates sequentially, acquiring only the features requested by the classifier at each traversal step.

Specifically, starting from the root node of the tree, the system acquires the feature associated with the current decision node, evaluates the corresponding test, and selects the next branch accordingly. Each decision not only refines the path toward a final classification, but also implicitly defines a subset of candidate classes consistent with the observed measurement, while excluding all classes associated with the opposite branch of the tree. As a consequence, even intermediate decisions provide partial but meaningful information about the substance under analysis.

In several monitoring, control, or surveillance scenarios, this progressive exclusion of classes may already be sufficient to trigger an action or raise an alert, without requiring the inference process to reach a final leaf node. In such cases, the system can intentionally terminate the acquisition early, further reducing sensing time and energy consumption. This process is repeated until a leaf node is reached, or until the available partial classification information is deemed sufficient for the application at hand. As a result, the total acquisition time for a given sample is proportional to the depth of the traversed path rather than to the dimensionality of the full feature space.

Let t_{acq} denote the time required by the sensing electronics to acquire a single feature. The total acquisition time for classifying a sample associated with leaf ℓ can then be expressed as:

$$T_\ell = |\mathcal{P}_\ell| \cdot t_{acq}$$

which satisfies:

$$T_\ell \leq d_\ell \cdot t_{acq}$$

This formulation highlights how the inference process directly determines the sensing effort, enabling adaptive acquisition strategies that depend on the structure of the trained classifier.

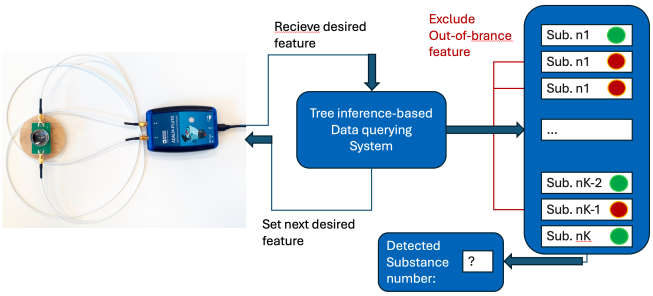

Fig. 1. Conceptual representation of the proposed query-driven sensing system. The decision tree sequentially selects which radio-frequency features to acquire, and classification is performed once a leaf node is reached.

C. System Overview and Conceptual Architecture

Figure 1 illustrates the conceptual architecture of the proposed sensing and inference pipeline. The decision tree serves as a controller, sequentially querying the sensing hardware for specific radio-frequency features. Each acquired feature is immediately evaluated to determine the next acquisition step until a classification decision is reached.

This tight coupling between inference logic and feature acquisition enables substantial reductions in both sensing time and energy consumption while preserving classification accuracy. Moreover, since the selected features retain their physical meaning, the resulting decision paths provide insight into the most informative frequency regions for substance discrimination.

The described methodology serves as the basis for the subsequent analysis, in which feature selection and classification performance are evaluated across different downsampling levels of the acquired radio-frequency data.

IV. Experimental Results

A. Dataset Description

The dataset used in this work was acquired following the measurement procedure described in [5]. The dataset consists of measurements of simple chemical solutions, as listed in Table I.

To minimize temporal variability in the dataset, only solutions exhibiting long-term chemical stability were considered, thereby excluding reactive species, acidic media, and supersaturated or gas-evolving systems For each substance, multiple acquisitions were performed to account for measurement variability.

Id.	Substance	Quantity
sub 1	Empty (no container)	–
sub 2	Empty container	–
sub 3	Distilled Water	0.5 ml
sub 4	Distilled Water	2.0 ml
sub 5	Distilled Water	1.5 ml
sub 6	Distilled Water	1.0 ml
sub 7	Distilled Water + Salt 0.8%	2.0 ml
sub 8	Distilled Water + Salt 1.6%	2.0 ml
sub 9	Distilled Water + Salt 2.4%	2.0 ml
sub 10	Distilled Water + Sugar 0.8%	2.0 ml
sub 11	Distilled Water + Sugar 1.6%	2.0 ml
sub 12	Distilled Water + Sugar 2.4%	2.0 ml
sub 13	Vinegar	2.0 ml
sub 14	Corn Oil	2.0 ml

TABLE I

CHEMICAL SOLUTION CLASSES IN VARIABLE QUANTITIES AND CONCENTRATIONS.

Each acquisition is represented by a vector of 1000 radio-frequency features, corresponding to the sensor's excitation over the frequency range from 3 GHz to 6 GHz with a uniform frequency step of 3 MHz. The complete dataset has been organized into CSV format and processed within a Python-based analysis environment.

B. Feature Selection and Spectral Analysis

The first step of the analysis focused on identifying the most informative features within the high-dimensional frequency-domain representation. To this end, an embedded feature selection method based on XGBoost was employed to provide qualitative information on the relative importance of each feature in the classification task.

Given the very dense character of the original dataset, several downsampling configurations were evaluated to assess whether reducing the number of acquired features would significantly change the distribution of informative content across the frequency spectrum. Specifically, the dataset was progressively downsampled by simply discarding samples at a given frequency step, while preserving the same overall frequency span.

Figures 2 and 3 show the distribution of relevant features across the frequency band for different downsampling levels, as identified respectively by CART-based and XGBoost-based analyses. The XGBoost results indicate that informative content tends to maintain consistent frequency localization across multiple downsampling factors. Only when the information content is severely degraded by excessive feature removal the relevance distribution shift significantly among the remaining features. In the CART-based one instead, the downsampling action significantly affects feature importance, probably due to the XGboost capability to produce more generalizable models. This behavior suggests the presence of specific frequency regions that are particularly descriptive for substance discrimination within the proposed dataset, supporting the feasibility of targeted frequency acquisition strategies.

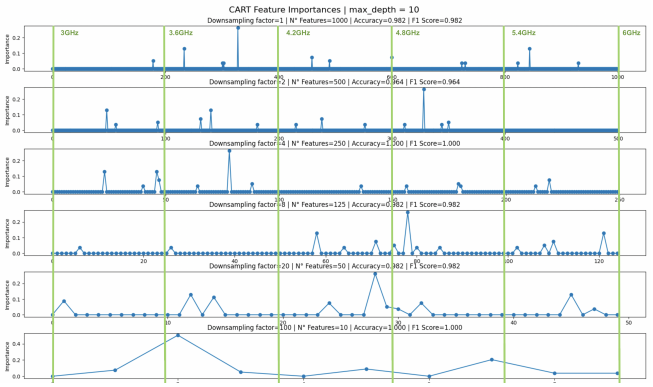

Fig. 2. Distribution of relevant features across the frequency spectrum identified using CART-based analysis at different downsampling levels.

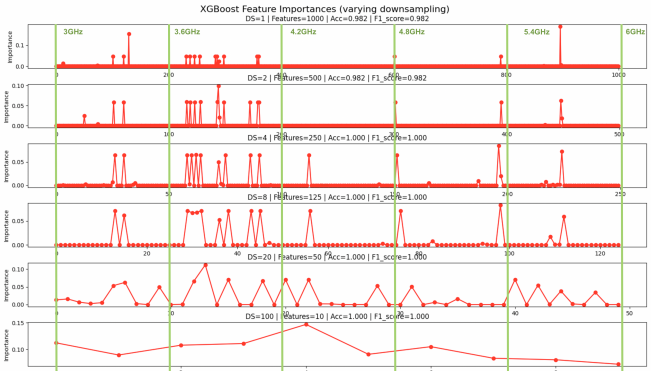

Fig. 3. Distribution of relevant features across the frequency spectrum identified using XGBoost embedded feature selection at different downsampling levels.

C. Tree-Based Inference and Acquisition Efficiency

Based on the XGBoost feature selection analysis, the top 10 most informative features were selected to train a classification tree. Figure 5 depicts the resulting decision tree.

The trained tree achieves 98.18% accuracy and 98.21% F1-score with a maximum depth of 8. In the worst case, inference requires acquisition of at most $8 \cdot t_{acq}$, where t_{acq} is the time to acquire a single feature. However, due to the sequential and conditional nature of decision-tree inference, the average number of features acquired can be significantly lower, as discussed in the previous section.

The effective inference time depends on the balance of the test dataset. In a fully balanced scenario, despite minor variations due to the specific structure learned during training, simulation results show an average reduction in inference time of approximately 40% compared to full-feature acquisition. For generalization, the method was also validated on an additional chemical dataset presented in [2].

D. Impact of Tree Depth and Downsampling

A further analysis was conducted to investigate the impact of two key parameters on classification performance and inference efficiency: the downsampling factor and the maximum depth of the decision tree. The maximum depth parameter directly limits the worst-case inference time, while the

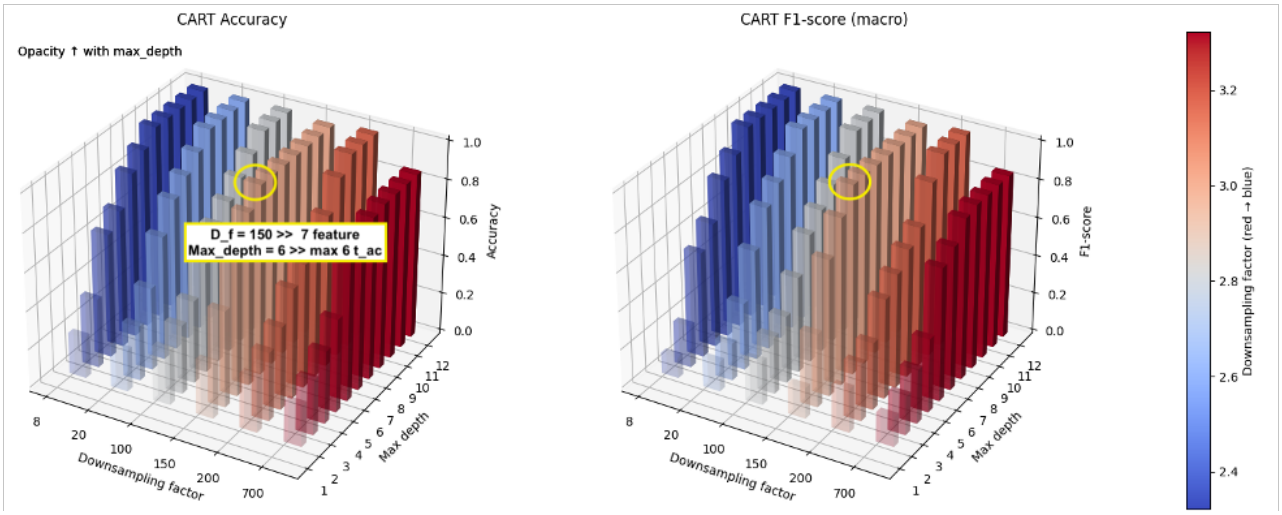

Fig. 4. Classification performance (accuracy and F1-score) as a function of downsampling factor and maximum tree depth.

a

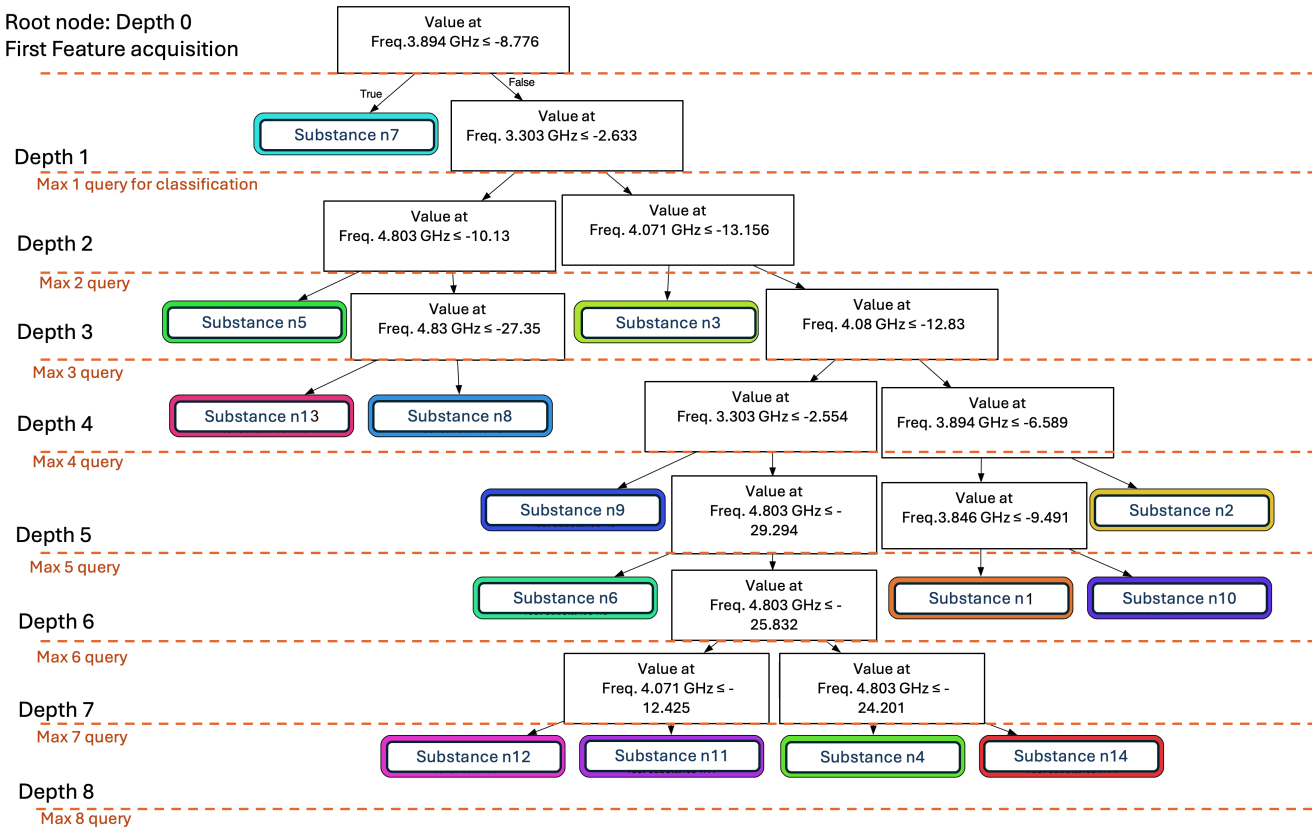

Fig. 5. Decision tree trained on the top 10 selected features. Note that the feature corresponding to Freq. 4.803 appears 4 times.

downsampling factor controls the density of the frequency-domain representation.

Figure 4 reports the evolution of classification performance metrics, including accuracy and F1-score, as a function of both the downsampling factor and the maximum tree depth. The results highlight a clear trade-off between acquisition efficiency and classification performance.

This analysis is useful for identifying, within the proposed methodology, the optimal trade-off between the number of features supplied to the classifier and the maximum inference time, without significantly sacrificing performance. In the considered experimental setting, it is observed that a small subset of approximately 10 features is sufficient to correctly classify (98% acc.) all 14 substances in an 80–20 training/test split, while maintaining a substantially reduced acquisition and inference time.

V. CONCLUSIONS AND FUTURE WORK

This work investigated the operational constraints of a radio-frequency chemical sensing technology based on software-defined radio measurements, with a specific focus on scenarios where *feature acquisition time* dominates the end-to-end latency and energy budget. Building on the sensing principles and preliminary discriminability evidence reported in the previous work [5], we target two practical objectives: (i) reducing the number of RF features that must be acquired to reach a reliable decision, and (ii) identifying frequency regions that maximize information content while avoiding redundant measurements.

On a preliminary dataset comprising 14 substances (denoted as sub1, ..., sub14), 20 repetitions per substance, and 1000 candidate RF features per acquisition, an embedded feature selection stage based on XGBoost provided qualitative evidence that only a small subset of features drives separability. Using the top-ranked features, we then implemented a CART decision policy that supports *sequential* inference: at run time, the sensor acquires only the features along the visited decision path, so the number of acquired features is upper-bounded by the reached depth. Under balanced operating conditions, this strategy yields a substantial reduction in expected acquisition/inference time (e.g.,$> 30\%$ in our preliminary tests, with larger gains when class priors are available and the decision policy can prioritize more likely substances).

From a healthcare perspective, the main value of the proposed framework is that it turns an RF/SDR chemical sensor into a time- and energy-aware decision system, aligning substance identification with point-of-care and home-care constraints. Concretely, we (*i*) identified that separability is driven by a small subset of spectral features, enabling targeted frequency regions of interest, and (*ii*) implemented a tree-based, query-driven inference strategy in which features are acquired only along the visited decision path, bounding the required acquisitions by the tree depth rather than by the full feature dimensionality. On our preliminary dataset (14 substances, 3–6 GHz band, 1000 candidate features), this yields a substantial reduction in expected acquisition/inference time under balanced conditions (on the order of tens of percent) while maintaining accurate classification, thereby supporting time-critical sensing of biological or otherwise time-evolving samples.

Future work will address the current limitations of not considering class priority or of not ensuring faster inference for specific classes. This will be explored using the Cascades approach or by integrating sequential models with reinforcement learning. Expanding the dataset in size and diversity will enable a more refined selection of frequency ranges and targeted excitation of high-information carrier frequencies. Additionally, future studies will explicitly address time-evolving samples, such as biological or highly reactive substances, where prolonged acquisition windows may degrade label consistency and insufficient temporal sampling may impair classifier generalizability. The results achieved, specifically the identified frequency bands and the spacing between spectral features, will play a key role in guiding decisions about expanding the dataset to include additional liquids and in informing the design of new, smaller sensors.

ACKNOWLEDGMENT

The work was partially supported by the Italian Ministry of Economic Development, EADAPTIVE Project N. F/310261/01-02/X56.

REFERENCES

[1] F. Kriebel, S. Rehman, M. A. Hanif, F. Khalid, and M. Shafique, "Robustness for smart cyber physical systems and internet-of-things: From adaptive robustness methods to reliability and security for machine learning," in *2018 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, 2018, pp. 581–586.

[2] M. Pavone, N. Epicoco, G. Pola, and A. Manno, "Modelling sensors degradation for water quality monitoring: A data-driven approach," in *2024 IEEE 20th International Conference on Automation Science and Engineering (CASE)*, 2024, pp. 1696–1701.

[3] G. Stojanovic, T. Kojic, M. Simic, A. Jovanovic-Galovic, B. Pavlovic, A. Zurutuza, L. Anzi, and R. Sordan, "Rapid selective detection of ascorbic acid using graphene-based microfluidic platform," *IEEE Sensors Journal*, vol. 21, pp. 16 744–16 753, Aug. 2021.

[4] J. La, N. Fillmore, and N. Do, "Factors associated with the speed and scope of diffusion of covid-19 therapeutics in a nationwide healthcare setting: a mixed-methods investigation," *Health Research Policy and Systems*, vol. 20, p. 134, 2022.

[5] C. D. Natale, R. Alesii, and S. Tennina, "Liquid identification through sdr: System design and performance analysis," in *2024 IEEE Workshop on Complexity in Engineering (COMPENG)*, 2024, pp. 1–5.

[6] I. Birs, C. Muresan, I. Nascu, and C. Ionescu, "A survey of recent advances in fractional order control for time delay systems," *IEEE Access*, vol. 7, pp. 30 951–30 965, 2019.

[7] K. Xu, M. Qin, F. Sun, Y. Wang, Y.-K. Chen, and F. Ren, "Learning in the frequency domain," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[8] M. Ragnoli, M. Pavone, N. Epicoco, G. Pola, E. De Santis, G. Barile, and V. Stornelli, "A condition and fault prevention monitoring system for industrial computer numerical control machinery," *IEEE Access*, vol. 12, pp. 20 919–20 930, 2024.

[9] M. Pavone, N. Epicoco, F. Magliocca, and G. Pola, "A generalized approach for feature selection in water quality monitoring," in *2023 31st Mediterranean Conference on Control and Automation (MED)*, 2023.

[10] F. Nan, J. Wang, and V. Saligrama, "Feature-budgeted random forest," in *Proceedings of the 32nd International Conference on Machine Learning*, F. Bach and D. Blei, Eds., vol. 37. PMLR, 07–09 Jul 2015, pp. 1983–1991.

[11] K. Trapeznikov and V. Saligrama, "Sequential decision making and staged classification," in *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013, pp. 899–907. [Online]. Available: https://proceedings.mlr.press/v28/trapeznikov13.html

[12] S. Lomax and S. Vadera, "A survey of cost-sensitive decision tree induction algorithms," *ACM Computing Surveys (CSUR)*, vol. 45, no. 2, pp. 1–35, 2013.

[13] J. Fürnkranz and T. Kliegr, "A brief overview of rule learning," *Foundations of Intelligent Systems. ISMIS 2015. Lecture Notes in Computer Science*, vol. 9384, pp. 54–69, 2015.

[14] X. Zhang, J. Zou, K. He, and J. Sun, "Accelerating very deep convolutional networks for classification and detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 10, pp. 1943–1955, 2016.

[15] P. T. . L. V. Janisch, J., "Classification with costly features in hierarchical deep sets," *Mach Learn*, vol. 113, pp. 4487–4522, May 2022.