Interpretable Water Leakage Detection Using Federated Prototype-Based Learning

Diego Perdigão Sousa
Department of Teleinformatics Engineering
Federal University of Ceara
Fortaleza, Brazil
0000-0001-6408-2760

Polycarpo Souza Neto
Department of Teleinformatics Engineering
Federal University of Ceara
Fortaleza, Brazil
0000-0001-5057-1942

José Mairton Barros da Silva Jr.
Division of Computer Systems
Uppsala University
Uppsala, Sweden
0000-0002-4503-4242

Charles Casimiro Cavalcante
Department of Teleinformatics Engineering
Federal University of Ceara
Fortaleza, Brazil
0000-0002-4198-4064

Carlo Fischione
Division of Network and Systems Engineering
KTH Royal Institute of Technology
Stockholm, Sweden
0000-0001-9810-3478

***Abstract*—This work presents an interpretable and privacy-aware solution for leakage detection in water distribution networks using federated prototype-based learning. Real data from pumping stations in Stockholm expose a non-independent and identically distributed data scenario, where each client reflects distinct operational conditions. Despite data heterogeneity, the model achieves consistently high performance. Interpretability is achieved via Voronoi-based prototypes, while privacy emerges from client-specific decision boundaries. The approach shows that combining federated learning and prototype-based models enables scalable, explainable, and secure anomaly detection in critical infrastructure.**

***Index Terms*—federated learning, leakage detection, prototype-based models, water distribution network.**

I. Introduction

The use of pipelines and distribution networks for transporting water and other fluids has seen substantial technological advancements over the past century, significantly improving the reliability of this mode of transport [1]. Despite the well-established benefits of pressurized pipeline systems [2], maintaining their secure and sustainable operation remains a persistent challenge due to the frequent occurrence of leaks and bursts. One of the most effective strategies for mitigating these issues is the early detection of anomalies, which plays a crucial role in minimizing resource loss.

According to the report by the Institute for Research and Economic Strategy of Ceará (IPECE) [3], Brazil's water loss rate is approximately 40%. In contrast, the government's target, under the current legal framework for basic sanitation, is to reduce this rate to 25% by 2034. Accordingly to IPECE [3], the State of Ceará notably reports a significantly lower water loss rate, currently around 27%.

The authors in [4] classified state-of-the-art strategies for water leakage management into pre-operational and post-operational approaches. According to [4], pre-operational approaches include hydraulic model evaluation and structural interventions, while post-operational approaches rely on leak assessment, prevention, and detection.

Within the context of leak detection, as reviewed in [5], several methodologies have been proposed in the literature leveraging machine learning techniques, including random forests, neural networks, and convolutional neural networks (CNNs). Furthermore, the study presented in [6] employed hydraulic simulations of real-world water distribution networks (WDNs) using the environmental protection agency network (EPANET) software to design data-driven solutions grounded in machine learning. In addition, the investigation carried out in [7] offers a comprehensive overview of the results achieved through the emerging benchmarking framework titled the battle of leakage detection and isolation methods (BattLeDIM). Lastly, the authors in [8] introduced the open-source Python package EPyT-Flow, which facilitates access to widely adopted benchmark datasets and exposes low-level functionalities of the EPANET simulation environment.

Despite the benefits obtained when using deep networks, these techniques commonly suffer from their black box characteristics [9]. This creates a critical trade-off between performance and interpretability, especially in high-stakes infrastructure monitoring. Recent advances have proposed interpretable alternatives, such as counterfactual-based event fingerprints, which improve transparency and user trust in automated diagnostics [10]. In line with this direction, our approach integrates prototype-based learning into a federated framework,

The work of D. P. S. was partially supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Proc. No. 172998/2023-9, and in part by Instituto Gridtech de Pesquisa e Desenvolvimento. The work of J. M. B. da S. Jr. was partially supported by the European Union's Horizon Europe Research and Innovation Programme under the Marie Skłodowska-Curie Project FLASH (Grant Agreement No. 101067652), and in part by the Ericsson Research Foundation, the Hans Werthén Foundation, and the Mistra InfraMaint Programme. The work of C. C. C. was partially supported by the CNPq, Proc. No. 308512/2023-5, and in part by the Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico (FUNCAP), Grant No. PS-0186-00103.01.00/21. The work of C. F. was partially supported by the SSF project SAICOM and the VR project WIDCOMP.

enabling local interpretability and privacy-preserving modeling while addressing the need for clarity and confidentiality in distributed real-world water systems.

Motivated by the aforementioned research challenges, we proposed an efficient and low-complexity distributed modeling approach to identify potential leakages in real-world WDNs in municipal areas, while ensuring the privacy of hydraulic data. The methodology relies on prototype-based models (PBMs), which balance interpretability and computational efficiency, and incorporates fundamental concepts from federated learning (FL) to enable a privacy-preserving solution capable of handling data from spatially distributed sensing devices.

In [11], we compared the performance of multiple centralized PBMs and obtained promising classification outcomes in highly resource-constrained scenarios, thereby validating the hypothesis that a limited set of prototypes can effectively capture the variability of observed water pressure measurements.

Subsequently, in the study [12], we extended the previous analysis by incorporating water flow measurements and conducting a performance comparison between the federated and centralized versions of the proposed model.

Then, in [13], we presented findings related to a proposed device-oriented learning rate designed to address the challenges posed by non-independent and identically distributed (non-IID) data in federated settings.

Finally, we further investigate the interpretability of the proposed federated PBM (FPBM) for detecting water leakages in WDNs, highlighting its potential in practical anomaly detection scenarios. Hence, this study constitutes a substantial extension of our prior research, advancing the understanding of how prototype-based learning (PBL) and FL can be jointly leveraged to enable secure and interpretable distributed learning in critical infrastructure environments.

The remainder of this paper is organized as follows. Section II introduces the fundamental concepts of PBL and FL. Section III presents the FPBM modeling approach. Section IV describes the case study adopted for evaluation. Section V discusses the experimental results, and Section VI concludes the paper with final remarks.

II. Background Knowledge

A. Prototype-based Learning

PBL refers to a class of machine learning methods that construct models by leveraging representative examples, known as prototypes. In [14], PBMs are also referred to as competitive learning algorithms in the context of artificial neural networks. The term "competitive" pertains to the fundamental mechanism of PBMs, wherein the reference units (called prototypes) compete to represent different partitions of the input data.

The PBL framework encompasses both supervised learning models, such as the learning vector quantization (LVQ) family of algorithms [15], and unsupervised models, such as the self-organizing map (SOM) [16].

Typically, prototypes are defined as a selected subset of instances that are representative of specific classes or categories within the dataset. During training, these prototypes are iteratively updated. Once trained, they are used to assign a class label (in supervised learning) or to determine cluster membership (in unsupervised learning) for new input instances, based on similarity to the learned prototypes.

Similarity between data instances is generally assessed using a dissimilarity metric, such as Euclidean, Manhattan, Chebyshev, or Minkowski distance. The prototype exhibiting the smallest dissimilarity to a new instance is selected to perform prediction. Due to this direct comparison capability, PBMs are recognized in machine learning theory for their explicit and interpretable representation of input data [17].

Learning in PBMs is realized through a mapping (or projection) from the input space $\chi \in \mathbb{R}^p$ onto the set of R reference units, where p denotes the dimensionality of the feature space. The training procedure of competitive algorithms is governed by the competition among the column vectors of the prototype matrix $\mathbf{W}_{[p \times R]}$. In this process, each reference vector competes to represent regions of the input space covered by the data matrix $\mathbf{X}_{[p \times N]} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$, where N denotes the total number of input samples, and it is typically assumed that $R \ll N$.

The formal structure of PBL methods relies on several definitions. Let $\mathcal{A}$ denote a finite set of R reference units, $\mathcal{A} = c_1, c_2, \ldots, c_R$, where each unit $c_r \in \mathcal{A}$, $r = 1, \ldots, R$, is associated with a reference vector $\mathbf{w}_r \in \mathbb{R}^p$, representing its location in the input space, referred to as receptive field center. The complete set of prototypes is represented by the matrix $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_R]$, where each column $\mathbf{w}_r$ corresponds to a prototype vector. The objective of prototype construction is to determine the vectors $\mathbf{w}_r$ that most effectively represent the input data matrix $\mathbf{X}$.

The assignment of an input vector $\mathbf{x} \in \mathbf{X}$ to a prototype initiates a competitive process among the units in $\mathcal{A}$, wherein the closest prototype is selected to represent the input. The learning rule in PBL algorithms typically follows the winner-takes-all (WTA) principle [14]: only the winning prototype is updated, while the remaining prototypes preserve their learned representations.

Once the prototype matrix $\mathbf{W}$ is trained, a fundamental concept from computational geometry, the Voronoi region, becomes applicable [18]. For each reference unit $i \in \mathcal{A}$, its Voronoi region V_i is defined as the set of points $\mathbf{x} \in \mathbb{R}^p$ for which $\mathbf{w}_i$ is the nearest prototype:

$$V_i = \{\mathbf{x} \in \mathbb{R}^p \mid i = \underset{j=1,\ldots,R}{\arg\min}\, d(\mathbf{x}, \mathbf{w}_j)\}. \tag{1}$$

We define the Voronoi set $\mathcal{R}_i$ as the collection of data points in $\mathbf{X}$ for which i is the closest reference unit. As a result, the input space is partitioned into R mutually exclusive regions:

$$\mathbb{R}^p = V_1 \cup V_2 \cup \cdots \cup V_R \quad \text{with} \quad V_i \cap V_j = \emptyset \quad \text{for } i \neq j, \tag{2}$$

where a partitioning holds under both supervised and unsupervised learning paradigms.

Since the PBL training procedures are established on iterative learning by means of the reference units competition, we initiate by defining the concept of iteration. In unsupervised

PBL, we denote as iteration a single stimulus provoked over the set of reference units when we present an input sample $\mathbf{x}_n$, $n = 1, \ldots, N$, to this set. For a given t-th iteration, the competition is based on the following decision criterion:

$$c_r(\mathbf{x}_n, \mathbf{w}_r(t)) = \underset{i=1,\ldots,R}{\arg\min}\, d(\mathbf{x}_n, \mathbf{w}_i(t)), \tag{3}$$

in which $d(\cdot;\cdot)$ denotes a dissimilarity measure specific to the PBL algorithm used, and $c_r(\cdot) : \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}$ is the reference unit of the nearest, known as winner, prototype among the R available.

In general, unsupervised PBL algorithms are extensions of the WTA learning rule [19]:

$$\begin{aligned} \mathbf{w}_r(t+1) &= \mathbf{w}_r(t) + \eta(t)[\mathbf{x}(t) - \mathbf{w}_r(t)], \\ \mathbf{w}_i(t+1) &= \mathbf{w}_i(t), \quad \text{if} \quad i \neq r, \end{aligned} \tag{4}$$

where $0 < \eta(t) < 1$ is the learning rate.

B. Federated Learning

FL is an emerging machine learning paradigm that enables multiple decentralized entities, commonly referred to as devices, to collaboratively train a global model without sharing devices' local datasets. Its primary objective is to enable model training across distributed networks while preserving data privacy and minimizing communication overhead [20]. This decentralized approach offers several advantages, such as enhanced privacy preservation, lower communication costs, and improved model generalization by leveraging heterogeneous data distributions across clients. FL is particularly beneficial in scenarios where data centralization is infeasible due to privacy concerns or regulatory compliance, as observed in cloud and edge security [21] and healthcare applications [22]. For a comprehensive review of FL applications across various domains, we refer the reader to [23].

In a typical FL scenario, each participating device $k \in 1, \ldots, K$, e.g., smart sensors, retains its own private dataset and conducts local model training. These locally trained models are then aggregated to update a global model, which generally outperforms any individual model while maintaining data privacy. The FL process at each global iteration t typically comprises the following steps:

1) Initialization: A central server initializes the global model and transmits it to a subset S_t of m randomly selected devices, where $1 \leq m \leq K$.
2) Local Training: Each selected device trains locally the model using its local data for E local training epochs.
3) Model Aggregation: The locally updated models are transmitted back to the central server, where they are aggregated to refine the global model.
4) Model Update: The updated global model is redistributed to participating devices, and the process repeats until convergence.

The first FL algorithm introduced in the literature was federated averaging (FedAvg) [24]. Since its introduction, numerous extensions have been proposed, including approaches that incorporate differential privacy [25], fairness considerations [26], and convergence analysis under non-IID data distributions [27].

Despite its benefits, FL faces several inherent challenges. These include:

1) Non-IID Data: Data on each device often reflects the behavior of a specific user, resulting in local distributions that deviate significantly from the global data distribution. This heterogeneity hinders convergence and generalization [24];
2) Data Imbalance: The volume of data per device may vary widely, leading to unequal influence during model aggregation;
3) High Distribution Degree: In certain configurations, the number of participating devices in a training round may exceed the average number of samples per device, complicating model updates;
4) Limited Communication: Mobile and edge devices may exhibit intermittent connectivity, heterogeneous responsiveness, or costly communication channels, posing additional constraints on the training process.

Within the context of models settings, FL frameworks are categorized regarding their scale of participating devices, typically distinguished as cross-silo and cross-device settings [28]. Cross-silo FL involves a limited number of stable and high-capacity nodes (e.g., institutions and organizations), that participate consistently in training. In contrast, cross-device FL engages a massive number of personal or edge devices that may join or leave the network dynamically. The principal characteristics of both settings are summarized in Table I and discussed in [29].

In conclusion, the survey by [30] outlines the principal research directions in FL, including strategies for data partitioning, privacy-preserving mechanisms, FL applications to various machine learning models, and methodologies for addressing data heterogeneity.

III. FedWTA

Federated PBMs (FPBMs) comprise a class of machine learning algorithms that enable the decentralized training of PBMs [31]. In the context of unsupervised FPBMs, the work presented in [32] demonstrates a decentralized implementation of the SOM network. For supervised FPBMs, the term federated LVQ was first introduced in [31].

In FL setups, K represent the total number of devices, and N_k the number of training samples available on the k-th device. In addition, T represents the total number of global

TABLE I
Main characteristics of FL-oriented solutions

Characteristic	**Cross-silo**	**Cross-device**
Processing and storage capacities	High	Low
Scalability	Typically 2-100 devices	Up to $10^{10} devices$
Stability	High	Low
Data distribution	Typically non-IID	Typically IID

iterations, $\alpha = \frac{m}{R}$ denotes the fraction of devices participating in each communication round, and η is the learning rate.

In this work, we extend the formulation introduced in Section II-A to define the federated WTA (FedWTA) algorithm. Let E denote the number of local training epochs, and I_k the number of local iterations at the k-th device. A local iteration corresponds to a single update step of the local model based on one input sample. Thus, one local epoch consists of N_k iterations, and a complete local training procedure with E epochs entails a total of $E \times N_k$ iterations.

The local dataset on device k is denoted by ${}_{[p\times N_k]}\mathbf{X}_k$. The global prototype matrix is denoted by $\mathbf{W}_{[p\times R]}$, while the local prototype matrix for device k is represented by ${}_{[p\times R]}\mathbf{W}_k$.

The FPBM training procedure begins with the initialization of a global prototype matrix $\mathbf{W}(t=0)$, which serves as a set of reference units covering distinct regions of the input space. Each participating client then trains a local model $\mathbf{W}_k(t)$ using its private dataset $\mathbf{X}_k$, updating the prototypes based on local information. The resulting updated prototypes, $\mathbf{W}_k(t+1)$, are sent back to the central server, where they are aggregated to construct the global prototype matrix $\mathbf{W}(t+1)$ for the next iteration.

Our proposed modeling framework adopts the WTA learning rule for local training, as shown in Eq. (4). Furthermore, we employ a weighted mean of the local models as the aggregation strategy to update the global model across global rounds.

This process generates the FedWTA algorithm and minimizes the following cost function:

$$\mathcal{C}(t) = \min_{\nu_{knr},\mathbf{w}_r} \sum_{k\in K}\sum_{n\in N_k}\sum_{r\in R}\frac{1}{2}\nu_{knr} \parallel \mathbf{x}_{kn} - \mathbf{w}_r(t) \parallel^2$$
$$\text{s.t.} \sum_{k\in K}\nu_{knr} = 1, k\in K, n\in N_k, r\in R, \nu_{knr}\in[0,1], \tag{5}$$

where the objective is to minimize the sum of squared errors between the input samples and their associated prototypes across all devices. Here, $\mathbf{x}_{kn}$ denotes the n-th sample from the k-th device, and ν_{knr} is the membership coefficient that quantifies the degree to which the sample $\mathbf{x}_{kn}$ is assigned to the r-th global Voronoi region.

IV. Case Study Description

The case study in this work utilizes water pressure and flow data collected from four pumping stations located within a district-metered area (DMA) in Stockholm, Sweden, spanning the period from January 2018 to March 2019. The dataset, provided by Stockholm Vatten och Avfall (SVOA), encompasses both normal and faulty (leakage) operating conditions, as identified through historical maintenance records. Due to privacy constraints, detailed information regarding the network topology and the labeling procedure is withheld, and the pumping stations are anonymized as A, H, K, and S.

Specifically, the dataset comprises one-minute resolution time series of pressure and flow data for entire days, yielding a total of 448 valid daily samples after excluding days with excessive missing values. Each daily sample includes 1,440 measurements for both pressure and flow, captured across the four pumping stations. A pronounced class imbalance is observed in the dataset, with approximately 88% of the samples corresponding to normal operating conditions. Despite the anonymization and associated limitations, exploratory analysis reveals substantial differences in the data distributions across stations and between operational states, underscoring the importance of effective feature extraction for downstream modeling tasks.

As the samples are acquired from heterogeneous sources, i.e., different pumping stations, a key characteristic of the SVOA dataset is the presence of distinct and non-IID data across sources.

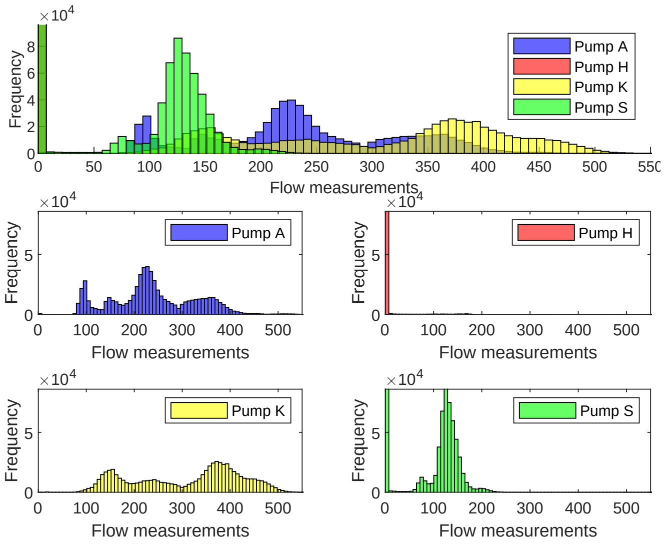

Fig. 1. Histograms of the observed water flow measurements (in m^3/h) at each pump.

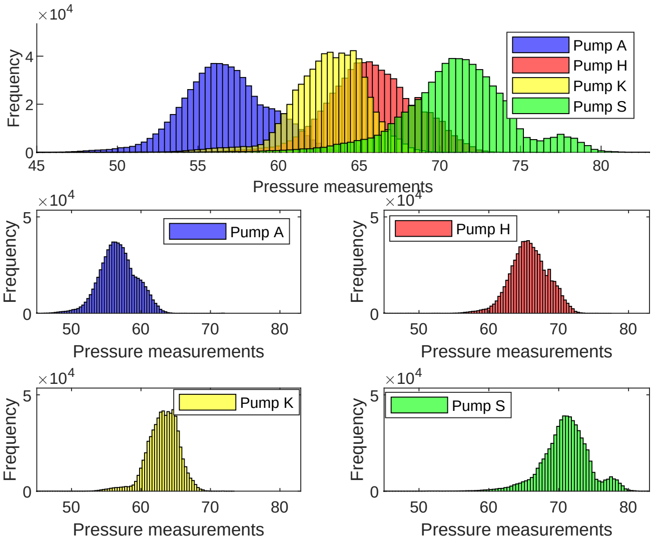

Fig. 2. Histograms of the observed water pressure measurements (in meters of water column) at each pump.

This phenomenon is illustrated in Figs. 1 and 2, which show the distributional differences observed in flow and pressure measurements across the stations.

V. Results and Discussions

A. Contextualization

In this section, we evaluate the proposed federated machine learning approach for leakage detection using a real-world hydraulic dataset. The dataset comprises two operational conditions: normal (N) and leakage (L). To assess the effectiveness of the proposed method, we analyze both the performance and interpretability of the obtained results. Table II summarizes the main characteristics of the proposed distributed federated learning solution.

We performed 100 independent runs of the federated clustering algorithm. Each run adhered to the proposed methodological framework, which consists of three sequential steps: (a) application of canonical discriminant analysis (CDA) [33] to the training set of each pumping station to obtain linear combinations of the interval variables, known as *canonical variables*, that summarize between-class variation; (b) training of the prototype matrix; and (c) evaluation of the resulting clustering models. After each execution, the cluster purity was computed. Specifically, cluster labels were assigned based on the majority class of samples within each cluster. Accordingly, the purity rate is defined as the proportion of data samples whose assigned cluster labels match their ground-truth class labels, relative to the total number of samples.

The empirical evaluation assumes full device participation, where all pumping stations contribute to each round of aggregation. The training process is executed for a maximum of $T = 800$ global iterations, after which the global prototype matrix $\mathbf{W}(T)$ is returned as the final solution. Additionally, the number of local training epochs is set to $E = 10$, and the total number of prototypes is defined as $R = \frac{\sqrt{N_k}}{2} \approx 10$, where N_k denotes the number of samples associated with the k-th pumping station. The local learning rate for the k-th client is defined as $\eta_k(i) = \frac{2}{(E+i)\times N_k}$, where $i = 1, \dots, I_k$ is the iteration index within the local WTA procedure. Note that $I_k = E \times N_k$ represents the total number of WTA iterations executed locally at the k-th device during a single global iteration. Furthermore, the initial prototype matrix $\mathbf{W}_k(t = 0)$ for each local model is constructed by randomly selecting five samples from class N and five from class L at the k-th pumping station.

TABLE II
Main characteristics of the proposed distributed federated solution

Local model	FL setup	Data partitioning	Privacy mechanism	Method for solving heterogeneity
WTA	Cross-silo	Horizontal FL	Model aggregation	Synchronous communication

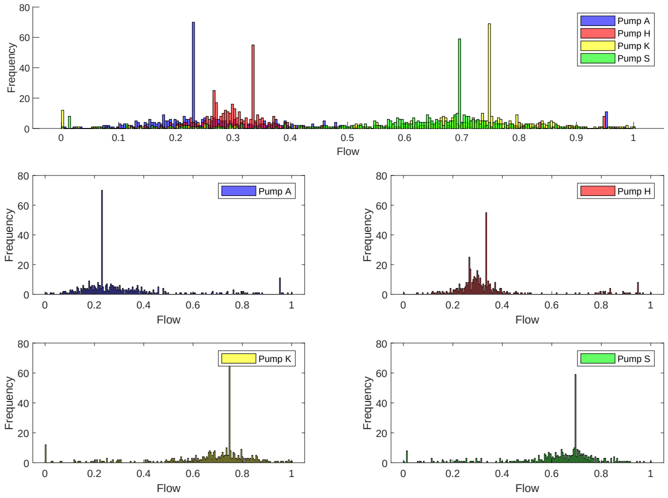

Fig. 3. Histograms of the observed water flow measurements at each pump projected to canonical units.

B. Evaluating the Feature Extraction

We start our analysis by applying the CDA through the following steps: (i) read the pressure and flow signals from the pumping station k; (ii) separate the samples according to their corresponding labels, such as normal and leakage; (iii) calculate the within-group $\mathbf{W_g}$ and between-group $\mathbf{B_g}$ scatter matrices of both hydraulic signals, separately; (iv) find the eigenvector $\mathbf{v}_{1_{[1\times\rho]}}$ associated to the largest eigenvalue of the matrix $\mathbf{W_g}^{-1}\mathbf{B_g}$, in which $\rho = 1440$ denotes the number of components of the raw hydraulic signal; (v) obtain the projected data by applying the inner product between $\mathbf{v}_1$ and the raw local hydraulic dataset; (vi) concatenate the projected pressure and flow data to obtain the processed local dataset $_{[p\times N_k]}\mathbf{X}_k$, where $p = 2$ and $\mathbf{X}_k = 448$ (viii) repeat the steps i to vi for the remaining pumping stations.

Therefore, the pressure and flow measurements from each pumping station are used to construct their processed hydraulic dataset. The resulting histograms of the processed flow and pressure features are illustrated in Figs. 3 and 4, respectively.

In addition to the dimensionality compression of the daily hydraulic time series, these results reveal that the histograms of the treated SVOA dataset are also non-IID, maintaining this characteristic that the FedWTA will manage.

C. Evaluating the Cost Function

Subsequently, we train and validate the performance of the proposed FedWTA algorithm by evaluating its effectiveness in minimizing the cost function defined in Eq. (5). Fig. 5 presents the cost function of the learning model along independent training runs.

In this figure, the solid blue line represents the median values of the mean squared error (MSE) throughout the training process. At the same time, the interquartile range among the first and third quartiles is indicated in the blue-shaded area, and the dotted lines indicate the upper and lower Tukey limits of the MSE values.

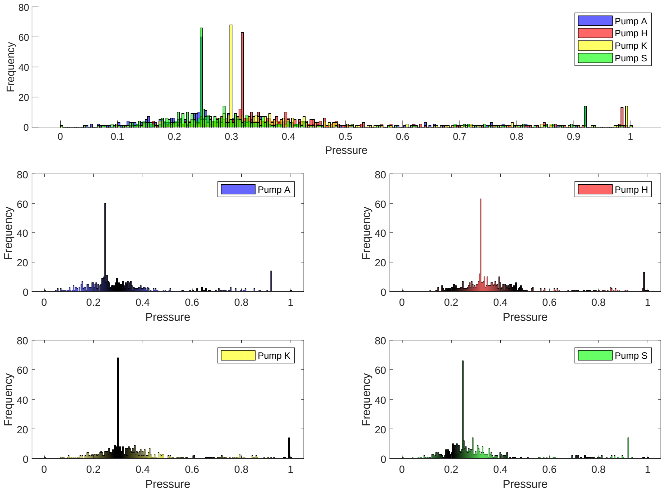

Fig. 4. Histograms of the observed water pressure measurements at each pump projected to canonical units.

Furthermore, Fig. 6 shows the median MSE values obtained for each pumping station throughout the global training process. As can be observed, Pump S consistently exhibited the lowest MSE values, whereas Pumps A and K showed the highest cost function levels. However, the MSE associated with Pump A decreased more significantly over time, eventually converging to values comparable to those of the other pumping stations.

D. Performance of the FedWTA

The statistical purity performance rate of the FedWTA method for each pumping station is shown in Fig. 7.

A closer look at these metrics reveals that all pumping stations achieved high performance through the global model. Additionally, there is a substantial performance difference between Pumps A and K. While both obtained very similar maximum and median values, their mean and minimum values show a minimum variation. Moreover, Pump H showed the lowest performance results and was the only one that failed to reach a maximum purity rate of 100%.

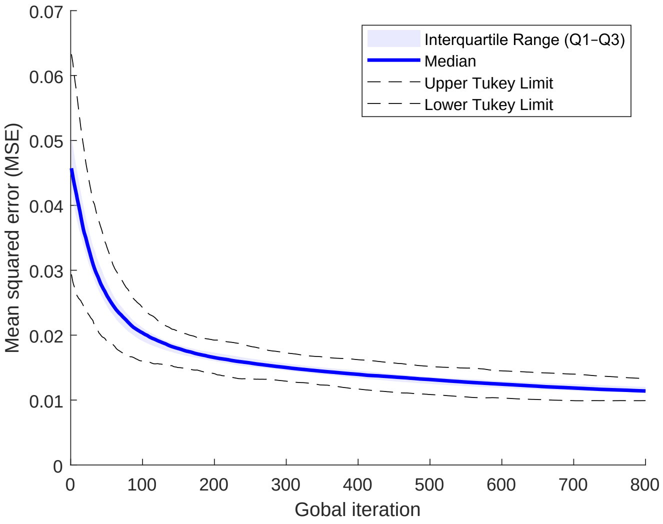

Fig. 5. MSE obtained along the global iterations.

A final aspect worth highlighting concerns the consistently high purity rates, which suggest that the entire day data acquisition period effectively characterizes the samples. This observation raises two key insights: (i) the leakage detection problem may have become linearly separable due to the extended observation window, and (ii) it may be possible to reduce the required acquisition period while maintaining attractive predictive performance. These insights suggest a potential trade-off between the time required for data characterization and the resulting clustering quality, measured by purity.

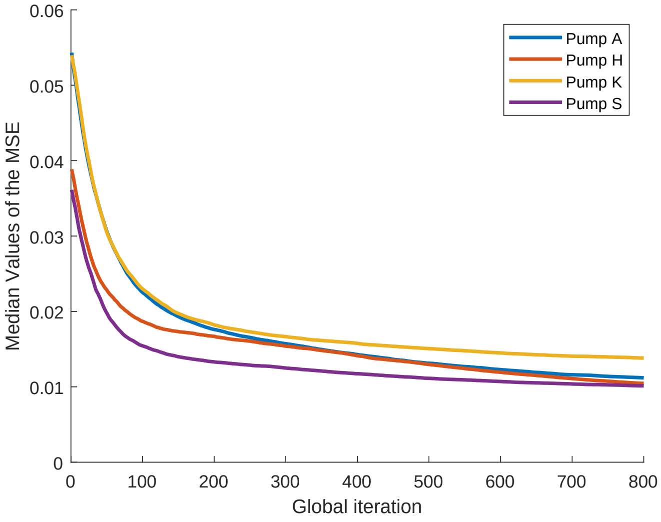

Fig. 6. MSE per pump obtained along the global iterations.

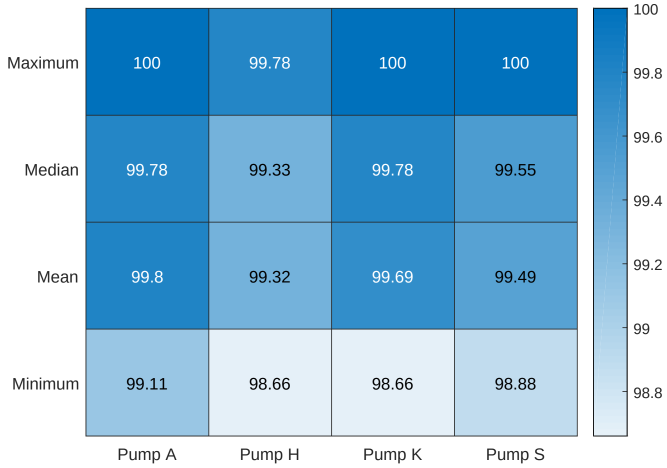

Fig. 7. Purity rate perfomance (in %).

E. Interpretability of the FedWTA

Finally, we assess the interpretability of the results through visual inspection of the global models.

The obtained models, represented by Voronoi cells, allowed us to analyze how the local data was clustered in their respective feature spaces. In the following visualizations, blue dots indicate the N samples, red crosses correspond to the L samples, black pentagrams mark the prototypes, and the hyperplanes delineate the boundaries of the Voronoi cells.

In detail, during a given run of the algorithm with purity rate of $\text{Purity}[A, H, K, S] = [100\%, 98.88\%, 99.78\%, 99.78\%]$, Fig. 8 illustrates the local models generated at the end of the first local training phase using the WTA algorithm. As shown in the figure, each pumping station produced its own local model based on its respective data, resulting in the formation of different Voronoi cells. An important observation is that only Pump A generated data partitions that achieved a purity rate of 100%, indicating a highly effective local separation of the data in that device.

Then, Fig. 9 presents the initial global model obtained by the first aggregation the local models. Specifically, this figure illustrates the first use of a shared global model across all pumping stations. Moreover, it can be observed that class separation at Pump K has improved, whereas Pump A no longer achieves purity rate of 100% .

Finally, Fig. 10 shows the positioning of the Voronoi cells at the end of the training of the FedWTA algorithm. A particularly relevant observation is the presence of dead units, i.e., Voronoi regions that remain unoccupied by data in some clients while actively representing samples in others. This heterogeneity enhances data privacy, as it becomes challenging for external observers or adversaries to determine which regions of the space are truly representative for a given client. Consequently, the unique operational characteristics of each pumping station are better preserved throughout the federated training process.

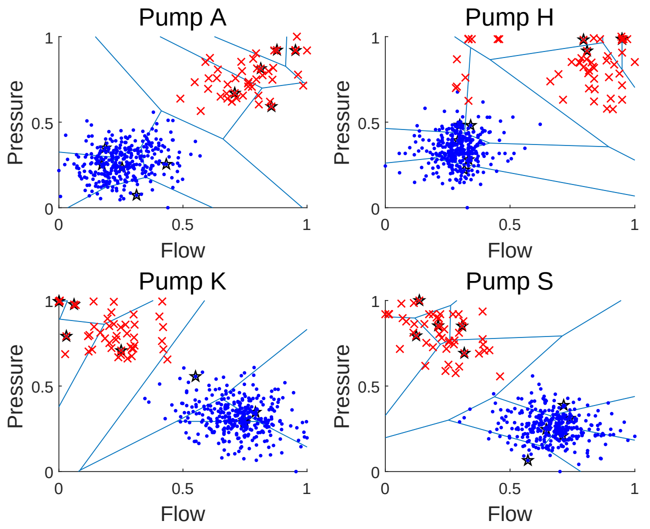

Fig. 8. Voronoi cells generated along the local models training ($T = 1$).

VI. Conclusions

In this work, we focused on modeling a federated solution to address the water leakage detection challenge in WNDs. To this goal, we considered a relevant real-world study case to analyze interpretative concepts from PBL and FL paradigms to efficiently explore non-IID data.

The resulting global models offered valuable insights into the distribution and separation of patterns across devices, reinforcing the model's ability to produce meaningful and interpretable decision boundaries in a privacy-aware federated learning setting.

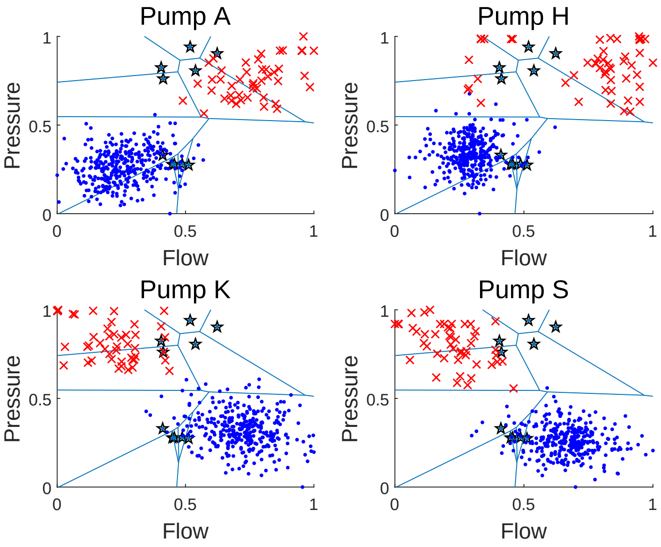

Fig. 9. Voronoi cells generated at the fisrt model aggregation ($T = 1$).

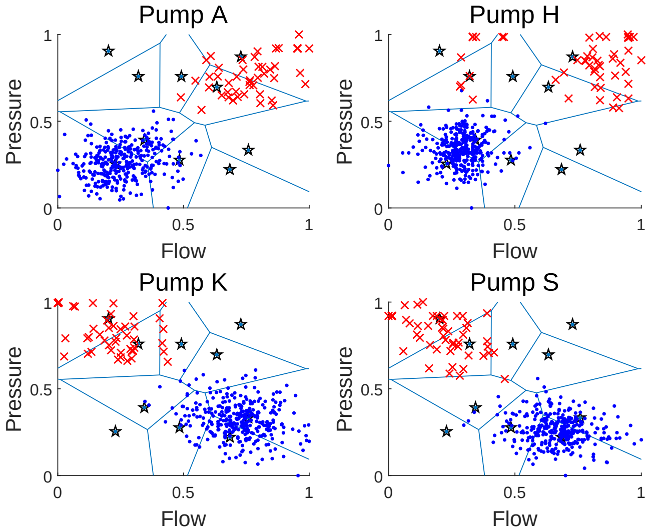

Fig. 10. Voronoi cells generated when using the proposed federated solution at the end of a given run ($T = 800$).

Hence, the observed findings showed the viability and potential benefits of FPBMs. Since our modeling focused on the FedWTA algorithm can be adjusted and applied to several problems in anomaly detection, we hope our insights can inspire future works on other FPBMs.

With regard to the design and improvement of WDNs, we plan on evaluating the proposed FL solution under varying sampling periods. This analysis is particularly important in anomaly detection scenarios, where the required duration of data acquisition prior to prediction plays a critical role.

To conclude, we highlight the main contributions of this work: (i) promoting the practice of initially investigating the problem using interpretable methods rather than directly applying black-box strategies; (ii) demonstrating the potential of FPMs; and (iii) presenting a feasible operational design for water leakage detection, aiming to encourage improvements in Brazilian water utilities within the context of advanced metering infrastructures.

Acknowledgment

The authors thank the Stockholm Vatten och Avfall company, Stockholm, Sweden, for providing the data used in this study.

References

[1] M. Lawal, "Historical development of the pipeline as a mode of transportation," *Geograph Bull*, vol. 43, no. 2, pp. 91–99, 2001.

[2] S. K. Sharma and S. Maheshwari, "A review on welding of high strength oil and gas pipeline steels," *Journal of Natural Gas Science and Engineering*, vol. 38, pp. 203–217, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1875510016309398

[3] Instituto de Pesquisa e Estratégia Econômica do Ceará (IPECE), "Relatório de ações ceará e os objetivos de desenvolvimento sustentável," Governo do Estado do Ceará, Tech. Rep., jun 2022. [Online]. Available: https://www.ipece.ce.gov.br/wp-content/uploads/sites/45/2022/05/Relatorio_de_Acoes_Ceara_ODS_062022.pdf

[4] D. Zaman, M. K. Tiwari, A. K. Gupta, and D. Sen, "A review of leakage detection strategies for pressurised pipeline in steady-state," *Engineering Failure Analysis*, vol. 109, p. 104264, 2020.

[5] T. K. Chan, C. S. Chin, and X. Zhong, "Review of current technologies and proposed intelligent methodologies for water distributed network leakage detection," *IEEE Access*, vol. 6, pp. 78 846–78 867, 2018.

[6] D. Barros, I. Almeida, A. Zanfei, G. Meirelles, E. Luvizotto Jr, and B. Brentan, "An investigation on the effect of leakages on the water quality parameters in distribution networks," *Water*, vol. 15, no. 2, p. 324, 2023.

[7] S. G. Vrachimis, D. G. Eliades, R. Taormina, Z. Kapelan, A. Ostfeld, S. Liu, M. Kyriakou, P. Pavlou, M. Qiu, and M. M. Polycarpou, "Battle of the leakage detection and isolation methods," *Journal of Water Resources Planning and Management*, vol. 148, no. 12, p. 04022068, 2022.

[8] A. Artelt, M. S. Kyriakou, S. G. Vrachimis, D. G. Eliades, B. Hammer, and M. M. Polycarpou, "Epyt-flow: A toolkit for generating water distribution network data," *Journal of Open Source Software*, vol. 9, no. 103, p. 7104, 2024. [Online]. Available: https://doi.org/10.21105/joss.07104

[9] J. Brinkrolf and B. Hammer, "Interpretable machine learning with reject option," *at - Automatisierungstechnik*, vol. 66, no. 4, pp. 283–290, 2018. [Online]. Available: https://doi.org/10.1515/auto-2017-0123

[10] A. Artelt, S. G. Vrachimis, D. G. Eliades, U. Kuhl, B. Hammer, and M. M. Polycarpou, "Interpretable event diagnosis in water distribution networks," 2025. [Online]. Available: https://arxiv.org/abs/2505.07299

[11] D. P. Sousa, R. Du, J. Mairton Barros da Silva Jr, C. C. Cavalcante, and C. Fischione, "Leakage detection in water distribution networks using machine-learning strategies," *Water Supply*, vol. 23, no. 3, pp. 1115–1126, 02 2023. [Online]. Available: https://doi.org/10.2166/ws.2023.054

[12] D. P. Sousa, J. Mairton Barros da Silva Jr, C. C. Cavalcante, and C. Fischione, *A Federated Prototype-Based Model for IoT Systems: A Study Case for Leakage Detection in a Real Water Distribution Network*. John Wiley & Sons, 2025.

[13] D. P. Sousa, "Leakage detection in a real water distribution network through a federated prototype-based model," Tese (Doutorado em Engenharia de Teleinformática), Universidade Federal do Ceará, Fortaleza, Brasil, 2024. [Online]. Available: http://repositorio.ufc.br/handle/riufc/80329

[14] T. Kohonen, "Essentials of the self-organizing map," *Neural Networks*, vol. 37, pp. 52–65, 2013, twenty-fifth Anniversay Commemorative Issue.

[15] D. Nova and P. A. Estévez, "A review of learning vector quantization classifiers," *Neural Computing and Applications*, vol. 25, no. 3-4, pp. 511–524, 2014.

[16] D. Miljković, "Brief review of self-organizing maps," in *40th international convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE, 2017, pp. 1061–1066.

[17] M. Biehl, B. Hammer, and T. Villmann, "Prototype-based models in machine learning," *WIREs Cognitive Science*, vol. 7, no. 2, pp. 92–111, 2016.

[18] B. Boots, K. Sugihara, S. N. Chiu, and A. Okabe, *Spatial tessellations: concepts and applications of Voronoi diagrams*. John Wiley & Sons, 2009.

[19] T. Kohonen, "Improved versions of learning vector quantization," in *IEEE IJCNN International Joint Conference on Neural Networks*, 1990, pp. 545–550.

[20] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[21] L. Albshaier, S. Almarri, and A. Albuali, "Federated learning for cloud and edge security: A systematic review of challenges and ai opportunities," *Electronics*, vol. 14, no. 5, p. 1019, 2025.

[22] A. S. Fathima, S. M. Basha, S. T. Ahmed, S. B. Khan, F. Asiri, S. Basheer, and M. Shukla, "Empowering consumer healthcare through sensor-rich devices using federated learning for secure resource recommendation," *IEEE Transactions on Consumer Electronics*, 2025.

[23] L. Li, Y. Fan, M. Tse, and K.-Y. Lin, "A review of applications in federated learning," *Computers & Industrial Engineering*, vol. 149, p. 106854, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360835220305532

[24] Z. Lu, H. Pan, Y. Dai, X. Si, and Y. Zhang, "Federated learning with non-iid data: A survey," *IEEE Internet of Things Journal*, 2024.

[25] A. El Ouadrhiri and A. Abdelhadi, "Differential privacy for deep and federated learning: A survey," *IEEE access*, vol. 10, pp. 22 359–22 380, 2022.

[26] T. Li, M. Sanjabi, A. Beirami, and V. Smith, "Fair resource allocation in federated learning," *arXiv preprint arXiv:1905.10497*, 2019.

[27] L. Chen, X. Ding, Z. Bao, P. Zhou, and H. Jin, "Differentially private federated learning on non-iid data: Convergence analysis and adaptive optimization," *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[28] Q. Li, Z. Wen, Z. Wu, S. Hu, N. Wang, Y. Li, X. Liu, and B. He, "A survey on federated learning systems: Vision, hype and reality for data privacy and protection," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[29] I. Kholod, E. Yanaki, D. Fomichev, E. Shalugin, E. Novikova, E. Filippov, and M. Nordlund, "Open-source federated learning frameworks for IoT: A comparative review and analysis," *Sensors*, vol. 21, no. 1, p. 167, 2020.

[30] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, and Y. Gao, "A survey on federated learning," *Knowledge-Based Systems*, vol. 216, p. 106775, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705121000381

[31] J. Brinkrolf and B. Hammer, "Federated learning vector quantization." in *ESANN European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2021.

[32] M. Servetnyk, C. C. Fung, and Z. Han, "Unsupervised federated learning for unbalanced data," in *IEEE Global Communications Conference*, 2020, pp. 1–6.

[33] A. C. Rencher, "Interpretation of canonical discriminant functions, canonical variates, and principal components," *The American Statistician*, vol. 46, no. 3, pp. 217–225, 1992.