Improving Water Quality Prediction by Eliminating the Impact of Water Background in ML Models

Sarah Instenes, Nehpal S. Shekhawat, Ivan Micanovic, Cristinel Ababei, Richard Povinelli, Chung Hoon Lee

Electrical and Computer Engineering, Marquette University
Email: {sarah.instenes,nehpal.shekhawat,ivan.micanovic,cristinel.ababei,richard.povinelli,chunghoon.lee}@marquette.edu

***Abstract*—We present a method to investigate how water source (i.e., water background) impacts the performance of machine learning (ML) models developed to predict concentration of metal ions in drinking water. Our hypothesis is that portions of the input data in the ML model may be responsible for or capture information about the source itself. If such portions could be identified—and provided that such portions do not affect or their impact is minimal on the actual prediction of concentration of metal ions—then, they could be eliminated (1) during model development and/or (2) when the ML model is used for real-time inference/prediction. The proposed method partitions the input data range into a number of sub-ranges and then investigates which such sub-ranges are the most relevant for capturing the background. For this investigation, we develop a new separate ML model to classify water background, and employ a SHAP technique to identify the most relevant sub-ranges that impact this classification. It is these sub-ranges that indicate what portions of the input data should safely be removed from the development/optimization of the main ML model used for prediction of metal ion concentration. The simulation results indicate that the proposed method successfully identifies portions of the input data that are responsible mainly for capturing water background. We conclude that the proposed method thus provides an effective technique to develop high prediction accuracy ML models (for predicting metal ion concentration) that can be robust/agnostic to factors related to water backgrounds.**

***Index Terms*—machine learning; prediction; ppb-level metal ions; water background; water source; input relevance;**

I. Introduction and Motivation

Monitoring drinking water to detect the presence of pollutants (chemicals, heavy metal ions, etc.) is crucial in ensuring safety of the population and security of critical infrastructure. Given the increase of water pollution and increasing attempts to control or compromise water treatment infrastructure, there is a growing need for cost-effective and highly accurate water sensor systems to be deployed in households to continuously monitor for the presence of pollutants and their concentration. It is in this context that in our previous recent work, we proposed machine learning (ML) models for the prediction of both Cu and Pb ion concentrations (measured as ppb) in drinking water. These ML models were used in a novel contact-less water monitoring system built using a novel microwave block loop gap resonator (BLGR) coupled with a vector network analyzer (VNA).

An important challenge we faced with our previous work was that our ML models developed/optimized using data collected with water samples originating form a given location or source (also referred to as water background) suffered from significantly lower prediction accuracy when employed for real-time inference on new water samples (never seen by the model before during model development and optimization) originating from different sources. This is precisely the main motivation for the work presented in this paper. We present a method to address this challenge.

Our method is based on the hypothesis that portions of the input data used in our ML models developed for prediction of metal ion concentrations may be solely responsible for or capture information about the water background/source itself. If such portions could be identified—and provided that such portions do not affect or their impact is minimal on the actual prediction of concentration of metal ions—then, they could be eliminated (1) during model development and/or (2) when the ML model is used for real-time inference. In other words, portions of the input data that capture/characterize the background aspect of the data could be removed while the remaining portions of the input data would suffice for the development of ML models. These models would provide the same high prediction accuracy on both data collected from the first water source (used for model development) and from additional water sources (used for model deployment for real-time inference in the field).

In this paper, we investigate how we could identify specifically those portions. The proposed method is based on the SHAP (SHapley Additive exPlanations) technique, which is a popular, game theory-based method used in ML to explain the output of complex black-box models. It quantifies how much each individual input feature contributes to a specific prediction. That is, SHAP can be used to conduct a sensitivity analysis to determine the importance (or the extent of contribution) of input features towards the output prediction. Therefore, SHAP analysis is used as a filtering process in order to reduce the input feature space for our ML model used to classify lead levels in drinking water.

A. Background

In order to provide the reader with information where the work in this paper fits in the big picture and to make this paper self-contained, in this section, we provide a brief description of our contact-less water monitoring system [1] and the ML models for the prediction of both Cu and Pb ion concentrations [2]. A simplified block diagram of our sensor system from [1]

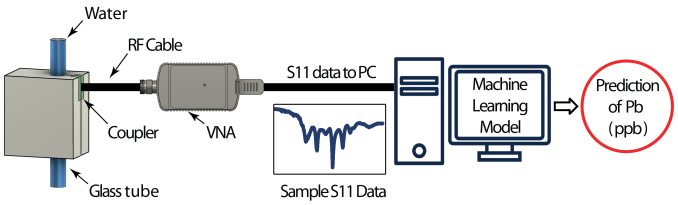

Figure 1. System level diagram of sensor system from [1].

is shown in Fig.1. The system combines a block loop gap resonator (BLGR) with a vector network analyzer (VNA). An inductive coupling loop fabricated on a printed circuit board and integrated with the resonator is connected to Port 1 of the VNA, which collects the reflection coefficient, S11, over a 100 MHz - 6 GHz frequency range.

The S11 amplitude values are then used as input features into the ML models that we reported in [2]. While in [2] we studied several different ML models, including convolutional neural networks (CNN) and deep neural networks (DNN) models, in this paper we will focus only on the use of DNNs as our primary ML model for prediction of Pb ion concentration (ppb levels) in drinking water. The original model demonstrated strengths in predicting the ppb levels when the dataset was confined to one water source. However, the efficacy of this model deteriorated as the dataset expanded to include additional water sources. While the accuracy was high for all ppb levels for the Milwaukee water samples, for instance, it was not as strong when Brookfield water samples and samples from other areas were introduced for real-time inference or testing.

II. Proposed Method

A. Description of SHAP

The method proposed in this paper is based on the SHAP (SHapley Additive exPlanations) technique. SHAP is an instance of explainable AI that is helpful in understanding the motivations behind a machine learning model's classification decisions. It is especially apt for discrete models, which cannot use Layerwise Relevance Propagation (LRP) to interpret decisions because of the absence of gradient descent that LRP relies on. This approach is also useful for gathering information on local and global levels as it shows where each sample falls into the decision making process, in addition to the larger trends in classification of the data.

The development of SHAP was a major catalyst for explanatory artificial intelligence (XAI). Lundberg and Lee developed this novel approach in 2017 [3] by creating a bridge among the six popular methods of model explainability at the time: LIME, DeepLift, Layer-Wise Relevance Propagation, Shapely Sampling Values, Shapely Regression Values, and Quantitative Input Influence. They proved that the underlying model for the explanation of each method was actually the same.

Marcillio and Eller [4] took these developments a step further by demonstrating the efficacy of SHAP for feature-selection purposes and demonstrating its general advantage over other popular feature-selection methods, including Analysis of Variance (ANOVA), Mutual Information (MI), and Recursive Feature Elimination (RFE). On the three main datasets examined for accuracy, SHAP performed the best in two out of three of them, with ANOVA only being slightly more accurate than SHAP for one of the datasets (Boston), though that difference was almost negligible (0.02 difference), showing that SHAP was better in most areas of accuracy or at least comparable. However, in terms of execution runtime, their study found that SHAP was slower, especially when compared to ANOVA in all datasets used. They attributed this to the greater depth involved in SHAP, making it ideal for a deep analysis and guaranties of relevant feature selection, but less competitive for highly scalable projects at this point.

Due to its performance, SHAP has become a popular technique for extracting the most relevant features, with many studies using it recently [5]–[10]. However, in this paper, we invert this approach by using it to remove the most relevant features from the input data, relevance in terms of capturing the background of the input features, in order to retaininin those input features that are less relevant to identify water background, but which are relevant and essential to discerning various levels of concentrations of Pb metal ions in drinking water for water data from multiple different water sources.

B. Description of the Proposed Method

The proposed approach can be described in two main stages (see Fig.2).

Stage 1: In the first stage, we employ SHAP analysis together with a new ML model developed for the purpose of classification for the three different water sources/backgrounds that we work with in this paper: Milwaukee water, DI water, and Pure water. The objective here is to identify input features that are relevant or capture water background. Once those most salient features are identified, we eliminate them from the input data. The resulting filtered data will be used in Stage 2 (described later) as reduced input data into the main ML models used for prediction of Pb metal ion concentration (ppb levels). Our hypothesis is that if the most salient features for determining/capturing water source's classification were removed, then the ML models from Stage 2 would have several benefits: (i) First, they would work with smaller size inputs, thereby reducing the ML model size, which should be more computationally efficient; their development (training) would be faster too; (ii) Second, they could be more likely to fit on resource constrained devices for IoT deployment. This hypothesis relies on the assumption that the features most crucial to the identification of water sources are distinct from the features necessary to predict metal ion concentrations. If these features are shared, then performance of ML models from Stage 2 may further erode, rather than improve or remain the same. The justification for this distinction assumption is the initial drop in performance of the model when multiple water sources were introduced.

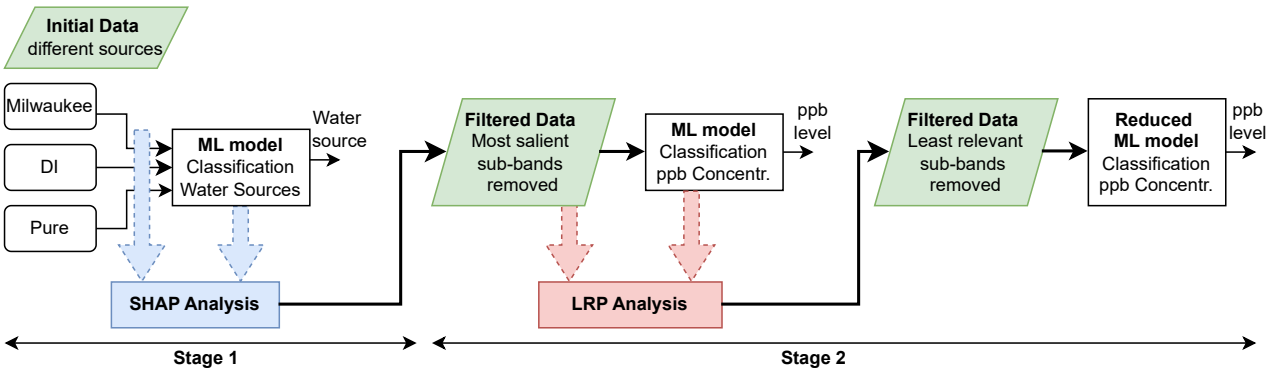

Figure 2. The proposed method has two main stages.

For the ML model in Stage 1, we use a Random Forest Classifier from the *sklearn* library, with the hyperparameter number of estimators set to n=300. Once this model is trained and tested on the initial data, SHAP values are determined with the help of *shap* library. In our investigation, we first partition the input data into bands as follows: the initial input data range of 0-6 GHz (comprising of 20,000 S11 coefficient amplitude values) is split into six sub-bands: 0-1 GHz, 1-2 GHz, 2-3 GHz, 3-4 GHz, 4-5 GHz, and 5-6 GHz.

After retrieving the SHAP values for each water source shown in Fig.2, we identify patterns in the most relevant frequency sub-bands. These sub-bands are in the top-half of the model's classification of all three water sources, indicating their salience for the classification ML model. Our objective is to find such top-half sub-bands that are the same (i.e., shared) across all three different water sources because they would be our top candidates for removal from the initial data. For example, if data from all three water sources show that the 4-5 GHz sub-band is highly relevant for classification with ML model from the left-hand side of Fig.2, then that sub-band will be removed for the main model from Stage 2. To continue pruning beyond full consensus, we apply a majority vote strategy for selecting additional frequency sub-bands for removal. Thus, if two of the three water sources indicate that a frequency sub-band is highly relevant to classification, then that sub-band will also get removed. We continue this process until accuracy drops below an acceptable threshold, initially defined as 3%.

Because we want our primary metal concentration model from Stage 2 to be water source agnostic, we remove all data values from the frequency sub-bands identified for removal as described above. In this way, the metal concentration model will not be "distracted" by superfluous information. Here, the explainable AI, SHAP, is used as a filter to refine the original dataset and reduce the input size.

Stage 2: In the second stage, we evaluate the main DNN model used to predict Pb level concentrations. This is done now with input features being the size-reduced "Filtered Data Most salient sub-bands removed" in Fig.2. The DNN model is first trained using the reduced dataset, then it is evaluated using this smaller feature space. If our hypothesis and assumption discussed earlier hold true, then we expect the prediction accuracy of the ML model to be as good as the accuracy of the same ML model when it was developed with the entire "Initial Data". Performance of the DNN is evaluated by accuracy and F1 score, while its efficiency and size are measured by the total number of parameters and the storage size of the model with this set of features (.keras file, FP32).

We further attempt to reduce the input data by conducting a Layerwise Relevance Propagation (LRP) analysis to identify and eliminate from the input data the least relevant input features from the already reduced data entering Stage 2. The number of features eliminated in this step is selected by trial and error such that it is as large as possible with a minimal penalty in prediction accuracy of the ML model from the right hand side in Fig.2.

III. Datasets Description

In developing and optimizing the ML models described in this paper, we needed datasets that are used for supervised training, validation and testing. For this purpose, we have experimentally collected three datasets (referred to as Dataset1, Dataset2, and Dataset3) as described next. The three datasets were collected for three different water samples, each representing three different sources (e.g. backgrounds), specifically (Milwaukee, DI, and Pure). In all three cases, we are studying several different levels of Pb ion metal concentrations (measured as ppb, parts per billion).

For example, to generate and collect data for Dataset1, we use drinking water samples from Milwaukee. These water samples are prepared into four different batches before our experiments in order to control the actual Pb concentration in each batch: (0ppb, 1ppb, 5ppb, 10ppb). Note that the 0ppb batch is simply the water sample as collected from the source, without any Pb added to it. The other batches have added a controlled amount of Pb in the desired amounts. Using the experimental setup described in Fig.1 for each of the four different batches, we collect and store 283 S11 measurements from the VNA. Thus, we collect a total of 1,132 measurements and store as Dataset1. Dataset2 and Dataset3 are generated

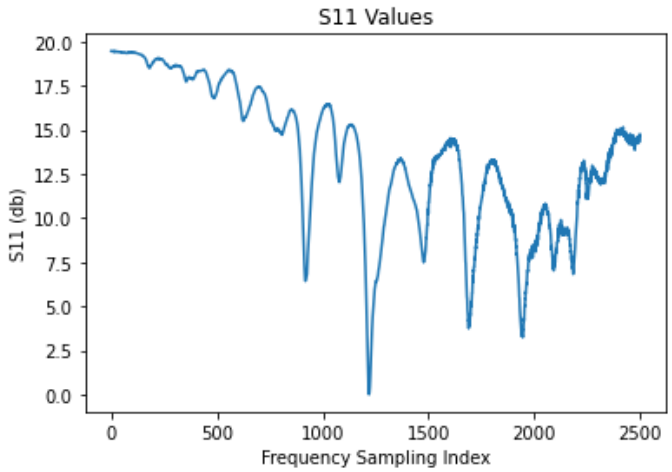

Figure 3. Example of input features used for ML models investigated in this paper: S11 magnitude values collected from the VNA, over a frequency range 0-6 GHz.

using the same procedure. Each S11 measurement collected from the VNA and stored into the corresponding dataset includes 20,000 numerical values for S11 magnitude and 20,000 values for S11 phase values; these values correspond to 20,000 different frequencies spread across the VNA's working frequency range 0-6 GHz. Note that in this paper, we only use the S11 magnitude values for ML development. Thus, the 20,000 S11 values represent the input features of a *datapoint* of the form {*Input Features*, *Output Label*} that is recorded in the dataset. An example of a such a datapoint is shown in Fig. 3, where we downsampled the 20,000 values by 8. Finally, we note that each of these three datasets is split into 70/30 for training/testing during model development.

IV. Experiments, Results and Discussion

A. Stage 1 - SHAP Analysis

As discussed in Section II-B, the ML model in Stage 1 from Fig.2 is a Random Forest (RF) classifier. Because we are looking at three different water sources, this is a 3-class classification problem. The RF model has a single final output, which is the predicted class label (e.g., Class1, Class2, or Class3 corresponding to water sources Milwaukee, DI, or Pure) for a given input. Hence, in a first step at this stage, we defined the RF model and trained it. Upon standard training, the RF model resulted in an accuracy of 99% when predicting or classifying the water source.

Next, we employed the SHAP technique. In framing the problem for our SHAP analysis, we first partition input features into sub-bands as follows: the initial input data range of 0-6 GHz (spanned by all 20,000 S11 coefficient amplitude values) is split into six sub-bands: 0-1 GHz, 1-2 GHz, 2-3 GHz, 3-4 GHz, 4-5 GHz, and 5-6 GHz. It is precisely these sub-bands that the SHAP technique deals with in order to figure out which sub-bands are the most important or critical in determining the classification output for each class. SHAP technique provides its analysis result in the form of specialized

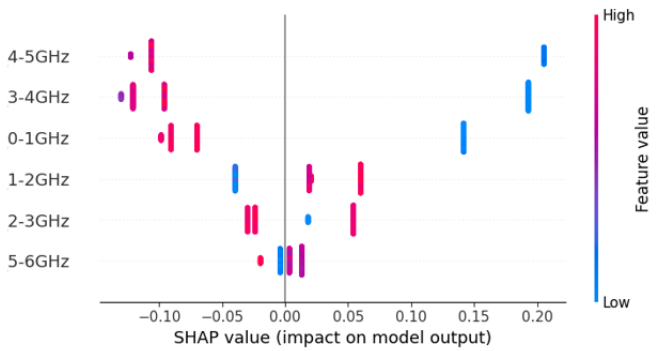

Figure 4. SHAP summary plot of Class1, i.e., Milwaukee water source.

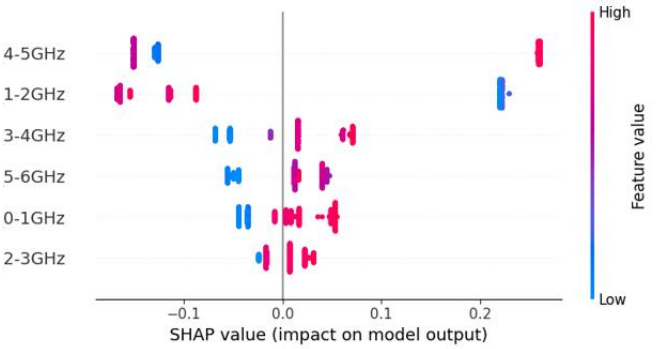

Figure 5. SHAP summary plot of Class2, i.e., DI water source.

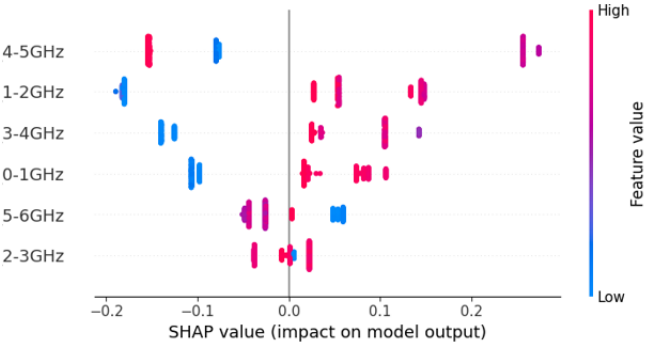

Figure 6. SHAP summary plot of Class3, i.e., Pure water source.

plots. For example, Figs. 4, 5, and 6 show such SHAP plots for all three classes. On these plots, the y-axis represents the *importance* of the frequency sub-bands, which is ranked from the most important at the top of the y-axis to least important at the bottom. The x-axis of these plots indicates the relative *influence* or *strength* of the input feature values within each sub-band or group; this influence is color-coded, and it is shown in red if the influence is high and in blue if the influence is low.

Looking at Fig. 4, the SHAP plot indicates that frequency sub-band 4-5 GHz proved to be the most important in determining the water source classification for Class1. It also

indicates that the sub-band 5-6 GHz band was the least important. The plots from Figs. 5 and 6 indicate that the same band 4-5 GHz is again the most important in determining the classification output for classes Class2 and Class3 as well, which supports the ultimate objective in this paper: to find whether there exist such sub-bands, common across all classified labels. We observe from the pattern emerging from these SHAP plots that sub-bands 4-5 GHz, 3-4 GHz, and 1-2 GHz are the most important bands for classification. These sub-bands will be eliminated from the input features used in the model development from Stage 2 in Fig.2. In other words, the input features used to develop ML models in Stage 2 from Fig.2 will only be the data from sub-bands 0-2 GHz and 2-3 GHz.

B. Alternative Methods

Alternative dimensionality reduction methods were explored, such as PCA and Mutual Information (MI). PCA reduced the feature space to 3 components that explain 95%. This means that there will be 7,500 multiplication operations per sample (2500 downsampled features x 3). MI reduced the feature space to 5 features, but these were spread across a broader spectrum of frequency bands and inconsistent, which limits the ability to reduce the inputs at the collection stage during inference. Therefore, these approaches are not optimal for embedded devices with more hardware restrictions.

C. Stage 2 - ML Models for Pb Concentration Prediction

Based on the SHAP analysis from the previous section, in Stage 2 in Fig.2, we develop the DNN model for the prediction of the Pb ion concentration by using as input features only the data from bands 0-2 GHz and 2-3 GHz. Specifically, using only data from these two sub-bands and employing a downsampling by 8, the total number of actual S11 coefficient amplitude values used as one input into the DNN model is 830, which is reduced from the initial value of 2,500 that represents our base/reference for comparison. For example, Fig.7 shows one such impact with this number of features to the DNN model.

We conducted further model optimization with the goal of reducing its architecture without degrading performance. First, we reduced the number of hidden layers to one only, which did not affect accuracy of the base model (e.g., the model developed using all 2,500 values as input features). The accuracy remained at 0.9999, as did the precision and recall. Training of this model was performed with Adam for optimization over 250 epochs, resulting in acceptably low loss as seen in Fig.8. The F1 score and accuracy remained high during testing, which can be seen in Fig.9, and the comparisons among width variations in the architecture are listed in Fig.I. The model size (storage size needed to store the trained model) decreased to 34.52 KB. Table 1 shows a summary of our results for several additional model changes where the number of units/neurons on the hidden layer (the width) is varied for values 32, 16, 8, and 4 units. We observe that the accuracy degraded (0.6775) beyond an acceptable threshold when the number of units on the hidden layer was 4. Therefore, the smallest layer width that achieved the best accuracy was 8, with an accuracy of 0.9961. The number of model parameters was reduced to 1108, and the storage size was reduced to 4.33 KB. This size change is a reduction of more than 87% compared to the 64 layer width single hidden layer version of the base/reference DNN model, and a reduction of more than 97% compared to the initial base/reference DNN model that has two hidden layers.

Table I
SUMMARY OF DNN MODEL PERFORMANCE WHEN THE NUMBER OF NEURONS ON THE HIDDEN LAYER IS VARIED.

Width	**Accuracy**	**F1 Score**	**Params.**	**Size (FP32)**
64	0.9999	0.999	8836	34.52 KB
32	0.9980	0.9980	4420	17.27 KB
16	0.9971	0.9971	2212	8.64 KB
8	0.9961	0.9961	1108	4.33 KB
4	0.6775	0.6718	556	2.17 KB

These results confirm our initial hypothesis that if the most salient input features, which would capture mostly information about the water source/background, can be identified and removed, then the ML models from Stage 2 in Fig.2 could be developed using the reduced input features for the purpose of accurately predicting the Pb ion concentration. This is possible because the portions of the input features that were responsible for the water background are removed, and the ML models developed with the reduced input features are now more *robust* against variety in water sources.

Our hypothesis relied on the assumption that the features most crucial to identifying the water sources were distinct from the remaining features necessary in predicting metal ion concentrations. However, this assumption is only partially confirmed: our findings do indicate that some of the properties associated with the frequency sub-bands that are most salient in identifying water sources are not necessarily as crucial/important in identifying Pb levels in the water. While we do not have direct evidence for a complete independence or separation of the properties of input features that capture background on the one hand and Pb levels on the other hand, the fact that the prediction accuracy of the DNN models developed with the reduced set of input features indirectly suggests that such independence or separation is very likely. Further analysis is required to investigate the impact of other possible impacting factors, such as climate, season, region, and other environmental factors; this is left for our future work.

V. CONCLUSION

We presented a method to identify and remove the inessential input features of a ML model used to predict the concentration of Pb metal ions in drinking water. We identified and isolated features that are directly related to the water source or background, thereby reducing the input size, and consequently reducing the model size and improving efficiency. This procedure improved the robustness of the model against data measurements generated on samples of water from various

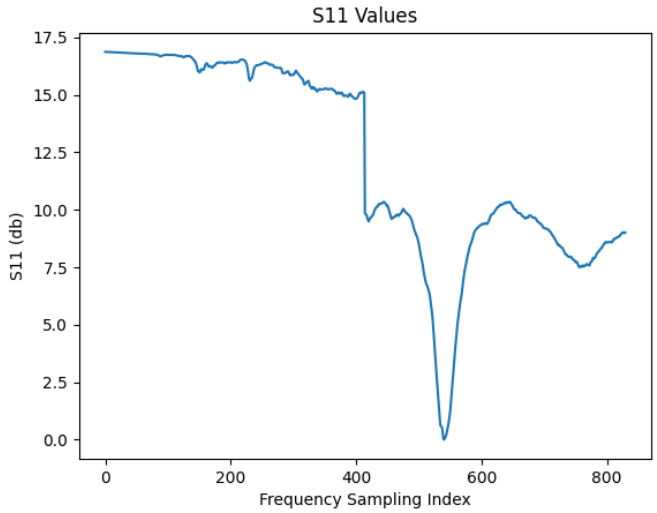

Figure 7. Example of input used for the DNN model: this S11 plot represents a flattened sequence of 830 values of S11 sampled from the selected frequency bands, 0-1 GHz and 2-3 GHz.

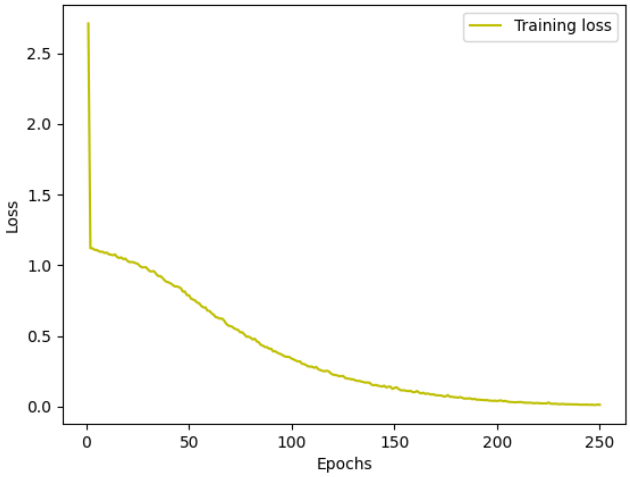

Figure 8. Training loss of DNN on Filtered Data.

sources. The proposed method is based on the SHAP technique that helps reduce feature space. This approach removes the most salient input features in identifying water sources under the hypothesis that some of the features that are useful in identifying water sources may not be necessary to identify Pb levels in the water. Our hypothesis rested on the assumption that these features may be independent, or at least sufficiently independent to allow us to remove input features responsible for determining the water source while keeping those that are responsible mainly for Pb level classification. The simulation results confirmed our hypothesis, resulting in a reduction in model size of 97% and an 89.73% reduction in the total number of parameters, while losing less than 1% accuracy

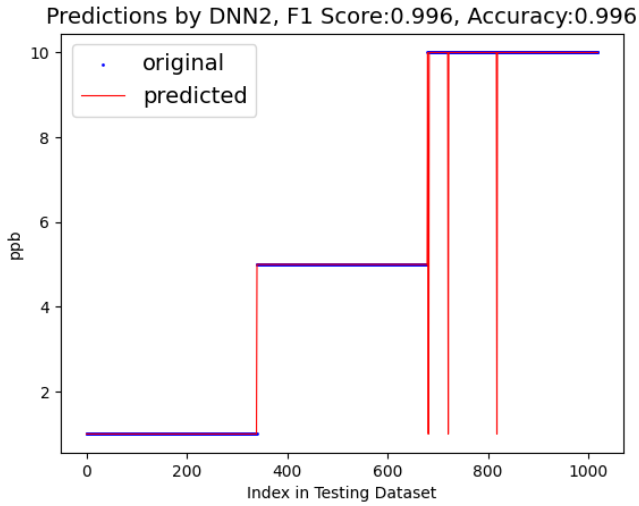

Figure 9. Testing Accuracy of DNN on Filtered Data.

(99.61%).

In our future work, we plan to further investigate the SHAP analysis by looking at finer frequency sub-bands for more precise feature extractions. Additionally, we will conduct experiments with water measurements generated from a larger number of water sources. More experiments will be conducted to determine necessary specifications for model performance, such as the minimum quantity of water sources needed for certain regional applications.

REFERENCES

[1] S. Oh, I. Hossen, J. Luglio, G. A. Justin, J.E. Richie, H. Medeiros,and C.H. Lee, "On-site/in situ continuous detecting ppb-Level metal ions in drinking water using block loop-gap resonators and machine learning," *IEEE Trans. on Instrumentation and Measurement*, vol. 70, pp. 1-9, 2021.

[2] N.S. Shekhawat, S. Oh, C. Ababei, C.H. Lee, and D.H. Ye, "Machine learning models for prediction of metal ion concentrations in drinking water," *IEEE Int. Electro/Information Technology Conference (EIT)*, Eau Claire, WI, May 30 - June 1, 2024.

[3] Lundberg, Scott M. and Lee, Su-In, "A unified approach to interpreting model predictions," *International Conference on Neural Information Processing Systems*, pp. 4768–4777, 2017.

[4] Wilson E. Marcílio and Danilo M. Eler, "From explanations to feature selection: assessing SHAP values as feature selection mechanism," *Conference on Graphics, Patterns and Images (SIBGRAPI)*, pp. 340-347, 2020.

[5] L. Grbi, Ivana Luin, L. Kranjevi, Sinia Drueta, "A Machine Learning-Based Algorithm for Water Network Contamination Source Localization," *Italian National Conference on Sensors*, 2020.

[6] Seong-Yun Hwang, Byung-Woong Choi, Jong-hwan Park, Dong-seok Shin, Hyeonsu Chung, Mi-Sun Son, Chae-Hong Lim, Hyeon-Mi Chae, D. Ha, K. Jung, "Evaluating Statistical Machine Learning Algorithms for Classifying Dominant Algae in Juam Lake and Tamjin Lake, Republic of Korea," *Water*, pp. 340-347, 2023.

[7] Seok Hyun Ahn, Do Hwan Jeong, MoonSu Kim, Tae Kwon Lee, Hyun-Koo Kim, "Prediction of groundwater quality index to assess suitability for drinking purpose using averaged neural network and geospatial analysis," *Elsevier BV*, 2023.

[8] Enas E. Hussein, Bilel Zerouali, N. Bailek, A. Derdour, S. Ghoneim, C. Santos, Mofreh A. Hashim, "Harnessing Explainable AI for Sustainable Agriculture: SHAP-Based Feature Selection in Multi-Model Evaluation of Irrigation Water Quality Indices," *Water*, 2024.

[9] Annisa Permata Sari, Billy, Denanda Aufadlan Tsaqif, B. Sartono, Aulia Rizki Firdawanti, "Classification of Drinking Water Source Suitability in West Java Using XGBoost and Cluster Analysis Based on SHAP Values," *Indonesian Journal of Statistics and Its Applications*, 2024.

[10] Vinita Sangwan, Rashmi Bhardwaj, "Machine learning framework for predicting water quality classification," *Water Practice & Technology*, 2024.