Combining Explainable Artificial Intelligence (XAI) with Blockchain towards Trustworthy Data-driven Policies

Konstantinos Mavrogiorgos[1], Shlomit Gur[2], Nikolaos Kalantzis[3], Konstantinos Tzelaptsis[3], Xanthi S. Papageorgiou[3], Andreas Karabetian[1], Georgios Manias[1], Argyro Mavrogiorgou[1], Dimosthenis Kyriazis[1], Celia Parralejo Cano[4]
[1]*University of Piraeus*, Piraeus, Greece
[2]*IBM Research,* Haifa, Israel
[3]*Ubitech Limited,* Limassol, Cyprus
[4]*Diputación de Badajoz,* Badajoz, Spain

komav@unipi.gr, Shlomit.Gur@ibm.com, nkalantzis@ubitech.eu, ktzelaptsis@ubitech.eu, xpapageorgiou@ubitech.eu, adreaskar@unipi.gr, gmanias@unipi.gr, margy@unipi.gr, dimos@unipi.gr, cparralejo@dip-badajoz.es

***Abstract*—Data-driven policy making is considered one of the most important aspects of decision-making systems and, as such, a lot of research is being carried out to provide the right tools and techniques to support it and make it more efficient. Towards this direction, the utilization of Artificial Intelligence (AI) approaches is widely being adopted to enhance current policies, create new ones and provide more accurate results. However, the way that those results are generated is often considered a black box, since the AI models are complicated, and the policy makers are not able to understand the reasoning behind the extracted results. To this context, Explainable AI (XAI) models have made their way into decision-making systems to address the aforementioned challenge. XAI, as its name suggests, provides the explanations with regards to the way that an AI model generates its results, thus enabling the end users to better understand its output. Nevertheless, XAI is not able to ensure the trustworthiness of an AI model's output on its own, since the provided explanations, as well as the output of an AI model could potentially be malformed by a third party. To this end, in this paper the authors propose an approach for data-driven policy making that combines XAI with blockchain technology in order to not only provide explanations for the output of an AI model, but also ensure this output, and the corresponding explanations, are reliable.**

Keywords—explainable artificial intelligence, blockchain, machine learning, policy making

I. Introduction

According to a very recent survey [1], the AI market is around 244 billion US dollars in 2025 and is expected to reach and exceed the astonishing amount of 800 billion US dollars by 2030, thus highlighting the adaptation of AI technologies in different domains and for a wide range of tasks. With regards to the policy making domain, AI models are utilized to foster the whole process and provide data-driven insights. The utilization of such models is expected to grow exponentially, thus it is imperative that their output should be explained to the policy makers and not remain a black box [2]. Towards this direction, XAI approaches are usually adopted to tackle this challenge and provide insights with regards to the output of AI models.

Even though XAI approaches are a significant step towards trustworthy data-driven policies, they are not capable of ensuring the reliability of neither the AI models output, nor the integrity of the explanations that they provide. Systems that incorporate AI models to support policy makers, are usually deployed on a cloud infrastructure to which the policy makers have access over the Internet, through the corresponding User Interface (UI). This poses a great risk to the trustworthiness of the models' output and the corresponding explanations (given that XAI is also adopted), since they can be malformed by any third-party. Ensuring the trustworthiness of the predictions and the explanations could be achieved by utilizing technologies like blockchain, which has already been used for validating data integrity [3].

To this end, in this paper the authors propose an approach for combining XAI with blockchain technology towards the formulation of trustworthy data-driven policies. The approach is incorporated into a developed policy making platform with a corresponding UI and is validated through specific use cases with regards to water management policies, thus highlighting its effectiveness and adaptability to real-life scenarios.

The rest of this paper is organized as follows. Section I briefly describes the problem statement and the proposed approach. In Section II, a literature review is conducted, focused on the utilization of XAI and blockchain technology in the context of policy making. In Section III the proposed approach is analyzed along with the corresponding architecture, whilst in Section IV results from the deployment of the proposed approach in real-life use cases with regards to water management policy making are presented. Finally, Section V summarizes this research work and provides insights regarding potential future steps.

II. Literature Review

A. Explainable Artificial Intelligence (XAI)

XAI consists of a set of algorithms and techniques that aim to support humans in better understanding and thus, trusting Machine Learning (ML) models. XAI has gained popularity over the last year, being applied in a plethora of different use cases, including policy making. For example, the authors in [4] develop an explainable and regulatory compliant approach for

public policymaking in the context of smart parking and infrastructure maintenance. Similarly, XAI can also be adopted in developing evidence-based policies in finance [5], health [6] and education [7]. The abovementioned highlight the applicability of XAI approaches in the formulation of new policies and the enrichment of existing ones.

XAI approaches can be classified in one of the categories of data explainability or model explainability [8]. With regards to data explainability, this can be achieved by performing exploratory analysis and visualizations on the data, as well as dimensionality reduction techniques. As for model explainability, there exist the white box models that are self-explanatory, such as the Decision Tree (DT) algorithm or linear models [9], and black box models that are inherently complex and difficult to explain, such as Support Vector Machine (SVM) [10]. As for the latter, those can be explained either by featured-based techniques or by example-based techniques.

Feature-based techniques include feature importance, Partial Dependence Plots (PDPs), Individual Conditional Expectation (ICE) plots, Accumulated Local Effects (ALEs), Global Surrogates, Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP). More specifically, feature importance is a type of measure that identifies the way that the change of a feature's value affects a model's outcome with comparison to other features [11]. Similar to feature importance, PDPs are graphs that allow end users to visually inspect and understand possible complicated connections between feature values and a model's output [12]. Another type of plots that are used in XAI are ICE plots. The key difference between PDPs and ICE plots is the fact that the latter focus on individual data instances, while PDPs identify the average effect that data instances have on the corresponding predictions [13]. ALEs are also a type of plots used for explainability that are preferred over PDPs when it comes to highly correlated data [14]. As for Global Surrogates, those are interpretable models that are built across an entire data domain in order to provide a high-level view of the given task that the corresponding black box model aims to solve [15]. LIME is a type of local surrogate model that locally approximates a model with an interpretable one, thus explaining the corresponding predictions [16]. On the other hand, SHAP assigns values to each feature to indicate its contribution to a model's output. As a result, it can be used to provide both global and local insights on a given model [17].

With regards to example-based techniques, those mainly consist of Anchors, Counterfactuals, Contrastive Explanations Method (CEM), Kernel and Tree SHAP, as well as Integrated Gradients (IG). Anchors focuses on producing conditions (i.e., anchors) for explaining a model's prediction for a specific data instance, thus providing local explanations [18]. Counterfactual Explanations describe an outcome by also considering alternative events that could potentially occur. They provide explanations regarding the way that an outcome of an automated decision could be changed [19]. CEM is another type of XAI approach for local explanations that focuses mostly on classification models and features that are preferable (i.e., pertinent positives) and unwanted (i.e., pertinent negatives) [20]. Tree and Kernel SHAP are both variations of the SHAP. Tree SHAP is faster but can only be used in tree-based algorithms, whilst Kernel SHAP is model-agnostic and can explain anomalies by using Shapley values [21]. Lastly, IG computes the integral of gradients along the path from a baseline input to the actual one in order to calculate the importance of each feature to a model's prediction [22].

B. Blockchain Technology

Blockchain is a shared and decentralized ledger that is immutable and is based on hashing [23]. This means that any transactions that are recorded cannot be altered without changing all the other subsequent blocks and the consensus of the corresponding network. Blockchain technology is the foundation of all the cryptocurrency-related platforms and platforms that utilize smart contracts. However, in recent years blockchain has found its way into different domains that are not associated with cryptocurrencies [24].

In deeper detail, blockchain is widely used in applications in financial services. Indicative examples include fraud prevention, credit score calculations, transactions monitoring and identity management and management of digital assets [25]. Furthermore, blockchain can be applied to the education domain in terms of data protection and scalability, since it allows the creation of decentralized education ecosystems while it supports credential issuance and management. Moreover, through blockchain, it is possible to digitalize and decentralize educational certifications, as well as manage learning records [26]. With regards to the healthcare domain, blockchain has been adopted in the corresponding systems in order to facilitate the sharing of records and images and support the monitoring of patients through Internet of Things (i.e., IoT) devices, whilst ensuring the secure transfer of the corresponding information [27]. It is also worth mentioning the exploitation of blockchain in digital data marketplaces [28], [29] to support decentralized and programmable data assets' trading and pricing.

As for the environment, blockchain has also been adopted in related use cases such as waste management and water management. More specifically, blockchain can be combined with several waste management methods to encourage efficiency and accountability when handling plastic and electronic waste [30]. What is more, blockchain can support the recycling process in smart cities, by using digital asset tokens that provide traceability to the generated waste and the way that they are being processed, thus ensuring the protection of the environment for pollution [31]. As for water management, blockchain has been used to monitor shared water resources and ensure that they are effectively coordinated between different communities [32]. Blockchain-related approaches have also been adopted to support the management of both agricultural and urban water and ensure the corresponding quality, as well as contribute to achieving Sustainable Development Goals (SDGs) related to hydrology [33].

As for combining XAI with blockchain to ensure trustworthiness, there exist several approaches in the literature, focusing on domains such as finance [34] and healthcare [35]. However, there seems to be a lack of blockchain approaches that utilize state-of-the-art XAI methods to ensure the trustworthiness of generated predictions and explanations, especially in environment-related domain such as water management. To this end, the authors of this paper propose a

platform that utilizes self-explaining Recurrent Neural Networks (RNNs) [36] to make predictions about different aspects of water management and blockchain technology to ensure the trustworthiness of the said predictions and the corresponding explanations.

III. Proposed Approach

The architecture of the proposed approached is depicted in Fig. 1 and analyzed below.

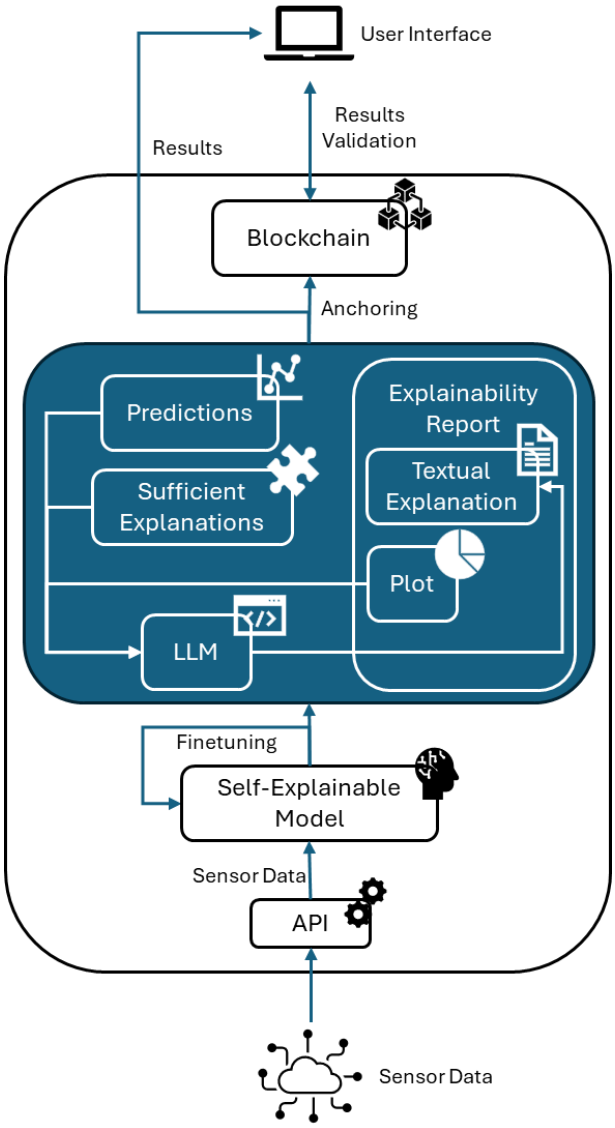

Fig. 1. Proposed approach architecture

More specifically, the sensor data coming from the appropriate smart devices (e.g., sensors located inside smart bins) are gathered through the corresponding Application Programming Interface (API) and are being preprocessed and cleaned to ensure their reliability, exploiting the approaches of [37] and [38]. Afterwards, the data are used to train the self-explainable model while they are also being used to finetune it, whenever new data are available. In the context of this manuscript, the self-explainable model is a Long Short-Term Memory (LSTM) Recurrent Neural Network (RNN) [39] that consists of two (2) LSTM layers and a third component called "explanation component", which is responsible for providing sufficient explanations, as presented in [40]. The output of the model (i.e., the predictions, the sufficient explanations and a plot (i.e., heatmap) showcasing the explanations) are also fed into a Large Language Model (LLM) to generate a textual explanation of them that is easily understandable by the end users. The explainability report (i.e., the textual explanation and the heatmap), as well as the predictions generated by the model are then anchored to the blockchain. Those are also available to the end users through the corresponding UI, through which the end users can validate the illustrated results with the results that were previously anchored to the blockchain. If the validation fails, this means that someone has altered the predictions and/or the explainability report that are available through the UI, thus they should not be trusted. On the other hand, if the validation succeeds, the end users can trust the illustrated results, since those are identical to the ones that are anchored to the blockchain, meaning that they have not been tampered with by any third-party.

A. *Self-explainable Neural Network*

As mentioned above, the model used for generating the predictions is an LSTM RNN which consists of two (2) LSTM layers and a third component called "explanation component", which is responsible for providing sufficient explanations. Both the input and the output of the model are based on the needs of the corresponding use case. More specifically, in order to generate forecasts, the LSTM RNN has a predefined input size, based on the number of features present in the dataset, as well as the number of past records needed to perform the corresponding forecast. As for the output size, this is also related to the number of features that need to be predicted, as well as the forecast horizon (i.e., the future time steps for which the predictions are made). The "explanation component" makes the model self-explainable since it enables it to generate the explanations along with the predictions. This particular XAI method is ante-hoc, since the model is intrinsically explainable, compared to other XAI approaches that are post-hoc, meaning that they are applied after the model training is performed. It is also worth mentioning that common local post-hoc XAI methods, such as LIME and SHAP, have a significant drawback when compared to the XAI method that is used in this proposed approach. This drawback lies in their implicit assumption of near-linear behavior of the model around the analyzed input [41], which may not apply to highly non-linear contemporary deep learning (i.e., DL) models like the LSTMs that are trained in this manuscript.

The predictions and the explanations are incorporated into a heatmap and are also fed into an LLM to provide a textual explanation of the heatmap and the corresponding predictions so that the end users can easily understand them. The LLM that is used is the Microsoft Phi-3-Mini-4K-Instruct [42], which is finetuned specifically for memory/compute constrained environments. The generated heatmap and the textual explanation are then incorporated into the explainability report, which, along with the generated predictions, are included into a JavaScript Object Notation (JSON) file that is both anchored to the blockchain and returned to the end users and visualized through the corresponding UI.

B. Blockchain

To ensure the validity, and immutability of the model's output with a transparent way, a data anchoring mechanism is leveraged using blockchain technology, as mentioned above. Specifically, the anchoring process involves storing a cryptographic hash as a digital fingerprint of the model's output on the blockchain. This approach allows verification of the data integrity without exposing the content itself. The blockchain infrastructure adopted in this work is Hyperledger Fabric (HLF). HLF is an open-source, permissioned Distributed Ledger Technology (DLT) tailored for enterprise applications. Its modular and highly configurable architecture supports Smart Contracts, referred to as "chain code" within the HLF ecosystem [43].

Two (2) categories of data are relevant in this context: on-chain and off-chain. On-chain data is stored directly within the blockchain ledger, while off-chain data resides in external storage systems, such as databases or file servers. In the off-chain approach, a reference which could be a hash (e.g., SHA-256) is written to the blockchain, linking it to the externally stored data. The off-chain model is particularly advantageous in terms of scalability and resource efficiency, as it minimizes the volume of data stored on-chain. Despite being stored externally, the integrity of the off-chain data can be reliably verified by comparing it with its corresponding on-chain hash. If the recalculated hash of the retrieved off-chain report does not match with the blockchain record, it indicates that the data has been tampered. In the case of the model's output, this mechanism ensures their authenticity and provenance. Every time an explainability report and a prediction are generated, their hash is recorded on the blockchain. This immutable hash then serves as a verification anchor, allowing any end user to confirm that the report has not been malformed since its original creation.

IV. Experimental Results

A. Performance Evaluation

The proposed approach was tested and validated in two (2) use cases related to water management, in the context of the European-funded project AI4Gov [44]. The first use case is related to drinking water management and the goal is to predict the quality of the water based on data received from sensors that are installed in Drinking Water Treatment Plants (DWTPs). The data consist of six (6) features, namely observation date, entity ID, pH, chlorides, water level and instant output quantity. The observation date corresponds to the date that the corresponding measurement refers to, and entity ID refers to the ID of the treatment plant, since the data originate from sensors installed in three (3) different treatment plants. The other four (4) features are related to the quality of the water and are the ones that need to be predicted by the RNN. The data are timeseries data ranging from 2022 up to 2023, whilst the frequency is one (1) hour. An overview of the dataset is shown in Table I, including the number of values per feature (i.e., Count), the mean and the standard deviation (i.e., Std) of each feature, as well as the minimum (i.e., Min) and the maximum (i.e., Max) value of each one.

TABLE I. Description of drinking water management dataset

Feature	Count	Mean	Min	Max	Std
observation date	20235	-	01/06/2022	08/03/2023	-
entity ID	3	-	-	-	-
pH	20235	7.65	7.03	8.29	0.30
chlorides	20235	0.65	0.01	5.00	0.32
water level	20235	91.23	0.07	115.64	7.88
instant output quantity	20235	13.07	0.60	50.40	12.68

With regards to the second use case, this is related to sewage water management and the goal is to predict the energy consumption of Wastewater Treatment Plants (WWTPs) based on data received from sensors that are installed in the WWTPs. The data also consist of six (6) features, namely observation date, entity ID, consumed energy, reactive energy consumed, total active power and total reactive power. The observation date corresponds to the date that the corresponding measurement refers to and entity ID refers to the ID of the treatment plant, since the data originate from sensors installed in three (3) WWTPs. The other four (4) features are related to energy consumption of the WWTPs and are the ones that need to be predicted by the RNN. The data are timeseries data ranging from 2019 up to 2020, whilst the frequency is one (1) day. An overview of the dataset is shown in Table II, including the number of values (i.e., Count), the mean and the standard deviation (i.e., Std), as well as the minimum (i.e., Min) and the maximum (i.e., Max) value of each feature.

TABLE II. Description of Sewage water management dataset

Feature	Count	Mean	Min	Max	Std
observation date	522	-	05/12/2019	01/06/2020	-
entity ID	3	-	-	-	-
consumed energy	522	307284	7661	859364	355311
reactive energy consumed	522	297595	6939	784176	313010
total active power	522	3406	2	20118	4171
total reactive power	522	3735	2	16865	4080

In the context of this manuscript, several self-explainable LSTMs have been trained, one for each DWTP and WWTP. All LSTMs have similar architectures, however, there exist some differences that are analyzed below.

In deeper detail, regarding the drinking water use case, the input of the corresponding LSTM is twenty-four (24) past observations, and the output is the predicted values per each of the four (4) features (i.e., pH, chlorides, water level and instant output quantity) for a time horizon of six (6) hours into the

future. Apart from the predicted values, the LSTM also provides the corresponding sufficient explanations that describe which of the input values, including the corresponding time points, directly affect the provided predictions.

As for the sewage water use case, the input of the LSTM is seven (7) past observations, and the output is the predicted values per each of the four (4) features (i.e., pH, chlorides, water level and instant output quantity) for a time horizon of one (1) day into the future. Apart from the predicted values, the LSTM also provides the corresponding sufficient explanations that describe which of the input values, including the corresponding time points, directly affect the provided predictions.

There existed two (2) options for training the corresponding LSTMs for both the use cases. The first approach was to train a global model per use case based on the three (3) entities that exist in each use case. The other option was to train a separate LSTM per entity. In order to select the best approach, rigorous experiments were conducted so that the best approach was selected in terms of Mean Square Error (MSE). A summarization of the results of the testing datasets for the drinking water and the sewage water are depicted in Table III and Table IV respectively. All the LSTM architectures were implemented using PyTorch [45] and in each experiment different learning rate and gamma value were used. The gamma value is used in the "explanation component" and acts as a weight for the regularization term that influences the sparsity of the explanations. As for the optimizer that was utilized, this was the Adam optimizer, since it is a widely used one in LSTMs for time series forecasting [46].

TABLE III. CALCULATED MSE FOR DRINKING WATER FORECASTING MODELS

Entity ID	Learning rate	Gamma value	Epochs	MSE
ALL	1.00E-04	1.00E-06	7	0.3839
	1.00E-04	1.00E-07	4	0.3854
	1.00E-04	1.00E-08	3	0.3877
	1.00E-05	1.00E-06	1	0.3984
	1.00E-05	1.00E-07	3	0.3926
	1.00E-05	1.00E-08	1	0.3967
DWTP1	1.00E-04	1.00E-06	3	0.0350
	1.00E-04	1.00E-07	5	0.0061
	1.00E-04	1.00E-08	1	0.0146
	1.00E-05	1.00E-06	13	0.0018
	1.00E-05	1.00E-07	26	0.0042
	1.00E-05	1.00E-08	24	0.0034
DWTP2	1.00E-04	1.00E-06	1	0.1940
	1.00E-04	1.00E-07	4	0.1434
	1.00E-04	1.00E-08	1	0.3096
	1.00E-05	1.00E-06	10	0.0707
	1.00E-05	1.00E-07	38	0.0306
	1.00E-05	1.00E-08	11	0.0343
DWTP3	1.00E-04	1.00E-06	1	0.7485
	1.00E-04	1.00E-07	2	0.5751
	1.00E-04	1.00E-08	4	0.4493
	1.00E-05	1.00E-06	15	0.5492
	1.00E-05	1.00E-07	14	0.3919
	1.00E-05	1.00E-08	89	0.0303

TABLE IV. CALCULATED MSE FOR SEWAGE WATER FORECASTING MODELS

Entity ID	Learning rate	Gamma value	Epochs	MSE
ALL	1.00E-04	1.00E-06	9	0.7899
	1.00E-04	1.00E-07	3	0.7919
	1.00E-04	1.00E-08	1	0.8111
	1.00E-05	1.00E-06	14	0.7861
	1.00E-05	1.00E-07	2	0.7915
	1.00E-05	1.00E-08	4	0.7982
WWTP1	1.00E-04	1.00E-06	17	1.4561
	1.00E-04	1.00E-07	1	1.3187
	1.00E-04	1.00E-08	5	0.1806
	1.00E-05	1.00E-06	7	1.6138
	1.00E-05	1.00E-07	1	1.4599
	1.00E-05	1.00E-08	3	1.8775
WWTP2	1.00E-04	1.00E-06	3	0.8389
	1.00E-04	1.00E-07	7	0.8148
	1.00E-04	1.00E-08	3	0.9392
	1.00E-05	1.00E-06	7	1.0009
	1.00E-05	1.00E-07	4	0.9704
	1.00E-05	1.00E-08	2	0.8957
WWTP3	1.00E-04	1.00E-06	1	0.0979
	1.00E-04	1.00E-07	9	0.0862
	1.00E-04	1.00E-08	3	0.8596
	1.00E-05	1.00E-06	17	0.8474
	1.00E-05	1.00E-07	1	0.8334
	1.00E-05	1.00E-08	1	0.8453

According to the results of the experiments, a global model is underperforming when compared to custom models that are trained on each DWTP and WWTP. As a result, based on the corresponding MSE values, the models that are highlighted with grey color in Table III and Table IV were selected and deployed in the context of the proposed approach. A summarization of the developed models is depicted in Table V.

TABLE V. SELECTED MODEL PER DWTP AND WWTP

Entity ID	Learning rate	Gamma value	Epochs	MSE

DWTP1	1.00E-05	1.00E-06	13	0.0018
DWTP2	1.00E-05	1.00E-07	38	0.0306
DWTP3	1.00E-05	1.00E-08	89	0.0303
WWTP1	1.00E-04	1.00E-08	5	0.1806
WWTP2	1.00E-04	1.00E-07	7	0.8148
WWTP3	1.00E-04	1.00E-07	9	0.0862

B. *Functional Evaluation*

As mentioned above, the proposed approach and the trained models have also been validated by the corresponding end users in the context of the AI4Gov project in terms of the quality and the reliability of predictions and explanations, as well as the usability of the UI. More specifically, a validation workshop was carried out, where the end users tested the proposed approach and answered a user experience questionnaire (i.e., UEQ). The results of the UEQ showcased an overall positive user experience both with regards to the users' trust in the predictions provided and the usability of the proposed approach.

The end users are public workers in the water management department of the municipality of Badajoz in Spain. More specifically, the end users can gain access to the UI through a web browser. There, they are able to upload their data based on which the corresponding trained model will provide predictions and explanations. Regarding the drinking water use case, the end users select from a given map the DWTP for which they would like to predict the future quality of the drinking water, as shown in Fig. 2. Similarly, as for the sewage water, the end users select from a given map the WWTP for which they would like to predict the future energy consumption.

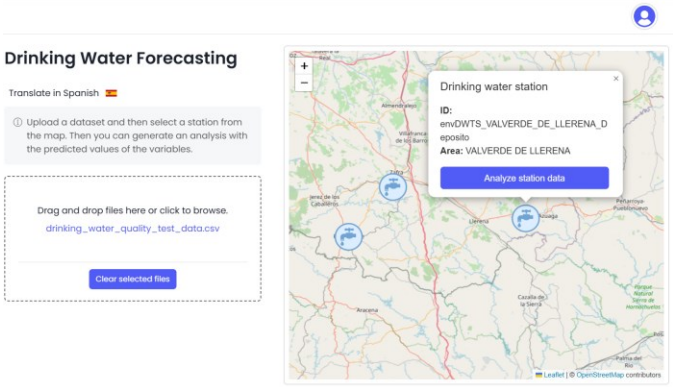

Fig. 2. Selection of DWTP through the UI

After the model generates the predictions and the explainability report is also formulated, those are simultaneously anchored to the blockchain and returned to the UI. This process is the same for both the drinking water and the sewage water use cases. In deeper detail, the information that are illustrated in the UI are: (i) the predicted value per feature and (ii) the textual explanation that is generated by the LLM, as well as the sufficient explanations heatmap. An example of the above-mentioned related to the drinking water use case are depicted in Fig. 3 and Fig. 4 respectively.

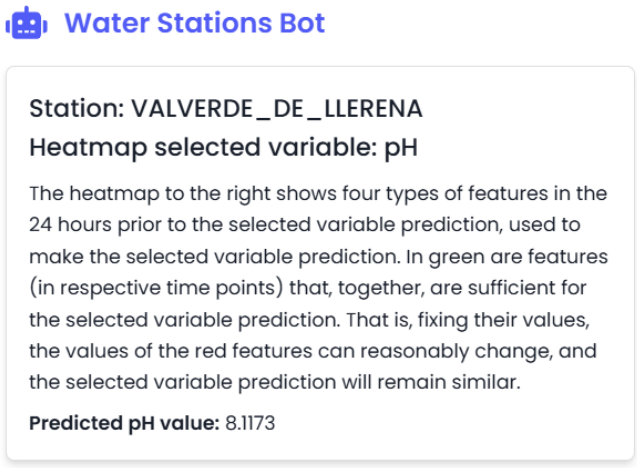

Fig. 3. Indicative example of textual explanation for the prediction of pH value

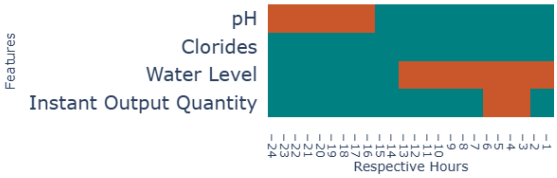

Fig. 4. Indicative example of sufficient explanations heatmap for the prediction of pH value

In the example shown above, the predicted value for the pH is depicted in bold and is accompanied by the textual explanation that is related to the heatmap that is shown to the end users. The heatmap visualizes the output of the explanation component of the LSTM RNN and, in the context of the drinking water use case, it shows four (4) types of features in the 24 hours prior to the pH prediction, used to make the pH prediction. In green are the features (in respective time points) that, together, are sufficient for the pH prediction. That is, fixing their values, the values of the red features can reasonably change, and the pH prediction will remain similar. The end users are able to view the explainability reports and the predictions for all of the four (4) variables, whilst the same functionalities are available for the sewage water use case.

Finally, the end users can validate the predictions and the explainability report based on the JSON data that are anchored to the blockchain, by pressing the corresponding button. If the validation of the prediction and the explainability report is successful, the end user is notified that those have not been tampered by any third-party entity. Otherwise, if the validation fails, this means that the results have been altered and thus, they should be discarded. An indicative example of a successful validation through the blockchain is shown in Fig. 5, where the end users are notified through the UI.

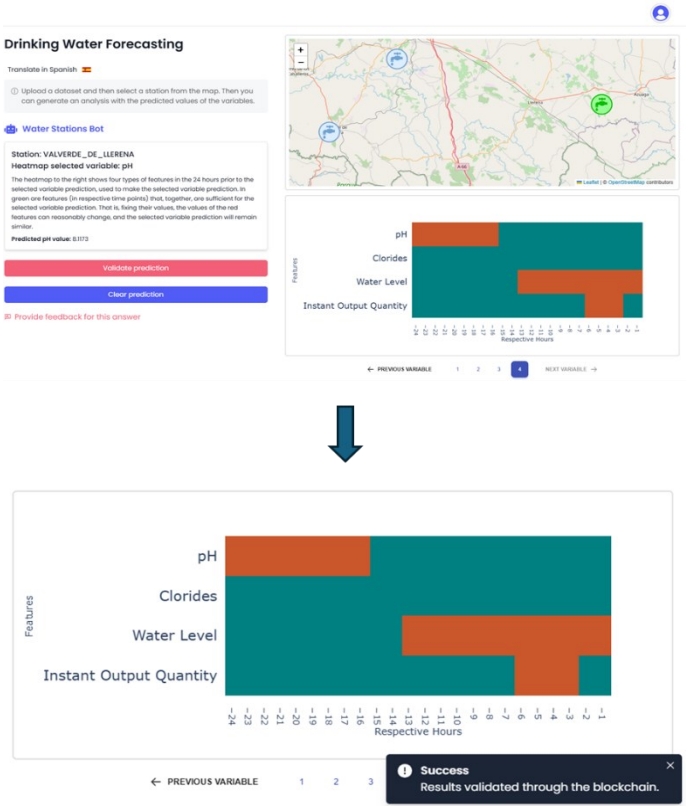

Fig. 5. Successful validation of predictions and explainability report through the blockchain

Overall, the proposed approach combines state-of-the-art methods and technologies, such as self-explainable neural networks, LLMs and blockchain, to provide accurate, explainable and interpretable predictions whilst ensuring their reliability. However, it should be noted that further improvements could be made, especially with regards to the optimizer that is being utilized in the LSTMs, since the selection of an optimizer can affect a model in terms of bias [47]. Moreover, in order to ensure the integrity of the trained models and the corresponding data, blockchain-based auditing mechanisms should also be integrated.

V. Conclusions

Overall, data-driven policy making is a key aspect of decision-making systems, thus numerous approaches have been proposed to make it more efficient. Those approaches mainly consist of AI algorithms. However, most of the time those algorithms are complicated and not easy to be interpreted by the end users. This highlights the need for XAI approaches that turn a black box AI model into an explainable one. However, the trustworthiness of the said model is not guaranteed since its predictions, as well as its explanations can potentially be altered by a third-party. To this end, in this paper the authors propose a method for combining a self-explainable neural network with an LLM and blockchain technology, in order to provide both explainable and trustworthy predictions. The proposed approach is validated in two (2) use case scenarios related to water management, thus being one of the first that combines such technologies and methods in the context of this domain. The generated results, as well as the feedback received from the corresponding experts, highlighted not only the accuracy of the predictions, but also the quality of the provided explanations and the trustworthiness of both the predictions and the explanations.

With regards to any potential future steps, the authors aim to experiment on other domains, such as waste management, focusing on the evaluation of the self-explained neural network architecture and the quality of the provided explanations on larger datasets. In this context, the authors also aim to experiment with other optimizers to assess them in terms of the bias that they might introduce to the model. What is more, it would be interesting to implement and review other types of self-explainable RNNs, such as Gated Recurrent Units (GRUs), with regards to the quality of the provided explanations. As for the integrity of the trained models and the corresponding data, blockchain-based auditing mechanisms could also be adopted. Moreover, additional research could be conducted in order to evaluate the utilization of blockchain technology in this type of use case scenarios and investigate potential points of improvement in terms of performance and computational resources, especially when it comes to supporting large number of public policy making transactions. To this end, a comparative study of the proposed approach with other similar approaches from other domains would also be highly beneficial. Lastly, further research will take place to enhance the overall environmental sustainability of the developed system and its underlying resources [48].

Acknowledgment

The research leading to the results presented in this paper has received funding from the European Union's funded Project AI4Gov under grant agreement no 101094905.

References

[1] Artificial intelligence (AI) worldwide - statistics & facts, Mar 18, 2025 [Online]. Available: https://www.statista.com/topics/3104/artificial-intelligence-ai-worldwide/#topicOverview

[2] T. Papadakis, I. T. Christou, C. Ipektsidis, J. Soldatos, A. Amicone, Explainable and transparent artificial intelligence for public policymaking. Data & Policy, 6, e10, 2024.

[3] D. Marijan, C. Lal, Blockchain verification and validation: Techniques, challenges, and research directions. Computer Science Review, 45, 100492, 2022.

[4] T. Papadakis, I. T. Christou, C. Ipektsidis, J. Soldatos, A. Amicone. Explainable and transparent artificial intelligence for public policymaking. Data & Policy, 6, p.e10, 2024.

[5] M. S. De Carvalho, G. L. Da Silva. Inside the black box: using Explainable AI to improve Evidence-Based Policies. In 2021 IEEE 23rd Conference on Business Informatics (CBI) (Vol. 2, pp. 57-64), September 2021.

[6] J. Gerlings, M. S. Jensen, A. Shollo, Explainable AI, but explainable to whom? An exploratory case study of xAI in healthcare. In Handbook of Artificial Intelligence in Healthcare: Vol 2: Practicalities and Prospects (pp. 169-198), 2021.

[7] H. Khosravi, et, al. Explainable artificial intelligence in education. Computers and education: artificial intelligence, 3, 100074, 2022.

[8] R. Dwivedi, D. Dave, H. Naik, et. al. Explainable AI (XAI): Core ideas, techniques, and solutions. ACM Computing Surveys, 55(9), 1-33, 2023.

[9] A. Blanco-Justicia, J. Domingo-Ferrer. Machine learning explainability through comprehensible decision trees. In Machine Learning and Knowledge Extraction: Third IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2019, Proceedings 3 (pp. 15-26), 2019.

[10] V. Hassija, A. Chamola, et. al. Interpreting black-box models: a review on explainable artificial intelligence. Cognitive Computation, 16(1), 45-74, 2024.

[11] K. Främling. Feature importance versus feature influence and what it signifies for explainable AI. In World Conference on Explainable Artificial Intelligence (pp. 241-259), July 2023.

[12] M. Muschalik, F. Fumagalli, R. Jagtani, et. al. iPDP: on partial dependence plots in dynamic modeling scenarios. In World Conference on Explainable Artificial Intelligence (pp. 177-194), July 2023.

[13] Y. Wang. A comparative analysis of model agnostic techniques for explainable artificial intelligence. Research Reports on Computer Science, 25-33, 2024.

[14] J. M. Darias, B. Díaz-Agudo, J. A. Recio-Garcia. A Systematic Review on Model-agnostic XAI Libraries. In ICCBR workshops (pp. 28-39), September 2021.

[15] E. Mariotti, A. Sivaprasad, J. M. Moral. Beyond prediction similarity: ShapGAP for evaluating faithful surrogate models in XAI. In World conference on explainable artificial intelligence (pp. 160-173), July 2023.

[16] V. Vimbi, N. Shaffi, M. Mahmud. Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer's disease detection. Brain Informatics, 11(1), 10, 2024.

[17] A. Dikshit, B. Pradhan. Interpretable and explainable AI (XAI) model for spatial drought prediction. Science of the Total Environment, 801, 149797, 2021.

[18] G. Casalino, G. Castellano, K. Kaczmarek-Majer, et. al. Explaining Predictions of Hypertension Disease through Anchors. In Proceedings of the First Workshop on Explainable Artificial Intelligence for the Medical Domain (EXPLIMED 2024) co-located with 27th European Conference on Artificial Intelligence (ECAI 2024), 2024.

[19] M. T. Keane, E. M. Kenny, E. Delaney, B. Smyth. If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques. arXiv preprint arXiv:2103.01035, 2021.

[20] J. Labaien, E. Zugasti, X. De Carlos. Contrastive explanations for a deep learning model on time-series data. In International Conference on Big Data Analytics and Knowledge Discovery (pp. 235-244), September 2020.

[21] K. Roshan, A. Zafar. Using kernel shap xai method to optimize the network anomaly detection model. In 2022 9th International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 74-80), March 2022.

[22] M. Sundararajan, A. Taly, Q. Yan. Axiomatic attribution for deep networks. In International conference on machine learning (pp. 3319-3328), July 2017.

[23] M. Di Pierro. What is the blockchain?. Computing in Science & Engineering, 19(5), 92-95, 2017.

[24] S. Nakamoto. Bitcoin: A peer-to-peer electronic cash system, 2008.

[25] M. Javaid, A. Haleem, R. P. Singh, et. al. A review of Blockchain Technology applications for financial services. BenchCouncil transactions on benchmarks, standards and evaluations, 2(3), 100073, 2022.

[26] A. El Koshiry, E. Eliwa, T. Abd El-Hafeez, M. Y. Shams. Unlocking the power of blockchain in education: An overview of innovations and outcomes. Blockchain: Research and Applications, 4(4), 100165, 2023.

[27] P. K. Ghosh, A. Chakraborty, M. Hasan, et. al. Blockchain application in healthcare systems: a review. Systems, 11(1), 38, 2023.

[28] A. Kiourtis, et. al. Data Marketplaces: Best Practices, Challenges, and Advancements for Embedded Finance. In 2023 19th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT) (pp. 533-540), June 2023.

[29] A. Mavrogiorgou, et. al. FAME: federated decentralized trusted data marketplace for embedded finance. In 2023 International Conference on Smart Applications, Communications and Networking (SmartNets) (pp. 1-6). July 2023.

[30] P. Jiang, L. Zhang, S. You, et. al. Blockchain technology applications in waste management: Overview, challenges and opportunities. Journal of Cleaner Production, 421, 138466, 2023.

[31] K. Bułkowska, M. Zielińska, M. Bułkowski. Implementation of blockchain technology in waste management. Energies, 16(23), 7742, 2023.

[32] H. Zeng, G. Dhiman, A. Sharma, et. al. An IoT and Blockchain‐based approach for the smart water management system in agriculture. Expert Systems, 40(4), e12892, 2023.

[33] T. K. Satilmisoglu, Y. Sermet, M. Kurt, I. Demir. Blockchain opportunities for water resources management: a comprehensive review. Sustainability, 16(6), 2403, 2024.

[34] J. Černevičienė, A. Kabašinskas. Explainable artificial intelligence (XAI) in finance: a systematic literature review. Artificial Intelligence Review, 57(8), 216, 2024.

[35] C. V. B. Murthy, M. L. Shri. Novel Architecture Integrating XAI with Blockchain and IoT Devices for Healthcare. In 2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE) (pp. 1-5), February 2024.

[36] S. Bassan, S. Gur, S. Zeltyn, K. Mavrogiorgos, R. Eliav, D. Kyriazis. Self-Explaining Neural Networks for Business Process Monitoring. arXiv preprint arXiv:2503.18067, 2025.

[37] K. Mavrogiorgos, et. al. Automated rule-based data cleaning using NLP. In 2022 32nd Conference of Open Innovations Association (FRUCT) (pp. 162-168), November 2022.

[38] A. Mavrogiorgou, et. al. Adjustable data cleaning towards extracting statistical information. In Public Health and Informatics (pp. 1013-1014), 2021.

[39] Y. Yu, et. al. A review of recurrent neural networks: LSTM cells and network architectures. Neural computation, 31(7), 1235-1270, 2019.

[40] S. Bassan, R. Eliav, S. Gur. Explain Yourself, Briefly! Self-Explaining Neural Networks with Concise Sufficient Reasons. arXiv preprint arXiv:2502.03391, 2025.

[41] C. K. Yeh, C. Y. Hsieh, et. al. On the (in) fidelity and sensitivity of explanations. Advances in neural information processing systems, 32, 2019.

[42] Huggingface - Microsoft Phi-3-Mini-4K-Instruct, Apr 3, 2025 [Online]. Available: https://huggingface.co/microsoft/Phi-3-mini-4k-instruct-gguf

[43] Hyperleder Fabric - A Blockchain Platform for the Enterprise, Apr 7, 2025 [Online]. Available: https://hyperledger-fabric.readthedocs.io/

[44] G. Manias, et. al. AI4Gov: Trusted AI for transparent public governance fostering democratic values. In 2023 19th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT) (pp. 548-555), June 2023.

[45] PyTorch – PyTorch Documentation, Apr 7, 2025 [Online]. Available: https://pytorch.org/docs/stable/index.html

[46] A. Makinde. Optimizing Time Series Forecasting: A Comparative Study of Adam and Nesterov Accelerated Gradient on LSTM and GRU networks Using Stock Market data. arXiv preprint arXiv:2410.01843, 2024.

[47] K. Mavrogiorgos, et. al. Bias in Machine Learning: A Literature Review. Applied Sciences, 14(19), 8860, 2024.

[48] A. Karabetian, et. al. An environmentally-sustainable dimensioning workbench towards dynamic resource allocation in cloud-computing environments. In 2022 13th international conference on information, intelligence, systems & applications (IISA) (pp. 1-4), July 2022.