XAI Tools in the Public Sector: A Case Study on Predicting Combined Sewer Overflows

Nicholas Maltbie
maltbind@mail.uc.edu
University of Cincinnati
Cincinnati, Ohio, USA

Nan Niu*
University of Cincinnati
Cincinnati, Ohio, USA
nan.niu@uc.edu

Matthew Van Doren
Metropolitan Sewer District of Greater Cincinnati
Cincinnati, Ohio, USA
matthew.vandoren@cincinnati-oh.gov

Reese Johnson
Metropolitan Sewer District of Greater Cincinnati
Cincinnati, Ohio, USA
reese.johnson@cincinnati-oh.gov

ABSTRACT

Artificial intelligence and deep learning are becoming increasingly prevalent in contemporary software solutions. Explainable artificial intelligence (XAI) tools attempt to address the black box nature of the deep learning models and make them more understandable to humans. In this work, we apply three state-of-the-art XAI tools in a real-world case study. Our study focuses on predicting combined sewer overflow events for a municipal wastewater treatment organization. Through a data driven inquiry, we collect both qualitative information via stakeholder interviews and quantitative measures. These help us assess the predictive accuracy of the XAI tools, as well as the simplicity, soundness, and insightfulness of the produced explanations. Our results not only show the varying degrees that the XAI tools meet the requirements, but also highlight that domain experts can draw new insights from complex explanations that may differ from their previous expectations.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; *Neural networks*; • **Software and its engineering** → *Process validation.*

KEYWORDS

explainability, AI, case study, goal-question-metric (GQM)

ACM Reference Format:
Nicholas Maltbie, Nan Niu, Matthew Van Doren, and Reese Johnson. 2021. XAI Tools in the Public Sector: A Case Study on Predicting Combined Sewer Overflows. In *Proceedings of the 29th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '21), August 23–28, 2021, Athens, Greece.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3468264.3468547

*Corresponding author.

ESEC/FSE '21, August 23–28, 2021, Athens, Greece

ACM ISBN 978-1-4503-8562-6/21/08…$15.00
https://doi.org/10.1145/3468264.3468547

1 INTRODUCTION

Artificial Intelligence (AI) has become so ubiquitous that many decisions nowadays in our daily life are shaped by it, e.g., news feed suggestions and shopping item recommendations. To put the flourish of AI into perspective, Adadi and Berrada [2] highlight the reports forecasting that from 2017 to 2021, the global investment on AI will increase from $12 billion US dollars to $52.2 billion, and the revenues from the AI enabled industries worldwide is expected to grow from $480 billion to $2.59 trillion.

The unstoppable penetration of AI also reaches into the public sector. For example, the European Commission envisions that AI could be used to serve citizens 24/7 in faster, more agile, and more accessible ways [22]. However, some public AI services have already shown harmful consequences. In the U.S., for instance, AI was used to allocate caregiver hours for people with disabilities, but dramatically lowered the number of hours in multiple cases without any explanation or meaningful opportunity to contest the decisions made by the proprietary algorithm [52].

Although AI mistakes are inevitable, the lack of *explainability* raises significant concerns from the citizens and public organizations about AI-based decision making's accountability, fairness, responsibility, and transparency. Explainable artificial intelligence (XAI) addresses these concerns by aiming to make AI more understandable to users, so as to increase the users' trust and reliance on the AI system.

Consequently, many XAI tools are built to automatically generate explanations from deep learning models. Deep learning models have been able to achieve near-human accuracy levels in various types of classification and prediction tasks including images, text, speech, and video data [7]. These deep learning models are often opaque in their nature and hence referred to as "black boxes" due to the difficulty in understanding how they operate [20]. An influential XAI tool is LIME (Local Interpretable Model-agnostic Explanations) proposed by Ribeiro and his colleagues [43]. LIME can approximate a black box model locally in the neighborhood of any prediction of interest. An illustration given by Ribeiro *et al.* [43] is that, once a model predicts a patient has the flu, LIME shows with relative weights that "sneeze" and "headache" contribute to this particular prediction whereas "no fatigue" is evidence against it. Identifying a few weighted features as in LIME is only one way to produce explanations. Others extract rules, visualize salience maps, or implement other methodologies [2].

Miller *et al.* [31] argue that most XAI researchers are currently building tools for themselves, rather than for the intended user. Even the seminal work on LIME scored the lowest on Miller *et al.*'s 'Data Driven' criteria, because Ribeiro *et al.* [43] constructed their own understanding of how people might evaluate explanations and recruited human subjects on Amazon Mechanical Turk to perform the experiments. Behavioral experiments conducted with lay persons are simplifications of, and therefore cannot replace, putting the explanation into the real-world application and letting the actual end user (typically a domain expert) test it [15].

To gain insights into application-grounded XAI evaluations, we conducted a case study on how public services might exploit deep learning to predict combined sewer overflows (CSOs). Combined sewer systems transport various sources of water from residential, industrial, and commercial customers as well as storm runoff. A problem with these systems is handling CSO events when the system is overwhelmed by surges of water and the combined sewer system is forced to discharge untreated water into the local environment. With infrastructures like sensor networks collecting real-time water flow data, along with the availability of contributing sources like the rainfall data, public sewer services show keen interests in deep learning techniques that are capable of offering high degrees of predictive accuracy as well as explainability.

This paper makes two main contributions. We perform a goal-question-metric analysis of explainability to quantitatively measure three state-of-the-art XAI tools, and we interview two domain experts to qualitatively assess the XAI results on 'Data Driven'[1] CSO predictions. Our study not only updates some commonly held beliefs about explainability, but also emphasizes the engineering considerations of incorporating explainability into the entire deep learning's development workflow. In what follows, we present background information in Section 2, detail our case design in Section 3, analyze the results in Section 4, discuss our work's implications in Section 5, and draw some concluding remarks in Section 6.

2 BACKGROUND

2.1 Explainable Artificial Intelligence (XAI) and Tools

In AI, the high level of difficulty for the system to provide a suitable explanation for how it arrived at an answer is referred to as the black box problem [2]. This difficulty is particularly prominent for deep learning models, because a deep neural network trained end-to-end can be as complex as an accurate explanation of why the model works [17]. The complexity can be illustrated by ResNet [21], which incorporates about 5×10^7 learned parameters and executes about 10^{10} floating point operations to classify a single image. XAI tries to demystify the black boxes as they begin making decisions previously entrusted to humans. Thus, explainability—the ability to interpret the inner workings or the logic of reasoning behind the decision making—helps to achieve an AI system's:

- **accountability**: justifying the decisions and actions,
- **fairness**: having impartial treatment and behavior,
- **responsibility**: answering for one's decisions and identifying errors or unexpected results, and
- **transparency**: describing, inspecting, and reproducing the mechanisms through which the decisions are made.

Adadi and Berrada [2] identified 17 XAI techniques by surveying 381 papers published between 2004 and 2018. According to the survey, most recent work done in the XAI field offers a *post-hoc*, *local* explanation. Because only a few models, such as linear regression or decision trees, are inherently interpretable, generating post-hoc explanations is necessary for complex models like deep neural networks. Post-hoc XAI tool can therefore be applied to any classifier or regressor that is appropriate for the application domain—even those that are yet to be proposed [43]. Local explanations justify why a single prediction was made, which are in contrast to global explanations trying to understand the entire reasoning leading to all possible outcomes.

What the XAI tools do can be classified by how they emulate the processing of the data to draw connections between the inputs and outputs. Gilpin *et al.*'s taxonomy [16] organizes XAI tools by their function to (1) extract rules to summarize decisions, (2) create a salience map to highlight a small portion of the computation which is most relevant, and (3) employ a simplified proxy that behaves similarly to the original model. For instance, Benítez *et al.* [5] transformed deep neural networks to fuzzy rules through an equivalence-by-approximation process, Simonyan *et al.* [45] produced a salience map by directly computing the input gradient, and Ribeiro *et al.* [43] used a local linear model in LIME as a simplified proxy for the full model.

With the increased usage of XAI techniques, evaluating their efficacy becomes important to inform practitioners about tool adoptions. Miller *et al.*'s survey of 23 XAI papers [31] shows that rigorous human behavioral experiments are not currently being undertaken. As the verb to explain is a three-place predicate: "Someone explains something to someone [23]", Miller and his colleagues [31] argue that most XAI tools explain things (e.g., feature or neuron importance) to the AI researchers but *not* to the intended users. Doshi-Velez and Kim [15] further argue that the best way to show how an XAI technique works is to evaluate the tool by consulting domain expert grounded in the exact application task. Although costly, the application-grounded evaluations provide direct and strong evidence (or lack thereof) of XAI's fulfillment of the requirements.

2.2 Explainability as a Non-Functional Requirement (NFR)

In software engineering, *functional requirements* describe what the system does, whereas *non-functional requirements* (NFRs) focus on how well the system does it [10, 38]. Making classifications, recommendations, and predictions are among the common functional requirements of an AI system [11], and doing so in an explainable way is often regarded as a non-functional concern [28]. Thus, researchers consider explainability to be an NFR.

In a survey study with 107 participants (90 from Brazil and 17 from Germany), Chazette and Schneider [9] elicited the participants' expectations from an explanation. Chazette and Schneider's online

[1] By 'Data Driven', we mean explicitly referencing articles on explanation in social science, and testing if the produced explanations are appropriate for the intended users [31].

questionnaire used a hypothetical scenario where the survey participants would use a vehicle's AI-based navigation system while driving on a route they had traveled before; however, AI suggested a different route than usual. Of the 103 codes analyzed from all the responses, 36 (35%) expressed desire in knowing *what* specific piece of information supported and influenced the suggestion, 12 (12%) wanted to know the *how* of the algorithm's inner reasoning, and 55 (53%) expressed willingness to understand *why* something happened (e.g., "why the [usual] route is not being suggested" and "benefits of the new route when compared to the usual" [9]).

The survey results clearly show that people's explainability requirements are different. Chazette and Schneider [9] further pointed out that eliciting explainability should also consider laws and norms, cultural and corporate values, domain aspects, and practical project constraints such as time and budget. The European Union, for instance, debated about a general "Right to Explanation" [18] which is partly enshrined in certain regulations [41]. Such policies, along with the globally emerging ethics guidelines [25], are making AI—especially AI in citizen services—more auditable.

NFRs may interact: the attempts to achieve one NFR can hurt or help the achievement of another [34]. For example, generating a post-hoc explanation imposes additional computational overhead, possibly hurting an AI service's responsiveness. Meanwhile, 35% of the codes in [9] corresponded to the responses in which the users perceived explanations as a way to reduce obscurity due to the more information about the AI system and its outcomes. Yet 15% cautioned that too technical or lengthy explanations might add more obscurity. Recognizing the trade-offs between explainability and other NFRs is therefore important for prioritizing requirements and making design choices.

In summary, the requirements engineering literature suggests that explainability is an NFR, or a softgoal whose satisfaction is a matter of degree without a clear-cut criterion [10]. Because explainability is not a technical concept but tightly coupled to human understanding, examining what XAI tools actually do and what they should do must be carried out with respect to the relevant aspect in relevant contexts. Understanding the degree to which existing XAI tools satisfy the explainability softgoal in an application-grounded task is precisely the focus of our research.

3 CASE STUDY DESIGN

3.1 Problem Context

Nearly 860 cities and towns across the U.S. have combined sewer systems, which manage stormwater as well as wastewater, creating what the U.S. Environmental Protection Agency (EPA) considers to be the largest unaddressed risk to human health from the water infrastructure [50]. According to an EPA report [50], about 850 billion gallons of untreated wastewater is discharged into waterways annually in the U.S. The excess water from storms carries dust, trash, and debris from developed regions and washes them into the combined sewer system. When these combined sewer systems are overwhelmed, they will discharge untreated wastewater into nearby waterways at an outfall site. This is defined as a combined sewer overflow (CSO) event. Figure 1 shows a simplified view of the causes of CSO events.

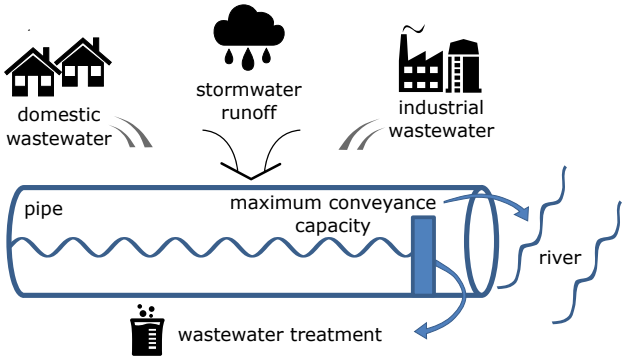


Figure 1: Illustration of a CSO (combined sewer overflow) site and how overflow can lead into nearby water sources.

In the U.S., the over 9,000 CSO outfall sites account for approximately 5,000 infections annually, damages habitats for animals in wetlands, killing fish in rivers, and closures of recreational waterways and beaches [50]. This problem is not unique to the U.S., but occurs all around the world. For example, an average of 39 million tons of untreated wastewater is dumped into the river Thames in London, UK annually due to the CSO events [13]. Even modern cities with combined sewer systems such as Shenzhen, China have to handle mitigation of pollution into rivers from the CSO events [47]. With increasing levels of urbanization and changes in weather patterns due to climate change, these problems are expected to become more severe and require new solutions to handle them in the future [13].

We worked with a wastewater treatment organization, Metropolitan Sewer District of Greater Cincinnati (MSDGC), that services an operating area of about 300 square miles, over 850,000 customers, and over 3,000 miles of combined sewers. MSDGC has set up a large scale sensor network to collect data and remotely operate their system. Some of the older outflow sites in their system can only hold a limited amount of water before they will overflow and cause a CSO event. The current practice of MSDGC is to reference weather forecast, then alert citizens if a CSO event may occur within the next day.

Since MSDGC is a public service, they need to be able to justify their reasoning for their decisions, especially when their decisions affect the safety of customers. This need for transparency is why their current system of mostly relying on weather forecasts is preferred. They can justify their decisions easily, quickly identify mistakes, and utilize this information for future warning. Ideally, alerting customers early before a CSO event occurs can help keep their customers safe. When using weather forecasts, a warning may be sent every time a large storm is expected. However, many of the alerts sent are false positives leading to customers simply ignoring them. Reducing the high false positives in predicting the CSO events ("predicting CSOs" for short) is the main reason why MSDGC is exploring deep learning solutions.

Drawing from prior experience [6, 36], we designed an exploratory case study [55] to investigate the use of deep learning and XAI tools to predict the CSOs within the real-life context of MSDGC. In particular, we worked with two domain experts from MSDGC:

Table 1: Sample of rainfall data which is sampled every minute from a sensor and returns the depth of rainfall measured in an area upstream of the CSO outfall site.

Timestamp	Rainfall (inches)
Oct 12, 2018 14:29	0.0006
Oct 12, 2018 14:30	0.0006
Oct 12, 2018 14:31	0.0006
Oct 12, 2018 14:32	0.0006
Oct 12, 2018 14:33	0.0015

Table 2: Sample of level, velocity, and flow data from the manhole site upstream of the outflow sensor. This data was sampled at a rate of once every 5 minutes.

Timestamp	Level	Velocity	Flow
Aug 20, 2019 2:15:00	1.706	0.920	0.066
Aug 20, 2019 2:20:00	1.673	0.861	0.060
Aug 20, 2019 2:25:00	1.648	0.789	0.054
Aug 20, 2019 2:30:00	1.634	0.753	0.051
Aug 20, 2019 2:35:00	1.618	0.779	0.052

a hydrologist and an operational manager. They were points of contact for our case study and were the employees that managed the data. They are representative of our stakeholders due to their organizational roles and working experiences. We communicated via emails as well as online meetings and discussions throughout the course of the case study due to restrictions relating to COVID-19. These meetings were conducted in an informal interview style where we co-designed the goal-question-metric framework[2] with the domain experts and we presented results and made revisions as needed. When presenting results of the XAI tools to the domain experts, we used the visuals and interactive elements provided directly from the selected XAI tools to observe how well the tool could provide insight to domain experts more familiar with the data but not familiar with AI research. Our work seeks to provide some examples of successes and problems of conducting further research into how domain experts can use XAI tools to better apply and utilize deep learning models.

The data collected by our wastewater treatment organization was taken from various sensors at a CSO outflow site, a manhole approximately 450 ft upstream of the outflow site, and a rainfall sensor for the area. The site is considered to be "overflowing" whenever the level of water at the CSO site exceeds the site's capacity. Each of these sites collects data independently from each other at different rates. The slowest sampling rate is one sample for every 5 minutes while the fastest is every minute. In order to handle the inconsistency and variations in real data, we used linear interpolation to handle variations in sampling time to synchronize the samples from each of our sources. As illustrations with fictitious data, Table 1 shows a sample of the rainfall data, Table 2 shows samples from sensors in a manhole a few minutes upstream in a pipe upstream, and Table 3 shows a sample of the synchronized and interpolated dataset.

[2]The framework will be further discussed in Section 3.3.

Table 3: From our dataset, we collected three features (flow, level, velocity) from the manhole upstream of the outflow site, one feature (outfall) from the outfall site itself, and one feature (rainfall) from the rainfall sensor. This is a sample of the synchronized and interpolated data points from our dataset.

Timestamp	Flow	Level	Velocity	Rainfall	Outfall
Aug 17, 2019 7:35	0.038	1.441	0.673	0.0	45.78
Aug 17, 2019 7:40	0.032	1.424	0.590	0.0	45.78
Aug 17, 2019 7:45	0.035	1.395	0.654	0.1	45.79
Aug 17, 2019 7:50	0.032	1.366	0.624	0.1	45.80

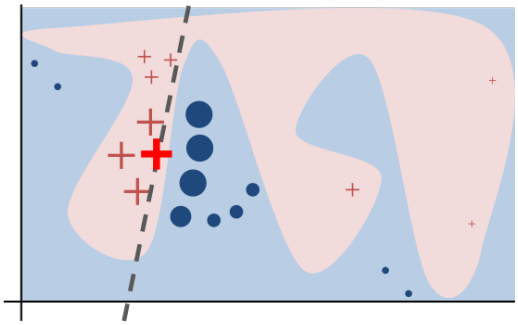


Figure 2: Figure from [43] showing an abstraction of how LIME forms a local, linear decision boundary from the more complex decision space.

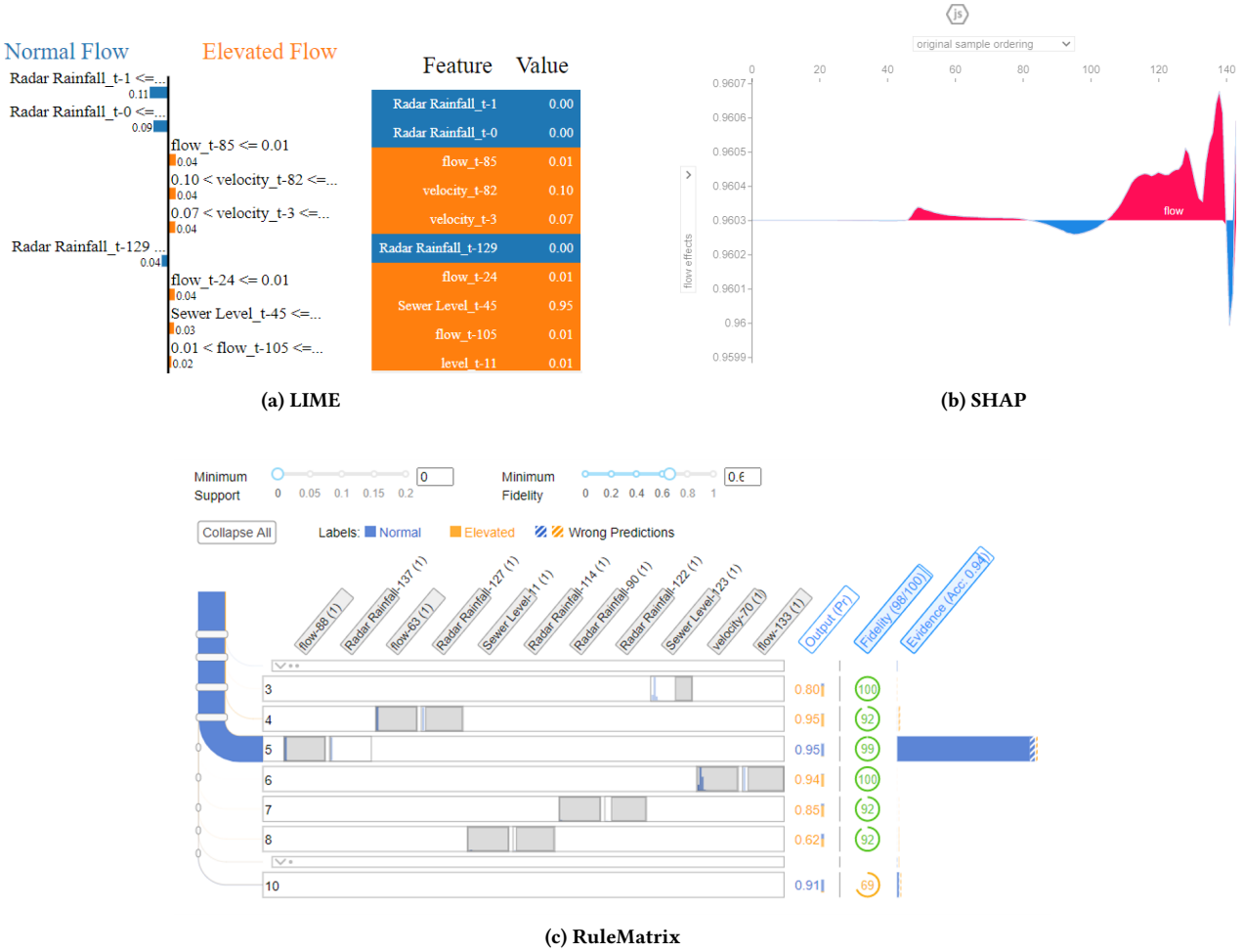
3.2 Deep Learning Solution and XAI Tools

We are using a deep learning model to take advantage of the year of continuous data collected by our wastewater treatment organization and the smart network they have developed. Since our data is sequential in nature, we are using a Long Short Term Memory (LSTM) cell structure in our model, which is in line with our recent work on deep learning based CSO predictions [8, 19]. In order to simplify the problem, we first check when all of the CSO events occur, then pass the LSTM the features from Table 3 for 12 hours of data. We utilized an Adam optimizer [27] and tensorflow [1] to create and train our model[3].

Simply giving our deep learning model to the wastewater treatment organization is not sufficient to meet their needs for transparency and justification. Therefore, we applied various XAI tools to our deep learning solution. There are quite a few tools that provide explanations for LSTM-based models [2]. In order to select and compare tools for our work, the tools we used must be available to the engineers at the wastewater treatment organization even without our direct input so they can continue to use and expand on our work. Given this constraint, the tool should be open source, compatible with our solution, and easy to use.

In addition to the above selection criteria, we wanted to use the XAI tools to explain how the LSTM has made a decision of overflow for the CSO site. Ideally this can be used to help inform future decisions for our stakeholders at the MSDGC. These tools should

[3]Our source code and results are shared at https://doi.org/10.5281/zenodo.4818970.



(a) LIME

(b) SHAP

(c) RuleMatrix

Figure 3: Illustrating the explanations generated from the XAI tools: (a) LIME's explanation displays the most influential features supporting or opposing a prediction decision, (b) SHAP's explanation visualizes how much input features affect the CSO predictions, and (c) RuleMatrix's explanation generates a hierarchy of rules by using the input features.

increase transparency and accountability of the deep learning model by better understanding how it operates.

From these requirements, we selected the XAI tools of LIME [43], SHAP [30] based on DeepLIFT [44], and RuleMatrix [32]. Each of these state-of-the-art tools can take an LSTM-based model and a given sample from our dataset, and then produce an explanation for how the deep learning model made a decision. The tools have different assumptions and make different explanations.

- **LIME** creates a local approximation of the deep learning model's output space by sampling various inputs from our dataset. LIME then uses this approximation of the output space to determine which features in the input space are the most significant to determine the model's prediction. Figure 2 shows a representation of how a linear boundary is created for a given sample of interest. This identifies the most significant features and which classes these features support.
- **SHAP** uses backpropagation and computes shapely values to determine how much influence the inputs of each layer have on the next layer. Through backpropogation, these values are progressed from each layer starting with the output back towards the input of the deep learning model. This is then used to create a significance map of how much each individual input influenced the final prediction. These values of influence can range dynamically on a gradient.
- **RuleMatrix** creates a global approximation of the deep learning model's decisions. This is done through a set of rules organized hierarchically where each rule in sequence divides the dataset based on a threshold for a given input feature. These rules, though only approximations, are inherently explainable to humans.

Figure 3 provides sample explanations illustrating the outputs from our chosen XAI tools. Some key differences of these tools can be seen through how they represent their explanations. For

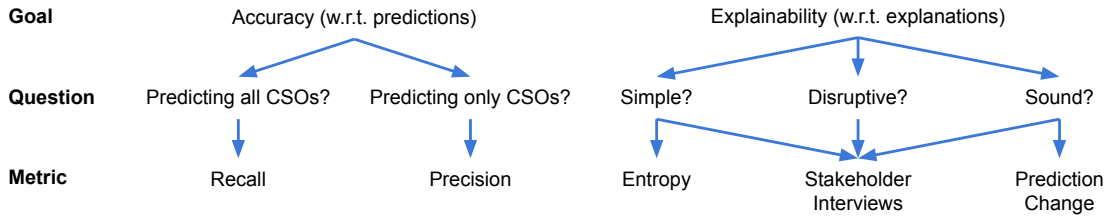


Figure 4: The goal-question-metric (GQM) framework guides our case study. Each arrow represents defining a concept in more detail. The diagram progresses from high level goals into more concrete questions, then finally to the measurable metrics.

example, between LIME and SHAP, SHAP gives a gradient of values to each feature as to how much it helped or hurt a prediction. LIME attributes influence to different features instead of a gradient for each feature to the prediction. SHAP uses backpropagation to assign influence from the prediction to the input space while LIME samples other data points in the local output space to generate explanations. RuleMatrix operates in a completely different manner as it creates a whole new model making it difficult to directly compare to SHAP and LIME. These differences motivate us to establish a coherent framework for evaluating the XAI tools.

3.3 Research Questions

We follow the goal-question-metric (GQM) approach [46] to critically understand the three chosen XAI tools in the context of deep learning based CSO prediction. The analysis is drawn from our GQM analysis of visual requirments analytics tools [37, 40]. Compared with many XAI studies that focused on evaluating the produced explanations with lay persons or AI researchers [31], we collected the feedback directly from two domain experts, i.e., a hydrologist and an operational manager at the MSDGC. We were in constant email communications with the two domain experts. Furthermore, we held three one-hour virtual meetings with these two experts to understand the data shared with the research team, to elicit their explainability related concerns, and to interview them while presenting the explanation results from the XAI tools.

The structure of our GQM analysis is presented in Figure 4 where two general goals of accuracy and explainability of CSO prediction are addressed. Relevant questions are used to refine the goals. While we measure accuracy by well-known metrics of recall and precision, the questions of explainability are explicitly built on human behavioral studies, making our case study directly 'Data Driven' according to Miller *et al.* [31]. In particular, we consider two studies on explainability from cognitive psychology and behavioral sciences.

Lombrozo [29] conducted human subject experiments to decide what caused a given event from a set of possible choices then justify these decisions, and showed that people disproportionately preferred simpler explanations over more likely ones, indicating some trade-off between the simplicity and soundness of explanations. Moreover, Thagard [49], in developing his well-known ECHO model to characterize the cognitive processes responsible for selecting between competing explanatory hypotheses, reported that people preferred the explanations consistent with their prior knowledge. Therefore, we also investigate how disruptive the XAI tools' explanations are, compared to the domain experts' existing CSO understandings. As shown in Figure 4, our case study addresses five research questions (RQs):

- **RQ$_1$**: How complete does the XAI-enabled deep learning solution predict CSOs?

  We measured this through *recall* which is the number of correctly identified CSO events out of all CSO events in the dataset. LIME and SHAP make post-hoc predictions directly from the LSTM-based deep learning model. Consequently, LIME and SHAP have the same recall value as the LSTM. RuleMatrix, on the other hand, requires a re-computation of recall according to the generated rules.

- **RQ$_2$**: How much noise is there as the XAI-enabled deep learning solution predicts CSO events?

  We measured this through *precision*, the number of correctly identified CSO events out of all predicted events by the deep learning model. Just as with recall, LIME and SHAP make post-hoc predictions and thus have the same precision as the original LSTM, whereas this metric will need to be re-computed for the rule-based model created by RuleMatrix.

- **RQ$_3$**: How simple are the explanations generated by the XAI tools?

  In order to evaluate the simplicity of these tools, we used both numeric metrics as well as interviews. For a quantitative metric, we computed the *entropy* of the explanations produced by the XAI tools. Entropy measures the uncertainty, or disorder, of a distribution [42] which can be used to approximate how much unique information and variability is in the explanation. In addition, we interviewed the two experts and showed them explanations from the various XAI tools.

- **RQ$_4$**: How sound are the explanations generated by the XAI tools?

  Soundness of an explanation can be difficult to investigate since it depends on the background of a stakeholder, as discussed by Gilpin *et al.* [16]. However, XAI tools such as DeepLIFT [44], SHAP [43], and Layer-wise Relevance Propagation [4] all attempt to assess the "correctness" of the generated explanations by masking the most significant data values identified by an XAI tool from a sample to examine

the corresponding prediction changes of a given deep learning model. Thus, we applied this "prediction change" metric as a quantitative measure of the soundness of an explanation. We also interviewed the two domain experts to assess their confidence in the plausibility of the XAI tools' explanations.

- **RQ$_5$**: How much new insight, if any, do the XAI tools' explanations offer?

  To our stakeholders from the MSDGC, performance is an important aspect as an AI system must both perform better and be as interpretable as their existing system to justify its use. It is clear to us that our stakeholders want to know if deep learning can provide a new perspective on the problem. However, having a model deviate too much from their expectations may make it difficult to trust. Through interviewing the stakeholders, we assessed how deep learning equipped with the explanations can help provide new insights toward identifying the CSO events, thereby potentially disrupting some aspects of the MSDGC's practice.

4 RESULTS AND ANALYSIS

As mentioned in Section 3, we designed an LSTM deep learning solution to predict CSOs for the MSDGC. We analyzed the results from this deep learning model as well as the XAI tools described in Section 3.2. These results are from an LSTM that predicts if a CSO event will occur within the next hour after being given the previous 12 hours of data.

4.1 Predictive Accuracy: Recall and Precision

The most important metric for model performance to the MSDGC is accurately identifying events. If the solution fails to identify events (i.e., recall is low), it will not be able to warn the citizens in the serving area. If there are too many false positives (i.e., precision is low), citizens will likely ignore the warnings.

As discussed in Section 3.3, LIME and SHAP had the same recall and precision values as the LSTM model. RuleMatrix's recall and precision needed to be re-computed once the resulting rules were generated. To calculate the accuracy measures of both LSTM and RuleMatrix, we used a 2-month long test subset of the dataset and then evaluated the predictions based on the given labels from the wastewater treatment organization: "elevated" means CSO events occurred; "normal" means otherwise.

The recall and precision results are plotted in Figure 5. The figure shows the results of predicting whether a CSO event will happen within the next hour for every 5 minute interval over a 2-month duration. The recall and precision of the deep learning model (and hence LIME and SHAP) are approximately 80% and 45% respectively. The recall and precision for RuleMatrix are only about 40% and 20% respectively, representing a 50% recall drop and a 55% precision drop.

While disappointed in the RuleMatrix's accuracy levels, the two domain experts believed LSTM's CSO predictions were encouraging and agreed with our suggestions of improving the model performances by incorporating data from more CSO sites. During the interviews, the experts were also interested in how the LSTM's performances would compare to the MSDGC's current practice of

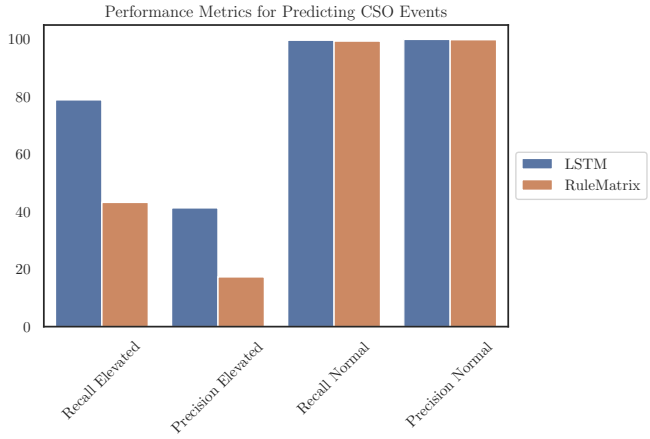


Figure 5: Answering RQ$_1$ and RQ$_2$. Recall of predicting CSOs ("elevated" class): LIME=SHAP=LSTM=78.9%, RuleMatrix=43.2%. Precision of predicting CSOs: LIME=SHAP=LSTM=41.3%, RuleMatrix=17.3%. Recall of predicting non-CSO events ("normal" class): LIME=SHAP=LSTM=99.6%, RuleMatrix=99.3%. Precision of predicting non-CSO events: LIME=SHAP=LSTM=99.9%, RuleMatrix=99.8%.

relying on weather forecast to inform the citizens about potential CSO events.

To investigate this, we further collected rainfall data from NOAA DIVER [14] for the area around the CSO site for the same date and time range that our deep learning model was tested on. When using a constant rainfall threshold for a given day (i.e., 0.5 inches of rainfall per day), a recall of 100% and a precision of 20% were obtained. The low precision level helped illustrate the exploration of the deep learning solutions. Although LSTM's 41.3% doubled the CSO predictions' precision, it is important to note that the NOAA dataset can only collect rainfall data for each day, whereas the LSTM-based deep learning model makes predictions continuously for a time range in the future. This continuous prediction of the deep learning model could lead to lower recall as a prediction one hour before an event may be correct but half an hour before an event may be incorrect. LSTM does not predict every interval before the CSO event correctly but it does identify 78.9% of the 5 minute intervals an hour before a CSO event from Figure 5. The deep learning model is more precise than only considering rainfall data and can be substantially improved in the future. The value added by the deep learning model redefines the problem and, with some improvement to the deep learning model, could give the wastewater treatment organization to proactively address CSO events before they occur. The above analyses suggest:

> **Finding 1:** While RuleMatrix's recall and precision are low, LSTM (and hence LIME and SHAP) achieves about 80% recall and performs more precisely than the current practice. In addition, LSTM provides new predictive capabilities for every five minutes before a CSO event as opposed to daily weather forecasts.

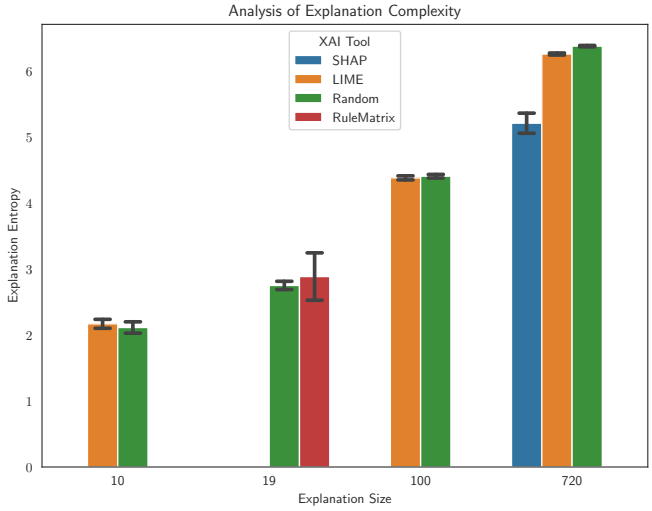


Figure 6: Explanation complexity measured by entropy. Random is sampled from a uniform distribution, representing a baseline entropy. Entropy is computed from the Scikit-Learn library [39] with a $\log_2$ scale for these results. Greater entropy values indicate more complex explanations (RQ_3).

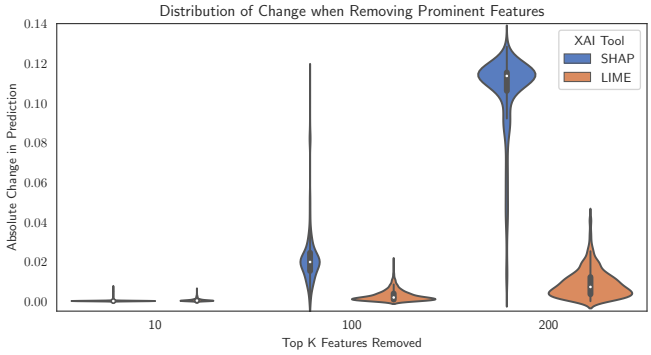


Figure 7: Explanation soundness generated by removing the most significant k features identified by SHAP and LIME from a sample and then evaluating how much the removal changed the prediction. Greater change in prediction implies more soundness of the explanations (RQ_4).

4.2 Explanation Simplicity and Soundness

Our RQ_3 and RQ_4 investigate to what extent, "simpler explanations favored over more likely ones" [29], holds in our case study. As illustrated in Table 3, our LSTM model had a total of 720 input parameters consisting of five features sampled at a rate of every 5 minutes for 12 hours. To quantitatively evaluate all three XAI tools, we computed SHAP for all input parameters, and experimented LIME for 10 (default), 100, and 720 (all) parameters. RuleMatrix had a fixed number of decisions so we used the 19 rules that were produced.

In order to quantify simplicity, we computed the entropy to measure the variation and uniqueness of the explanations produced by each tool. The explanation of each instance was a matrix sharing the same shape as the input dataset. The influence of each element on the final prediction was stored in this matrix. RuleMatrix created a set of rules, so to approximate this level of information we simply used a string of ones and zeros where a one indicates a rule was used and a zero indicates that a rule was skipped. When computing entropy, a set of numbers of the same value has an entropy of zero while a set of completely unique or random numbers would have much greater entropy. We repeated a random sampling 1,000 times to identify an accurate representation of random entropy. This is limited to a small sampling due to the time it takes to create an explanation.

Results in Figure 6 show an increase in entropy with increase in explanation size (i.e., the number of parameters in an explanation). This trend is expected as more unique values are being added to the explanations. As far as the XAI tools are concerned, LIME had complexity almost equivalent to a random sampling, and it could also be filtered to the most significant results. SHAP had much less entropy for the 720 parameters that it explained, and had a wider variance than a random sampling. We speculate that this is due to the majority of the features have little significance, potentially resulting from backpropagation and how the SHAP tool parses the LSTM structure. RuleMatrix's entropy was on par with a random sample of that size but the variation was wide. Note that RuleMatrix produced what is commonly believed to be inherently explainable and simple in terms of citing a small list of causes for each prediction.

Explanation soundness (RQ_4) evaluates how "correct" an explanation is at determining the decision made by the black box LSTM. We quantified how much the deep learning model's prediction would change when removing the most significant features at specific timestamps within the dataset. This measure gives us an experimental verification of the importance of the identified features in the explanations. Since RuleMatrix did not distinguish in the explanations which elements were the most significant for each individual prediction, we could not directly evaluate RuleMatrix by using the "prediction change" procedure.

Figure 7 shows the soundness results when removing the top k features for SHAP and LIME. For the top 10 features, SHAP and LIME have a change in prediction of about 0.0001 and 0.0007, respectively. When increasing k to 100, SHAP has a much wider range, centered at 0.0217 while LIME has a 0.0028 average. When k increases further to 200, SHAP and LIME's distributions cluster around 0.1050 and 0.0088 respectively. Figure 7's results suggest that LIME has the slightly more sound results for a smaller k, but a larger k seems to manifest SHAP's soundness better.

The interviews with the two domain experts confirmed our observations and provided new views. Although RuleMatrix is inherently more explainable than LIME and SHAP, the hydrologist and the operational manager at the MSDGC did not find the rule hierarchy captured the relevant knowledge in the CSO domain. Both LIME and SHAP were well received by the experts, with more preferences shifting to SHAP as this XAI tool exhibits more soundness when all the features are considered together. Despite being sound at smaller k and being flexible in terms of having a customizable number of

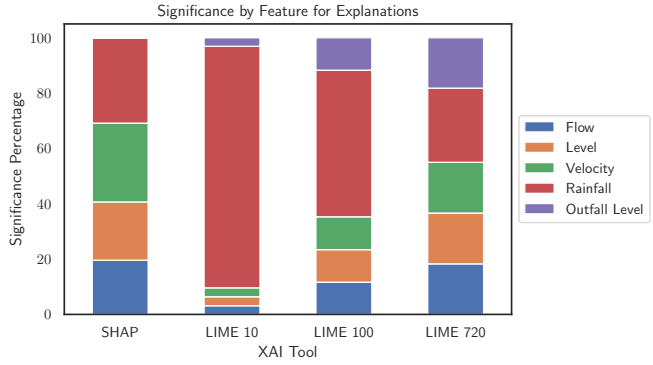


Figure 8: The average influence each feature had on the overall prediction. Observing where the influence comes from helps to establish both soundness of results (RQ_4) and to explore new insights (RQ_5).

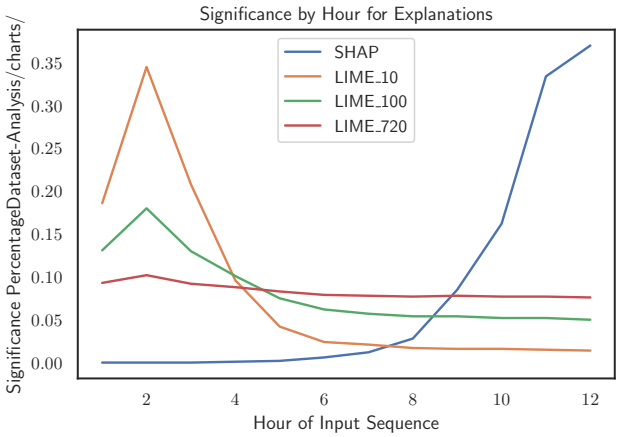


Figure 9: The average influence each hour of time before the predicted event had on the overall prediction. This distribution helps to identify the new insight about what specific features to focus on and *when* to focus on them (RQ_5).

parameters in an explanation, LIME could be confusing when incorporating many features. Surprisingly, from Figure 6's entropy perspective, SHAP is simpler than LIME when all the features are taken into account. Based on the analyses of RQ_3 and RQ_4, we summarize:

> **Finding 2:** Explanation's simplicity does *not* always come at the cost of soundness. Domain experts would clearly favor soundness over simplicity, and in our case study of CSO predictions, SHAP's more sound explanations with many parameters turn out to be also simpler.

4.3 New Insights from the Explanations

Our stakeholder interviews included an interactive session where the two domain experts could explore the XAI tools beyond the results that the research team had prepared. We highlight some concrete insights gained from the interactive session, which focused more on LIME and SHAP, due to RuleMatrix's low recall and precision levels.

Figure 8 shows the general results as to which features from the dataset are the most influential to the deep learning model across all predictions. SHAP distributed influence fairly evenly between the features while LIME heavily favored Rainfall. As LIME increased to include more features in the explanation, the distribution evened out more. In addition to this, a visualization of the influence by time is shown in Figure 9. SHAP heavily favored the time right before the CSO event while LIME favored the start of the sample. Similar to Figure 8, the influence of LIME in Figure 9 evened out as more features were included in the explanation.

The experts stated that LIME results were useful in identifying the most significant features when looking at the top 10 elements. Meanwhile, they were able to interpret new insights from the significance plots generated by SHAP such as Figure 3(b) and the summary plot of Figure 8: they found a pattern in the correlation between velocity, flow, and level and suggested why these attributes might be more significant in a few sample events. At first, they had not assumed the feature of velocity to be useful for predictions but realized and discussed how correlations between velocity and flow could help predict future CSO events. This provided a new insight to help guide future development and resources for their analysis that challenged their initial view.

Another major insight was the prominence of Rainfall in the decision making process of the model, as shown in Figure 8. This confirmed our stakeholders' expectations as excess storm water is the main cause of CSO events and helped them to trust the results. However, the dominant influence of Rainfall leveled off not only when LIME's explanations involved more features, but also when SHAP was applied. Given that LIME and SHAP achieved the same recall and precision levels, the patterns revealed in Figure 9 offers remarkable insights into *when* to focus on which features. While weather forecast's Rainfall could still be dependent upon in alerting CSOs to the relevant citizens 12 hours ahead of the time, paying additional and equal attention to sensor network data like velocity could potentially leave 2–4 hours for the MSDGC to dispatch engineers onsite to prevent, alleviate, or otherwise manage the CSO events. Based on the interviews, we conclude:

> **Finding 3:** XAI tools of our case study, especially LIME and SHAP, have the potential to disrupt stakeholder expectations of the influential factors of CSOs, as well as to take justifiable actions to ameliorate the CSO situation.

4.4 Threats to Validity

Our inquiry is an exploratory case study [55] aimed at investigating the contemporary CSO phenomenon in depth and within its real-life context. We discuss some of the most important aspects that must be considered when interpreting our case study's results. A threat to construct validity is our choice of the three XAI tools. As mentioned in Section 3.2, our tool selections were guided by the considerations of being open source, being compatible with our LSTM solution, and being easy to use for the stakeholders in the

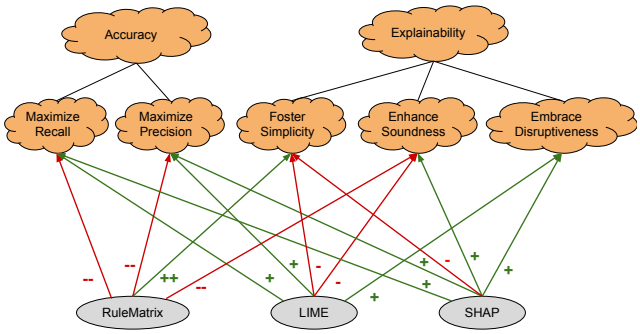


Figure 10: Softgoal Interdependence Graph (SIG) for the XAI tools. The undirected lines represent goal decompositions, informed by our GQM analysis (cf. Figure 4). The arrows represent softgoal contributions [10]: "−−" means breaks, "−" means hurts, "+" means helps, and "++" means makes.

wastewater treatment organization even without our assistance. To those ends, we limited our scope to evaluate the XAI tools *as is*, without any further adjustment or customization. Another construct validity relates to our use of entropy. Insights provided by the XAI tools provide added value to stakeholders and can guide future development and improvements for CSO prediction. Although entropy directly measures variance and uniqueness and thus implies more on information density, we found it to be useful when quantitatively measuring how complex a result was.

A threat to internal validity concerns the size of the real-world dataset shared with us by our stakeholder organization. Our dataset was limited in scale, e.g., the dataset was heavily biased towards the negative class (normal flow). We therefore augmented it with oversampling to effectively train the deep learning model. The augmentations might have had unintended consequences on the results of the XAI tools as this was not fully explored in our work. More historical data of the CSO site would help evaluate our assumptions of our augmentations, mitigating the threat to internal validity.

We believe our study's conclusion validity is high. First and foremost, we set out to overcome the "data driven" weakness of current XAI studies [31]. Referring explicitly to the relevant literature [29, 49] allowed our inquires to stay focused and our conclusions to be well grounded. Furthermore, bias is mitigated by investigating XAI tools not developed by the research team. Although the two domain experts are only a small sample in the field, they have real stake in the potential changes to be introduced by deep learning. Last but not least, we share our source code at https://doi.org/10.5281/zenodo.4818970 in order to facilitate replication and expansion of our results.

5 DISCUSSION

5.1 Satisficing Explainability

Explainability, as an NFR discussed in Section 2.2, is satisficed [10] in a matter of degrees. Based on our GQM analysis from Figure 4 and the quantitative and qualitative results presented in Section 4, we build a Softgoal Interdependence Graph (SIG) in Figure 10. In the SIG, each of the XAI tools contributes either positively or negatively to the softgoals. From Figure 10, we note that none of the XAI tools that we investigated makes all positive contributions, indicating the tools are all limited in some aspects. A tool cannot help meet some softgoal without hurting some others, suggesting the trade-offs among the softgoals. Interestingly, Figure 10 does not reveal the well-known trade-off between recall and precision. Instead, the LSTM (and hence LIME and SHAP) achieved higher recall and higher precision than RuleMatrix, showing that RuleMatrix is less fit for making CSO predictions. However, as of now, LIME and SHAP do not perform accurately enough to be adopted by our stakeholders.

One might argue the key to improving accuracy is with the underlying deep learning model (namely LSTM in our case), making explainability less of a concern. We argue considering explainability, even when accuracy levels are not ideal, is still valuable. For example, an interesting contrast between LIME and SHAP is where the influence for CSO predictions comes from. Although different, the explanations provided by LIME and SHAP are both valid and show how these XAI methodologies diverge in operation. The deep learning model predicts events one hour into the future. However, LIME prioritizes influence of rainfall from many hours before a CSO event, whereas SHAP prioritizes influence from all features evenly immediately preceding a CSO event. Deciphering the black-box deep learning models, though with varying degrees of satisficing explainability, is of vital importance for ensuring public sector's transparency and accountability.

5.2 Data Driven Explainability

As we drive our inquiry by explicitly referencing the explainability findings from [29, 49], we cast our case study's results in light of the relevant literature. Lombrozo [29] showed that people disproportionately prefer simpler explanations over more likely ones; however, Lombrozo's work was carried out with student subjects who were not domain experts in the field. Through our case study, we have seen that the hydrologist and the operational manager greatly favored the soundness of the explanations. It is also worth mentioning that simplicity does not necessarily correlate with size (i.e., the number of features an explanation has). By computing entropy to measure the randomness of the information contained in the explanations (cf. Figure 6), we observe that soundness and simplicity co-exist in SHAP's explanations.

Thagard [49] reported that people prefer the explanations that are consistent with their prior knowledge. Our domain experts conformed largely to Thagard's conjecture. They confirmed the XAI tools' outputs generally cohered with what they expected. In some occasions, we noticed that the XAI tools' explanations refuted our domain experts' expectations. The experts kept an open mindset, and were able to find new insights from the dataset. Specifically, the results of SHAP gave more influence to velocity than the experts initially expected (cf. Figure 8). Since SHAP mostly drew influence from right before the CSO event (cf. Figure 9), they were able to reason that this explanation drew upon information that they might have overlooked. Because of these observations, LIME or SHAP *alone* might not be able to uncover the new insights. In the SIG of Figure 10, therefore, it is the *synergy* of LIME and SHAP that contributes positively toward the "embrace disruptiveness" softgoal.

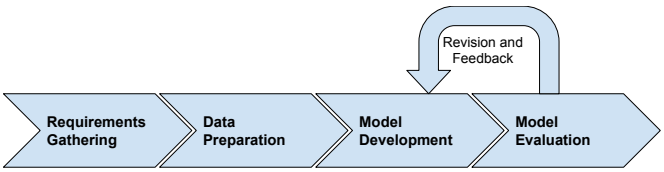


Figure 11: Simplified view of the software engineering process for AI-based systems proposed by Amershi *et al.* [3].

5.3 Software Engineering for AI-Based Systems

An important lesson learned from our case study is that one shall *not* treat explainability as something to add after a deep learning model is built. Our experience advocates strongly for explainability to be engineered throughout the deep learning project. Amershi *et al.* [3] breaks the software engineering process for machine learning into nine stages. A simplified view of this process into four phases is shown in Figure 11. Explainability is so broadly scoped that it influenced our decisions for each phase of the software engineering process.

- **Requirements Gathering**. We interviewed our stakeholders to identify what needs to be explained and why. The critical requirement of warning citizens about the CSO risks helped us to better understand the role that XAI might play in accountability, and to better build the deep learning model in making CSO predictions.
- **Data Preparation**. We made a few assumptions about the data and applied data cleaning and augmentation by interpolating data points to synchronize the various data sources (cf. Tables 1–3). Understanding the composition of data from each of the sensors and their preparation helped us better contextualize the results of the XAI tools' explanations.
- **Model Development**. Integrating XAI into the deep learning model not only required extra effort, but also led to performance decrements due to the resources required to generate and visualize explanations. For tools like RuleMatrix, an additional step was required to predict CSOs according to the generated rules.
- **Model Evaluation**. Explanations can be consumed by more than just AI researchers or lay persons; from our work, we have found that domain experts can contextualize these explanations or use them to gain new insights into the task at hand. The "revision and feedback" of Figure 11 may involve exploring different numbers of top features from explanations provided by LIME. Additional feedback could be linked to other phases. For example, the insight gained from SHAP's results discussed in Section 4.3 helped elicit a new requirement of using deep learning to inform engineer-dispatch decisions 2–4 hours prior to a likely CSO event.

XAI tools can be integrated into all phases of Figure 11 as shown through our case study. They helped inform decisions throughout software development and can be integrated into pipelines as a method of verifying model performance or to diagnose issues and their causes. As noted by Zhang *et al.* [56], there is a need for identifying how and why deep learning models make decisions to satisfy other broadly scoped requirements, such as fairness, privacy, and robustness. We believe these concerns must be incorporated into all the phases of machine learning development, and our case study has demonstrated the feasibility of engineering explainability with state-of-the-art XAI tools.

6 CONCLUDING REMARKS

Research on explainability of XAI tools with domain experts using real-world data is significant to ensuring the tool's ability to be used with new topics. Our work is a proof of concept for how a GQM analysis might be used to assess various XAI tools. Despite not developing an immediately applicable product for domain experts, we believe that this research can serve as a foundation to assist development of XAI tools that are more useful for a broader set of stakeholders by providing a framework based on GQM for potential analysis.

In our exploratory case study, we employed qualitative and quantitative methods to evaluate the predictions' accuracy and the explanations' simplicity, soundness, and disruptiveness. Through comparing the numeric metrics and reviewing the results of the stakeholder interviews, we are able to build upon existing psychological results and contextualize them with respect to the modern XAI tools. Domain experts welcome new insights and more complex explanations with multiple causes. Our findings do not directly refute the work of Lombrozo [29] and Thagard [49], but rather build upon their work in noting that the different levels of complexity may be appropriate for different stakeholders depending on their background [16].

To further expand upon this work, explanations from more XAI tools can be investigated for new insights and for supporting different software engineering tasks (e.g., [12, 24, 48, 51]). Future work can also explore how to effectively, efficiently, dynamically, and continuously present value-added explanations of deep learning model to stakeholders [33]. Additionally, we want to use more data to expand our case study and investigate how seasonal rainfall differences affect the XAI results. Last but not least, we seek to employ diverse empirical methods in our future work, such as case studies co-design with domain experts [53, 54] and theoretical replications [26, 35]. Investigating explainability as a non-functional requirement of AI-based systems is an open area of research. The individual metrics and interview questions of our study are only first steps to spark a discussion of how they can be improved further.

ACKNOWLEDGMENTS

We appreciate the anonymous reviewers for their constructive and insightful suggestions towards improving this manuscript. We also thank Richard Chiang and Dr. Cathy Maltbie for help editing and proofreading.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng.

2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Retrieved June 2021 from https://www.tensorflow.org/

[2] Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[3] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald C. Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. 2019. Software Engineering for Machine Learning: A Case Study. In *Proceedings of the 41st IEEE/ACM International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP'19)*. Montreal, Canada, 291–300. https://doi.org/10.1109/ICSE-SEIP.2019.00042

[4] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PloS one* 10, 7 (2015), e0130140. https://doi.org/10.1371/journal.pone.0130140

[5] José Manuel Benítez, Juan Luis Castro, and Ignacio Requena. 1997. Are Artificial Neural Networks Black Boxes? *IEEE Transactions on Neural Networks* 8, 5 (September 1997), 1156–1164. https://doi.org/10.1109/72.623216

[6] Tanmay Bhowmik, Vander Alves, and Nan Niu. 2014. An Exploratory Case Study on Exploiting Aspect Orientation in Mobile Game Porting. In *Integration of Reusable Systems*, Thouraya Bouabana-Tebibel and Stuart H. Rubin (Eds.). Springer, Chapter 11, 241–261. https://doi.org/10.1007/978-3-319-04717-1_11

[7] Supriyo Chakraborty, Richard Tomsett, Ramya Raghavendra, Daniel Harborne, Moustafa Alzantot, Federico Cerutti, Mani B. Srivastava, Alun D. Preece, Simon Julier, Raghuveer M. Rao, Troy D. Kelley, Dave Braines, Murat Sensoy, Christopher J. Willis, and Prudhvi Gurram. 2017. Interpretability of Deep Learning Models: A Survey of Results. In *Proceedings of the IEEE International Conference on Ubiquitous Intelligence and Computing (UIC'17)*. San Francisco, CA, USA, 1–6. https://doi.org/10.1109/UIC-ATC.2017.8397411

[8] Harshitha Challa, Nan Niu, and Reese Johnson. 2020. Faulty Requirements Made Valuable: On the Role of Data Quality in Deep Learning. In *Proceedings of the 7th IEEE International Workshop on Artificial Intelligence for Requirements Engineering (AIRE'20)*. Zurich, Switzerland, 61–69. https://doi.org/10.1109/AIRE51212.2020.00016

[9] Larissa Chazette and Kurt Schneider. 2020. Explainability as a Non-Functional Requirement: Challenges and Recommendations. *Requirements Engineering* 25, 4 (December 2020), 493–514. https://doi.org/10.1007/s00766-020-00333-1

[10] Lawrence Chung, Brian A. Nixon, Eric Yu, and John Mylopoulos. 1999. *Non-Functional Requirements in Software Engineering*. Springer.

[11] Fabiano Dalpiaz and Nan Niu. 2020. Requirements Engineering in the Days of Artificial Intelligence. *IEEE Software* 37, 4 (July/August 2020), 7–10. https://doi.org/10.1109/MS.2020.2986047

[12] Hoa Khanh Dam, Truyen Tran, and Aditya Ghose. 2018. Explainable Software Analytics. In *Proceedings of the 40th ACM/IEEE International Conference on Software Engineering: New Ideas and Emerging Results (ICSE'18)*. Gothenburg, Sweden, 53–56. https://doi.org/10.1145/3183399.3183424

[13] Department for Environment Food & Rural Affairs. 2015. Creating a River Thames fit for our future: An updated strategic and economic case for the Thames Tideway Tunnel. Retrieved June 2021 from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/471847/thames-tideway-tunnel-strategic-economic-case.pdf

[14] DIVER. 2020. *Web Application: Data Integration Visualization Exploration and Reporting Application, National Oceanic and Atmospheric Administration.* Retrieved June 2021 from https://www.diver.orr.noaa.gov

[15] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. (2017). arXiv:1702.08608

[16] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. In *Proceedings of the 5th IEEE International Conference on Data Science and Advanced Analytics (DSAA'18)*. Turin, Italy, 80–89. https://doi.org/10.1109/DSAA.2018.00018

[17] Xavier Glorot and Yoshua Bengio. 2010. Understanding the Difficulty of Training Deep Feedforward Neural Networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS'10)*. Sardinia, Italy, 249–256. http://proceedings.mlr.press/v9/glorot10a.html

[18] Bryce Goodman and Seth R. Flaxman. 2017. European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". *AI Magazine* 38, 3 (2017), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

[19] Hemanth Gudaparthi, Reese Johnson, Harshitha Challa, and Nan Niu. 2020. Deep Learning for Smart Sewer Systems: Assessing Nonfunctional Requirements. In *Proceedings of the 42nd IEEE/ACM International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS'20)*. Seoul, South Korea, 35–38. https://dl.acm.org/doi/10.1145/3377815.3381379

[20] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. A Survey of Methods for Explaining Black Box Models. *Comput. Surveys* 51, 5 (January 2019), 93:1–93:42. https://doi.org/10.1145/3236009

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*. Las Vegas, NV, USA, 770–778. https://doi.org/10.1109/CVPR.2016.90

[22] High-Level Expert Group on Artificial Intelligence, European Commission. 2019. *Policy and Investment Recommendations for Trustworthy AI.* Retrieved June 2021 from https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai

[23] Denis J. Hilton. 1990. Conversational Processes and Causal Explanation. *Psychological Bulletin* 107, 1 (January 1990), 65–81. https://doi.org/10.1037/0033-2909.107.1.65

[24] Jirayus Jiarpakdee, Chakkrit Tantithamthavorn, and John Grundy. 2021. Practitioners' Perceptions of the Goals and Visual Explanations of Defect Prediction Models. (2021). arXiv:2102.12007

[25] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence* 1 (September 2019), 389–399. https://doi.org/10.1038/s42256-019-0088-2

[26] Charu Khatwani, Xiaoyu Jin, Nan Niu, Amy Koshoffer, Linda Newman, and Juha Savolainen. 2017. Advancing Viewpoint Merging in Requirements Engineering: A Theoretical Replication and Explanatory Study. *Requirements Engineering* 22, 3 (September 2017), 317–338. https://doi.org/10.1007/s00766-017-0271-0

[27] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. (2014). arXiv:1412.6980

[28] Maximilian A. Köhl, Kevin Baum, Markus Langer, Daniel Oster, Timo Speith, and Dimitri Bohlender. 2019. Explainability as a Non-Functional Requirement. In *Proceedings of the 27th IEEE International Requirements Engineering Conference (RE'19)*. Jeju Island, South Korea, 363–368. https://doi.org/10.1109/RE.2019.00046

[29] Tania Lombrozo. 2007. Simplicity and probability in causal explanation. *Cognitive Psychology* 55, 3 (November 2007), 232–257. https://doi.org/10.1016/j.cogpsych.2006.09.006

[30] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Long Beach, CA, USA, 4765–4774. http://papers.nips.cc/paper/6930-a-universal-analysis-of-large-scale-regularized-least-squares-solutions

[31] Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of Inmates Running the Asylum. (2017). arXiv:1712.00547v2

[32] Yao Ming, Huamin Qu, and Enrico Bertini. 2019. RuleMatrix: Visualizing and Understanding Classifiers with Rules. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (January 2019), 342–352. https://doi.org/10.1109/TVCG.2018.2864812

[33] Nan Niu, Sjaak Brinkkemper, Xavier Franch, Jari Partanen, and Juha Savolainen. 2018. Requirements Engineering and Continuous Deployment. *IEEE Software* 35, 2 (March/April 2018), 86–90. https://doi.org/10.1109/MS.2018.1661332

[34] Nan Niu and Steve Easterbrook. 2007. So, You Think You Know Others' Goals? A Repertory Grid Study. *IEEE Software* 24, 2 (March/April 2007), 53–61. https://doi.org/10.1109/MS.2007.52

[35] Nan Niu, Amy Koshoffer, Linda Newman, Charu Khatwani, Chatura Samarasinghe, and Juha Savolainen. 2016. Advancing Repeated Research in Requirements Engineering: A Theoretical Replication of Viewpoint Merging. In *Proceedings of the 24th IEEE International Requirements Engineering Conference (RE'16)*. Beijing, China, 186–195. https://doi.org/10.1109/RE.2016.46

[36] Nan Niu, Alejandra Yepez Lopez, and Jing-Ru C. Cheng. 2011. Using Soft Systems Methodology to Improve Requirements Practices: An Exploratory Case Study. *IET Software* 5, 6 (December 2011), 487–495. https://doi.org/10.1049/iet-sen.2010.0096

[37] Nan Niu, Sandeep Reddivari, and Zhangji Chen. 2013. Keeping Requirements on Track via Visual Analytics. In *Proceedings of the 21st IEEE International Requirements Engineering Conference (RE'13)*. Rio de Janeiro, Brazil, 205–214. https://doi.org/10.1109/RE.2013.6636720

[38] Nan Niu, Li Da Xu, Jing-Ru C. Cheng, and Zhendong Niu. 2014. Analysis of Architecturally Significant Requirements for Enterprise Systems. *IEEE Systems Journal* 8, 3 (September 2014), 850–857. https://doi.org/10.1109/JSYST.2013.2249892

[39] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830. https://jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

[40] Sandeep Reddivari, Shirin Rad, Tanmay Bhowmik, Nisreen Cain, and Nan Niu. 2014. Visual Requirements Analytics: A Framework and Case Study. *Requirements Engineering* 19, 3 (September 2014), 257–279. https://doi.org/10.1007/s00766-013-0194-3

[41] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016. 2016. *General Data Protection Regulation.* Retrieved June 2021 from https://eur-lex.europa.eu/eli/reg/2016/679/oj

[42] Alfréd Rényi. 1961. On Measures of Entropy and Information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California.

[43] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*. San Francisco, CA, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[44] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning Important Features Through Propagating Activation Differences. (2017). arXiv:1704.02685

[45] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. (2014). arXiv:1312.6034v2

[46] Rini Van Solingen and Egon Berghout. 1999. *The Goal/Question/Metric Method: A Practical Guide for Quality Improvement of Software Development.* McGraw-Hill.

[47] Gianni Talamini, Di Shao, X. Su, X. Guo, and X. Ji. 2016. Combined Sewer Overflow In Shenzhen, China: The Case Study of Dasha River. *WIT Transactions on Ecology and the Environment* 210 (2016), 785–796. https://doi.org/10.2495/SDP160661

[48] Chakkrit Tantithamthavorn, Jirayus Jiarpakdee, and John Grundy. 2020. Explainable AI for Software Engineering. (2020). arXiv:2012.01614

[49] Paul Thagard. 1989. Explanatory Coherence. *Behavioral and Brain Sciences* 12, 3 (September 1989), 435–502. https://doi.org/10.1017/S0140525X00057046

[50] United States Environmental Protection Agency. 2004. Report to Congress: Impacts and control of CSOs and SSOs. Retrieved June 2021 from https://www.epa.gov/npdes/2004-npdes-cso-report-congress

[51] Wentao Wang, Nan Niu, Hui Liu, and Zhendong Niu. 2018. Enhancing Automated Requirements Traceability by Resolving Polysemy. In *Proceedings of the 26th IEEE International Requirements Engineering Conference (RE'18)*. Banff, Canada, 40–51. https://doi.org/10.1109/RE.2018.00-53

[52] Meredith Whittaker, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, Sarah Myers West, Rashida Richardson, Jason Schultz, and Oscar Schwartz. 2018. *AI Now Report.* Retrieved June 2021 from https://ainowinstitute.org/AI_Now_2018_Report.pdf

[53] Christine T. Wolf and Jeanette Blomberg. 2019. Evaluating the Promise of Human-Algorithm Collaborations in Everyday Work Practices. *Proceedings of the ACM on Human-Computer Interaction* 3, EICS (June 2019), 143:1–143:23. https://doi.org/10.1145/3359245

[54] Christine T. Wolf and Jeanette Blomberg. 2019. Explainability in Context: Lessons from an Intelligent System in the IT Services Domain. In *Joint Proceedings of the ACM IUI 2019 Workshops (IUI'19)*. Los Angeles, CA, USA. http://ceur-ws.org/Vol-2327/IUI19WS-ExSS2019-17.pdf

[55] Robert K. Yin. 2008. *Case Study Research: Design and Methods.* Sage Publications.

[56] Jie M. Zhang, Mark Harman, Lei Ma, and Yang Liu. 2020. Machine Learning Testing: Survey, Landscapes and Horizons. *IEEE Transactions on Software Engineering* (2020). https://doi.org/10.1109/TSE.2019.2962027