Università degli studi di Milano-Bicocca

Data Science Lab for Smart Cities

Final Essay

Smart Cities and Safe Drinking Water: A Data-Driven Approach to Detecting Drinking Water Contamination

Author:
Amadeus Beckmann - 921682 - a.beckmann@campus.unimib.it

June 4, 2024

Abstract

Clean and safe drinking water is essential for public health and the development of societies, especially in the context of Smart Cities. This essay examines a data-driven approach to detecting drinking water contamination, focusing on the need for efficient and affordable real-time monitoring. Traditional methods of water quality testing are often slow and involve manual tests, which makes detection in real-time challenging. By utilizing modern computational intelligence methods, such as AI and machine learning, we can improve water quality monitoring systems. These technologies offer significant benefits for both developed regions with existing infrastructure and developing regions facing limited resources. This essay discusses the sociological and ethical importance of water quality, evaluates different indicators for measuring water safety, and examines the potential of AI-based methods for real-time contamination detection. We demonstrate the effectiveness of machine learning models for contamination detection using a real-world dataset from the Thüringer Fernwasserversorgung (Thuringian water supply), which involves challenges like missing values, class imbalance, and collinearity. Finally, we propose policy suggestions to ensure sustainable and equitable access to safe drinking water in order to highlight the broader implications for public health and social equity in Smart Cities.

1 Introduction

Around 71 % of earth's surface is covered with water. Together with carbon, hydrogen, nitrogen, oxygen, phosphorus, and sulfur, it is one of the most important building blocks for all known forms of life [1]. Not only is it essential for sustaining all forms of life, it is fundamental for the growth and development of human civilization, as it plays a leading role in e.g. food production and economic development, and is relevant for general well being [2]. As humans are made up of 70 % water and lose a lot of fluids through sweat and other excretions throughout the day, steady hydration is essential for survival.

Here, the quality of the water plays an important role, as clean and safe drinking water has a major impact on health and life expectancy [3], [4]. For example, there is a clear relationship between contaminated water and cholera [5]. In regions where access to clean drinking water is scarce, infections with diarrhea become a life-threatening danger, as they cause approximately 1.9 million deaths among children globally per year [6], [7]. But even in Germany, where drinking water suppliers have to follow strict regulations regarding water quality [8], *Legionella spp.* which can cause severe pneumonia [9] are found within various water systems, e.g. in private housing, hospitals, or hotels [10]. But also contamination by inorganic substances such as lead, nitrate, fluoride, or aluminum poses a serious health concern [11].

This explains, why drinking water is subjected to strict controls [12]. To ensure the safety and quality of drinking water, the water that comes to people's homes through complex drinking water networks has to be treated by biological, physical or chemical purification in wastewater treatment plants [13], [14]. Because in centralized drinking water networks, impurities in the water can potentially affect several million people, contamination detection is a particular challenge to closely control the quality of the water. The water has to be tested for various pathogens and other contaminants in elaborate laboratory tests [15]. However, the evaluation of these tests takes a relatively long time and are therefore not suitable for online monitoring [16], [17]. Contaminants can sometimes only be detected after people may have already been exposed to them.

Therefore, there is a need for inexpensive, yet effective real-time contamination detection. Modern computational intelligence methods offer an attractive option here [18]. AI-based technologies can not only help people in Western cultures, where the supply of drinking water is already highly developed, but especially in countries in the Global South and improve their quality of life and life expectancy [19]. With the help of machine learning, the quality of drinking water can be reliably controlled with little effort, and the early recognition of unwanted substances in drinking water allows water suppliers to counteract in time [20].

This essay is organized as follows: Section 1.1 discusses the sociological importance of clean drinking water in general, but with an even greater focus in the context of Smart Cities. In section 1.2 we will introduce indicators to measure drinking water quality and its impact on society and discuss their ethical and social implications in section 1.3. Continuing in section 2.1-2.3 we will then utilize a real-world dataset to show, that a data-driven approach to a reliable online contamination detection can be both time- and cost-effective. Finally, in section 2.5 we will propose policy suggestions to ensure a continuous supply of safe and clean drinking water and highlight their implications in section 2.6.

1.1 Smart Cities and Safe Water

Although there is no precise definition of Smart Cities, there is consensus that smart cities aim to address specific challenges such as mobility, energy consumption, waste management, economical sustainability, and public safety [21]. By making data-driven decisions that improve urban living conditions, these cities leverage the possibilities

of modern *Information and Communication Technology* (ICT), such as *Internet of Things* (IoT), data analytics, and *Artificial Intelligence* (AI) to create more interactive, accessible, and efficient urban environments [22].

1.1.1 Smart Cities in the Global South

For countries of the Global South, the implementation of Smart Cities has a particular importance for promoting social equity. These regions often face significant challenges related to rapid urbanization, such as insufficient infrastructure, limited coverage of basic needs, and pronounced socio-economic disparities. By adopting modern ICT, these countries can bridge these gaps, ensuring that all citizens, regardless of their socio-economic status, benefit from essential urban services and enhanced quality of life.

While the benefits of smart technologies in developing countries are obvious, it is necessary to consider the constraints these countries face before they can implement such data-driven approaches. The main barriers are: limitations in education, budget constraints, and lack of policy commitments [23].

1.1.2 The Importance of Safe Water

In 2004, the WHO estimated that out of 6.3 billion people globally, 1.1 billion people drink unsafe water [4]. Therefore, even before the concept of Smart Cities existed, the need for reliable contamination detection was openly discussed and examined, as the demand for such equipment and an early warning system was high, to protect public health [20].

Existing evidence about the strong relation between clean drinking water and life expectancy emphasize the importance of proper water treatment. While the challenge of ensuring a constant supply of safe tap water has largely been resolved for the countries of the Western world, it is not the case that a renewed examination of the issue is only relevant for countries in the Global South. Aging infrastructure and surface level pollution make this challenge even more demanding due to its invisible nature [19].

However, maintaining high drinking water quality poses significant economic challenges, especially in developing countries. The costs associated with water treatment infrastructure, continuous monitoring, and contamination control can be prohibitive. A survey conducted by Dogo, Nwulu, Twala, *et al.* [24] in 2019 reported, that of 182 global water utilities, the annual cost of providing clean water is around USD 184 billion. Unfortunately, many regions struggle to allocate sufficient resources to these efforts, and the economic burden of waterborne diseases intensifies this problem even further, as communities face additional medical costs and lost productivity due to illness.

1.1.3 Advances in AI and Online Contamination Detection

Recent developments in the field of ICT, on the other hand, could accelerate this development of Smart Cities not only in Western countries, but also in developing countries [25]. This is based on the wide availability of data-driven algorithms for recognizing patterns in complex data, such as detecting contamination in drinking water composition. Several studies proved the cost- and time-effectiveness of AI-driven approaches to contamination detection [26]–[29].

Dogo, Nwulu, Twala, *et al.* [24] furthermore revealed, that 41 % of surveyed water utilities still rely on manual collection of water samples for analysis, with only 16 % relying on automated sampling. While 40 % of these utilities would like to have a real-time water quality monitoring system, only 17 % currently have one. The remaining utilities still rely on manual sample collection. Of the aforementioned USD 184 billion spent annually on clean water supply, switching to AI-driven methods could potentially save around USD 12.5 billion [24].

Summing up the above gathered evidence, the recent advances in the field of ICT heavily assist us in the creation and development of Smart Cities. Not only does the technological basis for detecting contamination in real time exist, but several studies have already shown that modern data-driven methods are perfectly capable of analyzing the data collected by this technology. Safe drinking water is key to protecting citizens from water-related diseases, thus making societies healthier and economically more productive. The need for reliable availability of drinking water from countries in both the Western world and the Global South demands that we create a sustainable and resource-efficient foundation for the future.

1.2 Indicators for Safe Water

In order to evaluate the quality of drinking water and its impact on society, various indicators can be evaluated. These indicators comprise the obvious chemical or physical parameters which can be measured using sensors or through laboratory test, but also socio-economic and behavioral indicators can provide information about the safety of drinking water.

1.2.1 Physical, Chemical and Biological Indicators

There are several physical and chemical parameters of drinking water that can serve as direct indicators of the quality and safety of the water. It has been found that a few parameters such as pH, chlorine, total organic carbon (TOC), conductivity, and temperature already provide the most reliable means by which changes in drinking water quality can be measured with no delay [25], [30]. There already exist many affordable sensors on the market, to reliably measure these parameters, which makes evaluating the water quality in real-time viable. Other parameters of online water quality measuring are: turbidity, color, hardness, disinfectants, metals, fluoride, nutrients, hydrocarbon, pesticides, algae, and many more [16].

Unfortunately, there are other many other organic contaminants, which can not be measured in real-time as they often involve the use of indicator organisms or other comprehensive laboratory tests. These kinds of tests allow to precisely determine contamination e.g. with bacteria such as *E. coli*, *Enterococci*, or *Salmonella* [16]. There is an ongoing endeavor towards more complex biological sensors to detect hazardous bio-molecules, but their greatest drawback lies in their disability to detect low concentrations of microorganisms [25]. Although the presence of bio-molecules in drinking water cannot yet be detected sufficiently well in real-time, it is still a valuable indicator of water quality, as it has a direct influence of public health and socio-economic growth [2].

1.2.2 Socio-Economic and Behavioral Indicators

In addition to the above indicators, incorporating socio-economic and behavioral data can provide a more holistic understanding of water safety. These indicators can also offer insights into the underlying causes of water quality issues and the effectiveness of policy suggestions.

One option is tracking the rates of waterborne diseases, such as cholera or dysentery [4], [5]. This can indicate potential issues with water quality. Analyzing correlations between disease outbreaks and water quality data can help identify contamination sources. This requires comprehensive and, above all, area-wide data collection, as drinking water treatment plants are often responsible for large areas. In addition, it is necessary to record which households are connected to which drinking water network in order to be able to draw these kinds of conclusions. Studies that focus on very small areas therefore do not allow any meaningful conclusions to be drawn about the possible causes of the diseases that have occurred.

Another indicator for the quality of drinking water is the socio-economic status of cities and their access to clean water. Disparities in water quality can often be linked to economic inequality, with poorer areas suffering from lower life quality and a higher environmental risk exposure [31]. While the differences within a country's borders may not be so prominent, the differences between The West and countries of the Global South are much more tangible.

Furthermore, the public perception of water quality can be assessed through surveys or social media analysis [32]. For example, public behavior, such as the use of bottled water over tap water serves as an indirect indicator of perceived water safety. Compared to tap water, bottled water is perceived as a more pure and safe alternative [33]. Although more difficult to measure, as these kinds of studies can be skewed by the perception bias of interviewed people [34], they can offer interesting insights into trust in local water authorities and the safety of water.

1.3 Ethical and Social Implications

The application of previously mentioned indicators to assess water quality in Smart Cities carries significant ethical and social implications. These implications span across public health, social equity, trust in public institutions, and the responsible use of technology.

Water is a fundamental human right [35], and failing to provide clean, safe water directly threatens public health [3], [4]. Utilizing physical, chemical, and biological indicators to monitor water quality helps fulfill this ethical duty by enabling timely detection of contaminants. However, the challenge lies in ensuring that these indicators are consistently and accurately monitored across all regions, especially in marginalized communities. If lower-income areas lack resources for comprehensive water quality monitoring, residents might face prolonged exposure to harmful contaminants, which leads to health issues and intensifies socio-economical disparities.

Incorporating socio-economic and behavioral indicators brings additional ethical considerations related to equity and justice. Disparities in water quality often correlate with socio-economic inequalities which suggests that economically disadvantaged populations are more likely to suffer from contaminated water sources [31]. This creates a moral obligation for policymakers to prioritize investments in water infrastructure and quality monitoring in these areas.

Furthermore, the effective implementation of water quality measures requires public trust in water authorities. When residents resort to bottled water due to mistrust in safe tap water, it reflects a failure in communication

and transparency from the authorities [32] or even worse, a lack of governing policies and regulations. This failure can lead to an increased ecological burden due to the increase of waste production, but also to an increased economic burden on households, particularly in low-income areas where the cost of bottled water can be high. This is especially visible in countries of the Global South where the supply of safe tap water is usually scarce, and bottled water is the primary source of hydration [36].

In conclusion, the lack of safe and clean drinking water has many different implications. Besides the obvious consequences on the physical health of residents of cities, there is a much larger and more multifaceted set of socio-economic implications on how contamination of drinking water affects society. Solving the problem of safe and clean drinking water will therefore not only have a beneficial impact on life quality and life expectancy due to the increase in public health, it will furthermore have a strong impact on social equity and social justice.

2 Data Analytics, Optimization and Policy Suggestions

In the second part of this essay, we will implement an actual strategy for detecting drinking water contamination. To do so, we will first investigate a real-world dataset provided by the Thüringer Fernwasserversorgung. We will then identify the necessary steps to prepare the data for building an online monitor to detect contamination. Afterwards, we will compare the performance of different machine learning models to find a suitable candidate for reliably detecting contamination and try to improve this candidate using hyper-parameter optimization. Finally, we will propose policy suggestions based on our findings and discuss their ethical and social implications.

2.1 Dataset

The *Genetic and Evolutionary Computation Conference* (GECCO) is the premier conference in the area of genetic and evolutionary computation and has been held every year since 1999 [37]. For several years now, the organizers have been inviting various industry partners to organize challenges across diverse sectors to engage both *Computational Intelligence* (CI) researchers and practitioners. The goal of the GECCO 2019 Industrial Challenge was to develop an online drinking water contamination detector to accurately predict any kinds of changes in time series of drinking water composition data [18].

For this purpose, the Thüringer Fernwasserversorgung published real-world drinking water composition data from its waterworks in Germany [38]. The data at hand is a time series dataset, containing 218,880 records of drinking water composition data, which were recorded every minute over the time span of almost five months from 2017-07-01 up to and including 2017-11-29 (152 days). These measurements were performed at significant points throughout the water distribution system, in particular at the outflow of the waterworks and the in- and outflow of the water towers [18].

Records comprise a timestamp and various sensor values, i.e. the water's temperature, its pH value, the electric conductivity, the water's turbidity, the spectral absorption coefficient, and the pulse-frequency modulation. However, temperature and the pulse-frequency modulation value are considered operational data, and therefore, changes in those values may indicate changes in the water quality but are not considered events themselves [18]. Finally, the data contains another variable `EVENT`, which describes whether the corresponding entry in the dataset should be considered a contamination or not. Table 1 gives an overview of the structure of the data and figure 1 shows a snapshot of the sensor readings for a randomly selected date, in this case November 13, 2017, with an area highlighted in red where contamination has occurred. The incidents which occurred on this date highlight the difficulty of contamination detection only using drinking water composition data. The first incident around 3:30 PM might be detectable through the sudden increase in turbidity. However, this decision is not that easy for the second incident around 10 PM.

During this five-month period, only 628 anomalies occurred. Figure 2 shows the huge differences in the quantity of data per class. This low amount of contamination events is fortunate for the Thüringer Fernwasserversorgung and their consumers, however, this means a huge class imbalance which has to considered during training, as merely 0.29 % of the dataset contains information about identifying a possibly harmful event.

Due to the nature of technical devices such as sensors, the dataset contains missing values. Especially in the case of the dataset at hand, the level of outliers and missing values is particularly high, as the data originates from sensors that were not continuously maintained for test purposes [18]. There are 2610 rows which contain at least one missing value, or in other words there are in total 15,019 sensor values missing, which makes 1.14 % of the entire dataset.

Finally, figure 3 shows, that the target `EVENT` variable has no correlation with any other variable from the dataset. However, we can observe that `Tp`, `Cond`, and `SAC` have significant pairwise correlations. Correlated variables cause collinearity and can have a negative impact on many predictive models and has to be considered in modeling.

Column	Description	Unit
`Time`	Time of measurement	timestamp
`Tp`	The temperature of the water	°C
`pH`	pH value of the water	pH
`Cond`	Electric conductivity of the water	S/m
`Turb`	Turbidity of the water	FNU
`SAC`	Spectral absorption coefficient	1/m
`PFM`	Pulse-Frequency-Modulation	Hz
`EVENT`	Marker if this entry should be considered as a remarkable change resp. event	boolean

Table 1: Description of drinking water composition data published by Thüringer Fernwasserversorgung for the GECCO 2019 Industrial Challenge: Monitoring of drinking-water quality [38]. The presented data was measured at significant points throughout the water distribution system.

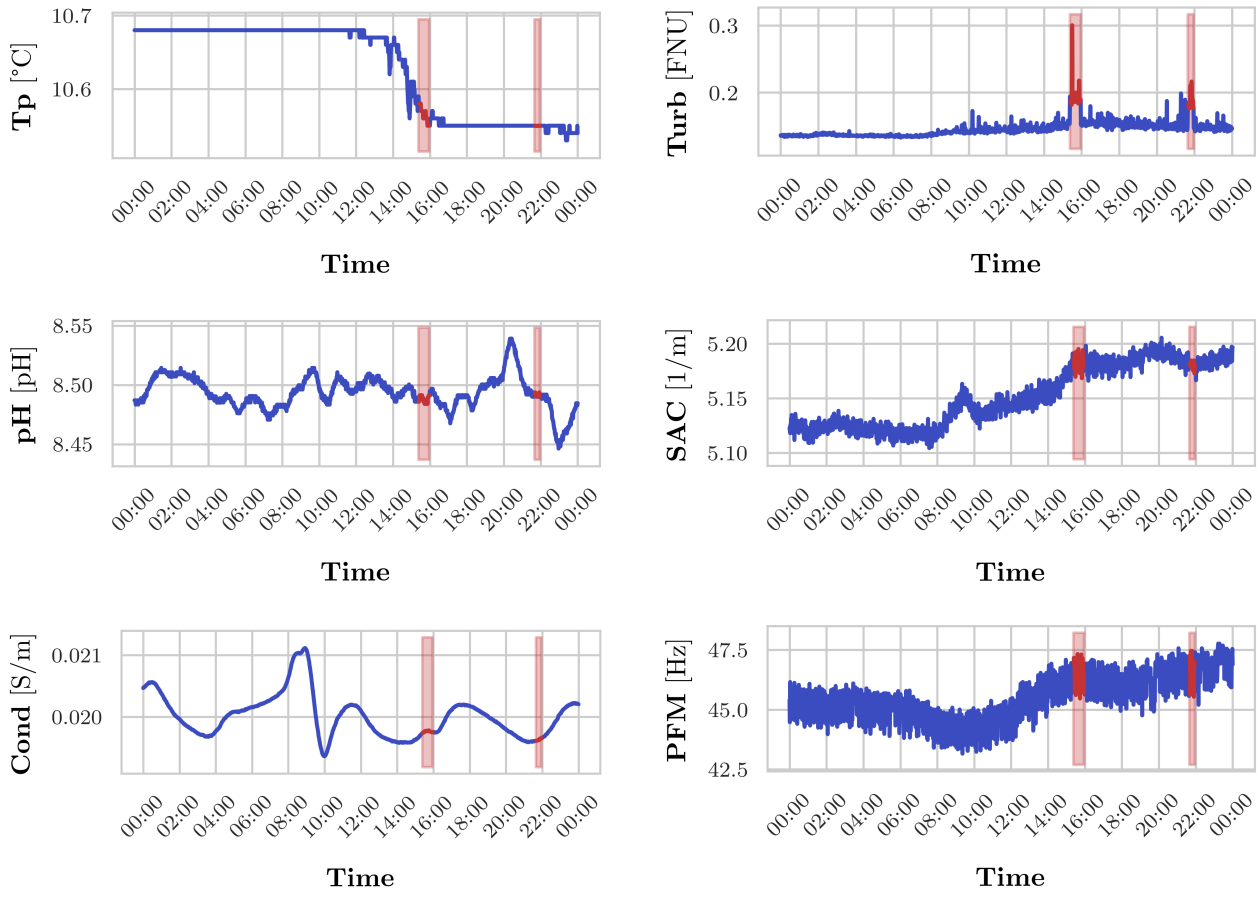

Figure 1: Sensor values for November 13, 2017 with areas of contamination highlighted in red. While the first incident around 3:30 PM looks like it can be detected through the sudden increase in turbidity, this decision is not that easy for the second incident around 10 PM.

2.2 Data Preprocessing

With regard to the continuous measurement of the water composition, it is to be expected that the sensors will not always function properly [39], which can result in missing values. There are many different techniques to deal with this issue, e.g. deleting rows containing missing values, imputing missing values, or using classifiers, which are capable of working with missing values in the first place, e.g. histogram-based gradient boosting.

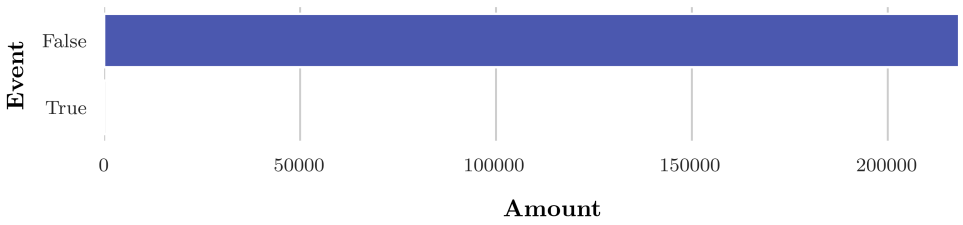

Figure 2: Class imbalance of within the dataset. Out of 218,880 records, only 628 describe the composition of contaminated water. In 218,252 cases, the water was safe. The dataset contains so few entries with contamination that the bar is barely visible.

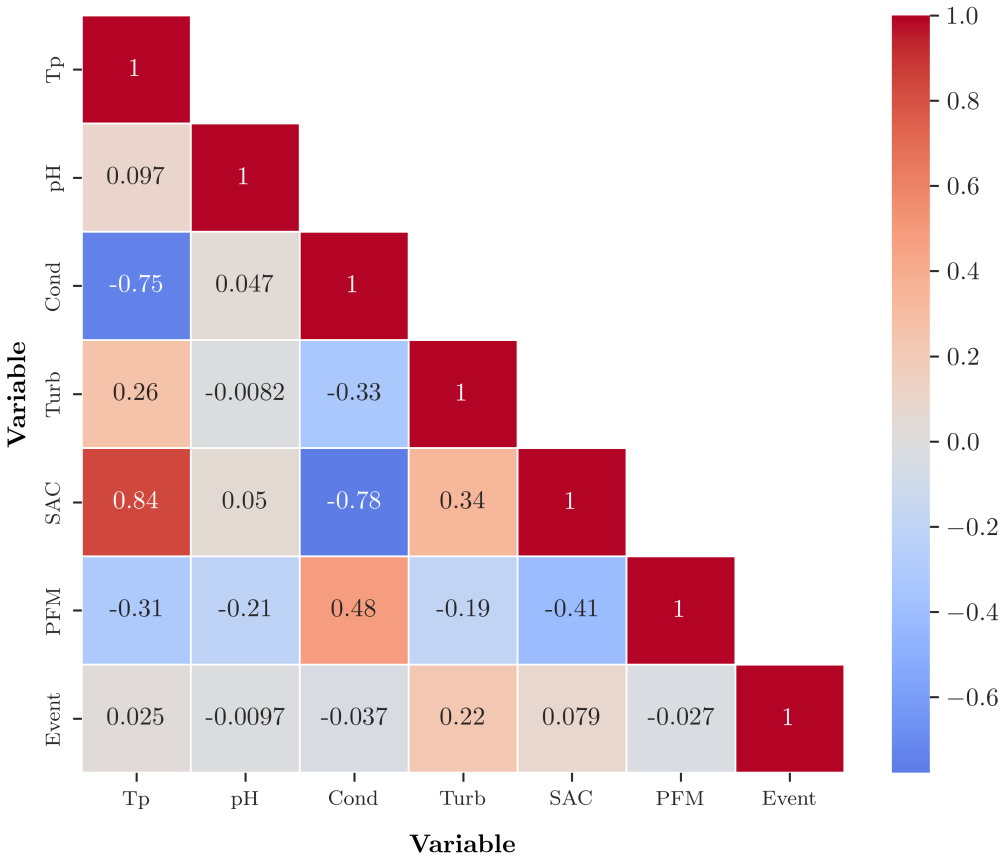

Figure 3: Pairwise correlation matrix. There are significant pairwise correlations between `Tp` and `Cond`, `Tp` and `SAC`, and `Cond` and `SAC`.

But, in order not to go beyond the the scope of this work, the implications and benefits of different imputation techniques are not further elaborated, as they have already been thoroughly studied [40]. Instead, missing values will simply be forward-filled, where a last valid value will be propagated forward. This has already led to sufficiently good results in the past [27].

2.3 Training Procedure

2.3.1 Model Selection

Tree ensembles perform quite well when used for anomaly detection [27]. For evaluating the most suitable tree-based classifier, we examine the following models: gradient boosting (`GradientBoosting`), histogram-based gradient boosting (`HistGradientBoosting`), decision trees (`DecisionTree`), adaptive boosting (`AdaBoost`),

extremely randomized trees (`ExtraTrees`), random forests (`RandomForest`), and extreme gradient boosting with and without dropout (`xgbTree` and `xgbDART` respectively).

2.3.2 Evaluation

The above mentioned tree ensembles will be evaluated in terms of their performance in predicting water contamination by utilizing 5-fold cross-validation with shuffled stratified folds and a 4:1 ratio between training data and test data, as it is done in comparable studies [40]. Despite the fact that we are dealing with time series data, due the short time span of the training data, it is not effective to use time series splits for cross-validation [27].

The models' performance is assessed by comparing their F_1 scores, which is not only the recommended performance indicator according to the editors of the challenge [18], but is also an accurate metric when dealing with high class imbalance, as it penalizes classifiers performing poorly on minority classes, considering the balance between recall and precision [41]. Higher F_1 scores mean better performance.

Given the amount of *True Positives* (TP), *False Positives* (FP), and *False Negatives* (FN), the F_1 score can be computed as

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

with

$$Precision = \frac{TP}{TP + FP}$$
$$Recall = \frac{TP}{TP + FN}.$$

2.3.3 Class Imbalance

Machine learning algorithms by design assume a balanced class distribution [42]. Therefore, in many experiments with unequal amount of records per class it is necessary to apply resampling methods such as over-sampling the minority class or under-sampling the majority class. However, with the exception of AdaBoost, the performance of tree-based classifiers does not significantly benefit from resampling methods, at least not regarding the problem domain of contamination detection in drinking water networks, as it has been shown that predictive performance of decision trees and random forests remains the same or even worsens when applying resampling methods [40].

2.3.4 Normalization and Standardization

Another great advantage of tree-based classifiers is, that they don't require any kind of normalization or standardization [43]. This is due to the fact, that, in contrast to other machine learning methods, they don't rely on the distance between data points, but instead on the order of the data. As normalization is a monotonous transformation, it preserves the order of values and therefore does not provide any benefit. This drastically simplifies the training pipeline.

2.3.5 Collinearity

Our analysis of the dataset revealed that three variables are significantly pairwise correlated: `Tp`, `Cond`, and `SAC`. There are many different ways to deal with collinearity in datasets. For example, the dimension of the existing data can be reduced by *Principal Component Analysis* (PCA) and thus strongly correlated information can be merged. However, the experiments conducted as part of this essay have shown that any measures against collinearity have only led to a reduction in prediction performance.

2.3.6 Results

In table 2 you can see the results of 5-fold cross-validation. Overall, there are significant differences in the performance of reliable contamination detection for the examined classifiers. While some models fail to accurately predict contamination, the `ExtraTrees` classifier seems to detect anomalies in drinking water composition with ease. It reached a mean F_1 score of 0.9311. Another very important metric in this context is the amount of false positives and false negatives. From 43,776 predictions in total, only 2.6 predictions were false positives, which means that the classifier detected contamination where there was none. The much more important metric, however, is probably are probably the amount of false negatives, as it measures the amount of undetected contamination. A high number of false negatives indicates, that the model is unable to detect ongoing contagion.

While the `ExtraTrees` classifier is not perfect in this aspect, it is still the best performing model. Together with an efficient training time, this makes the `ExtraTrees` model our favored candidate. Without any hyper-parameter optimization, the results from this evaluation already allow for a relatively cost- and time-efficient online contamination detection.

Algorithm	Training	TP	TN	FP	FN	F_1
GradientBoosting	22.86s	77.8	43,638.8	11.6	47.8	0.7228 ± 0.0482
HistGradientBoosting	**0.20s**	78.0	43,619.6	30.8	47.6	0.6568 ± 0.0984
DecisionTree	1.12s	107.8	43,633.8	16.6	17.8	0.8633 ± 0.0343
AdaBoost	4.82s	75.6	43,641.6	8.8	50.0	0.7193 ± 0.0242
ExtraTrees	3.01s	**111.8**	**43,647.8**	**2.6**	**13.8**	$\mathbf{0.9311} \pm 0.0260$
RandomForest	20.64s	104.6	**43,647.8**	**2.6**	21.0	0.8986 ± 0.0075
xgbDART	33.77s	109.0	43,645.0	5.4	16.6	0.9082 ± 0.0125
xgbTree	**0.20s**	109.8	43,644.6	5.8	15.8	0.9102 ± 0.0152

Table 2: Performance of various tree-based classifiers for 5-fold cross-validation with shuffled stratified folds.

2.4 Feature Importance

The use of tree-based classifiers allows us to determine the relative importance of drinking water composition parameters without much effort, as this information is a by-product of the training process. Figure 4 visualizes this and indicates that the spectral absorption coefficient and the turbidity of the water are most important for successful contamination detection. This also provides valuable insights into which sensors should be treated with particular care. Outage of these sensors poses a higher risk of false predictions, as the information density in these parameters is higher as for example in the water's pH value or the pulse-frequency modulation value.

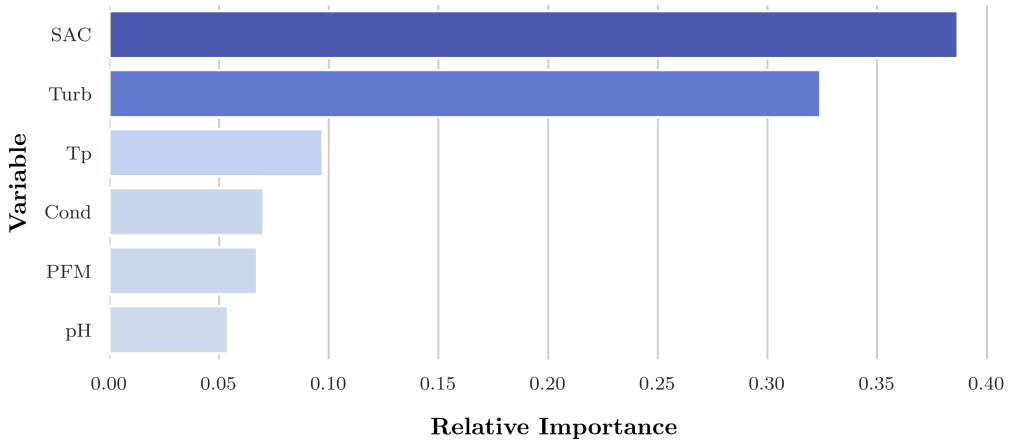

Figure 4: The relative importance of water composition parameters. Most important for contamination detection are `SAC` and `Turb`.

2.4.1 Hyper-parameter Optimization

Training a model and evaluating its performance usually also comprises hyperparameter optimization, as this results in a better performing classifier. However, the ideal configuration of hyper-parameters is often tightly related to the actual problem at hand and the underlying dataset [44].

While not being the most efficient way to optimize the hyper-parameters of a machine learning model, exhaustive grid search cross-validation will most likely offer the ideal configuration when given the right search space [44]. After several iterations of optimization, the following parameters were found to yield the best performance for contamination detection using an `ExtraTrees` classifier:

- `n_estimators`: 1000

- `criterion`: `"entropy"`
- `max_depth`: `None`
- `min_samples_split`: 2
- `min_samples_leaf`: 1
- `min_weight_fraction_leaf`: 0
- `max_leaf_nodes`: `None`
- `min_impurity_decrease`: 0.

The above specified hyper-parameters have a significant negative impact on the training time of the `ExtraTrees` classifier. While training previously took around 3 seconds to finish, the mean training time for 5-fold cross-validation now is over 31 seconds. This makes sense as the default number of estimators is 100, and as we increased the number of estimators by a factor of 10, the training also took almost 10 times as long.

However, the increase in performance achieved is rather disappointing. The F_1 score for the optimized parameters is 0.9369 which is an increase of only 0.10 %. This increase is in no way proportionate to the increased resource requirements and is therefore not advisable.

2.5 Policy Suggestions

In the previous sections we have seen that an effective online detector of drinking water contamination can be achieved with relatively little effort. Based on these findings, we will propose several policy suggestions to enhance drinking water safety and address contamination issues effectively. These policies aim to improve the supply of clean water for the population and thereby protect public health.

The first policy suggestion is the governmental requirement of comprehensively deploying smart sensors and their maintenance. These sensors should be installed at critical points within the water distribution network, including water treatment plants, storage facilities, and ideally in the pipes of the distribution network. Continuous monitoring through these sensors will provide comprehensive data on water quality in real-time. Additionally, predictive maintenance is necessary to minimize outages of these sensors and therefore reduce the amount of missing values.

A centralized data platform should be established to collect, integrate, and store data from all sensors. This platform must support interoperability with existing water management systems and allow for easy integration of new sensor technologies. By centralizing data collection, we can utilize the collected data for further improving the performance of the contamination detector.

The third suggestion involves the utilization of machine learning models to detect contamination. Machine learning models should be developed and trained using both historical and real-time data to identify patterns and anomalies indicative of contamination. Implementing real-time anomaly detection algorithms will enable the system to detect potential contamination events as soon as possible and allow for immediate investigation and response.

Especially in the beginning, where a lot of data has first to be collected and the contamination detector has to be properly trained, there might be some false predictions. To build trust for the deployed systems, it is therefore recommended to have a tried and tested backup system running simultaneously and first assess the performance of the novel technology before deriving any kind of action from its predictions.

Finally, in the case of a detected contamination, relevant authorities have to be notified immediately through a robust alert system. For this, a standardized response protocol is required to quickly investigate and mitigate contamination events. For example, this protocol may include shutting down affected water lines, issuing public advisories, or deploying emergency water treatment measures.

The above mentioned steps allow the Smart City to ensure a continuous supply of safe and clean drinking water. Furthermore, the suggested policy protects public health by reducing the risk of waterborne diseases and other health issues associated with contaminated drinking water. Operational efficiency is enhanced through predictive maintenance. This not only reduces downtime and resource wastage but also results in cost savings by preventing large-scale contamination incidents.

2.6 Ethical and Social Implications of Policy Suggestions

The above suggested policies carry several ethical and social implications that must be considered carefully before actually implementing them.

One primary ethical concern is data security. Due to the extensive use of smart sensors and centralized data platforms, it is necessary to ensure that this data is protected against breaches or tampering. Unauthorized access to such data could lead to malicious actions, such as dispatching false alarms or hiding an ongoing contamination. The infrastructure introduced by these regulations must be classified as critical infrastructure and secured and guarded accordingly to mitigate these risks.

Another ethical question concerns equity and access. Wealthier areas might receive more immediate and thorough implementation of these technologies, while economically disadvantaged communities could be left with outdated or insufficient water monitoring systems. This disparity could exacerbate existing inequalities in access to safe drinking water. Therefore, it has to be ensured, that the deployed technologies are evenly distributed across different regions and communities, regardless of their economic status.

Transparency and public trust are also significant social implications of this policy. False predictions can lead to panic among the population if not handled correctly. Thus, every prediction made by the online contamination detector has to be verified by a human, to also build trust in this novel technology.

The policy suggestions also raise questions about the potential impact on employment and the reduction of workforce within the water treatment plant. The integration of advanced technologies for monitoring and maintenance may reduce the need for certain manual labor tasks. Of course, they may create new job opportunities in areas such as data analysis, cybersecurity, or technology maintenance, but these opportunities do not exist for the currently employed workers. It is therefore an ethical responsibility to ensure that transition programs exist for displaced workers.

Last but not least, the suggested policies raise ethical concerns regarding responsibility and liability in the event of an incorrect prediction. If the system falsely predicts contamination where none exists, it could lead to unnecessary panic, costly interventions, and a loss of public trust in the water management system. Conversely, a false negative, where the system fails to predict an actual contamination event, poses severe risks to public health. It is therefore essential that the question of liability is clear at all times.

In conclusion, when utilizing modern CI techniques for drinking water contamination detection, there are numerous benefits for public health and how fast incidents can be detected. However, they also bring with them several ethical and social implications. These include concerns about data security, equity and access, public trust, impact on the workforce, and liability in the event of an incorrect prediction. These issues need to be addressed first to ensure that these measures not only achieve their technological and health-related objectives, but also uphold social justice and public safety.

3 Conclusion

In this essay we found that ensuring access to clean and safe drinking water is a critical challenge that Smart Cities must address to enhance public health and societal well-being. Furthermore, we highlighted the potential of data-driven approaches, particularly the use of AI and machine learning, to improve the efficiency and effectiveness of real-time water contamination detection. By utilizing a real-world dataset by the Thüringer Fernwasserversorgung, we demonstrated that these technologies can offer cost-effective solutions that benefit both developed and developing regions.

Moving forward, it is essential for policymakers to support the integration of these technologies into water management, while always taking into account their ethical and social implications, in order to ensure sustainable and equitable access to safe drinking water.

References

[1] NASA. "Ingredients for life." (2024), [Online]. Available: https://europa.nasa.gov/why-europa/ingredients-for-life/ (visited on 05/21/2024).

[2] P. K. Goel, *Water Pollution: Causes, Effects and Control*. New age international, 2006.

[3] E. Hertz, J. R. Hebert, and J. Landon, "Social and environmental factors and life expectancy, infant mortality, and maternal mortality rates: Results of a cross-national comparison," *Social Science & Medicine*, vol. 39, no. 1, pp. 105–114, 1994, ISSN: 0277-9536. DOI: https://doi.org/10.1016/0277-9536(94)90170-8. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0277953694901708.

[4] N. J. Ashbolt, "Microbial contamination of drinking water and disease outcomes in developing regions," *Toxicology*, vol. 198, no. 1-3, pp. 229–238, 2004.

[5] S. Gundry, J. Wright, and R. Conroy, "A systematic review of the health outcomes related to household water quality in developing countries," *Journal of water and health*, vol. 2, no. 1, pp. 1–13, 2004.

[6] T. S. Steiner, A. Samie, and R. L. Guerrant, "Infectious diarrhea: New pathogens and new challenges in developed and developing areas," *Clinical Infectious Diseases*, vol. 43, no. 4, pp. 408–410, Aug. 2006, ISSN: 1058-4838. DOI: 10.1086/505874. eprint: https://academic.oup.com/cid/article-pdf/43/4/408/1095680/43-4-408.pdf. [Online]. Available: https://doi.org/10.1086/505874.

[7] J. Bryce, C. Boschi-Pinto, K. Shibuya, and R. E. Black, "Who estimates of the causes of death in children," *The lancet*, vol. 365, no. 9465, pp. 1147–1152, 2005.

[8] K. Lauterbach, B. Pistorius, C. Özdemir, V. Wissing, and S. Lemke. "Zweite verordnung zur novellierung der trinkwasserverordnung." (2023), [Online]. Available: https://www.recht.bund.de/eli/bund/bgbl_1/2023/159 (visited on 05/17/2024).

[9] B. S. Fields, R. F. Benson, and R. E. Besser, "*Legionella* and legionnaires' disease: 25 years of investigation," *Clinical Microbiology Reviews*, vol. 15, no. 3, pp. 506–526, 2002. DOI: 10.1128/cmr.15.3.506-526.2002. eprint: https://journals.asm.org/doi/pdf/10.1128/cmr.15.3.506-526.2002. [Online]. Available: https://journals.asm.org/doi/abs/10.1128/cmr.15.3.506-526.2002.

[10] E.-B. Kruse, A. Wehner, and H. Wisplinghoff, "Prevalence and distribution of legionella spp in potable water systems in germany, risk factors associated with contamination, and effectiveness of thermal disinfection," *American Journal of Infection Control*, vol. 44, no. 4, pp. 470–474, 2016, ISSN: 0196-6553. DOI: https://doi.org/10.1016/j.ajic.2015.10.025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0196655315011116.

[11] R. M. Harrison, *Pollution: Causes, Effects and Control*. Royal society of chemistry, 2001.

[12] WHO *et al.*, *Guidelines for Drinking-Water Quality: Incorporating the First and Second Addenda*. World Health Organization, 2022.

[13] C. P. L. Grady Jr., G. T. Daigger, N. G. Love, and C. D. M. Filipe, *Biological Wastewater Treatment*. CRC press, 2011.

[14] S. Sharma and A. J. A. W. S. Bhattacharya, "Drinking water contamination and treatment techniques," *Applied water science*, vol. 7, no. 3, pp. 1043–1067, 2017.

[15] S. D. Richardson, "Environmental mass spectrometry: Emerging contaminants and current issues," *Analytical chemistry*, vol. 80, no. 12, pp. 4373–4402, 2008.

[16] S. N. Zulkifli, H. A. Rahim, and W.-J. Lau, "Detection of contaminants in water supply: A review on state-of-the-art monitoring technologies and their applications," *Sensors and Actuators B: Chemical*, vol. 255, pp. 2657–2689, 2018.

[17] F. Y. Ramírez-Castillo, A. Loera-Muro, M. Jacques, *et al.*, "Waterborne pathogens: Detection methods and challenges," *Pathogens*, vol. 4, no. 2, pp. 307–334, 2015, ISSN: 2076-0817. DOI: 10.3390/pathogens4020307. [Online]. Available: https://www.mdpi.com/2076-0817/4/2/307.

[18] F. Rehbach, S. Moritz, and T. Bartz-Beielstein. "Gecco 2019 industrial challenge: Monitoring of drinking-water quality." (2019), [Online]. Available: https://www.th-koeln.de/mam/downloads/deutsch/hochschule/fakultaeten/informatik_und_ingenieurwissenschaften/rulesgeccoic2019.pdf (visited on 05/09/2024).

[19] A. N. Angelakis, H. S. Vuorinen, C. Nikolaidis, *et al.*, "Water quality and life expectancy: Parallel courses in time," *Water*, vol. 13, no. 6, 2021, ISSN: 2073-4441. DOI: 10.3390/w13060752. [Online]. Available: https://www.mdpi.com/2073-4441/13/6/752.

[20] D. Byer and K. H. Carlson, "Real-time detection of intentional chemical contamination in the distribution system," *Journal-American Water Works Association*, vol. 97, no. 7, 2005.

[21] H. Chourabi, T. Nam, S. Walker, *et al.*, "Understanding smart cities: An integrative framework," in *2012 45th Hawaii international conference on system sciences*, IEEE, 2012, pp. 2289–2297.

[22] S. Pellicer, G. Santa, A. L. Bleda, R. Maestre, A. J. Jara, and A. G. Skarmeta, "A global perspective of smart cities: A survey," in *2013 Seventh International conference on innovative mobile and internet services in ubiquitous computing*, IEEE, 2013, pp. 439–444.

[23] S. Chatterjee and A. K. Kar, "Smart cities in developing economies: A literature review and policy insights," in *2015 international conference on advances in computing, communications and informatics (ICACCI)*, IEEE, 2015, pp. 2335–2340.

[24] E. M. Dogo, N. I. Nwulu, B. Twala, and C. Aigbavboa, "A survey of machine learning methods applied to anomaly tetection on drinking-water quality data," *Urban Water Journal*, vol. 16, no. 3, pp. 235–248, 2019.

[25] M. V. Storey, B. Van der Gaag, and B. P. Burns, "Advances in on-line drinking water quality monitoring and early warning systems," *Water research*, vol. 45, no. 2, pp. 741–747, 2011.

[26] G. Kang, J. Z. Gao, and G. Xie, "Data-driven water quality analysis and prediction: A survey," in *2017 IEEE third international conference on big data computing service and applications (BigDataService)*, IEEE, 2017, pp. 224–232.

[27] M. Nguyen and D. Logofătu, "Applying tree ensemble to detect anomalies in real-world water composition dataset," in *Intelligent Data Engineering and Automated Learning – IDEAL 2018*, H. Yin, D. Camacho, P. Novais, and A. J. Tallón-Ballesteros, Eds., Cham: Springer International Publishing, 2018, pp. 429–438, ISBN: 978-3-030-03493-1.

[28] K. Qian, J. Jiang, Y. Ding, and S. Yang, "Deep learning based anomaly detection in water distribution systems," in *2020 IEEE International Conference on Networking, Sensing and Control (ICNSC)*, IEEE, 2020, pp. 1–6.

[29] F. Muharemi, D. Logofătu, and F. Leon, "Machine learning approaches for anomaly detection of water quality on a real-world data set," *Journal of Information and Telecommunication*, vol. 3, no. 3, pp. 294–307, 2019. DOI: 10.1080/24751839.2019.1565653. eprint: https://doi.org/10.1080/24751839.2019.1565653. [Online]. Available: https://doi.org/10.1080/24751839.2019.1565653.

[30] U. EPA, "Watersentinel online water quality monitoring as an indicator of drinking water contamination," 2005.

[31] G. W. Evans and E. Kantrowitz, "Socioeconomic status and health: The potential role of environmental risk exposure," *Annual review of public health*, vol. 23, no. 1, pp. 303–331, 2002.

[32] M. de França Doria, N. Pidgeon, and P. R. Hunter, "Perceptions of drinking water quality and risk and its effect on behaviour: A cross-national study," *Science of the total environment*, vol. 407, no. 21, pp. 5455–5464, 2009.

[33] C. Ferrier, "Bottled water: Understanding a social phenomenon," *AMBIO: A journal of the Human Environment*, vol. 30, no. 2, pp. 118–119, 2001.

[34] S. Turgeon, M. J. Rodriguez, M. Thériault, and P. Levallois, "Perception of drinking water in the quebec city region (canada): The influence of water quality and consumer location in the distribution system," *Journal of environmental management*, vol. 70, no. 4, pp. 363–373, 2004.

[35] P. H. Gleick, "The human right to water," *Water policy*, vol. 1, no. 5, pp. 487–503, 1998.

[36] B. Rani, R. Maheshwari, A. Garg, and M. Prasad, "Bottled water–a global market overview," *Bulletin of Environment, Pharmacology and Life Sciences*, vol. 1, no. 6, pp. 1–4, 2012.

[37] Various. "Genetic and evolutionary computation conference." (2024), [Online]. Available: https://en.wikipedia.org/wiki/Genetic_and_Evolutionary_Computation_Conference (visited on 05/21/2024).

[38] F. Rehbach, S. Moritz, and T. Bartz-Beielstein, *Gecco industrial challenge 2019 dataset: A water quality dataset for the 'internet of things: Online event detection for drinking water quality control' competition at the genetic and evolutionary computation conference 2019, prague, czech republic.* Feb. 2019. DOI: 10.5281/zenodo.4304080. [Online]. Available: https://doi.org/10.5281/zenodo.4304080.

[39] S. Panguluri, G. Meiners, J. Hall, and J. G. Szabo, "Distribution system water quality monitoring: Sensor technology evaluation methodology and results," *US Environ. Protection Agency, Washington, DC, USA, Tech. Rep. EPA/600/R-09/076*, vol. 2772, 2009.

[40] E. M. Dogo, N. I. Nwulu, B. Twala, and C. O. Aigbavboa, "Empirical comparison of approaches for mitigating effects of class imbalances in water quality anomaly detection," *IEEE Access*, vol. 8, pp. 218 015–218 036, 2020. DOI: 10.1109/ACCESS.2020.3038658.

[41] M. Bekkar, H. K. Djemaa, and T. A. Alitouche, "Evaluation measures for models assessment over imbalanced data sets," *J Inf Eng Appl*, vol. 3, no. 10, 2013.

[42] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge university press, 2014.

[43] L. B. V. de Amorim, G. D. C. Cavalcanti, and R. M. O. Cruz, "The choice of scaling technique matters for classification performance," *Applied Soft Computing*, vol. 133, p. 109 924, 2023.

[44] T. Yu and H. Zhu, "Hyper-parameter optimization: A review of algorithms and applications," *arXiv preprint arXiv:2003.05689*, 2020.