The application of artificial intelligence in water transportation systems

Lučin, Ivana

Doctoral thesis / Disertacija

2022

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Rijeka, Faculty of Engineering / Sveučilište u Rijeci, Tehnički fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:190:564948

Download date / Datum preuzimanja: **2024-07-16**

Repository / Repozitorij:

Repository of the University of Rijeka, Faculty of Engineering

UNIVERSITY OF RIJEKA
FACULTY OF ENGINEERING

Ivana Lučin

THE APPLICATION OF ARTIFICIAL INTELLIGENCE IN WATER TRANSPORTATION SYSTEMS

DOCTORAL DISSERTATION

Rijeka, 2022.

UNIVERSITY OF RIJEKA
FACULTY OF ENGINEERING

Ivana Lučin

THE APPLICATION OF ARTIFICIAL INTELLIGENCE IN WATER TRANSPORTATION SYSTEMS

DOCTORAL DISSERTATION

Thesis Supervisor: Prof. D. Sc. Zoran Čarija
Thesis Co-supervisor: Prof. D. Sc. Siniša Družeta

Rijeka, 2022.

SVEUČILIŠTE U RIJECI
TEHNIČKI FAKULTET

Ivana Lučin

PRIMJENA UMJETNE INTELIGENCIJE U SUSTAVIMA TRANSPORTA VODE

DOKTORSKA DISERTACIJA

Mentor: prof. dr. sc. Zoran Čarija
Komentor: prof. dr. sc. Siniša Družeta

Rijeka, 2022.

Thesis Supervisor: Prof. D. Sc. Zoran Čarija, University of Rijeka, Faculty of Engineering

Thesis Co-supervisor: Prof. D. Sc. Siniša Družeta, University of Rijeka, Faculty of Engineering

This doctoral dissertation was discussed on _____ at the University of Rijeka, Croatia, Faculty of Engineering in front of the following Evaluation Committee:

1.

2.

3.

Abstract

Water distribution systems are designed to assure safe water transportation to the end-users. Since the water needs to have required quality and hydraulic characteristics, these systems are regularly monitored, controlled, and improved. In this doctoral dissertation, an investigation of different applications of artificial intelligence methods for the purpose of improving water distribution systems was conducted. Firstly, the optimization procedure coupled with numerical simulations is used for improving the design of the parts of the water system intake structure. In the further investigation of optimization applications, pollution detection strategy is developed, where novel optimization approach based on search space reduction method and independent optimizations conducted for each possible source node is proposed. Machine learning has been applied in the prediction of a number of pollution sources, based on a wide range of pollution scenarios with a various number of pollution sources. Additionally, machine learning has been used for leak localization, based on a wide range of leak scenarios. As a further development of leak localization methodology, pipe segmentation approach was proposed in which additional divisions of pipes were introduced to simulate a more realistic scenario where leaks can occur not only at pipe junctions but at any point of pipe. The conducted research showed several new possible utilizations of artificial intelligence methods which were previously not considered mainly due to their considerable computational demand. These applications need to be further explored since with the rapid increase of computational power these methods could provide valuable insight into water system behavior and improve water transportation system operation.

Keywords: water distribution systems, shape optimization, leak localization, pollution localization, Random Forest

Sažetak

Sustavi transporta vode služe za opskrbu različitih korisnika pri čemu je glavna funkcija sustava osiguranje želje kvalitete vode i njenih hidrauličkih karakteristika. Problemi u sustavu mogu uzrokovati značajne gubitke, trajna oštećenja, a u konačnosti mogu predstavljati opasnost za ljudske živote, te se zbog toga sustavi transporta vode redovito prate i reguliraju. S povećanjem količine dostupnih mjerenja kao i s povećanjem računalnih resursa, primjena umjetne inteligencije prilikom dizajniranja i kontrole sustava transporta vode postala je sve zastupljenija. U ovoj doktorskoj disertaciji predloženo je nekoliko novih smjerova primjene umjetne inteligencije u svrhu poboljšanja sustava transporta vode. Prvi od istraženih smjerova je optimizacijski pristup koji je primjenjen za poboljšanje dizajna djelova ulazne strukture sustava transporta vode, konkretno zaštitne rešetke. Primjenom optimizacijskih metoda moguće je prilagoditi geometriju poprečnog presjeka kako bi se minimizirali hidraulički gubitci uz zadovoljenje ekoloških i inženjerskih zahtjeva. Optimizacijski pristup je primjenjen i na problem detekcije mjesta unosa onečišćenja u sustav transporta vode. U slučaju pojave onečišćenja u sustavu potrebno je brzo odrediti lokaciju i parametre onečišćenja, u cilju upozorenja korisnika i poduzimanja potrebnih zaštitnih radnji. Primjenom nove metode koja smanjuje broj potencijalnih čvorova unosa onečišćenja, za svaki preostali potencijalni čvor proveden je zaseban optimizacijski postupak, čime je smanjena dimenzionalnost problema što pojednostavljuje i ubrzava optimizacijski postupak. Nadalje, strojno učenje primjenjeno je za predviđanje nekoliko mogućih lokacija unosa onečišćenja na temelju ograničenih senzorskih mjerenja. Predikcijski model je istreniran na sintetičkim mjerenjima dobivenim iz većeg broja numeričkih simulacija provedenih za varijabilni broj lokacija onečišćenja i za varijabilne parametre unosa onečišćenja. Slična metodologija provedena je i za određivanje mjesta oštećenja cjevovoda, gdje je predikcijski model istreniran na sintetičkim podatcima o izmjerenim tlakovima, koji su dobiveni

iz većeg broja simulacija sa varijabilnim mjestom i veličinom oštećenja. Za razliku od standardne metodologije, u kojoj se pretpostavlja da se oštećenje dogodilo u nekom od čvorova vodovodne mreže, u ovom radu predlaže se novi pristup u kojem se nakon preliminarne lokalizacije oštećenja provodi dodatna segmentacija cijevi kako bi se lokacija oštećenja mogla točnije odrediti. Provedeno istraživanje pokazalo je da se metode umjetne inteligencije danas mogu uspješno primjeniti na probleme koji se prethodno nisu rješavali na ovaj način, ponajviše zbog prevelikih računalnih zahtjeva. Metode predložene u ovom radu pokazuju da se povećanjem računalnih resursa i korištenjem poboljšanih tehnika umjetne inteligencije može poboljšati rad i kontrola sustava transporta vode.

Contents

Part I

Introduction

Chapter 1

Introduction

Water transportation systems are designed for the safe distribution of water from the catchment area to the end-users, e.g. industry or households. Depending on the designated purpose (technical water, drinking water etc.), water needs to have the required quality and hydraulic characteristics. To obtain these attributes, water transportation systems are regularly monitored and controlled to provide optimal system operation [46, 85, 93].

The first step for achieving this goal is the appropriate design of water intake, where trash racks or screens are installed so as to prevent the entrance of debris or fish in the water distribution system [11, 89, 37]. Trash racks or screens can cause additional clogging due to debris accumulation or can reduce system efficiency due to losses caused by flow disturbance, hence they should be optimally designed [73, 33, 98]. Subsequent water treatment can be conducted using filters and water purification systems if water is to be used as drinking water [10, 23]. Although precautions are being taken, unexpected events such as accidental or intentional pollution intrusion in water distribution network can occur [84, 72]. Intrusions of pollution can also occur in pipe leak locations where contaminated soil may enter the pipe under certain conditions [53, 9]. These intrusions can cause serious health problems to the end-users, thus sensors are installed in water distribution networks for water quality monitoring [32, 65]. In the case of an accident event, various mathematical and statistical techniques based on the sensor measurements can be used to identify conditions under which the accident occurred.

With growing technological trends such as Smart Cities and the Internet of Things, the complexity of water distribution systems is continually increasing and consequently

available amount of data which can provide valuable insight in system operation. Additionally, improvements to the existing infrastructure are constantly being implemented. These trends indicate a strong need for computational techniques that can process and analyze obtained data and consequently provide engineers with useful knowledge which can enhance water transportation systems operation.

1.1   Accident events

1.1.1   Structural failures

Structural failures in water transport systems can be caused by various factors, e.g. vibrations due to fluid-structure interaction [44, 80, 86], impact of large debris collision with trash-racks, or smaller debris entering water distribution systems and causing damage to system parts [17, 21]. For this reason, the geometry of intake structures needs to be carefully considered to provide good protection and sustain debris load while at the same time produce minimal disturbance of fluid flow. This presents an optimization problem since opposing goals need to be satisfied [14].

Additionally, structural failures can occur due to material deterioration [42, 67, 15]. Various factors can influence pipe material deterioration, such as temperature variation, soil influence, age, stress due to pressure changes, etc. Due to expensive installation, water distribution network pipes are not regularly substituted, which can cause corrosion and cracks in the material over the long term. The problem is that when a leak occurs, it can precipitate material deterioration, which can ultimately lead to pipe bursts. Pipe bursts are the greatest problem since they cause serious damage and losses, thus a number of papers have considered methods for predicting pipe failures [78, 22, 7, 87, 19]. A greater pipe burst can cause flooding of a populated or industrialized area, which can cause considerable material losses, while the consequent water supply outage can last for an extended period of time. Therefore, different methods for detecting and localizing leak and burst locations have been explored [16, 88, 45, 35].

1.1.2   Water quality failure

Water quality failures can occur because of deliberate contamination injection in water distribution networks or due to accident events. Pollution intrusion through leak location is serious issue, since the soil at the location of the leak may be contaminated with harmful micro-organisms and pathogens [41, 39, 9, 24]. Additionally, pipe corrosion can lead to reduced water quality, i.e. occurrence of "red water" [54]. In the case of these incidents, it is of utmost importance to rapidly determine the location of intrusion, starting time, duration of intrusion, and contamination concentration. With these parameters identified, simulations of pollution spreading through water distribution networks can be conducted. Simulation of contamination spreading identifies which parts of the water distribution network, and in what amount, have been contaminated. On the basis of this knowledge, required actions can be taken, such as prevention of further contamination and warning of users in the contaminated area.

The presented problem is difficult to solve since it is an inverse problem, where based on sensor measurements, different analytical techniques need to be employed to determine causal factors of the event. For solving this problem optimization methods coupled with numerical simulations are predominantly used, which is known as the simulation-optimization approach [62, 63, 2, 91, 92, 81, 90]. In this approach, optimization methods try to find contamination parameters for the numerical simulation of pollution scenario which will produce results most similar to the sensor measurements obtained from the real event. In recent years machine learning approach is also often employed, where machine learning algorithms for prediction of pollution parameters are trained on a wide range of simulated contamination scenarios [51, 28, 29, 30].

1.2   Numerical simulations in water transportation systems

Due to large scale and great complexity of water systems, model testing is usually not the feasible approach for analyzing existing systems or investigating possible improvements. In-field measurements usually provide only limited information, thus these measurements are typically used for calibration of the numerical model, which is then

used for obtaining detailed information regarding fluid flow. Therefore, numerical simulations are increasingly being used to enhance the existing design, to provide better insight into system behavior under different conditions, or for designing a new infrastructure. Numerical simulations can be used for one-dimensional, two-dimensional, and three-dimensional fluid flow analysis, depending on the considered problem. One-dimensional simulation represent flows through long pipes, preferably circular pipes, which are mostly used for water distribution systems. The most widely used software for this purpose is EPANET, a public domain software developed by US Environmental Protection Agency [71]. This software enables fast simulation of system behavior even for most complex networks. However, this speed is due to considerable simplifications of its flow model, such as the assumption of complete mixing at junctions, the assumption of constant pressure and velocity values along a pipe, etc. However, wrong results can be obtained as a result of these simplifications. Thus, two-dimensional and three-dimensional fluid flow analysis is conducted when more detailed and more precise information regarding fluid flow is needed. For example, pressure and velocity distribution at intake structures, identification of recirculation zones that occur due to trash-rack and screen installation, mixing at junctions, etc.

Numerical simulations are the basis for using artificial intelligence methods such as optimization algorithm or machine learning algorithms. Optimization methods can be used to enhance existing system design if system design variations are evaluated by numerical simulations of system behavior. Optimal sensor placements can be determined if synthetic measurements obtained from the simulations are compared for various sensor layouts. Fault events can be detected by optimizing simulation parameters of accident events by comparing simulation results to observed values. With the increasing amount of sensor data, machine learning algorithms, which are based on finding patterns and underlying correlations in the data, can be used to extract meaningful information from sensor measurements, which can then provide better insight into existing systems and enhance its monitoring. With increasing computational resources, the importance and applications of these methods are growing, especially machine learning methods. In this thesis, multiple novel areas of applications of proposed methods in water systems are presented, and further research possibilities are proposed.

1.3   Organization of the thesis

In the presented thesis, several applications of optimization and machine learning techniques are presented for known problems in water transportation systems, namely reduction of hydraulic losses, pollution localization, and leak localization. For each of these directions, limitations of conducted research are discussed and future work is proposed. The thesis is organized as follows.

In the second chapter, an overview of optimization methods used for water transportation systems analysis is presented. Numerical and optimization approaches are applied for the evaluation of novel trash-rack geometry designs. Previous research of improved designs was mostly relying on experimental testing, whereby the proposed artificial intelligence based methods provide an investigation of novel designs which were previously not considered in experimental testing. The future application of the proposed approach is discussed. Additionally, in the same chapter, the application of optimization techniques for the detection of the pollution source is presented. It is known that the pollution localization problem is very complex due to the categorical variable which represents the source node. Additionally, it is a multi-modal problem, since multiple equally good solutions can exist. Therefore, a novel approach that reduces the search space of the considered problem is presented, with the addition of the novel optimization approach. In this optimization approach, separate optimizations are conducted for each suspect source node that survived the search space reduction. This reduces the dimensionality and complexity of the considered problem, since it eliminates the categorical variable which is also the most problematic from the optimization standpoint. It also enables obtaining the best solution for each potential source node, i.e. deals with the multi-modality of the problem. Limitations of the proposed method are mentioned, and future work is proposed.

In the third chapter, an overview of machine learning methods used in water transportation systems is presented. A novel machine learning approach based on a Random Forest classifier employed for the prediction of a limited number of pollution sources in water distribution network in the case of accident event is presented. Since the efficiency of search space reduction methods and optimization methods depends on the number of pollution injection locations, it is greatly beneficial to have information about

the number of pollution sources. Assumptions used for the proposed method are explained, and further areas of investigation are explored. Machine learning application was also used for the identification of possible leak locations. Random Forest classifier is employed to detect source location based on large number of synthetic pressure data obtained from the leak simulations conducted for various node demands and leak sizes. As the further improvement of the proposed approach pipe segmentation is introduced. Since the leaks can occur anywhere in the pipe segment, the proposed approach simulates a more realistic case, where prediction model accuracy is estimated for leaks that can occur anywhere in pipe segment, not only in network nodes. Further development of the proposed idea is also briefly mentioned. In the fourth chapter, a conclusion is provided where a summary of main contributions is presented and a proposal of future research is given.

Chapter 2

Optimization methods in water transportation systems

Optimization methods are aimed at finding the best solution for a defined goal function, which can be formulated as a maximization or minimization problem. During the iterative process of the optimization method, values of input variables are varied, typically within a predefined range, until stopping criteria are satisfied. The most straightforward optimization problem in water transportation systems is finding optimal geometry design for defined criteria. For example, the design of trash-racks or screens within provided limits to assure needed blockage while producing minimial hydraulic losses. In this case, an optimization algorithm is used to iteratively provide values for chosen geometry parameters, e.g. bar width, bar spacing, and bar length, and numerical simulations are conducted to evaluate fitness function value, i.e. define hydraulic losses for the chosen design. An increase in the number of considered geometry parameters increases geometry flexibility and enables finding improved designs; however, a greater number of optimization parameters considerably widen the search space. Therefore, enhanced optimization methods need to be used to reduce the probability of obtaining only local optima and to increase convergence speed, since numerical simulations for evaluation of structure design usually require 2D or 3D numerical simulations which are often computationally quite expensive.

Another type of optimization problem in water transport systems are inverse problems. In case of an incident event, it is of main importance to find conditions that caused the incident. Optimization techniques and numerical simulations are jointly used where

incident event parameters used for simulation are changed by optimization algorithm, with results obtained by simulation being compared to the true sensor measurements. Parameters that provide the best agreement between computed "measurements" from the simulated scenario and real sensor measurements are the optimal solution of the optimization problem solved. This is considered as an inverse problem since based on the recorded output (sensor measurements) the algorithm tries to find optimal input (incident event parameters). Detection of pollution event parameters is considered as an inverse type of problem, where based on sensor measurements pollution source location and pollution parameters need to be obtained. The main problem is the node variable which is the categorical variable that makes considered optimization problem of the most complex type since the mixture of continuous and categorical values is present in the optimization problem. Additional problem is that in case of pollution event rapid reaction time is needed to minimize the harm for end-users, thus fast optimization methods are needed.

Additional optimization problems in water distribution networks include optimal sensor placement, calibration of numerical models based on in-field measurements, optimization of pipe diameters and lengths for new infrastructure, etc. However, these optimization problems will not be covered in this thesis.

2.1   Optimization of hydraulic element design

To determine the geometry of various infrastructure segments engineering practice and model testings are mostly used. For example protection racks or trash-racks can be chosen in accordance with known guidelines regarding hydraulic losses [36]. However, with growing ecological concern, it is observed that classic designs can cause injuries and increase mortality of fish species [5, 37]. Therefore, novel designs are being considered where turbulence zones are being induced, which fish naturally avoid. However, these novel designs, which usually consist of angled trash racks and angled bars, increase hydraulic losses, thus a compromise between engineering and ecological concern needs to be made. In recent years, model testing is being conducted to investigate different trash-rack and screen designs, bar spacing, length and inclination [95, 97, 68, 69, 4, 96]. However, cross-sections are usually kept rectangular or

with a streamlined shape. Only recently novel cross-section designs are being considered [8, 55]. The main limitation of model testing is that designs used in investigation need to be defined before testing, which makes it difficult to pinpoint the true optimal design. Additionally, in model testing intake design is defined by testing facility equipment, which is usually a straight channel and unique specifics of each intake cannot be considered. This is a considerable simplification, since it is reasonable to believe that each intake has a unique optimal design that ideally needs to be defined. Papers dealing with numerical simulations of intake structures are rather sparse [70, 59, 3, 40]. In all of these papers, a limited number of pre-defined designs was investigated, thus coupling numerical simulations with optimization techniques would enable exploration of these new, innovative cross-sections specifically designed for the considered intake.
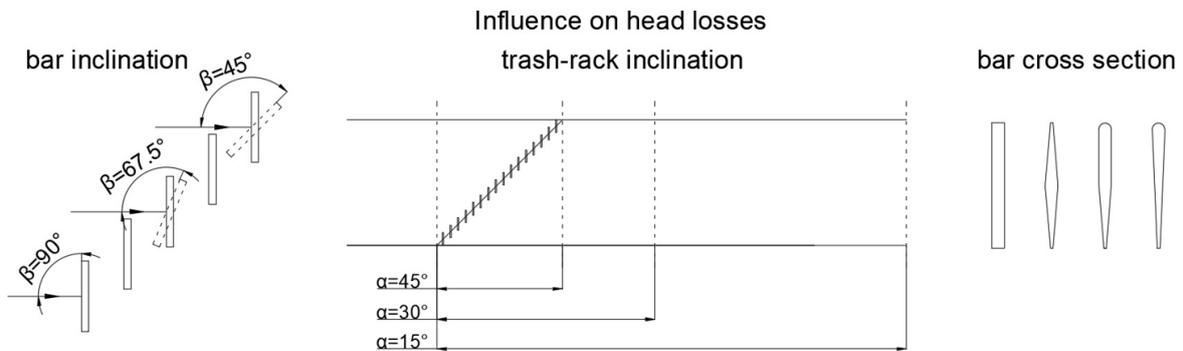
Figure 2.1: Numerically investigated bar inclinations, trash-rack inclinations, and bar cross sections [48].

In this thesis, numerical analysis was conducted on four different types of bar cross-sections: rhombus, rectangular, rounded front edge with inclined back in the lower half, and rounded front edge with inclination starting after rounded edge, for different bar and trash rack inclinations (Figure 2.1). Numerical analysis was conducted in the ANSYS Fluent using 2D $k - \epsilon$ turbulent flow model. The validation of numerical setup was made by comparing with experimental results obtained from [4] for different bar inclinations ($45°$, $67.5°$, and $90°$) and trash-rack inclinations ($15°$, $30°$, and $45°$). It was observed that depending on thrash rack and bar inclination angles different cross-sections produce an optimal solution, i.e. yield smallest head losses. This indicated that optimization of bar cross-section could be beneficial for specific intake geometry. Therefore, an optimization procedure using Particle Swarm Optimization (PSO) was

conducted for a real intake structure of Hydroelectric power plant Senj, Croatia, where three different bar cross-sections were considered: with all rounded edges, with all inclined edges, and design with rounded front edges and inclined back edges. In the considered case, fluid flow was adjacent to the bars, so the optimization for all considered cases converged in cross-section with the lowest cross-section area since it produced the smallest disturbance in fluid flow (Figure 2.2).
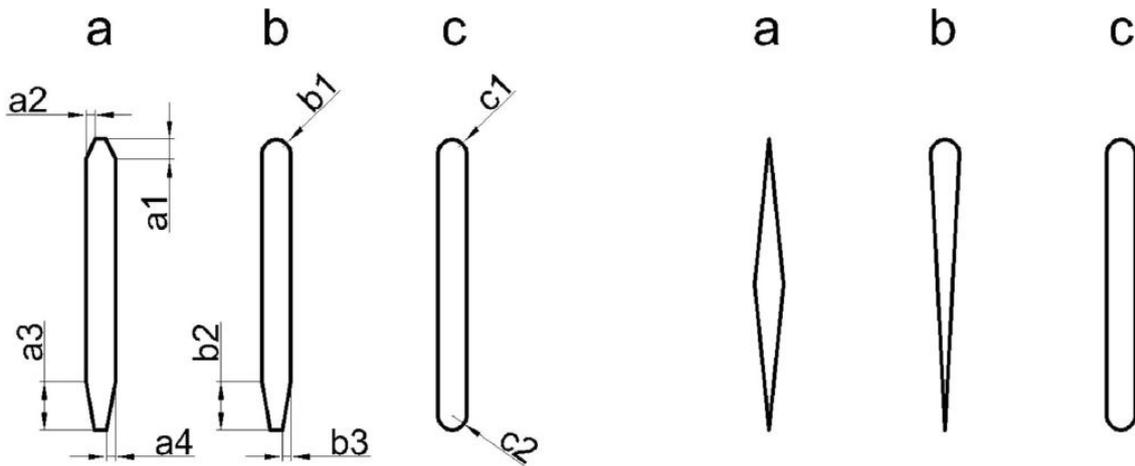
Figure 2.2: Investigated cross sections for optimization approach (left) with final optimal design (right) [48].

The obtained results give a strong indication that further investigation of numerical and optimization methods needs to be conducted. It must be noted that in the current optimization procedure parametric shape was considered which is often the case in optimization methods since it provides a smaller search space due to the smaller number of optimization variables. However, it limits the optimization possibilities since innovative designs, such as curved bars [8] cannot be obtained. However, with growing computational power, optimization procedures in which cross-sections are described with the set of points where coordinate values are optimization variables could enable obtaining more innovative designs such as curved bars with irregular cross-sections. Although such shapes could be hydraulically more efficient, construction constraints must be also taken into consideration such as the strength of the material, manufacturing demands, method of cleaning of the trash-rack structures, etc. These types of problems need to satisfy multiple constraints and have multiple goals which are often opposed, therefore coupling of computational fluid dynamics (CFD) and structural

(FEM) numerical analysis with optimization methods can provide improved solutions while considering all present limitations.

2.2 Pollution localization with optimization - simulation approach

Monitoring of complex systems such as water transportation systems can be a challenging task due to various accident events that can occur. Accidental or intentional pollution intrusions need a rapid reaction, where identification of pollution source, contamination concentration, starting time, and duration need to be identified as soon as possible. Based on this information, simulation of pollution spreading can be conducted so polluted water distribution network areas can be identified, warnings can be given to the affected users, and the source of pollution can be eliminated. Since the reaction time is of main importance, different methods and techniques are being used to simplify the considered problem and narrow down the search space to provide faster response time [20, 43, 64]. When using stochastic optimization methods, such as PSO, multiple optimization runs are needed. Keeping in mind the necessity of rapid intervention, multiple runs of optimization cycles can be time extensive and still do not assure obtaining the optimal solution. This is especially important when discrepancy in sensor measurements [63] and water demand uncertainties [81, 90] are included. Additionally, it is known that the considered problem is a multimodal, i.e. multiple solutions exist. As a solution to this, niching algorithm has been proposed in work by [34, 91] where during the optimization run the best solution for each network node is stored in its niche. This approach produces multiple solutions from a single optimization run.

In this thesis, a search space reduction technique has been proposed in which, prior to conducting the optimization process, preliminary evaluation of possible source nodes is conducted. For each network node, an unrealistically high contamination value was injected during the entire simulation time. Simulations were conducted using EPANET2 software. If sensors did not register contamination for this extreme case it is concluded that they would not be able to detect contamination for any other, less severe, contamination scenario parameters (e.g. smaller pollution concentration value, shorter injection time, etc.), thus these nodes are excluded from the search space. In this way,

a considerable percentage of network nodes can be eliminated before commencing the optimization procedure. The main benefit of this forward approach is that number of needed simulations is equal to the number of network nodes and the execution of these simulations can be run in parallel. The proposed approach is not computationally demanding but can provide considerable search space reduction. An extensive investigation of the proposed method was conducted for 5 different sized benchmark water distribution networks which are shown in Figure 2.3. Details of considered networks and sensor layouts can be found in Table 2.1.

Table 2.1: Overview of investigated networks and sensor layouts in [52].

Network	No. of network nodes	Simulation time	Sensor placement
Anytown	19	24 h	70, 160
			90, 110, 140
Net3	92	24 h	117, 143, 181, 213
			115, 119, 187, 209
			113, 120, 147, 211
			117, 149, 167, 213, 253
			117, 173
Richmond	865	72 h	123, 219, 305, 393, 589
			93, 352, 428, 600, 672
BWSN Network 1	126	96 h	10, 31, 45, 83, 118
			10, 83
BWSN Network 2	12523	48 h	871, 1334, ..., 11519

Anytown network

Net3 network

Richmond network

BWSN Network 1

BWSN Network 2

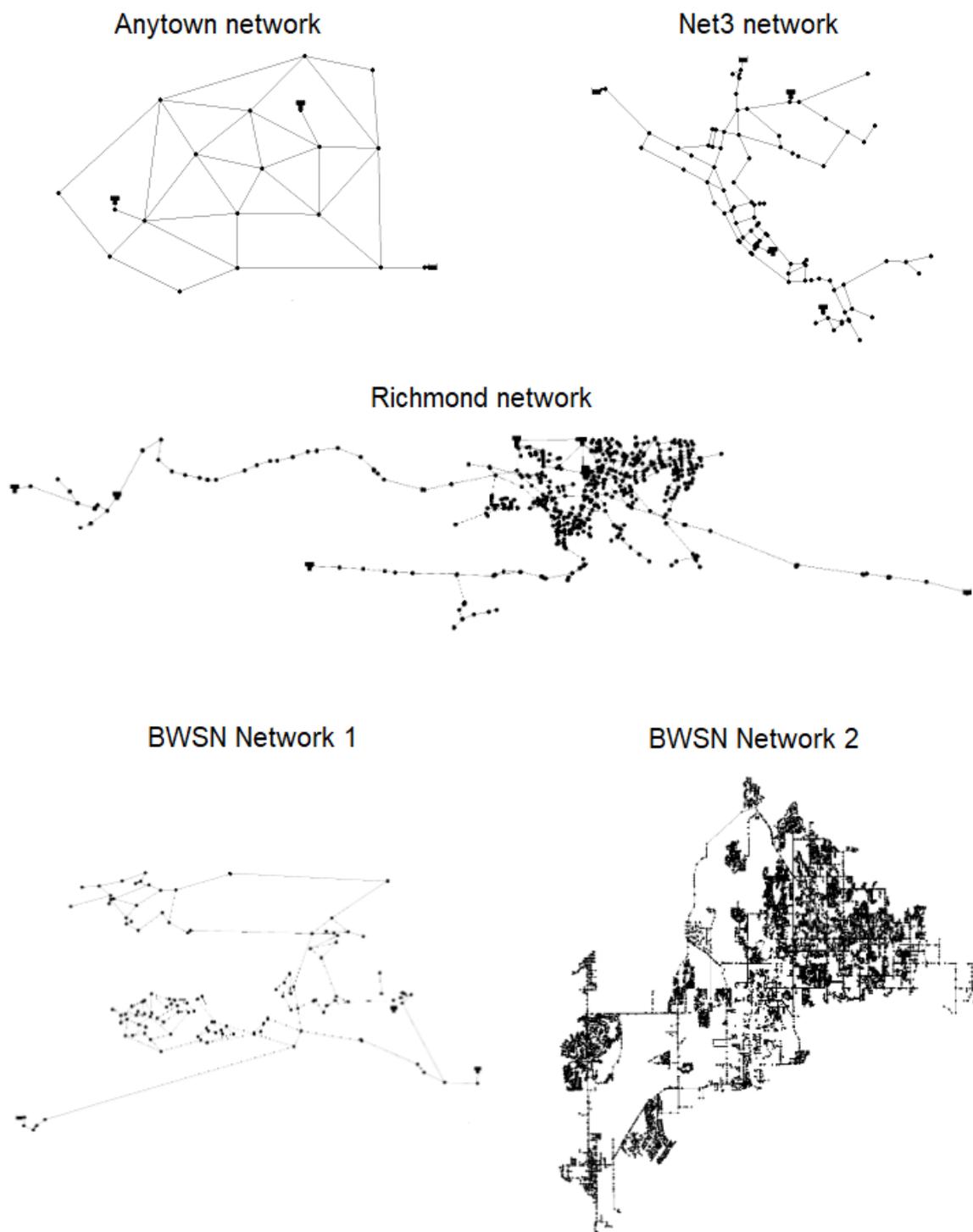Figure 2.3: Networks investigated in [52] for search space reduction method.

Additionally, two different approaches were considered, one in which multiple injection locations are possible and one in which only a single pollution source is considered. If multiple sources of pollution are considered, network nodes were eliminated only if all sensors did not record pollution in the extreme scenario. If only a single source of

pollution is considered, the condition was that exactly all sensors that detected pollution in the real pollution event must detect pollution in the case of an extreme scenario. Greater reduction of network nodes was obtained for the approach for a single source of pollution, however, the assumption of only one pollution source can cause wrong results if multiple sources are present.

Based on these findings, it was observed that if a reasonable number of suspect nodes remained, independent optimization procedures for each suspect node can be conducted in reasonable time. In this way, for each optimization the remaining optimization variables are pollution injection starting time, injection duration, and concentration value. This considerably reduces optimization complexity since the most problematic categorical variable (pollution source location) is eliminated. Additionally, the proposed approach eliminates the problem of obtaining only local optima. The niching algorithm [34, 91] also provides multiple solutions, but the advantage of the approach proposed in this thesis is that for each injection the optimum fitness is obtained, since independent optimization runs performed for each injection node. In the case of the niching algorithm, it is expected that the optimization algorithm investigates the most in the vicinity of the optimal solution, where it is not given that obtained solutions from other niches are the best solutions that can be obtained for these source nodes.

The performance of the proposed search space reduction method should be investigated under demand uncertainties and sensor measurement imperfection, which is a more realistic case. The research conducted in this thesis, and many other pollution detection techniques and methods found in literature, are tested with the measurements from an extended period. However, in a real case scenario, methods for pollution detection will be utilized from the first positive sensor detection. Therefore, the proposed approach should be evaluated for more realistic conditions. The independent optimization approach was further explored in work by [30] where based on machine learning prediction several most suspect nodes are detected and are then used in independent optimization procedures. However, these independent optimizations assume a single injection location, where multiple injection locations should be also explored.

Chapter 3

Machine learning in water transportation systems

Machine learning algorithms are designed to find underlying patterns and correlations in the data. Their best performance is achieved when a big amount of data is available but at the expense of increased model complexity. With recent advances in technology, the competence of machine learning models is increasing for various problems which were previously not considered for machine learning application. Machine learning models can be used as a substitute for extensive experimental or numerical testing when a considerable number of parameters is investigated. For example, when a large number of parameters need to be considered, such as pipe diameters, pipe lengths, temperature, water quality parameters, etc., a substantial number of experiments or simulations would be required to obtain valid conclusions about the investigated phenomena. However, if a limited but significant number of experiments or simulations are conducted, a machine learning model can be used to predict the desired output variable for various input parameter combinations which were not evaluated through the experiment. Additionally, machine learning methods are also often used for finding anomalies in sensor readings which can indicate accident events such as contamination occurrence or pipe bursts. Machine learning was previously used for evaluation of mixing in double pipe junctions [27], water quality monitoring [56, 26], prediction of possible sources of pollution intrusion [83, 29, 28, 30], anomaly detection [82], prediction of pipe failures [74, 25], leakage localization [47, 13], detection of cyber-attacks [58, 1] etc.

In case of accident events in which reaction time is the most important, the main advantage of machine learning methods over optimization methods is considerable reduction in needed computational time. In the case of machine learning applications, the majority of computational time is used for data preparation and model training. This can be conducted prior to the accident event, so later the prepared prediction model can be used with only a small computational effort needed. However, the main disadvantage of the machine learning approach is that real conditions of water distribution networks in the case of the accident event cannot be known, and can be considerably different from those used for the construction of the machine learning model. This is especially important for water network demands which can vary considerably on a daily or hourly basis, where for the machine learning approach an estimate of system behavior is used. Therefore, the advantage of the optimization approach is that it is prepared after the accident event is observed, thus calibrated water distribution network model, based on observed sensor measurements during the accident event, can be used.

It should be noted that the recorded accident events in water distribution networks are rather rare, therefore the data for prediction model training is limited. However, simulations of the wide range of different conditions can be utilized to consider various uncertainties that can occur to gather a considerable amount of synthetic data. Ultimately, multiple prediction models can be prepared for utilization in the case of an accident event. In this thesis, machine learning algorithms have been used to predict the number of pollution locations and to determine possible leak locations based on sensor measurements.

3.1 Machine learning application for determination of number of pollution sources

Based on a larger number of pollution simulations with various pollution parameters, machine learning models can be used to provide the most probable pollution source with the prediction of injection time, injection duration and contamination concentration [28, 29]. After a machine learning based source localization, further finer determination of pollution parameters can be conducted with optimization techniques [30]. The main problem is that the majority of both optimization and machine learning techniques

in the literature only consider a single pollution location. In [83] Bayesian approach coupled with Support Vector Regression was used for the probability distribution of possible contaminant sources with the assumption of a single injection location. In [75] the efficiencies of the Bayesian probability-based method, backtracking method, and optimization-based method were evaluated, where it was noted that the Bayesian method was designed only for a single contamination location. Machine learning predictions of pollution scenario parameters based on Random Forest algorithm [28, 29, 30] all assume a single pollution location. As mentioned in the previous chapter, the pollution localization technique proposed in [52] showed better pollution localization in the case of a single pollution source, although, it can be extended to multiple injection sources. If several pollution locations are assumed, optimization variables need to be assured for additional source locations which considerably increases the search space. If these variables are ultimately not needed, the reaction time is prolonged due to an unnecessary increase in problem complexity.

In this thesis, the machine learning approach is presented where the number of contamination sources is predicted based on a large number of simulations for various pollution scenario parameters. A considerable number of pollution scenarios were generated with a randomly chosen number of pollution sources, concentration value, injection time, and duration time. The number of injection locations varied from 1 to 4, and investigated networks were Net3 and Richmond network. When multiple injection locations are chosen, simplification was made and the same pollution parameters (injection time, duration, and concentration value) were used for all locations. Similarly as in [52], it was observed that random pollution parameters can cause pollution scenario which is undetected by all sensors in the water distribution network, thus the prior check of prepared data is conducted. Network nodes for which pollution is not detected are eliminated, and simulation with multiple locations was conducted only for source nodes that contribute to pollution readings. The example of conducted data preparation can be observed in Figure 3.1 where 3 different source nodes were randomly chosen; however, ultimately only 2 network nodes were considered since for the one source node (node 151) sensors did not detect contamination.
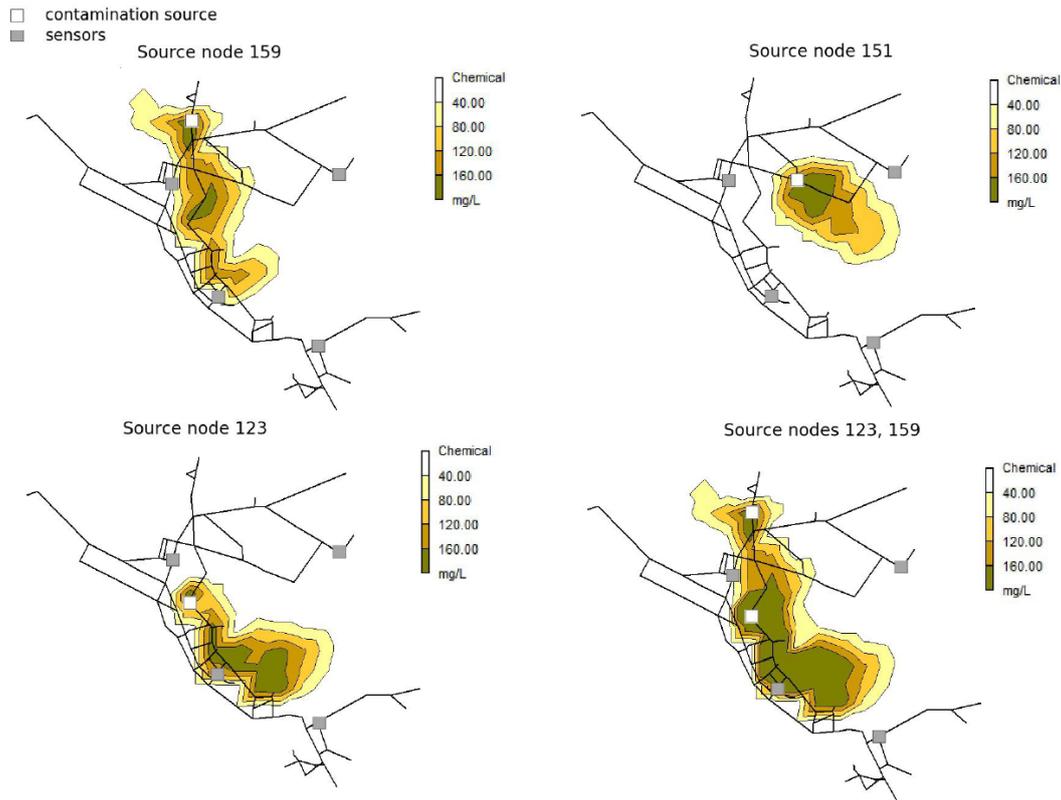
Figure 3.1: Example of single and multiple sources pollution scenarios of Net3 considered in [51].

Random Forest classifier implementation in the Python library Scikit-learn [60] is used to determine the number of pollution sources. This is important since the number of optimization variables needs to be defined before the optimization process and some space reduction techniques have better performance in the case of the single pollution source. Based on the right prediction of the number of contamination sources only needed optimization variables can be used for the optimization problem. It was observed that good accuracy can be obtained when an exact number of sources is predicted, however considerable improvement is obtained when it is predicted if single or multiple numbers of sources are present. This information is also important since some pollution localization techniques can be used only in the case of a single source or are more efficient when that is assumed, as was shown in [52].

In the case of the proposed machine learning approach, a considerable assumption was made where all sources of pollution had the same parameters. This could be realistic in the case of simultaneous intentional intrusions; however, even then exactly equal parameters are hard to expect. Therefore, further investigation should be

20

conducted for the evaluation of the number of pollution sources but with various pollution scenario parameters. Additionally, due to a large amount of considered data, especially when uncertainties are incorporated, data reduction techniques should be explored and other machine learning algorithms which could provide better model accuracy. It must be noted that the reaction time is most important in this case, thus all possible improvements that could reduce the computational time should be explored.

3.2 Machine learning application for leak localization

The presence of leak locations in water distribution networks can be a considerable problem due to substantial water losses. The main problems are small leaks that can be often hard to detect, and over time can cause considerable cumulative water losses. Additionally, for some types of terrain, water is absorbed in the soil, thus leak presence is not evident on the surface and even greater leaks can remain undetected for longer periods. Leak locations can also be hazardous due to possible contamination intrusion from surrounding contaminated soil. Therefore, the detection, localization, and repair of even small leaks is important. Usual methods for leak localization consider hardware-based methods and software-based methods. Hardware-based methods use in-situ measurements, for example, infrared thermography, acoustic methods, or ground-penetrating radar. The main problem is that these methods require an experienced operator and are time and money-consuming. Software-based methods use various software for simulation of water systems or analysis of measured data from sensor measurements. The main problem is that many software based methods rely on residual-based analysis where pressure sensor measurements are compared to expected (predicted) values, which can considerably differ from true pressure measurements.Thus, these methods can have a considerable number of false positive predictions since an unexpected surge in water demand can be interpreted as leak occurrence. If numerical simulations are being conducted for the evaluation of expected system behavior, the numerical model needs to be a good representative of the real network, which is often the problem, due to various uncertainties. For example, the accumulation of corrosion byproducts and suspended particles with time can cause a reduction in pipe diameter and can change pipe roughness. These values can be

calibrated with in-field measurements; however, considerable estimations still remain present such as unknown valve opening status. An overview of some of leak detection and localization methods, with their advantages and limitations can be found in [18, 31, 6, 94, 38, 88, 12, 94].

Recently different machine learning algorithms have been used for leak detection and localization such as principal component analysis (PCA) [66], convolutional neural network (CNN) [99], artificial neural network (ANN) [61, 57], k-nearest neighbours [76], Bayesian classifier [77], deep learning [99], linear discriminant analysis (LDA) and neural network classifiers [79]. However, the main problem is a sparse number of data for actual leak and burst events, which is the requirement for high prediction model accuracy. However, a considerable amount of data can be obtained if simulations are conducted with the variation of leak location, leak size, and node demands. The idea is similar as in [29] and [51] where a large number of simulations for pollution scenarios were conducted. Additionally, uncertainties can also be addressed by simulating leak scenarios under various conditions, such as node demands, pipe diameters, etc.

For the leak localization using big data Random Forest algorithm implementation in the Python library Scikit-learn [60] was employed. Two different-sized networks were considered, Hanoi (Figure 3.2) and Net3 network. An overview of considered networks, sensor placements, and simulation parameters is given in Table 3.1. The prediction model accuracy was assessed for different base node variations, leak sizes, and sensor layouts. It was observed that better prediction accuracy can be obtained for greater leaks, which was expected since they cause greater disturbance in the pressure measurements. Additionally, with the increase of network complexity and a wider range of demand uncertainties, a possible number of leak scenarios rapidly increases, thus prediction model accuracy for the same number of training data is reduced. Although for these cases true leak node is often not detected, if several top nodes with the greatest prediction model certainty are considered, a significant leak localization can be obtained. Additionally, the prediction model can be prepared for a specific range, such as nighttime when there are smaller water demands and with that smaller demand variations. It must be noted that considerable simplification is made, since the proposed classification method is trained, and thus can only predict leak locations that are network nodes. In reality, this is not the case since the leak can occur anywhere in the

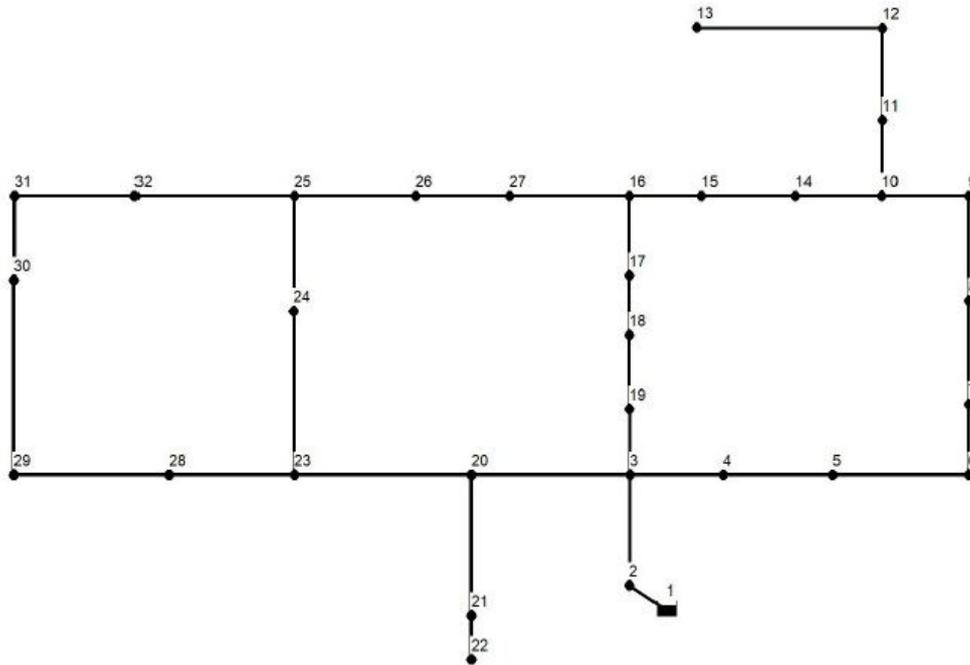pipe segment. These issues are further investigated.

Figure 3.2: Hanoi network.

Table 3.1: Overview of investigated networks and sensor layouts in [49].

Network	No. of network nodes	Sensor placement	Simulation time
Hanoi	31	14, 30	24 h
Net3	92	117, 143, 181, 213	24 h
		115, 119, 187, 209	
		117,181	
		119,209	

Further investigation was conducted for the pipe segmentation approach. Each pipe is divided into additional segments, where three different machine learning frameworks are considered and compared. The first model was trained and tested on leak scenarios with leak locations in original network nodes, the same approach as in [49]. In the second approach, the model was trained on all original and segmentation nodes and the prediction of leak node is original or segmentation node. The third model was trained on original network nodes, and prediction is made for scenarios with leak locations in both original and segmentation nodes. Since the number of classes, in this case, is the

23

same as the number of network nodes, segmentation nodes are associated with their nearest original network nodes. If a prediction of that nearest original network node is made, it is considered as the correct prediction for the segmentation node. It was observed that the second approach considerably increases computational demands and as such is currently not a feasible approach for larger networks. However, the last approach simulates the most realistic case and as such can successfully localize the area of leak location if several top nodes with the greatest prediction model certainty are considered.

To further localize the leak location, sequential prediction models are used. After the initial leak localization is made, pipe segmentation around most suspect nodes is conducted (Figure 3.3). Sequential prediction models were trained with simulations conducted with possible leak locations in the original most suspect network nodes and segmentation nodes. However, it was observed that machine learning models have a problem with detecting fine differences in pressure sensor measurements for different leak scenarios, and although true leak location is always in several top nodes, it is not always the prediction with the greatest certainty. Optimization methods could provide finer parameter tuning, therefore, coupling of machine learning approach for general localization with optimization approach for finer detection of leak location should be explored.
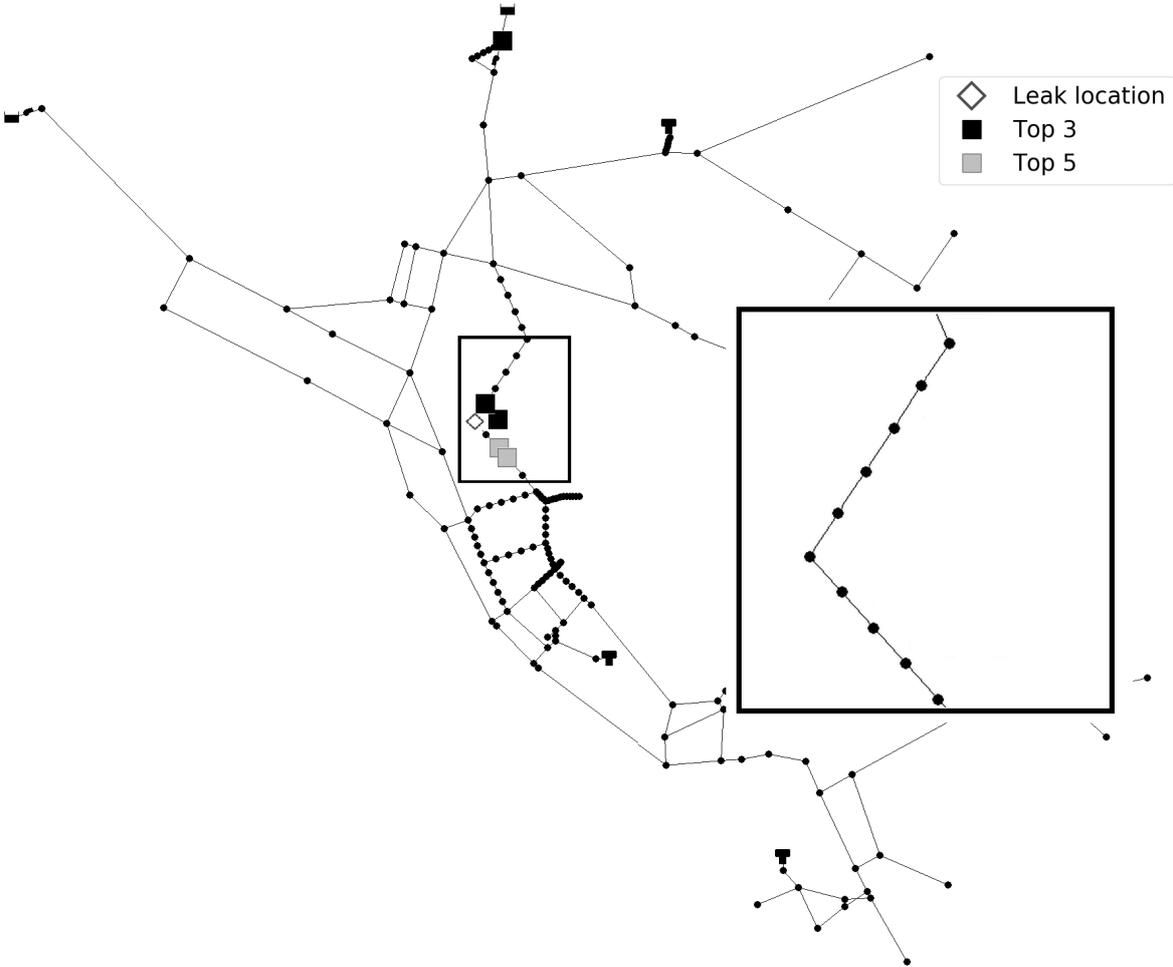
Figure 3.3: Pipe segmentation after initial machine learning localization as conducted in [50].

Chapter 4

Conclusion

4.1   Main contributions

In the presented thesis, multiple novel applications of artificial intelligence in water transportation systems are presented and discussed. The enclosed papers show the in-depth methodology of proposed applications, where multiple directions of possible further utilization are discussed in the present thesis. It was shown that with growing computational resources utilization of novel approaches is greatly beneficial, where previously used methods were less efficient.

The main contributions from the presented research are as follows:

- Numerical analysis and optimization methods have been used to determine optimal cross-section shape for the defined bar and trash-rack inclinations of the water intake structure. Conducted research showed that optimal cross-section shape varied for different trash-rack configurations, which showed the importance of optimization methods that can freely adjust geometry-shape for specific intake structures.

- A novel search space reduction method in pollution detection was presented, which considerably reduced the number of potential pollution sources. Based on the reduced number of solutions, a novel pollution localization technique was presented, in which for each remaining suspect node an independent optimization procedure was conducted to obtain pollution starting time, duration of injection, and concentration value. This approach significantly reduced the complexity of

the optimization problem since the categorical variable was removed. Additionally, since the considered problem is multi-modal, multiple solutions are simultaneously obtained which is overall computationaly more efficient.

• A novel machine learning approach that identifies the number of pollution sources in the water distribution network is proposed. The presented method show good accuracy when the exact number of pollution locations are predicted, which is important information for reducing the number of unnecessary optimization variables. Additionally, when it is predicted only whether single or multiple pollution sources are present, the accuracy of the technique increases. This is important since some space reduction methods are more efficient and others are specialized for a single pollution location.

• Random Forest prediction model trained on synthetic pressure data obtained by simulated leak scenarios was utilized for the leak localization problem. It was shown that the proposed approach can incorporate various uncertainties regarding water distribution network behavior and provide considerable leak localization. That represents a strong benefit since software-based methods that use residuals as the criteria for anomalies detection use an estimated water distribution network model and thus can cause wrong results.

• Further investigation of machine learning application in leak localization was conducted for the pipe segmentation approach. Since the usage of 1D numerical simulations enables detecting only network nodes as leak locations, and leaks can occur anywhere in pipe segment, additional network nodes were created to further localize leak location. Sequential machine learning models were used, where the first Random Forest model was used to identify leak area for pipe segmentation, and the second model was used to try to identify exact leak location. It was observed that the proposed approach narrows down the leak area for leaks occurring both in network nodes and in pipe segments with great accuracy; however, the exact location cannot be determined since multiple leak locations with various leak sizes can produce similar pressure sensor measurements.

4.2 Future work

As a continuation of conducted research possible future research areas are:

- The design optimization of the trash-rack cross-section should be conducted with more parameters where profiles would be described with a large number of points that would be able to converge into any shape, such as curved cross-sections. Additionally, the expansion of fitness function for design optimization with ecological goals should be explored, such as including fish avoidance ability considering the proposed design.

- Search space reduction technique coupled with independent optimization methods proposed for the detection of pollution scenario parameters should be explored in the context of rapid reaction time, where the search for the pollution parameters will start immediately after pollution detection, not after a longer measurement time which was the case in this study. Additionally, time-varying pollution injection should be considered since the assumption of constant injection concentration is assumed.

- Other machine learning algorithms should be investigated to achieve improved accuracy for the identification of a number of pollution sources and leak location. Additionally, dimensionality reduction methods should be also explored to reduce prediction model complexity.

- The machine learning application that identifies the number of pollution sources should be further explored with various pollution scenario parameters for each injection node.

- Coupling of machine learning algorithm that would identify the leak area and optimization methods which would find exact leak location should be explored.

- The segmentation approach should be applied to pollution location detection since leak locations are also possible pollution intrusions locations. Thus, the current assumption of pollution locations only in network nodes is also a considerable simplification of the problem, which should be avoided if possible.

Chapter 5

Summary of papers

A   Assessment of Head Loss Coefficients for Water Turbine Intake Trash-Racks by Numerical Modeling

In this work, numerical simulations of fluid flow around trash-rack for different bar cross sections are conducted to investigate cross section influence on head losses. Comparison with experimental data is conducted to validate the usage of numerical simulations which enable investigation of great number of trash-rack configurations. In previous experimental studies researchers mostly focused on trash-rack parameters (bar spacing, bar length, inclinations etc.) where bar cross section was mainly rectangular or streamlined shape. Therefore, 2D simulations for different cross sections are carried out for a range of trash-rack configurations in order to provide better insight how it affects energy losses. It is shown that head loss reduction due to change in cross section is greatly dependent on trash-rack configuration, therefore optimization of simplified real water turbine trash-rack is also conducted to produce the cross section that generates smallest head losses for given configuration.

B   Source Contamination Detection Using Novel Search Space Reduction Coupled with Optimization Technique

Contaminant intrusion in a water distribution network is an important concern because it can have hazardous consequences for the population. Reacting immediately is crucial to prevent or reduce the further propagation of contamination. In terms of contamination scenario characteristics, optimization is researched extensively as a valuable methodology to provide information. This work presented a procedure preceding the optimization which considerably reduces the search space for a potential contaminant source location. For each suspect node, a simulation is conducted with unrealistically high contaminant concentration injected throughout the whole simulation. If the sensors do not register contamination in a subsequent scenario, then that node can be eliminated as a possible contaminant source. The methodology is applicable for both single and multiple contaminant injection nodes. This approach was investigated in multiple benchmark networks and for different sensor placements in the literature. By coupling the proposed search space reduction method with an optimization approach, a novel efficient methodology for contamination source detection was presented.

C   Machine-Learning Classification of a Number of Contaminant Sources in an Urban Water Network

In the case of a contamination event in water distribution networks, several studies have considered different methods to determine contamination scenario information. It would be greatly beneficial to know the exact number of contaminant injection locations since some methods can only be applied in the case of a single injection location and others have greater efficiency. In this work, the Neural Network and Random Forest classifying algorithms are used to predict the number of contaminant injection locations. The prediction model is trained with data obtained from simulated contamination event scenarios with random injection starting time, duration, concentration value, and the number of injection locations which varies from 1 to 4. Classification is made to determine if single or multiple injection locations occurred, and to predict the exact number of injection locations. Data was obtained for two different benchmark networks, medium-sized network Net3 and large-sized Richmond network. Additionally, an investigation of sensor layouts, demand uncertainty, and fuzzy sensors on model accuracy is conducted. The proposed approach shows excellent accuracy in predicting if single or multiple contaminant injections in a water supply network occurred and good accuracy for the exact number of injection locations.

D Data-Driven Leak Localization in Urban Water Distribution Networks Using Big Data for Random Forest Classifier

In the present paper, a Random Forest classifier is used to detect leak locations on two different sized water distribution networks with sparse sensor placement. A great number of leak scenarios were simulated with Monte Carlo determined leak parameters (leak location and emitter coefficient). In order to account for demand variations that occur on a daily basis and to obtain a larger dataset, scenarios were simulated with random base demand increments or reductions for each network node. Classifier accuracy was assessed for different sensor layouts and numbers of sensors. Multiple prediction models were constructed for differently sized leakage and demand range variations in order to investigate model accuracy under various conditions. Results indicate that the prediction model provides the greatest accuracy for the largest leaks, with the smallest variation in base demand (62% accuracy for greater- and 82% for smaller-sized networks, for the largest considered leak size and a base demand variation of $\pm 2.5\%$). However, even for small leaks and the greatest base demand variations, the prediction model provided considerable accuracy, especially when localizing the sources of leaks when the true leak node and neighbor nodes were considered (for a smaller-sized network and a base demand of variation $\pm 20\%$ the model accuracy increased from 44% to 89% when top five nodes with greatest probability were considered, and for a greater-sized network with a base demand variation of $\pm 10\%$ the accuracy increased from 36% to 77%).

E   Detailed Leak Localization in Water Distribution Networks Using Random Forest Classifier and Pipe Segmentation

In this paper, a Random Forest classifier was used to predict leak locations for two differently sized water distribution networks based on pressure sensor measurements. The prediction model is trained on simulated leak scenarios with randomly chosen parameters - leak location, leak size, and base node demand uncertainty. Leak localization methods found in literature that rely on numerical simulations can only predict network nodes as leak nodes; however, since a leak can occur at any point along a pipe segment, additional spatial discretization of suspect pipe is proposed in this paper. It was observed that pipe segmentation of the whole network is a non-feasible approach since it rapidly increases the number of potential leak locations, consequently increasing the complexity of the prediction model. Therefore, a novel approach is proposed, in which a prediction model is trained on scenarios with leaks occurring in original network nodes only, but with its accuracy assessed against pressure sensor measurements from scenarios in which leaks occur in points between network nodes. It was observed that this approach can successfully narrow down the suspect leak area and, followed by additional segmentation of that network area and subsequent prediction, a precise leak localization can be achieved. The proposed approach enables incorporation of various uncertainties by simulating leak scenarios under different conditions. Investigation of leak size uncertainty and base demand variation showed that several different scenarios can produce similar sensor measurements which makes it difficult to unambiguously determine leak location using the prediction model. Therefore, future approaches of coupling prediction modeling with optimization methods are proposed.

Bibliography

[1] A. A. Abokifa et al. Real-time identification of cyber-physical attacks on water distribution systems via machine learning–based anomaly detection techniques. *Journal of Water Resources Planning and Management*, 145(1):04018089, 2019.

[2] O. S. Adedoja et al. Towards development of an optimization model to identify contamination source in a water distribution network. *Water*, 10(5):579, 2018.

[3] H. O. Åkerstedt et al. Numerical investigation of turbulent flow through rectangular and biconvex shaped trash racks. *Engineering*, 9(05):412, 2017.

[4] I. Albayrak et al. An experimental investigation on louvres and angled bar racks. *Journal of Hydraulic Research*, 56(1):59–75, 2018.

[5] S. V. Amaral et al. Survival of fish passing downstream at a small hydropower facility. *Marine and Freshwater Research*, 69(12):1870–1881, 2018.

[6] U. Baroudi et al. Pipeline leak detection systems and data fusion: A survey. *IEEE Access*, 7:97426–97439, 2019.

[7] N. A. Barton et al. Improving pipe failure predictions: Factors affecting pipe failure in drinking water networks. *Water Research*, 164:114926, 2019.

[8] C. Beck et al. Improved hydraulic performance of fish guidance structures with innovative bar design. In *12th International symposium on ecohydraulics (ISE 2018)*, 2018.

[9] M.-C. Besner et al. Assessing the public health risk of microbial intrusion events in distribution systems: conceptual model, available data, and challenges. *Water Research*, 45(3):961–979, 2011.

[10] C. Binnie et al. *Basic water treatment*. Royal society of chemistry, 2002.

[11] M. G. Carleton and J. S. Nielsen. A study of trash and trash interception devices. *Water Science and Technology*, 22(10-11):287–290, 1990.

[12] T. K. Chan et al. Review of current technologies and proposed intelligent methodologies for water distributed network leakage detection. *IEEE Access*, 6:78846–78867, 2018.

[13] J. Chen et al. An iterative method for leakage zone identification in water distribution networks based on machine learning. *Structural Health Monitoring*, 20(4):1938–1956, 2021.

[14] S. P. Clark et al. Experimental study of energy loss through submerged trashracks. *Journal of Hydraulic Research*, 48(1):113–118, 2010.

[15] I. S. Cole and D. Marney. The science of pipe corrosion: A review of the literature on the corrosion of ferrous metals in soils. *Corrosion Science*, 56:5–16, 2012.

[16] A. F. Colombo et al. A selective literature review of transient-based leak detection methods. *Journal of Hydro-environment Research*, 2(4):212–227, 2009.

[17] S. H. Crandall et al. Destructive vibration of trashracks due to fluid-structure interaction. *Journal of Engineering for Industry*, 97(4):1359–1365, 1975.

[18] S. Datta and S. Sarkar. A review on different pipeline fault detection methods. *Journal of Loss Prevention in the Process Industries*, 41:97–106, 2016.

[19] T. Dawood et al. Water pipe failure prediction and risk models: state-of-the-art review. *Canadian Journal of Civil Engineering*, 47(10):1117–1127, 2020.

[20] A. E. De Sanctis et al. Real-time identification of possible contamination sources using network backtracking methods. *Journal of Water Resources Planning and Management*, 136(4):444–453, 2010.

[21] E. Egusquiza et al. Failures due to ingested bodies in hydraulic turbines. *Engineering Failure Analysis*, 18(1):464–473, 2011.

[22] R. Farmani et al. Pipe failure prediction in water distribution systems considering static and dynamic factors. *Procedia Engineering*, 186:117–126, 2017.

[23] S. D. Faust and O. M. Aly. *Chemistry of water treatment*. CRC press, 2018.

[24] J. Gibson et al. Predicting health risks from intrusion into drinking water pipes over time. *Journal of Water Resources Planning and Management*, 145(3):04019001, 2019.

[25] M. M. Giraldo-González and J. P. Rodríguez. Comparison of statistical and machine learning models for pipe failure modeling in water distribution networks. *Water*, 12(4):1153, 2020.

[26] V. Gomez-Alvarez and R. P. Revetta. Monitoring of nitrification in chloraminated drinking water distribution systems with microbiome bioindicators using supervised machine learning. *Frontiers in Microbiology*, 11:2254, 2020.

[27] L. Grbčić et al. Efficient double-tee junction mixing assessment by machine learning. *Water*, 12(1):238, 2020.

[28] L. Grbčić et al. A machine learning-based algorithm for water network contamination source localization. *Sensors*, 20(9):2613, 2020.

[29] L. Grbčić et al. Water supply network pollution source identification by random forest algorithm. *Journal of Hydroinformatics*, 22(6):1521–1535, 2020.

[30] L. Grbčić et al. Machine learning and simulation-optimization coupling for water distribution network contamination source detection. *Sensors*, 21(4):1157, 2021.

[31] A. Gupta and K. D. Kulat. A selective literature review on leak management techniques for water distribution system. *Water Resources Management*, 32(10):3247–3269, 2018.

[32] J. Hall et al. On–line water quality parameters as indicators of distribution system contamination. *Journal-American Water Works Association*, 99(1):66–77, 2007.

[33] D. Honingh et al. Urban river water level increase through plastic waste accumulation at a rack structure. *Frontiers in Earth Science*, 8:28, 2020.

[34] C. Hu et al. A mapreduce based parallel niche genetic algorithm for contaminant source identification in water distribution network. *Ad Hoc Networks*, 35:116–126, 2015.

[35] Z. Hu et al. A comprehensive review of acoustic based leak localization method in pressurized pipelines. *Mechanical Systems and Signal Processing*, 161:107994, 2021.

[36] I. E. Idelchik. Handbook of hydraulic resistance. *Washington, DC, Hemisphere Publishing Corp.*, 1986.

[37] M. L. Inglis et al. Testing the effectiveness of fish screens for hydropower intakes. Report Project SC120079. https://www.gov.uk/government/publications/testing-the-effectiveness-of-fish-screens-for-hydropower-intakes, 2016. Accessed: 16-July-2019.

[38] M. I. M. Ismail et al. A review of vibration detection methods using accelerometer sensors for water pipeline leakage. *IEEE Access*, 7:51965–51981, 2019.

[39] M. R. Karim et al. Potential for pathogen intrusion during pressure transients. *Journal-American Water Works Association*, 95(5):134–146, 2003.

[40] L. A. Khan et al. Computational fluid dynamics modeling of turbine intake hydraulics at a hydropower plant. *Journal of Hydraulic Research*, 42(1):61–69, 2004.

[41] G. J. Kirmeyer and K. Martel. *Pathogen intrusion into the distribution system*. American Water Works Association, 2001.

[42] Y. Kleiner and B. Rajani. Comprehensive review of structural deterioration of water mains: statistical models. *Urban Water*, 3(3):131–150, 2001.

[43] K. A. Klise et al. Two-tiered sensor placement for large water distribution network models. *Journal of Infrastructure Systems*, 19(4):465–473, 2013.

[44] C. S. W. Lavooij and A. S. Tusseling. Fluid-structure interaction in liquid-filled piping systems. *Journal of Fluids and Structures*, 5(5):573–595, 1991.

[45] R. Li et al. A review of methods for burst/leakage detection and location in water distribution systems. *Water Science and Technology: Water Supply*, 15(3):429–441, 2015.

[46] M. Lin et al. Wireless sensor network: Water distribution monitoring system. In *2008 IEEE radio and wireless symposium*, pages 775–778. IEEE, 2008.

[47] Y. Liu et al. Water pipeline leakage detection based on machine learning and wireless sensor networks. *Sensors*, 19(23):5086, 2019.

[48] I. Lučin et al. Assessment of head loss coefficients for water turbine intake trashracks by numerical modeling. *Journal of advanced research*, 21:109–119, 2020.

[49] I. Lučin et al. Data-driven leak localization in urban water distribution networks using big data for random forest classifier. *Mathematics*, 9(6):672, 2021.

[50] I. Lučin et al. Detailed leak localization in water distribution networks using random forest classifier and pipe segmentation. *IEEE Access*, 9:155113–155122, 2021.

[51] I. Lučin et al. Machine-learning classification of a number of contaminant sources in an urban water network. *Sensors*, 21(1):245, 2021.

[52] I. Lučin et al. Source contamination detection using novel search space reduction coupled with optimization technique. *Journal of Water Resources Planning and Management*, 147(2):04020100, 2021.

[53] D. McInnis. A relative-risk framework for evaluating transient pathogen intrusion in distribution systems. *Urban Water Journal*, 1(2):113–127, 2004.

[54] L. S. McNeill and M. Edwards. Iron pipe corrosion in distribution systems. *Journal-American Water Works Association*, 93(7):88–100, 2001.

[55] J. Meister et al. Hydraulics of horizontal bar racks for fish downstream migration. In *38th International Association for Hydro-Environmental Engineering and Research World Congress (IAHR 2019)*, 2019.

[56] H. Mohammed et al. Machine learning: based detection of water contamination in water distribution systems. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pages 1664–1671, 2018.

[57] S. R. Mounce et al. Development and verification of an online artificial intelligence system for detection of bursts and other abnormal flows. *Journal of Water Resources Planning and Management*, 136(3):309–318, 2010.

[58] P. Nader et al. Detection of cyberattacks in a water distribution system using machine learning techniques. In *2016 Sixth International Conference on Digital Information Processing and Communications (ICDIPC)*, pages 25–30. IEEE, 2016.

[59] S. S. Paul and M. S. Adaramola. Analysis of turbulent flow past bar-racks. In *ASME international mechanical engineering congress and exposition*, volume 46545, page V007T09A029. American Society of Mechanical Engineers, 2014.

[60] F. Pedregosa et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[61] E. J. Pérez-Pérez et al. Leak diagnosis in pipelines using a combined artificial neural network approach. *Control Engineering Practice*, 107:104677, 2021.

[62] A. Preis and A. Ostfeld. A contamination source identification model for water distribution system security. *Engineering Optimization*, 39(8):941–947, 2007.

[63] A. Preis and A. Ostfeld. Genetic algorithm for contaminant source characterization using imperfect sensors. *Civil Engineering and Environmental Systems*, 25(1):29–39, 2008.

[64] T. Qin and D. L. Boccelli. Grouping water-demand nodes by similarity among flow paths in water-distribution systems. *Journal of Water Resources Planning and Management*, 143(8):04017033, 2017.

[65] Y. Qin et al. Microfabricated electrochemical pH and free chlorine sensors for water quality monitoring: recent advances and research challenges. *RSC Advances*, 5(85):69086–69109, 2015.

[66] M. Quiñones-Grueiro et al. An unsupervised approach to leak detection and location in water distribution networks. *International Journal of Applied Mathematics and Computer Science*, 28(2):283–295, 2018.

[67] B. Rajani and Y. Kleiner. Comprehensive review of structural deterioration of water mains: physically based models. *Urban Water*, 3(3):151–164, 2001.

[68] S. Raynal et al. An experimental study on fish-friendly trashracks–part 1. inclined trashracks. *Journal of Hydraulic Research*, 51(1):56–66, 2013.

[69] S. Raynal et al. An experimental study on fish-friendly trashracks–part 2. angled trashracks. *Journal of Hydraulic Research*, 51(1):67–75, 2013.

[70] S. Raynal et al. Numerical simulations of fish-friendly angled trashracks at model and real scale. In *35th World Congress of the International Association for Hydro-Environment Engineering and Research*, 2013.

[71] L. A. Rossman. Epanet 2: Users manual. Tech. Rep. EPA/600/R-00/057. 2000.

[72] R. Sadiq et al. Water quality failures in distribution networks—risk analysis using fuzzy logic and evidential reasoning. *Risk Analysis: An International Journal*, 27(5):1381–1394, 2007.

[73] M. Salah Abd Elmoaty. An experimental investigation of the impact of aquatic weeds trash racks on water surface profile in open channels. *Water Science*, 33(1):84–92, 2019.

[74] A. M. A. Sattar et al. Extreme learning machine model for water network management. *Neural Computing and Applications*, 31(1):157–169, 2019.

[75] A. Seth et al. Testing contamination source identification methods for water distribution networks. *Journal of Water Resources Planning and Management*, 142(4):04016001, 2016.

[76] A. Soldevila et al. Leak localization in water distribution networks using a mixed model-based/data-driven approach. *Control Engineering Practice*, 55:162–173, 2016.

[77] A. Soldevila et al. Leak localization in water distribution networks using bayesian classifiers. *Journal of Process Control*, 55:1–9, 2017.

[78] A. M. St. Clair and S. Sinha. State-of-the-technology review on water pipe condition, deterioration and failure rate prediction models! *Urban Water Journal*, 9(2):85–112, 2012.

[79] C. Sun et al. Leak localization in water distribution networks using pressure and data-driven classifier approach. *Water*, 12(1):54, 2020.

[80] A. S. Tijsseling. *Fluid-structure interaction in the case of waterhammer with cavitation.* PhD thesis, Technische Universiteit Delft (Netherlands), 1995.

[81] P. Vankayala et al. Contaminant source identification in water distribution networks under conditions of demand uncertainty. *Environmental Forensics*, 10(3):253–263, 2009.

[82] D. Vries et al. Application of machine learning techniques to predict anomalies in water supply networks. *Water Science and Technology: Water Supply*, 16(6):1528–1535, 2016.

[83] H. Wang and K. W. Harrison. Improving efficiency of the bayesian approach to water distribution contaminant source characterization with support vector regression. *Journal of Water Resources Planning and Management*, 140(1):3–11, 2014.

[84] T. Westrell et al. A theoretical approach to assess microbial risks due to failures in drinking water systems. *International Journal of Environmental Health Research*, 13(2):181–197, 2003.

[85] A. J. Whittle et al. Sensor networks for monitoring and control of water distribution systems. In *6th International Conference on Structural Health Monitoring of Intelligent Infrastructure (SHMII 2013)*. International Society for Structural Health Monitoring of Intelligent Infrastructure, 2013.

[86] D. C. Wiggert and A. S. Tijsseling. Fluid transients and fluid-structure interaction in flexible liquid-filled piping. *Applied Mechanics Reviews*, 54(5):455–481, 2001.

[87] D. Wilson et al. State-of-the-art review of water pipe failure prediction models and applicability to large-diameter mains. *Urban Water Journal*, 14(2):173–184, 2017.

[88] Y. Wu and S. Liu. A review of data-driven approaches for burst detection in water distribution systems. *Urban Water Journal*, 14(9):972–983, 2017.

[89] F. Xiang et al. Experimental study of debris capture efficiency of trash racks. *Journal of Hydro-environment Research*, 3(3):138–147, 2009.

[90] X. Yan et al. Research on contaminant sources identification of uncertainty water demand using genetic algorithm. *Cluster Computing*, 20(2):1007–1016, 2017.

[91] X. Yan et al. Multimodal optimization problem in contamination source determination of water supply networks. *Swarm and Evolutionary Computation*, 47:66–71, 2019.

[92] X. Yan et al. Pollution source localization in an urban water supply network based on dynamic water demand. *Environmental Science and Pollution Research*, 26(18):17901–17910, 2019.

[93] A. Yazdani and P. Jeffrey. Complex network analysis of water distribution systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 21(1):016111, 2011.

[94] D. Zaman et al. A review of leakage detection strategies for pressurised pipeline in steady-state. *Engineering Failure Analysis*, 109:104264, 2020.

[95] M. Zayed et al. An experimental investigation of head loss through a triangular "v-shaped" screen. *Journal of Advanced Research*, 10:69–76, 2018.

[96] M. Zayed et al. An experimental study on angled trash screen in open channels. *Alexandria Engineering Journal*, 57(4):3067–3074, 2018.

[97] M. Zayed et al. Experimental investigation of curved trash screens. *Journal of Irrigation and Drainage Engineering*, 146(6):06020003, 2020.

[98] M. Zayed and E. Farouk. Effect of blocked trash rack on open channel infrastructure. *Water Practice and Technology*, 16(1):247–262, 2021.

[99] X. Zhou et al. Deep learning identifies accurate burst locations in water distribution networks. *Water Research*, 166:115058, 2019.

List of Figures

List of Tables

Curriculum Vitae

Ivana Lučin was born in Rijeka, Croatia in 1991. She received mag. ing. mech. title in 2015. at Faculty of Engineering, University of Rijeka. In 2016. she enrolled in Postgraduate University Doctoral Study in Computational Mechanics at Faculty of Engineering, University of Rijeka under the supervision of Prof. D. Sc. Zoran Čarija. Since 2016. she is working as research and teaching assistant at Faculty of Engineering, University of Rijeka at Department of Fluid Mechanics and Computational Engineering. She is teaching assistant on the courses Computer Applications in Engineering, Computational Methods, Hydraulic Machines, Programming: Scripting Languages, Computational Fluid Dynamics, and Visualisation and Setup of Computer Simulations at the undergraduate and graduate university study of mechanical engineering and computer science. Her research areas are application of optimization and machine learning methods in engineering. She is a member of the Center for Advanced Computing and Modeling at the University of Rijeka.

52

List of Publications

Scientific papers in peer-reviewed journals:

- Lučin, I., Čarija, Z., Lučin, B. and Družeta, S., 2021. Detailed Leak Localization in Water Distribution Networks Using Random Forest Classifier and Pipe Segmentation. IEEE Access , 9, pp. 155113-155122.

- Grbčić, L., Kranjčević, L., Lučin, I. and Sikirica, A., 2021. Large Eddy Simulation of turbulent fluid mixing in double-tee junctions. Ain Shams Engineering Journal, 12(1), pp.789-797.

- Lučin, I., Grbčić, L., Družeta, S. and Čarija, Z., 2021. Source Contamination Detection Using Novel Search Space Reduction Coupled with Optimization Technique. Journal of Water Resources Planning and Management, 147 (2), p. 04020100.

- Lučin, I., Lučin, B., Čarija, Z. and Sikirica, A., 2021. Data-Driven Leak Localization in Urban Water Distribution Networks Using Big Data for Random Forest Classifier. Mathematics, 9 (6), p. 672.

- Lučin, I., Grbčić, L., Čarija, Z. and Kranjčević, L., 2021. Machine-Learning Classification of a Number of Contaminant Sources in an Urban Water Network. Sensors, 21 (1), p. 245.

- Sikirica, A., Čarija, Z., Lučin, I., Grbčić, L. and Kranjčević, L., 2020. Cavitation Model Calibration Using Machine Learning Assisted Workflow. Mathematics, 8(12), p.2107.

- Grbčić, L., Lučin, I., Kranjčević, L. and Družeta, S., 2020. Water supply network pollution source identification by random forest algorithm. Journal of Hydroinformatics, 22(6), pp.1521-1535.

- Družeta, S., Ivić, S., Grbčić, L. and Lučin, I., 2020. Introducing languid particle dynamics to a selection of PSO variants. Egyptian Informatics Journal, 21(2), pp.119-129.

- Grbčić, L., Lučin, I., Kranjčević, L. and Družeta, S., 2020. A machine learning-based algorithm for water network contamination source localization. Sensors, 20(9), p.2613.

- Grbčić, L., Kranjčević, L., Družeta, S. and Lučin, I., 2020. Efficient double-Tee junction mixing assessment by machine learning. Water, 12(1), p.238.

- Lučin, I. , Čarija, Z., Grbčić, L., Kranjčević, L., 2020. Assessment of Head Loss Coefficients for Water Turbine Intake Trash-Racks by Numerical Modeling. Journal of Advanced Research, 21, pp. 109-119.

- Sikirica, A., Čarija, Z., Kranjčević, L. and Lučin, I., 2019. Grid type and turbulence model influence on propeller characteristics prediction. Journal of Marine Science and Engineering, 7(10), p.374.

- Grbčić, L., Kranjčević, L., Lučin, I. and Čarija, Z., 2019. Experimental and numerical investigation of mixing phenomena in double-Tee junctions. Water, 11(6), p.1198.

- Ivić, S., Družeta, S., Hreljac, I., 2017. S-Lay pipe laying optimization using specialized PSO method. Structural and Multidisciplinary Optimization, 56(2), pp.297-313.

Scientific papers in conferences:

- Sikirica, A., Lučin, I., Čarija, Z. and Lučin, B., 2020. CFD Analysis of Marine Propeller Configurations in Cavitating Conditions. Pomorski zbornik, (3), pp.251-264.

- Lučin, I., Kranjčević, L., Čarija, Z. and Mogorović, A., 2018. Experimental Setup of Fluid Mixing in Double Tee-Junctions. In Proceedings of the 29th DAAAM international symposium (Vol. 29, pp. 1046-52).

- Čarija, Z., Lučin, I., Lučin, B. and Grbčić, L., 2018, January. Investigation of numerical simulation parameters on fluid flow around trash-racks. In Proceedings of the 29th DAAAM international symposium (Vol. 29, pp. 1046-52).

- Zeng, H., Grbčić, L., Lučin, I. and Kranjčević, L., 2018. Mesh creation for realistic terrain cases for shallowfoam-2D OpenFoam solver. In Proceedings of the 29th DAAAM international symposium (Vol. 29, pp. 1046-52).

Part II

Included Publications
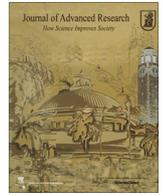
Assessment of head loss coefficients for water turbine intake trash-racks by numerical modeling

Ivana Lučin [a], Zoran Čarija [a,b,*], Luka Grbčić [a], Lado Kranjčević [a,b]

[a] *Faculty of Engineering, University of Rijeka, Vukovarska 58, 51000 Rijeka, Croatia*
[b] *Center for Advanced Computing and Modelling, University of Rijeka, Radmile Matejčić 2, 51000 Rijeka, Croatia*

HIGHLIGHTS

- Numerical modeling can be used to evaluate head losses for different trash-racks.
- Rectangular bar cross section mostly generates greatest head-losses.
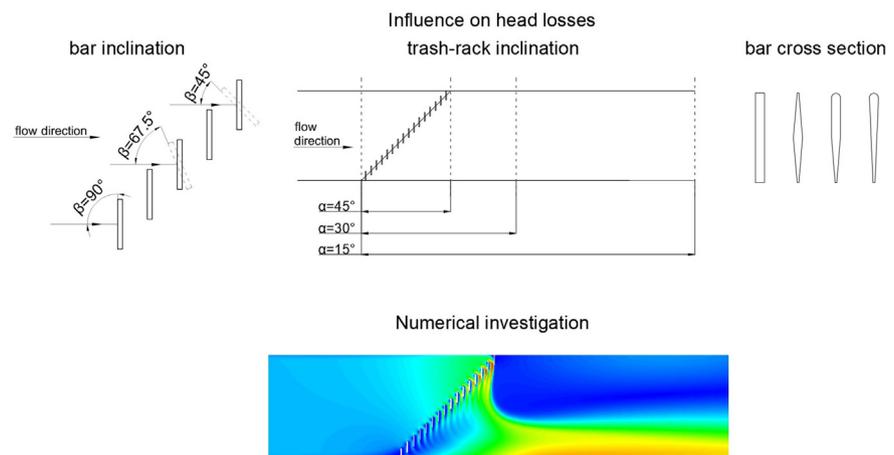- Change in bar cross sections can lead to considerable head-loss reduction.
- Optimization can be conducted to provide innovative trash-rack design.

GRAPHICAL ABSTRACT

ARTICLE INFO

ABSTRACT

In this work, numerical simulations of fluid flow around trash-rack for different bar cross sections are conducted to investigate cross section influence on head losses. Comparison with experimental data is conducted to validate the usage of numerical simulations which enable investigation of great number of trash-rack configurations. In previous experimental studies researchers mostly focused on trash-rack parameters (bar spacing, bar length, inclinations etc.) where bar cross section was mainly rectangular or streamlined shape. Therefore, 2D simulations for different cross sections are carried out for a range of trash-rack configurations in order to provide better insight how it affects energy losses. It is shown that head loss reduction due to change in cross section is greatly dependent on trash-rack configuration, therefore optimization of simplified real water turbine trash-rack is also conducted to produce the cross section that generates smallest head losses for given configuration.

Introduction

Trash-racks are installed in the intake system of hydroelectric power plants to prevent entrance of large debris which can damage turbine parts and cause serious problems in power plant operation. Installation of trash-rack causes disturbance in fluid flow with

inevitable energy losses which should be minimized. To reduce these losses and to keep the design simple for manufacturing and cleaning, trash-racks, oriented perpendicular to fluid flow, usually consist of many rectangular bars directed parallel to fluid flow. Another main purpose of trash-rack is to prevent fish species from entering the intake system [1]. With growing ecological concern [2], influence of trash-rack design on fish migration and fish mortality is increasingly taken into consideration [3,4]. Trash-rack is not a suitable obstacle for some fish species, especially for juvenile fish, which could be entrapped in turbine parts. Furthermore, in case of large approaching velocities, some fish species are incapable of avoiding trash-rack which can cause fatal injuries when colliding with bars. Increased awareness of these problems prompted a change in the design of hydroelectric power plants intake system. Inclined trash-racks in combination with angled bars are increasingly considered to provide better fish guidance toward fishways which are being installed to provide safe passageway for upstream or downstream migration considering fish behaviour [5,6]. Due to site specifications and fish species characteristics, great number of case studies regarding fishway efficiency are being conducted [7,8]. Multidisciplinary approach is also considered to improve current knowledge and practice of fishways [9].

To determine energy losses, a number of experimental investigations on trash-racks were conducted. Idel'chik [10] proposed empirical relationship regarding different bar cross sections, bar spacing and rack angles that estimates head loss for bars parallel to fluid flow. The United States Army Corps of Engineers [11] proposed head loss coefficient values based on summarized open channel tests with racks perpendicular to fluid flow for different bar designs and spacings. Tsikata et al. [12] experimentaly investigated influence of bar spacing and bar length on head losses where it was shown that bar length reduction and increasement in bar spacing reduce head losses. Furthermore, fluid flow around angled bar racks and influence of different cross sections (rectangular, bar with rounded leading edge and streamlined bar) were analysed in [13]. Significant reduction in head losses was observed when rectangular cross section edges were rounded or cross section was replaced with streamlined shape. Bar inclination to the approaching flow was investigated only for rectangular cross section while for other cross sections bars remained parallel to the fluid flow, where head loss value increased when bar inclination increased. Asymmetric flow behind inclined bars was also reported. Vortex shedding behind trash-rack bars induce vibrations which can interfere with natural frequency of the trash-rack and cause damage to bars. Therefore, structural aspect of trash-rack exploitation must be also taken into consideration [14]. Since design of trash-rack varies greatly and trash-racks are used in wide range of operating conditions, number of experimental and numerical studies investigated this problem [15–17]. Clark et al. [18] analysed head losses for six different cross sections (rectangular, rounded, commercially available bar and variants of NACA airfoil) for bars parallel to fluid flow and reported increase in head loss when channel inclination before trash-rack i.e. approach flow inclination increases. In Raynal et al. [19] different trash-rack inclinations with regards to channel flume bottom were investigated where new head loss equation considering blockage ratio, bar shape and rack inclination was proposed. Additionally [20], rectangular and hydrodynamic bar shapes were analyzed for various trash-rack to flume wall angles (while bars were kept perpendicular to the trash-rack). In more recent research, Albayrak et al. [21] investigated a wide range of angled trash-rack configurations for rectangular and rounded bars and other geometry parameters and proposed new head loss equation which included relation between bar spacing, rack and bar angles (primary parameters) and bar length, relative rack submergence and bar shape (secondary parameters). Szabo-Meszaros et al. [4]

examined six different configurations of streamlined and rectangular bar profiles for different bar-setups; in four configurations trash-rack was inclined against the channel wall for various bar angles while two configurations had horizontally oriented bars. Horizontal trash-racks and vertical with streamlined bars were suggested as the best candidates for fish-friendly trash-racks. Zayed et al. investigated influence of screen angle from the trapezoidal open channel wall for angled trash-racks [22] and V-shaped trash-rack [23], both with circular bars where new head loss equations were proposed. In Beck et al. [24] a new innovative curved bar design was investigated. Böttcher et al. [25] compared trash-rack with circular bars and new fish protection and guidance system - flexible fish fence where common head loss equations were adapted for new proposed design.

Few numerical studies investigated flow around trash-racks. Raynal et al. [26] validated two-dimensional fluid flow analysis for bars angled at 45° using their previous experimental results, under-estimating head loss experimental results. In work by Paul et al. [27], 3D analysis of fluid flow around 3 and 7 submerged bar-racks was conducted, where numerical analysis overestimated experimental head loss coefficient. Åkerstedt et al. [28] conducted an investigation for rectangular and biconvex bars for different inclinations of fully submerged trash-rack, where simplification was made with periodic boundary conditions and two-dimensional fluid flow domain.

It can be noticed that most experiments from previous studies considered two bar cross section shapes at most, whereas the proposition of different cross sections could provide more favourable hydraulic conditions, especially considering configurations with angled trash-racks and angled bars. The main problem with innovative designs (e.g. V shape trash-rack in [23] and curved bar shapes in [24]) is that researchers usually define trash-rack geometry a priori, hence optimal solution could be overlooked. The uniqueness of power plant intake geometry must also be taken into consideration since channel geometry after trash-rack is usually not regular as in experimental setups. Geometry changes in the intake channel, inclination or narrowing, are important since they affect head losses, especially if analyzing bars with greater angle of inclination. In those cases, recirculating zones are longer with the possibility of geometry interference in the wake zone which may also lead to turbine efficiency reduction. Numerical studies provide a solution for a number of presented problems. In the numerical approach, the whole turbine geometry can be modelled in full scale, the influence of all geometry parameters can be evaluated and fluid flow can be investigated in more detail [29]. Furthermore, an optimization procedure can be conducted to provide optimal trash-rack configuration for specific turbine that is investigated.

In this work, numerical simulations are conducted for four different cross sections with different configurations of trash-rack and bar inclinations. To validate the numerical results, trash-rack configurations are chosen according to experiments conducted by Albayrak et al. [21]. Following the numerical model validation, cross section influence on head loss reduction for different configurations is further investigated. Finally, optimization of simplified trash-rack geometry for a 50 years old hydroelectric power plant HE Senj (Senj, Croatia) is conducted in order to provide optimal cross-section regarding the head losses.

Materials and methods

Geometry definition

Numerical simulations are conducted for trash-rack inserted in 1 m wide, 12 m long and 0.1 m deep flume with constant rectangu-

lar cross section (Fig. 1). Flume and bar dimensions are chosen to validate numerical simulation with full scale trash-rack model investigated in Albayrak et al. [21], for the trash-rack inclination of 45° and rectangular bars with inclinations of 45°, 67.5° and 90°. Bars are considered completely submerged. Flow velocity ranges from 0.13 to 0.43 m/s, in accordance to experiment. Trash-rack bars are 0.1 m long with the greatest cross section width of 0.01 m and with bar spacing 0.05 m. Reynolds bar number $Rs = Us/v$ where U is approaching velocity and s bar width ranges from 1295 to 4285. After validation, further investigation is conducted for trash-rack inclinations of (α angle) 15°, 30° and 45° with bar inclinations of (β angle) 45°, 67.5° and 90° for different cross sections. Influence of bar spacing on head losses for different bar and trash-rack inclinations is investigated in Albayrak et al. [21] so this parameter is kept constant for all conducted simulations and only influence of cross section change was considered.

Four different cross section geometries are considered – rectangular, rhombus, rounded front edge with inclined back in the lower half and rounded front edge with inclination starting right after rounded edge (Fig. 1). Hereinafter considered cross sections will be referred to as cross section A, B, C and D, respectively. In cross sections B, C and D, sharp edges are avoided due to production reasons. Consequently, 2 mm straight segments can be seen in cross section profiles.

Geometry and trash-rack placement in the channel can be seen in Fig. 1. The trash-rack origin for all geometries is set at 3 m downstream from the inlet. Cross sections considered for head loss measurements for numerical model validation are defined 3 m upstream (inlet) and 3 m downstream from the trash-rack

origin. For all other configurations, head loss measurements were conducted on inlet and outlet cross sections.

Number of bars on trash-rack depends on α and β angles, which leads to different blockage of fluid flow on flume sides for different configurations, e.g. for configuration $\beta = 90°$ and $\alpha = 45°$ bars can be spaced on trash-rack in a way there is no clearance on flume sides or with clearance on both flume sides if one bar is removed. Numerical investigation of both configurations shows around 15% difference in head loss coefficient. Considering this information is usually not mentioned when the experiment is described to avoid influence of side clearance, outer bars were extended to completely block the fluid flow. A similar method can be seen in Raynal [26] where sides of the trash-rack domain were cut off.

For the configuration with greatest fluid flow blockage ($\alpha = 45°, \beta = 90°$), 3D multiphase, 3D single phase and 2D simulations are conducted. 3D multiphase fluid flow simulation best describes the open channel nature of the experiment but requires considerable computational resources, thus simplification is made to reduce computational time. A 3D geometry is created where domain height is set as an estimation of free surface level which was constant throughout the whole domain. This simplification allowed usage of a single phase fluid flow model which significantly reduced computational time. Since cross section along the vertical axis remained constant, 2D simulations are also considered. All three simulations provide similar results - both 3D models underestimate the head loss coefficient for around 14% while 2D single phase model underestimation is around 15%. Consequently, for all configurations, 2D simulation is chosen in order to reduce computational time.
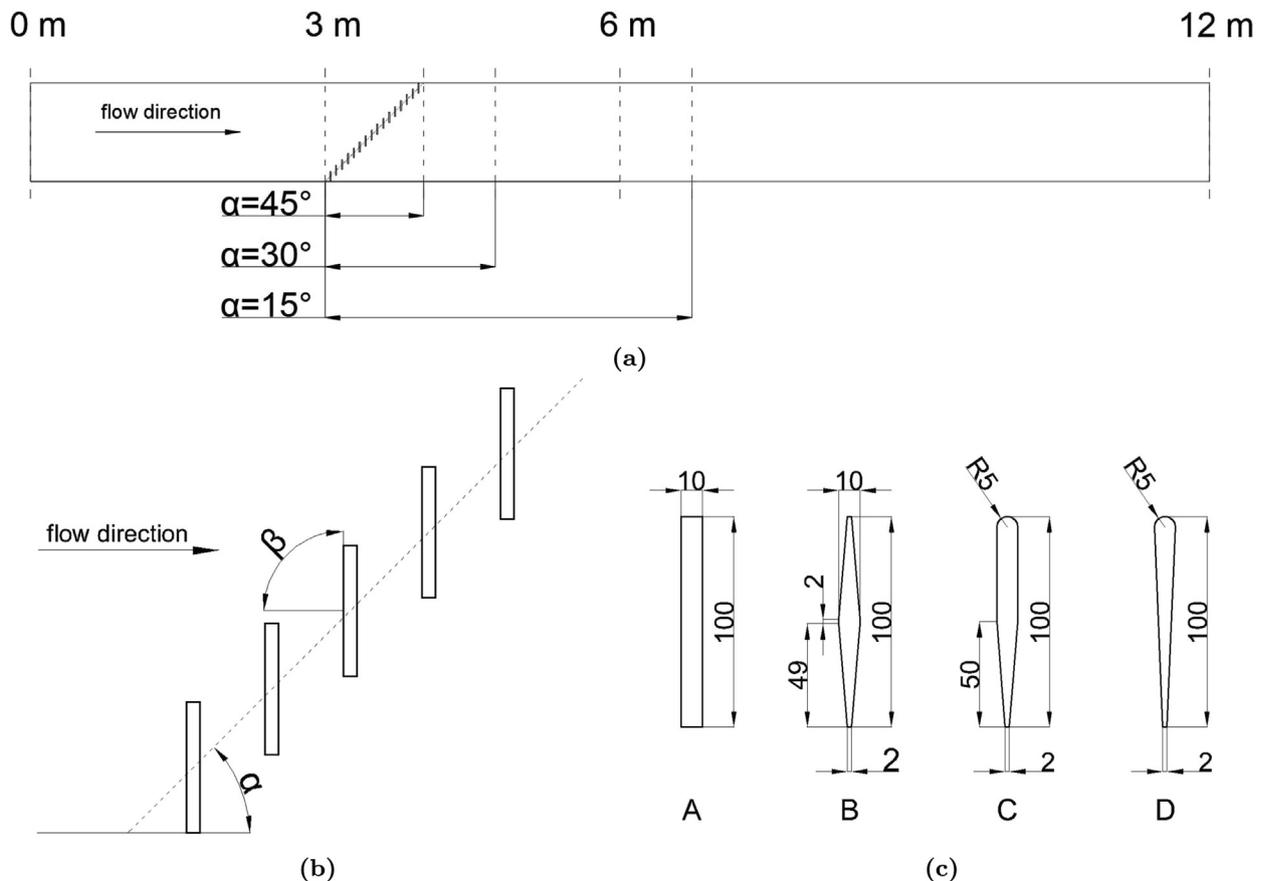

Fig. 1. (a) Numerical domain with trash-rack position and measurement locations (plan view). (b) Trash-rack detail with trash-rack inclination α and bar inclination β (plan view). (c) Bar cross sections with indicated dimensions (in mm) used for fluid flow simulations (plan view).

Table 1
Head loss coefficient relative error for different mesh sizes with different global element edge size.

number of elements	230 000	413 000	723 000	920 000
element size	0.007 m	0.005 m	0.004 m	0.0035 m
$\varepsilon\,(k_n)$	2.02%	0.07%	0%	0%

Numerical model

Simulations are conducted in ANSYS-Fluent for an unstructured mesh with local refinement around trash-rack and channel walls. Considering that changes in trash-rack configuration greatly influence fluid flow field (e.g. width and length of recirculation zone) and keeping in mind that optimization should be conducted with automated meshing, i.e. cannot be further refined according to simulation results, meshing parameters are kept the same for all considered configurations. First layer height is defined to maintain $y^+ > 30$ and scalable wall functions are used. Global element edge size is defined to be within 0.016 m and 0.0001 m with prescribed value of 0.004 m. Maximum size of the element edge for bar edges is defined as 0.003 m and for channel wall 0.005 m. Mesh independence study is conducted for configuration $\alpha = 45°, \beta = 90°$ with numerical meshes sizing 230 000, 413 000, 723 000 and 920 000 elements (Table 1). Values of head loss coefficient became constant for numerical mesh consisting of 723 000 elements, which prompted the choice of the grid with around 800 000 elements (depending on trash-rack configuration) for all simulations. Detailed investigation of turbulent models for numerical simulations of fluid flow around trash-rack was conducted in previous study [30], where it was observed that k-ϵ standard turbulence model generates greatest head loss values showing the best agreement with experimental results at the same time. In general, when greater recirculation zone is present behind trash-rack, k-ϵ standard turbulence model shows stability in results, while other models tend to oscillate. Due to these observations, k-ϵ standard turbulence model is chosen for all simulations in this study. Overview of boundary conditions can be seen in Table 2.

Numerical simulation is done by solving the steady-state incompressible isothermal Navier–Stokes (NS) equations which describe the fluid flow:

$$\nabla \cdot \mathbf{u} = 0 \tag{1}$$

$$(\mathbf{u} \cdot \nabla)\mathbf{u} = -\frac{1}{\rho}\nabla p + v\nabla^2\mathbf{u} + \mathbf{f} \tag{2}$$

where $\mathbf{u}$ is the velocity vector, p represents the pressure, ρ is the fluid density, v is the fluid kinematic viscosity and $\mathbf{f}$ represents the external forces acting upon the fluid (e.g. gravity). Eq. (1) represents the conservation of mass while Eq. (2) defines the conservation of momentum of fluid flow. Reynolds averaging is additionally applied to the NS equations for turbulence modeling.

Chosen fluid is water with properties for temperature of 20° (Table 3). Pressure-velocity coupling SIMPLE algorithm is used and discretization scheme for the convection terms of governing equations is second order upwind. Convergence criteria is assumed if all residuals drop below 10^{-5} and additionally no change of head loss coefficient is observed with further iterations.

Results and discussion

Validation of simulation

Validation is conducted for rectangular bars with α angle 45° and β angles 45°, 67.5° and 90° for four different velocities, 0.13, 0.23, 0.33 and 0.43 m/s. Head loss coefficient is calculated as (to match the head loss coefficient in Albayrak [21]):

$$k = \Delta p \frac{2g}{U_0^2} \tag{3}$$

In Eq. (3) U_0 is inlet velocity and Δp is the pressure difference between upstream and downstream cross sections. Pressure difference represents an approximation of water level difference (Δh) present in experiments. This assumption is validated with aforementioned comparison with multiphase simulation results where small variation was present for both considered models.

A greater recirculation zone for trash-racks with greater bar inclination is noticed in the simulations (Fig. 2). Highly turbulent flow behind trash-rack was also observed in experiments [21]. For β angle 45° recirculation zone accounts for around one third of channel cross section, which is in agreement with Raynal [26]. For β angle 67.5° recirculation zone is present in around half of the channel, while for β angle 90° recirculation zone increases even more and with that suppresses fluid flow and increases head losses (pressure drop). For the same inlet velocities, with the change in trash-rack configuration, greater recirculation zone leads to higher magnitudes of velocities due to the reduced cross sectional area available for fluid flow. This produces a greater variance in downstream velocity profiles.

Measurement locations must be placed at an adequate distance where fluid flow is undisturbed in order to obtain precise data. That is often a problem, due to the space limitation of the experiment. Mean velocities at observed cross-sections, that are needed to determine head loss coefficient in the experiment, are calculated with water height measurements at a given number of points or combined with flow rate measurements - depending on available instruments. For example, in Albayrak [21] three points in the measurement cross section were considered. This is especially a problem if measurements are made in a recirculation zone where great velocity variation in the cross section is present. Therefore, the average of measurements with a smaller number of points and measurements with a greater number of points can produce significantly different results.

With the increase in β angle, a greater deviation in results is observed, where simulation underestimates head loss coefficient with maximum deviation of 15%. Geometry simplifications must be taken into consideration regarding this deviation since the trash-rack structure is simplified, e.g. spacers are omitted from the geometry. Design of trash-rack sides is not defined in the

Table 2
Boundary conditions used for fluid flow simulation.

boundary	inlet	outlet	channel walls	bar walls	top	bottom
type	velocity inlet	pressure outlet	wall	wall	symmetry	symmetry
value	0.13–0.43 m/s	atmospheric pressure	no slip	no slip	–	–

Table 3
Fluid properties used for fluid flow simulation.

fluid	water
temperature [°C]	20
density [kg/m³]	998.2
viscosity [kg/m-s]	0.001

experiment description and is thus chosen arbitrarily for simulation, as mentioned previously in Section b. Albayrak [21] reported a head loss difference of 15% for some configurations due to scale effects. Free surface measurement can also generate errors, with a deviation of around 5% as reported in Raynal [20]. Also, when considering experiments which have low water heights, the bottom has a greater influence on head loss coefficient due to friction, while in real turbine intakes, these water heights are always greater. Water depth to channel width ratio in the experiment is always considerably smaller than 1, while in real intakes it is greater, making the influence of bottom surface negligible, thus resulting in an overestimation of head loss coefficient measurements in experimental studies. To avoid uncertainty regarding aforementioned issues, head loss coefficient is normalized as:

$$k_n = \frac{k_e}{k_{max}} \quad (4)$$

In Eq. (4) k_e represents experimental head loss coefficients for given trash-rack configuration and k_{max} represents maximum head loss coefficient observed in all considered experiments. Normalization of head loss coefficients will be used in the course of this study.

Validation of numerical results can be seen in Fig. 3. Normalized values of head loss coefficient obtained from simulations show good agreement with normalized values of experiment results. Greatest discrepancy is 4% for β=90° where for other configurations it is under 2%. Numerical analysis shows very small variation

in head loss coefficient due to change in inlet velocity, contrary to the experiment which is subjected to measurement errors. This behaviour is expected, because head loss coefficient equation is chosen to be invariant of the inlet velocity.

Numerical shape investigation

Numerical investigations are conducted for 4 different cross sections with 9 different combinations of α and β angles for inlet velocity 0.43 m/s. Measurement locations for verification are set at inlet and 6 m downstream from inlet. At these measurement locations for some configurations, large recirculation zone is observed and for configurations with α = 15° if trash-rack starts 3 m downstream, measurement location at 6 m is not behind the trash-rack (in experiment trash-rack location varied due to space limitation where in this study it is set 3 m downstream from the inlet). Therefore, numerical shape investigation measurements are conducted at inlet and outlet cross sections. Trash-rack position for different α angles and influence on fluid flow field can be seen in Fig. 4.

Investigations conducted for different cross sections with different ranges of bar and trash-rack inclinations showed that for most configurations, cross section A provides the greatest head loss coefficient (since the A area is the largest when compared to other bar types) with the exception of configuration α = 45°, β = 90° and α = 30°, β = 90° where cross section B generates the greatest head loss coefficient. This could be explained with cross section A creating better fluid flow guidance (smaller turbulence zones) when fluid flow is perpendicular to bar orientation. The smallest head loss coefficient was observed mostly for cross section C, with the exception of configuration α = 15°, β = 45° where cross section B generated the smallest head loss. For greater α and β angles, selection of cross section is more relevant, whereas for smaller angles, the value of head loss coefficient is similar for all cross sections. These results are presented in Fig. 5 where values of normalized head loss coefficient (normalized with value of greatest head loss,
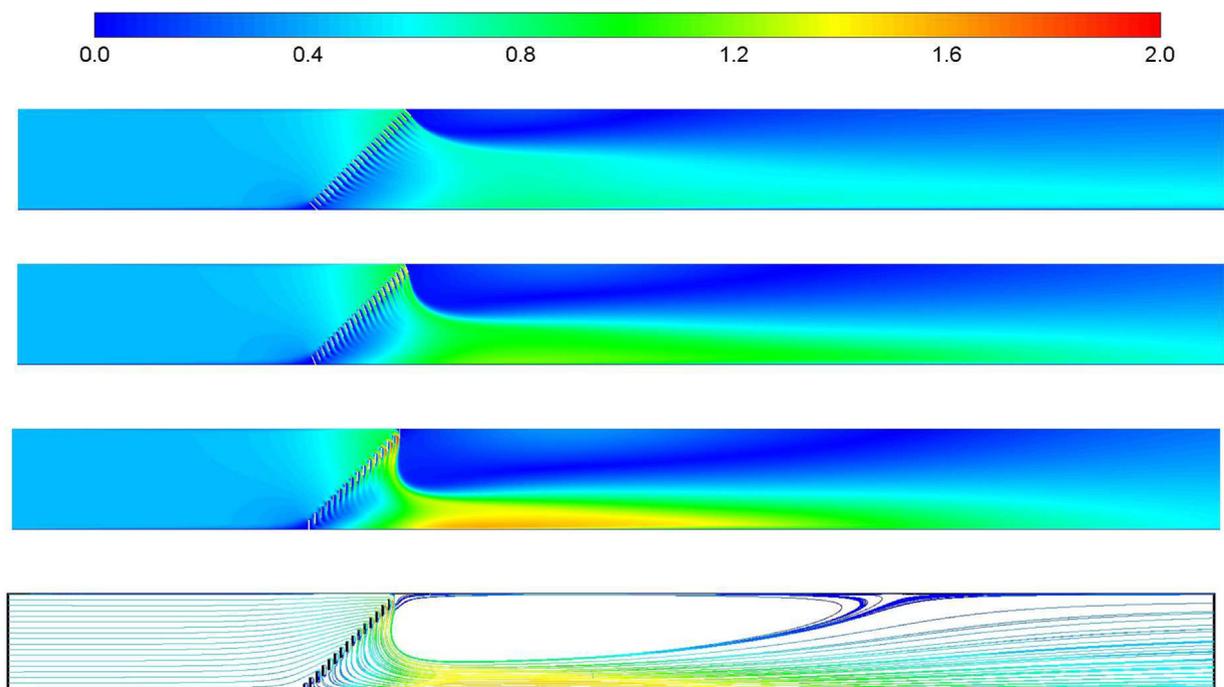
Fig. 2. Velocity magnitude (in m/s) for trash-rack configuration α = 45° and for β angles 45°, 67.5° and 90° (top to bottom) and pathlines coloured by velocity magnitude for trash-rack configuration α = 45° and β = 90°.
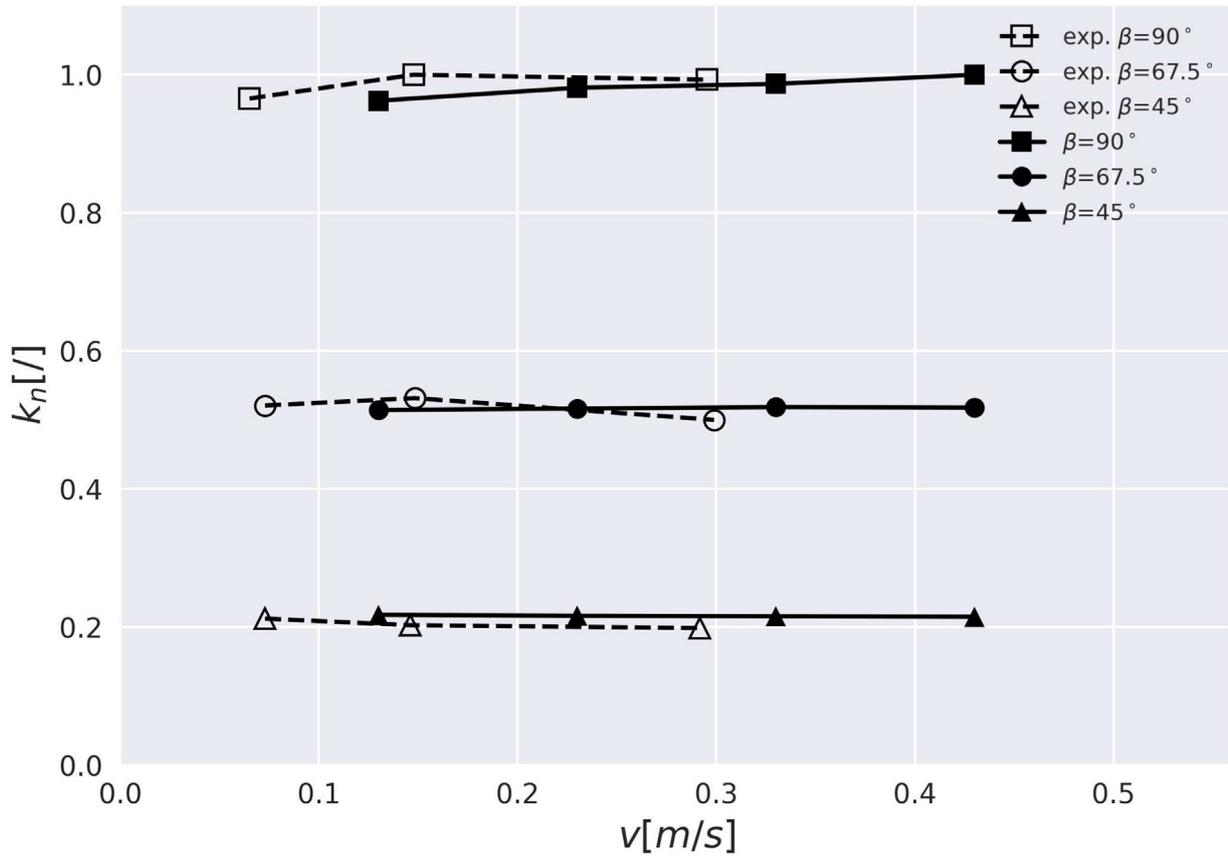
Fig. 3. Experimental and numerical results of normalized head loss coefficient for trash-rack configurations for angles $\alpha = 45°$ and $\beta = 45°$, $67.5°$ and $90°$.
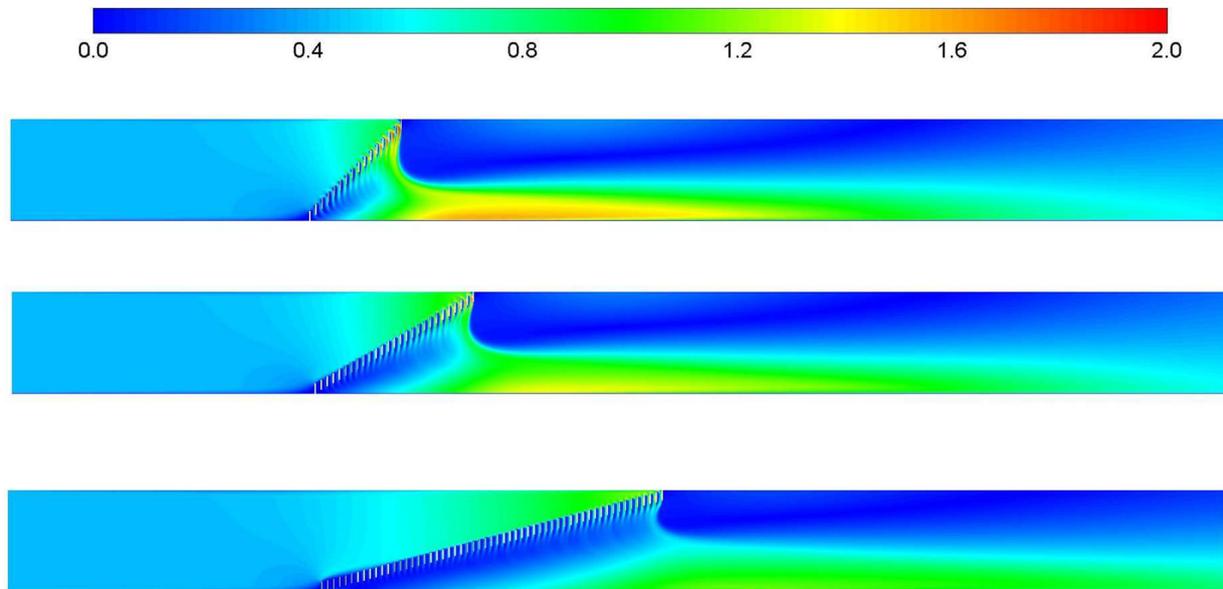
Fig. 4. Velocity magnitude for trash-rack configuration $\beta = 90°$ and for α angles (top to bottom) $45°$, $30°$ and $15°$.

i.e. cross section B for configuration $\alpha = 45°$, $\beta = 90°$), for cross sections that generate greatest and smallest head loss, are presented for all configurations.

It can be observed that trash-rack configuration (inclinations of trash-rack and bars) has the greatest influence on head loss. Simulation results show that for greatest bar inclination ($\beta = 90°$) reduc-tion of trash-rack inclination (α) leads to a reduction of head loss greater than 40%. For greatest considered trash-rack inclina-tion ($\alpha = 45°$), reduction of bar inclination leads up to head loss reduction of around 80%. For smaller inclinations (for exam-ple $\beta= 45°$ where α is changed or $\alpha=15°$ where β is changed) lesser reductions in head loss can be observed, which is expected due
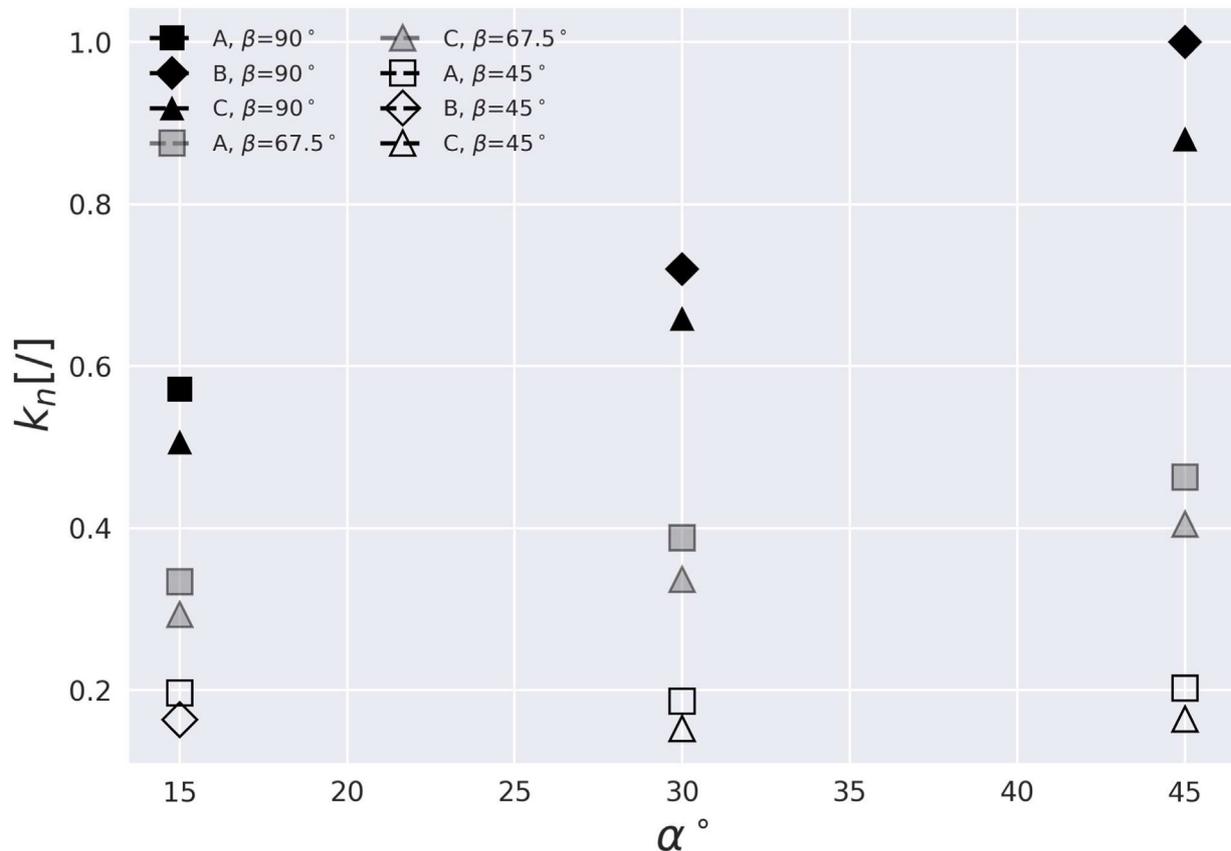
Fig. 5. Normalized head loss coefficient values for considered trash-rack configurations. Presented data shows only cross sections that generate minimum and maximum losses.

to the fact that values of head loss coefficient are generally smaller for these configurations. Configurations that provide better fish avoidance are increasingly being installed, but since they cause more losses, influence of cross section becomes more prominent.

In Fig. 6 normalized head loss values are presented for all considered configurations and for all cross sections. Reduction of head losses due to change in cross section accounts mostly for around 10%. Results for configuration $\alpha = 15°$, $\beta = 45°$ are not aligned with the trend of other configurations which could be explained due to small head loss coefficients for configurations with $\beta = 45°$ (seen in Fig. 5). For these configurations, a reduction of α angle or change in cross section geometry generates a very small reduction of head loss. For some configurations, different cross sections provide very similar results, where if the configuration is changed, the head loss coefficient difference becomes greater i.e. cross section selection is more prominent. For example, for trash-rack configuration $\alpha = 45°$, $\beta = 90°$ both cross section A and D generate similar head loss coefficient, where if α angle is decreased to $15°$ cross section A generates the greatest head loss coefficient. This shows that generalization of the optimal cross section cannot be made, hence it must be optimized for every trash-rack configuration, especially when new designs such as V-shaped trash-rack [23] start being implemented.

Cross section optimization

Optimization of bar cross section is conducted for turbine intake system of 50 years old hydroelectric power plant HE Senj (Senj, Croatia) (Fig. 7a). Since the power plant is in the need of reconstruction, a new trash-rack design is being considered also.

In the time of power plant construction there was no concern for fish species so trash-rack consisted of rectangular bars installed parallel and trash-rack perpendicular to fluid flow.

The optimization process is conducted for simplified geometry; trash-rack remained perpendicular and bars parallel to fluid flow. Distance between bars and their length is kept the same and only cross sections are changed. Validation of numerical simulation was conducted for rectangular cross section. Results showed good agreement with available empirical results [10] and with in situ measurements with error around 4%. Three different cross sections, which are chosen due to easy machining, are considered: cross section with front and back inclinations, cross section with curvature at front and back and cross section with front curvature and back inclination. For the first cross section, four optimization variables defining width and length of inclination are considered. The second cross section has three optimization variables which define the radius of front curvature and inclination width and length in the back. For the last cross section, two optimization variables which define the front and back curvature are considered (Fig. 7b). There are no limitations imposed on optimization variables due to construction reasons, thus considered shapes present theoretical solution. Overview of optimization variables for each optimization case is presented in Table 4.

Optimization is done using Particle Swarm Optimization (PSO) which is a population based search algorithm that is inspired by swarm intelligence, such as birds flock or fish school movements [31]. The starting point of PSO is to initially randomly generate, within certain bounds, a set of solutions (swarm) to a problem and iteratively evaluate the quality (fitness) of every candidate solution (particle). After every evaluation, the position of every particle is adjusted towards the local or global optimal position.
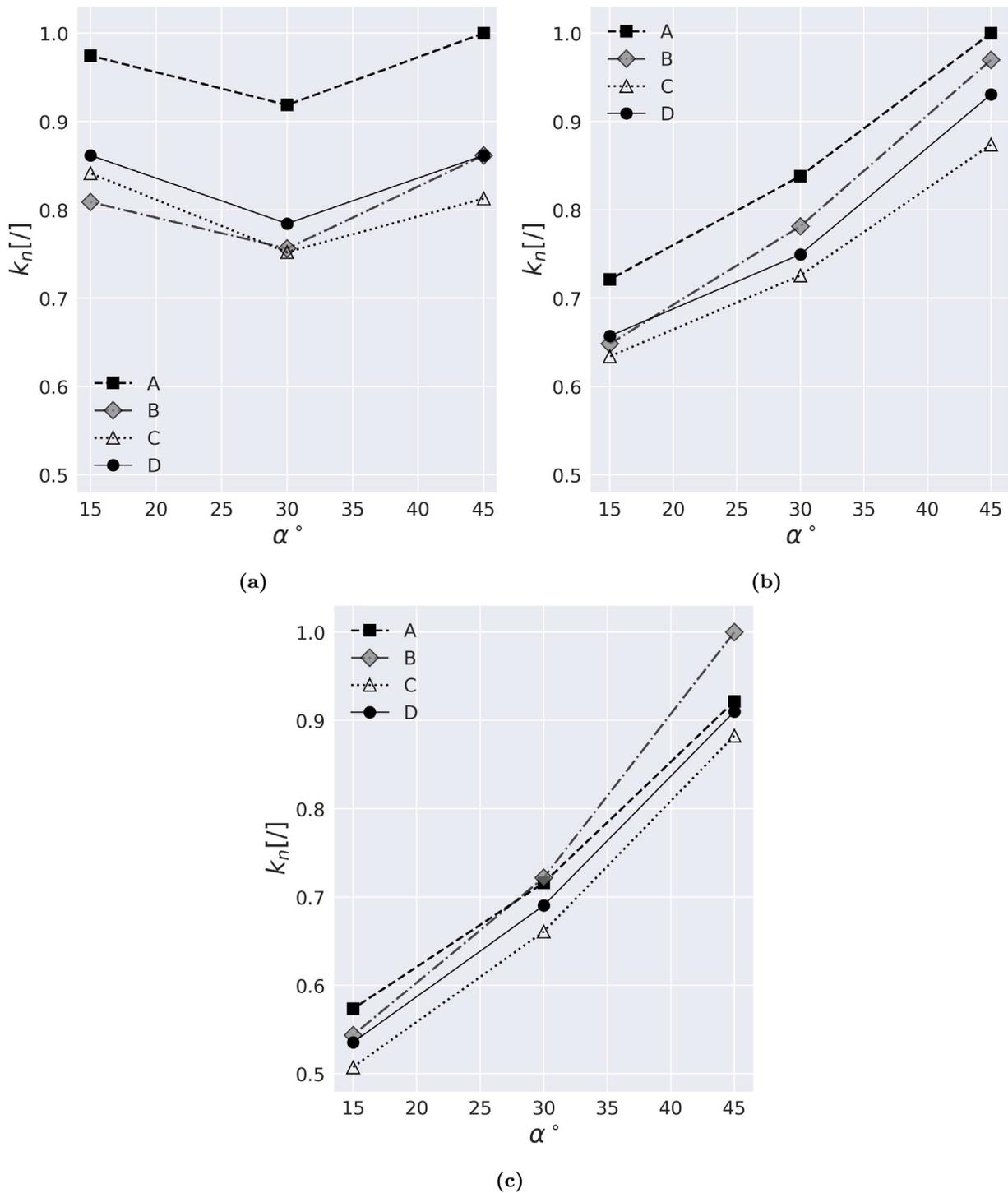
Fig. 6. Normalized head loss coefficients for (a) $\beta = 45°$, (b) $\beta = 67.5°$ and (c) $\beta = 90°$.

Movement of every particle through the problem space is influenced both by its own best solution and swarm's best solution. This process continues until values converge into a satisfactory and/or steady set of solutions. Factors such as particle cognitive rate, social rate, and problem space movement inertia greatly influence the optimal position convergence. The PSO algorithm implemented in the python optimization package inspyred is used with swarm size of 10 particles, inertia factor 0.75, cognitive rate 1 and social rate 1.

Goal functions for all considered optimization cases are defined as:

$$
\begin{aligned}
minf_a(\boldsymbol{x_a}) &= \Delta p(\boldsymbol{x_a}) \\
minf_b(\boldsymbol{x_b}) &= \Delta p(\boldsymbol{x_b}) \\
minf_c(\boldsymbol{x_c}) &= \Delta p(\boldsymbol{x_c})
\end{aligned}
\tag{5}
$$

In Eq. (5) Δp represents result of numerical simulation conducted for optimization variables $\boldsymbol{x_a}, \boldsymbol{x_b}$ or $\boldsymbol{x_c}$ which denotes vectors of optimization variables dependent on the case:

$$
\begin{aligned}
\boldsymbol{x_a} &= [a_1, a_2, a_3, a_4] \\
\boldsymbol{x_b} &= [b_1, b_2, b_3] \\
\boldsymbol{x_c} &= [c_1, c_2]
\end{aligned}
\tag{6}
$$

Fig. 7. (a) Intake structure of HE Senj with detail of current bar design. (b) Cross sections considered for optimization cases with optimization parameters (left) and their optimized shape (right).

Table 4
List of optimization parameters for optimization cases with parameter constraints (L denotes bar length and s bar width).

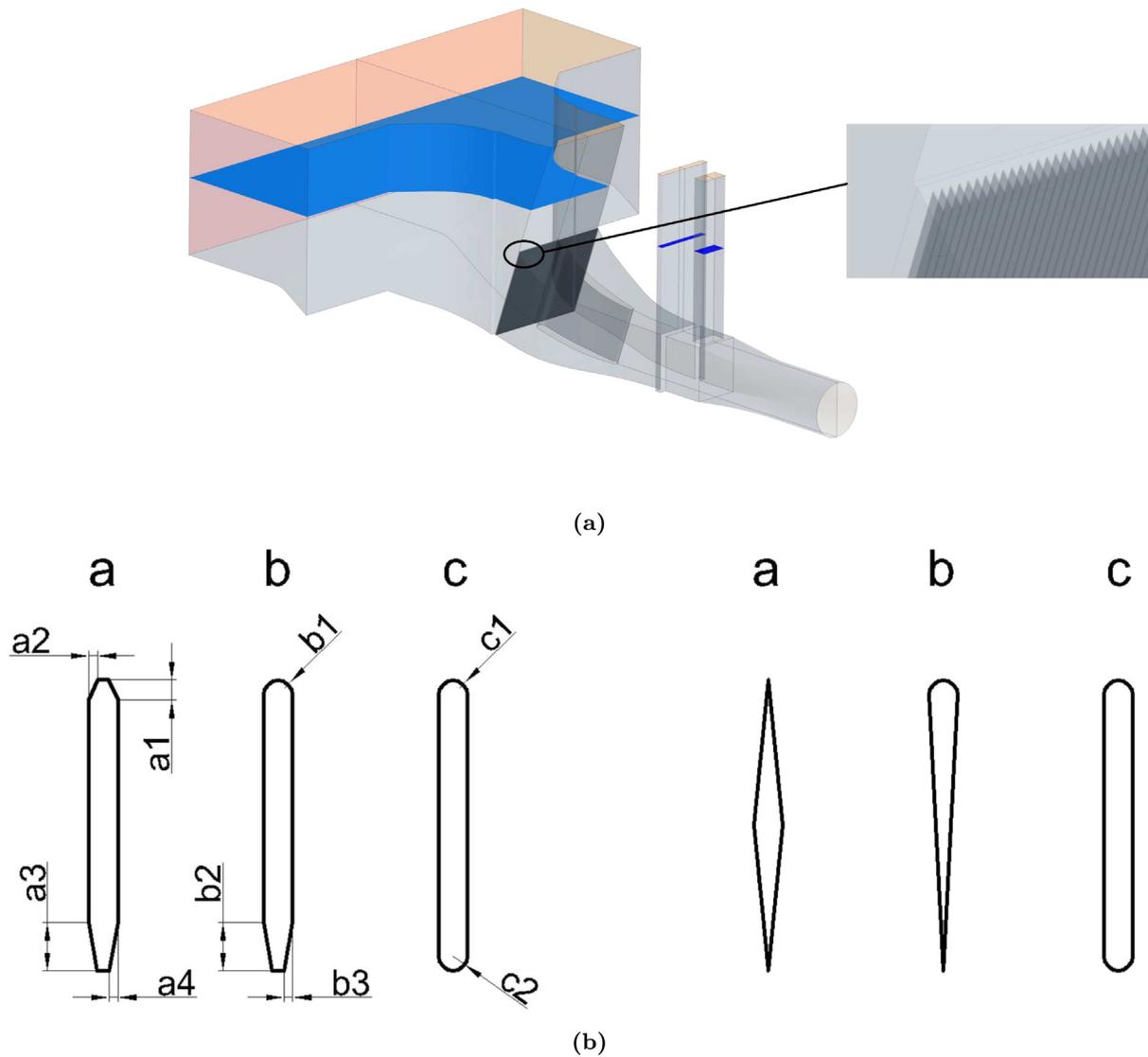| Optimization variables		Constraints [lower limit, upper limit]
Case a	x_a	
front inclination length	a1	[0, L/2]
front inclination width	a2	[0, s/2]
back inclination length	a3	[0, L/2]
back inclination width	a4	[0, s/2]
Case b	x_b	
front curvature radius	b1	[0, s/2]
back inclination length	b2	[0, L - s/2]
back inclination width	b3	[0, s/2]
Case c	x_c	
front curvature radius	c1	[0, s/2]
back curvature radius	c2	[0, s/2]

Details of optimization variables in Eq. (6) can be seen in Table 4.

Optimization is conducted several times to verify results. Results converged to identical solutions for every cross section separately. Particle swarm optimization is used where for all three cross sections optimization variables converged in their upper limits. For case (a) optimization generated a cross section with maximum front and back inclinations which generated rhombus shaped cross section. For case (b) front edge has maximum curvature with maximum inclination in the back, which generated streamlined shaped bar and for case (c) optimization generated cross section with front and back edges with maximum curvature. These results are expected since all considered profiles converged in cross section with minimal cross section area; they generated the smallest head loss which validated this optimization process. Initial and optimized cross sections with indicated optimization parameters can be seen in Fig. 7b.

Sharp edges in cross sections must be carefully considered due to production, exploitation and safety reasons. During the process of trash rack cleaning considerable forces can be induced on bars, especially when removing debris stuck between bars, which If trash rack cleaning system is in direct contact with the trash-rack, forces induced during interaction can cause structural damage. Also, depending on the hydroelectric power plant location, different intakes are subjected to different type of debris. Considered HE Senj mainly deals with smaller debris (weed or branches) so rhombus shaped bars that have thinner front edge can be considered for

installation. However, if greater debris (logs) is frequently present at intake it can cause damage to construction if that type of cross section is chosen. Also, depending on configuration, sharp edges can cause fish injuries when interacting with trash-rack. This problem is also present with rectangular cross section, where for some bar inclinations (e.g. 90°) rhombus shape provides a safer solution. Considering these problems are problem specific, edge thickness constraints must be defined in accordance.

For different intake geometries, shape optimization of the bar cross section can be conducted to provide the optimal solution. Hence, cross sections that are usually not used, can be derived as an optimal result for specific intake (e.g. innovative design considered in [24]). In this study, only parameters defining cross section are included as optimization parameters, however, other geometry parameters such as bar spacing, bar length, bar inclination etc. can also be included. Cross section optimization for HE Senj was done to reduce the losses without changing the bar spacing (which was proved to be valid during exploitation). As it was mentioned in [19,20] head loss reduction due to cross section change enables reduction in bar spacing, but that must be carefully evaluated due to its influence on other criteria such as structural aspect, debris accumulation, vibrations and velocity filed that can influence the fish movement. With new innovative designs, such as V-shaped trash-rack [23] optimization value becomes more prominent because it can reduce the time necessary for conducting experiments that vary different geometry parameters. Also, since vortex shedding that influences vibrations and can cause damages to trash-rack structure is known for standard trash-rack design, when considering new innovative designs this aspect must also be taken into consideration. More detailed numerical analysis (LES) of unsteady fluid behaviour should be conducted [16] with encompassing structural (FEM) numerical analysis [17].

Conclusion

In this study, the influence of trash-rack and bar geometry on head losses is examined. Validation of numerical results is conducted with experimental results from previous studies. A numerical investigation of four bar cross sections for nine different trash-rack configurations, where trash-rack and bar inclinations are varied, is performed. Additionally, optimization of trash-rack bar cross section is conducted using the PSO algorithm.

For a given experiment, where the channel cross section is constant along the vertical axis, similar results are obtained with 3D multiphase, 3D single phase and 2D simulations. Since difference in 2D and 3D results were around 1%, 2D simulations are conducted for all considered cases to save computational time. For greatest bar (90°) and trash-rack (45°) inclinations greatest variation in the result is observed with numerical simulation underestimating head loss coefficient by 15%. Rectangular cross section, which is mainly present in turbine intakes, causes the greatest head loss for almost all configurations which suggests there is an area for improvement in current designs. For greater bar and trash-rack inclinations greater turbulence zones can be observed which cause greater head loss coefficient. Also, in case of low-head turbines where the turbine is positioned rather close to the trash-rack, the non-uniformity of flow may cause a reduction of turbine efficiency. For these configurations influence of cross section is greater than for configurations with smaller inclinations. Optimization conducted for trash-rack perpendicular and bars parallel to fluid flow, generated geometry with minimal bar cross section area.

In future work, possibilities of optimization should be explored and validated with the experiment. Optimization can be conducted for real intake geometries where the influence of channel before and after trash-rack should also be also included. To decide on the optimal cross section, apart from head losses, other flow field parameters which influence the fish behaviour near trash-rack can be included in the optimization goal function to encompass both ecological and engineering approach. Construction and stability aspect must also be taken into consideration, where constraints or penalties for designs that induce vibrations that could lead to construction failure should be included. Currently this optimization procedure would include expensive goal function evaluation since it would include both LES simulation and structural (FEM) numerical analysis, but with growing computational power it would provide comprehensive study of trash-rack design.

Declaration of Competing Interest

The authors have declared no conflict of interest.

Compliance with Ethics Requirements

This article does not contain any studies with human or animal subjects.

References

[1] Coutant CC, Whitney RR. Fish behavior in relation to passage through hydropower turbines: a review. Trans Am Fish Soc 2000;129(2):351–80.
[2] Amaral SV, Coleman BS, Rackovan JL, Withers K, Mater B. Survival of fish passing downstream at a small hydropower facility. Mar Freshwater Res 2018;69(12):1870–81.
[3] Inglis ML, McCoy GL, Robson M. Testing the effectiveness of fish screens for hydropower intakes. Report Project SC120079. <https://www.gov.uk/government/publications/testing-the-effectiveness-of-fish-screens-for-hydropower-intakes>; 2016. [Accessed: 16-July-2019].
[4] Szabo-Meszaros M et al. Experimental hydraulics on fish-friendly trash-racks: an ecological approach. Ecol Eng 2018;113:11–20.
[5] Katopodis C. Developing a toolkit for fish passage, ecological flow management and fish habitat works. J Hydraul Res 2005;43(5):451–67.
[6] Castro-Santos T, Haro A. Fish guidance and passage at barriers. In: Domenici P, Kapoor B, editors. Fish locomotion: an eco-ethological perspective. Enfield, NH: Science Publishers; 2010. p. 62–89.
[7] Bunt C, Castro-Santos T, Haro A. Performance of fish passage structures at upstream barriers to migration. River Res Appl 2012;28(4):457–78.
[8] Fjeldstad HP, Pulg U, Forseth T. Safe two-way migration for salmonids and eel past hydropower structures in Europe: a review and recommendations for best-practice solutions. Mar Freshwater Res 2018;69(12):1834–47.
[9] Silva AT et al. The future of fish passage science, engineering, and practice. Fish Fish 2018;19(2):340–62.
[10] Idelchik IE. Handbook of hydraulic resistance. Washington, DC: Hemisphere Publishing Corporation; 1986.
[11] of Engineers USAC. Corps of Engineers Hydraulic Design Criteria. v. 1; Waterways Experiment Station Vicksburg, MS; 1977.
[12] Tsikata JM, Katopodis C, Tachie MF. Experimental study of turbulent flow near model trashracks. J Hydraul Res 2009;47(2):275–80.
[13] Tsikata JM, Tachie MF, Katopodis C. Open-channel turbulent flow through bar racks. J Hydraul Res 2014;52(5):630–43.
[14] Naudascher E, Rockwell D. Flow-induced vibrations: an engineering guide. Rotterdam, Netherlands: A.A. Balkema; 1994.
[15] Nascimento L, Silva J, Di Giunta V. Damage of hydroelectric power plant trash-racks due to fluid-dynamic exciting frequencies. Latin Am J Solids Struct 2006;3(3):223–43.
[16] Nakayama A, Hisasue N. Large eddy simulation of vortex flow in intake channel of hydropower facility. J Hydraul Res 2010;48(4):415–27.
[17] Huang X, Valero C, Egusquiza E, Presas A, Guardo A. Numerical and experimental analysis of the dynamic response of large submerged trash-racks. Comput Fluids 2013;71:54–64.
[18] Clark SP, Tsikata JM, Haresign M. Experimental study of energy loss through submerged trashracks. J Hydraul Res 2010;48(1):113–8.
[19] Raynal S, Courret D, Chatellier L, Larinier M, David L. An experimental study on fish-friendly trashracks–part 1. Inclined trashracks. J Hydraul Res 2013;51(1):56–66.
[20] Raynal S, Chatellier L, Courret D, Larinier M, David L. An experimental study on fish-friendly trashracks–part 2. Angled trashracks. J Hydraul Res 2013;51(1):67–75.
[21] Albayrak I, Kriewitz CR, Hager WH, Boes RM. An experimental investigation on louvres and angled bar racks. J Hydraul Res 2018;56(1):59–75.
[22] Zayed M, El Molla A, Sallah M. An experimental study on angled trash screen in open channels. Alexandr Eng J 2018;57(4):3067–74.

[23] Zayed M, El Molla A, Sallah M. An experimental investigation of head loss through a triangular "v-shaped" screen. J Adv Res 2018;10:69–76.

[24] Beck C, Albayrak I, Boes RM. Improved hydraulic performance of fish guidance structures with innovative bar design. In: 12th International symposium on ecohydraulics (ISE 2018).

[25] Böttcher H, Gabl R, Aufleger M. Experimental hydraulic investigation of angled fish protection systems-comparison of circular bars and cables. Water 2019;11 (5).

[26] Raynal S, Chatellier L, David L, Courret D, Larinier M. Numerical simulations of fish-friendly angled trashracks at model and real scale. In: 35th IAHR World Congress.

[27] Paul SS, Adaramola MS. Analysis of turbulent flow past bar-racks. In: ASME international mechanical engineering congress and exposition; 2014.

[28] Åkerstedt HO, Eller S, Lundström TS. Numerical investigation of turbulent flow through rectangular and biconvex shaped trash racks. Engineering 2017;9 (05):412.

[29] Khan LA, Wicklein EA, Rashid M, Ebner LL, Richards NA. Computational fluid dynamics modeling of turbine intake hydraulics at a hydropower plant. J Hydraul Res 2004;42(1):61–9.

[30] Čarija Z, Lučin I, Lučin B, Grbčić L. Investigation of numerical simulation parameters on fluid flow around trash-racks. In: Proceedings of the 29th DAAAM international symposium; vol. 29; 2018. p. 1046–52.

[31] Kennedy J, Eberhart R. Particle swarm optimization. In: Proceedings of ICNN'95 - international conference on neural networks, vol. 4; 1995. p. 1942–8.

Source Contamination Detection Using Novel Search Space Reduction Coupled with Optimization Technique

Ivana Lučin[1], Luka Grbčić[2], Siniša Družeta[3], and Zoran Čarija[4]

[1]Ph.D. Student, Dept. of Fluid Mechanics and Computational Engineering, Faculty of Engineering, University of Rijeka, Vukovarska 58, 51000 Rijeka, Croatia and Center for Advanced Computing and Modelling, University of Rijeka, Radmile Matejčić 2, 51000 Rijeka, Croatia (corresponding author). Email: ilucin@riteh.hr

[2]Ph.D. Student, Dept. of Fluid Mechanics and Computational Engineering, Faculty of Engineering, University of Rijeka, Vukovarska 58, 51000 Rijeka, Croatia and Center for Advanced Computing and Modelling, University of Rijeka, Radmile Matejčić 2, 51000 Rijeka, Croatia

[3]Associate Professor, Dept. of Fluid Mechanics and Computational Engineering, Faculty of Engineering, University of Rijeka, Vukovarska 58, 51000 Rijeka, Croatia and Center for Advanced Computing and Modelling, University of Rijeka, Radmile Matejčić 2, 51000 Rijeka, Croatia

[4]Professor, Dept. of Fluid Mechanics and Computational Engineering, Faculty of Engineering, University of Rijeka, Vukovarska 58, 51000 Rijeka, Croatia and Center for Advanced Computing and Modelling, University of Rijeka, Radmile Matejčić 2, 51000 Rijeka, Croatia

ABSTRACT

Contaminant intrusion in a water distribution network is an important concern since it can have hazardous consequences for the population. Reacting immediately is crucial to prevent or reduce the further propagation of contamination. In terms of contamination scenario characteristics, optimization is extensively researched as a valuable methodology to provide information. In this work, a procedure preceding the optimization which considerably reduces the search space for a potential contaminant source location is presented. For each suspect node, a simulation is conducted

with unrealistically high contaminant concentration injected throughout the whole simulation. If the sensors do not register contamination in a later scenario, then that node can be eliminated as a possible contaminant source. The methodology is applicable for both single and multiple contaminant injection nodes. This approach is investigated on multiple benchmark networks and different sensors placements, as given in the literature. By coupling the proposed search space reduction method with an optimization approach, a novel efficient methodology for contamination source detection is presented.

INTRODUCTION

In the event of a contaminant intrusion in a water distribution network, that can quickly affect a great number of network users, restoring water quality is a main concern. In order to give a fast response, it is needed to rapidly identify the contaminant injection location; thus, a number of studies have investigated emergency reaction and effects on human health in contamination events (Davis et al. 2014; Rasekh and Brumbelow 2015; Shafiee and Berglund 2017; Shafiee et al. 2018; Strickling et al. 2020). Due to a limited number of sensors in real water distribution networks, optimal sensor placement is particularly relevant in the fast and accurate detection of the pollution event parameters. Therefore, different optimization approaches and algorithms have been extensively investigated (Ostfeld et al. 2008; Hart and Murray 2010; Mukherjee et al. 2017; Zhao et al. 2016; Palleti et al. 2016; Ung et al. 2017).

In order to determine the number of network users which are potentially affected by a contamination event in a water distribution network, the following needs to be known: the source node of contamination, the starting time of the injection, the duration of the injection and the injection concentration value. A number of different approaches for the identification of contamination scenario characteristics have been investigated, such as data mining (Huang and McBean 2009), backtracking method (Laird et al. 2005; De Sanctis et al. 2009), Bayesian approach (Yang and Boccelli 2014) and simulation-optimization approach (Preis and Ostfeld 2007; Zechman and Ranjithan 2009). When considering an optimization approach, the discrepancy in sensor measurements must be taken into consideration (Preis and Ostfeld 2008), as well as the uncertainty of water demand

(Vankayala et al. 2009; Xuesong et al. 2017; Yan et al. 2019) since both factors have an impact on the efficiency of the optimization method. Since source identification is an inverse problem, which are generally considered to be ill-posed, multiple solutions can yield similar results i.e. different contamination sources can cause similar contamination measurements in sensors. This is especially the case when sensor measurement and network demand uncertainties are considered. Consequently, the source identification problem can be considered as a non-unique, multimodal problem. This was investigated by using a niching algorithm which provided multiple potential optimization solutions (Hu et al. 2015; Yan et al. 2017). A detailed overview of optimization methods for identification of contamination source was presented in Adedoja et al. (2018).

When considering larger networks, the complexity of the proposed problem increases and different methods for reducing complexity of the problem have been introduced. Aggregation (Qin and Boccelli 2017) and network skeletonization (Klise et al. 2013) approaches reduced the water distribution network model by simplifying it, albeit while deteriorating the accuracy of the water quality simulations. De Sanctis et al. (2009) investigated a reduction of suspect nodes using a backtracking method extended with contaminant status algorithm, where possible injection nodes and injection times pairs were identified by iterating over all sensor measurements. However, in the said method, the assumption of ideal sensors was made which could cause the false identification of suspect nodes. Further examination of the proposed method was conducted in Seth et al. (2016), where the efficiency of three different approaches was investigated for source identification; Bayesian-probability, optimization approach and contaminant status algorithm strategies. Liu et al. (2012b) and (2012a) used logistic regression analysis to identify potential source locations in order to narrow down the search space for the optimization methods which were then used to identify other contamination scenario characteristics. Sankary and Ostfeld (2019) investigated Bayesian localization of contamination intrusion using mobile sensor data.

The authors of this study observed that during the optimization process, in which an algorithm searches for optimal contamination scenario parameters, in a considerable number of simulations there is no detected contamination whatsoever. Therefore, a new search space reduction method,

reducing the number of suspect nodes for the optimization process, is proposed. The said search space reduction is achieved by simulating an extreme contamination scenario for each water network node which is a potential source, with an unrealistically large contamination amount constantly being injected at the analyzed node. This can be considered as an upper bound of a contamination scenario simulation which was chosen to assure that the true source node is not eliminated. If the sensors do not detect contamination for such an extreme scenario, this node can be excluded from the list of suspect nodes. Contrary to the backtracking method, the method proposed in this work uses a forward simulation model with a simple sensor measurement comparison. Due to the fast execution of the water network flow simulations, the proposed forward approach can easily be conducted even on large scale networks.

Since the described pre-optimization procedure only reduces the number of suspect nodes, it needs to be coupled with an optimization algorithm to identify the contamination scenario parameters (injection time, duration, and concentration) and the true contamination source node. In this study, the proposed search space reduction method is coupled with an optimization approach in which an independent optimization is conducted for each remaining suspect node. A similar approach can be observed in Liu et al. (2012b) where local search methods (coupled with logistic regression) were conducted for each remaining suspect node. The independent optimizations reduce the number of optimization variables, in turn reducing the complexity of the given problem. The optimization run with the smallest fitness value identifies the optimal source node with corresponding pollution event parameters. A major issue with the standard optimization approach is the source node parameter being a categorical variable and which makes the optimization task a combinatorial optimization problem, that is a very difficult problem to solve. Also, due to the multimodal nature of the problem, the optimization is highly dependent on initial conditions, as it can easily converge to local optima without ever discovering the global optimum. The proposed optimization approach successfully avoids the problem of obtaining only local optima solutions by providing both the optimal and approximate solutions.

This paper is organised in the following manner: considered contamination scenarios and

investigated benchmark water distribution networks are presented, followed by a description of the pre-optimization procedure and the used optimization approach in which Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) are independently employed for each remaining suspect node. In the results section, the search-space reductions are presented for all networks, contamination scenarios and sensor placement layouts, in order to inspect the robustness of the proposed method. Also, the results of the optimization approach for the chosen contamination scenarios are presented. In the discussion section, the benefits and limitations of the proposed method are examined.

MATERIALS AND METHODS

Test cases

Hypothetical contamination scenarios are simulated in EPANET2 (Rossman et al. 2000) version 2.0.12 for small, medium and large sized benchmark water distribution networks. All networks are obtained from The Centre for Water Systems (CWS) at the University of Exeter (Centre for Water Systems), with the exception of EPANET2 example network Net3. All simulation parameters (duration, hydraulic, quality and pattern time step) are kept at default values, with the exception of the pattern time step which is set to 10 minutes for the small and medium sized networks. The contamination starting time, duration of injection and concentration value are chosen arbitrarily and are equal for every scenario within the considered network. The EPANET2 flow paced method is used for the contaminant injection for all conducted simulations. In all simulations, a single injection node and a constant value of concentration is defined. The basic assumptions are that the considered network demand multipliers are calibrated and the sensors are considered ideal. Different sensor placement is taken into consideration for each pipe network and presented in Table 1 with the corresponding literature. In order to investigate the robustness of the proposed method, for each considered network, the contamination scenarios are simulated for each network node, e.g. for a network consisting of 100 nodes, a search space reduction for 100 contamination scenarios is investigated.

Small network

For the small sized network, hypothetical network Anytown, USA consisting of 19 nodes, proposed by Walski et al. (1987), is used (Fig 2). The simulation time is 24 h with a hydraulic time step of 1 min, a quality time step of 1 min and a pattern time step of 10 min. Two different sensor layouts are examined. In the first layout, sensors are placed in network nodes 70 and 160, while in the second one, in nodes 90, 110 and 140 (Table 1). The contaminant injection starts at $t = 5$ h, with a duration of 120 min with a constant injection concentration of 500 mg/L.

Medium networks

The first medium sized network used is the EPANET2 example Net3 consisting of 92 nodes (Fig 3). The simulation time is 24 h with a hydraulic time step of 10 min, a quality time step of 5 min and a pattern time step of 10 min. The investigation of four different sensor layouts is performed (Table 1). For all considered cases, the contamination injection starts at $t = 5$ h with a duration of 100 min and with a constant injection concentration of 300 mg/L.

The second medium sized network analyzed is the Richmond network (Fig. 4a) with 865 nodes and two different sensor layouts (Table 1). The simulation time is 72 h with a hydraulic time step of 1 h, a quality time step of 5 min and a pattern time step of 1 h. For all considered cases, the contaminant injection starts at $t = 5$ h with a duration of 5 h and with a constant injection concentration of 500 mg/L.

The third medium sized network considered is the BWSN Network 1 (Ostfeld et al. 2008) (Fig. 4b) with 126 nodes. The simulation time is 96 h with a hydraulic time step of 30 min, a quality time step of 5 min and a pattern time step of 30 min. Two different sensor layouts are examined (Table 1). For all considered cases, the contaminant injection starts at $t = 5$ h with a duration of 5 h and with a constant injection concentration of 300 mg/L.

Large network

For a large network, BWSN Network 2 consisting of 12523 nodes is considered (Fig 4c). Due to the significant number of nodes, only the sensor layout by Wu and Walski (2008) is considered because of its greatest detection likelihood (Table 1). Originally, the simulation time is set to 48 h,

but for the purpose of this research, it was set to 24 h with a hydraulic time step of 1 h, a quality time step of 5 min and a pattern time step of 1 h. For all considered cases, the contaminant injection starts at $t = 5$ h with a duration of 5 h and with a constant injection concentration of 500 mg/L.

Search space reduction method

For each network, a group of extreme contamination scenarios is conducted. Extreme contamination scenario simulations are conducted for each node independently, with the contamination injection starting at the beginning of the simulation and lasting throughout the whole simulation. Concentration is kept constant with an unrealistic value of contamination (1 kg/L). The idea is to test whether the contamination originating from the examined node will ever be registered in the used pollution detection sensors. If the relevant sensors are unable to detect contamination for this extreme scenario, it is safe to assume that this node is not a possible contaminant source and does not need to be considered as one in the optimization process. The search space reduction method is conducted with a Python script which executes the EPANET2 extreme scenario simulation for each node, collects the sensor measurements from the report file and compares them with the sensor measurements from the contamination event (in this case the hypothetical contamination scenarios which were previously simulated). The flowchart of the proposed method can be seen in Fig. 1.

The elimination of the true source node can be done due to the demand uncertainty or sensor measurement imperfections. However, if the extreme scenarios are simulated on a calibrated hydraulic model, which is also to be used later in the optimization phase, then these uncertainties remain intrinsic to the entire optimization approach and as such cannot be avoided. In other words, even if the optimization would still take into account the suspect node which was eliminated by the search space reduction procedure, the sensors would always measure zeros when that node was chosen as a source node, regardless of the variation of other optimization parameters (injection time, duration, and concentration). Consequently, the optimization method would also eliminate that node as a possible solution, i.e. it would concentrate on exploring other suspect nodes (although with less efficiency).

In light of the above, two different methods of analysis of sensor detection in extreme scenario

₁₈₅ are used. In the conservative method, only those nodes that did not register contamination in the

₁₈₆ extreme scenario, i.e. for which all sensor measurements are zero, are removed. This assures

₁₈₇ that the elimination of the true source node cannot be done because of the search space reduction

₁₈₈ method. In the case of multiple contaminant sources, one node can influence detection in one sensor

₁₈₉ and another node can influence detection in a different sensor. Thus, in those scenarios only nodes

₁₉₀ for which sensors do not detect contamination at all can be safely eliminated, i.e. the conservative

₁₉₁ search space reduction method is to be used. The non-conservative method is based on matching

₁₉₂ sensor detection, i.e. where it is required that the sensors which jointly detected pollution in a real

₁₉₃ pollution event must also jointly detect pollution in the extreme scenario. This method assumes a

₁₉₄ single injection location and enables greater search space reduction. However, in this method, the

₁₉₅ sensor measurement imperfections must be taken into consideration.

Optimization procedure

₁₉₇ For the optimization phase of the source contamination detection, an optimization approach

₁₉₈ which removes the source node variable from the optimization process is used. In order to do

₁₉₉ so, the assumption of a single contaminant source node is adopted. In this approach, a separate

₂₀₀ optimization process is independently conducted for each network node remaining after the search

₂₀₁ space reduction. The independent optimization processes are conducted with the contamination

₂₀₂ starting time, duration and concentration as the optimization variables. In this manner, the com-

₂₀₃ plexity of the given problem decreases by reducing the optimization problem from four mixed

₂₀₄ (categorical/continuous) variables to three continuous variables.

₂₀₅ A great number of studies of the optimization-based contamination detection used Genetic

₂₀₆ Algorithm (GA) and its variants, thus the GA implementation in the Python optimization package

₂₀₇ DEAP (2012) is used. Since in the proposed approach, the search space is reduced to continuous

₂₀₈ variables, Particle Swarm Optimization (PSO) algorithm, as implemented in Python optimization

₂₀₉ package inspyred (2017), is also applied to this problem. The goal function is defined as cumulative

₂₁₀ squared difference in measurement for each sensor and each time step:

$$f = \sum_{i=1}^{n} \sum_{t=1}^{T} (c_i^o(t) - c_i^s(t))^2 \qquad (1)$$

where n denotes the number of sensors, T represents the simulation duration with discrete time step t, c^o represents the observed concentrations in real contamination event and c^s represents the concentrations obtained by the simulation. Since sensors are considered ideal, for the exact solution the fitness function must yield zero.

The proposed procedure is applied to a contamination scenario from Net3 network with four sensors as given in Preis and Ostfeld (2007). The contamination scenario is defined as it was explained in subsection 2, with node 10 being the contaminant source. For BWSN Network 2, the contamination scenario with injection node 12500 is considered with other parameters being the same as given in subsection 2. For both networks, the starting time and duration variables range from 0 h to 24 h and the concentration value ranges from 10 to 2000 mg/L. The overview of the optimization parameters is presented in Table 2.

RESULTS

In this section, the search space reduction results are presented for the small, medium and large sized water distribution networks with different sensor placements to investigate the efficiency of the proposed method. A detailed presentation of the search space reduction process is shown for the small sized network. Due to a great number of network nodes in the medium and large sized networks, i.e. a great number of investigated contamination scenarios, only a summary of the node reductions is presented for those networks.

Investigation of injection time and location for small network

In order to enable the definition of the appropriate test-case contamination scenarios, an investigation of the injection time and location is conducted for Anytown, USA network for scenarios with an injection starting time at 5 h (Scenario A), 10 h (Scenario B) and 20 h (Scenario C). The sensors are placed in nodes 90, 110 and 140. Table 3 shows chosen contamination scenario conditions and simulation results with information on the ability of sensors to detect the contamination. The

scenarios in which none of the sensors reported contamination are omitted from the table since the awareness of contamination event is non existent. The changes in sensors detection of contamination for scenarios B and C, comparing with scenario A, are given in bold typeface. Results show that fluid flow in the pipe network greatly influences the sensors' registration of contaminant. For example, the contamination detection in sensors can change just by varying the injection time, even though the injection duration and concentration value are kept constant. As an illustration, when contamination is injected in node 20 at 5 h, all three sensors detect contamination. If, for the same injection node, the starting time is changed to 20 h, the sensor in node 140 does not report contamination anymore.

When the injection node changes, different detection in sensors can be observed as well. For example, for the scenario A when the injection node is node 20, all three sensors detect contamination, while for the simulation with the injection at node 140 only one sensor detects contamination. This shows that the investigation of different injection nodes is sufficient to display the benefit of the proposed method for different scenario conditions. Therefore, the injection starting time, duration and concentration are equal for all scenarios within the investigations of the same network.

Search space reduction for small network

Injection nodes 20 and 30 are chosen for a detailed presentation of search space reduction for scenario A. Extreme scenarios are conducted for all network nodes and sensor detection is presented in Table 4. For the extreme scenarios with contaminant injected in nodes 10, 65, 120, 130, 165 and 170 none of the sensors register contamination and those nodes can be removed from the optimization process when the conservative method is applied (underlined in Table 4). If the non-conservative method is used, a positive detection of contamination in scenario A must be matched with a positive sensor detection in the extreme scenario. For injection node 20, all sensors report contamination, hence all sensors must also report contamination in the extreme scenario. That is valid only for the extreme scenarios with injection nodes 20 and 110 (given in bold typeface) and these nodes should be further considered for the optimization. All other nodes can be excluded

from the optimization search space since at least one sensor detection does not match detection in the contamination event (marked with star).

In Table 5, an overview of nodes that can be eliminated for scenarios with different source nodes is presented with the indicated reductions of suspect nodes. At least 30% of suspect nodes can be eliminated for all considered cases, i.e. no matter where the real source is located. This can be observed for multiple injections scenario (conservative method), where the same removed nodes are listed as in case of single injection scenario (non-conservative method) for nodes 140, 150 and 160. However, reduction can be even around 90% for some contamination scenarios, i.e. injection locations. The achieved range of suspect nodes reductions indicates that, depending on the real contamination event, the optimization process can greatly benefit from the proposed method. Even the smallest reduction obtained (32% in this case) is still significantly reducing the number of suspect contamination nodes. It must be noted that these results are true only for the considered network, sensor layout and considered scenario (injection time, duration and concentration value). A different contamination scenario can yield different reduction ranges, hence presented results serve only as an illustration of possible benefits of the proposed method.

Influence of sensors layout

The efficiency of the proposed source node reduction method depends also on the number of sensors and their placement in a network. This is investigated for all networks presented in section 2. Since the number of investigated contamination scenarios is equal to the number of network nodes for all considered networks and sensor layouts, a wide range of different search space reductions is obtained. For that reason, only the greatest and smallest search space reduction is presented for each sensor layout. A single injection scenario result (yielding the greatest achieved reduction of suspect nodes) and multiple injection scenario result (yielding the smallest reduction obtained using conservative method) are presented in the Table 6. For all other scenarios with different injection nodes, the search space reduction efficiency is somewhere in a range between the presented smallest and greatest search space reduction values.

The analysis displays that a wide range of reductions can be observed, but a considerable

reduction of search space is present for all cases. For example, for Net3 with five sensors, the search space reduction ranges from 23% to 83%. However, for the sensor layout with two sensors, these reductions increase and are in the range of 60% to 80%. In the case of multiple sources, with an increase of the number of sensors, a smaller reduction of search space can be observed (Anytown, Net3, BWSN Network1), as expected. With a greater number of sensors, the likelihood that at least one sensor will detect contamination increases and there is a smaller number of nodes for which contamination event in extreme scenario is not detected. For the majority of the considered networks and sensor placements, at least around 25% of suspect nodes can be eliminated, the only exception being the BWSN Network 1 with 13% reduction. However, the sensor placement in question is in fact the optimal sensor layout with the maximum detection likelihood, as reported in Ostfeld et al. (2008) so it is only reasonable that most of the time at least one sensor will detect pollution in extreme scenarios. For the largest investigated network, BWSN Network 2, the minimum reduction of 64% is achieved which removes 8036 out of 12523 suspect nodes. For some scenarios, the reduction efficiency is much higher than the minimum of 64%. For instance, in the case of the injection node 12500, the number of remaining suspect nodes was reduced to 23.

Although a great number of simulations is needed, the overall computation time of the proposed search space reduction is fairly short. The results presented in this paper are obtained by utilizing the HPC resources of the Center for Advanced Computing and Modelling at the University of Rijeka, specifically one INTEL E7 fat node. For the Net3 network, 4 cores are used for conducting proposed search space reduction method which took several seconds to execute. For the BWSN Network 2 with 24 cores used, the reduced suspect nodes are obtained within 12 minutes, while with 48 cores it takes 7 minutes for the nodes to be obtained.

Optimization results

The optimization procedure is conducted for one Net3 and one BWSN Network 2 contamination scenario. For the Net3 network scenario with injection node 10, 18 suspect nodes remain (the number of suspect nodes is reduced by 80%), namely 18 independent optimizations are conducted to determine the starting time, duration and concentration of the contamination injection. The

graphical overview of suspect nodes is presented in Fig 5 where it can be observed that the proposed search space reduction method localizes region where contamination occurs. All 18 independent optimizations are run in parallel, where each optimization is conducted on one processor core of $INTEL\ E7$ fat node, totally employing 18 processor cores. One independent optimization lasts 9 minutes, and due to parallel optimization execution, the solutions are obtained in the same time. If a smaller number of processor cores are available, e.g. 6, optimizations would need to be run sequentially, resulting in 18 minutes required to obtain the solution. To investigate and compare success rates of GA and PSO optimizations, a total of 50 optimization runs are repeated.

For the optimization with injection node 10 (which is a true contaminant source node), the GA optimization manages to find the exact solution in 25 runs out of 50, which is only a 50% success rate. To obtain the optimal solution, the GA needs 40 generations on average. PSO optimization is also used, where the optimal solution is obtained in 41 runs out of 50, which is a 82% success rate. To obtain the optimal solution PSO needs 35 iterations on average.

For the other suspect nodes, the minimum fitness from 50 runs is obtained and compared. It is observed that the contamination scenarios with the source node 101 and 105 are the second and third best solutions (marked in Fig 5). This is expected since those nodes are topologically near the real contaminant source node. It is interesting to note that node 103 (placed just below node 101) is also topologically near the source of contamination, but due to hydraulics of the system, in extreme scenario for that node contamination is not registered in the sensors at nodes 117 and 143 so it is not considered as a potential source node. The sensor measurements for the optimal scenarios with nodes 101 and 105 as source nodes are presented in Fig. 6. The sensor measurements for the optimal solutions of source node 10 and 101 show a very similar trend, thus indicating multimodal nature of the observed problem, while for source node 105, a greater difference is evident.

For the BWSN Network 2 contamination scenario with injection node 12500, 23 suspect nodes remain after pre-optimization reduction procedure, hence 23 independent optimizations are conducted in parallel, with 50 repeated PSO runs. For injection node 12500, which is the true injection location, the optimization manages to find the exact solution in all runs. To obtain the

optimal solution, 30 iterations are needed on average. Three other nodes provide a similar solution, all of them being topologically near the true source node (Fig 7). A single optimization run conducted on one processor core of $INTEL\ E7$ fat node takes 4 hours, but the optimum is on average obtained in 30 iterations, for which not more than half the time is needed. However, the simulation parameters greatly influence the optimization time, e.g. for a hydraulic time step of 10 minutes, the optimal solution is obtained on average in 35 generations, which lasts 5 hours. Therefore, the given computation times should only be understood as rough estimates, and further study should be made for investigating the computational efficiency on multiple processor cores and with fine-tuned optimization parameters.

DISCUSSION

The above presented tests show that the proposed search space reduction method can successfully be applied on different size networks with different sensor layouts. Since hydraulic simulation parameters vary for each network, and contamination scenario parameters are chosen arbitrarily, it is also shown that they do not influence the method's performance. Furthermore, for the largest investigated network consisting of 12523 nodes and 20 sensors, a considerable reduction is observed (ranging from 64% to 99%) indicating that this kind of realistic problems would benefit a lot from the proposed method, since most real water distribution networks have a great number of nodes and a sparse sensor placement.

The proposed method can require a considerable number of simulation runs, which is always equal to the number of network nodes. However, the proposed method quickly obtains results even for large scale networks due to the fast EPANET2 execution of hydraulic simulations. Parallel computation of extreme scenarios can be made, accelerating the gathering of required data for search space reduction and making the time needed for this process practically negligible even for the largest network. Extending the proposed method for parallel computation on multiple processors would be fairly straightforward, due to SIMD nature of the method's computational process.

In this study, the sensor measurements are considered ideal, thus the impact of false positive

and false negative sensor detection must be further explored. This problem is also present in the backtracking method and discussed in De Sanctis et al. (2009) and Seth et al. (2016). However, considering that the presented methodology serves only as a preparation for the optimization-based search procedure, using the conservative version of the suspect node reduction method (where only nodes for which in extreme scenario contamination is not detected are eliminated) these uncertainties exist only inasmuch as they are intrinsic to the entire optimization approach. However, as a trade-off for security that true source node is not wrongly eliminated, a smaller reduction of suspect nodes is obtained. Until the influence of uncertainties in the single injection approach is further investigated, the use of conservative approach should be encouraged, as it removes any uncertainty issues associated with the search space reduction method.

The examples of independently run optimizations for each suspect node display that the source identification problem has multiple solutions in terms of fitness, as was previously observed in the literature, and that this approach successfully provides those solutions. GA algorithm and its variants are usually used for this problem; however, it is observed that for problem formulation without source node as optimization variable PSO algorithm outperforms GA. The further exploration of PSO variants and optimization parameters should be conducted to improve the optimization procedure, i.e. decrease the number of needed EPANET simulations.

It is shown that, for medium sized network, the results can be obtained under 10 minutes when each independent optimization is conducted on a single CPU and optimizations are conducted in parallel (the number of employed processor cores is equal to the remaining number of suspect nodes). As the network size increases, several hours are needed to obtain results, implying that all independent optimizations are run in parallel with single processor core dedicated for each optimization. However, since the optimization procedure is specific for each network node, it can easily be distributed on multiple processors simultaneously, employing a greater number of processor cores per optimization and further accelerating the discovery of the optimal result. With the use of supercomputers and a simple parallel execution, a rapid response to a water network contamination problem would be guaranteed. The main drawback is that the reasonable number

of remaining suspect nodes is needed. This is especially the case when large scale networks are investigated since a limitation of computational resources could be reached. The coupling of the proposed search reduction with a probability based method using the reduced number of suspect nodes could be explored, which would further refine the list of suspect nodes, resulting in a reasonable number of processors needed for parallel utilization of the proposed approach.

CONCLUSION

In this work, the detection of source contamination is investigated by the use of suspect source nodes reduction as a preparation for the optimization approach detection process. The proposed search space reduction method is tested on five benchmark networks of varying size and with different sensor placement layouts that are taken from the literature. The results show that a considerable reduction in search space can be achieved, thus greatly reducing the complexity of the contamination source identification optimization problem. This is especially important considering the necessity of the rapid reaction time given the severity of the considered problem of water network contamination.

The main observations are:

- The proposed method is applicable for both single and multiple contaminant source scenarios, where a considerable reduction is present for both cases with a greater benefit in the case of single injection location.
- With an increase in network size, where sensor placement becomes more sparse, a greater reduction of search space is observed. This indicates that the proposed method can be highly beneficial when applied to a real water distribution network.
- For some contamination scenarios, a larger number of suspect nodes is eliminated and the source of pollution is localized. For these scenarios, a new optimization approach for single injection scenario is proposed where independent optimizations are conducted for every remaining suspect node as a contamination source.
- The proposed optimization procedure successfully obtains both the optimal and approximate

16 Lučin, December 14, 2021

solutions in a small number of optimization iterations.

Further study of this work should include the following:

- The influence of demand uncertainty and imperfect sensors on the efficiency of the proposed search space reduction method should be investigated.
- Enhancing the proposed methodology with probabilistic methods should be explored, as it could enable the detection of remaining suspect nodes which are more likely to be the sources of contamination.
- Different optimization algorithms and optimization parameters should be investigated in order to minimize the computational effort needed for reaching the optimal solution.

DATA AVAILABILITY STATEMENT

Some or all data, models, or code that support the findings of this study are available from the corresponding author upon reasonable request. (Available data: Python script for search space reduction method, Python script for PSO, Python script for GA, Epanet2 benchmark networks.)

Published version:

https://ascelibrary.org/doi/abs/10.1061/%28ASCE%29WR.1943-5452.0001308?af=R

REFERENCES

Adedoja, O., Hamam, Y., Khalaf, B., and Sadiku, R. (2018). "Towards development of an optimization model to identify contamination source in a water distribution network." *Water*, 10(5), 579.

Centre for Water Systems, U. o. E. "Benchmarks. Accessed: 6-November-2019.

Davis, M. J., Janke, R., and Magnuson, M. L. (2014). "A framework for estimating the adverse health effects of contamination events in water distribution systems and its application." *Risk Analysis*, 34(3), 498–513.

De Sanctis, A. E., Shang, F., and Uber, J. G. (2009). "Real-time identification of possible contamination sources using network backtracking methods." *Journal of Water Resources Planning and Management*, 136(4), 444–453.

Fortin, F.-A., De Rainville, F.-M., Gardner, M.-A., Parizeau, M., and Gagné, C. (2012). "DEAP: Evolutionary algorithms made easy." *Journal of Machine Learning Research*, 13, 2171–2175.

Garrett, A. (2017). "inspyred: Bio-inspired algorithms in python." *GitHub Repository*, <http://aarongarrett.github.io/inspyred/> (June).

Hart, W. E. and Murray, R. (2010). "Review of sensor placement strategies for contamination warning systems in drinking water distribution systems." *Journal of Water Resources Planning and Management*, 136(6), 611–619.

Hu, C., Zhao, J., Yan, X., Zeng, D., and Guo, S. (2015). "A mapreduce based parallel niche genetic algorithm for contaminant source identification in water distribution network." *Ad Hoc Networks*, 35, 116–126.

Huang, J. J. and McBean, E. A. (2009). "Data mining to identify contaminant event locations in water distribution systems." *Journal of Water Resources Planning and Management*, 135(6), 466–474.

Klise, K. A., Phillips, C. A., and Janke, R. J. (2013). "Two-tiered sensor placement for large water distribution network models." *Journal of Infrastructure Systems*, 19(4), 465–473.

Laird, C. D., Biegler, L. T., van Bloemen Waanders, B. G., and Bartlett, R. A. (2005). "Contamination source determination for water networks." *Journal of Water Resources Planning and Management*, 131(2), 125–134.

Liu, L., Zechman, E. M., Mahinthakumar, G., and Ranji Ranjithan, S. (2012a). "Identifying contaminant sources for water distribution systems using a hybrid method." *Civil Engineering and Environmental Systems*, 29(2), 123–136.

Liu, L., Zechman, E. M., Mahinthakumar, G., and Ranjithan, S. R. (2012b). "Coupling of logistic regression analysis and local search methods for characterization of water distribution system contaminant source." *Engineering Applications of Artificial Intelligence*, 25(2), 309–316.

Mukherjee, R., Diwekar, U. M., and Vaseashta, A. (2017). "Optimal sensor placement with mitigation strategy for water network systems under uncertainty." *Computers & Chemical Engineering*, 103, 91–102.

Ostfeld, A. and Salomons, E. (2004). "Optimal layout of early warning detection stations for water distribution systems security." *Journal of Water Resources Planning and Management*, 130(5), 377–385.

Ostfeld, A., Uber, J. G., Salomons, E., Berry, J. W., Hart, W. E., Phillips, C. A., Watson, J.-P., Dorini, G., Jonkergouw, P., Kapelan, Z., et al. (2008). "The battle of the water sensor networks (bwsn): A design challenge for engineers and algorithms." *Journal of Water Resources Planning and Management*, 134(6), 556–568.

Palleti, V. R., Narasimhan, S., Rengaswamy, R., Teja, R., and Bhallamudi, S. M. (2016). "Sensor network design for contaminant detection and identification in water distribution networks." *Computers & Chemical Engineering*, 87, 246–256.

Preis, A. and Ostfeld, A. (2007). "A contamination source identification model for water distribution system security." *Engineering optimization*, 39(8), 941–947.

Preis, A. and Ostfeld, A. (2008). "Genetic algorithm for contaminant source characterization using imperfect sensors." *Civil Engineering and Environmental Systems*, 25(1), 29–39.

Qin, T. and Boccelli, D. L. (2017). "Grouping water-demand nodes by similarity among flow paths in water-distribution systems." *Journal of Water Resources Planning and Management*, 143(8), 04017033.

Rasekh, A. and Brumbelow, K. (2015). "A dynamic simulation–optimization model for adaptive management of urban water distribution system contamination threats." *Applied Soft Computing*, 32, 59–71.

Rossman, L. A. et al. (2000). "Epanet 2: users manual.

Sankary, N. and Ostfeld, A. (2019). "Bayesian localization of water distribution system contamination intrusion events using inline mobile sensor data." *Journal of Water Resources Planning and Management*, 145(8), 04019029.

Seth, A., Klise, K. A., Siirola, J. D., Haxton, T., and Laird, C. D. (2016). "Testing contamination source identification methods for water distribution networks." *Journal of Water Resources Planning and Management*, 142(4), 04016001.

Shafiee, M. E. and Berglund, E. Z. (2017). "Complex adaptive systems framework to simulate the performance of hydrant flushing rules and broadcasts during a water distribution system contamination event." *Journal of Water Resources Planning and Management*, 143(4), 04017001.

Shafiee, M. E., Berglund, E. Z., and Lindell, M. K. (2018). "An agent-based modeling framework for assessing the public health protection of water advisories." *Water resources management*, 32(6), 2033–2059.

Strickling, H., DiCarlo, M. F., Shafiee, M. E., and Berglund, E. (2020). "Simulation of containment and wireless emergency alerts within targeted pressure zones for water contamination management." *Sustainable Cities and Society*, 52, 101820.

Ung, H., Piller, O., Gilbert, D., and Mortazavi, I. (2017). "Accurate and optimal sensor placement for source identification of water distribution networks." *Journal of Water Resources Planning and Management*, 143(8), 04017032.

Vankayala, P., Sankarasubramanian, A., Ranjithan, S. R., and Mahinthakumar, G. (2009). "Contaminant source identification in water distribution networks under conditions of demand uncertainty." *Environmental Forensics*, 10(3), 253–263.

Walski, T. M., Brill Jr, E. D., Gessler, J., Goulter, I. C., Jeppson, R. M., Lansey, K., Lee, H.-L., Liebman, J. C., Mays, L., Morgan, D. R., et al. (1987). "Battle of the network models: Epilogue." *Journal of Water Resources Planning and Management*, 113(2), 191–203.

Xuesong, Y., Jie, S., and Chengyu, H. (2017). "Research on contaminant sources identification of uncertainty water demand using genetic algorithm." *Cluster Computing*, 20(2), 1007–1016.

Yan, X., Zhao, J., Hu, C., and Zeng, D. (2017). "Multimodal optimization problem in contamination source determination of water supply networks." *Swarm and Evolutionary Computation*.

Yan, X., Zhu, Z., and Li, T. (2019). "Pollution source localization in an urban water supply network based on dynamic water demand." *Environmental Science and Pollution Research*, 26(18),

17901–17910.

Yang, X. and Boccelli, D. L. (2014). "Bayesian approach for real-time probabilistic contamina-
tion source identification." *Journal of Water Resources Planning and Management*, 140(8),
04014019.

Zechman, E. M. and Ranjithan, S. R. (2009). "Evolutionary computation-based methods for char-
acterizing contaminant sources in a water distribution system." *Journal of Water Resources
Planning and Management*, 135(5), 334–343.

Zhao, Y., Schwartz, R., Salomons, E., Ostfeld, A., and Poor, H. V. (2016). "New formulation and
optimization methods for water sensor placement." *Environmental Modelling & Software*, 76,
128 – 136.

TABLE 1. Overview of sensor layouts for investigated networks.

Network	Sensor placement	Reference
Anytown	70, 160	Ostfeld and Salomons (2004)
		Preis and Ostfeld (2007)
	90, 110, 140	Ostfeld and Salomons (2004)
Net3	117, 143, 181, 213	Preis and Ostfeld (2007)
	115, 119, 187, 209	Zechman and Ranjithan (2009)
	113, 120, 147, 211	Liu et al. (2012a)
	117, 149, 167, 213, 253	Seth et al. (2016)
	117, 173	Yan et al. (2017)
Richmond	123, 219, 305, 393, 589	Preis and Ostfeld (2007)
	93, 352, 428, 600, 672	Preis et al. (2008)
BWSN Network 1	10, 31, 45, 83, 118	Preis et al. (2008)
	10, 83	Yan et al. (2017)
BWSN Network 2	871, 1334, 2589, 3115, 3640	Wu and Walski (2008)
	3719, 4247, 4990, 5630, 6733	
	7442, 7714, 8387, 8394, 9778	
	10290, 10522, 10680, 11151, 11519	

TABLE 2. Optimization parameters for Net3 and BWSN network 2 optimization.

GA			
Generations	Population	Crossover rate	Mutation probability
100	100	0.95	0.1

PSO				
Generations	Population	Inertia	Cognitive rate	Social rate
100	100	0.75	1	1

TABLE 3. Sensors contamination detection for different injection locations of contamination and for different scenarios for Anytown network.

Injection node	Scenario A (starting time 5 h)			Scenario B (starting time 10 h)			Scenario C (starting time 20 h)		
	Contamination detection in sensors								
	S90	S110	S140	S90	S110	S140	S90	S110	S140
20	yes	yes	yes	yes	yes	yes	yes	yes	**no**
30	yes	no	yes	yes	no	yes	**no**	no	yes
40	yes	no	yes	yes	no	yes	**no**	no	yes
50	yes	no	yes	yes	no	yes	**no**	no	yes
60	yes	no	yes	yes	no	yes	yes	no	yes
70	yes	no	yes	yes	no	yes	yes	no	**no**
80	yes	no	yes	yes	no	yes	**no**	no	yes
90	yes	no	yes	yes	no	yes	yes	no	**no**
100	yes	no	yes	yes	no	yes	**no**	no	**no**
110	yes	yes	yes	**no**	yes	yes	**no**	yes	**no**
140	no	no	yes	no	no	yes	no	no	yes
150	no	no	yes	no	no	yes	no	no	yes
160	no	no	yes	no	no	yes	no	no	yes

TABLE 4. Comparison of sensor detection of contamination for scenario A injection in nodes 20 and 30 and for extreme scenarios for Anytown network.

	Scenario A Injection node 20			Scenario A Injection node 30		
	Sensor 90	Sensor 110	Sensor 140	Sensor 90	Sensor 110	Sensor 140
	yes	yes	yes	yes	no	yes

Injection node	Extreme scenario					
	Sensor 90	Sensor 110	Sensor 140	Sensor 90	Sensor 110	Sensor 140
10	no	no	no	no	no	no
20	**yes**	**yes**	**yes**	**yes**	**yes**	**yes**
30	yes	*no**	yes	**yes**	**no**	**yes**
40	yes	*no**	yes	**yes**	**no**	**yes**
50	yes	*no**	yes	**yes**	**no**	**yes**
60	yes	*no**	yes	**yes**	**no**	**yes**
65	no	no	no	no	no	no
70	yes	*no**	yes	**yes**	**no**	**yes**
80	yes	*no**	yes	**yes**	**no**	**yes**
90	yes	*no**	yes	**yes**	**no**	**yes**
100	yes	*no**	yes	**yes**	**no**	**yes**
110	**yes**	**yes**	**yes**	**yes**	**yes**	**yes**
120	no	no	no	no	no	no
130	no	no	no	no	no	no
140	*no**	*no**	yes	*no**	no	yes
150	*no**	*no**	yes	*no**	no	yes
160	*no**	*no**	yes	*no**	no	yes
165	no	no	no	no	no	no
170	no	no	no	no	no	no

TABLE 5. Node reductions for Anytown network for scenario A (contamination starting time at 5 h).

	Single source	
Injection node	Deleted nodes	Search space reduction
20	10 30 40 50 60 65 70 80 90 100 120 130 140 150 160 165 170	89%
30	10 65 120 130 140 150 160 165 170	47%
40	10 65 120 130 140 150 160 165 170	47%
50	10 65 120 130 140 150 160 165 170	47%
60	10 65 120 130 140 150 160 165 170	47%
70	10 65 120 130 140 150 160 165 170	47%
80	10 65 120 130 140 150 160 165 170	47%
90	10 65 120 130 140 150 160 165 170	47%
100	10 65 120 130 140 150 160 165 170	47%
110	10 30 40 50 60 65 70 80 90 100 120 130 140 150 160 165 170	89%
140	10 65 120 130 165 170	32%
150	10 65 120 130 165 170	32%
160	10 65 120 130 165 170	32%

	Multiple sources	
	10 65 120 130 165 170	32%

TABLE 6. Overview of node reductions for investigated benchmark networks and sensor layouts.

Network	Sensor placement	Single injection		Multiple injections
		Injection node	Reduction	Reduction
Anytown	70, 160	20	89%	74%
	90, 110, 140	20	89%	32%
Net3	117, 143, 181, 213	10	80%	25%
	115, 119, 187, 209	10	80%	38%
	113, 120, 147, 211	10	80%	32%
	117, 149, 167, 213, 253	10	83%	23%
	117, 173	10	80%	60%
Richmond	123, 219, 305, 393, 589	1	96%	82%
	93, 352, 428, 600, 672	1	94%	59%
BWSN Network 1	10, 31, 45, 83, 118	15	86%	13%
	10, 83	5	81%	60%
BWSN Network 2	871, 1334 ... 11519	87	99%	64%

Lučin, December 14, 2021

Fig. 1. Flowchart of proposed pre-optimization search space reduction method.
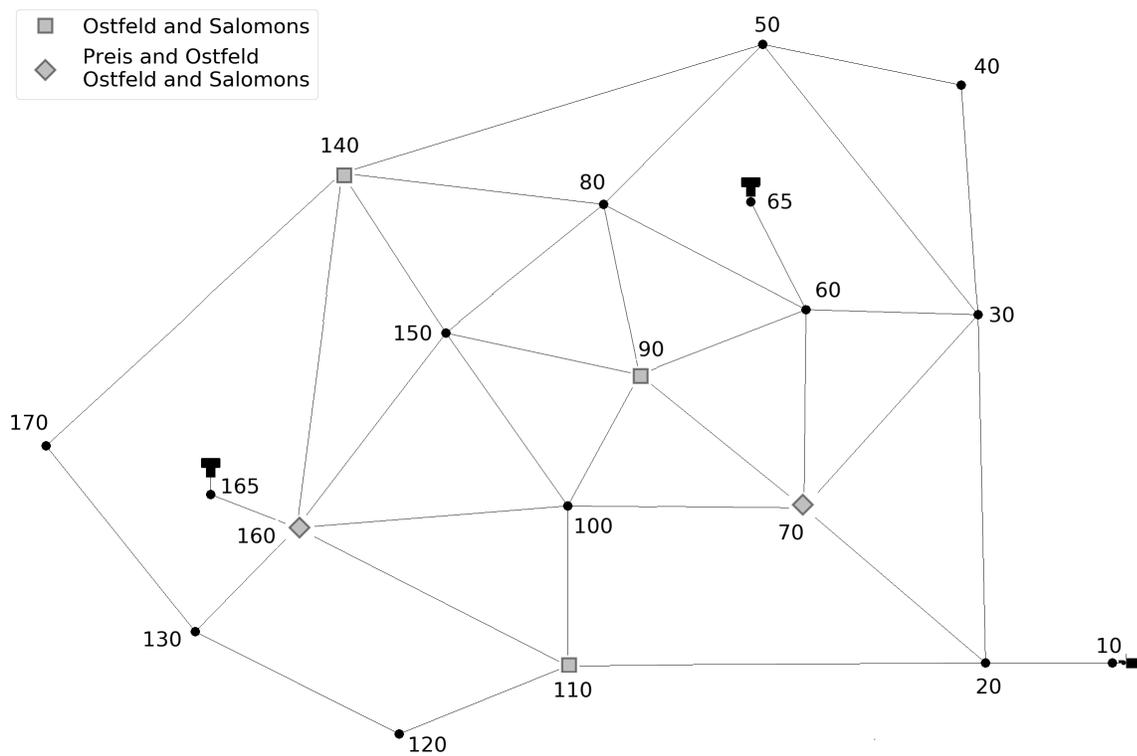
Fig. 2. Anytown network with indicated two considered sensor layouts.
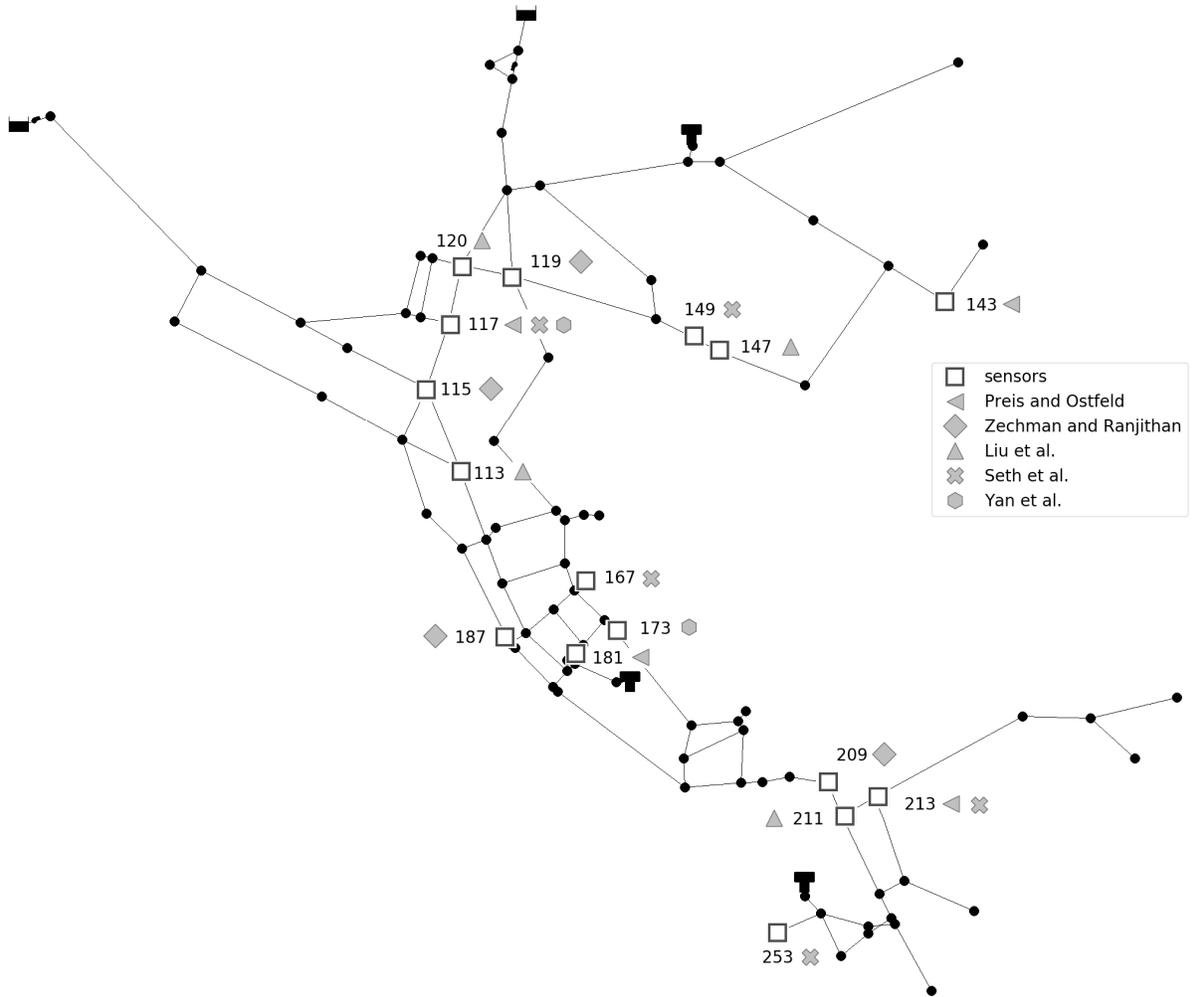
Fig. 3. Net3 EPANET example with investigated sensor layouts. Markers next to sensor numbers indicate literature which used those sensors.
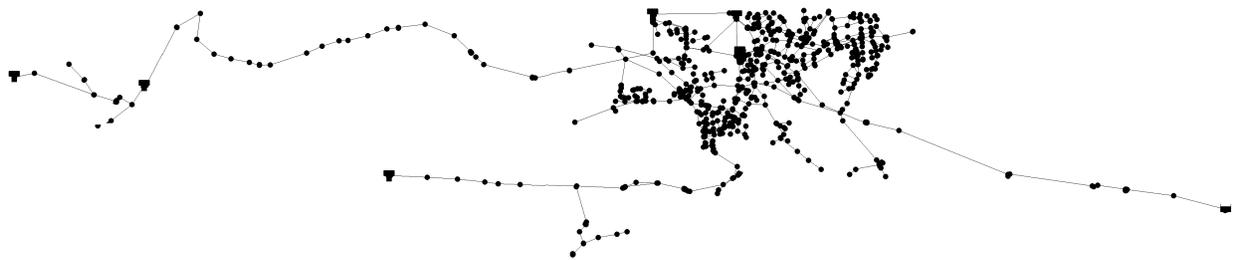
(a)
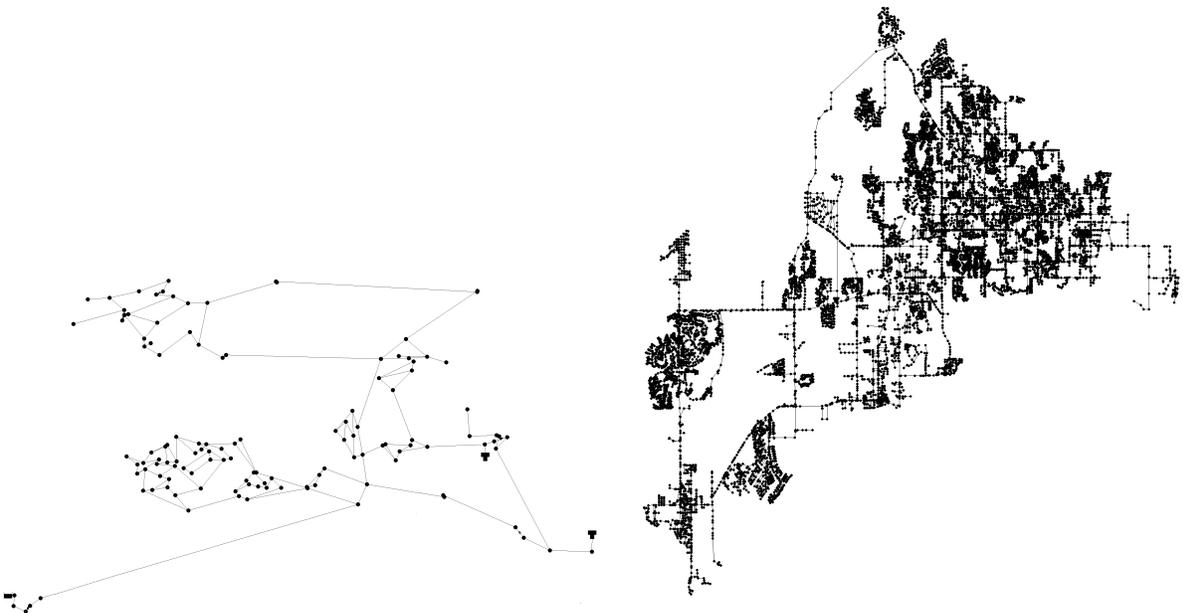
(b)

(c)

Fig. 4. Investigated networks (a) Richmond network, BWSN (b) small network and (c) large network.
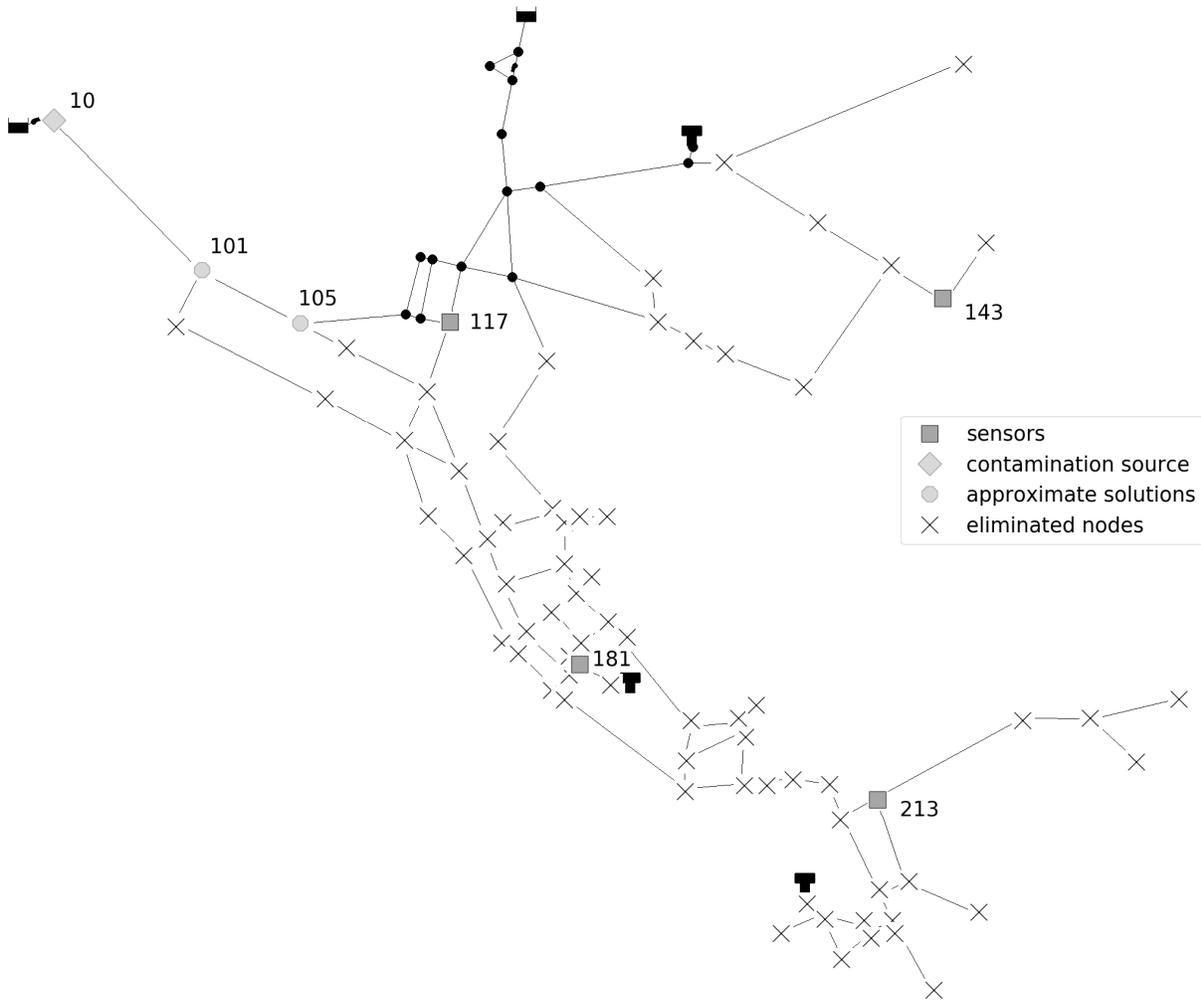
Fig. 5. Sensor locations, real source of contamination, approximate solutions and eliminated nodes for Net3 optimization problem.

Lučin, December 14, 2021
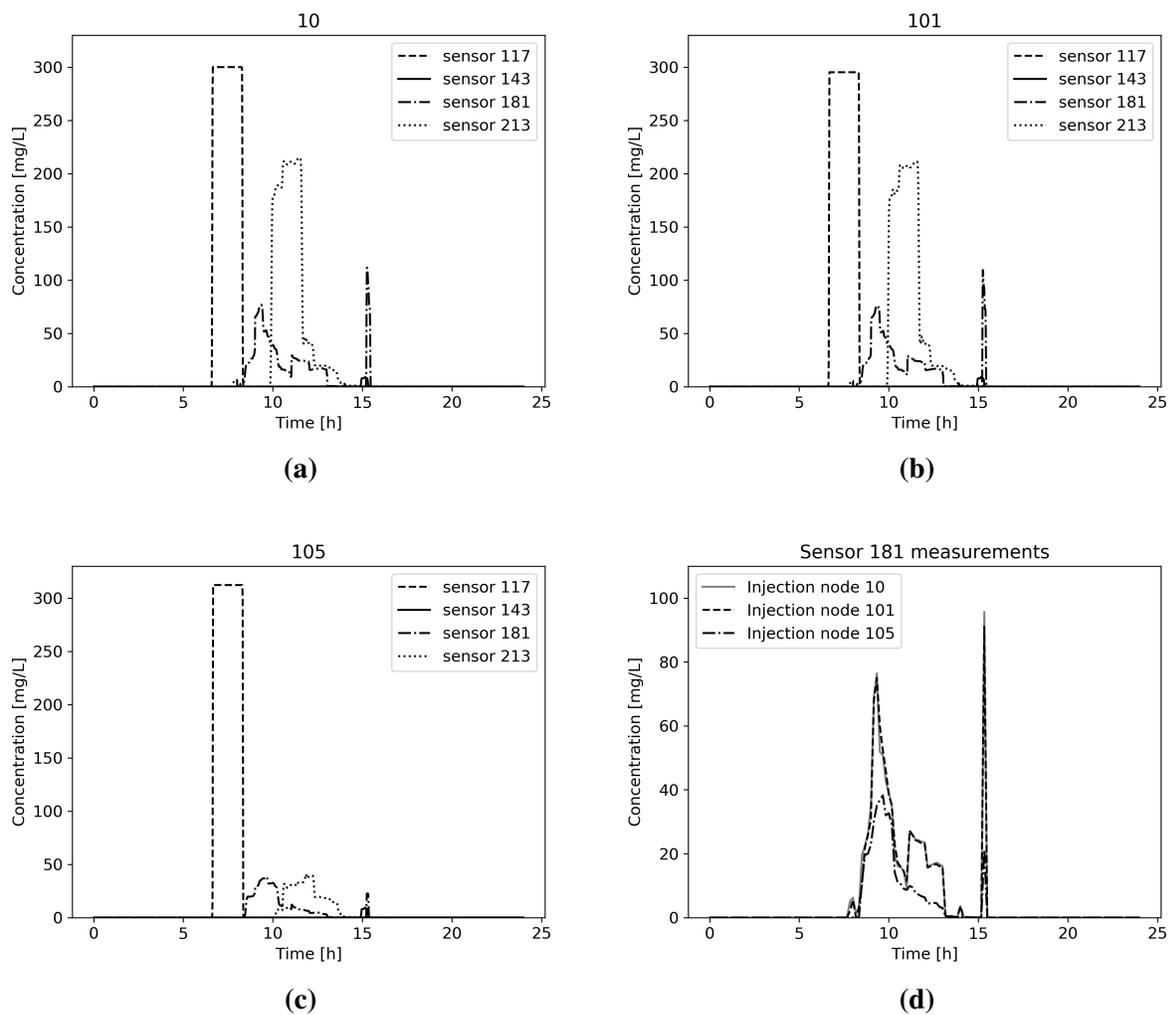
Fig. 6. Net3 measured concentrations in sensors for (a) exact solution, (b) second best and (c) third best solution. (d) Comparison of sensor 181 measurements for different injection nodes.
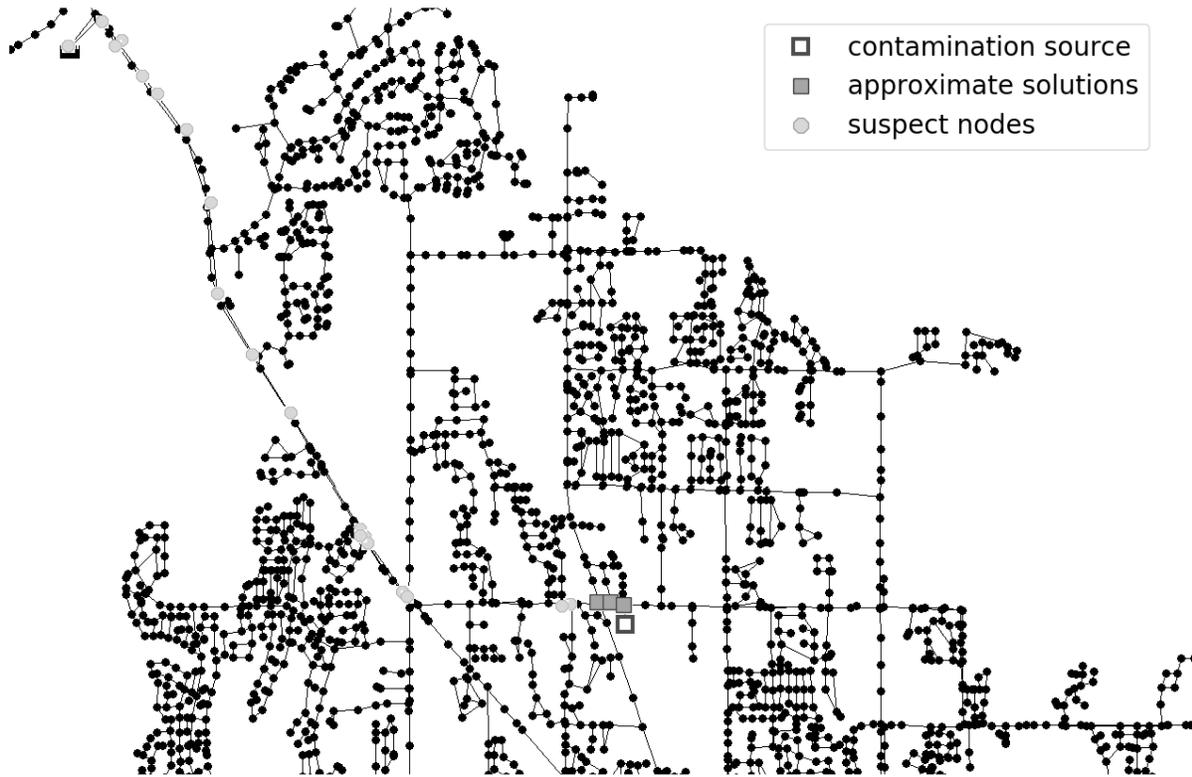
(a)
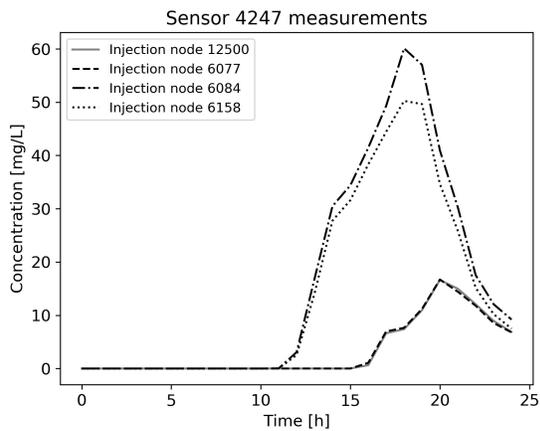
(b)
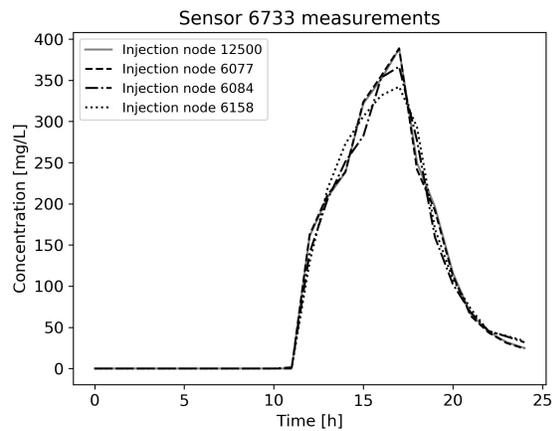
(c)

Fig. 7. (a) Real source of contamination, approximate solutions and suspect nodes for BWSN Network 2 optimization problem. Comparison of (b) sensor 4247 measurements and (c) sensor 6733 for real source and approximate solutions.
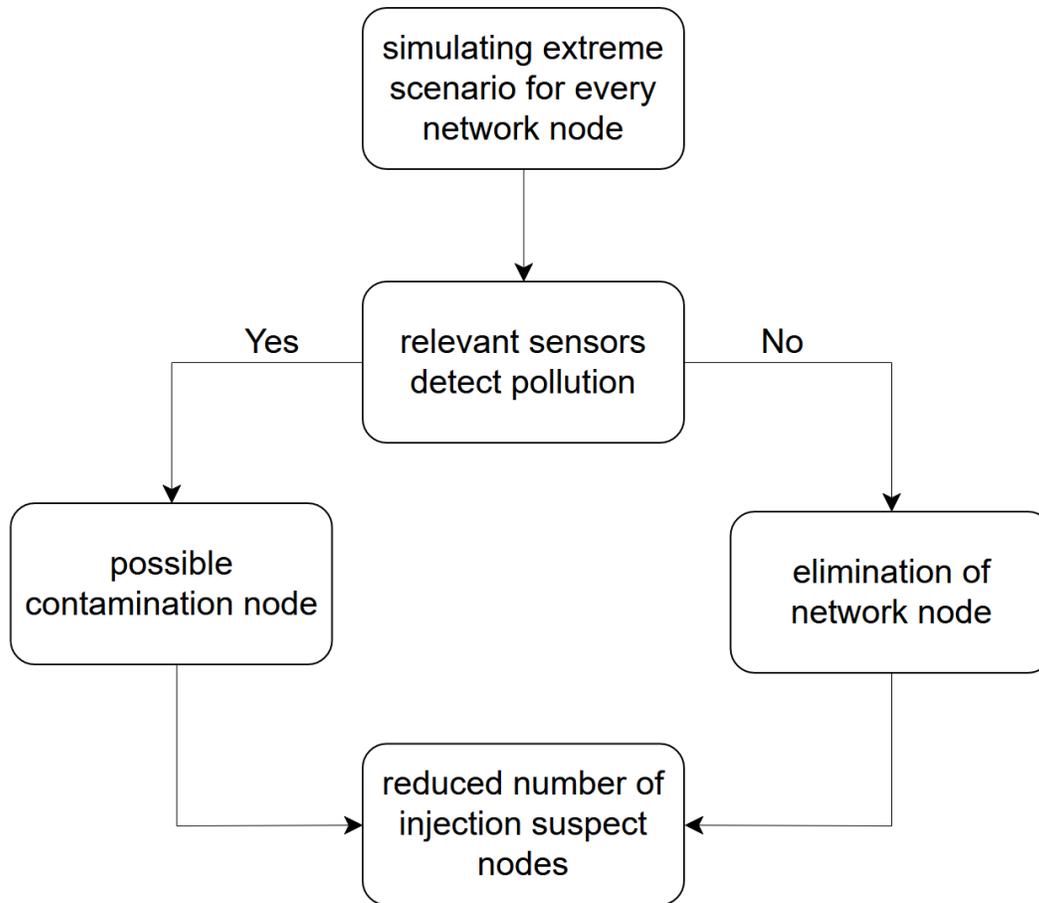
Lučin, December 14, 2021

Machine-Learning Classification of a Number of Contaminant Sources in an Urban Water Network

Ivana Lučin [1,2,*] , Luka Grbčić [1,2] , Zoran Čarija [1,2] and Lado Kranjčević [1,2]

[1] Faculty of Engineering, University of Rijeka, Vukovarska 58, 51000 Rijeka, Croatia; lgrbcic@riteh.hr (L.G.); zcarija@riteh.hr (Z.Č.); lado.kranjcevic@riteh.hr (L.K.)
[2] Center for Advanced Computing and Modelling, University of Rijeka, Radmile Matejčić 2, 51000 Rijeka, Croatia
* Correspondence: ilucin@riteh.hr; Tel.: +385-51-651-418

Abstract: In the case of a contamination event in water distribution networks, several studies have considered different methods to determine contamination scenario information. It would be greatly beneficial to know the exact number of contaminant injection locations since some methods can only be applied in the case of a single injection location and others have greater efficiency. In this work, the Neural Network and Random Forest classifying algorithms are used to predict the number of contaminant injection locations. The prediction model is trained with data obtained from simulated contamination event scenarios with random injection starting time, duration, concentration value, and the number of injection locations which varies from 1 to 4. Classification is made to determine if single or multiple injection locations occurred, and to predict the exact number of injection locations. Data was obtained for two different benchmark networks, medium-sized network Net3 and large-sized Richmond network. Additionally, an investigation of sensor layouts, demand uncertainty, and fuzzy sensors on model accuracy is conducted. The proposed approach shows excellent accuracy in predicting if single or multiple contaminant injections in a water supply network occurred and good accuracy for the exact number of injection locations.

Keywords: water distribution networks; water network contamination; machine learning; random forest; neural network

1. Introduction

Contamination in water distribution networks can occur due to deliberate or unintentional intrusions and it is of extreme importance to determine the contamination event parameters so it can be detected which parts of water distribution networks have been exposed to the contaminant and needed measures can be conducted. This is considered to be an inverse problem since injection location, injection starting time, injection duration, and contaminant chemical concentration value needs to be predicted based on sensor measurements. Numerical simulations are used to determine these parameters, but model limitations need to be taken into consideration. EPANET [1] is the most commonly used software for water distribution network simulations and uses an advective approach which cannot efficiently analyze contaminant dispersion in the networks. Piazza et al. [2] conducted experiments where it was shown that dispersive and diffusive processes must be incorporated in the transport model for less turbulent fluid flows to achieve more accurate results than the pure advection model. Also, EPANET assumes complete mixing in all network junctions, which can be valid only in the case of a single outlet or if there is considerable distance between two junctions. Therefore, EPANET extension EPANET-BAM [3] was proposed which uses experimentally calibrated mixing model parameter to more accurately model mixing in network junctions. A number of studies investigated mixing behavior for different conditions, both experimentally and numerically, to further enhance these simpler 1D numerical models [4–9].

Huang and McBean [10] investigated a data mining approach for identifying possible sources of intrusion where single and multiple injection scenarios were considered. In the case of multiple injection scenario, the method provided a limited number of nodes with the probability of them being the true contamination source. However, in their work, it is not predicted what is the true number of injection locations. In Wang and Harrison [11] a Bayesian approach was coupled with Support Vector Regression to provide a probability distribution of water network nodes being contaminant sources. However, a single injection is assumed, and it is noted that multiple contaminant sources should be considered in future work where the likelihood evaluation needs to be adjusted. Seth et al. [12] investigated the efficiency of three different methods for source detection; Bayesian probability-based method, backtracking method (using contaminant status algorithm), and optimization-based method where accuracy in case of multiple injection locations was investigated for two and three contamination injection locations. It was noted that the Bayesian method is designed only for a single contamination location while the contaminant status algorithm used in De Sanctis et al. [13] provides a list of possible solutions that narrow down search space for the optimization method; however, it also does not identify the possible number of injection locations. In Lučin et al. [14] a new search space reduction method was proposed, which can eliminate a considerable number of source nodes for both single and multiple injection locations, but with considerably greater reduction for single injection scenario. A number of different optimization approaches were considered to determine the contamination source, an overview of proposed methods can be found in Adedoja et al. [15]. Optimization approach can be easily extended to consider multiple contamination sources, as mentioned in [16–18].

If considering the optimization approach with multiple injection locations, with each additional source of contamination, the complexity of search space increases with an increase of optimization variables. Since the number of injection locations is not known, as a precaution, multiple injection locations should be allowed, since optimization can set variables to zero (which eliminates that source node and eliminates the number of injection locations), but it cannot add additional variables (injection locations) during the optimization process. In this way, in the case of a single injection location, optimization can eliminate other source nodes (all contamination parameters would be set to 0). However, this considerably increases the complexity of the considered problem since unnecessary fitness function evaluations would be conducted due to greater search space. Thus, it would be greatly beneficial to determine the number of injection locations before the optimization algorithm is employed. Also, if it is known that a single injection event occurred, a number of methods can be used more efficiently to reduce the complexity of the problem. For example, the machine learning approach provides probabilities for each network node being the true contamination scenario, which greatly reduces the number of suspect nodes and helps in quicker detection of true contamination location. However, in the case of multiple injections, different likelihood evaluation is needed which increases the complexity of the machine learning approach. Prediction of the number of contamination sources has previously been conducted for air pollution in Wade and Senocak [19], but to authors knowledge was not conducted for water distribution network contamination scenarios.

Machine learning tools have been increasingly used in contamination detection, where Random Forest has been used for groundwater source of contamination detection [20] and source detection in a river [21]. In Grbčić et al. [22] Random Forest algorithm was used to predict contamination event parameters in water distribution networks and in Grbčić et al. [23] new machine learning-based algorithm was proposed. A great advantage of prediction models is that they can be constructed before an accident occurs, so when a contamination event is detected prediction can be made even for large networks in a computationally efficient way. Thus, the proposed model which predicts number of injection locations can be used prior to conducting approaches that search for contamination parameters, without influencing the reaction time needed to contain the contamination event. However, in accident situations hydraulic conditions can greatly differ from those

on which model was trained, thus, a wrong prediction could be made. This can be handled with the preparation of multiple prediction models with different hydraulic conditions or by using a prediction model that achieves great accuracy with the small number of inputs so time for prediction also becomes negligible considering the benefit of search space reduction when redundant optimization parameters are not used.

In this paper, the Random Forest and Artificial Neural Network classifier are used to predict the number of contamination sources based on contamination sensor measurements in the water distribution network. Sensor measurements of contamination needed for model teaching are obtained from contamination scenarios simulated using EPANET2 with Monte Carlo generated contamination parameters. An investigation was conducted for two different sized benchmark water distribution networks with different sensor layouts, to examine the efficiency of the proposed machine learning approach. Investigation of demand uncertainty and fuzzy sensors is also estimated.

2. Materials and Methods

2.1. Benchmark Water Supply Networks

Prediction of the number of injection sources is conducted for two benchmark different sized networks. Investigated networks are Net3 EPANET2 example consisting of 92 nodes and Richmond network consisting of 865 nodes, obtained from The Centre for Water Systems (CWS) at the University of Exeter [24]. For the Net3 network, two different sensor layouts are investigated. In first layout four sensors were placed in network nodes 117, 143, 181, and 213 as in [25] and in second layout four sensor were placed in network nodes 115, 119, 187, and 209 as in [26]. Additionally, an investigation of the number of sensors was conducted. For the first layout, two sensors were placed in network nodes 117 and 181, and for the second layout sensors were placed in network nodes 119 and 209. For Richmond network five sensors were placed in network nodes 93, 352, 428, 600, and 672 where sensor layout was taken from [27]. Layout with three sensors placed in network nodes 93, 428, and 672 was also considered. Considered networks with sensor layouts can be seen in Figures 1 and 2.

Contamination scenarios are simulated using EPANET2 version 2.0.12. where for both networks, simulation time is 24 h with a hydraulic time step of 10 min, quality time step 5 min, pattern time step 10 min and report time step 1 h. For all conducted simulations, the EPANET2 flow paced method is used for the contaminant injection. Contamination scenario parameters are chosen randomly. The number of injection locations is chosen from 1 to 4 nodes. The starting time and duration of contamination injection are chosen from 0 to 24 h. Concentration was randomly chosen from 10 to 2000 mg/L. For contamination scenarios with multiple injection locations starting time, duration, and concentration was kept the same for every injection location.

Prior to simulating multiple injection scenario, independent simulations for each randomly chosen node as a source of contamination are conducted. If contamination is not registered for the investigated node with chosen contamination parameters, that node is eliminated as source location and only nodes for which contamination was detected in at least one sensor are kept as a source of contaminant. For example, if four source nodes are randomly chosen to be the source of contamination, but only two source nodes influence sensor detection of contaminant, the same time series of sensor measurements would be obtained for two, three, and four injection locations since the latter two do not influence contamination measurements. If four sources are given to the prediction model as input, where contamination can be measured only from two sources, that would significantly reduce the accuracy of the prediction model. Thus, only nodes which contribute to the contamination measurements in sensors are considered for multiple injection scenario.
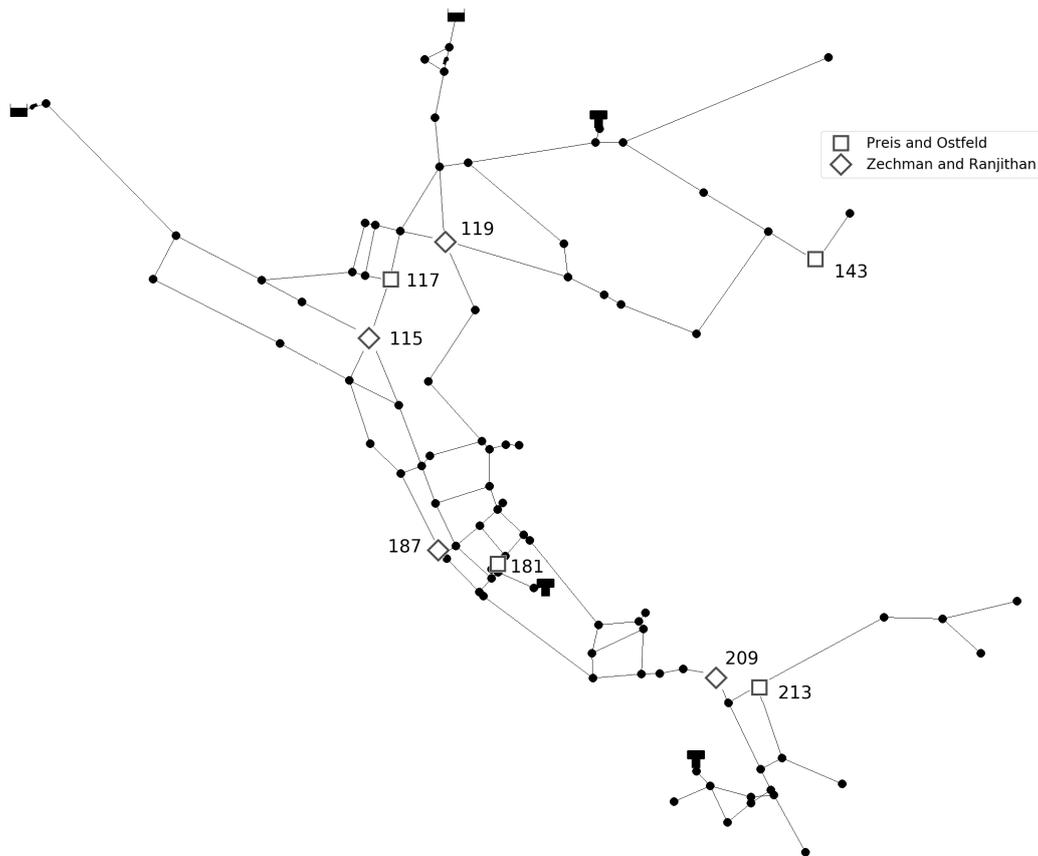
Figure 1. Net3 network with sensor layouts.
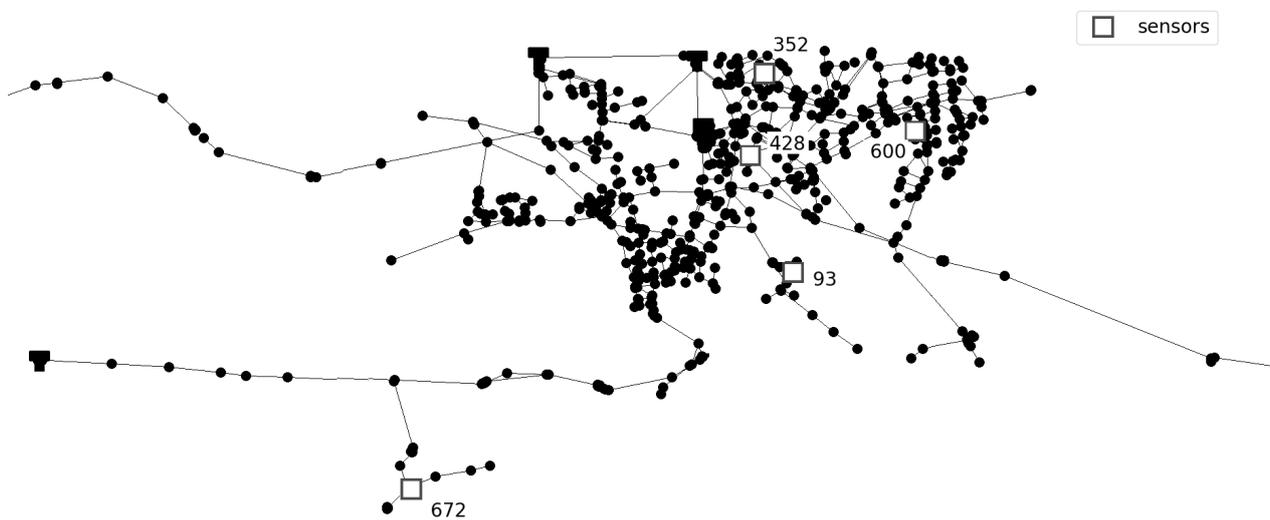
Figure 2. Richmond network detail with sensor layout.

An example of the proposed methodology can be seen for arbitrarily chosen Net3 contamination scenario in Figure 3. Randomly chosen contamination scenario parameters are 3 source nodes (159, 151 and 123), with contamination value of 200 mg/L, starting time 13 h and 20 min and injection duration 2 h. Sensor measurements for chosen contamination scenario can be seen in Figure 4. It can be observed that for source node 151 contamination scenario remains undetected in all sensors placed in the water distribution network, thus for multiple sources scenario only source nodes 123 and 159 are further considered.

☐ sensors
☐ contamination source

Source node 159

Source node 151

Source node 123

Source nodes 123, 159
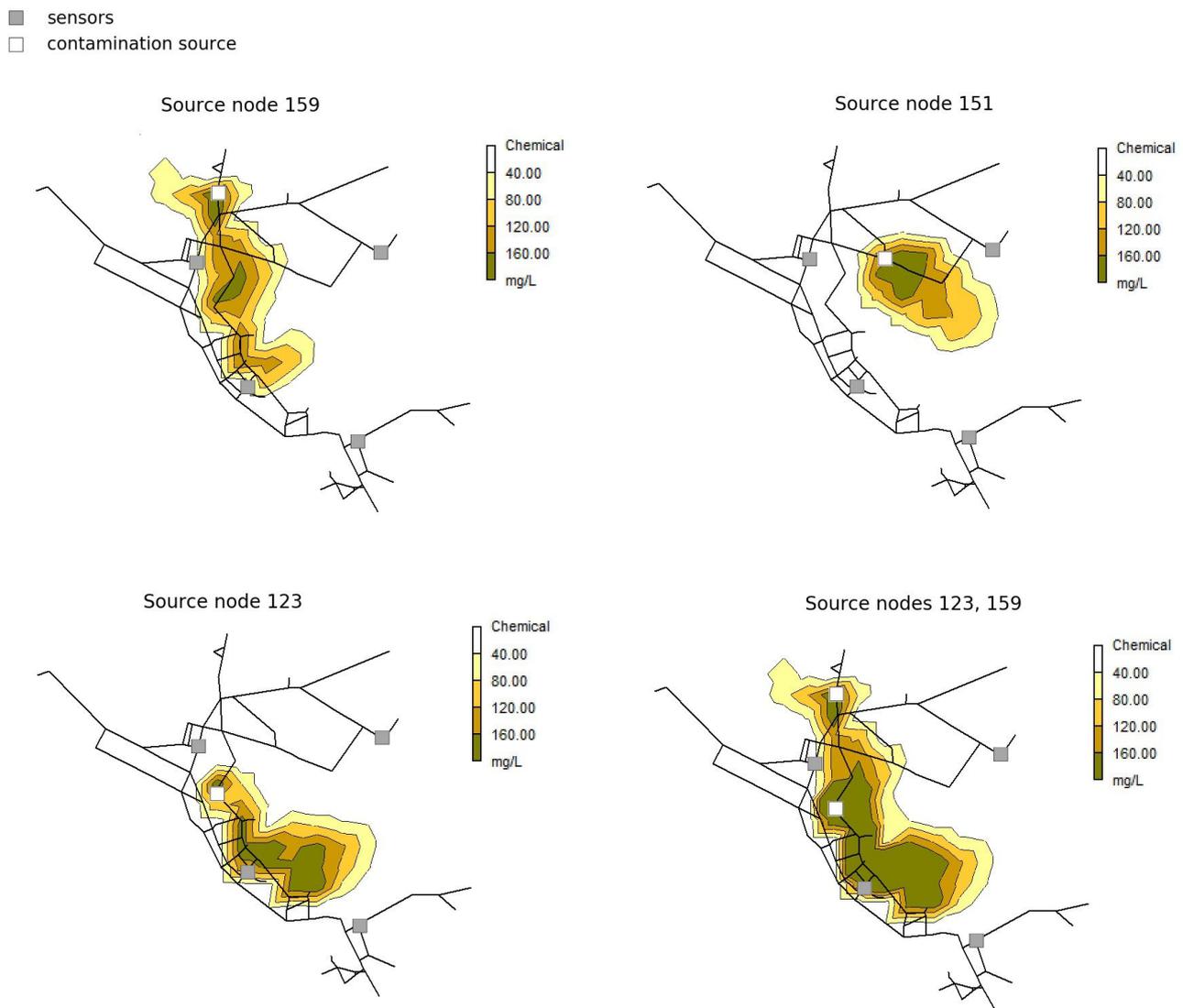
Figure 3. Contours of chemical for randomly chosen Net3 contamination scenario 90 min after injection starting time. Contamination from source node 151 remains undetected, so the source node is not included for multiple injections scenario.
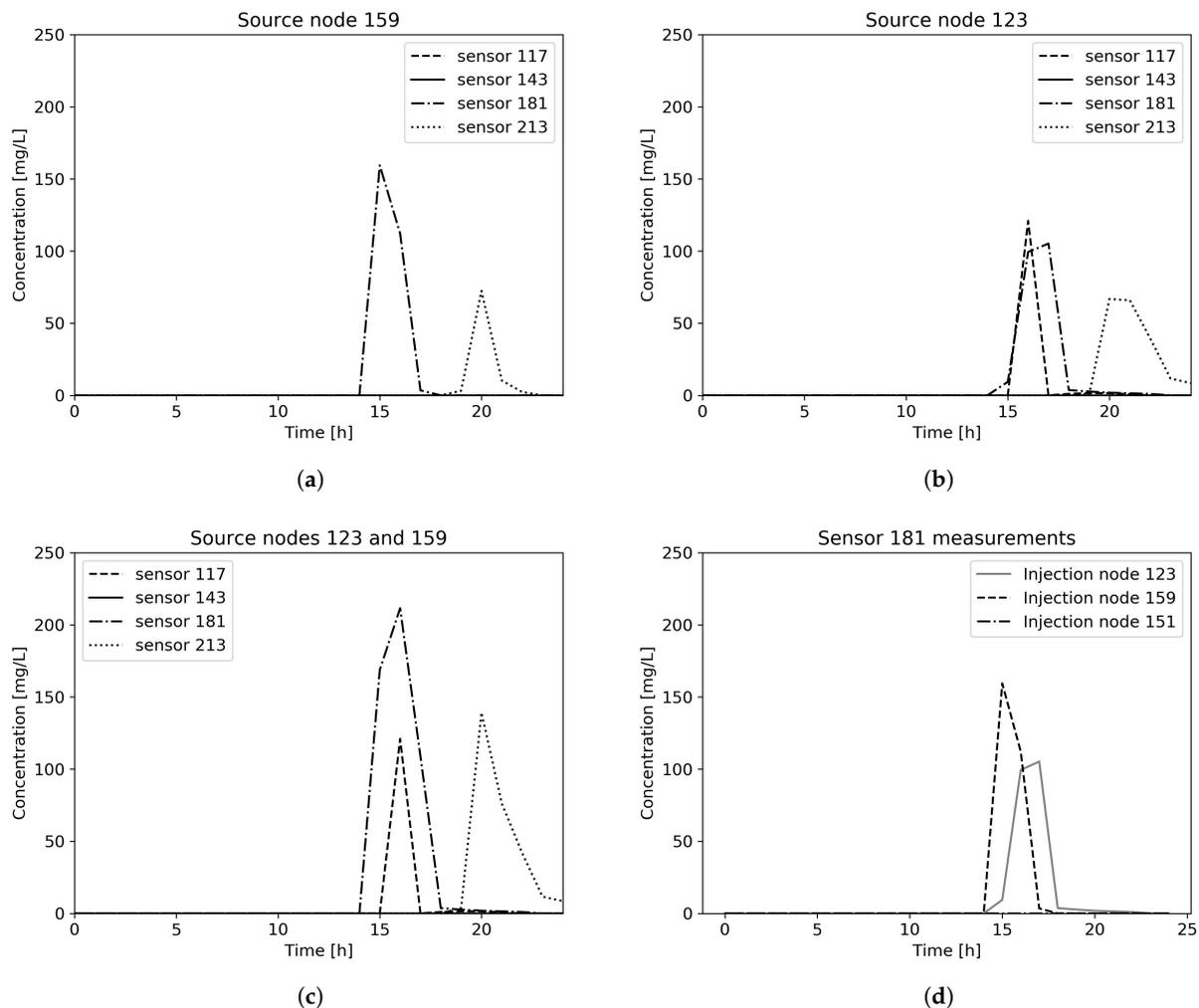
Figure 4. Sensor measurements for Net3 contamination scenario with (**a**) injection node 159, (**b**) injection node 123, (**c**) injection nodes 123 and 159 and (**d**) contamination measurements in the sensor in node 181.

2.2. Demand Uncertainty and Sensor Type

To investigate demand uncertainty, for both Net3 and Richmond networks, for every network node first it was randomly chosen if demand will be altered or not. If node base demand was to be altered, the percentage from 0–5% is randomly chosen for each network node, to reduce or increase base demand by the chosen percentage, resulting in a random demand span of 10%. To further investigate influence of demand uncertainty, the percentage from 0–10% is randomly chosen to reduce or increase base demand, resulting in a random demand span of 20%. All network demand patterns were kept the same, only base demand was changed. This method was conducted for every contamination scenario, thus resulting in different hydraulic conditions for each contamination scenario.

For sensor type influence, fuzzy sensor measurements were made where sensor detection was considered either low, medium, or high. Chemical concentration value C in range $0 < C < 300$ mg/L was considered low, in range $300 < C < 1000$ mg/L was considered medium and high if $C > 1000$ mg/L. Prediction model input features were defined as 0 if no contaminant was detected, 1 for low measurements, 2 and 3 for medium and high measurements, respectively.

2.3. Machine Learning Classifiers

Two different machine learning classifiers, Random Forest and Artificial Neural Network were used to compare the efficiency of the proposed method. Random Forest al-

gorithm [28], based on multiple decision trees is used, with 250 estimators (trees) with a maximum depth of 30 and the minimum number of samples required to split an internal node 8. An artificial neural network with three hidden layers with 100 nodes in each layer, with hyperbolic tangent activation function and Adam solver for weight optimization is used. Proposed parameters were chosen with the grid search hyperparameter optimization method, while other parameters, which are not mentioned, are kept constant. Implementation in the Python library Scikit-learn [29] version 0.20.3 is used for both classifiers. Obtained data was split 70% for teaching and 30% for model testing. Flowchart of the prediction model can be seen in Figure 5. Data generation and prediction model training was done using the supercomputing resources at the Center for Advanced Computing and Modelling, University of Rijeka.
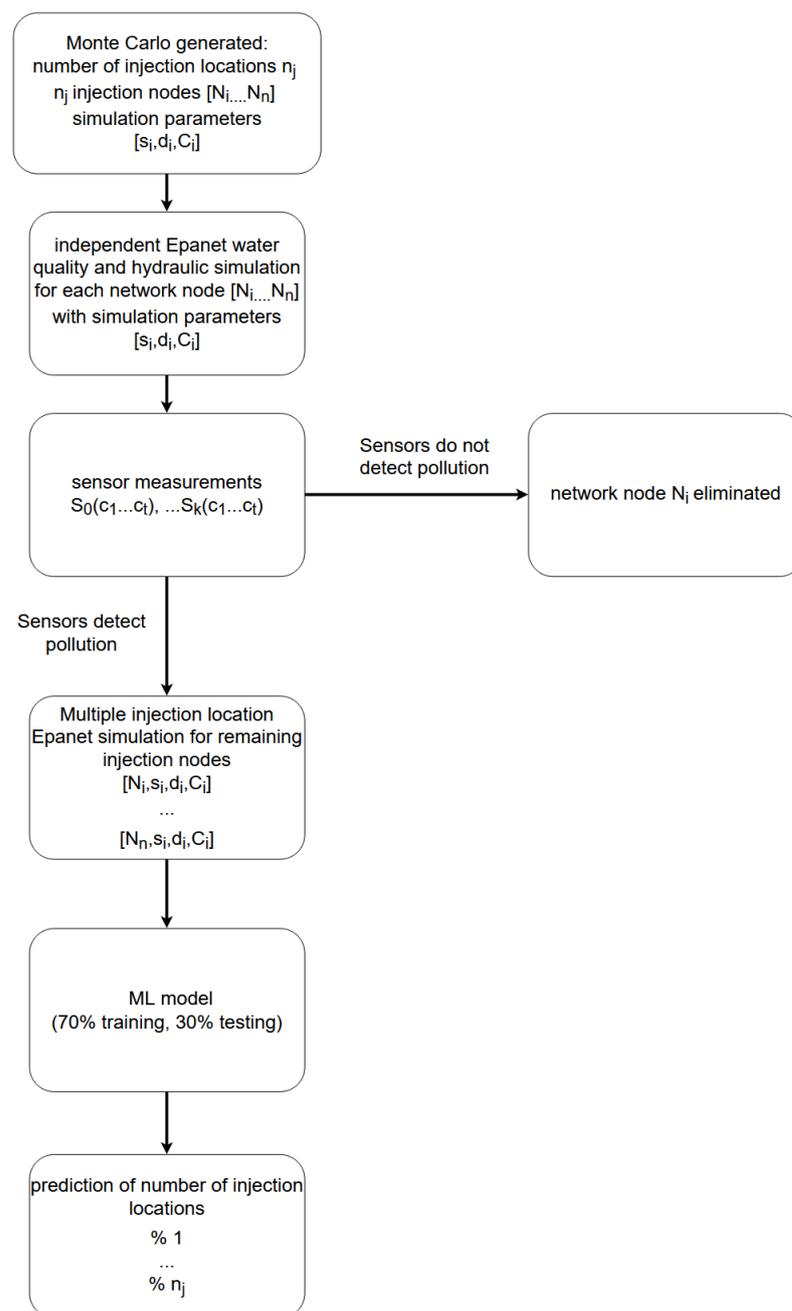
Figure 5. Flowchart of Machine Learning algorithm for prediction of number of contamination sources.

Input data for the prediction model is the time series of sensor measurements. For both Net3 and Richmond network, 25 features per sensor are obtained, which resulted in 100 features for Net3 and 125 features for Richmond network. The output of the machine learning model is the number of injection locations where two different prediction models are used. The first prediction model was used to predict the exact number of injection locations, i.e., 4 different classes are predicted. In the second model it is predicted only if single or multiple injections occurred, i.e., 2, 3 and 4 injection locations are treated as same, multiple injections class, thus only 2 different classes are predicted (single and multiple injections). To further increase the accuracy of the latter prediction model, the threshold value is introduced. Only if the model predicts a single source scenario with a probability greater than the chosen threshold value, single source prediction is made. In other cases, the scenario is treated as multiple sources. Threshold values of 50%, 60%, 70%, 80%, 90%, and 95% are investigated.

3. Results

3.1. Model Accuracy

The influence of input data on prediction model accuracy is investigated for both benchmark networks where data ranged from 50,000 to 500,000 inputs (Figure 6). An investigation is conducted for prediction model with 2 categories (model predicts only if single or multiple injection locations are present) and with 4 categories (model predicts an exact number of injection locations). For each model and each number of inputs, 20 runs were conducted to take into consideration the influence of random seed. For the Net3 network second sensor layout with sensors placed in nodes 115, 119, 187, and 209 was considered. For Net 3 results are presented for both RF and NN prediction models. Standard deviation ranged from 0.63% for 50,000 to 0.33% for 500,000 inputs for NN model, and from 0.33% for 50,000 to 0.1% for 500,000 inputs. It can be observed that the RF model has slightly better accuracy for all investigated models. Also, due to the faster execution time of the RF model, for all further analyses, only RF results will be presented. For Richmond network, standard deviation ranged from 0.28% for 50,000 inputs to 0.12% for 500,000 inputs which indicates the stability of the model. Presented results are an average of all 20 runs.

(a)

Figure 6. *Cont.*

(**b**)
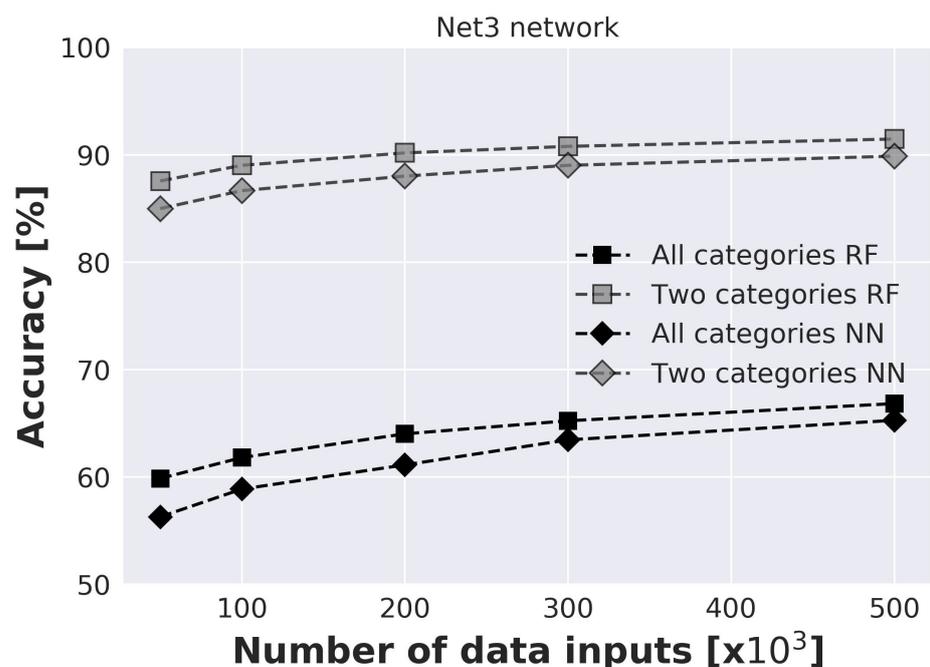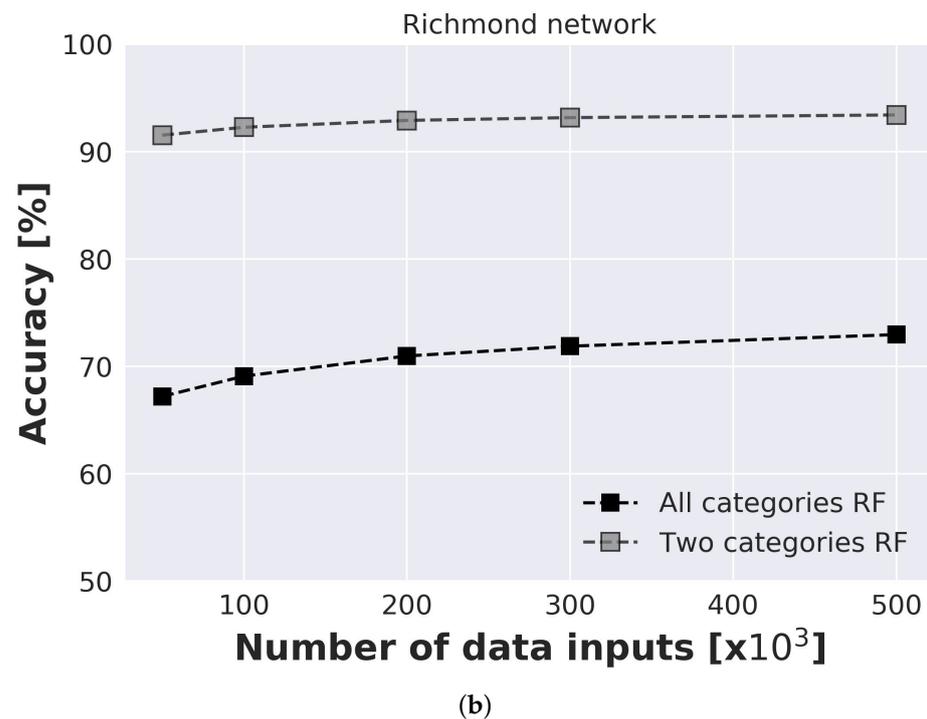
Figure 6. Accuracy of prediction models for different number of inputs for (**a**) Net3 network and (**b**) Richmond network.

It can be observed that even for a small number of input data considerable accuracy can be achieved. For model with 2 categories even with 50,000 inputs accuracy of the model is above 85% for both considered networks. After 200,000 inputs accuracy of the models for both networks tend to only slightly increase with the further increase of the number of input data. For 500,000 inputs accuracy of the Net3 network is 66.83% and for Richmond network 72.96%. When simplification is made, and the model only needs to predict single or multiple injection locations, accuracy significantly increases and for 500,000 inputs for the Net3 network is 91.46% and for the Richmond network 93.4%.

3.2. Threshold Influence

To further increase the accuracy of the prediction model, the threshold value is introduced for the model which predicts 2 categories. Detailed results are presented for models with 500,000 inputs for Net3 (Tables 1 and 2) and Richmond network (Tables 3 and 4). Presented results are the average of values obtained from 20 runs. As expected, with the increase in threshold value accuracy of the prediction model increases. However, with a greater threshold value, a greater number of single injection scenarios, as a precaution, are classified as multiple sources, thus a smaller number of true single injection scenarios are detected. For both networks, when the threshold value is 95%, a very low percentage of correct prediction of single source scenarios can be observed when prediction model parameters chosen with grid search optimization method (250 estimators, maximum depth 30, minimum samples for split 8) were used (Tables 1 and 3). Thus, different prediction model parameters (180 estimators, maximum depth 80, minimum samples for split 10) were also investigated to test its influence on model accuracy when threshold values are considered. In Tables 2 and 4 it can be observed that for the greatest threshold value (95%) correct prediction of single sources scenarios greatly increases, and is around 30% of the total number of single source scenarios. As threshold value decreases, similar percentages are observed for both models, which indicates that model accuracy is similar for different RF parameters. However, when greater prediction certainty is expected, model parameters must be carefully considered.

For both networks, accuracy with threshold value 95% is above 99.5%. It can be observed from Table 2 that for Net3 only 36% of total number of single source scenarios are correctly predicted where for Richmond network (Table 4) that value is 37%. For threshold value 50% for Net3 94.5% of single injection scenarios are correctly predicted; however, the number of wrong predictions increases. The same can be observed for the Richmond network where for threshold value 50%, 97.8% of single injection scenarios are correctly predicted but the percentage of wrong single injection scenarios increases from 0.8% to 12.7%.

The problem remains with scenarios that are wrongly predicted even for a threshold value of 95%. With further increase of threshold value, the number of wrongly predicted scenarios would decrease, but only because ultimately all scenarios would be classified as multiple sources (this can also be observed in Tables 1 and 3 for first chosen RF parameters). Thus, optimum threshold value should be chosen to both provide a reasonable number of single injection scenario predictions but with a high model accuracy. In-depth analysis of scenarios where the model wrongly predicts a single injection scenario with a high threshold value should be conducted. Also, it should be investigated how much accuracy of the model can be further increased with a larger number of inputs and with the usage of different classifiers.

Table 1. Influence of threshold value on model accuracy for Net3 network (250 estimators, maximum depth 30, minimum samples for split 8). Percentage indicates number of predicted simulations based on total number of single source scenarios.

Threshold Value	Accuracy	Single Source Scenarios	Correct Prediction	Wrong Prediction
95%	99.98%	48,682	3307 (6.8%)	36 (0.07%)
90%	99.73%	48,682	15,388 (31.6%)	405 (0.8%)
80%	98.6%	48,682	34,717 (71.3%)	2085 (4.3%)
70%	97.5%	48,682	41,204 (84.6%)	3683 (7.6%)
60%	96.7%	48,682	44,334 (91.1%)	4914 (10.1%)
50%	95.7%	48,682	46,388 (95.3%)	6390 (13.1%)

Table 2. Influence of threshold value on model accuracy for Net3 network (180 estimators, maximum depth 80, minimum samples for split 10). Percentage indicates number of predicted simulations based on total number of single source scenarios.

Threshold Value	Accuracy	Single Source Scenarios	Correct Prediction	Wrong Prediction
95%	99.7%	48,783	17,458 (35.8%)	508 (1%)
90%	99.4%	48,783	25,426 (52.1%)	863 (1.8%)
80%	98.9%	48,783	35,197 (72.2%)	1667 (3.4%)
70%	98.2%	48,783	40,640 (83.3%)	2636 (5.4%)
60%	96.7%	48,783	43,977 (90.2%)	3737 (7.7%)
50%	95.7%	48,783	46,091 (94.5%)	5072 (10.4%)

Table 3. Influence of threshold value on model accuracy for Richmond network (250 estimators, maximum depth 30, minimum samples for split 8). Percentage indicates number of predicted simulations based on total number of single source scenarios.

Threshold Value	Accuracy	Single Source Scenarios	Correct Prediction	Wrong Prediction
95%	99.9%	52,911	375 (0.7%)	5 (0.001%)
90%	99.8%	52,911	10,889 (20.6%)	303 (0.6%)
80%	97.9%	52,911	37,463 (70.8%)	3076 (5.8%)
70%	95.8%	52,911	49,149 (92.9%)	6269 (11.9%)
60%	94.8%	52,911	51,427 (97.2%)	7819 (14.8%)
50%	93.9%	52,911	52,198 (98.65%)	9178 (17.3%)

Table 4. Influence of threshold value on model accuracy for Richmond network (180 estimators, maximum depth 80, minimum samples for split 10). Percentage indicates number of predicted simulations based on total number of single source scenarios.

Threshold Value	Accuracy	Single Source Scenarios	Correct Prediction	Wrong Prediction
95%	99.7%	52,941	19,499 (36.8%)	435 (0.8%)
90%	99.3%	52,941	30,305 (57.2%)	1085 (2.1%)
80%	98.3%	52,941	42,000 (79.3%)	2567 (4.9%)
70%	97.3%	52,941	47,654 (90%)	4061 (7.7%)
60%	96.4%	52,941	50,433 (95.3%)	5433 (10.3%)
50%	95.5%	52,941	51,775 (97.8%)	6703 (12.7%)

3.3. Sensor Layout

The influence of sensor layout was tested for both Net3 and Richmond networks. 20 runs were conducted for the model with 500,000 inputs and average accuracy for all runs can be seen in Table 5. It can be observed that for the same number of sensors, their layout influences the accuracy of prediction models. This is expected, since the same behavior can be seen when the detection rate of contamination event is investigated for different sensor layouts. In the paper by Ostfeld et al. [30] for the same network and the same number of sensors detection likelihood of contamination event greatly differs for different sensor layouts. Results show that the prediction model for 2 categories (predicts single or multiple injections) is less influenced by sensor layout and all sensor layouts have accuracy around 90% or higher.

Interestingly, greater model accuracy can be observed when a smaller number of sensors is placed for Net3 layout with sensors in nodes 117, 143, 181, and 213 and for Richmond network. However, it can be explained with the fact that a greater number of contamination events remain undetected. i.e., with the greater number of sensors, contamination events from the greater number of network nodes are detected, resulting in more combinations when considering multiple injection locations. When sensor placement is sparser, a smaller number of network nodes can be detected when the contamination event occurs, resulting in a smaller number of combinations for multiple injection locations and consequently providing better model accuracy with 500,000 inputs.

Table 5. Influence of sensor layout for Net3 and Richmond networks on prediction model accuracy.

	Sensors Locations	Accuracy	
		4 Categories	2 Categories
Net3	117, 143, 181, 213	71%	94%
	115, 119, 187, 209	67%	91%
	117, 181	75%	89%
	119, 209	63%	89%
Richmond	93, 352, 428, 600, 672	73%	93%
	93, 428, 672	83%	92%

3.4. Demand Uncertainty and Fuzzy Sensors

Influence of demand uncertainty and fuzzy sensors was investigated for Net3 network with 4 sensors in nodes 117, 143, 181 and 213 and for Richmond network with 5 sensors in nodes 93, 352, 428, 600 and 672. 20 runs were conducted for RF models with 500,000 inputs and average accuracy can be observed in Table 6. When demand uncertainty is considered the accuracy of RF models slightly decreases for both networks. The influence of fuzzy sensors is more prominent, where the greater reduction in prediction accuracy can be observed for the Net3 network. When considering both demand uncertainty and fuzzy sensors in the same model, accuracy further slightly decreases. However, it can be

observed that for both networks model which predicts 2 categories has accuracy above 90% for all cases. This shows that the proposed model could be applied in a real case scenario.

Table 6. Influence of demand uncertainty and fuzzy sensors for Net3 and Richmond network on prediction model accuracy.

	Net3	
	4 Categories	**2 Categories**
perfect sensors	71%	94%
demand uncertainty ($\pm5\%$)	69%	93%
demand uncertainty ($\pm10\%$)	69%	93%
fuzzy sensors	65%	91%
demand uncertainty ($\pm5\%$) and fuzzy sensors	64%	90%
demand uncertainty ($\pm10\%$) and fuzzy sensors	63%	90%
	Richmond	
	4 Categories	**2 Categories**
perfect sensors	73%	93%
demand uncertainty ($\pm5\%$)	72%	93%
demand uncertainty ($\pm10\%$)	72%	93%
fuzzy sensors	72%	93%
demand uncertainty ($\pm5\%$) and fuzzy sensors	71%	93%
demand uncertainty ($\pm10\%$) and fuzzy sensors	71%	92%

4. Discussion

Accuracy of prediction models for both networks has similar results with small differences, which shows that the proposed methodology could be successfully applied to other networks. Further investigation should be conducted for large size water distribution networks and different sensor placements, to fully investigate the robustness of the proposed method. Also, it must be noted that simplification was used in this study, where all source nodes had the same parameters (injection starting time, duration, and concentration value), thus, it should be investigated how the model predicts if those parameters are different for each injection node.

Although slightly, with the increase of input data model accuracy still increases, so in further study a greater number of data inputs should be investigated. Also, in the proposed scenarios report time step was chosen to be 1 h, resulting in 25 features per sensor. It should be investigated if a greater number of features, i.e., smaller report time step would increase model accuracy and if similar model accuracy could be achieved with a smaller number of contamination readings. The optimal number of features and inputs should be investigated to achieve great accuracy but with reasonable execution time. However, to obtain a greater number of inputs a greater amount of time is needed, so the model should be trained before the actual contamination event occurs. In that case, the model would be trained with simulation results with average demand patterns. This surely would mean that true contamination event will have different demands which would influence the accuracy of the prediction model. Investigation of demand uncertainty with arbitrarily chosen demand variation spans showed that small differences of base demands slightly influence prediction model accuracy. However, it must be taken into consideration that when base demand variation is defined with percentage, small demand variation is achieved when base demand is small and greater demand variation only when base demand is greater. Greater difference in demands should be further investigated since the usual variability of consumption can be greater than considered in this paper. Different machine learning models, with different expected demand patterns, can be prepared for contamination event so prediction can be obtained instantaneously. However, in case of contamination event, greater oscillations in the hydraulics of water distribution network could occur, such

as pipe burst or some other unplanned event, which would greatly influence change in demand patterns. Thus, it would be beneficial to investigate other algorithms that could increase accuracy with a smaller number of input data. In that case, input data can be obtained after the contamination event occurred, in a reasonable amount of time. That would be greatly beneficial since the simulation model can then be calibrated with sensor measurements from the field and input data would be more precise. The proposed method can be easily coupled with other machine learning approaches since inputs obtained for this model can also be used for teaching model that predicts injection location.

Investigation of different sensor layouts, demand uncertainty, and fuzzy sensors showed that sensor layout and type of sensors have the greatest impact on prediction model accuracy. Demand uncertainty slightly decreases model accuracy. However, model accuracy can be greatly reduced when a real case event is considered since both demand uncertainty and measurement errors can be greater than considered in this work. Thus, a threshold value is introduced which can help increase model accuracy. Greater threshold value increases model accuracy; however, it also leads to a greater number of single injection scenarios classified as multiple injections. It is also observed that prediction models are not very sensitive to model parameters; however, when threshold value is used, i.e., model prediction certainty is evaluated, model parameters are very important for method efficiency. Thus, the investigation of different machine learning approaches should be further investigated to increase model accuracy.

When observing presented results it must be taken into consideration that numerical model simplifications are made, where EPANET was used which assumes complete mixing in all network junctions and uses pure advection transport model. Also, in the presented study benchmark networks are used, and numerical simulations are conducted for only 24 h, where more than 24 h are needed to obtain stable contamination scenario results. However, the functionality of the presented machine learning approach is not dependent on the numerical model setup, and it is assumed that the same numerical approach that is chosen for the optimization process is to be also chosen for the prediction model preparation. In this way, all discrepancies due to numerical model simplifications would be also present in the optimization and as such are not the result of using the proposed machine learning approach. Furthermore, network uncertainties were not considered regarding internal pipe diameter and pipe roughness which should be considered in the further research.

5. Conclusions

In this paper, the machine learning approach is presented which helps identify the number of injection locations based on sensor measurements. Random Forest classifier and Neural Network classifier are used on medium-sized benchmark network, where Random Forest classifier provided better accuracy and faster execution time, thus is used for all other investigations. Two different sized benchmark networks are considered, where it is shown that the machine learning approach can be successfully used to predict the number of injection locations. This can help define the number of optimization parameters, where redundant parameters can be avoided which needlessly increase the complexity of the problem. The prediction model shows great accuracy when it predicts only if single or multiple injection locations occurred. The threshold value is proposed which further increases model accuracy since the single injection scenario is assumed only if the model predicts with certainty greater than the threshold value. Lower accuracy is obtained when the exact number of injection locations is predicted. The accuracy of the prediction model is investigated for different sensor layouts and in case of demand uncertainties and fuzzy sensors. Conducted research showed promising results, where exploration of other algorithms and increased number of input data should be investigated to further increase the accuracy of both models.

Author Contributions: Conceptualization, I.L. and L.G.; Data curation, I.L.; Formal analysis, I.L; Investigation, I.L. and L.G.; Methodology, I.L. and L.G.; Resources, Z.Č. and L.K.; Software, I.L.; Supervision Z.Č. and L.K.; Validation, I.L.; Visualization; I.L.; Writing—original draft, I.L.; Writing—

review and editing, L.G., Z.Č. and L.K. All authors have read and agreed to the published version of the manuscript.

References

1. Rossman, L.A. EPANET 2: Users Manual. 2000. Available online: https://epanet.es/wp-content/uploads/2012/10/EPANET_User_Guide.pdf (accessed on 6 September 2020).
2. Piazza, S.; Blokker, E.M.; Freni, G.; Puleo, V.; Sambito, M. Impact of diffusion and dispersion of contaminants in water distribution networks modelling and monitoring. *Water Supply* **2020**, *20*, 46–58. [CrossRef]
3. Ho, C.K.; O'Rear, L., Jr. Evaluation of solute mixing in water distribution pipe junctions. *J. Am. Water Work. Assoc.* **2009**, *101*, 116–127. [CrossRef]
4. Yu, T.; Tao, L.; Shao, Y.; Zhang, T. Experimental study of solute mixing at double-Tee junctions in water distribution systems. *Water Sci. Technol. Water Supply* **2015**, *15*, 474–482. [CrossRef]
5. Yu, T.; Qiu, H.; Yang, J.; Shao, Y.; Tao, L. Mixing at double-Tee junctions with unequal pipe sizes in water distribution systems. *Water Sci. Technol. Water Supply* **2016**, *16*, 1595–1602. [CrossRef]
6. Song, I.; Romero-Gomez, P.; Andrade, M.A.; Mondaca, M.; Choi, C.Y. Mixing at junctions in water distribution systems: An experimental study. *Urban Water J.* **2018**, *15*, 32–38. [CrossRef]
7. Grbčić, L.; Kranjčević, L.; Lučin, I.; Čarija, Z. Experimental and Numerical Investigation of Mixing Phenomena in Double-Tee Junctions. *Water* **2019**, *11*, 1198. [CrossRef]
8. Grbčić, L.; Kranjčević, L.; Družeta, S.; Lučin, I. Efficient Double-Tee Junction Mixing Assessment by Machine Learning. *Water* **2020**, *12*, 238. [CrossRef]
9. Grbčić, L.; Kranjčević, L.; Lučin, I.; Sikirica, A. Large Eddy Simulation of turbulent fluid mixing in double-tee junctions. *Ain Shams Eng. J.* **2020**. [CrossRef]
10. Huang, J.J.; McBean, E.A. Data mining to identify contaminant event locations in water distribution systems. *J. Water Resour. Plan. Manag.* **2009**, *135*, 466–474. [CrossRef]
11. Wang, H.; Harrison, K.W. Improving efficiency of the Bayesian approach to water distribution contaminant source characterization with support vector regression. *J. Water Resour. Plan. Manag.* **2014**, *140*, 3–11. [CrossRef]
12. Seth, A.; Klise, K.A.; Siirola, J.D.; Haxton, T.; Laird, C.D. Testing contamination source identification methods for water distribution networks. *J. Water Resour. Plan. Manag.* **2016**, *142*, 04016001. [CrossRef]
13. De Sanctis, A.E.; Shang, F.; Uber, J.G. Real-time identification of possible contamination sources using network backtracking methods. *J. Water Resour. Plan. Manag.* **2009**, *136*, 444–453. [CrossRef]
14. Lučin, I.; Grbčić, L.; Družeta, S.; Čarija, Z. Source Contamination Detection Using Novel Search Space Reduction Coupled with Optimization Technique. *J. Water Resour. Plan. Manag.* **2020**, *147*, 04020100. [CrossRef]
15. Adedoja, O.; Hamam, Y.; Khalaf, B.; Sadiku, R. Towards development of an optimization model to identify contamination source in a water distribution network. *Water* **2018**, *10*, 579. [CrossRef]
16. Vankayala, P.; Sankarasubramanian, A.; Ranjithan, S.R.; Mahinthakumar, G. Contaminant source identification in water distribution networks under conditions of demand uncertainty. *Environ. Forensics* **2009**, *10*, 253–263. [CrossRef]
17. Liu, L.; Ranjithan, S.R.; Mahinthakumar, G. Contamination source identification in water distribution systems using an adaptive dynamic optimization procedure. *J. Water Resour. Plan. Manag.* **2011**, *137*, 183–192. [CrossRef]
18. Liu, L.; Zechman, E.M.; Mahinthakumar, G.; Ranji Ranjithan, S. Identifying contaminant sources for water distribution systems using a hybrid method. *Civ. Eng. Environ. Syst.* **2012**, *29*, 123–136. [CrossRef]
19. Wade, D.; Senocak, I. Stochastic reconstruction of multiple source atmospheric contaminant dispersion events. *Atmos. Environ.* **2013**, *74*, 45–51. [CrossRef]
20. Rodriguez-Galiano, V.; Mendes, M.P.; Garcia-Soldado, M.J.; Chica-Olmo, M.; Ribeiro, L. Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability: A case study in an agricultural setting (Southern Spain). *Sci. Total Environ.* **2014**, *476*, 189–206. [CrossRef]
21. Lee, Y.J.; Park, C.; Lee, M.L. Identification of a contaminant source location in a river system using random forest models. *Water* **2018**, *10*, 391. [CrossRef]
22. Grbčić, L.; Lučin, I.; Kranjčević, L.; Družeta, S. Water supply network pollution source identification by random forest algorithm. *J. Hydroinf.* **2020**, *22*, 1521–1535. [CrossRef]
23. Grbčić, L.; Lučin, I.; Kranjčević, L.; Družeta, S. A Machine Learning-based Algorithm for Water Network Contamination Source Localization. *Sensors* **2020**, *20*, 2613. [CrossRef] [PubMed]
24. Centre for Water Systems, U.o.E. Benchmarks. Available online: http://emps.exeter.ac.uk/engineering/research/cws/downloads/benchmarks/ (accessed on 6 November 2019).

25. Preis, A.; Ostfeld, A. A contamination source identification model for water distribution system security. *Eng. Optim.* **2007**, *39*, 941–947. [CrossRef]
26. Zechman, E.M.; Ranjithan, S.R. Evolutionary computation-based methods for characterizing contaminant sources in a water distribution system. *J. Water Resour. Plan. Manag.* **2009**, *135*, 334–343. [CrossRef]
27. Preis, A.; Ostfeld, A. Genetic algorithm for contaminant source characterization using imperfect sensors. *Civ. Eng. Environ. Syst.* **2008**, *25*, 29–39. [CrossRef]
28. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
29. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
30. Ostfeld, A.; Uber, J.G.; Salomons, E.; Berry, J.W.; Hart, W.E.; Phillips, C.A.; Watson, J.P.; Dorini, G.; Jonkergouw, P.; Kapelan, Z.; et al. The battle of the water sensor networks (BWSN): A design challenge for engineers and algorithms. *J. Water Resour. Plan. Manag.* **2008**, *134*, 556–568. [CrossRef]

MDPI

Article

Data-Driven Leak Localization in Urban Water Distribution Networks Using Big Data for Random Forest Classifier

Ivana Lučin [1,2,*] , Bože Lučin [1], Zoran Čarija [1,2] and Ante Sikirica [1,2]

[1] Faculty of Engineering, University of Rijeka, Vukovarska 58, 51000 Rijeka, Croatia; blucin@riteh.hr (B.L.); zcarija@riteh.hr (Z.Č.); asikirica@riteh.hr (A.S.)
[2] Center for Advanced Computing and Modelling, University of Rijeka, Radmile Matejčić 2, 51000 Rijeka, Croatia
* Correspondence: ilucin@riteh.hr; Tel.: +385-51-651-418

Abstract: In the present paper, a Random Forest classifier is used to detect leak locations on two different sized water distribution networks with sparse sensor placement. A great number of leak scenarios were simulated with Monte Carlo determined leak parameters (leak location and emitter coefficient). In order to account for demand variations that occur on a daily basis and to obtain a larger dataset, scenarios were simulated with random base demand increments or reductions for each network node. Classifier accuracy was assessed for different sensor layouts and numbers of sensors. Multiple prediction models were constructed for differently sized leakage and demand range variations in order to investigate model accuracy under various conditions. Results indicate that the prediction model provides the greatest accuracy for the largest leaks, with the smallest variation in base demand (62% accuracy for greater- and 82% for smaller-sized networks, for the largest considered leak size and a base demand variation of ±2.5%). However, even for small leaks and the greatest base demand variations, the prediction model provided considerable accuracy, especially when localizing the sources of leaks when the true leak node and neighbor nodes were considered (for a smaller-sized network and a base demand of variation ±20% the model accuracy increased from 44% to 89% when top five nodes with greatest probability were considered, and for a greater-sized network with a base demand variation of ±10% the accuracy increased from 36% to 77%).

Keywords: leak localization; water distribution network; random forest; prediction modeling; big data

1. Introduction

Leakages in water distribution networks can cause great cumulative losses as small leakages can remain undetected for long periods of time. Direct losses are typically followed by the overall reduction of the functionality of the water distribution network, which usually manifests as a pressure drop on the user end. Moreover, leakages can potentially cause health hazards since microbiological contamination can enter the water distribution network and reach end users. Porous soil introduces additional difficulties, as even greater leakages can remain undetected since water is absorbed in the soil, and there is no evidence of the leakage on the surface. Thus, different technologies and methodologies have been proposed for leakage detection and localization. In a work by Jacobsz and Jahnke [1], leak detection using discrete fiber optic sensing was investigated. In a recent study by Nkemeni et al. [2], a wireless sensor network application was investigated, where processing for leak detection is performed at the sensor nodes. In a work by Wu et al. [3], a two-stage method was proposed, which first detects outliers from flow measurements using a clustering algorithm and then detects whether burst occurred. In the work of Rajeswaran et al. [4], a multi-stage graph partitioning algorithm was presented, which uses flow measurements to indicate a minimum number of additional measuring locations needed to narrow down

leak location in large-size networks. In the work by Cody et al. [5], a linear prediction signal processing technique was used to extract features from acoustic data, which can detect and localize pipe leaks. In a work by Bohorquez et al. [6], an artificial neural network was applied to detect leak size and location in a single water pipeline.

Problems with leak detection and localization in pipelines that are used for transportation of hydrocarbon fluids are also extensively explored, since leaks can cause serious damage to people and the environment due to often hazardous fluid that is transported. A number of investigations were conducted, including a data-driven approach using the Kantorovich distance [7], feature extraction from acoustic signals [8], application of a least squares twin support vector machine [9], and a multi-layer perceptron neural network (MLPNN) [10]. A detailed overview of leak detection technologies in pipelines can be found in a review paper by Adegboye et al. [11].

Additionally, a number of studies considered strategies for optimal sensor placement since it greatly influences leak detection and localization methods efficiency. The optimization approach is most widely used, and thus different enhancements were considered, such as a clustering process prior to optimization [12], hybrid feature selection method [13], methods that reduce the optimization search space [14], and an investigation of the influence of measurement uncertainty [15]. A detailed overview of leakage detection methodologies can be found in review papers by Wu and Liu [16], Chan et al. [17], and Zaman et al. [18].

Software-based leakage detection methods can be divided into transient-based, model-based, and data-driven approaches. The transient-based approach is based on various analyses of pressure signals; the model-based approach analyzes residuals, i.e., compares pressure measurements with the pressure estimation based on a hydraulic network model; and the data-driven approach relies on collected data and mathematical operations in order to determine anomalies in pressure. In recent years, machine learning methods have been increasingly used for leakage detection and localization. Zhou et al. [19] and Pérez-Pérez et al. [20] investigated leak detection in a single pipeline. In the Zhou et al. [19]'s work, a convolutional neural network (CNN) was used to pinpoint leak locations in a 1500 m long pipe segment for different leak sizes, where the better prediction was obtained for greater leakages. Pérez-Pérez et al. [20] used a combined artificial neural network (ANN), where the ANN is first used to estimate the friction factor of the pipe and then to localize leak location. Tests were conducted for a 64.48 m pipe, for which it was reported that an average percentage error of 0.47% was achieved. Mounce et al. [21] proposed a system using an artificial neural network for online detection of bursts in water distribution networks that was shown to have 44% of alarms when burst really occurred, 32% of alarms in cases of unusual short-term increased demand, 9% of alarms due to industrial events and only 15% were false alarms, indicating the applicability of the proposed method. In the work of Jensen et al. [22], a sensitivity analysis of pressure residuals was performed to isolate possible leakage locations. The proposed methodology was applied to the actual water distribution network, where only a few false alarms occurred, and frequent alarms occured during the leakage. It was observed that the proposed methodology can isolate a limited set of candidate nodes, where better performance was observed for greater flows in the system. In the work of Zhang et al. [23], a data-driven and model-based approach was utilized, where large-scale water distribution networks were divided into leakage zones that were categories for multi-class support vector machine prediction. Large-scale networks were divided into up to 25 zones, with a classification accuracy of above 90% for a division into 25 zones, which further increased with smaller divisions into leakage zones. However, it must be taken into consideration that further leak localization needs to be conducted after the leak zone is determined to provide the exact leak location. Soldevila et al. [24] used a mixed model-based and data-driven approach in which the K-nearest neighbors (k-NN) algorithm is used to localize leaks. The proposed methodology was applied to three different sized networks, with leak, demand, and sensor measurement uncertainties. For the Hanoi benchmark network, for all considered uncertainties in the study, an accuracy

greater than 90% was reported for the time horizon of one day using pressure sensor measurements from two sensors. Additionally, some network nodes were grouped since leaks from those nodes cannot be distinguished due to similar pressure measurements. Further study was presented by Soldevila et al. [25], where Bayesian classifiers were applied and greater accuracy than when using a k-NN approach was obtained. Both proposed methods were successfully applied to real water distribution network case studies where leak locations detected by the proposed methods were in the vicinity of the real leak locations. In the work of Quiñones-Grueiro et al. [26], an unsupervised approach to leak detection was conducted for the Hanoi distribution network using three pressure sensors, where the average reported classification accuracy was 85% for leak magnitudes smaller than 2.5% of the total demand of the network for leaks detected within a time interval of one day. In the work of Zhou et al. [27], after the burst was detected, additional pressure sensors were placed at optimal locations and deep learning was employed to identify burst locations. The proposed methodology was applied to 58 synthetic burst cases, where in 57 cases the top five most probable pipes were correctly identified, and in 37 cases the top pipe was correctly located. However, it must be noted that the requirement for additional measurements can extend reaction time in case of a pipe burst. In the work of Sun et al. [28], a classification approach was utilized where pressure measurements in network nodes with no pressure sensors were estimated using the Kriging method. The Hanoi water distribution network was considered with a wide range of sensors, and it was reported that in the average case 70% accuracy was achieved; however, for some sensor layouts the reported accuracy was below 20%, which is believed to be due to the Kriging interpolation error. Javadiha et al. [29] used a convolutional neural network with pressure measurements for the Hanoi network, where for a one day time horizon the model accuracy varied from 56% for four sensors to 94% for 12 sensors considering leak size uncertainty, sensor noise, and base demand uncertainty. The Kriging method was also used in work by Soldevila et al. [30] with satisfactory leak localization in a real water distribution network case; however, when compared with their previous work, the Kriging method did not provide better results.

The main drawback of the machine learning approach is that only a small amount of real data measurements can be obtained for leak events. Additionally, when a new installation is made in the water distribution network, all previous records are not valid, consequently reducing the number of inputs for the prediction model. This is a common problem in rapidly developing urban areas. Thus, in this paper, a machine learning approach is presented, in which a great number of leak scenarios for randomly chosen network nodes and with different leak sizes under different demand conditions were conducted, to obtain a database of pressure sensor measurements that are inputs for the prediction model. This idea is similar to that proposed by Grbčić et al. [31] and Lučin et al. [32], where a number of Monte Carlo simulations were conducted to obtain a large number of inputs for a machine learning prediction model that successfully detects the location of contamination source and determines the number of contamination sources. To the authors' knowledge, the currently proposed methodology has not been previously applied to the leak localization problem to obtain a large amount of synthetic data.

Model-based methods' accuracy is greatly dependent on model calibration, where model uncertainties can decrease the method's efficiency. In this work, model uncertainties are taken into consideration by including randomness for leak and demand values, so as to describe as many possible combinations of different leak scenarios. Machine learning classification is then utilized to detect the most appropriate leak scenario, which will be utilized to determine leak location. A random forest classifier was tested for leak localization on two different sized benchmark networks. Investigation of the influence of sensor layout and number of sensors on model accuracy was conducted. Different prediction models were constructed for different sizes of leaks and for different ranges of demand uncertainty to estimate model accuracy. This approach allows for a large number of varying measurements to be simulated in a short amount of time, thus providing

relatively quick localization, which is suitable for use in real conditions. Additional model uncertainties such as pipe diameters, node elevations, etc. can easily be incorporated into the presented methodology.

The rest of the paper is organized as follows. In Section 2, the problem statement is defined with a description of the used benchmark water distribution networks and a description of the proposed methodology using a random forest classifier. In Section 3, results are presented for both benchmark networks investigating the influence of a different number of prediction model inputs and features, of different ranges of demand uncertainties and leak sizes, and of sensor layout on model accuracy. Additionally, an example of the application of the prediction model is presented. In Section 4, the main observations regarding the obtained results are presented with proposed further research. In Section 5, final remarks are presented.

2. Materials and Methods

2.1. Problem Statement

Model-based leakage detection methods rely on residuals obtained as a difference between measured and expected results from the simulation of a calibrated water distribution network model. Unfortunately, water distribution models used for simulation typically have estimated nodal demands, which greatly influences the accuracy of residual values, hence resulting in modeling errors that are the main drawback of this approach. Additionally, if sensitivity analysis is used with nominal leak values, further uncertainties are introduced. The basic premise of the currently proposed methodology is that the prediction model can be constructed from a large database of simulated measurement data, which should describe a variety of possible leak scenarios. Consequently, if a considerable amount of data is generated, with a set range of considered uncertainties, it is reasonable to assume that the real measurements can be determined from simulated events with randomly chosen leak parameters by the prediction model.

Leak scenarios were simulated using EPANET2 version 2.0.12. [33]. Simulation results were obtained with randomly chosen leak locations, leak size, and random demands of end users. Basic assumptions used in this work are that leaks can occur only in network nodes and that a single leak is present in the water distribution network. Sensor measurements are considered ideal. The used water distribution network models were considered to be calibrated, i.e., pipe diameter and roughness were considered to be known and well-calibrated. However, these uncertainties can easily be incorporated into the data generation stage and further investigation of these uncertainties is to be evaluated in future work.

The prediction model was constructed using raw pressure sensor measurements obtained every 15 min for a period of 24 h where different ranges of base demand variation were investigated. Although in a real case scenario base demands vary greatly on daily basis, several prediction models can be created with characteristic demand patterns, e.g., one for summer weekdays, one for winter weekends, etc. Additionally, a prediction model can be created specifically for night scenarios where smaller demand variations occur. This methodology is already used for leak detection when differences in flows are measured during the night to detect if the leak is present in the network. Prediction model random forest implementation in the Python library Scikit-learn [34] version 0.20.3 was used. Data generation and prediction model training were performed using the supercomputing resources at the Center for Advanced Computing and Modelling, University of Rijeka.

2.2. Benchmark Water Supply Networks

Prediction of the leak location was conducted on two differently sized benchmark networks. The investigated networks are the Hanoi (Vietnam) network with 31 nodes, obtained from The Centre for Water Systems (CWS) at the University of Exeter [35], and the Net3 EPANET2 example consisting of 92 nodes. Both benchmark networks were considered to be calibrated. To achieve unsteady simulation, demand patterns for the Hanoi network were taken as in [26] and are presented in the Figure 1. For the Hanoi

network, two pressure sensors were placed at nodes 14 and 30, as depicted in [24]. In the Net3 network, two different sensor layouts were considered, with four pressure sensors placed at network nodes 117, 143, 181, and 213, and for the second layout, four sensors were placed at network nodes 115, 119, 187, and 209. The considered networks with sensor placements can be seen in Figures 2 and 3.
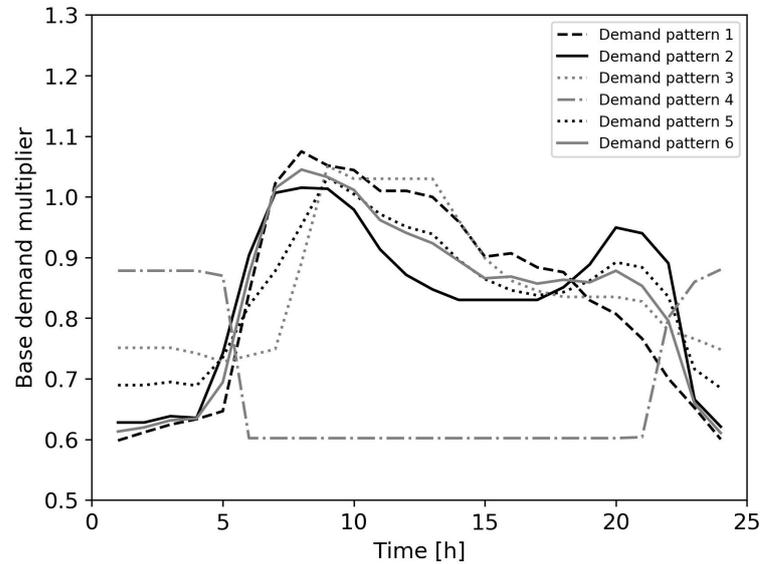
Figure 1. Hanoi network pattern demands.
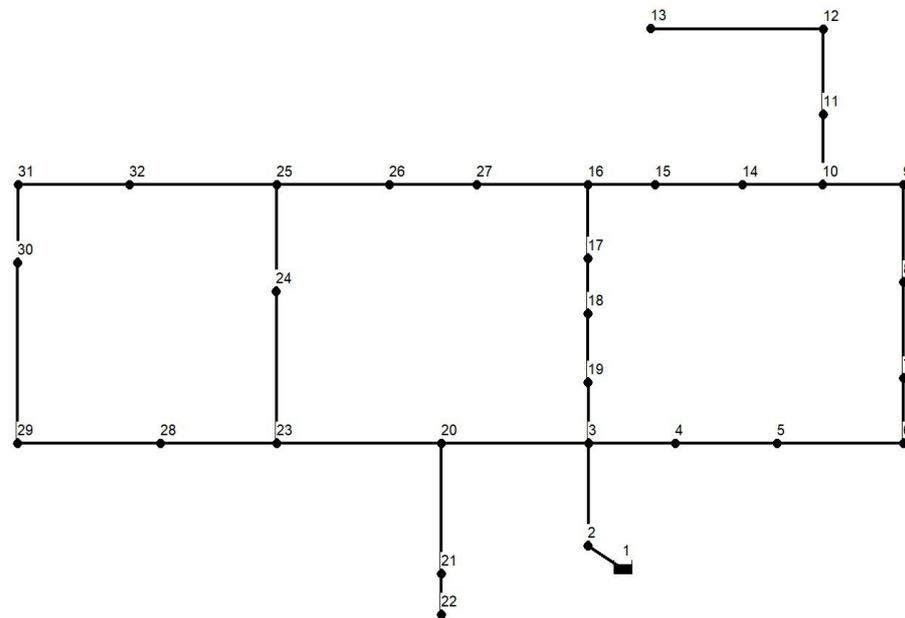
Demand pattern	Nodes
1	4, 5, 6, 7, 8, 9, 10, 14, 15
2	20, 26, 27
3	28, 29, 30, 31, 32
4	11, 12, 13, 21, 22
5	16, 17, 18, 19
6	23, 24, 25

Figure 2. Hanoi network with pattern demands in nodes.
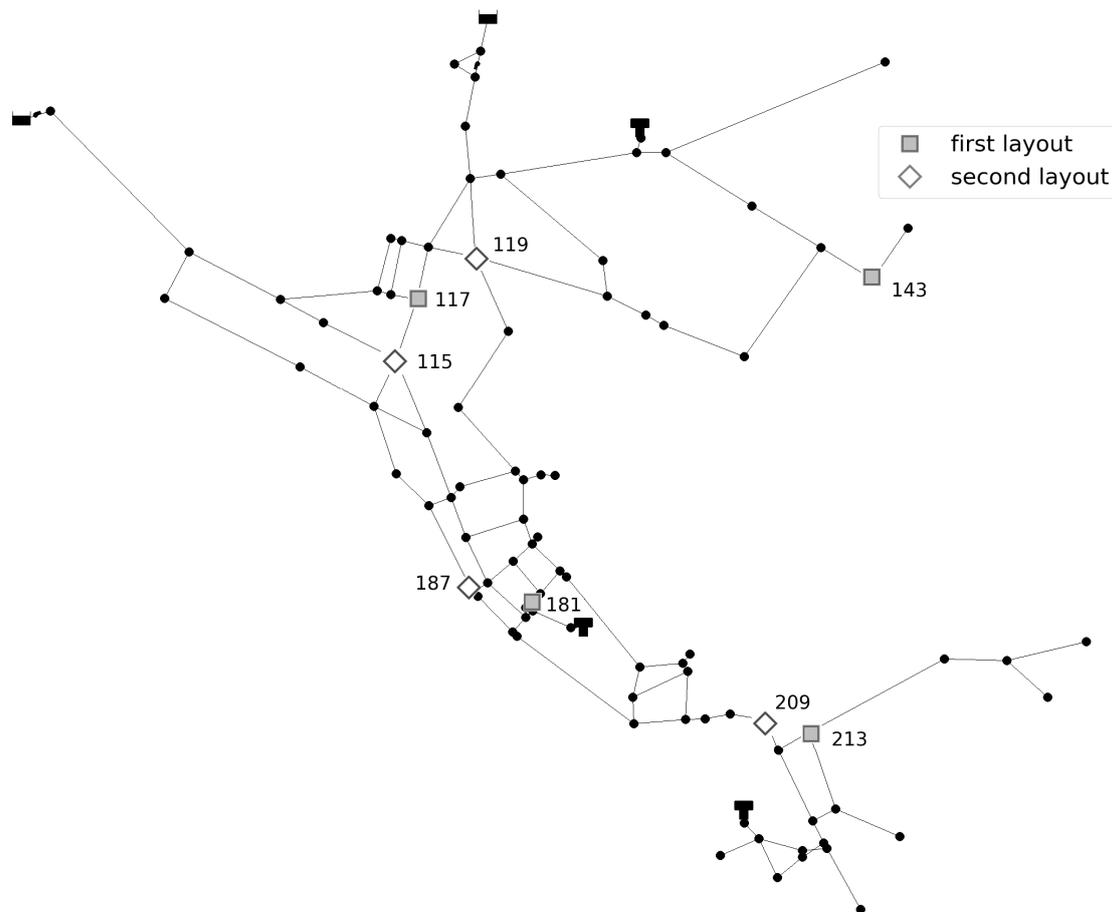
Figure 3. Net3 network with two sensor layouts.

For the Hanoi network, simulation time was 24 h with a hydraulic time step of 1 h and report time step of 15 min. For the Net3 network, simulation time was 24 h with a hydraulic time step of 10 min and report time step of 15 min. To obtain data for the machine learning model, leak scenarios were simulated using different emitter coefficients on randomly chosen leak nodes. For both networks, all network nodes were assumed as a potential location of the burst. The first dataset was constructed with no variation of base demand and only leak location and emitter coefficients were varied. Different ranges of emitter coefficients were considered, ranging from 5 to 15. To consider the variation of base demand, first, it was randomly chosen whether base demand is to be altered or not. If the base demand was to be altered it was randomly increased or decreased by randomly chosen percentages of 2.5, 5, 10, 15, and 20%. Scenarios with no base demand variation were considered to investigate the influence of different ranges of emitter coefficients on prediction model accuracy.

2.3. Random Forest Classifier

Machine learning algorithms build a model on sample data where the underlying correlation in the data is found and a prediction can be made for a new set of inputs. Machine learning algorithms can be divided into regression and classification, where regression provides information about continuous output values, whereas classification algorithms return discrete values, i.e., class labels. Since the problem considered in this paper is a classification problem, the machine learning classifier random forest was used. The random forest algorithm introduced by Breiman [36] is an ensemble learning algorithm that consists of multiple decision trees where each decision tree is trained independently on a random subset of data. Bootstrapping ensures that each decision tree in the random forest has a different subset of the training data, providing unique decision trees. Followed

by aggregation, a classification with the most occurrences is chosen by the random forest and is considered as the class prediction.

Random forest parameters used in this study were chosen with the grid search hyperparameter optimization method, which was conducted to optimize the number of estimators (trees), maximum depth, and a minimum number of samples required to split an internal node while other parameters were kept constant. The Net3 network with four sensors placed at network nodes 117, 143, 181, and 213, an emitter coefficient ranging from 5 to 15 and with no demand uncertainty was considered for the hyperparameter optimization method. Resulting machine learning parameters chosen for further study include 200 estimators, a maximum depth of 60 and a minimum number of samples required to split an internal node equal to 2. Obtained data were split into 70% for learning and 30% for model testing. A flowchart of the proposed method can be seen in Figure 4.
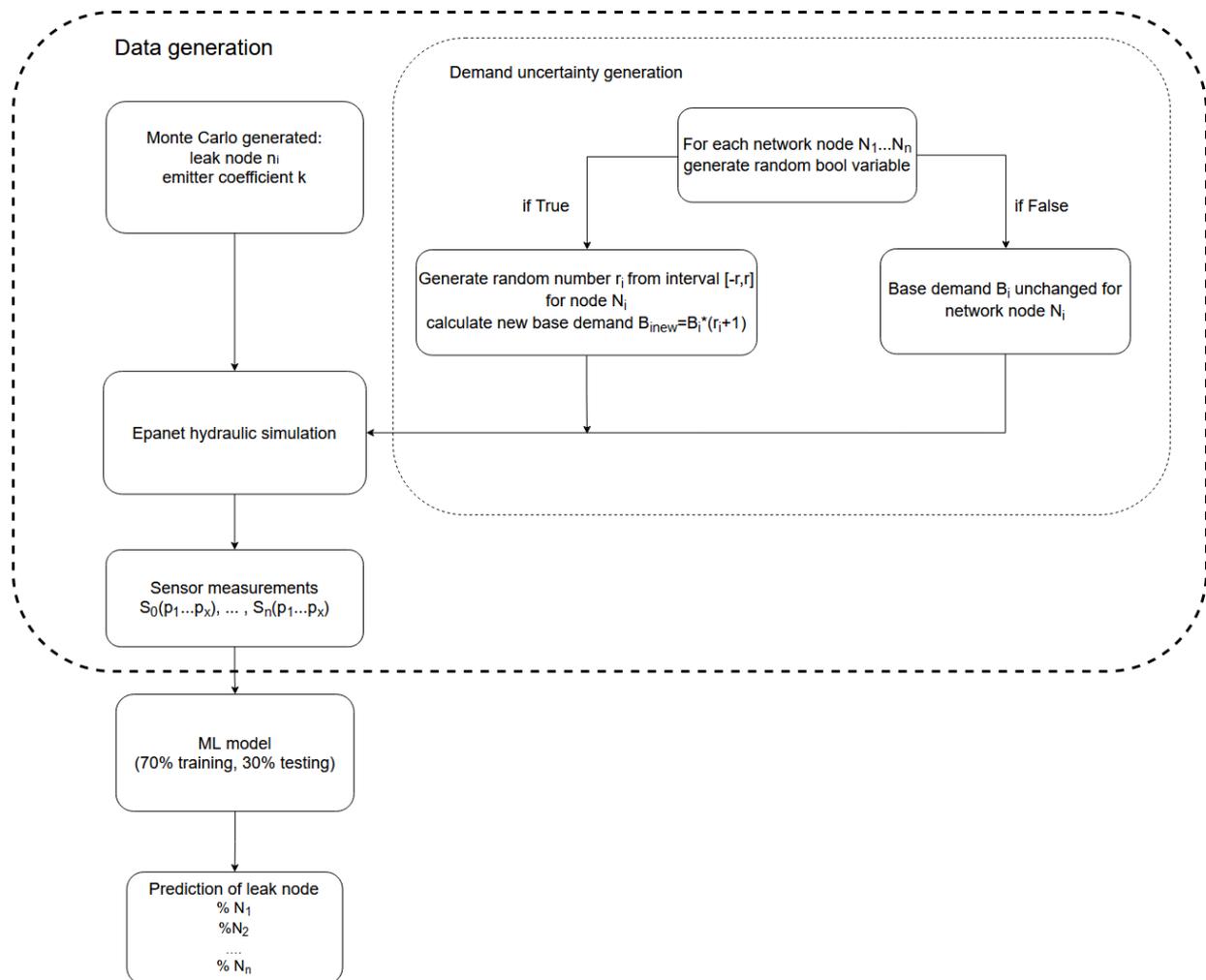
Figure 4. Flowchart of data generation and the machine learning algorithm used for the prediction of leak location.

3. Results

3.1. Data Influence

For both networks, the influence of the number of data inputs was investigated when only the emitter coefficient varied with no change in base demand. The emitter coefficient was chosen to be in a range from 5 to 15. For the Hanoi network with 100,000 inputs, 100% accuracy was achieved. For the Net3 network with the first sensor layout, results can be seen in the Table 1. For each model, 10 runs were conducted with a random training–test split to consider the influence of the random seed. Standard deviation ranged from 0.17%

for a model with 100,000 inputs to 0.03% for 500,000 inputs. It can be observed that the accuracy of the model was 98% for 500,000 inputs. Thus, all further results are with models with 500,000 inputs, and due to the small standard deviation, the presented results were calculated as an average of 5 runs.

Table 1 additionally includes results for the top three nodes with the greatest probability of being the true leak node. As evidenced by the results, an accuracy of 99% was achieved with merely 200,000 inputs. Considering the top three nodes can be greatly beneficial for big networks with dense network node placement where a small distance between network nodes is present. The prediction model can successfully localize leak location, where further procedures can be used to exactly detect which network node is a true leak location.

Table 1. Influence of data inputs on model accuracy without base demand variation for the Net3 network with an emitter coefficient range of 5–15.

Data Inputs	100,000	200,000	300,000	450,000	500,000
Accuracy	88%	93%	96%	97%	98%
Top 3	98%	99%	99%	99%	99%

3.2. Variation of Base Demand and Emitter Coefficient

For the Hanoi network, the influence of variation of base demand was investigated for the model with an emitter coefficient range of 10–15. Results are presented in Table 2. When demand variation was ±2.5%, the model accuracy was above 80%. It can be observed that with the greater demand variation for the same number of inputs, the model accuracy considerably decreased; however, if the top three and five nodes were considered, model accuracy greatly increased, where for the top five nodes accuracy was above 90% for the models with a demand variation of up to ±15%. It is important to note, however, that the top five nodes for such a small network do not provide a considerable localization, and thus further study of this approach must be conducted on larger networks.

Table 2. Influence of base demand variation on model accuracy for the Hanoi network with an emitter coefficient range of 10–15 with 500,000 inputs.

Base Demand Variation	±2.5%	±5%	±10%	±15%	±20%
Accuracy	82%	69%	57%	49%	44%
Top 3	98%	93%	86%	81%	76%
Top 5	99%	98%	95%	92%	89%

Results for the Net3 network with variations of base demand for the emitter coefficient range of 10–15 are presented in the Table 3. It can be observed that for the same base demand variation and same emitter coefficient range, the model accuracy for the Net3 network decreased by roughly 20% when compared to the Hanoi network. This is to be expected, since the Net3 network has a considerably larger number of network nodes. However, it is evident that the approach that considers the top three and five network nodes significantly increased model accuracy, which would make it possible to successfully localize leak location even in large networks. The results indicate that for greater variation in base demand more data inputs are needed when considering large networks.

Table 3. Influence of base demand variation on model accuracy for the Net3 network with an emitter coefficient range of 10–15 with 500,000 inputs.

Base Demand Variation	$\pm2.5\%$	$\pm5\%$	$\pm10\%$
Accuracy	62%	49%	36%
Top 3	92%	80%	65%
Top 5	98%	90%	77%

For the investigation of emitter coefficient variation a $\pm2.5\%$ base demand variation was chosen, and the results for the Hanoi network are presented in Table 4. It can be observed that for the smaller emitter coefficient values, model accuracy was considerably smaller than for cases with greater emitter coefficient values. This is to be expected, since a greater emitter coefficient value represents a greater leak where a greater discrepancy in sensor measurements is present, which is easier to detect with the prediction model. Additionally, when the emitter coefficient range was narrowed down from 10 (emitter coefficient range 5–15) to 5 (emitter coefficient range 10–15) it was also observed that model accuracy increased. This is also to be expected since the smaller emitter coefficient range has a smaller number of leak combinations and the same number of inputs better describes the prediction model in that case.

Table 4. Influence of emitter coefficient variation on model accuracy for the Hanoi network with a demand variation of $\pm2.5\%$ with 500,000 inputs.

Emitter Coefficient Range	1–5	5–10	5–15	10–15
Accuracy	38%	67%	71%	82%
Top 3	67%	92%	93%	98%
Top 5	81%	98%	98%	99%

Results for the Net3 network are presented in the Table 5. The accuracy of the model, similarly, decreased when smaller emitter coefficients were used. However, it is interesting to observe that for greater emitter coefficient values the difference in model accuracy between the two models increased—e.g., for the emitter coefficient range of 1–5, both models had low accuracy with a difference around 6%. When the emitter coefficient was in the range of 10–15, model accuracy increased with the Hanoi network yielding improved accuracy by around 20% when compared to the Net3 network. When considering the top five leak candidates, both models achieved high accuracy.

Table 5. Influence of emitter coefficient variation on model accuracy for the Net3 network with a demand variation of $\pm2.5\%$ with 500,000 inputs.

Emitter Coefficient Range	1–5	5–10	5–15	10–15
Accuracy	32%	51%	52%	62%
Top 3	59%	83%	84%	92%
Top 5	72%	92%	94%	98%

3.3. Sensor Layout Influence

The influence of sensor layout was investigated for the Net3 network where two different sensor layouts with four sensors and two different sensor layouts with 2 sensors were considered. Results are presented in the Table 6. It has been shown that with no demand variation all sensor layouts achieved exceptional accuracy. When demand variation was introduced, sensor layout slightly influenced model accuracy. Smaller number of sensors led to a reduction in model accuracy of around 10%. This is to be expected and indicates that for greater model accuracy a greater number of sensors should be used.

Table 6. Influence of sensor layout on model accuracy for the Net3 network for an emitter coefficient range of 5–10 with 500,000 inputs.

Sensor Locations		Demand Variation		
		No Variation	±2.5%	±5%
117, 143, 181, 213	Accuracy	98%	51%	37%
	Top 3	99%	83%	69%
	Top 5	99%	92%	79%
117, 181	Accuracy	96%	41%	27%
	Top 3	99%	71%	52%
	Top 5	99%	83%	64%
115, 119, 187, 209	Accuracy	98%	54%	37%
	Top 3	99%	84%	70%
	Top 5	99%	94%	83%
119, 209	Accuracy	97%	40%	27%
	Top 3	99%	71%	53%
	Top 5	99%	85%	67%

3.4. Feature Influence

To investigate the influence of a number of features, two different report time steps were considered. For all prediction models and for both water distribution networks, a report time step of 15 min was used, resulting in 97 features per sensor for each leak scenario. For the Hanoi network with two sensors, this resulted in 194 features, and for the Net3 network with four sensors, this resulted in 388 features. In [18], it was reported that sampling data typically vary between 1 min and 15 min; however, to reduce prediction model complexity, for both the Hanoi and Net3 networks simulations were conducted for an emitter coefficient range of 10–15, with ±2.5% demand variation, with a report time step of 1 h resulting in 25 features per sensor per leak scenario. Results are presented in the Table 7. It is evident that with a smaller number of features, model accuracy slightly decreased. This indicates that prediction models with a greater number of inputs but with a smaller number of features should be investigated to see if greater accuracy could be achieved with the same computational expense.

Table 7. Influence of number of features on model accuracy for the Hanoi and Net3 networks with an emitter coefficient range of 10–15 and a demand variation of ±2.5% with 500,000 inputs.

	Hanoi Network		Net3 Network	
Number of Features	194	50	388	100
Accuracy	82%	81%	62%	60%
Top 3	98%	98%	92%	91%
Top 5	99%	99%	98%	97%

3.5. Application of the Prediction Method

To investigate the possibility of application of the proposed method on real case events, 30 simulations were conducted to simulate daily measurements during a one month period for the same leak location and same leak emitter coefficient. The Hanoi network was chosen with a leak at network node 26 and with an emitter coefficient value of 10. For each simulation, if the network node was chosen to be altered, demand was randomly changed in the range of ±10%. In this way, daily demand variation was simulated. The machine learning model for the Hanoi network, with an emitter coefficient range of 10–15 and with a demand variation of ±10% with a model accuracy 57% was used to predict leak location. Results of the predictions can be observed in Figure 5. It can be seen that for the majority

of days (16 out of 30) true leak location was successfully detected, which roughly matches the overall model accuracy. For the remaining days, adjacent network nodes were detected as leak locations. This shows that the proposed methodology can be successfully used to approximately localize and detect leak location.
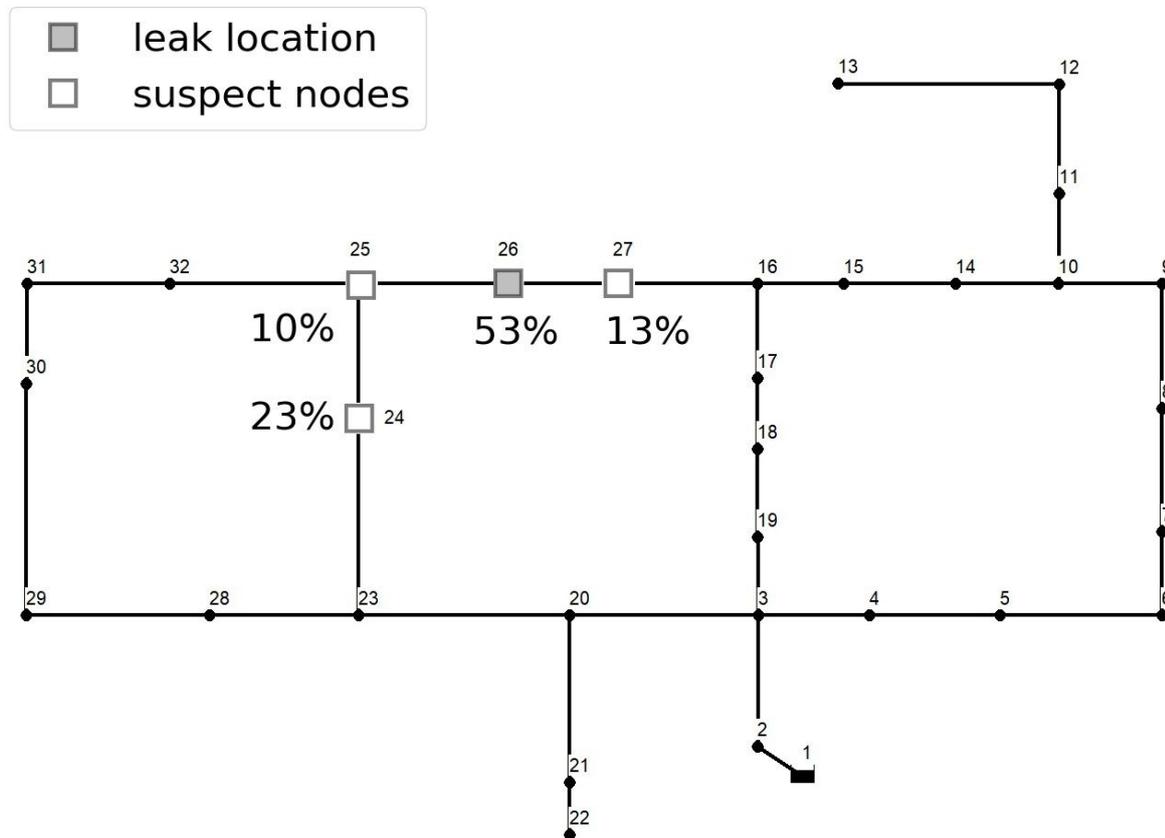

Figure 5. Prediction of leak location for 30 day measurements with percentage of predicted leak nodes.

4. Discussion

Based on the presented results, it can be concluded that the proposed methodology can be successfully applied to small-sized and medium-sized networks. With the increase in network size, model accuracy considerably decreases. It is important to note, however, that this behavior is not unexpected as the same assumptions and amount of data (i.e., inputs) are utilized for small and larger cases. Despite this, meaningful results and adequate localization can be achieved when the top three and five network nodes based on leak location probability are considered. This indicates that leak nodes can be localized on a more general water distribution network with one prediction model, where exact location can be detected if coupled with another prediction model that focuses on a specific network zone or employs a different leak localization methodology. Both approaches should be further investigated.

It can be observed that the greatest accuracy was achieved for prediction models trained with smaller demand variation and greater leak coefficients. This is expected since in the case of no demand variation, 500,000 simulations provide a considerable amount of combinations of leak events, where the prediction model simply chooses from the most similar event. When demand variation is introduced, model accuracy decreases as the demand variation range increases. However, several prediction models can be constructed with different demand patterns, e.g., a night demand model, a workday demand model, etc., where in case of a leak event, the prediction model with the most similar demand pattern can be chosen for leak localization. When considering leak coefficients, independent

prediction models can also be built; for example, prediction models to detect small, medium, and large leaks. The greatest accuracy is achieved for the large leaks, which can be greatly beneficial in the case of large bursts in the water distribution network. This kind of events needs quick intervention since the water supply to end users is usually interrupted until the burst is repaired. A prediction model can indicate leak location so rapid intervention can be achieved.

Another potential issue stems from the fact that the calibrated model relies on data that might already incorporate leaks. Consequently, predominantly new leaks can be predicted, as existing leaks are incorporated in the calibrated model itself. As existing leaks become larger with time and due to the material deterioration, older leak locations can eventually be detected as well, although they would appear as a comparatively smaller leak than they actually are; this is not a crucial problem, however. This drawback can be mitigated by coupling or employing as standalone older calibrated models that predate the current one.

The study of the influence of the report time step indicates that with a smaller amount of features, similar accuracy can be achieved, and thus a greater number of inputs can be considered to achieve better model accuracy. The optimum number of features and inputs should be further investigated to provide the best accuracy and model complexity ratio. This is especially important if the proposed methodology is to be used on more complex water distribution networks. This approach is valid if existing leaks that are undetected for a longer period of time are to be found. However, in the case of a pipe burst event, the prediction model with a smaller time step should be considered to reduce reaction time in case of the event. This should be further investigated since larger pipe bursts considerably change water distribution network dynamics and the measurement period should be considerably smaller than one day (as is in the current paper) to provide rapid reaction. Additionally, techniques for data dimensionality reduction should be explored to possibly reduce the model complexity.

The sensor layouts considered in this paper can be considered sparse. Improvement in prediction model accuracy can be achieved if additional sensors are installed in the water distribution network. Additionally, a combination of pressure and flow sensors should be investigated since additional data could be beneficial to the model and water distribution networks have a combination of both types of sensors. Further study should be focused on investigating other classification models, for example K-NN and ANN, which were successfully applied in previous literature using model-based approaches, which could possibly provide greater model accuracy. The coupling of multiple prediction models should also be investigated, where one model would provide coarse leak localization and the second model would provide the exact location of the leak. Moreover, future studies should account for uncertainties such as pipe diameter and pipe roughness with the methodology tested on real water distribution networks.

Although computationally demanding, the proposed methodology with introduced randomness can successfully describe a wide range of operating conditions, thus providing a considerable amount of data that cannot be obtained from field measurements. With growing computational power, the proposed methodology could be successfully utilized, as once they are generated, prediction models can be employed to evaluate a network with a considerably lower amount of computational resources and time.

5. Conclusions

In this paper, a machine learning approach was presented that helps identify leak locations based on pressure sensor measurements. A random forest classifier is used for small-sized and medium-sized benchmark networks. The presented results show that the proposed methodology can be successfully used for leak localization using data obtained from numerical simulations even for sparse sensor placement. The discrepancy between synthetic data obtained from numerical simulations and real data can be compensated for with randomness in the model simulation. Using Monte Carlo random parameters

of leak events and demands, a significant amount of data can be obtained, which can be successfully used for building a machine learning prediction model.

Our main findings include:

- Greatest prediction model accuracy was achieved for the largest leaks, with the smallest demand variation. With the increase in demand variation, prediction model accuracy considerably decreased.
- Model accuracy increased significantly when the top three and five network nodes with the greatest certainty of being leak nodes were considered to narrow down the leak location region.
- Investigation of the application of the proposed methodology on a small-sized network showed that in the majority of records, true leak location was detected, where in other cases neighbor nodes were chosen.

The obtained results indicate that the proposed methodology could be successfully applied to real water distribution networks; however further study should include the following:

- Investigation of a greater number of inputs should be conducted to increase model accuracy under greater demand variation, or multiple prediction models should be used for different demand ranges.
- Validation of the proposed methodology should be conducted on real water distribution networks.
- Randomness should be incorporated into other model uncertainties, such as pipe diameter and pipe roughness.
- Further investigation should be conducted to explore other algorithms with an increased number of inputs and an optimized number of features to further increase model accuracy.

Author Contributions: Conceptualization, I.L., B.L. and A.S.; data curation, I.L.; formal analysis, I.L.; investigation, I.L. and B.L.; methodology, I.L., B.L. and A.S.; resources, Z.Č.; software, I.L.; supervision Z.Č.; validation, I.L; visualization; I.L.; writing—original draft, I.L. and A.S.; writing—review and editing, B.L., Z.Č. and A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jacobsz, S.W.; Jahnke, S.I. Leak detection on water pipelines in unsaturated ground by discrete fiber optic sensing. *Struct. Health Monit.* **2020**, *19*, 1219–1236. [CrossRef]
2. Nkemeni, V.; Mieyeville, F.; Tsafack, P. A Distributed Computing Solution Based on Distributed Kalman Filter for Leak Detection in WSN-Based Water Pipeline Monitoring. *Sensors* **2020**, *20*, 5204. [CrossRef]
3. Wu, Y.; Liu, S.; Wu, X.; Liu, Y.; Guan, Y. Burst detection in district metering areas using a data driven clustering algorithm. *Water Res.* **2016**, *100*, 28–37. [CrossRef] [PubMed]
4. Rajeswaran, A.; Narasimhan, S.; Narasimhan, S. A graph partitioning algorithm for leak detection in water distribution networks. *Comput. Chem. Eng.* **2018**, *108*, 11–23. [CrossRef]
5. Cody, R.A.; Dey, P.; Narasimhan, S. Linear prediction for leak detection in water distribution networks. *J. Pipeline Syst. Eng. Pract.* **2020**, *11*, 04019043. [CrossRef]
6. Bohorquez, J.; Alexander, B.; Simpson, A.R.; Lambert, M.F. Leak detection and topology identification in pipelines using fluid transients and artificial neural networks. *J. Water Resour. Plan. Manag.* **2020**, *146*, 04020040. [CrossRef]
7. Arifin, B.; Li, Z.; Shah, S.L.; Meyer, G.A.; Colin, A. A novel data-driven leak detection and localization algorithm using the Kantorovich distance. *Comput. Chem. Eng.* **2018**, *108*, 300–313. [CrossRef]

8. Wang, F.; Lin, W.; Liu, Z.; Wu, S.; Qiu, X. Pipeline leak detection by using time-domain statistical features. *IEEE Sens. J.* **2017**, *17*, 6431–6442. [CrossRef]

9. Lang, X.; Li, P.; Hu, Z.; Ren, H.; Li, Y. Leak detection and location of pipelines based on LMD and least squares twin support vector machine. *IEEE Access* **2017**, *5*, 8659–8668. [CrossRef]

10. Zadkarami, M.; Shahbazian, M.; Salahshoor, K. Pipeline leakage detection and isolation: An integrated approach of statistical and wavelet feature extraction with multi-layer perceptron neural network (MLPNN). *J. Loss Prev. Process. Ind.* **2016**, *43*, 479–487. [CrossRef]

11. Adegboye, M.A.; Fung, W.K.; Karnik, A. Recent advances in pipeline monitoring and oil leakage detection technologies: Principles and approaches. *Sensors* **2019**, *19*, 2548. [CrossRef]

12. Blesa, J.; Nejjari, F.; Sarrate, R. Robust sensor placement for leak location: Analysis and design. *J. Hydroinform.* **2015**, *18*, 136–148. [CrossRef]

13. Soldevila, A.; Blesa, J.; Tornil-Sin, S.; Fernandez-Canti, R.M.; Puig, V. Sensor placement for classifier-based leak localization in water distribution networks using hybrid feature selection. *Comput. Chem. Eng.* **2018**, *108*, 152–162. [CrossRef]

14. Khorshidi, M.S.; Nikoo, M.R.; Taravatrooy, N.; Sadegh, M.; Al-Wardy, M.; Al-Rawas, G.A. Pressure sensor placement in water distribution networks for leak detection using a hybrid information-entropy approach. *Inf. Sci.* **2020**, *516*, 56–71. [CrossRef]

15. Raei, E.; Shafiee, M.E.; Nikoo, M.R.; Berglund, E. Placing an ensemble of pressure sensors for leak detection in water distribution networks under measurement uncertainty. *J. Hydroinform.* **2019**, *21*, 223–239. [CrossRef]

16. Wu, Y.; Liu, S. A review of data-driven approaches for burst detection in water distribution systems. *Urban Water J.* **2017**, *14*, 972–983. [CrossRef]

17. Chan, T.K.; Chin, C.S.; Zhong, X. Review of current technologies and proposed intelligent methodologies for water distributed network leakage detection. *IEEE Access* **2018**, *6*, 78846–78867. [CrossRef]

18. Zaman, D.; Tiwari, M.K.; Gupta, A.K.; Sen, D. A review of leakage detection strategies for pressurised pipeline in steady-state. *Eng. Fail. Anal.* **2020**, *109*, 104264. [CrossRef]

19. Zhou, M.; Pan, Z.; Liu, Y.; Zhang, Q.; Cai, Y.; Pan, H. Leak Detection and Location Based on ISLMD and CNN in a Pipeline. *IEEE Access* **2019**, *7*, 30457–30464. [CrossRef]

20. Pérez-Pérez, E.; López-Estrada, F.; Valencia-Palomo, G.; Torres, L.; Puig, V.; Mina-Antonio, J. Leak diagnosis in pipelines using a combined artificial neural network approach. *Control Eng. Pract.* **2021**, *107*, 104677. [CrossRef]

21. Mounce, S.; Boxall, J.; Machell, J. Development and verification of an online artificial intelligence system for detection of bursts and other abnormal flows. *J. Water Resour. Plan. Manag.* **2010**, *136*, 309–318. [CrossRef]

22. Jensen, T.N.; Puig, V.; Romera, J.; Kallesøe, C.S.; Wisniewski, R.; Bendtsen, J.D. Leakage localization in water distribution using data-driven models and sensitivity analysis. *Ifac Pap.* **2018**, *51*, 736–741. [CrossRef]

23. Zhang, Q.; Wu, Z.Y.; Zhao, M.; Qi, J.; Huang, Y.; Zhao, H. Leakage zone identification in large-scale water distribution systems using multiclass support vector machines. *J. Water Resour. Plan. Manag.* **2016**, *142*, 04016042. [CrossRef]

24. Soldevila, A.; Blesa, J.; Tornil-Sin, S.; Duviella, E.; Fernandez-Canti, R.M.; Puig, V. Leak localization in water distribution networks using a mixed model-based/data-driven approach. *Control Eng. Pract.* **2016**, *55*, 162–173. [CrossRef]

25. Soldevila, A.; Fernandez-Canti, R.M.; Blesa, J.; Tornil-Sin, S.; Puig, V. Leak localization in water distribution networks using Bayesian classifiers. *J. Process Control* **2017**, *55*, 1–9. [CrossRef]

26. Quiñones-Grueiro, M.; Verde, C.; Prieto-Moreno, A.; Llanes-Santiago, O. An unsupervised approach to leak detection and location in water distribution networks. *Int. J. Appl. Math. Comput. Sci.* **2018**, *28*, 283–295. [CrossRef]

27. Zhou, X.; Tang, Z.; Xu, W.; Meng, F.; Chu, X.; Xin, K.; Fu, G. Deep learning identifies accurate burst locations in water distribution networks. *Water Res.* **2019**, *166*, 115058. [CrossRef] [PubMed]

28. Sun, C.; Parellada, B.; Puig, V.; Cembrano, G. Leak localization in water distribution networks using pressure and data-driven classifier approach. *Water* **2020**, *12*, 54. [CrossRef]

29. Javadiha, M.; Blesa, J.; Soldevila, A.; Puig, V. Leak localization in water distribution networks using deep learning. In Proceedings of the 2019 6th International Conference on Control, Decision and Information Technologies (CoDIT), Paris, France, 23–26 April 2019; pp. 1426–1431.

30. Soldevila, A.; Blesa, J.; Fernandez-Canti, R.M.; Tornil-Sin, S.; Puig, V. Data-driven approach for leak localization in water distribution networks using pressure sensors and spatial interpolation. *Water* **2019**, *11*, 1500. [CrossRef]

31. Grbčić, L.; Lučin, I.; Kranjčević, L.; Družeta, S. Water supply network pollution source identification by random forest algorithm. *J. Hydroinform.* **2020**, *22*, 1521–1535. [CrossRef]

32. Lučin, I.; Grbčić, L.; Čarija, Z.; Kranjčević, L. Machine-Learning Classification of a Number of Contaminant Sources in an Urban Water Network. *Sensors* **2021**, *21*, 245. [CrossRef] [PubMed]

33. Rossman, L.A. EPANET 2: Users Manual. 2000. Available online: https://epanet.es/wp-content/uploads/2012/10/EPANET_User_Guide.pdf (accessed on 6 September 2020).

34. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

35. Centre for Water Systems, University of Exeter. Benchmarks. Available online: http://emps.exeter.ac.uk/engineering/research/cws/downloads/benchmarks/ (accessed on 6 November 2020).

36. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]

Detailed Leak Localization in Water Distribution Networks Using Random Forest Classifier and Pipe Segmentation

IVANA LUČIN[1,2], ZORAN ČARIJA[1,2], SINIŠA DRUŽETA[1,2], AND BOŽE LUČIN[1,3]

[1]Department of Fluid Mechanics and Computational Engineering, Faculty of Engineering, Faculty of Engineering, University of Rijeka, 51000 Rijeka, Croatia
[2]Center for Advanced Computing and Modelling, University of Rijeka, 51000 Rijeka, Croatia
[3]Flowtech d.o.o., 51000 Rijeka, Croatia

Corresponding author: Ivana Lučin (ilucin@riteh.hr)

ABSTRACT In this paper, a Random Forest classifier was used to predict leak locations for two differently sized water distribution networks based on pressure sensor measurements. The prediction model is trained on simulated leak scenarios with randomly chosen parameters - leak location, leak size, and base node demand uncertainty. Leak localization methods found in literature that rely on numerical simulations can only predict network nodes as leak nodes; however, since a leak can occur at any point along a pipe segment, additional spatial discretization of suspect pipe is proposed in this paper. It was observed that pipe segmentation of the whole network is a non-feasible approach since it rapidly increases the number of potential leak locations, consequently increasing the complexity of the prediction model. Therefore, a novel approach is proposed, in which a prediction model is trained on scenarios with leaks occurring in original network nodes only, but with its accuracy assessed against pressure sensor measurements from scenarios in which leaks occur in points between network nodes. It was observed that this approach can successfully narrow down the suspect leak area and, followed by additional segmentation of that network area and subsequent prediction, a precise leak localization can be achieved. The proposed approach enables incorporation of various uncertainties by simulating leak scenarios under different conditions. Investigation of leak size uncertainty and base demand variation showed that several different scenarios can produce similar sensor measurements which makes it difficult to unambiguously determine leak location using the prediction model. Therefore, future approaches of coupling prediction modeling with optimization methods are proposed.

INDEX TERMS Leak localization, pipe segmentation, prediction modeling, random forest, water distribution networks.

I. INTRODUCTION

Leaks in water distribution networks can cause considerable losses, especially in older water distribution networks where considerable investments are needed for restoration. Smaller leaks can remain undetected for longer periods causing considerable water losses over time. Also, in the case of older water distribution networks rapid progression of leak size can eventually cause pipe burst which leads to water outages for end users. Therefore, a number of different techniques are being used to detect and localize leaks. These methods can be divided into hardware-based and software-based methods. Hardware-based methods use in situ visual observations

or measurements. Software-based methods rely on different software for leak detection analysis. Since some methods have been developed for specialized applications depending on the transporting fluid (water, oil, gas, etc.) and location of the pipeline (water distribution network, facility, housing, etc.), a number of papers analyzed the advantages and limitations of the proposed methods and an overview of some of these methods is given in papers [1]–[5].

Software-based methods can be further divided into transient-based methods, model-based methods, and data-driven methods. Transient-based methods rely on analysis of transient pressure wave that occurs when leakage happens. For model-based methods, estimated pressure values are obtained from simulation with no leaks and in-field measured pressure values are compared, i.e. subtracted from

The associate editor coordinating the review of this manuscript and approving it for publication was Ahmed Farouk.

estimated pressure values. Obtained residuals are evaluated and if residuals are above the chosen threshold it is considered that a leak is present. Data-driven methods rely on statistical analysis and processing of raw sensor measurement data to obtain information about the presence of leaks and possible locations.

The main problem with the model-based approach is the assumption of the model being a good representation of the network. Water distribution networks have a lot of uncertainties that need to be taken into consideration, such as demand uncertainties, sensor measurement imperfections, pipe diameter uncertainties, etc. Thus, the model-based approach cannot capture all these parameters. The data-driven approach using raw sensor measurements could incorporate all these variations, but the main problem is the number of leak events which are rather sparse. Since the amount of data is small compared to the amount of data needed for efficiently employing machine learning algorithms, models can be advanced by incorporating uncertainties through simulations with varying parameters which can produce additional data.

Machine learning has been used for a variety of water distribution system applications. Prediction of failure of water mains was investigated in [6] where artificial neural network (ANN), ridge regression, and ensemble decision tree were used. Different machine learning algorithms have been explored for the prediction of leak locations in pipelines, such as convolutional neural network (CNN) [7], [8] and ANN [9], [10]. In [11] support vector machine (SVM) method was used to predict leaks in wall-mounted pipelines.

When considering water distribution networks, in [12], a deep learning model based on additional pressure meters installed on optimal places was used to identify pipe burst locations. In [13], SVM was used for prediction of leak size and location based on pressure sensors gathered from EPANET simulations for small size leakages. In [14], leakage detection was conducted for 1500 m × 1500 m experimental network using principal component analysis (PCA) and SVM. In [15], model-based method was used to identify leak event and data-driven approach using k-Nearest Neighbors (k-NN) classifier was used in the second stage to determine leak location. In the further study [16] Bayesian classifier was used with improved localization accuracy. Both methods were applied to real water distribution network case studies. In [17], unsupervised principal component analysis (PCA) approach for leak detection was conducted for the Hanoi distribution network. In [18], Kriging method was used to estimate pressure measurements in the whole network based on the limited number of sensor measurements and classification methods were used to determine leak location. It was shown that the accuracy of the proposed method was very low for some sensor layouts due to Kriging interpolation error. In [19], detection and localization of multiple leak locations were explored. SVM was used as a classifier for leak detection using the residual method and a statistical method was used for leak localization in the Hanoi network.

All mentioned papers assume possible leak locations only in network nodes.

In order to increase the number of input data, in previous work [20] it was proposed that a great number of leak scenarios can be generated by simulating different leak locations and leak sizes under different demand uncertanties. The machine learning approach for leak localization was investigated for variously sized water distribution networks, various demand ranges, and various sensor placements. However, considerable simplification was made insomuch that the prediction model was trained with simulated scenarios in which leak locations occur only in network nodes while in reality leaks can occur at any point along a pipe segment. Thus, in this paper, an approach with pipe segmentation in suspect areas is investigated. The idea is taken from the adaptive mesh refinement approach used in computational fluid dynamics (CFD) simulations, where the area of interest is refined with additional numerical nodes in order to increase the accuracy of results. An alternative approach of fault zone identification has been used in work by [21] and [22]. However, that approach could be problematic for leak locations at the borders of leak zones since water distribution network needs to be divided into zones before using leak localization method. The approach proposed in this paper identifies suspect nodes from machine learning prediction model, which then serve as indicators for pipes that need to be further explored using segmentation. Therefore a possible leak area is adjusted for each leak event based on prediction results.

In the first part of this paper, it is investigated whether a prediction model trained only on simulations with leak locations in network nodes can successfully predict leaks that occur in-between network nodes. Two differently-sized water distribution networks, Hanoi and Net3 were used for this, coupled with various sensor layouts, leak sizes, and demands. Furthermore, the accuracy of sequential prediction models in predicting leak location was investigated. The prediction model performance is investigated when several most-suspect nodes are considered and segmentation of pipes near those suspect nodes is performed. The subsequent prediction model is trained on scenarios with leak locations in most-suspect network nodes and in nodes added through pipe segmentation from the previous stage. Limitations of the proposed method and future work are presented in the discussion section.

II. METHODOLOGY
A. PROBLEM STATEMENT
Leak localization methods based on machine learning methods require considerable amount of data for model training. Since the measurements for real leak events are rather sparse, additional data can be obtained by simulating different leak scenarios. For this purpose, leak scenarios were simulated using EPANET version 2.0.12. [23] with various leak scenario parameters. Leak location, leak size, and node demands were chosen randomly to cover a wide range of possible leak events. Typically it is assumed that water distribution network models are calibrated and that leaks can occur only
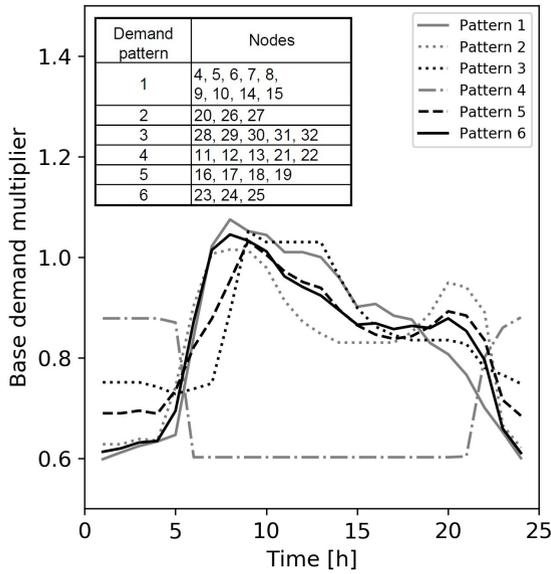
FIGURE 1. Demand patterns used for Hanoi network.

FIGURE 2. Hanoi network, original with indicated sensor locations (above), and after pipe segmentation with 5 segments per pipe (below).

in network nodes. The latter assumption can be problematic for water distribution networks with longer pipe segments since localization will be a very rough estimate. Therefore, additional pipe segmentation is introduced which divides a pipe into smaller sections, allowing better leak localization. Random Forest machine learning algorithm is trained with pressure sensor measurements from simulated scenarios and is then employed to determine most suspect leak locations.

B. WATER DISTRIBUTION NETWORKS

The investigated water distribution networks are small-sized Hanoi network and medium-sized Net3 network. Hanoi (Vietnam) network with 31 nodes was obtained from The Centre for Water Systems (CWS) at the University of Exeter [24]. For Hanoi network, demand patterns as described in [17] are adopted (Figure 1). Net3 network is an EPANET example network for dual-source system that changes over time, consisting of 92 nodes. For both networks simulation time was 24 h, hydraulic time step was 10 min and report time step 1 h. To generate a wide range of possible leak scenarios, emitter coefficient and leak location were chosen randomly. Additionally, to incorporate demand variation, it was randomly decided whether node base demand was to be changed or not. If it was chosen to be changed, base demand was increased or decreased by randomly chosen percentage in the range ±2.5% or ±5%.

For each water distribution network, two different sensor layouts were considered. For Hanoi network, the first layout has two sensors located in network nodes 14 and 30, as given in [15], and the second layout has three sensors located in network nodes 8, 20, and 31, as given in [17] (Figure 2). For Net3 network, the first layout has four sensors located in network nodes 117, 143, 181, and 213, and the second layout has two sensors located in network nodes 117 and 181 (Figure 3).
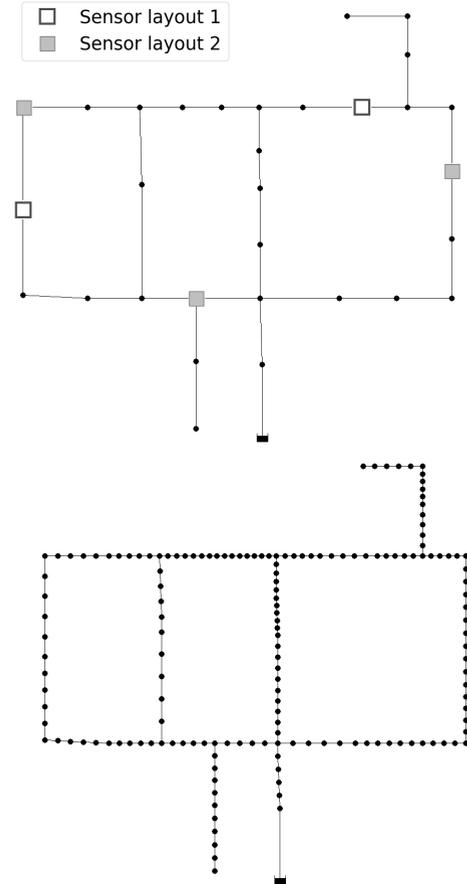
C. PIPE SEGMENTATION

Discretization of water distribution network pipes was achieved by inserting additional network nodes, where each pipe was split on 5 segments of equal length, resulting in additional 4 nodes for each pipe (Figure 2). Although it would be more beneficial to define a fixed segment length, a fixed number of segments was used as a methodological simplification.

To investigate machine learning efficiency in the localization of leak locations in pipe segments, three different models were analyzed. Model 1 was trained and tested on leak scenarios with leak locations in original network nodes. Model 2 was trained and tested on leak scenarios with leaks located both in network nodes and refinement nodes, resulting in a significantly increased number of ML output classes. Finally, Model 3 was trained on scenarios with leaks in original network nodes, but it was then tested for scenarios in which leak locations can be both in network nodes and refinement nodes.

Flowcharts of the proposed models can be observed in Figure 4. Depending on considered model leak node N_i is chosen from original network nodes $N_i \in \{N_0^o, \ldots, N_{no}^o\}$ where superscript o denotes original network nodes, or from original network nodes and additional nodes generated from
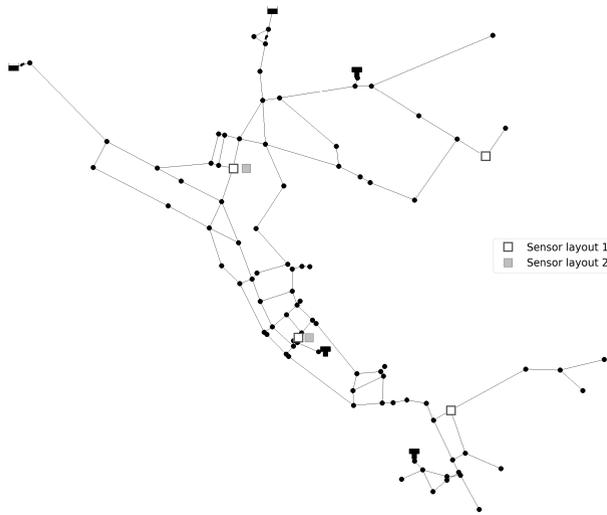
FIGURE 3. Net3 network with indicated sensor locations for considered sensor layouts.

segmentation $N_i \in \{N_0^o, \ldots, N_{no}^o, N_0^s, \ldots, N_{ns}^s\}$ where superscript s denotes segmentation nodes, subscript no denotes total number of original network nodes and ns total number of segmentation nodes. The sensor measurements $S_i \in \{S_0(t), \ldots, S_n(t)\}$, were n indicates total number of sensors for considered sensor layout, were recorded through time t, namely 25 timesteps in all considered cases. Since the model 3 is trained only on the original network nodes it cannot possibly predict a refinement node. Thus the refinement nodes are considered to be predicted correctly if their nearest original network node $N_i \in \{N_0^o, \ldots, N_{no}^o\}$ was predicted. This simulates a most realistic scenario where leaks can occur anywhere in the pipe segment, however, the model can be trained only with scenarios with leaks in network nodes we have in the model.

D. RANDOM FOREST CLASSIFIER
Machine learning (ML) algorithms are being used to find underlying correlations or patterns from obtained data. This ability enables machine learning algorithms to provide a prediction for unseen data, which can be categorized into regression and classification problems. Regression algorithms are designed to provide a prediction of the exact value of the output variable, while classification algorithms separate data into logical groups, i.e. classes.

Random Forest classifier was first proposed by [25] and is an ensemble type of algorithm based on multiple decision trees which are created as independent prediction models. Decision trees (DT) are constructed in a form of flowchart structure, where nodes represent attributes used for outcome prediction. Based on feature values a decision is made at each node and ultimately based on these decisions classification is reached. Each tree is defined with tree depth parameter which defines how many splits can be made before making a prediction. Random Forest uses bootstrap and aggregation methods to obtain unique data subsets for the training of each decision tree and to ultimately count the class with the most

predictions. Increased number of trees increases the precision of the classifier, albeit also increasing its complexity. The problem considered in this paper is the classification problem since each potential leak node represents one class, thus Random Forest classifier was adopted as a suitable ML method. Random Forest classifier implementation in the Python library Scikit-learn [26] version 0.20.3 was used.

The dataset is composed of 500 000 inputs, with training-testing split 70%-30%, resulting in 350 000 training records and 150 000 testing records. It was observed in [20] that a smaller timestep only slightly increases prediction accuracy so timestep of 1 hr was adopted in order to reduce number of features and reduce computational time.

Grid search optimization of Random Forest parameters was conducted for Hanoi network with 100 000 inputs with leak coefficient range $10 \ldots 15$ and with $\pm 2.5\%$ demand variation in order to find optimal number of estimators (trees), maximum depth, and minimum number of samples required to split an internal node. It was found that the optimal minimum number of samples required to split an internal node is 2, the optimal maximum depth of the tree is 20, and the optimal number of estimators (i.e. trees) is 200. These parameters are kept constant for all investigated prediction models. Other Random Forest parameters were kept at default values of the Scikit-learn implementation. For each prediction model, five runs were conducted to consider the influence of prediction model parameter randomness and average accuracy values are reported. Additionally, model accuracy was measured for true leak node presence in top 3 and top 5 suspect network nodes with greatest prediction certainties. Even if true leak node is not correctly predicted, presence of true leak node in top 3 or top 5 most suspect nodes considerably narrows down the area of leak location.

III. RESULTS
A. EFFECT OF PIPE SEGMENTATION
Hanoi network with two sensors, emitter coefficient range $10 \ldots 15$ and different demand variations was investigated first. In Model 1, where leaks can occur only in the original network nodes, 31 prediction classes were obtained. For Model 2 each pipe segment is divided into 5 segments of equal length, resulting in 163 prediction classes. Although Model 3 was used for predicting leak scenarios on segmented network of Model 2, it was trained on leak scenarios used for Model 1. Thus, in Model 3 the 31 prediction classes corresponding to the original network nodes were used, with leaks in the 135 segmentation nodes expected to be classified as leaks in their nearest original network nodes.

Results for the conducted investigation are presented in Table 1. It can be observed that with the increase of demand variation, model accuracy considerably decreases; indicating a rapid increase of possible scenarios which are consequently difficult to predict. However, when top 3 and top 5 suspect network nodes with the greatest certainties are considered, model accuracy is high. For Model 2, where 163 network nodes are possible prediction classes, model accuracy is very
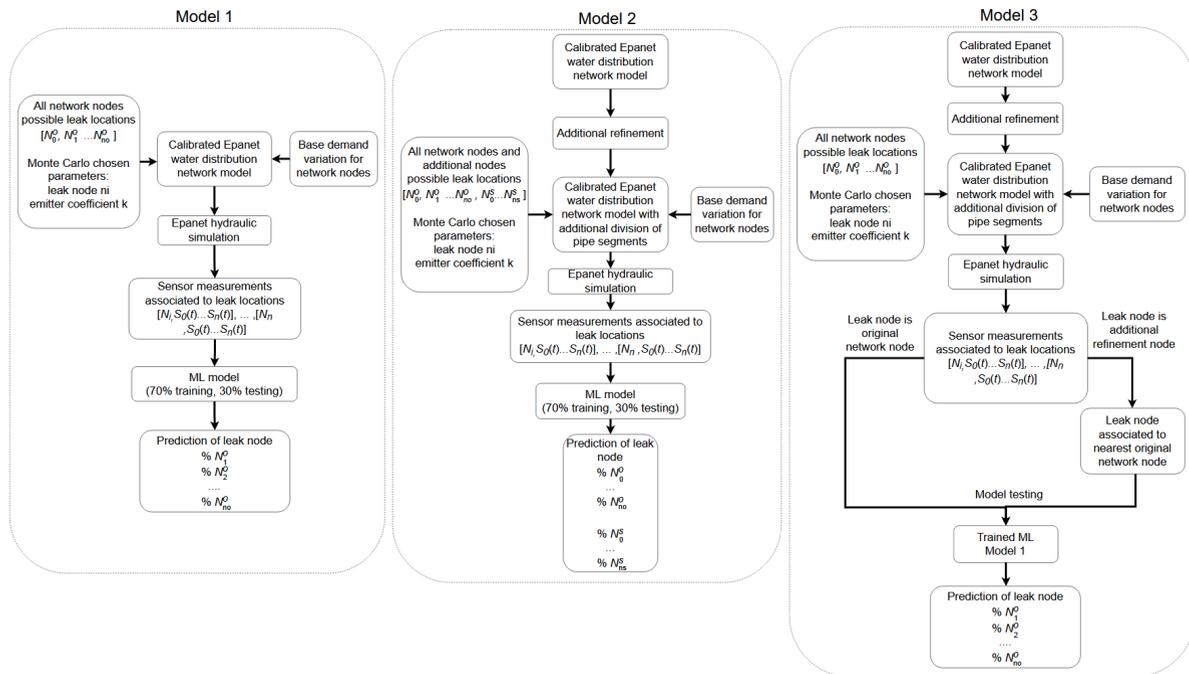
FIGURE 4. Flowcharts of the considered machine learning approaches.

low, indicating that for greater networks this approach would require even more data and computational resources, which is currently not feasible. Model 3 accuracy is reduced compared to the Model 1 approach, which is expected as segmentation nodes increase the total number of possible leak locations. Furthermore, leaking in the segmentation nodes in the middle of the pipe could provide flow patterns that could be equally similar to flow patterns produced by leaking on one or the other edge node of that pipe, thus also contributing to reduced accuracy. However, when top 3 and top 5 suspect nodes are considered, the difference in prediction accuracy for Model 1 and Model 3 shrinks to only around a couple of percents. Although the proposed ML approach demonstrates modest accuracy in predicting the exact leak locations, the proposed approach can be successfully used to narrow down the leak location.

The same investigation was conducted for Net3 network with 4 sensors, emitter coefficient range $10 \ldots 15$, and for different demand variation ranges. Model 1 and Model 3 are created with 92 classes, while Model 2 was also created with 5 additional segments per pipe, resulting in 544 classes altogether.

Results for Net3 are reported in Table 2. It can be observed that prediction model accuracy for the Net3 network is significantly lower than for the Hanoi network. For a model with no demand variation, it is around 7% lower than for the Hanoi network and with an increase in demand variation this decline is over 20%. This is expected, since the Net3 network has a greater number of network nodes and consequently a greater number of possible leak locations. Model 2 accuracy is very small, especially for the strongest variation of demand, as it was observed for the Hanoi network, confirming this

TABLE 1. Influence of Hanoi network refinement on prediction model accuracy for emitter coefficient range $10 \ldots 15$ for different ranges of demand variation. Results are average of 5 runs with 500 000 inputs (350 000 training records).

| | Demand variation | | |
	None	±2.5%	±5%
	Model 1		
Accuracy	100%	82%	69%
Top 3	100%	96%	91%
Top 5	100%	99%	97%
	Model 2		
Accuracy	99%	36%	21%
Top 3	99%	68%	49%
Top 5	99%	80%	64%
	Model 3		
Accuracy	82%	68%	57%
Top 3	97%	94%	88%
Top 5	99%	98%	96%

approach is not feasible. However, although Model 3 has reduced accuracy when compared with Model 1, when considering top 3 and top 5 nodes the accuracy of Model 3 comes very close to the accuracy of Model 1, indicating that the Model 3 approach could be successfully used in a real leak scenario.

Considering these results, only Model 1 and Model 3 will be considered in further research.

B. SENSOR AND EMITTER COEFFICIENT INFLUENCE

The investigation was conducted for various sensor placements, number of sensors and emitter coefficient ranges. The results for the Hanoi network are presented in Table 3. It can be observed that overall prediction model accuracy decreases with greater coefficient range. This is expected since a greater coefficient range increases the size of the problem solution space. On the other hand, with a greater number of sensors,

TABLE 2. Influence of Net3 network refinement on prediction model accuracy for emitter coefficient range 10 . . . 15 for different ranges of demand variation. Results are average of 5 runs with 500 000 inputs (350 000 training records).

	Demand variation		
	None	±2.5%	±5%
	Model 1		
Accuracy	93%	58%	46%
Top 3	99%	89%	77%
Top 5	100%	97%	88%
	Model 2		
Accuracy	77%	22%	13%
Top 3	87%	42%	28%
Top 5	92%	54%	38%
	Model 3		
Accuracy	69%	45%	36%
Top 3	93%	83%	70%
Top 5	98%	93%	84%

TABLE 3. Prediction model accuracy for Hanoi network for various emitter coefficient ranges, sensor layouts and demand variations for model 3.

		Demand variation		
		None	±2.5%	±5%
Emitter coeff.	2 sensors			
10..15	Model 1	100%	80%	67%
	Model 3	81%	66%	56%
5..15	Model 1	100%	68%	53%
	Model 3	85%	57%	45%
Emitter coeff.	3 sensors			
10..15	Model 1	100%	82%	69%
	Model 3	82%	68%	57%
5..15	Model 1	100%	73%	57%
	Model 3	86%	60%	48%

prediction model accuracy slightly increases. Additionally, the greatest difference in Model 1 and Model 3 accuracy appears for scenarios with no demand variation, ranging from 15% to 19%. However, as demand variation increases, the accuracy difference falls to 8 . . . 12%.

The results for Net3 network are presented in Table 4. Same as in the Hanoi network case, with a greater range of emitter coefficient both Model 1 and Model 3 accuracy decrease, for both sensor layouts. Same as in the Hanoi case, as demand variation increases the difference between Model 1 and Model 3 accuracy decreases and again the greatest difference in model accuracy is for no demand variation.

C. PIPE SEGMENT SEGMENTATION INFLUENCE

In order to investigate pipe segmentation influence in the Model 3 approach, three different discretizations are considered for the Net3 network with 4 sensors. Pipes were divided into 3, 5, and 11 segments, resulting in 318, 544, and 1222 possible leak locations, respectively. The results are presented in Table 5. It can be observed that a finer network segmentation slightly reduces model accuracy, which is entirely expected since the number of prediction classes rises with greater refinement. Also, it is expected that at some point further refinement would lead to scenarios with different leak nodes but almost identical pressure readings, since these nodes may happen to be situated very close to each other. However, the rather small decline in accuracy indicates that the proposed approach can be successfully used to narrow down a leak location.

TABLE 4. Prediction model accuracy for Net3 network for various emitter coefficient ranges, sensor layouts and demand variations.

		Demand variation		
		None	±2.5%	±5%
Emitter coeff.	2 sensors			
10..15	Model 1	90%	50%	36%
	Model 3	64%	41%	30%
5..15	Model 1	85%	41%	28%
	Model 3	63%	34%	23%
Emitter coeff.	4 sensors			
10..15	Model 1	93%	58%	46%
	Model 3	69%	48%	37%
5..15	Model 1	89%	49%	36%
	Model 3	67%	41%	30%

TABLE 5. Prediction model 3 accuracy for Net3 network for various number of pipe segments and emitter coefficient ranges.

	Segments	Demand variation		
Emitter coeff.	per pipe	None	±2.5%	±5%
10..15	3	70%	48%	38%
	5	69%	48%	37%
	11	67%	48%	37%
5..15	3	68%	41%	30%
	5	67%	41%	30%
	11	66%	41%	30%

D. ACCURACY IMPROVEMENT

The number of top suspect nodes which need to be considered to achieve 99% model accuracy was investigated to increase accuracy of the prediction model. This approach was already used in [27] to localize the source of pollution and similarly in [13] where the correlation between accuracy and distance between predicted and actual leak node was presented. In this way, a considerable number of network nodes is eliminated, thus the leak area can be localized with considerable certainty even for sparse sensor placement and greatest demand variation.

Number of needed top nodes for Hanoi network is presented in Table 6. It can be observed for Model 1 that with the increase in demand variation, a greater number of top nodes needs to be considered to achieve 99% accuracy; however, considerable localization is achieved even for the strongest demand variation. Similar behavior can be observed with Model 3, where the greatest number of top nodes needs to be considered for the greatest demand variation. Also, a number of top nodes comparing to Model 1 is slightly greater, which is expected. In Figure 5, the increase of model accuracy with the increase of considered top nodes is illustrated. It can be observed that for all models the accuracy of 90% is surpassed when using only top 4 nodes. Additionally, a rapid increase in prediction model accuracy is observed when including the first several top nodes. However, after some threshold the additional nodes in the top list only slightly improve the overall model accuracy.

This kind of investigation has also been conducted for the Net3 network, and the results are presented in Table 7. The number of top nodes is greatest for Model 3 and for stronger demand variation, which is expected and consistent with Hanoi results. It must be noted that even for the worst performing model, with emitter coefficient range 5 . . . 15 and demand variation ±5%, 32 top nodes represent only 35% of

TABLE 6. Number of top nodes needed to achieve 99% accuracy for Hanoi network with various emitter coefficient ranges, sensor layouts and demand variations.

		Demand variation	
		±2.5%	±5%
Emitter coeff.	2 sensors		
10..15	Model 1	5	6
	Model 3	5	7
5..15	Model 1	7	9
	Model 3	8	10
Emitter coeff.	3 sensors		
10..15	Model 1	5	7
	Model 3	6	8
5..15	Model 1	7	9
	Model 3	8	10

TABLE 7. Number of top nodes needed to achieve 99% accuracy for Net3 network for various emitter coefficient ranges and demand variations.

		Demand variation	
		±2.5%	±5%
Emitter coeff.	4 sensors		
10..15	Model 1	8	21
	Model 3	12	24
5..15	Model 1	15	29
	Model 3	19	32

all network nodes, which is still a considerable localization. Additionally, it must be taken into consideration that the chosen 99% accuracy threshold is very high, where the strong model accuracy manifests even for the smaller number of top nodes (Figure 6). To further evaluate the proposed model, the sequential prediction modeling approach is evaluated in the next section.

E. REALISTIC SCENARIO TESTING

To further evaluate the proposed ML approach, an investigation was conducted for a simulated case on Net3 network with 30 records which represent 30 different days. Scenarios are generated with fixed leak location and leak coefficient, but with different demands in network nodes obtained through base demand variation of ±2.5%. Two different leak locations were chosen, first with leak location in network node 159 (Figure 7) with emitter coefficient set to 10, and second with leak location in a pipe segment between nodes 205 and 207 (Figure 8) and with emitter coefficient set to 15. The initial prediction was made using Model 1 with emitter coefficient range 10...15 and base demand variation of ±2.5%. From previous investigation (Table 7) it was observed that when leak locations in pipe segment nodes are allowed, the top 12 nodes achieve 99% accuracy, thus 12 nodes with the greatest prediction model certainty are considered for further segmentation and secondary Model 3 predictions.

For each of the 30 records different certainties are obtained, i.e. the top 12 nodes could be different for each record. Therefore, the average value of all 30 certainties for each node was chosen as a measure for choosing the top 12 nodes with the greatest certainty. For leak node 159, the greatest model certainty is obtained for true leak location, where for leak node in pipe segment between nodes 205 and 207 the greatest certainty is obtained for leak location 207 which is the edge node of the considered pipe segment. Suspect
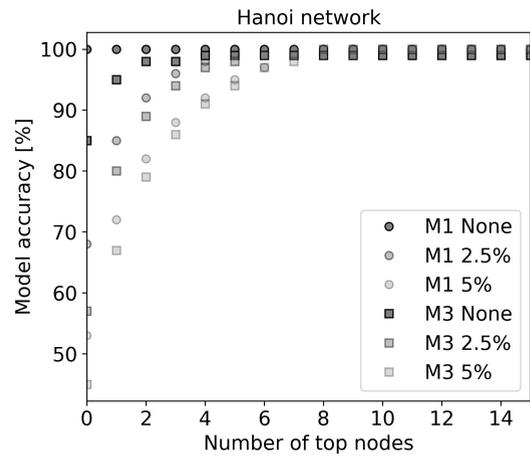

FIGURE 5. Influence of the number of top nodes on prediction model accuracy for Hanoi network with two sensors and emitter coefficient in range 5...15.
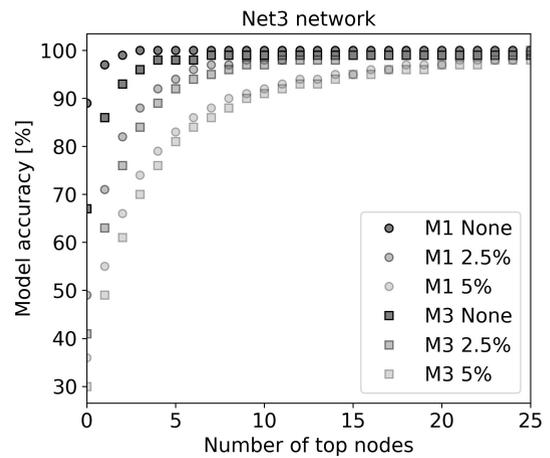

FIGURE 6. Influence of the number of top nodes on prediction model accuracy for Net3 network with 4 sensors layout and emitter coefficient range 5...15.

nodes for both considered cases are presented in Figures 7 and 8, with indicated top 3 nodes with greatest certainty. It can be observed that the top 3 nodes always include true leak location, together with network nodes in the immediate vicinity of the true leak location.

For the next stage, additional pipe segmentation was performed around these top 12 nodes and a prediction model was created where possible leak locations were the top 12 network nodes plus the newly inserted nodes. At this stage, for leak location 159, the most suspect node was node 60, and the second candidate node was node 159 which is the true leak node. For leak location in pipe segment between nodes 205 and 207, the most suspect node was node right next to the true leak node and the second candidate was the true leak node. Top 3 most suspect nodes for both considered cases can be observed in Figures 9 and 10.

The third sequential prediction model was trained also on the top 12 nodes with the greatest average certainty from the previous stage. Both considered cases have true leak location as the second most suspect node. Additionally, from
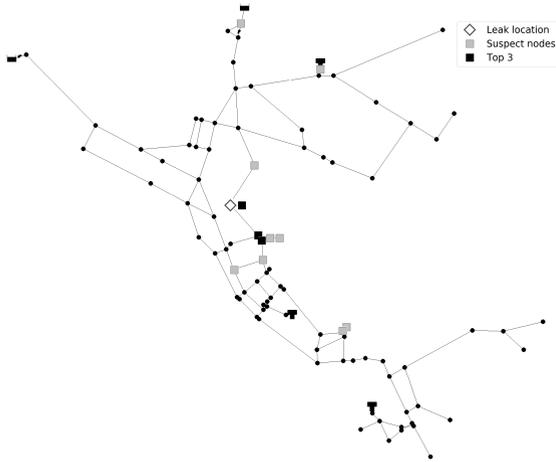
FIGURE 7. Net3 realistic scenario testing for source node 159 with indicated suspect nodes and top 3 nodes at first stage.
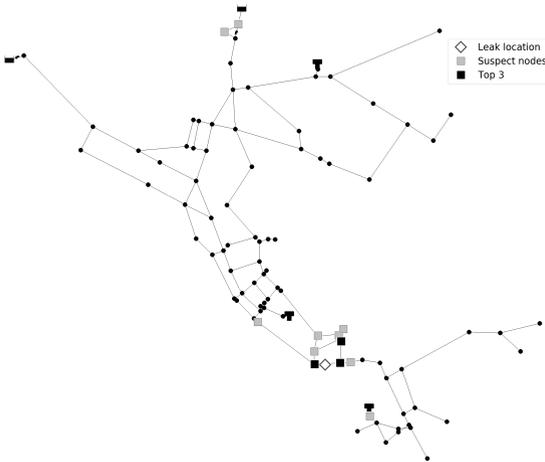

FIGURE 8. Net3 realistic scenario testing for source node between network nodes 205 and 207 with indicated suspect nodes and top 3 nodes at first stage.

Figures 9 and 10 variation in top 3 most suspect nodes can be observed, showing that an unambiguous solution cannot be obtained. This indicates that for different leak locations, demands, and emitter coefficients, still a very similar pressure measurement can be obtained. In other words, there are multiple solutions to the problem. It is shown that the prediction model can efficiently localize leak areas for sparse sensor placement for leak locations which can occur anywhere in pipe segments. However, due to wide range of leak scenarios that are used for prediction model training, a prediction model for fine localization may not be able to provide a single solution.

IV. DISCUSSION

It is shown that the proposed ML approach can be successfully used for localization of leak area under demand uncertainty, for different sized networks, and for different sensor placement layouts. ML model for segmented network pipes was investigated to take into consideration that leaks can occur anywhere along a pipe, but it was shown to be an unfeasible approach. Any pipe segmentation considerably
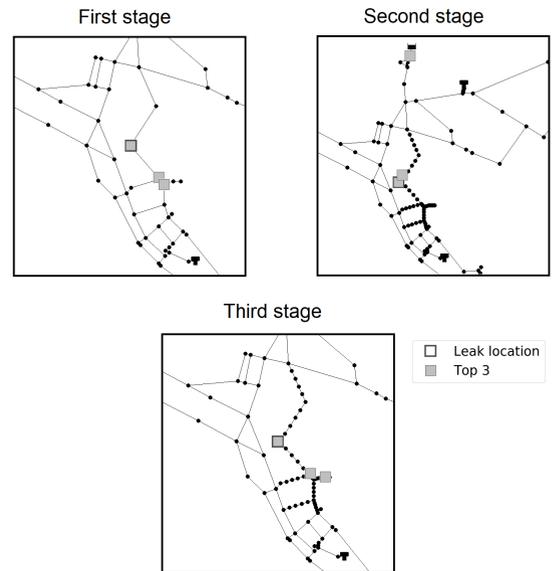

FIGURE 9. Realistic scenario testing for Net3 network for leak location in network node 159 with indicated top 3 nodes through refinement stages.

increases the number of network nodes, i.e. number of prediction classes, with consequently rapidly increasing computational complexity. Additionally, a greater number of inputs is required, which is a considerable problem for greater networks. However, it seems that leaks that occur in pipe segments can be successfully localized with a prediction model trained only on scenarios generated for original network nodes, especially when several top most suspect nodes are considered. It was also observed that regardless of pipe refinement, similar prediction accuracy can be obtained. However, as was mentioned before, a simplification was made where all pipes, regardless of their length, had the same number of divisions. Therefore in future work, fixed lengths for additional refinement nodes should be explored to further explore the presented approach and align the proposed technique with practical purposes.

Sequential prediction models were tested, where the first prediction model specified area for further segmentation, and subsequent models were used to find the exact leak location. It was observed that ML has a problem with detecting fine differences in leak scenarios; the true leak location was always in top nodes but was not always the node with the greatest model certainty. This can be explained by the fact that machine learning models need to cover a large span of scenarios (different demands, different leak sizes, etc.), thus it is reasonably expected that several equally good solutions exist. Similar observation was made in [15] where some leak locations were grouped in single classes, since distinction between locations could not be made.

In further research, coupling of ML and optimization methods needs to be explored. Genetic algorithm (GA) was explored in [28] for leak localization using the inverse transient method for a network with 7 nodes. The main problem with optimization methods in water distribution networks is the network node variable, which is a categorical variable
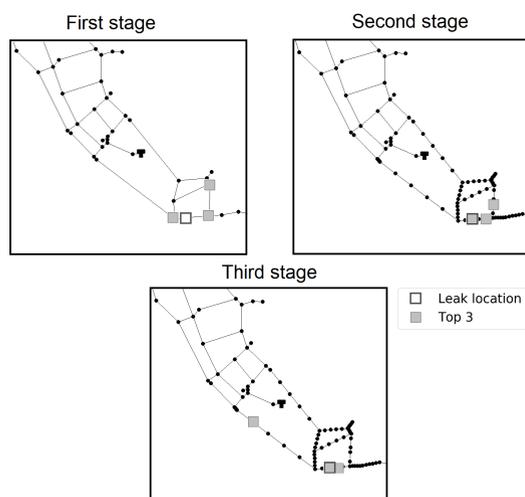
FIGURE 10. Realistic scenario testing for Net3 network for leak location between network nodes 205 and 207 with indicated top 3 nodes through refinement stages.

and as such makes the optimization problem very complex and computationally demanding. However, if ML is used to localize a leak area, independent optimizations for suspect nodes can be conducted and thus reduce the optimization complexity. This was successfully applied in [29] where the pollution source was localized and independent optimizations were conducted to obtain a true pollution source. Additionally, if the optimization method is to be employed, network demands could be more carefully monitored for some period, for example from 2 to 3 AM as proposed in [13], to eliminate or reduce demand variation which is shown to considerably decrease prediction accuracy.

It must be noted that Random Forest classifier was chosen due to its simplicity and since it allows for a reasonably reliable prediction without method parameter fine tuning. For example in [30] RF classifier outperformed SVM, ANN, k-NN and DT for leak detection using acoustic signals, however extensive analysis of classifier parameters was not shown. In [31] six deep neural networks structures and three RF classifier were compared for source tracking of chemical leaks and best accuracy was achieved with RF classifier. Additionally, in [32] Gradient Boosting, DT, RF, SVM and ANN models were investigated for detection of leaks in natural gas pipelines where models were tuned to ensure no false alarm. ANN and SVM showed best performance, however RF and DT were most sensitive to detect small leaks. Therefore, it can be concluded that other models such as ANN may outperform Random Forest algorithm if fine-tuning of hyper-parameters is conducted. Novel ANN methods which deal with this ANN complexity are being developed such as quantum-inspired neural network Autonomous Percep-tron Model [33] which showed better performance than other algorithms, including classic ANN and RF. Therefore, extensive investigation of other machine learning algorithms should be conducted in future work to determine which classifier can provide best model accuracy for leak localiza-tion problem in water distribution networks. Dimensionality

reduction methods should also be explored to reduce the number of features, consequently reducing prediction model complexity which could be important for bigger water distri-bution networks.

The proposed methodology could provide real-time support in water distribution network surveillance. The pre-diction model can be prepared with incorporated demand uncertainties, and can therefore be continuously used to detect when a single leak location is repeatedly reported. However, future work should investigate the possibility of identification of multiple leak locations, which is also most often the case. Other uncertainties should also be incorpo-rated, such as sensor measurement uncertainties and model uncertainties such as pipe diameter and pipe roughness. Ultimately, the proposed methodology should be tested on real-world water distribution network data where all these uncertainties are present.

V. CONCLUSION
In this paper, machine learning approach using big data obtained from computer simulations was investigated for leak localization in water distribution networks. In previous research, a simplification was made in which leaks were only occurring in network nodes and here the methodology is enhanced by allowing for leaks to occur anywhere on any network pipe. It was observed that global refinement of the network in which segmentation is performed on all pipes is not a feasible approach, since the number of potential leak locations rapidly increases and construction of a capa-ble machine learning model is currently computationally too demanding.

However, only a small reduction in model accuracy is observed when the prediction model is trained exclusively on scenarios with leaks appearing in network nodes, while the prediction is then given for leak scenarios with leaks in pipe segments. Further investigation showed that this reduction in model accuracy can be compensated by considering sev-eral most suspect nodes. This approach significantly narrows down the leak area, especially if larger water distribution networks are considered. These observations indicate that the proposed approach could be applicable in real-world water distribution networks and further study of the proposed approach should be conducted.

In future research, additional model uncertainties regard-ing pipe roughness and pipe lengths should be included. Since it is observed that increasing demand uncertainty rapidly decreases model accuracy, an additional approach should also include dimensionality reduction of input data. Sequential prediction models were also explored, where further pre-diction models were trained using only leak scenarios for most suspect leak nodes from the previous prediction model. This approach was shown not to be beneficial since predic-tion models provide a generalized model, and further leak localization needs a specific solution. Coupling the proposed methodology with an optimization procedure could provide better results, which should be explored in future work.

REFERENCES

[1] S. Datta and S. Sarkar, "A review on different pipeline fault detection methods," *J. Loss Prevention Process Ind.*, vol. 41, pp. 97–106, May 2016.

[2] A. Gupta and K. D. Kulat, "A selective literature review on leak management techniques for water distribution system," *Water Resour. Manage.*, vol. 32, no. 10, pp. 3247–3269, Aug. 2018.

[3] U. Baroudi, A. A. Al-Roubaiey, and A. Devendiran, "Pipeline leak detection systems and data fusion: A survey," *IEEE Access*, vol. 7, pp. 97426–97439, 2019.

[4] D. Zaman, M. K. Tiwari, A. K. Gupta, and D. Sen, "A review of leakage detection strategies for pressurised pipeline in steady-state," *Eng. Failure Anal.*, vol. 109, Jan. 2020, Art. no. 104264.

[5] M. I. Mohd Ismail, R. A. Dziyauddin, N. A. Ahmad Salleh, F. Muhammad-Sukki, N. Aini Bani, M. A. Mohd Izhar, and L. A. Latiff, "A review of vibration detection methods using accelerometer sensors for water pipeline leakage," *IEEE Access*, vol. 7, pp. 51965–51981, 2019.

[6] Z. Almheiri, M. Meguid, and T. Zayed, "Intelligent approaches for predicting failure of water mains," *J. Pipeline Syst. Eng. Pract.*, vol. 11, no. 4, Nov. 2020, Art. no. 04020044.

[7] M. Zhou, Z. Pan, Y. Liu, Q. Zhang, Y. Cai, and H. Pan, "Leak detection and location based on ISLMD and CNN in a pipeline," *IEEE Access*, vol. 7, pp. 30457–30464, 2019.

[8] H. Shukla and K. Piratla, "Leakage detection in water pipelines using supervised classification of acceleration signals," *Autom. Construction*, vol. 117, Sep. 2020, Art. no. 103256.

[9] J. Bohorquez, B. Alexander, A. R. Simpson, and M. F. Lambert, "Leak detection and topology identification in pipelines using fluid transients and artificial neural networks," *J. Water Resour. Planning Manage.*, vol. 146, no. 6, Jun. 2020, Art. no. 04020040.

[10] E. J. Pérez-Pérez, F. R. López-Estrada, G. Valencia-Palomo, L. Torres, V. Puig, and J. D. Mina-Antonio, "Leak diagnosis in pipelines using a combined artificial neural network approach," *Control Eng. Pract.*, vol. 107, Feb. 2021, Art. no. 104677.

[11] M.-U.-R.-A. Virk, M. F. Mysorewala, L. Cheded, and I. M. Ali, "Leak detection using flow-induced vibrations in pressurized wall-mounted water pipelines," *IEEE Access*, vol. 8, pp. 188673–188687, 2020.

[12] X. Zhou, Z. Tang, W. Xu, F. Meng, X. Chu, K. Xin, and G. Fu, "Deep learning identifies accurate burst locations in water distribution networks," *Water Res.*, vol. 166, Dec. 2019, Art. no. 115058.

[13] J. Mashford, D. De Silva, S. Burn, and D. Marney, "Leak detection in simulated water pipe networks using SVM," *Appl. Artif. Intell.*, vol. 26, no. 5, pp. 429–444, May 2012.

[14] Y. Liu, X. Ma, Y. Li, Y. Tie, Y. Zhang, and J. Gao, "Water pipeline leakage detection based on machine learning and wireless sensor networks," *Sensors*, vol. 19, no. 23, p. 5086, Nov. 2019.

[15] A. Soldevila, J. Blesa, S. Tornil-Sin, E. Duviella, R. M. Fernandez-Canti, and V. Puig, "Leak localization in water distribution networks using a mixed model-based/data-driven approach," *Control Eng. Pract.*, vol. 55, pp. 162–173, Oct. 2016.

[16] A. Soldevila, R. M. Fernandez-Canti, J. Blesa, S. Tornil-Sin, and V. Puig, "Leak localization in water distribution networks using Bayesian classifiers," *J. Process Control*, vol. 55, pp. 1–9, Jul. 2017.

[17] M. Quiñones-Grueiro, C. Verde, A. Prieto-Moreno, and O. Llanes-Santiago, "An unsupervised approach to leak detection and location in water distribution networks," *Int. J. Appl. Math. Comput. Sci.*, vol. 28, no. 2, pp. 283–295, Jun. 2018.

[18] C. Sun, B. Parellada, V. Puig, and G. Cembrano, "Leak localization in water distribution networks using pressure and data-driven classifier approach," *Water*, vol. 12, no. 1, p. 54, Dec. 2019.

[19] E. G. Mohammed, E. B. Zeleke, and S. L. Abebe, "Water leakage detection and localization using hydraulic modeling and classification," *J. Hydroinformatics*, vol. 23, no. 4, pp. 782–794, Jul. 2021.

[20] I. Lučin, B. Lučin, Z. Čarija, and A. Sikirica, "Data-driven leak localization in urban water distribution networks using big data for random forest classifier," *Mathematics*, vol. 9, no. 6, p. 672, Mar. 2021.

[21] W. Moczulski, R. Wyczółkowski, K. Ciupke, P. Przystałka, P. Tomasik, and D. Wachla, "A methodology of leakage detection and location in water distribution networks—The case study," in *Proc. 3rd Conf. Control Fault-Tolerant Syst. (SysTol)*, Sep. 2016, pp. 331–336.

[22] Q. Zhang, Z. Y. Wu, M. Zhao, J. Qi, Y. Huang, and H. Zhao, "Leakage zone identification in large-scale water distribution systems using multiclass support vector machines," *J. Water Resour. Planning Manage.*, vol. 142, no. 11, Nov. 2016, Art. no. 04016042.

[23] L. A. Rossman, "Epanet 2: Users manual," Nat. Risk Manage. Res. Lab., Office Res. Develop., U.S. Environ. Protection Agency, Cincinnati, OH, USA, Tech. Rep. EPA/600/R-00/057, 2000.

[24] *University of Exeter Centre for Water Systems*. Benchmarks. Accessed: Nov. 6, 2020. [Online]. Available: http://emps.exeter.ac.U.K./engineering/research/cws/downloads/benchmarks/

[25] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, and M. Blondel, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[27] L. Grbčić, I. Lučin, L. Kranjčević, and S. Družeta, "Water supply network pollution source identification by random forest algorithm," *J. Hydroinformatics*, vol. 22, no. 6, pp. 1521–1535, Nov. 2020.

[28] J. P. Vítkovský, A. R. Simpson, and M. F. Lambert, "Leak detection and calibration using transients and genetic algorithms," *J. Water Resour. Planning Manage.*, vol. 126, no. 4, pp. 262–265, 2000.

[29] I. Lučin, L. Grbčić, S. Družeta, and Z. Čarija, "Source contamination detection using novel search space reduction coupled with optimization technique," *J. Water Resour. Planning Manage.*, vol. 147, no. 2, Feb. 2021, Art. no. 04020100.

[30] Z. Chi, Y. Li, W. Wang, C. Xu, and R. Yuan, "Detection of water pipeline leakage based on random forest," in *Proc. J. Phys., Conf.*, vol. 1978, 2021, Art. no. 012044.

[31] J. Cho, H. Kim, A. L. Gebreselassie, and D. Shin, "Deep neural network and random forest classifier for source tracking of chemical leaks using fence monitoring data," *J. Loss Prevention Process Ind.*, vol. 56, pp. 548–558, Nov. 2018.

[32] O. Akinsete and A. Oshingbesan, "Leak detection in natural gas pipelines using intelligent models," in *Proc. SPE Nigeria Annu. Int. Conf. Exhib.*, 2019, pp. 573–583.

[33] A. Sagheer, M. Zidan, and M. M. Abdelsamea, "A novel autonomous perceptron model for pattern classification applications," *Entropy*, vol. 21, no. 8, p. 763, Aug. 2019.

IVANA LUČIN received the B.S. and M.S. degrees in mechanical engineering from the Faculty of Engineering, University of Rijeka, in 2013 and 2015, respectively, where she is currently pursuing the Ph.D. degree in computational mechanics.

She is currently a Teaching Assistant with the Faculty of Engineering, University of Rijeka. Her research interests include hydraulic systems analysis, machine learning, and optimization methods.

ZORAN ČARIJA received the Ph.D. degree from the Faculty of Engineering, University of Rijeka, in 2007, with a dissertation in the field of computational mechanics.

He is currently a Professor with the Department of Fluid Mechanics and Computation Engineering, Faculty of Engineering, University of Rijeka. He is also the Head of the Section of Fluid Mechanics and Hydraulic Turbomachinery. His research interests include computational fluid dynamics, fluid mechanics, turbomachinery, water turbines, and renewable energy.

SINIŠA DRUŽETA received the Ph.D. degree from the Faculty of Engineering, University of Rijeka, in 2007, with a dissertation in the field of free surface flow modeling.

He is currently a Professor with the Department of Fluid Mechanics and Computation Engineering, Faculty of Engineering, University of Rijeka. He is the also the Head of the Section of Computational Engineering. His research interests include hydraulic systems analysis and optimization, open channel flow, and optimization methods.

BOŽE LUČIN received the B.S. and M.S. degrees in mechanical engineering from the Faculty of Engineering, University of Rijeka, in 2013 and 2016, respectively, where he is currently pursuing the Ph.D. degree in computational mechanics.

He is currently employed as a Project Engineer at Flowtech d.o.o., and he is also an External Associate with the Faculty of Engineering, University of Rijeka. His research interests include computational fluid dynamics and renewable energy.

• • •