Jyotirmoy Bhardwaj; Joshin P. Krishnan; Diego F. Larios Marin;
Baltasar Beferull-Lozano; Linga Reddy Cenkeramaddi;
Christopher Harman. 2021.
Cyber-Physical Systems for Smart Water Networks: A Review.
IEEE Sensors Journal. Vol 21 (23): 26447-26469.

The article has been published in final form at
http://dx.doi.org/10.1109/JSEN.2021.3121506
by Institute of Electrical and Electronics Engineers.

Cyber-Physical Systems for Smart Water Networks: A Review

Jyotirmoy Bhardwaj, Joshin P. Krishnan, Diego F. Larios Marin, Baltasar B. Lozano, Linga R. Cenkeramaddi, and Christopher Harman

***Abstract*—There is a growing demand to equip Smart Water Networks (SWN) with advanced sensing and computation capabilities in order to detect anomalies and apply autonomous event-triggered control. Cyber-Physical Systems (CPSs) have emerged as an important research area capable of intelligently sensing the state of SWN and reacting autonomously in scenarios of unexpected crisis development. Through computational algorithms, CPSs can integrate physical components of SWN, such as sensors and actuators, and provide technological frameworks for data analytics, pertinent decision making, and control. The development of CPSs in SWN requires the collaboration of diverse scientific disciplines such as civil, hydraulics, electronics, environment, computer science, optimization, communication, and control theory. For efficient and successful deployment of CPS in SWN, there is a need for a common methodology in terms of design approaches that can involve various scientific disciplines. This paper reviews the state of the art, challenges, and opportunities for CPSs, that could be explored to design the intelligent sensing, communication, and control capabilities of CPS for SWN. In addition, we look at the challenges and solutions in developing a computational framework from the perspectives of machine learning, optimization, and control theory for SWN.**

***Index Terms* —Cyber-Physical Systems, Smart Water Networks, Internet-of-Things, Machine Learning, Water Quality, and Optimal Control.**

I. INTRODUCTION

Water is an essential resource for both the natural environment and human life. Protecting water from contamination and ensuring the availability of high-quality pure water are widely recognized as critical societal goals around the world. Furthermore, the right to safe water is one of the United Nations' top priorities, as reaffirmed in several official reports [1]. Traditional methods and techniques for monitoring and controlling water networks are being replaced by new methods and techniques. Sensors installed at pumping stations or water treatment plants collect data on a variety of chemical, biological, physical, and hydraulic parameters. However, once water enters the network, it becomes difficult to perform water quality assessment and event-triggered control in an online manner over distributed locations. Real-world applications such as urban, industrial, and household water networks highlight this issue. Some of the realistic challenges of water networks include water demand management, online contamination detection, autonomous control, pressure and flow management, and real-time leakage detection. To address these issues, several experimental studies suggest that intelligent monitoring and control capabilities be implemented in Smart Water Networks (SWN)[1] such as water distribution network (WDN), wastewater networks, Aquaponics, fish farms, Recirculating Aquaculture, etc. These studies also demonstrated that traditional offline methods are out of date and incapable of meeting the current challenges of SWN. As a result, it is critical to develop a system of various components capable of integrating sensing, computing, and communication in order to address the challenges of SWN [2]. With the ever-increasing expansion of water infrastructure, these systems are also expected to be re-configurable and adaptive.

Cyber Physical Systems (CPSs) have recently received a great deal of attention due to their application in a wide range of real-time networks such as smart-grid networks, water/gas distribution networks, etc [3]. CPSs are an extended version of embedded systems with feedback capabilities that can integrate sensing, communication, and control capabilities to observe and control the physical process state. Furthermore, CPSs are designed in such a way that they can react autonomously in the event of an unexpected crisis development while keeping users informed. CPS, in conjunction with multiple sensors (electronic, voltammetry, optical) and transducers, can sense and interact with the physical environment in an online fashion [4]. CPSs can also learn from the SWN in order to extract observations and inference patterns. CPSs offer scalable and reconfigurable properties, which can be modified based on the volume of

This work is supported in part by IKTPLUSS funded Project "Data-driven cyber-physical networked systems for autonomous cognitive control and adaptive learning in industrial urban water environments (INDURB)", led by WISENET Center, University of Agder, Norway and in part by Norwegian Institute for Water Research, Oslo, Norway (Corresponding author: jyotirmoy.bhardwaj@uia.no).

Jyotirmoy Bhardwaj is with WISENET Center, University of Agder, Grimstad, 4879, Norway and Norwegian Institute for Water Research, Oslo, 0349, Norway (e-mail: jyotirmoy.bhardwaj@uia.no).

Baltasar B. Lozano and Joshin P. Krishnan are with WISENET Center, University of Agder, Grimstad, 4879, Norway (e-mail: {baltasar.beferull, joshin.krishnan}@uia.no).

Diego F. Larios Marin is with Department of Electronic Technology, University of Seville, Sevilla, 41004, Spain (e-mail:dlarios@us.es).

Linga R. Cenkeramaddi is with ACPS group, University of Agder, Grimstad, 4879, Norway (e-mail:linga.cenkeramaddi@uia.no).

Christopher Harman is with Norwegian Institute for Water Research, Oslo, 0379, Norway (e-mail: christopher.harman@niva.no).

Digital Object Identifier: XXXXXXXXXXXX

[1]Henceforth, throughout the paper, whenever we refer to term SWN, we refer to the Industrial and Urban water networks, such as water distribution network (WDN), wastewater networks, Aquaponics, fish farms, Recirculating Aquaculture, etc.

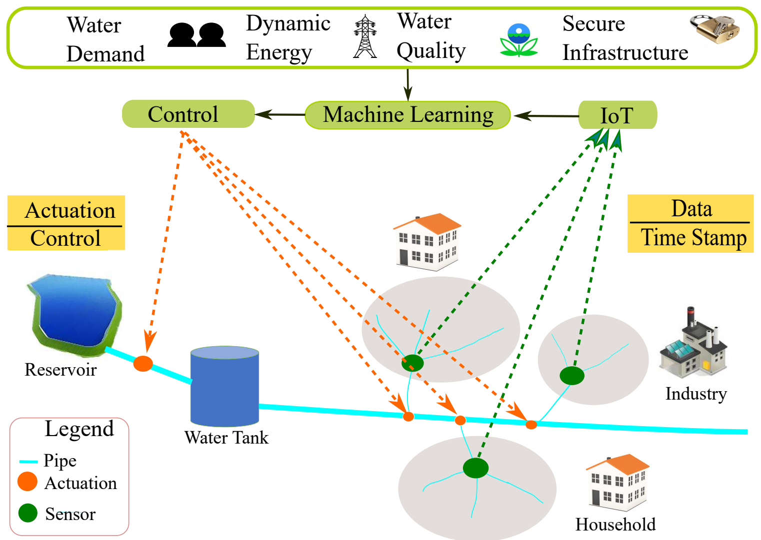


Fig. 1: Cyber-physical systems for WDN

TABLE I: List of Abbreviations

Abbreviation	Description
ANN	Artificial Neural Network
CPS	Cyber Physical System
DP	Dynamic Programming
DNN	Deep Neural Networks
DRL	Deep Reinforcement Learning
DO	Dissolved Oxygen
EPA	Environmental Protection Agency
GSM	Global System for Mobile
IoT	Internet of Things
kNN	k Nearest Neighbours
LPWAN	Low Power Wide Area Network
LTE	Long-Term Evolution
MEC	Mobile Edge Computing
MI	Mixed Integer
ML	Machine Learning
MPC	Model Predictive Control
PCA	Principal Component Analysis
RL	Reinforcement Learning
RF	Random Forest
SQL	Structured Query Language
SoS	System of Systems
SVM	Support Vector Machine
LoRa	Long Range
LPWAN	Low Power Wide Area Networks
WDN	Water Distribution Network
SWN	Smart Water Networks
WSN	Wireless Sensor Network

data, available bandwidth, power, and sensing requirements. CPS for WDN is depicted in Fig. 1.

The majority of the existing review studies in the literature focus on the methods of design and development of CPS for SWN [5]-[6]. For example, [5] presents a theoretical framework of CPS development for SWN and [6] presents the CPS challenges and roadmaps for WDN management. Similarly, in [7], a comprehensive review of communication technologies, such as Internet-of-Things for SWN management is provided. However, in the event of unexpected anomaly detection, the CPSs are expected to take control of the SWN autonomously. To the best of our knowledge, no comprehensive survey has been conducted that addresses the fundamental issue of integrating computation and autonomous control capabilities in such CPSs. Because of the various *nonlinear*, *non-convex*, and *integer* constraints posed by flow, pump, and tank operations, integrating autonomous control in such SWN is a complex task. The *non-convex* constraints imposed by the flow and pump operations make this problem $\mathcal{NP}$-Hard. Solving $\mathcal{NP}$-Hard problems is computationally expensive, both in terms of memory and time [8]. Therefore, in addition to covering the data observation and acquisition framework for SWN, we discuss how we can integrate challenging computation and control capabilities via the Internet of Things (IoT). Furthermore, we present how data-driven Machine Learning (ML) techniques can be used to address the challenges posed by complex problems in SWN. The structure of this paper is given in Fig. 2 and the main contributions of this survey paper can be enumerated as follows:

- A review of the literature on how to perform data acquisition in water CPSs via IoT (Section III).
- We present a comprehensive review of ML techniques aimed at SWN (Section IV).
- We present a detailed overview of the algorithmic challenges posed by the $\mathcal{NP}$ hydraulic constraints to control algorithms. We also look at how machine learning (ML) techniques like Deep Learning(DL), Reinforcement Learning (RL), and Deep Reinforcement Learning (RL) can be used to address the challenges posed by such

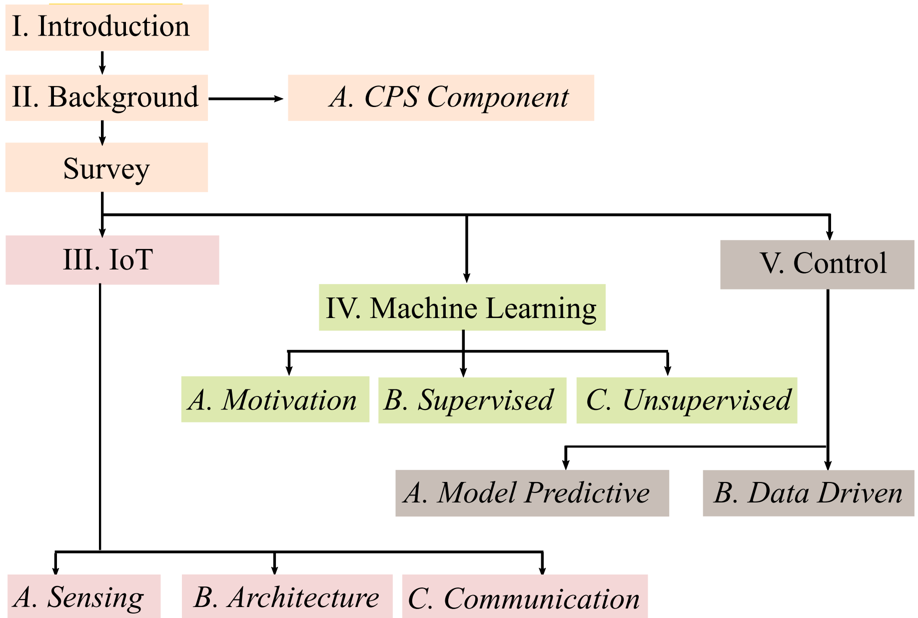


Fig. 2: Structure of this paper

constraints (Section V).

II. Background

Water quality monitoring is the first step in the management of any SWN because it provides the necessary evidence to make intelligent decisions. With the introduction of glass electrodes in the early 1920s, scientific efforts to develop water quality monitoring began. Such electrodes used voltammetry or amperometry measurement techniques to determine an individiual water quality parameter such as pH [9]. Overall, water quality monitoring, however, remains a complex task because water can contain a wide range of chemical and biological parameters that can indicate the presence of contamination in the SWN. The main limitation of individual sensing instruments is that they cannot detect a wide range of chemical and biological parameters. As a result, a more cooperative integrated approach has been followed to detect multiple parameters of water simultaneously by integrating heterogeneous water quality sensors. This combination of heterogeneous sensors in a single system is expected to provide superior sensitivity and selectivity, as well as the ability to analyze data in real-time [10]. The spatial coverage of SWN presents another challenge in water quality monitoring. Since SWN has extensive spatial coverage, wired systems are incapable of providing an adequate flow of information transmission between user and source. As a result, Wireless Sensor Networks (WSN) emerged as a potential tool for the online transfer of relevant water quality information. Online monitoring of WDN in Singapore, for example, proposes an end-to-end solution using WSN for monitoring, analyzing, and modeling urban water distribution networks [11].

However, such advancements were limited to observing the state of the SWN using distributed sensor nodes linked by WSN, with control issues left to the discretion of the controlling authorities. Manual control is a cumbersome task in such a complex SWN because the SWN may be distributed over a large geographical region. As a result, autonomous and event-triggered control strategies for the operational management of such SWN are required. CPS is important in this case because it can monitor the state of the SWN using sensors and apply desired autonomous control. CPSs-based monitoring and control approaches have already been tested for the management of oil pipelines and autonomous cars [12], and they are gaining popularity for the operational management of SWN. The most recent developments in CPSs for SWN can be found in Table II.

A. CPS Components

CPS are designed to achieve autonomous end-to-end control, i.e from sensing to control. We can classify the key components of CPS as follows:

- **Advanced sensing and networking technologies**, such as the Internet of Things (IoT), to capture and store data of physical, chemical, and hydraulic parameters.
- **Computing Technologies** to perform several (centralized or decentralized) tasks such as data pre-processing or filtering, as well as various data-driven ML techniques, in order to address the challenges posed by several SWN-related application use cases, such as anomaly detection and prediction of relevant events.
- **Control**, that is, autonomous real-time event-triggered control capabilities to achieve tightly coordinated control actions [18] towards maintaining desirable properties or behavior in the SWN.

TABLE II: Studies proposing Cyber physical systems for SWN

References	SWN			Overview	Implementation method
	Drinking Water	Waste water	Aqua-ponics		
[6], 2020	✓	✓		Multi-layer CPS framework.	Barcelona water supply system.
[5], 2015	✓	✓		Proposed theoretical architecture of water CPSs.	-
[13], 2019			✓	Use IoT and CPS for Aquaponics system management.	Authors integrated sensor units, networking units, and computational units using microcontrollers.
[14], 2019		✓		CPS designed for real-time sensing and actuation for urine diversion.	Testbed using sensors, actuators and pumps.
[15], 2014	✓			CPS using mobile sensors in WDN infrastructure.	Envision a CPS with mobile sensors.
[16], 2015	✓			Connectivity in CPS subsystems.	Virtual Shanghai water distribution network.
[17], 2016	✓			Five-layer CPS architecture.	The study proposed a CPS framework using data mining, data fusion, hydraulics, and modelling.
[4], 2018	✓			CPS architecture.	Developed a testbed, and decision support system.

Through its interaction with SWN, sensing generates time-series data. Because the sensors may be distributed in geographically dispersed locations, intelligent communication techniques that provide a common data acquisition framework are required. Through nodes, storage servers, and intelligent algorithms, IoT provides an intelligent framework for data communication, data storage, and data analytics [19]. Once the time-series data is collected via IoT, we need intelligent algorithms to detect patterns in the data set and assist the user with predictive analytics and decision making. As a result, we require intelligent computing techniques such as machine learning (ML) to detect inferences from patterns and identify anomalies in the high volume of complex data streams [20]. These inferences are required for the development of advanced control capabilities in SWN. In the following sections (Section III-Section V), we look at IoT, ML, and Control techniques for the design of CPS in the context of overall SWN management.

III. Internet of Things

With the advancements in communication technologies, we are moving towards an era of ubiquitous connectivity, where a wide range of applications are connected to the Internet. Internet of Things (IoT) is a new technology paradigm, where the sensors, embedded processors, and actuators are deeply intertwined through advanced communication technologies to monitor the state of a physical process in real-time. According to Vermesen et al. [21], IoT is an interaction between the physical and digital worlds, where the digital world interacts with the physical world through a plethora of sensors and actuators. We would like to emphasize that IoT is not a single and stand-alone technology, but it is a collection of different technologies, which work together to monitor the state of a physical environment such as SWN. In addition, IoT can be seen as an enabling technology for CPS, as IoT is expected to link the diverse elements (Sensing, ML, and Control) of CPS to the internet [22]. IoT can be used for various applications such as healthcare, education, energy management, home automation, and smart city management. In the context of SWN, some of the use cases for IoT are water quality monitoring, WDN Management [23], Aquaponics [24], and Hydroponics [25]. IoT is necessary to construct the data management and communication infrastructure of CPSs as emphasized in [26]. Therefore, in this section, we present the major components of IoT, mainly Sensing, Architecture, and Communication in the context of SWN.

A. Sensing

Sensing is an important component of SWN and IoT architecture. Sensors interact with the SWN and monitor various physical, chemical, and hydraulic parameters. In addition, these sensors provide valuable data from aforementioned parameters. For instance, a pH sensor determines the acidity and alkalinity of the water. Total Dissolved Solid measurement determines the presence of organic salt and inorganic matter. A dissolved oxygen sensor determines the presence of oxygen in water, which is an important criterion for drinking purposes and aquatic life. However, monitoring various physical, chemical, and hydraulic parameters requires a diverse range of measurements from heterogeneous sensors, and therefore water utilities install heterogeneous sensors to monitor the overall state of SWN.

The selection of the type of heterogeneous sensors are application-specific in SWN, which is based on empirical evidences and on the recommendation of environmental monitoring agencies such as the United States Environmental Protection Agency (EPA) [30]. For instance, Hall et al. recommend WDN water quality monitoring by measuring

TABLE III: Heterogeneous sensor for different SWN

References	SWN			Heterogeneous sensors
	Drinking Water	Waste Water	Aqua-ponics	
[27], 2007	✓			pH, Free Chlorine, ORP, DO, EC, Turbidity, Total Organic Carbon, Chloride, Ammonia, and Nitrate.
[28], 2019			✓	pH, Temperature, DO, Nitrate, Ammonia, and EC.
[10], 2014	✓			Turbidity, Free Residual Chlorine, ORP, Nitrates, Temperature, pH, EC, and DO.
[29], 2006		✓		Total Organic Carbon, Chemical Oxygen Demand, Biological Oxygen Demand, Total Suspended Solids, Nitrogenous, and Phosphorous compounds.

heterogeneous parameters such as pH, dissolved oxygen (DO), electrical conductivity (EC), and oxygen reduction potential (ORP) [27], whereas [28] recommends measuring pH, Temperature, DO, Nitrate, Ammonia, and EC for Aquaponics application. Table III summarizes the important research studies, which integrated heterogeneous sensors for different SWNs. These evidences also suggest that some specific water parameters, mainly pH, EC, DO and ORP, are the most sensitive indicators of contaminants such as *nicotine, arsenic trioxide* and *Escherichia coli* [27]. Therefore, instead of direct detection of any specific contaminant, monitoring these specific parameters through selected heterogeneous sensors is a feasible and low-cost alternative for overall water quality monitoring. This approach of integrating heterogeneous sensors offers a broad contamination coverage and is sometimes also termed as sensor fusion [31].

These heterogeneous sensors have distinct manufacturing properties, different throughput and, distinct measurement cycles. The Low-level layer of IoT architecture plays a crucial role in data acquisition from such heterogeneous sensors by synchronizing different throughput and measurement cycles. In the next subsection (III-B), we review the IoT architectures for smooth and efficient data acquisition from heterogeneous sensors.

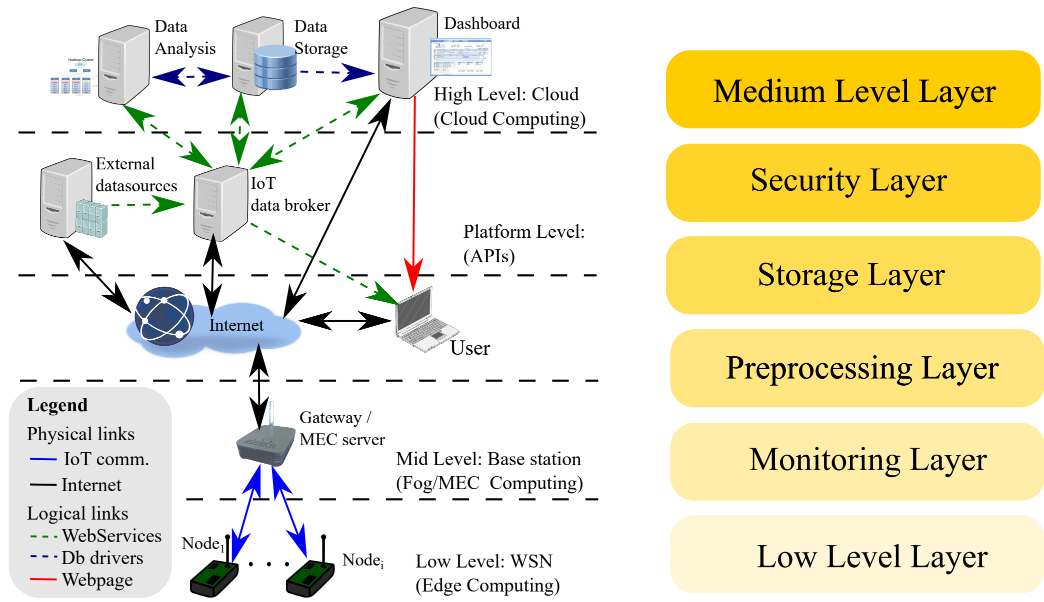
B. IoT Architecture

IoT Architecture can be described as an environment that supports data acquisition, data storage, data visualization, and computing in a distributed fashion over the Internet. Recently, IoT architectures have received great attention for smooth data acquisition and analysis; see, e.g., [32]. In the context of SWN, IoT architecture facilitates smooth data acquisition from heterogeneous sensors in an online fashion. We can classify the IoT architectures as Layered Architecture or Cloud/Fog based Architecture [33]. In the following subsections, we present different IoT architectures in the context of SWN.

1) Layered Architectures

This class of architecture consists of multiple layers for smooth data acquisition and processing. Although, there is no universally agreed consensus over the number of layers, different researchers propose *Three-*, *Four-*, *Five-* or even *Seven*-Layer IoT architectures. For instance, *Three-* and *Five-* layered IoT architectures are presented in [34]. For smooth data acquisition in SWN, we present a *Four*-layered architecture as shown in Fig. 3a, and the function of each layer is described as follows:

- The Low-level layer, also known as the perception layer, is composed of distributed and heterogeneous sensors to collect the data from SWN. This layer senses physical and chemical parameters to obtain observations representing the state of the environment.
- The Medium-level layer, also termed as Network layer, directs the data from the Low-Level layer to the Platform layer. The Medium-Level layer determines the path of data transfer using devices (such as gateways, routing devices, hubs, etc), which are connected through various networks (such as wireless, 3G, LAN, Bluetooth, RFID, and NFC) [35].
- The Platform layer consists of mainly databases, data, and data pre-processing modules. This layer accumulates and processes the data streams acquired from the Low-Level layer. Generally, this layer is composed of two major stages: (i) the Data accumulation stage and (ii) the Data abstraction stage. The data accumulation stage captures the real-time data from various sources (such as an Application Programming Interface) in a structured manner. SQL and NoSQL are the most popular and powerful data accumulation servers. Whereas, the Data abstraction stage performs data pre-processing.
- The High-level layer, also known as the Application layer, is responsible for data visualization and analytics. This layer consists of (i) User Interface and (ii) Data analytics section. The User Interface displays the time-series information of sensor data and subsequently presents an analysis in a user-friendly way. Grafana is one such User Interface platform commonly used in IoT. The Data Analytics section performs computing over dataset and may consist of an advance statistical algorithm (such as ML, discussed in Section IV) for data analysis



(a) (b)

Fig. 3: (a) Architecture of IoT (Four layers), (b) Fog architecture of an IoT gateway.

and anomaly detection. It is expected that this layer should have high computational capabilities to address the challenges posed by the high volume of dataset [36].

In layered architectures, data from the sensors are usually sent to a fusion center, which may be, e.g., a storage server over the cloud or some other secure server in a control center supervising operations. Such architectures are mostly designed in a way that one could store, process, and perform computation over the entire dataset in a centralized manner. In addition, the layers of IoT infrastructure are to be designed in an ad-hoc manner and need careful planning for future restructuring as per the expected requirements of SWN.

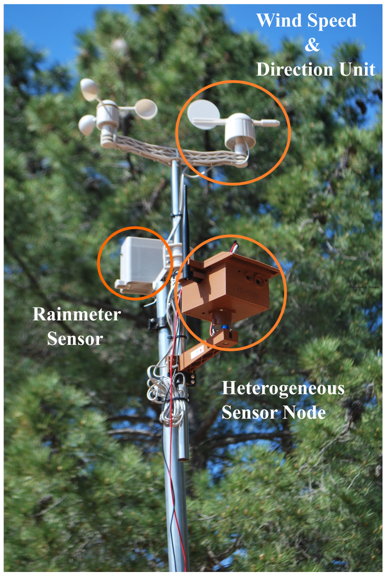
2) *Cloud/Fog-based Architectures*

Cloud and Fog based architectures are system-based architectures [33], composed by integrating different elements, which work together to achieve a specific goal. Cloud-based architectures are scalable and flexible system-based architecture, where the Cloud refers to the host server over the internet. The elements of a cloud server are data storage, software tools, ML, and user interfaces. In Cloud-based architectures, the sensor communicates to the cloud, and the cloud performs the data processing and analytics tasks in a centralized fashion. Unfortunately, such an approach may be slow and time-inefficient for large-scale SWN, as the sensors even transfer the redundant and repetitive data to the central server. Instead, one can process the dataset locally and transfer only the relevant sensor data to the central server. Fog-based architectures are designed to process the dataset locally, where the sensors and gateways can be used to perform part of the data processing and communicate only relevant sensor data to the cloud [38]. The Fog-based architecture is composed of multiple layers and is depicted in Fig. 3b. Such architectures are constructed by inserting four additional layers between the Low-Level layer and Medium-Level layers (discussed in Section III-B1). The four layers can be classified as; *Monitoring layer* - to monitors the resources and power consumption; *Preprocessing layer*- for filtering and analytics of data; *Storage layer*- for the temporary storage of data, and *Security layer*-to ensure privacy and data integrity.

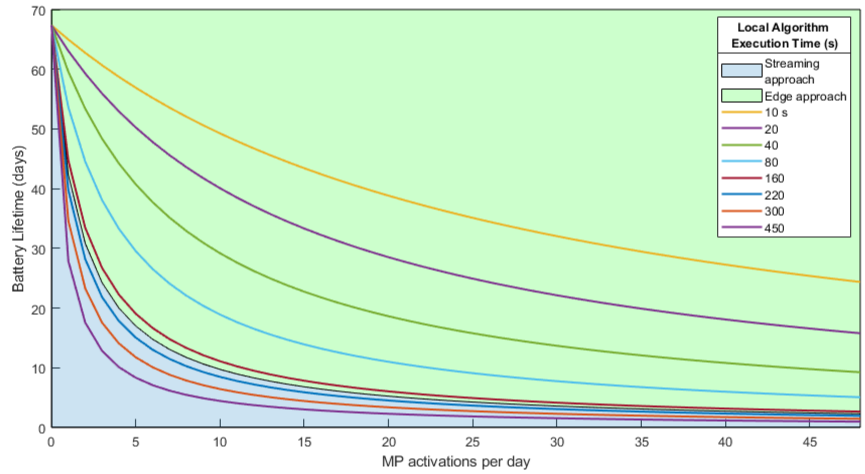
Edge computing can be seen as an extension of Fog-based architectures. *Edge computing* envisions that users can improve the performance of IoT by introducing smart data preprocessing capabilities. This technology pushes cloud services to the end-user, and is often deployed at the gateways to perform analytics, and minimize the power and bandwidth consumption of the network. In brief, this approach discards the redundant data, and transfer only the selective and essential data to the host server over the cloud; which results in better energy management, improved data transfer rates, and improved data processing capacity in an IoT Network [39]. Reference [37] presented the advantages of Edge computing in terms of energy management as shown in Fig. 4, where Fig. 4a depicts the assembled and deployed heterogeneous sensor node in the Doñana National Park (Spain), and Fig. 4b demonstrate the advantages of Edge Computing for improved battery lifetime.

C. *Communication*

Since heterogeneous sensors are geographically distributed, the wired data acquisition is an infeasible and not preferred choice. In such a scenario, data acquisition through wireless communication technologies emerges as a natural choice. Bluetooth, WiFi, ZigBee, LoRA, Narrow Band-IoT, and Sigfox are the leading wireless communication technologies for efficient IoT deployment. The selection of deployed technology depends on the factors such as communication



(a)



(b)

Fig. 4: Work by Garcia et al. [37], where (a) depicts assembled and deployed wireless sensor node in the Doñana National Park (Spain), and (b) demonstrate the advantages of edge computing over continuous data streaming for environment monitoring. Approved to be used by the original authors.

coverage, power consumption, data transmission latency, and bandwidth offered. We cover some of the major wireless communication technologies for IoT as follows:

- *Short Range Communication*: RFID and NFC (Near field communication) are some of the short-range communication technologies, which can communicate to the devices located in close proximity.
- *Wireless Sensor Networks (WSN)*: The use of short-range communication is constrained for applications that cover a large geographical area. WSN consists of distributed senor nodes, deployed over a small or vast geographical region and are connected in a wireless fashion through gateways. WSN can be deployed through diverse topologies such as star, delta, or mesh [40]. Communication through WSN is based on several standards, the most popular one being IEEE 801.15.4. WSN is an efficient and robust technology and has been utilized in a diverse range of applications in SWN such as water quality monitoring [41], Aquaponics, and WDN.
- *Low Power WiFi*: Traditional WiFi provides a substantial data rate (up to 9.6 Gbps); however, it consumes a significant amount of power. The WiFi Alliance has developed WiFi HaLow, which is a low-power long-range alternative to WiFi. This technology offers a communication range nearly double of traditional WiFi and relies on standard IEEE 802.11a.
- *Wireless Personal Area Network (WPAN)*: WPAN is a low power, short-distance, and low data rate wireless communication technology. The coverage of such technology ranges from a few centimeters to a few meters. Bluetooth, ZigBee, and Helium are some of the examples of WPAN. This technology is based on standard IEEE 802.15.
- *Low Power Wide Area Networks* (LPWAN): Power-hungry short-range wireless communication technologies (such as WiFi) is not suitable for long-range communication. LPWAN is a low-bit long-range communication technology, which is useful for power-constrained long-range IoT environments. Some of the examples of LPWAN are Narrow Band IoT, Sigfox, Neul, and LoRaWAN. In SWN, one of the use cases of LPWAN is WDN monitoring [42].

Table IV presents a comparison of various wireless communication technologies in terms of coverage, bandwidth, power consumption, etc. From comparative analysis, Bluetooth, Zig-Bee and WiFi are intended for short-range; whereas, LPWAN technologies are useful for long-range applications. Zig-Bee and WiFi offer better robustness as they support higher channel bandwidth compared to LoRA. Readers can refer to [43] for a comprehensive study of Wireless technologies, mainly, Bluetooth, UWB, ZigBee, and WiFi. The selection of proper wireless technology ensures a timely response with high reliability. Therefore, it is essential to deploy a suitable wireless technology for data acquisition, as per the application requirements to address the challenges posed by the voluminous data matrices of heterogeneous sensors.

D. Challenges

1) Heterogeneity

Major challenge of sensing unit in IoT arises due to heterogeneity of sensors. In SWN, sensor measurements methods are based on different approaches such as Voltammetry, Amperometry, Electro-optical, Biosensing, UV Spectrometry [44], etc. In brief, voltammetry is suitable for pH

TABLE IV: Comparison of Selected Wireless Communication Technologies*

Technology	Bluetooth	ZigBee	WiFi	LoRa
IEEE Spec	802.15.1	802.15.4	802.11a/b/g/n/ac/ax	802.11ah
Frequency Band	2.4 GHz	868/915 MHz; 2.4 GHz	2.4 GHz; 5 GHz	423 MHz, 868 MHz, 915 MHz, 923 MHz
Max Signal Rate	1 Mb/s	250 kb/s	9.6 Gb/s	50 kb/s
Nominal Range	50-80 m	10-100 m	100 m	10-12 km LOS
Nominal Tx power	(-)20 to (+)20 dBm	(-25)-0 dBm	15-20 dBm	0-13 dBm
Channel Bandwidth	1 MHz	0.3/0.6 MHz; 2MHz	22 MHz	125 kHz; 500 kHz

* Information in Table IV is subjected to change over time with improvements in technology. We advise readers to follow state-of-the-art specifications from relevant sources.

measurements, Electro-optical method is efficient for turbidity measurements, and biosensing is suitable to detect bacterial contaminants such as E.Coli [45]. Due to heterogeneous measurement methods, water quality sensors have different measurement cycles and time stamps, which makes the data acquisition process a challenging task. Such challenges can be addressed by introducing advanced microcontrollers. For example, in [46], authors introduce Arduino Mega 2560 microcontroller for integrating heterogeneous water quality sensors, mainly, pH, Temperature, Turbidity, EC, Light, and ORP. Similarly, in [47], heterogeneous water quality sensors are integrated using Raspberry Pi microcontroller.

2) Sensor Calibration

SWN requires multiple heterogeneous sensors and involves a geographically distributed set of dense sensors. The sensors in such systems tend to deviate from the actual measurements over time and require maintenance and periodic calibration. In general, calibration process is an offline method and may require physical interaction with the sensors. Physical interaction is a time-consuming and cost-inefficient process to resolve calibration issues. Therefore, the scientific community is exploring various ways to develop remote and online auto-calibration approaches. Auto-calibration can be defined as a method of online calibration without physical intervention, while leaving the sensors deployed in the field. Reference [48] proposes an ML-based method ML4CREST for the auto-calibration of the water flow sensor. Similarly, in [49], authors propose a method of auto-calibration for a Turbidity sensor.

3) Interoperability

Interoperability can be considered as a key for efficient management of SWN [50]. The heterogeneous IoT devices may operate over diverse protocols; having different data formats and structures, which require smooth cooperation and coordination. Interoperability facilitates smooth cooperation and coordination between heterogeneous devices of an IoT environment. However, Interoperability is a challenge for such IoT application, which is preventing the wide acceptance of IoT ecosystem.

4) Edge Intelligence

In a water CPS, ML performs analytics over data obtained from heterogeneous sensors. This analytics is performed in a High-level layer (as described in Section III-B) over the cloud; however, uploading such data over the cloud using IoT is inefficient in terms of bandwidth and resources. In contrast, Edge Intelligence process and analyze the data locally, and provide a platform to train and deploy an ML model in a local environment rather than cloud through embedded systems. For instance, embedded devices such as NVIDIA Jetson TX2 can be used to deploy an ML algorithm locally [51]. This approach may save important resources; however, it is still a major challenge for such embedded devices to run a large-scale complex ML model over the edge. Data scarcity, bad adaptability, and security issues are other major challenges of such devices.

5) Scalability and Reconfigurability

With ever-increasing expansion of SWN due to factors such as population growth, industrial demand, and environmental challenges, It is expected that IoT networks are scalable and reconfigurable. Here, we refer to scalable and reconfigurable as the adaptive ability of the network to evolve as per the changes in the SWN. The growth in industrial and urban water infrastructure goes through progressive stages, and therefore the IoT architecture is expected to be scalable and reconfigurable to address the challenges.

6) Limitations of wireless communication modules

Water CPSs are essentially data-driven systems. For timely operation, we require an efficient wireless communication module; however, wireless communication is constrained by power uses and data transmission capabilities.

7) Security

With the proliferation of communication networks, the IoT/CPS coverage is expanding to a wide geographical area. Such an IoT ecosystem frequently connects critical infrastructure such as WDN and Wastewater networks. Reference [77] points out the possible areas of CPS (Sensing, Communication, and Control), which are prone to attacks. Therefore, CPS is expected to have built-in mechanisms to tackle security challenges.

Summary:

The IoT can be seen as an enabling technology for CPS for the management of efficient data acquisition, and the merging of IoT with CPS into closed-loop is an important future challenge [18]. In this section, we reviewed the layers of IoT infrastructure and covered the IoT use cases for SWN.

TABLE V: Some of the use cases of supervised and unsupervised ML in SWN

ML	Algorithm	Applications
Supervised	k-NN	Water quality [52], pipe leakage [53], nutrient control in Aquaponics [54]
	SVM	Water demand forecasting [55], water quality [56], Aquaponics [57]
	Naïve Bayes	DO in Aquaculture [58], toxic compounds [59]
	Logistic Regression	Water contamination [60], pipeline failure [61]
	Decision Trees	Water quality prediction [62]
	Random Forest	Leak detection [63], water consumption monitoring [64], contamination detection [65]
	Bayesian Ridge Regression	Pipeline burst detection [66]
	Gradient Boosting	Water demand forecasting [67], Biological oxygen demand prediction [68], and flood level detection [69]
	Artificial Neural Networks	Water quality forecasting [70], water pollution estimation [71], DO prediction in aquaponics [72], water demand forecasting[73]
Unsupervised	k-means	Water quality analysis [74], wastewater treatment plant [75]
	Fuzzy C-means	DO control in a wastewater treatment plant [76]

Once, the CPS acquires data from IoT Infrastructure, it is expected to perform data analytics through advanced statistical techniques such as ML. In the next section, we review various ML techniques in the context of water CPS.

IV. Machine Learning

One of CPS's goals is to interact with the SWN via heterogeneous sensors and detect the presence of anomalies (such as contamination or leakages) in the system. The CPS observes real-time heterogeneous SWN parameters (such as water quality, physical, and chemical parameters) and detects unexpected changes in the parameters. Such unexpected changes may indicate the presence of an anomaly. The benefits of such observations include improved water quality monitoring, better control over nutrient presence, timely leak detection, improved pressure/flow management, and secure infrastructure. Despite significant advancements in online anomaly detection systems [10], controlling authorities require improved prediction models [78] to obtain inferences from the high volume of heterogeneous sensor data.

A. Motivation

According to Hawkins, "an anomaly is an observation that deviates so much from other observations as to arouse suspicion that a different mechanism generated it" [79]. Formally, given a sequence of observed data points $\boldsymbol{x}_t \in \mathbb{R}^n$, the objective of anomaly detection is to differentiate between normal and abnormal states, which can be denoted as $y_t \in \{0, 1\}$, where $t \in \{1, \cdots, T\}$ is the sample index in the time domain. Traditionally, the anomaly detection process was done in a lab. A user collects water samples from bodies of water and processes them using traditional lab-based techniques. The work presented in [80] summarizes these traditional lab-based techniques. These techniques are, however, not very effective for monitoring dynamic SWN, such as geographically distributed WDN, Aquaponics, and industrial water networks. An anomaly in such networks can occur for a variety of reasons, including contamination incidents, leakage incidents, and so on. There is a need to develop appropriate inference methods that can detect anomalies in such dynamic networks in real-time and then learn models from the data to explain why an anomaly exists.

Machine learning (ML) techniques are specialized computing methods that can be used to predict and detect anomalies in such SWNs. ML works by utilizing the statistical properties of data from heterogeneous sensors to generate intelligent inferences. Anomalies can be predicted and detected using such inferences. ML could also capture the nonlinear dynamics of the water environment, which are posed by flow, and pump constraints. Some of the recent applications of ML algorithms in SWN are contamination detection, water quality analysis, identifying the correlation of physical and chemical parameters, development of a decision support system, detecting pressure-flow inconsistencies, real-time leakage detection, dissolved oxygen control, and nutrient monitoring. In addition, ML can also be used to develop predictive and autonomous event-triggered pressure and flow control algorithms. ML algorithm can be classified as supervised, unsupervised, or reinforcement learning [81]. In the following section, we provide an overview of various ML algorithms that can be used in the context of a water CPS.

B. Supervised ML

Supervised ML is the most common ML methodology to detect anomalies by using a set of labeled data. In supervised ML, the objective of the algorithm is to learn a mapping function between input variables $\boldsymbol{x} \in \mathcal{X}$ and output variable $y \in Y$ such that $f{:}\mathcal{X} \rightarrow \mathcal{Y}$, where the output variable y can be predicted. *Classification* and *Regression* are two main subcategories of supervised ML techniques. In *Classification* the output variable y is categorical (discrete), whereas in *Regression* the output variable is continuous. There are various supervised ML algorithms available in the literature, and readers can refer to [82]. The following are the most important supervised ML algorithms in the context of water CPS:

- k**-nearest neighbours (**k**NN) -** The k-nearest neighbor algorithm is a well-known class of ML algorithms for classification and regression. The underlying assumption of kNN is that similar data points occur adjacent to each other. For each unlabelled query sample, the algorithm finds k number of nearest training samples, which are labeled. The most frequent label from these k neighbors is assigned as the label of the query sample in classification problems, whereas the average of the neighbor labels is assigned to the query point in regression problems. The optimal value of k can be specified by the user or learned. For example, [83] proposes a tenfold approach for cross-validation to obtain optimal k values. The applications of kNN in SWN are to classify drinking water quality, predict water pollution index [52], detect water pipe leakage [53], and so on. Reference [54] uses kNN to control nutrient levels in aquaponics. One of the limitations of a traditional kNN algorithm is the time-consuming process of manually setting k values. Furthermore, as the volume of data increases, this algorithm becomes computationally expensive in terms of time and memory.
- **Support Vector Machine (SVM) -** SVM is a robust supervised ML algorithm for classification and regression, developed based on Vapnik–Chervonenkis (VC) theory. SVM is commonly known as a large-margin classifier as it relies on the decision boundaries, which are hyperplanes having the largest distance to the support vector (the nearest training sample) of any class (see Fig. 5), resulting in low generalization error. Although the original algorithm is proposed to develop linear classifiers, the key attractiveness of SVM is that the idea of the maximum-margin hyperplane can be extended to construct nonlinear decision boundaries by invoking kernels. The typical procedure involves mapping the original finite-dimensional space of data points to a higher-dimensional feature space using a suitable kernel function such that the nonlinear classification can be performed by constructing a hyperplane-based linear classifier in the transformed feature space.
  Some of the typical SVM kernel functions are linear, polynomial, sigmoid and radial basis function (RBF). RBF is the most commonly used kernel, given by $\boldsymbol{k}(\boldsymbol{x}, \boldsymbol{x}_i) = exp(-\gamma||\boldsymbol{x} - \boldsymbol{x}_j||^2)$ where $\boldsymbol{x}$ is the data vector that belongs to a binary class y and the parameter γ controls the over-fitting or under-fitting [84].
  SVM is a leading pattern classification and function approximation technique because it reduces estimation error, and is less prone to overfitting. SVM is used in [55] for hourly water demand forecasting. In [56], authors use SVM to classify water quality, and in [57], it is used to evaluate observation sensors in an Aquaponics plant.
- **Naive Bayes -** kNN and SVM are discriminative ML models, whereas Naive Bayes is a generative ML model [85]. For a given input $\boldsymbol{x}$ and the corresponding label y, the discriminative models are designed to learn the

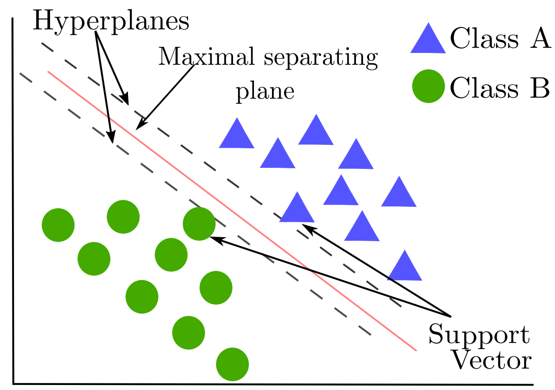


Fig. 5: Binary SVM Classification

probability distribution $\Pr(y|\boldsymbol{x})$. Whereas the generative ML model estimates the joint probability $\Pr(\boldsymbol{x},y)$, and applies the Bayes theorem to obtain $\Pr(y|\boldsymbol{x})$. This algorithm is based on the assumption that features are independent of one another. Reference [58] predicts the DO in an aquaculture plant using Naive Bayes. In [59], authors predict the presence of lead components using Naive Bayes.
- **Logistic Regression -** Logistic regression is a supervised ML technique based on logistic function. This ML technique indicates the presence of anomaly through binary decision variables such as 0/1 or yes/no. Detection contamination [60], pipeline failure [61], etc., are some of the WDN applications of logistic regression.
- **Decision Trees -** Due to its efficiency in addressing large scale regression tasks, the decision tree is one of the most widely used class of supervised ML. Decision tree consists of two main elements: *nodes*, representing features and *branches*, representing division rules. Typically in a decision tree, starting with the first *node* i, features of the training data $\{\boldsymbol{d}_i\}$ is evaluated to split the observation into two *branches*, which ends at child nodes. This process is followed recursively [86]. In [62], authors used hybrid decision tree for water quality prediction. .
- **Random Forest -** Random forest (RF) is an extension of the decision tree supervised ML approach. Decision trees are sensitive to minor changes in data sets, which can result in an inaccurate prediction. RF compensates for this shortcoming by combining multiple decision trees and producing an average of involved decision trees. RF addresses the issue of missing data [87], overfitting, and is noise immune [62]. Paper [88] evaluates the performance of 179 classifiers and concludes that by parallelizing RF implementation, users can achieve significantly higher classification accuracy than their counterparts. RF applications include leak detection [63] and contamination detection [65].
- **Bayesian Ridge Regression -** Bayesian ridge regression merges the foundation of Bayesian probabilistic method with ridge L_2 regularization. This approach

is particularly suitable to address challenges that arise from multicollinearity issues. Multicollinearity refers to a situation in which explanatory variables are linearly dependent. The author of [66] present a use case of Bayesian Ridge regression in order to detect bursts in a pipeline. Also, the authors estimate the short-term water demand using this approach.

- **Gradient Boosting -** Gradient boosting is a ML technique for classification and regression. Boosting in this context refers to a method of combining a group of weak learners (e.g., decision trees). The underlying assumption is that weak learner performance is marginally better than random guess and that an ensemble of weak learners can significantly improve ML model performance for classification and regression tasks. Gradient Boosting algorithms are greedy, and they tend to overfit the training dataset. To avoid overfitting, various regularization methods can be used to penalize the parts of the algorithm that perform poorly. Water demand forecasting [67], Biological Oxygen Demand prediction [68], and flood detection [69] are some of the applications of Gradient Boosting.
- **Artificial Neural Networks -** ANN models are highly flexible function approximators that can be used to solve a wide range of classification and regression problems. ANN is inspired by the human brain's structure, and its processing and learning abilities. The mathematical model of an artificial neuron is presented in Fig. 6a. As shown in Fig. 6a, the synapses provide weights w_i to the inputs x_i for $i = 1, 2, \cdots, m$. Adder generates $v = w_0 + \sum_{i=1}^{m} w_i x_i$. At the output, $g(v)$ maps (typically, using a nonlinear function) the sum of weighted inputs v to the output of the neuron.

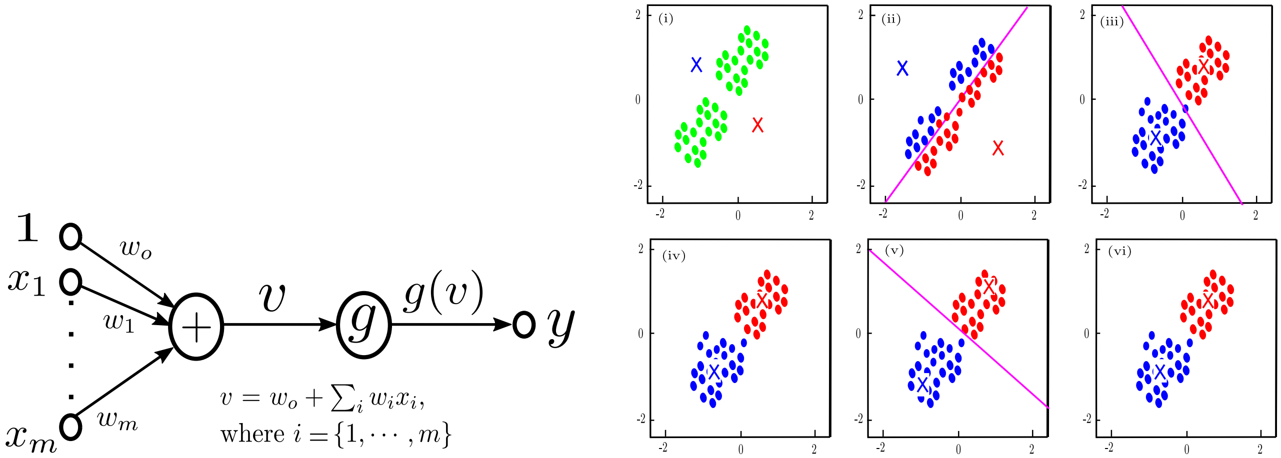
  Water quality forecasting [70], water pollution estimation [71], dissolved oxygen prediction in aquaponics [72], etc., are some applications of ANN. Authors of [73] use ANN to model the short-term water demand. The authors conclude that the proposed ANN-based method outperformes the other short-term demand forecasting methods such as regression and time series models. Author of [89] performed a comparative analysis of ANN against SVM for predicting time-series of water demand and concluded that the ANN has significantly better generalization capability compared to SVM. For a detailed review of ANNs for SWN applications, readers can refer to [90].

C. Unsupervised Learning

Supervised ML methods are efficient and robust, but they require 'labeled' data for training. However, data labeling is a time-consuming and laborious process. For example, labeling the presence of *e.coli* is a time-consuming analytical measurement process because *e.coli* detection is only possible based on bacterial growth. Furthermore, as the network's dimensions and the number of distributed heterogeneous sensors grow, the various sensor data matrices grow voluminous, making the labeling process prohibitively inconvenient. Unsupervised ML is an alternative choice to learn the underlying structure in a dataset.

Clustering is the most important type of unsupervised learning, with the goal of classifying data using a finite set of clusters [91]. Clustering is based on the assumption that normal data instances belong to a large or dense cluster, whereas anomalies do not belong to any cluster. Clustering has been extensively tested in the evaluation of water quality analysis [92]. In the following subsection, we discuss some of the most common clustering algorithms and their applications in industrial and urban water environments such as WDN and Aquaponics.

- **K-means -** K-means algorithm is used to partition n data samples into K clusters such that the inter-cluster variance is high and intra-cluster variance is low. The algorithm iteratively computes K centroids (means) corresponding to K clusters, and in each iteration, the samples are clustered by computing the closest centroids. Figure 6b shows a graphical representation of K-Means clustering. When the clusters in the dataset are distinct or well separated, K-means clustering performs well. Furthermore, in terms of computational complexity, K-means is efficient. This method is useful for applications such as enhanced water quality analysis [74] and decision support for wastewater treatment plant development [75].
- **Fuzzy C-means -** The K-means algorithm performs well when the dataset is distinct; however, the K-means algorithm fails to find overlapping clusters. This issue can be addressed by modifying the K-means algorithm by adopting a 'soft' strategy for the cluster membership, which is referred to as fuzzy C-means or soft K-means. If a data object is associated with overlapping clusters, a fuzzy parameter is assigned to determine the degree of associativity to a cluster. Since the water quality parameters are correlated, this approach provides the degree of data point associativity to a cluster. In [93], authors use this approach for water quality analysis in the Niharu dam reservoir. Another application of fuzzy C-means can be found in the predictive control of the dissolved oxygen model in wastewater treatment plants [76].
- **Manifold Learning -** Dataset from geographically distributed SWN may contain irrelevant and correlated features. Dimensionality reduction improves the performance of an ML model by extracting relevant features from the dataset and discarding the irrelevant and correlated features. Traditionally, linear approaches (such as principal component analysis) were used for the dimensionality reduction; however linear dimensionality reduction approaches are inefficient, and can not extract the relevant features adequately from complex and nonlinear data. Manifold learning is an unsupervised ML approach to extract features from complex nonlinear datasets. Semidefinite Embedding, Isomap, Laplacian



(a) (b)

Fig. 6: (a) depicts the schematic representation of a single ANN neuron, and (b) presents the clustering through K Means algorithm.

Eigenmap, and Local linear Embedding are the major techniques of Manifold learning. More recently, Manifold learning is recommended to use along with K means clustering to improve the overall model accuracy [94]. The challenge associated with Manifold Learning is that it is prohibitively expensive in terms of computational time for large-scale problems.

Unsupervised learning provides valuable insights into data by identifying potential clusters or groups to which data points may belong. One significant disadvantage of this approach is that, while the algorithms are trained to detect clusters, they are not trained to detect anomalies. Furthermore, because unsupervised learning is prone to suboptimal solutions, it necessitates careful hyperparameter tuning. To avoid the challenges of unsupervised learning, researchers in some applications use a 'unlabeled' data set in conjunction with a small amount of 'labeled' data to improve the overall ML model accuracy. This method is referred to as the semi-supervised ML approach. In [95], authors used a semi-supervised ML approach to develop a risk warning system for chemical hazards in drinking water applications.

1) Performance Matrices

The accuracy of an ML model can be calculated using various performance matrices such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Mean Arctangent Absolute Percentage Error (MAAPE). The RMSE measures how well a regression model fits a data point. Furthermore, using RMSE, the user can examine the similarity of estimated values to actual data. MAE can be used to calculate the difference in predicted and observed data points. MAE is scale-dependent and does not provide information about the direction of error. MAPE is an alternative to MAE that provides an intuitive interpretation of the error between observed and estimated data points. The observed data x_t and estimated data $\hat{x}_t$ can be compared in terms of MAPE as follows:

$$\text{MAPE} = 100\% \times \frac{1}{n} \sum_{t=1}^{n} \Big|(x_t - \hat{x}_t)/\, x_t\Big| \qquad (1)$$

When time series have zero or near-zero values, it is preferable to use other metrics, such as Mean Arctangent Absolute Percentage Error (MAAPE) [96].

D. Challenges

1) Real-Time Adaptive Reconfigurability

Successful real-time implementation of ML methods and real-time adaptive reconfigurability for such CPS are still open challenges. Nowadays, the acquired data from sensor arrays are processed mostly offline, since the training of such ML models relies on data sets that are obtained offline, hence can be termed as offline methods. However, to represent a holistic development of water CPS, CPS are expected to be adaptive, and reconfigurable in real-time [5].

2) Online contamination detection

The existing ML algorithms provide an adequate framework of contamination detection in an offline fashion. Such ML algorithms process the data in batches. However, Water CPS are envisioned to exercise real-time control in application scenarios that require online detection of contamination, and therefore, ML algorithms are expected to acquire and process real-time data streams of water quality parameters. Acquiring the real-time data set from all the possible water quality parameters is complex [27], as not all the sensors provide real-time observation of targeted parameters (e.g. E.Coli sensors). Therefore, integrating online contamination detection capabilities in such water CPS requires real-time observation, and further research is required to develop state-of-the-art methods, which could observe the state of the targeted parameter in real-time.

Summary: CPSs are expected to be designed in such a way that they can detect anomalies and then apply control actions via actuators. This section discusses the various supervised

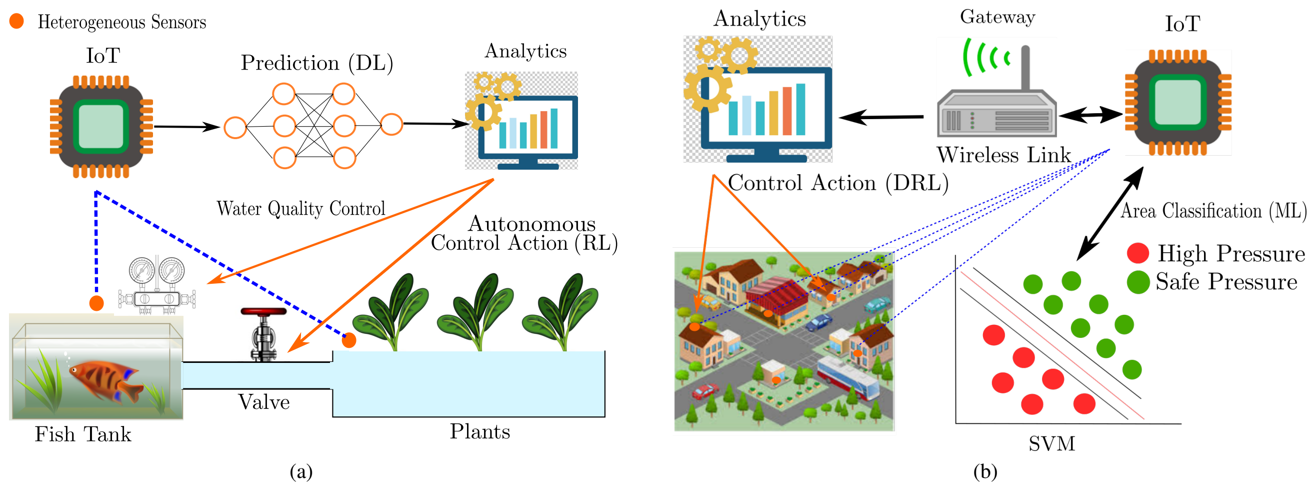


Fig. 7: (a) depicts the DL-RL Model for Autonomous Control in an Aquaponics system (b) presents the application of SVM-DRL model for autonomous water pressure management in a WDN.

and unsupervised ML algorithms that can be used to detect anomalies in industrial and urban water CPS applications scenarios. In the following section, we cover the challenges that must be addressed in the design of CPS in order to introduce autonomous control capabilities.

V. Control

Control is an integral component of any SWN, and coordinated control of pumps, valves, water quality parameters, etc., are highly desirable in SWN in order to prevent the unexpected occurrence of anomalies. For instance, due to accidental water pipe leakages in a WDN, it is estimated that water authorities lose a significant amount of water globally [97]. The authorities require an Intelligent control method to systematically detect an anomaly (such as leakages and contamination) in a WDN, and autonomously control the various elements of a WDN (such as valve states, water flow, pump speed, etc.), without human intervention. Figure 7a and Fig. 7b depict the use cases of control applications derived from ML in an Aquaponics and WDN, respectively.

However, integrating autonomous and intelligent control capabilities in a CPS is an important design challenge, as it requires close interaction between sensors, actuators, and parameters of the physical world [98]. The main contribution of this section is to capture the control aspect for such CPS by answering the questions: how the problem of control can be defined and how the autonomous control formulations can be integrated into targeted water CPS. In Section V-A, we cover the traditional offline model-predictive control formulations and the challenges posed by such formulations. In Section V-B, we cover the data-driven control methods; a promising approach to integrate autonomous control capability in SWN. We also intend to cover the existing works addressing the autonomous control approaches in order to optimize the hydraulic, physical, and chemical parameters of SWN.

A. Model-Predictive Control

MPC is one of the leading approaches for the operational management of water ecosystems for diverse applications, such as flow management, pipeline pressure management, chlorine management, nutrients management in Aquaponics, etc. MPC is a model-driven control approach, which relies on a *system model*, where the *system model* presents a mathematical and logic-based representation of the physical components of SWN. Some of the frequently used benchmarks of *system model* in a WDN are Anytown, New york city tunnel, and Two reservoir model; whereas, in Waste water networks, the commonly used system model benchmarks are Mays and Wenzel, and Li and Matthew [99]. In such SWN, the primary objectives of MPC are to (i) identify a set of optimal operating points for operational management, and (ii) compute a time-series control trajectory for pump and valve control through suitable optimization formulations. Constructing a suitable optimization formulation requires prior information of water flow distribution, physical dimensions, properties of various components, uncertainties caused by the parameters, optimization objectives, and network constraints.

1) Optimization Formulation

The goal of providing an optimization formulation is to identify an optimal set of points for a given Objective (e.g. minimization of different types of costs) of interest, under a given set of hydraulic constraints. A typical SWN optimization framework is given by:

$$\begin{aligned} \text{maximize/minimize} \quad & f(\boldsymbol{x}) \\ \text{s.t} \quad & \boldsymbol{x} \in \mathcal{X}, \end{aligned} \tag{2}$$

where $f(\boldsymbol{x})$ is an objective function and $\mathcal{X}$ is a constraint set. The SWN objective function $f(\boldsymbol{x})$ is usually formulated for (a) minimizing the pipe cost in a network, (b) minimizing or maximizing the flow and pressure in the network, (c) optimizing the consumer water demand, (d) scheduling optimal water dispatch, (e) minimizing the cost incurred due to dynamic energy pricing, (f) managing water quality

TABLE VI: Head loss equations*

Formula	*Head loss* $(h_i - h_i^{'})$	Coefficient
Hazen-Williams	$C_{\rm HW}\text{sign}(f_\ell)(f_\ell)^{1.852}$	$C_{\rm HW} = 4.727 {K_1}^{-1.852} {d_\ell}^{-4.871} l_\ell$
Darcy-Weisbach	$C_{\rm DW}\text{sign}(f_\ell)(f_\ell)^2$	$C_{\rm DW} = 0.02 K_2 {d_\ell}^{-5} l_\ell$
Chezy-Manning	$C_{\rm CM}\text{sign}(f_\ell)(f_\ell)^2$	$C_{\rm CM} = 4.66 K_3^2 {d_\ell}^{-5.33} l_\ell$

*Here, i and $i^{'}$ are the consecutive nodes in water distribution networks and ℓ is the physical connection (pipes) between consecutive nodes. $C_{\rm HW}$, $C_{\rm DW}$, $C_{\rm CM}$ are the coefficients of Hazen-Williams, Darcy-Weisbach and Chezy-Manning. d_ℓ (ft) is the pipe diameter, l_ℓ (ft) is the pipe length. f_ℓ is the flow rate. K_1, K_2, and K_3 are the friction factor of Hazen-Williams, Darcy-Weisbach and Chezy-Manning respectively.

parameters, etc. Readers can refer to [100], which covers a diverse range of SWN objective functions.

In the existing literature, optimization formulations have been proposed and solved to address different control objectives, such as pump scheduling [101], valve operations [102], chlorine dispatch [103] and operational management [104]. In order to solve such optimization formulations, which happen to be usually highly non-convex problems, heuristics-based solvers are a popular choice of methods, which search for an optimal solution by considering an initial guess over a set of control points. Genetic algorithms (GA), Simulated annealing (SA), Branch-and-bound, and Tabu search (TS) are examples of heuristic-based methods and have been experimented with in large-scale WDN [105]. However, the constraints posed by the components of SWN bring a significant challenge for such solvers. In the next subsection, we cover the major constraints in SWN optimization formulations.

2) *Constraints*

The water flow in a typical SWN is governed by the *hydraulic* constraints. Such hydraulic constraints are imposed by the integral components of SWN such as tank dynamics, head loss equations (Hazen-Williams, Darcy-Weisbach, and Chezy-Manning), valves state, variable and fixed speed pumps, etc. Often, the constraints imposed by the model components makes the control formulation non-convex and in most cases, $\mathcal{NP}$-Hard [8]. Finding an optimal solution or close-to-optimal solutions to these problems is computationally expensive in terms of memory and time. The challenges posed by the constraints are discussed below:

a) Non-convexity of head loss equations

Empirical head loss equations, presented in Table VI, are the commonly used equations to model the water flow rate with the physical dimensions (e.g., pipe capacity, tank capacity, etc.) of the circuit. However, solving an optimization problem involving empirical head loss constraints is challenging due to its non-convex nature. The non-convexity is attributed due to presence of the non-convex $sign$ function. Some of the techniques to address the non-convexity of head loss equations are linearization [106], Big-M [101] and Geometric Programming [107].

b) Computing the water flow distribution

Water flow management is a major control objective in SWN. Computing flow distribution is a necessary step for efficient pressure management in a network, which requires prior information of the *network type*. The *network types* can be characterized as *Branch or Loop networks.* In a *branched* network, the optimal water flow distribution can be computed uniquely, given the availability of water outflow at nodes, whereas, in *loop* network, the flow can take multiple paths to reach from source to destination [108]. In such *loop* networks, computing flow distribution requires iterative methods as described in [109]. Another parameter to compute flow distribution is based on whether the flow in the water network is assisted by gravity or by the pumps. In a *gravity-fed* SWN, the MPC control objective is to manage the optimal water flow and maintain the necessary pressure in the nodes given the constraints posed by the pipes, tanks, valves, etc. Whereas, in a *pump-fed* SWN, the control objective is to solve the flow distribution and identify the optimal trajectory of the pump and valve scheduling under network constraints. However, computing the flow distribution in a pump-fed SWN is more challenging than in the *gravity-fed* SWN, as the constraints posed by pump and valves are integers [101].

c) Network Layout

The control formulations require also precise information of additional model components, mainly dimensions of tanks and pipes, pump capacity, valve states, head (geographical elevation), flow rate, etc. The interconnection between these components is often modelled using a state-space model. Given a large sized network with numerous components, the major challenge is to develop a suitable state-space model, which reflects the complexity of the original physical process and could estimate the parameters of interest as realistic as possible.

d) Demand and supply stochasticity

The water demand poses an important constraint in a SWN optimization formulation. The water demand forecasting adds stochasticity in the optimization formulation, which is challenging to tackle for existing solvers. Considering the above aspects, various scientific efforts have proposed methods for water demand forecasting. Traditionally, water demand forecasting relies on regression and time series analysis. In [127], an optimization formulation is proposed to minimize the chlorine dispatch in a WDN, where the water demand is computed every six hours. Similarly, in [128], authors propose an optimization formulation to minimize the operational cost of pump switching, where the water demand forecasting is computed every twenty-four hours using a hybrid dynamic neural network.

e) Integer constraints imposed by the pumps and valves

In SWN, pump and valve management is crucial for optimal control. In such application scenarios, it is expected that the decision variable of a pump and valves' states are constrained to hold binary or integer values. For instance, valve states can

TABLE VII: Use cases of Data-Driven ML for control in SWN

ML	Reference	Architectures	Applications	Case Study
DL	[110], 2020	Hybrid CNN-LSTM	Short Term Water quality prediction	Prespa Basin, Europe
	[111], 2018	Hybrid CNN-SVM	Pipe leakage detection	Testbed, Seoul, South Korea
	[112], 2019	DenseNET (CNN)	Pipe Burst detection	Anytown Network, EPANET
	[113], 2021	-	Aquaponics	Review
	[114], 2018	Autoencoder-LSTM	Water Quality Prediction	-
	[115], 2021	LSTM	Optimal Pump control	Simulation
	[116], 2020	RBM	Dissolved Oxygen Prediction	Recirculating Aquaculture
	[117], 2020	CNN	Streamflow Projection	California, USA
	[118], 2017	Hybrid DNN-SVM	Anomaly Detection	Testbed (Water Treatment Plant), Singapore
RL	[119], 2020	Multi Critic	Control of Water Tanks	Simulation
	[120], 2007	Q Learning	Water demand management, and optimizing hydropower	Geum River Basin, South Korea
	[121], 2002	Q Learning	Operational Management of a Water System	Lake Como, Italy
	[122], 2017	SARSA	Irrigation Management	Multiple locations, USA, India and Australia
DRL	[123], 2020	Deep Q Network	Pump Speed Control	Anytown and D-Town, EPANET
	[124], 2012	RL	Dissolved Oxygen control	Wastewater Treatment Plant
	[125], 2020	Deep Q Network	Online control of storm SWN	Simulation
	[126], 2020	Soft Actor-Critic	Hydropower Production Scheduling	Simulation, Norwegian Power Stations

be formulated through binary variables on/off or 0/1 [129]. The optimization formulation having decision variables that are constrained to be integers are termed as Mixed-Integer Optimization (MIO) formulation. However, introducing integer decision variables in optimization problems makes the problem non-convex and usually it is an $\mathcal{NP}$-hard problem. Solving such $\mathcal{NP}$-hard non-convex optimization problems are computationally expansive in terms of time and memory requirements [130]. Mixed-integer (MI) solvers, such as GUROBI, CPLEX, and LPSOLVE [131] can be used to solve MIO formulations. However, the computational complexity of such MI solvers grows exponentially as the number of network components, such as pumps and the valves, grow as the size of network expands. Therefore, to address the challenges posed by the integer constraints, studies recommend various approximations, linearization, and relaxation techniques. For instance, in [132], a piece-wise linear approximation technique is used to relax the constraints and compute the optimal solution of MIO formulation. Another work that address the $\mathcal{NP}$-hard MIO formulation in SWN is provided in [133].

Formulating an MPC-driven approach for the operational control of SWN is a challenging task. The linearization and the relaxation techniques, which are often used to tackle the non-convexity of the SWN constraints, usually results in a sub-optimal performance. In addition, some of the optimization formulations are inefficient in terms of computation time and memory, which may lead to interoperability challenges among various components of a CPS. Therefore, the recent focus shifts nowadays to integrate data-driven control techniques for the operational control of SWN. In the next section, we review the Data-driven control techniques in SWN and discuss how such techniques can be integrated in a water CPS.

B. Data-Driven Control

Data-driven control (DDC) is an alternative technique for introducing control in SWN. In DDC, the controller's objective is to learn to apply a coordinated sequence of control actions from the acquired dataset of the targeted system. ML is an important element of DDC to detect the presence of an anomaly (discussed in Section IV) and introduce control capabilities in a given SWN. From the perspective of control, ML can be introduced in SWN to (a) detect the presence of an anomaly, (b) reduce the computational complexity, and (c) compute a sequence of optimal control actions from the input and output dataset [134]. State-of-the-art ML techniques such as Deep Learning (DL), Reinforcement Learning (RL), and Deep Reinforcement Learning (DRL) have been successfully applied to integrate the process from anomaly detection to compute optimal control actions for real-life applications such as Robotics, Finance, and Self-driving cars. Table VII presents the use cases of DL, RL, and DRL in different SWNs, and we describe the aforementioned techniques in the next subsections.

1) Deep Learning

DL is an ML technique based on ANN (described in Section IV). DL can be supervised, unsupervised or semisupervised. In the application scenarios of water CPS,

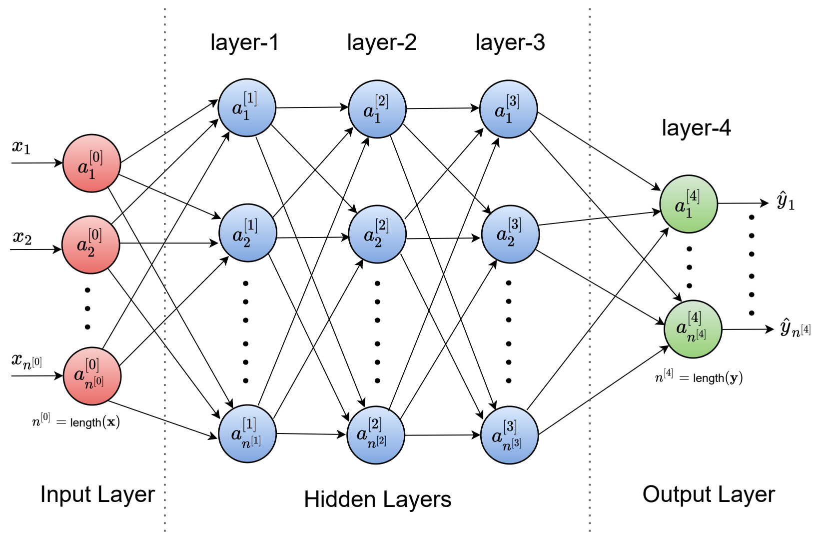


Fig. 8: Feedforward architecture

the most important DL networks are Feedforward Neural Networks and Recurrent Neural Networks. In the next subsections, we cover the description and use cases of such important DL networks:

a) Feedforward Neural Networks

Feedforward Neural Networks are one of the most commonly used DL network. Given the input-output training pairs $\{(\boldsymbol{x}^{(k)}, \boldsymbol{y}^{(k)})\}_{k=1}^m$, the network learns a function that maps the inputs to the outputs. Feedforward Neural Network is made up of several fully connected layers termed as an input layer, hidden layers, and an output layer. Each layer is made up of multiple nodes, and the output of one layer's nodes is connected to the input of the next layer's nodes. There are no feedbacks or loops in such networks, and information flows in only one direction through hidden nodes. Fig 8 shows the architecture of a Feedforward Neural Network, having $L = 4$ layers with $n^{[l]}$ number of neurons (nodes) in layer-l. In a typical neural network, each node in a layer receives inputs from the previous layer, computes a nonlinear activation, and then passes the activation to the next layer. Let $\boldsymbol{a}^{[l]} = (a_1^{[l]}, a_2^{[l]}, \dots, a_{n^{[l]}}^{[l]})^\top \in \mathbb{R}^{n^{[l]}\times 1}$ be the activation of layer-l, where $a_i^{[l]}$ represents the activation of node-i of layer-l, given by

$$\boldsymbol{a}_i^{[l]} = \sigma(\mathbf{w}_i^{[l]\top}\boldsymbol{a}^{[l-1]} + b_i^{[l]}). \tag{3}$$

In (3), the parameters $\mathbf{w}_i^{[l]} \in \mathbb{R}^{n^{[l-1]}\times 1}$ and $b_i^{[l]} \in \mathbb{R}$ are respectively the weight and the bias of the node-i of layer-l, and σ is a non-linear activation function. Some of the commonly used activation functions are rectified linear unit (ReLU), sigmoid, and tan hyperbolic.

The neural network training, i.e., the learning of $\mathbf{w}_i^{[l]}$ and $b_i^{[l]}$, are accomplished through forward and backward propagation steps. In forward propagation, given a training pair $(\boldsymbol{x}, \boldsymbol{y})$, the activations of all the nodes of the network are calculated using (3) with $\boldsymbol{a}^{[0]} = \boldsymbol{x}$, yielding an estimate of the output $\boldsymbol{a}^{[L]} = \widehat{\boldsymbol{y}}$. During backward propagation, a loss function $\sum_{k=1}^m \mathcal{L}(\boldsymbol{y}^{(k)}, \widehat{\boldsymbol{y}}^{(k)})$ is formulated by considering all the m training pairs and is optimized for the parameters $\mathbf{w}_i^{[l]}$ and $b_i^{[l]}$. The forward and the backward propagations are done iteratively until convergence by updating the parameter values with the optimized values in each iteration. Feedforward network architectures can be constructed in a diverse predefined methods such as Convolutional Neural Networks (CNN), Residual Networks, and Radial Boltzmann Machine (RBM). Next, we focus on the CNN and RBM architectures, which have been employed in the context of SWN.

CNN is an extension of feedforward neural networks. This architecture is particularly useful to extract the underlying special features from the datasets. CNN architecture consists of a sequence of layers, mainly, *Input layer*- to hold the data points, *Convolutional layer*- to extract the features, *Pooling layer*- to reduce the amount the parameters and computations in a network, and *Fully connected layer*- to assign dataset to the relevant class. CNN has a wide range of applications in scientific domains mainly image classification and computer vision. In SWN, the CNN has applications in water quality prediction [110], pipe leakages detection [111], streamflow projections [117], etc.

RBM is a generative ML model and is particularly useful to learn the probability distribution over the set of inputs. A key difference of RBM with its counterparts is that input nodes have connections among themselves. RBM has important applications in Aquaculture and Aquaponics. One such application can be found in [116], where the authors propose a prediction model of dissolved oxygen in aquaculture. Authors of [135] use a continuous deep belief network (a variant of RBM) to predict the hourly demand of water consumption.

b) Recurrent Neural Networks

Recurrent Neural Networks (RNN) architectures are typically designed to process sequential datasets. In contrast with Feedforward Neural Networks, RNN are trained to process a sequence of values such as $\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \cdots, \mathbf{x}_{(T)}$. Fig 9 shows the architecture of a RNN, $\mathbf{x}_t$ is the input at time

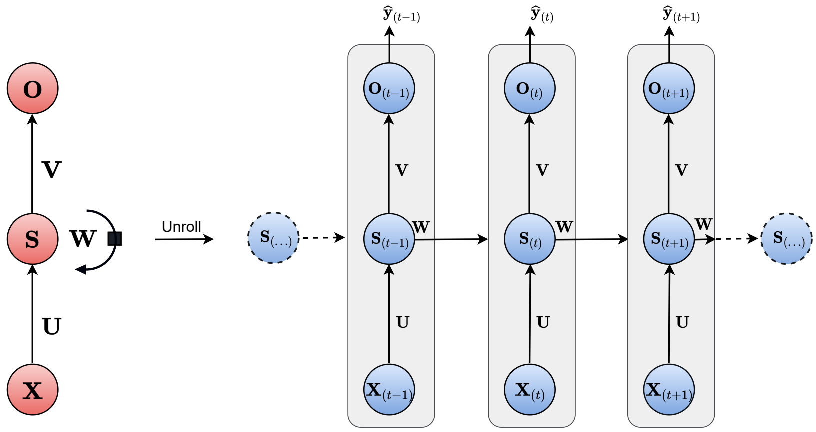


Fig. 9: RNN and its unrolling

t, $\mathbf{S}_{(t)}$ is the state, $<$ **U, V, W** $>$ are the learned parameters, and $\mathbf{o}_{(t)}$ is the output. To process such sequential information, RNN updates the current state based on past state and current input data. Long short term memory (LSTM) is a major architecture of RNN. LSTM is expected to improve the time-series water demand forecasting as compared to supervised ML techniques such as SVM and Random Forest. LSTM is used in [96] in conjunction with IoT to derive a time-series trajectory of water demand forecasting. Similarly, in [115], LSTM is used for water demand forecasting in order to develop a computational framework of pump scheduling.

DL have the capability to address a huge volume of dataset as compared to techniques addressed in Section IV. One such comparative study can be refereed from [118], which compares DL with one-class SVM over time-series data obtained from water CPS. The overall performance analysis evaluated using F-measures, and findings concludes that DL has better performance metrics as compared to one class SVM.

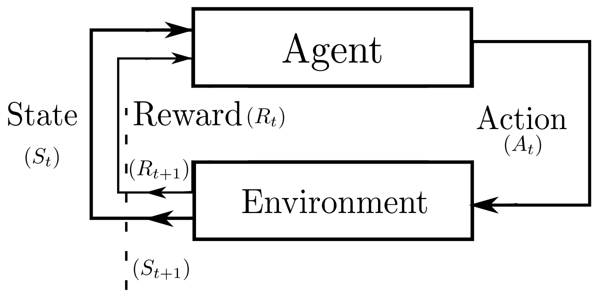
2) Reinforcement Learning

Reinforcement learning (RL) is a class of ML algorithms, primarily designed to integrate control capabilities in real-time applications (e.g., driverless cars, autonomous robotic control, etc.) [136], where there is an intelligent agent that learns from experience to take actions optimally to maximize some long-term objective function (optimal policy), called usually expected return. RL is distinct from supervised and unsupervised ML, as it does operate neither with 'labeled' nor 'unlabeled' data pairs, but only receiving only partial feedback signals from the environment. One of the motivating applications of RL in SWN is the management and control of a large-scale WDN in real-time, which may comprise several components such as tanks, reservoirs, pumps, valves, etc. For a given WDN, the controlling authorities aim to achieve optimal control of active hydraulic elements (pumps, valves) satisfying water demand and maximizing some utility function. For instance, the frequent switching of pumps is not desirable for water networks and active hydraulics. RL provides a control framework for such networks [119].

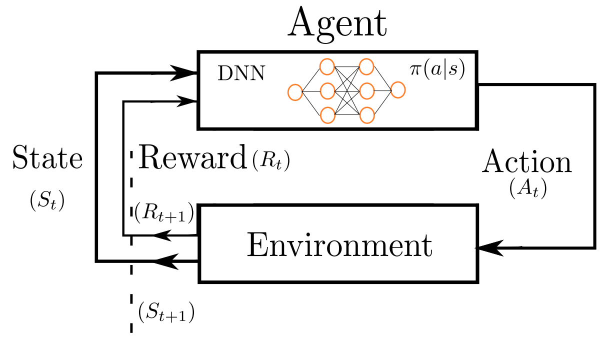
Previously to the recent wave of modern reinforcement learning, dynamic programming (DP) was exploited by the researchers to apply control over SWN [137]. DP is a *model-based* approach and requires exact knowledge of the environment to generate a sequence of optimal control actions. One of the essential aspects of DP is the requirement of discretization of continuous space. This discretization is feasible for uni-dimensional spaces (e.g. single reservoir operation); however, discretization is laboursom and computationally expensive for multidimensional continuous spaces (e.g. multiple interconnected tanks). This phenomenon is famously known as Bellman's "curse of dimensionality", which means that the volume of the computations increases exponentially by adding the extra dimensions to euclidean space [138]. This is an unrealistic requirement in the scenario of SWN given that such systems involve various interconnected tanks, pumps, valves, sensors, etc. Hence, research focus shifted to explore alternative *model-free* optimal control approaches for SWN, and RL promises to address such challenges.

The theory of RL is developed under the Markov decision process (MDP) assumption, which is a classical formulation for sequential decision making. MDPs can be considered as mathematically idealized form of RL. A typical MPD (or RL) cycle is depicted in Fig. 10a. The agent is the element of the RL formulation that interacts with the environment and learns to control the input components of the environment. An agent is "anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators" [139]. The agent learns the control actions by maximizing a reward function through an optimization formulation. In general, the objective of such formulations is to find a time-series sequence of control actions or optimal control policy, which produce an optimal action [140].

At each time step t, given that the agent is at a certain state $S_t \in \mathcal{S}$, it applies a certain action $A_t \in \mathcal{A}$, and as a consequence, it evolves to another state $S_{t+1} \in \mathcal{S}$ receiving a reward signal $R_{t+1} \subset \mathbb{R}$. In a finite MDP, the sets $\mathcal{S}$, $\mathcal{A}$, and $\mathcal{R}$ have finite number of elements. In a typical MDP framework, the random variable S_t has a well defined probability distribution (non necessarily known a-priori by the agent), which depends only on the preceding state and actions



(a)



(b)

Fig. 10: (a) Agent–Environment interaction in RL (b) DRL Architecture.

(Markov property):

$$p(s'|s,a) = Pr\left\{S_{t+1} = s'|S_t = s, A_t = a\right\}, \tag{4}$$

where $p : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \longrightarrow [0,1]$ is known as *state-transition probabilities*. Further, the policy followed by the agent is determined by the probability of taking an action in a specific state, which is defined as:

$$\pi(a|s) = Pr\left\{A_t = a|S_t = s\right\}, \tag{5}$$

where $\pi : \mathcal{S} \times \mathcal{A} \longrightarrow [0,1]$ is known as *policy* of the MDP. RL formulation typically involves two important functions: *state-value function* $v_\pi(s)$ and *state-action-value function* $q_\pi(s,a)$. The function $v_\pi(s)$ is the expected return when starting in a state s and following a policy π thereafter, whereas, $q_\pi(s,a)$ is the expected return when starting in a state s, taking an action a, and following a policy π thereafter:

$$v_\pi(s) := \mathbb{E}_\pi\left\{\sum_{k=0}^{\infty} \gamma^k \; R_{t+k+1}|S_t = s\right\}, \tag{6}$$

$$q_\pi(s,a) := \mathbb{E}_\pi\left\{\sum_{k=0}^{\infty} \gamma^k \; R_{t+k+1}|S_t = s, A_t = a\right\}, \tag{7}$$

where, $\gamma \in [0,1]$ is a parameter called *discount factor*, which decides how much weight is to be given to the future rewards. In RL formulations, the objective is to find an optimal policy $\pi^\star$ that maximizes (6) or (7), i.e.,

$$\pi_\star = \underset{\pi}{\operatorname{argmax}} \;\; v_\pi(s), \;\forall s \in \mathcal{S}, \tag{8}$$

or equivalently,

$$\pi_\star = \underset{\pi}{\operatorname{argmax}} \;\; q_\pi(s,a), \;\forall s \in \mathcal{S}, a \in \mathcal{A} \tag{9}$$

It is to be remarked that in most of the real-world problems, the model dynamics $p(s'|s,a)$ of the environment is not known or difficult to estimate; however, RL-based algorithms can find optimal policies without knowing p, by performing exploration and optimizing the policy iteratively. RL algorithms can be broadly classified as Value-based, Policy-Gradient, and Actor-Critic. We present these RL categories and their use cases in SWN below:

- **Value-based methods -** The value-based method estimates the expected return from each state for a given sequence of actions taken from the state thereafter. The value function (given in (6)) and action-value function (given in (7)) are estimated from observation data obtained through several trials of state-action pairs, learning and converging to the optimal state-action pairs given the availability of sufficient data samples. Q-learning is the most commonly used technique to learn the optimal action-value function. Agents update the action-value function as per the following update rule:

$$Q(S_t,A_t) \leftarrow Q(S_t,A_t) + \alpha\left[R_{t+1} + \max_{a\in\mathcal{A}} Q(S_{t+1},a) - Q(S_t,A_t)\right]. \tag{10}$$

  In (10), with a proper choice of the step-size parameter $\alpha \in (0,1]$, the learned state-action-value function Q converges to its optimal value q_*. In the context of SWN, such value-based methods have applications in water demand management, optimization of reservoir operations, operational management, and control of chemical and physical parameters. For instance, Q-learning has been applied to solve the stochastic optimization formulation for multi reservoir system in [120]. In addition, [120] computes optimal hydraulic water outflow from the reservoirs using Q-learning and it has been demonstrated that Q-learning outperforms other reservoir control formulations such as stochastic dynamic programming. Authors of [121] study Q-learning for the operational management of lake Como, Italy, which concludes that the control policies generated by Q-learning are more efficient as compared to that of stochastic dynamic programming.
- **Policy Gradient Methods -** The primary focus of value-based methods is to estimate cumulative rewards and devise a recommendation for the policy π to follow, based on those estimated values; however, in policy-gradient methods, the agent estimates the optimal policy directly by using tools from stochastic optimization. In general, in most practical applications, the number of states is large and the policy function is parameterized (π_θ) where θ are typically the weights of a neural network, with respect to which, the optimization takes place. In Policy Gradient Methods, actions are taken without consulting

the value-functions but using suitable optimizers (e.g., Gradient Descent) that ensure monotonic improvement of the policy, the most popular being Trust Region Policy Optimization and Proximal Policy Optimization. Although Policy Gradients are a popular class of RL algorithms for control formulations, the use cases of such methods are limited in SWN.

- **Actor-critic methods -** Actor-critic methods can be termed as a hybrid of value-based and policy-gradient methods. In such methods, the role of *critic* is to estimate the value function after each action selection and influence the next iteration of the policy gradient step, while the *actor* is the one applying the actions of the policy. Such methods have been found to be useful for the operational management, scheduling, and control in SWN. For instance, in [119], a multi-critic method (a variant of Actor-critic) for operational control of interconnected water storage tanks. Another use case of such methods is Hydropower production scheduling [126].

3) Deep Reinforcement Learning

DRL is an extension of RL and has been successful in addressing the challenges of various real-life applications (such as IoT, smart grid, and autonomous cars [141]), where the number of states and actions is very large. In particular, SWNs have various interconnected components, constituting a high-dimensional state and action space. As discussed in (Section V-A2), in addition, SWN have integer and non-convex constraints, and computing optimal solutions is expensive in terms of time and memory. Such high-dimensional state and action spaces are difficult to handle efficiently and sometimes intractable. For simplification, one may discretize the action space; however, a naive discretization leads to information loss. One of the approaches to avoid discretization of state or action space is to use function approximation which generalises the state and/or action spaces through *model-free* approaches. DRL, as shown in Fig. 10b, is a *model-free* method, which can be used to fit both the value function and the state-action Q function to perform the control.

DRL does not require either prior information of the targeted environment and has been successfully tested to various applications in real-time. Deep Q-Networks (DQN, a variant of DRL) is able to learn control policies without any prior information of the application environment. In Deep Q networks, the agent relies over a replay memory (a.k.a experience memory) matrix. Such reply memory stores the past experiences of agent, and it enables the agent to remember and reuse its experiences from past events. In [123], DRL is used to identify the optimal pump speed for given water demand in a WDN. In this method, an agent is proposed that relies over the dueling Deep Q-network concept (a variant of DQN) to control the pump operation. In [125], a control strategy is proposed as a real-time control strategy for the stormwater management, using DRL. Readers can refer to [142] for additional details of DRL and Deep Q-Networks.

4) Performance Metric

We have discuss the performance matrices of evaluating an ML model in section (Section IV-C1). On the other hand, in general, the control formulations and the various stochastic and non-convex constraints, makes it not possible for the algorithms to achieve exactly the optimal solution. Therefore, in general, to evaluate the accuracy of data-driven control models, sub-optimality metrics provids a satisfactory performance for model evaluation. By computing or estimating the sub-optimality of a solution obtained by a certain algorithm, we can measure the overall performance of a data-driven model in certain applications [130]. We can conclude that if the sub-optimality gap (i.e. deviation between obtained solution and optimal solution) is small, the proposed data-driven control approach is near-optimal and there is little room for further improvement. Let $f_o(\boldsymbol{z}^\star)$ denote the optimal value of a control formulation, and $f_o(\hat{\boldsymbol{z}})$ denote the estimated value of the control solution computed through a data-driven control algorithm. We can define the suboptimality Υ_o as:

$$\Upsilon_o = \frac{|f_o(\boldsymbol{z}^\star) - f_o(\hat{\boldsymbol{z}})|}{f_o(\boldsymbol{z}^\star)} \tag{11}$$

In addition, we can consider that the estimated solution is accurate if the sub-optimality $\Upsilon_o \leq \epsilon$, where ϵ is the error tolerance.

C. Challenges

1) Real-time event-triggered control

In most of the available literature, mainly two of the components of CPS are integrated, namely, communication and computation. Integration of real-time event-triggered control (e.g. autonomous pressure control, autonomous pump scheduling, autonomous valve control, etc.) and real-time anomaly (chemical hazard, toxicity, etc.) detection mechanisms is one of the main challenges of water CPS. Essentially, we are interested to design near real-time control formulations which are triggered by undesirable events. To achieve this objective, one needs to make a proper problem formulation with the various necessary constraints. However, in water CPS, the constraints which correlate the flow, pressure, pipe dimensions are in general non-convex. In addition, the fact that such formulations involve integer constraints make the problem even more challenging. In this paper, we discuss some of the techniques to address non-convexity in Section V-A2; however, in general, such techniques provide a near-optimal solution which is not possible to obtain typically in real-time. Further research efforts are required to improve and integrate computationally efficient and time-bounded techniques with real-time water CPS solving the necessary complex problems to achieve the next level of intelligent control.

2) Exploration and Choice of Appropriate Reward function

In RL formulations, optimal policies followed by an agent rely on a suitable reward trajectory and an appropriate exploration approach. In real-world formulations such as water CPS, devising an optimal exploration strategy and the appropriate reward function is a challenging task. Some of the approaches are *reward-shaping*, *curiosity*, and *experience replay* [141]. *Reward shaping* is the most common approach, which relies on providing additional reward points if the agent moves in the direction towards the optimal policy. In *curiosity*, the agent evaluates its own actions and predicts the consequences from the self-organized procedure. In *experience replay*, the agent relies on an experience memory and determines the future course of action. Although such approaches are useful to design a RL control formulation, further research is required to test the validity of such approaches in the context of water CPS.

3) Incorporating Safety

Water CPS are networked systems that are composed of diverse critical components. It is expected that proposed control formulations perform safe actions under pre-defined constraints. One of the approaches to ensure safe actions is to maintain the satisfaction of several hard safety constraints Alternatively, control formulations can also rely on constrained MDP and negative avoidance systems to integrate safe actions[143]. Readers can refer to [143] for comprehensive survey on safe RL.

Summary: In this section, we have covered the major control formulations, mainly MPC and DDC, for water CPS. DDC formulations are motivated by the fact that MPC is usually inefficient for large-scale systems which have various stochastic and non-convex constraints, and the fact that it is less capable to adapt to changing environments. We envision that water CPS should evolve in such a way that they could integrate autonomous control actions without human interventions and further improve the overall operational efficiency with minimum resource consumption, while being able to adapt to changing conditions in the environment where control is taking place. We have reviewed the challenges of control formulations for SWN, and have presented relatively new control approaches inspired by ML to introduce real-time control actions in SWN.

VI. Conclusion

Real-time end-to-end management of SWN is a major challenge. CPS are intelligent networked systems that can potentially manage SWN, preferably in an autonomous fashion, while keeping users in the loop. In SWN, CPS can integrate diverse physical components, through intelligent sensing and communication, and can apply event-triggered control in different types of scenarios, including crisis development. In this survey paper, we cover major components of CPS, mainly Internet of Things, Machine Learning, and Control formulations with the diverse applications of SWN. Along with presenting the challenges of SWN, we cover also the integration of the major components of CPS in a unified framework, and how the real-time computational challenges of control formulations can be addressed by using different state-of-the-art Machine Learning techniques such as DL, RL, and DRL. Given the various SWN challenges, and in order to develop fully-fledged water CPS, competent authorities, water public utilities and agencies are expected to adapt their future strategies with the best available technologies of data acquisition, data analytics, machine learning and autonomous control.

References

[1] R. Connor, *The United Nations world water development report 2015: water for a sustainable world*. UNESCO publishing, 2015, vol. 1.

[2] S. Kartakis, "Next generation cyber-physical water distribution systems," 2016.

[3] K.-D. Kim and P. Kumar, "An overview and some challenges in cyber-physical systems," *Journal of the Indian Institute of Science*, vol. 93, no. 3, pp. 341–352, 2013.

[4] J. Bhardwaj, K. K. Gupta, and R. Gupta, "Towards a cyber-physical era: soft computing framework based multi-sensor array for water quality monitoring," 2018.

[5] Z. Wang, H. Song, D. W. Watkins, K. G. Ong, P. Xue, Q. Yang, and X. Shi, "Cyber-physical systems for water sustainability: challenges and opportunities," *IEEE Communications Magazine*, vol. 53, no. 5, pp. 216–222, 2015.

[6] C. Sun, V. Puig, and G. Cembrano, "Real-time control of urban water cycle under cyber-physical systems framework," *Water*, vol. 12, no. 2, p. 406, 2020.

[7] M. Singh and S. Ahmed, "Iot based smart water management systems: A systematic review," *Materials Today: Proceedings*, 2020.

[8] D. Fooladivanda and J. A. Taylor, "Energy-optimal pump scheduling and water flow," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 3, pp. 1016–1026, 2017.

[9] A. K. Covington, R. Bates, and R. Durst, "Definition of ph scales, standard reference values, measurement of ph and related terminology," *Pure Appl. Chem*, vol. 57, no. 3, pp. 531–542, 1985.

[10] T. P. Lambrou, C. C. Anastasiou, C. G. Panayiotou, and M. M. Polycarpou, "A low-cost sensor network for real-time monitoring and contamination detection in drinking water distribution systems," *IEEE sensors journal*, vol. 14, no. 8, pp. 2765–2772, 2014.

[11] A. J. Whittle, L. Girod, A. Preis, M. Allen, H. B. Lim, M. Iqbal, S. Srirangarajan, C. Fu, K. J. Wong, and D. Goldsmith, "Waterwise@ sg: A testbed for continuous monitoring of the water distribution system in singapore," in *Water Distribution Systems Analysis 2010*, 2010, pp. 1362–1378.

[12] S. Soegijoko, "A brief review on existing cyber-physical systems for healthcare applications and their prospective national developments," in *2013 3rd International Conference on Instrumentation, Communications, Information Technology and Biomedical Engineering (ICICI-BME)*. IEEE, 2013, pp. 2–2.

[13] A. Ahmed, S. Zulfiqar, A. Ghandar, Y. Chen, M. Hanai, and G. Theodoropoulos, "Digital twin technology for

aquaponics: Towards optimizing food production with dynamic data driven application systems," in *Asian Simulation Conference*. Springer, 2019, pp. 3–14.
[14] D. Saetta, A. Padda, X. Li, C. Leyva, P. B. Mirchandani, D. Boscovic, and T. H. Boyer, "Water and wastewater building cps: Creation of cyber-physical wastewater collection system centered on urine diversion," *IEEE Access*, vol. 7, pp. 182 477–182 488, 2019.
[15] M. Suresh, U. Manohary, A. G. Ry, R. Stoleru, and M. K. M. Sy, "A cyber-physical system for continuous monitoring of water distribution systems," in *2014 IEEE 10th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, 2014, pp. 570–577.
[16] Y. Wei and S. Li, "Water supply networks as cyber-physical systems and controllability analysis," *IEEE/CAA Journal of Automatica Sinica*, vol. 2, no. 3, pp. 313–319, 2015.
[17] S. Imen and N.-B. Chang, "Developing a cyber-physical system for smart and sustainable drinking water infrastructure management," in *2016 IEEE 13th International Conference on Networking, Sensing, and Control (ICNSC)*. IEEE, 2016, pp. 1–6.
[18] E. R. Griffor, C. Greer, D. A. Wollman, and M. J. Burns, "Framework for cyber-physical systems: Volume 2, working group reports," 2017.
[19] S. Kumar, P. Tiwari, and M. Zymbler, "Internet of things is a revolutionary approach for future technology enhancement: a review," *Journal of Big Data*, vol. 6, no. 1, pp. 1–21, 2019.
[20] E. Alpaydin, *Introduction to machine learning*. MIT press, 2020.
[21] O. Vermesan, P. Friess, P. Guillemin, S. Gusmeroli, H. Sundmaeker, A. Bassi, I. S. Jubert, M. Mazura, M. Harrison, M. Eisenhauer *et al.*, "Internet of things strategic research roadmap," *Internet of things-global technological and societal trends*, vol. 1, no. 2011, pp. 9–52, 2011.
[22] H. Xu, W. Yu, D. Griffith, and N. Golmie, "A survey on industrial internet of things: A cyber-physical systems perspective," *IEEE Access*, vol. 6, pp. 78 238–78 259, 2018.
[23] R. Gonçalves, J. JM Soares, and R. MF Lima, "An iot-based framework for smart water supply systems management," *Future Internet*, vol. 12, no. 7, p. 114, 2020.
[24] R. A. Kjellby, L. R. Cenkeramaddi, A. Frøytlog, B. B. Lozano, J. Soumya, and M. Bhange, "Long-range & self-powered iot devices for agriculture & aquaponics based on multi-hop topology," in *2019 IEEE 5th World Forum on Internet of Things (WF-IoT)*. IEEE, 2019, pp. 545–549.
[25] J. Pitakphongmetha, N. Boonnam, S. Wongkoon, T. Horanont, D. Somkiadcharoen, and J. Prapakornpilai, "Internet of things for planting in smart farm hydroponics style," in *2016 International Computer Science and Engineering Conference (ICSEC)*, 2016, pp. 1–5.
[26] I. Dumitrache, "Intelligent cyber-energy-systems," in *invited paper on ICTSCC-18th International Conference on System Theory, Control and Computing*, 2014.
[27] J. Hall, A. D. Zaffiro, R. B. Marx, P. C. Kefauver, E. R. Krishnan, R. C. Haught, and J. G. Herrmann, "On–line water quality parameters as indicators of distribution system contamination," *Journal-American Water Works Association*, vol. 99, no. 1, pp. 66–77, 2007.
[28] Y. Wei, W. Li, D. An, D. Li, Y. Jiao, and Q. Wei, "Equipment and intelligent control system in aquaponics: A review," *IEEE Access*, vol. 7, pp. 169 306–169 326, 2019.
[29] P. Quevauviller, O. Thomas, and A. v. Derbeken, *Wastewater quality monitoring and treatment*. Wiley Online Library, 2006.
[30] J. DeZuane, *Handbook of drinking water quality*. John Wiley & Sons, 1997.
[31] E. Karami, F. M. Bui, and H. H. Nguyen, "Multisensor data fusion for water quality monitoring using wireless sensor networks," in *2012 Fourth International Conference on Communications and Electronics (ICCE)*, 2012, pp. 80–85.
[32] M. Weyrich and C. Ebert, "Reference architectures for the internet of things," *IEEE Software*, vol. 33, no. 1, pp. 112–116, 2015.
[33] P. Sethi and S. R. Sarangi, "Internet of things: architectures, protocols, and applications," *Journal of Electrical and Computer Engineering*, vol. 2017.
[34] O. Said and M. Masud, "Towards internet of things: Survey and future vision," *International Journal of Computer Networks*, vol. 5, no. 1, pp. 1–17, 2013.
[35] R. Shahzadi, A. Niaz, M. Ali, M. Naeem, J. J. Rodrigues, F. Qamar, and S. M. Anwar, "Three tier fog networks: Enabling iot/5g for latency sensitive applications," *China Communications*, vol. 16, no. 3, pp. 1–11, 2019.
[36] Y. Y. F. Panduman, S. Sukaridhoto, and A. Tjahjono, "A survey of iot platform comparison for building cyber-physical system architecture," in *2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*. IEEE, 2019, pp. 238–243.
[37] S. García, D. F. Larios, J. Barbancho, E. Personal, J. M. Mora-Merchán, and C. León, "Heterogeneous lora-based wireless multimedia sensor network multiprocessor platform for environmental monitoring," *Sensors*, vol. 19, no. 16, p. 3446, 2019.
[38] F. Bonomi, R. Milito, P. Natarajan, and J. Zhu, "Fog computing: A platform for internet of things and analytics," in *Big data and internet of things: A roadmap for smart environments*. Springer, 2014, pp. 169–186.
[39] W. Dong and Q. Yang, "Data-driven solution for optimal pumping units scheduling of smart water conservancy,"

IEEE Internet of Things Journal, vol. 7, no. 3, pp. 1919–1926, 2020.

[40] K. Sohraby, D. Minoli, and T. Znati, *Wireless sensor networks: technology, protocols, and applications*. John wiley & sons, 2007.

[41] B. O'Flynn, R. Martinez-Catala, S. Harte, C. O'Mathuna, J. Cleary, C. Slater, F. Regan, D. Diamond, and H. Murphy, "Smartcoast: A wireless sensor network for water quality monitoring," in *32nd IEEE Conference on Local Computer Networks (LCN 2007)*, 2007, pp. 815–816.

[42] G. Amarasinghe, M. D. de Assunçao, A. Harwood, and S. Karunasekera, "A data stream processing optimisation framework for edge computing applications," in *2018 IEEE 21st International Symposium on Real-Time Distributed Computing (ISORC)*. IEEE, 2018, pp. 91–98.

[43] J.-S. Lee, Y.-W. Su, and C.-C. Shen, "A comparative study of wireless protocols: Bluetooth, uwb, zigbee, and wi-fi," in *IECON 2007-33rd Annual Conference of the IEEE Industrial Electronics Society*. Ieee, 2007, pp. 46–51.

[44] J. Raich, "Review of sensors to monitor water quality," 2013.

[45] S. M. Radke and E. C. Alocilja, "A high density microelectrode array biosensor for detection of e. coli o157: H7," *Biosensors and Bioelectronics*, vol. 20, no. 8, pp. 1662–1667, 2005.

[46] A. S. Rao, S. Marshall, J. Gubbi, M. Palaniswami, R. Sinnott, and V. Pettigrovet, "Design of low-cost autonomous water quality monitoring system," in *2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2013, pp. 14–19.

[47] P. Khatri, K. K. Gupta, and R. K. Gupta, "Raspberry pi-based smart sensing platform for drinking-water quality monitoring system: a python framework approach," *Drinking Water Engineering and Science*, vol. 12, no. 1, pp. 31–37, 2019.

[48] S. Klikovits, A. Coet, and D. Buchs, "Ml4crest: Machine learning for cps models," in *Proceedings of MODELS 2018 Workshops: ModComp, MRT, OCL, FlexMDE, EXE, COMMitMDE, MDETools, GEMOC, MORSE, MDE4IoT, MDEbug, MoDeVVa, ME, MULTI, HuFaMo, AMMoRe, PAINS co-located with ACM/IEEE 21st International Conference on Model Driven Engineering Languages and Systems (MODELS 2018)*, 2018, pp. 515–520.

[49] O. A. Postolache, P. S. Girao, J. D. Pereira, and H. M. G. Ramos, "Multibeam optical system and neural processing for turbidity measurement," *IEEE Sensors Journal*, vol. 7, no. 5, pp. 677–684, 2007.

[50] A. Hauser and F. Roedler, "Interoperability: the key for smart water management," *Water Science and Technology: Water Supply*, vol. 15, no. 1, pp. 207–214, 2015.

[51] B. Blanco-Filgueira, D. Garcia-Lesta, M. Fernández-Sanjurjo, V. M. Brea, and P. López, "Deep learning-based multiple object visual tracking on embedded system for iot and mobile edge computing applications," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 5423–5431, 2019.

[52] S. Mohurle and M. Devare, "A study of knn classifier to predict water pollution index," *Computing in Engineering and Technology*, pp. 457–466, 2020.

[53] A. Rojik, Endroyono, and A. N. Irfansyah, "Water pipe leak detection using the k-nearest neighbor method," in *2019 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, 2019, pp. 393–398.

[54] D. Adidrana and N. Surantha, "Hydroponic nutrient control system based on internet of things and k-nearest neighbors," in *2019 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, 2019, pp. 166–171.

[55] Y. Bai, P. Wang, C. Li, J. Xie, and Y. Wang, "Dynamic forecast of daily urban water consumption using a variable-structure support vector regression model," *Journal of Water Resources Planning and Management*, vol. 141, no. 3, p. 04014058, 2015.

[56] A. H. Haghiabi, A. H. Nasrolahi, and A. Parsaie, "Water quality prediction using machine learning methods," *Water Quality Research Journal*, vol. 53, no. 1, pp. 3–13, 2018.

[57] Y. Haruo, H. Yamamoto, M. Arakawa, and I. Naka, "Development and evaluation of environmental / growth observation sensor network system for aquaponics," in *2020 IEEE International Conference on Consumer Electronics (ICCE)*, 2020, pp. 1–6.

[58] D. Li, J. Sun, H. Yang, and X. Wang, "An enhanced naive bayes model for dissolved oxygen forecasting in shellfish aquaculture," *IEEE Access*, vol. 8, pp. 217 917–217 927, 2020.

[59] M. A. K. Fasaee, E. Berglund, K. J. Pieper, E. Ling, B. Benham, and M. Edwards, "Developing a framework for classifying water lead levels at private drinking water systems: A bayesian belief network approach," *Water Research*, vol. 189, p. 116641, 2021.

[60] M. Saravanan, A. Das, and V. Iyer, "Smart water grid management using lpwan iot technology," in *2017 Global Internet of Things Summit (GIoTS)*, 2017, pp. 1–6.

[61] A. Robles-Velasco, P. Cortés, J. Muñuzuri, and L. Onieva, "Prediction of pipe failures in water supply networks using logistic regression and support vector classification," *Reliability Engineering & System Safety*, vol. 196, p. 106754, 2020.

[62] H. Lu and X. Ma, "Hybrid decision tree-based machine learning models for short-term water quality prediction," *Chemosphere*, vol. 249, p. 126169, 2020.

[63] L. Aymon, J. Decaix, F. Carrino, P. Mudry, E. Mugellini, O. Abou Khaled, and R. Baltensperger, "Leak detection using random forest and pressure

simulation," in *2019 6th Swiss Conference on Data Science (SDS)*, 2019, pp. 109–110.
[64] J. A. B. Somontina, F. Carlo C. Garcia, and E. Q. B. Macabebe, "Water consumption monitoring with fixture recognition using random forest," in *TENCON 2018 - 2018 IEEE Region 10 Conference*, 2018, pp. 0663–0667.
[65] H. Mohammed, I. A. Hameed, and R. Seidu, "Random forest tree for predicting fecal indicator organisms in drinking water supply," in *2017 International Conference on Behavioral, Economic, Socio-cultural Computing (BESC)*, 2017, pp. 1–6.
[66] C. V. Geelen, D. R. Yntema, J. Molenaar, and K. J. Keesman, "Burst detection by water demand nowcasting based on exogenous sensors," *Water Resources Management*, pp. 1–14, 2021.
[67] M. Xenochristou, C. Hutton, J. Hofman, and Z. Kapelan, "Water demand forecasting accuracy and influencing factors at different spatial scales using a gradient boosting machine," *Water Resources Research*, vol. 56, no. 8, p. e2019WR026304, 2020.
[68] Y. Su and Y. Zhao, "Prediction of downstream bod based on light gradient boosting machine method," in *2020 International Conference on Communications, Information System and Computer Engineering (CISCE)*, 2020, pp. 127–130.
[69] A. Y. Felix and T. Sasipraba, "Flood detection using gradient boost machine learning approach," in *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, 2019, pp. 779–783.
[70] S. Palani, S.-Y. Liong, and P. Tkalich, "An ann application for water quality forecasting," *Marine pollution bulletin*, vol. 56, no. 9, pp. 1586–1597, 2008.
[71] A. Astel, S. Tsakovski, V. Simeonov, E. Reisenhofer, S. Piselli, and P. Barbieri, "Multivariate classification and modeling in surface water pollution estimation," *Analytical and Bioanalytical Chemistry*, vol. 390, no. 5, pp. 1283–1292, 2008.
[72] Q. Ren, L. Zhang, Y. Wei, and D. Li, "A method for predicting dissolved oxygen in aquaculture water in an aquaponics system," *Computers and electronics in agriculture*, vol. 151, pp. 384–391, 2018.
[73] A. Jain, A. K. Varshney, and U. C. Joshi, "Short-term water demand forecast modelling at iit kanpur using artificial neural networks," *Water resources management*, vol. 15, no. 5, pp. 299–321, 2001.
[74] H. Zou, Z. Zou, and X. Wang, "An enhanced k-means algorithm for water quality analysis of the haihe river in china," *International journal of environmental research and public health*, vol. 12, no. 11, pp. 14 400–14 413, 2015.
[75] C. W. Chow, J. Liu, J. Li, N. Swain, K. Reid, and C. P. Saint, "Development of smart data analytics tools to support wastewater treatment plant operation," *Chemometrics and Intelligent Laboratory Systems*, vol. 177, pp. 140–150, 2018.
[76] M. Li, S. Hu, J. Xia, J. Wang, X. Song, and H. Shen, "Dissolved oxygen model predictive control for activated sludge process model based on the fuzzy c-means cluster algorithm," *International Journal of Control, Automation and Systems*, vol. 18, no. 9, pp. 2435–2444, 2020.
[77] E. K. Wang, Y. Ye, X. Xu, S.-M. Yiu, L. C. K. Hui, and K.-P. Chow, "Security issues and challenges for cyber physical system," in *2010 IEEE/ACM Int'l Conference on Green Computing and Communications & Int'l Conference on Cyber, Physical and Social Computing*. IEEE, 2010, pp. 733–738.
[78] M. V. Storey, B. Van der Gaag, and B. P. Burns, "Advances in on-line drinking water quality monitoring and early warning systems," *Water research*, vol. 45, no. 2, pp. 741–747, 2011.
[79] O. Maimon and L. Rokach, "Data mining and knowledge discovery handbook," 2005.
[80] J. Bhardwaj, K. K. Gupta, and R. Gupta, "A review of emerging trends on water quality measurement sensors," in *2015 International Conference on Technologies for Sustainable Development (ICTSD)*. IEEE, 2015, pp. 1–6.
[81] V. Chamola, V. Hassija, S. Gupta, A. Goyal, M. Guizani, and B. Sikdar, "Disaster and pandemic management using machine learning: A survey," *IEEE Internet of Things Journal*, pp. 1–1, 2020.
[82] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 161–168.
[83] X. Zhu, S. Zhang, Z. Jin, Z. Zhang, and Z. Xu, "Missing value estimation for mixed-attribute data sets," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 1, pp. 110–121, 2011.
[84] X. Xie, W. T. Liu, and B. Tang, "Spacebased estimation of moisture transport in marine atmosphere using support vector regression," *Remote Sensing of Environment*, vol. 112, no. 4, pp. 1846–1855, 2008.
[85] C. Rudin, "Naive bayes, mit 15.097 course notes," 2012.
[86] M. Norouzi, M. D. Collins, M. Johnson, D. J. Fleet, and P. Kohli, "Efficient non-greedy optimization of decision trees," *arXiv preprint arXiv:1511.04056*, 2015.
[87] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, 2016.
[88] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *The journal of machine learning research*, vol. 15, no. 1, pp. 3133–3181, 2014.
[89] I. S. Msiza, F. V. Nelwamondo, and T. Marwala, "Artificial neural networks and support vector machines for water demand time series forecasting," in *2007 IEEE International Conference on Systems, Man and Cybernetics*. IEEE, 2007, pp. 638–643.

[90] Y. Chen, L. Song, Y. Liu, L. Yang, and D. Li, "A review of the artificial neural network models for water quality prediction," *Applied Sciences*, vol. 10, no. 17, p. 5776, 2020.

[91] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Transactions on neural networks*, vol. 16, no. 3, pp. 645–678, 2005.

[92] K. Chang, J. L. Gao, W. Y. Wu, and Y. X. Yuan, "Water quality comprehensive evaluation method for large water distribution network based on clustering analysis," *Journal of Hydroinformatics*, vol. 13, no. 3, pp. 390–400, 2011.

[93] Y. Kageyama, K. Wakatabe, M. Ishikawa, B. Kobori, and D. Nagamoto, "Application of fuzzy regression analysis and fuzzy c-means technique using uav data to understand water quality in the miharu dam reservoir, japan," *IEEJ Transactions on Electrical and Electronic Engineering*, vol. 13, no. 12, pp. 1831–1832, 2018.

[94] A. Talwalkar, S. Kumar, and H. Rowley, "Large-scale manifold learning," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[95] S. Samanipour, S. Kaserzon, S. Vijayasarathy, H. Jiang, P. Choi, M. J. Reid, J. F. Mueller, and K. V. Thomas, "Machine learning combined with non-targeted lc-hrms analysis for a risk warning system of chemical hazards in drinking water: A proof of concept," *Talanta*, vol. 195, pp. 426–432, 2019.

[96] A. A. Nasser, M. Z. Rashad, and S. E. Hussein, "A two-layer water demand prediction system in urban areas based on micro-services and lstm neural networks," *IEEE Access*, vol. 8, pp. 147 647–147 661, 2020.

[97] M. Farley and S. Trow, *Losses in water distribution networks*. IWA publishing, 2003.

[98] V. Gunes, S. Peter, T. Givargis, and F. Vahid, "A survey on concepts, applications, and challenges in cyber-physical systems." *KSII Transactions on Internet & Information Systems*, vol. 8, no. 12, 2014.

[99] W. Zhao, T. H. Beach, and Y. Rezgui, "Optimization of potable water distribution and wastewater collection networks: A systematic review and future research directions," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 46, no. 5, pp. 659–681, 2015.

[100] H. Mala-Jetmarova, N. Sultanova, and D. Savic, "Lost in optimisation of water distribution systems? a literature review of system operation," *Environmental modelling & software*, vol. 93, pp. 209–254, 2017.

[101] M. K. Singh and V. Kekatos, "Optimal scheduling of water distribution systems," *IEEE Transactions on Control of Network Systems*, 2019.

[102] P. F. Boulos, Z. Wu, C. H. Orr, M. Moore, P. Hsiung, and D. Thomas, "Optimal pump operation of water distribution systems using genetic algorithms," in *Distribution system symposium*. Citeseer, 2001.

[103] I. Zimoch and E. Łobos, "The optimization of chlorine dose in water treatment process in order to reduce the formation of disinfection by-products," *Desalination and Water Treatment*, vol. 52, no. 19-21, pp. 3719–3724, 2014.

[104] A. Gleixner, H. Held, W. Huang, and S. Vigerske, "Towards globally optimal operation of water supply networks," 2012.

[105] A. Fanni, S. Liberatore, G. M. Sechi, M. Soro, and P. Zuddas, "Optimization of water distribution systems by a tabu search metaheuristic," in *Computing Tools for Modeling, Optimization and Simulation*. Springer, 2000, pp. 279–298.

[106] K. Oikonomou and M. Parvania, "Optimal coordination of water distribution energy flexibility with power systems operation," *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 1101–1110, 2018.

[107] S. Wang, A. Taha, N. Gatsis, and M. Giacomoni, "Receding horizon control for drinking water networks: The case for geometric programming," *IEEE Transactions on Control of Network Systems*, 2020.

[108] T. M. Walski, D. V. Chase, D. A. Savic, W. Grayman, S. Beckwith, and E. Koelle, "Advanced water distribution modeling and management," 2003.

[109] L. A. Rossman, "An overview of epanet version 3.0," *Water Distribution Systems Analysis 2010*, pp. 14–18, 2010.

[110] R. Barzegar, M. T. Aalami, and J. Adamowski, "Short-term water quality variable prediction using a hybrid cnn–lstm deep learning model," *Stochastic Environmental Research and Risk Assessment*, pp. 1–19, 2020.

[111] J. Kang, Y.-J. Park, J. Lee, S.-H. Wang, and D.-S. Eom, "Novel leakage detection by ensemble cnn-svm and graph-based localization in water distribution systems," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 5, pp. 4279–4289, 2017.

[112] 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 research*, vol. 166, p. 115058, 2019.

[113] X. Yang, S. Zhang, J. Liu, Q. Gao, S. Dong, and C. Zhou, "Deep learning for smart fish farming: applications, opportunities and challenges," *Reviews in Aquaculture*, vol. 13, no. 1, pp. 66–90, 2021.

[114] Z. Li, F. Peng, B. Niu, G. Li, J. Wu, and Z. Miao, "Water quality prediction model combining sparse auto-encoder and lstm network," *IFAC-PapersOnLine*, vol. 51, no. 17, pp. 831–836, 2018.

[115] C. Kühnert, N. M. Gonuguntla, H. Krieg, D. Nowak, and J. A. Thomas, "Application of lstm networks for water demand prediction in optimal pump control," *Water*, vol. 13, no. 5, p. 644, 2021.

[116] Q. Ren, X. Wang, W. Li, Y. Wei, and D. An, "Research of dissolved oxygen prediction in recirculating aquaculture systems based on deep belief network," *Aquacultural Engineering*, vol. 90, p. 102085, 2020.

[117] S. Duan, P. Ullrich, and L. Shu, "Using convolutional neural networks for streamflow projection in california," *Front. Water 2: 28. doi: 10.3389/frwa*, 2020.

[118] J. Inoue, Y. Yamagata, Y. Chen, C. M. Poskitt, and J. Sun, "Anomaly detection for a water treatment system using unsupervised machine learning," in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2017, pp. 1058–1065.

[119] J. Martinez-Piazuelo, D. E. Ochoa, N. Quijano, and L. F. Giraldo, "A multi-critic reinforcement learning method: An application to multi-tank water systems," *IEEE Access*, vol. 8, pp. 173 227–173 238, 2020.

[120] J.-H. Lee and J. W. Labadie, "Stochastic optimization of multireservoir systems via reinforcement learning," *Water resources research*, vol. 43, no. 11, 2007.

[121] A. Castelletti, G. Corani, and E. Weber, "Reinforcement learning in the operational management of a water system."

[122] L. Sun, Y. Yang, J. Hu, D. Porter, T. Marek, and C. Hillyer, "Reinforcement learning control for water-efficient agricultural irrigation," in *2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC)*. IEEE, 2017, pp. 1334–1341.

[123] G. Hajgató, G. Paál, and B. Gyires-Tóth, "Deep reinforcement learning for real-time optimization of pumps in water distribution systems," *Journal of Water Resources Planning and Management*, vol. 146, no. 11, p. 04020079, 2020.

[124] F. Hernandez-del Olmo, E. Gaudioso, and A. Nevado, "Autonomous adaptive and active tuning up of the dissolved oxygen setpoint in a wastewater treatment plant using reinforcement learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 5, pp. 768–774, 2012.

[125] A. Mullapudi, M. J. Lewis, C. L. Gruden, and B. Kerkez, "Deep reinforcement learning for the real time control of stormwater systems," *Advances in Water Resources*, vol. 140, p. 103600, 2020.

[126] S. Riemer-Sørensen and G. H. Rosenlund, "Deep reinforcement learning for long term hydropower production scheduling," in *2020 International Conference on Smart Energy Systems and Technologies (SEST)*. IEEE, 2020, pp. 1–6.

[127] S. Kim, J. Koo, H. Kim, and Y. Choi, "Optimization of pumping schedule based on forecasting the hourly water demand in seoul," *Water Science and Technology: Water Supply*, vol. 7, no. 5-6, pp. 85–93, 2007.

[128] F. K. Odan, L. F. Ribeiro Reis, and Z. Kapelan, "Real-time multiobjective optimization of operation of water supply systems," *Journal of Water Resources Planning and Management*, vol. 141, no. 9, p. 04015011, 2015.

[129] A. Morsi, B. Geißler, and A. Martin, "Mixed integer optimization of water supply networks," in *Mathematical optimization of water networks*. Springer, 2012, pp. 35–54.

[130] D. Bertsimas and B. Stellato, "Online mixed-integer optimization in milliseconds," *arXiv preprint arXiv:1907.02206*, 2019.

[131] F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Automated configuration of mixed integer programming solvers," in *International Conference on Integration of Artificial Intelligence (AI) and Operations Research (OR) Techniques in Constraint Programming*. Springer, 2010, pp. 186–202.

[132] B. Geißler, A. Martin, A. Morsi, and L. Schewe, "Using piecewise linear functions for solving minlp s," in *Mixed integer nonlinear programming*. Springer, 2012, pp. 287–314.

[133] D. Fooladivanda and J. A. Taylor, "Optimal pump scheduling and water flow in water distribution networks," in *2015 54th IEEE Conference on Decision and Control (CDC)*. IEEE, 2015, pp. 5265–5271.

[134] S. L. Brunton and J. N. Kutz, *Data-driven science and engineering: Machine learning, dynamical systems, and control*. Cambridge University Press, 2019.

[135] Y. Xu, J. Zhang, Z. Long, H. Tang, and X. Zhang, "Hourly urban water demand forecasting using the continuous deep belief echo state network," *Water*, vol. 11, no. 2, p. 351, 2019.

[136] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1238–1274, 2013.

[137] J. G. Bene and I. Selek, "Water network operational optimization: Utilizing symmetries in combinatorial problems by dynamic programming," *Periodica Polytechnica Civil Engineering*, vol. 56, no. 1, pp. 51–61, 2012.

[138] R. Bellman, "Dynamic programming: Princeton univ. press," *NJ*, vol. 95, 1957.

[139] J. Russell Stuart and P. Norvig, *Artificial intelligence: a modern approach*. Prentice Hall, 2009.

[140] A. G. Barto, "Reinforcement learning and dynamic programming," in *Analysis, Design and Evaluation of Man–Machine Systems 1995*. Elsevier, 1995, pp. 407–412.

[141] B. R. Kiran, I. Sobh, V. Talpaert, P. Mannion, A. A. A. Sallab, S. Yogamani, and P. Pérez, "Deep reinforcement learning for autonomous driving: A survey," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–18, 2021.

[142] L. Graesser and W. L. Keng, *Foundations of deep reinforcement learning: theory and practice in Python*. Addison-Wesley Professional, 2019.

[143] J. Garcıa and F. Fernández, "A comprehensive survey on safe reinforcement learning," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 1437–1480, 2015.

Jyotirmoy Bhardwaj is a Ph.D. research fellow (Major: *Artificial Intelligence*) at WISENET Center, University of Agder, Norway. In addition, Jyotirmoy is working with the Environmental Chemistry group, Norwegian Institute for Water Research, Norway. His research interest revolves around sensor fusion, data analytics, machine learning, and mathematical optimization. Presently, he is investigating data-driven machine learning approaches to predict the various events in smart water networks. Furthermore, he is working on various Electronics/ Communication methods in order to simplify the data aggregation process of heterogeneous water quality parameters. Jyotirmoy holds B.Engg (Bachelor of Engineering) and M.Engg (Master of Engineering - Birla Institute of Technology and Science, Pilani, India) in the field of Telecommunications Engineering.

Joshin P. Krishnan received the B.Tech. Degree in Electronics and Communication Engineering from College of Engineering, Thiruvananthapuram, (University of Kerala, India), in 2010, the M.E. degree in Telecommunications from Indian Institute of Science (IISc), Bengaluru, in 2014, and the Ph.D. degree in Electrical and Computer Engineering from the Instituto Superior Técnico, University of Lisbon, Portugal, in 2019. From 2015 to 2019, he was with the Instituto de Telecomunicações, Lisbon, as a Marie Curie Early-Stage Researcher of the Machine Sensing Training Network (MacSeNet). From 2019 to 2021, he worked as a Postdoctoral Researcher at WISENET Research Center, UiA, Norway. He is currently working as a Postdoctoral Researcher at SIMULA Research center, Norway. His research interests include image inverse problems, optimization, multivariate time-series analysis, graph signal processing, and machine learning.

Diego F. Larios Marin was born in Seville, Spain. He received the degree in "ingeniero técnico industrial en electrónica industrial" and the M.S. degree in industrial electronics and automatic control engineering from the University of Seville, Seville, in 2006 and 2009, respectively, where he is currently working as an Associate Professor in "Electronic Technology and Industrial Automation". He is currently a Member of the research group "Tecnología Electrónica e Informática Industrial" at the University of Seville. His main fields of interest are in low-power wireless sensor networks, ubiquitous and pervasive computing, automatic control, and industrial automation. In addition, he is a member of the Society for Modeling and Simulation International.

Baltasar Beferull-Lozano received his MSc in Physics from Universidad de Valencia, Spain, in 1995 (First in Class Honors) and the MSc and PhD degrees in Electrical Engineering from University of Southern California (USC), Los Angeles, in 1999 and 2002, respectively.

In October 2002, he joined the AudioVisual Communications Laboratory, Department of Communication Systems, at EPFL, as a Research Associate, where he spent around three years. In December 2005, he joined the School of Engineering at University of Valencia as Associate Professor. Since August 2014, he is a Professor at the Department of Information and Communication Technology and (by courtesy) the Department of Engineering, University of Agder, Norway, where he leads the Center Intelligent Signal Processing and Wireless Networks (WISENET). He serves as a Senior Area Editor for IEEE Transactions on Signal Processing since 2016 and has also served as a member of the Technical Program Committees for several ACM & IEEE International Conferences. His research interests are in the general areas of distributed in-network signal processing and collective intelligence, data science and machine learning, networked cyber-physical systems, optimization and artificial intelligence for next generation wireless networks.

At USC, Dr. Beferull-Lozano received several awards including the Best PhD Thesis paper Award in 2002 and the Outstanding Academic Achievement Award in 1999. He received the Best Paper Award at the IEEE DCOSS 2012 international conference. He has also received a TOPPFORSK Grant Award from the Research Council of Norway, 2015. He is a Senior Member of the IEEE and a member of the Norwegian Academy of Science and Technology.

Linga Reddy Cenkeramaddi (Senior Member, IEEE) received master's degree in electrical engineering from the Indian Institute of Technology, New Delhi, India, in 2004, and Ph.D. degree in electrical engineering from the Norwegian University of Science and Technology, Trondheim, Norway, in 2011. He worked for Texas Instruments in mixed-signal circuit design before joining the Ph.D. program at NTNU. After finishing his Ph.D., he worked in radiation imaging for an atmosphere space interaction monitor (ASIM mission to International Space Station) at the University of Bergen, Norway from 2010 to 2012. At present, he is the group leader of the autonomous and cyber-physical systems (ACPS) research group and working as an Associate professor at the University of Agder, Campus Grimstad, Norway. His main scientific interests are in Cyber-Physical Systems, Autonomous Systems, and Wireless Embedded Systems. He has co-authored over 80 research publications that have been published in prestigious international journals and standard conferences. He is a senior member of IEEE. He is also a member of the Editorial Boards of various international journals, as well as the technical program committees of several IEEE conferences.

Christopher Harman received his master's degree in Marine Science from the University of Plymouth, UK, and his Ph.D. in Environmental and Analytical Chemistry from the University of Oslo, Norway. Since 2017, he has been a Research Director at the Norwegian Institute for Water Research (NIVA), Oslo, Norway. He currently manages over 100 water professionals organized in research groups covering contaminants, ecotoxicology, oceanography, infrastructure, and environmental technologies. He has managed or supported a wide range of projects addressing the measurement and toxicological effects of chemical contaminants in aquatic systems, with a special focus on industry and technology. Christopher Harman develops, customizes, and applies novel sampling and sensing methodologies to measure a variety of chemicals such as oil constituents, classical POPs, emerging micro-pollutants, and mercury. He has extensive international project experience and has published over 70 scientific papers/ technical reports.