D4.8 – a. AI Inference tool for water efficiency problem detection and remedial action (intermediate)

WP4: Digital Twin with Smart Analytics and Cognitive Services for Water Efficiency

November 2023

Authors: Dimitra Pournara (ICCS), Nikos Papageorgiou (ICCS), Dimitris Apostolou (ICCS), Gregoris Mentzas (ICCS), Aziz Mousas (MAG), Robert Sanfeliu Prat (EUT)

The AquaSPICE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 958396.

Document Information

GRANT AGREEMENT NUMBER	958396	**ACRONYM**	AquaSPICE
FULL TITLE	Advancing Sustainability of Process Industries through Digital and Circular Water Use Innovations		
START DATE	1st December 2020	**DURATION**	51 months
PROJECT URL	www.AquaSPICE.eu		
DELIVERABLE	D4.8 – AI Inference tool for water efficiency problem detection and remedial action (intermediate)		
WORK PACKAGE	WP4 – Digital Twin with Smart Analytics and Cognitive Services for Water Efficiency		
DATE OF DELIVERY	**CONTRACTUAL** 11/2023	**ACTUAL**	11/2023
NATURE	Report	**DISSEMINATION LEVEL**	Public
LEAD BENEFICIARY	ICCS		
RESPONSIBLE AUTHOR	Dimitra Pournara (ICCS)		
CONTRIBUTIONS FROM	Nikos Papageorgiou (ICCS), Dimitris Apostolou (ICCS), Gregoris Mentzas (ICCS), Aziz Mousas (MAG), Robert Sanfeliu Prat (EUT), Athanasios Angelis-Dimakis (UOH – reviewer)		
ABSTRACT	This scope of the deliverable is to demonstrate the AquaSPICE Analytics Platform and the use of AI, ML and analytics techniques to produce descriptive analytics, predictive models for forecasting and anomaly detection on industrial water data. It includes a literature review of AI and ML in the water treatment sector, a presentation of the developed platform and its microservices, and an overview of selected methods, algorithms and models that have been developed to support the use case requirements. Given that it is an intermediate version of the deliverable, it does not report on the results obtained using the Analytics Platform in the use cases, which is work under way and will be reported in the final version of the same deliverable.		

Document History

VERSION	ISSUE DATE	DESCRIPTION	CONTRIBUTOR
0.1	01/05/2023	Table of Contents shared with partners	ICCS
0.2	16/11/2023	Final draft for review	ICCS
1.0	30/11/2023	Final for submission	ICCS

Disclaimer

Any dissemination of results reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains.

Copyright message

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS/ACRONYMS

Parameters

BOD	Biochemical Oxygen Demand
BOD_5	Biochemical Oxygen Demand test run for 5 days)
COD	Chemical Oxygen Demand
DO	Dissolved Oxygen
EC	Electrical Conductivity
F/M	Food to Microorganism
MLSS	Mixed Liquor Suspended Solids
MLVSS	Mixed Liquor Volatile Suspended Solids
N	Nitrogen
SS	Suspended Solids
T-N	Total Nitrogen
T-P	Total Phosphorus
TDS	Total Dissolved Solids
TS	Total Solids
TSS	Total Suspended Solid

Machine Learning & Statistics

AANN	Autoassociative Neural Networks
AMGA	Adaptive Merging and Growing Algorithm
ANN	Artificial Neutal Network
ANOVA	Analysis of Variance
CFL	Committee Fuzzy Logic
FL	Fuzzy Logic
FS-RBFN	Flexible Structure Radial Basis Function NN
GA	Genetic Algorithm
LFL	Larsen FL
LSTM	Long-Short Term Memory
MAPE	Mean Absolute Percentage Error
MFL	Mamdani FL
MLP	Multi Layer Perceptron
MLR	Multivariate Linear Regression
MRAN	Minimal Resource-Allocating Network
MSE	Mean Squared Error
NAR	Nonlinear Autoregressive
NSE	Nash-Sutcliff Efficiency
r	Coefficient of correlation
R2	Coefficient of determination
RBF	Radial basis function
RFNN	Recurrent Fuzzy Neural Network
RMSE	Root Mean Square Error

RVM	Relevant Vector Machine
SCFL	Supervised Committee Fuzzy Logic
SenV-RBF	Sensitivity-based RBF
SVM	Support Vector Machine
TSFL	Takagi-Sugeno FL
VBPCA	Variational Bayesian Principal Component Analysis
XGBoost	Extreme Gradient Boosting

Stages

BIO	Bioreactor
BT_C	Bioreactor Pit C
BT_N	Bioreactor Pit N
Clari	Clarifier
D	Discharge Pit
EQ	Equalizer
OxT	Oxydation Tank

Miscellaneous

MBR	Membrane Bioreactor
TMP	Transmembrane Pressure
WWTP	Wastewater Treatment Plant
WRC	Water Reclamation Plant
BSM 1	Benchmark Simulation Model 1
Type-I error	false alarm rate
Type-II error	faulty detection rate
KG	Knowledge Graph
API	Application Programming Interface

Publishers

IASE	Institute of Advanced Science Extension
IEEE	Institute of Electrical and Electronics Engineers
IOP	Institute of Physics
IWA	International Water Association
MDPI	Multidisciplinary Digital Publishing Institute
INFORMS	Institute for Operations Research and the Management Sciences

1. Executive summary

Industrial water management can benefit from the advances in AI, computational power, and IoT. In industrial processes it is increasingly usual to continuously generate different types of data at high velocity, implying that data can rapidly gain high volume. AI and ML techniques can be leveraged to effectively analyze high dimensional data to provide insights, find latent patterns, detect anomalies, predict future values, and prescribe actions to assist in the decision-making process. AI models can offer significant value by enhancing process monitoring & optimization, fault diagnosis, and prognosis in complex engineering systems, like water treatment plants. However, the process of creating and fine-tuning AI and ML models for real-world applications is a challenging and time-consuming task that typically includes numerous trials and experiments. Additionally, the experiment conduction, model selection and deployment require the collaboration of data scientists, domain experts, and software engineers.

This deliverable covers the development of the AquaSPICE Analytics Platform and the demonstration of the statistical and machine learning methods to produce descriptive analytics and the ML and AI techniques to create predictive models for forecasting and anomaly detection. It begins with a concise literature review of AI and data analytics techniques in the field of water treatment. Next, it demonstrates the Analytics Platform architecture, with an emphasis on the Data Analytics Workbench service, which enables analysts to execute ML experiments and manage models for deployment and use by the Data Analytics Service. Following that, the document presents the AI and analytics algorithms, and methods used in descriptive and predictive analytics along with visualizations of their results. Finally, it ends with the conclusions and next steps. Given that it is an intermediate version of the deliverable, it does not report on the results obtained using the Analytics Platform in the use cases, which is work under way and will be reported in the final version of the same deliverable.

2. Introduction

This deliverable reports in the intermediate results of task T4.5, which develops data-driven, data analytics platform for developing and deploying methods enabling the real-time analysis of water treatment process data for the detection of current and emerging situations that require attention specific remedial actions (e.g., anomalies). Typical cases of such situations include emerging water efficiency problems (e.g., in terms of both quantity and quality), posing the need for adaptation or optimization of the relevant processes, e.g., the production or water recovery-treatment-reuse processes.

The proposed data analytics platform supports both descriptive and predictive models. Descriptive analytics methods provide a statistical interpretation used to analyze historical data to identify patterns and relationships. Descriptive analytics seeks to describe an event, phenomenon, or outcome. It helps understand what has happened in the past and provides operators and engineers the base to track trends in the treatment process. Predictive analytics is the process of using data to forecast future outcomes. The process uses data analysis, machine learning, artificial intelligence, and statistical models to find patterns that might predict future behavior.

A specific but widespread case of data analytics is anomaly detection. Our work includes an anomaly detection service, which yields the following results: anomalous events, anomalies based on complex event processing, risk prediction, and root cause analysis for detected anomalies/problems.

The goals of the proposed analytics platform include:

- Facilitate the complete life-cycle of analytics operations, from design to deployment and monitoring.
- Allow retrieval and exchange of analytics models between data analysts and domain experts (engineers) during the phases of problem understanding, data understanding, and evaluation by creation of analytics models and methods.
- Allow experimentation and fine-tuning of analytics models and pave the way for integration data analytics methods and first-principle models.
- Leverage the evaluation of the impact of analytics methods on the key performance indicators (KPIs).

The AquaSPICE data analytics platform utilizes and exploits: (i) real-time data from sensors related to the production process and water recovery-treatment-reuse; (ii) historical data from legacy and operational systems; (iii) expert knowledge and real-life feedback from relevant stakeholders in the production value chain. Upon problem detection, expert knowledge is coupled with machine learning approaches to allow a data-driven understanding of the underlying causal relations.

The developed data analytics service design is being encapsulated as a module for the Digital Twin’s KG, with an API to allow exporting the services to other modules. Given that it is an intermediate version of the deliverable, it does not report on the results obtained using the Analytics Platform in the use cases, which is work under way and will be reported in the final version of the same deliverable.

3. Review of AI and analytics approaches in the water treatment domain

3.1. Introduction

In this section we review prominent approaches that utilize data analytics in water treatment processes. The four rudimentary types of data analytics that are also applied and benefit water treatment processes are descriptive, diagnostic, predictive, and prescriptive. Data analytics provide methods to understand what happened, why something happened in the past, what could happen next, and what should happen in the future (Figure 1).

Descriptive analysis explains events over time; this type of data analytics produces reports by analyzing information to determine the current status of a process in a way that highlights patterns and exceptions. Diagnostic analytics focus on the root cause of the occurrence of an event, trying to provide answers to questions like “why it happened”. Predictive Analytics focuses on events that are expected to happen in the future, framing answers to questions such as what is likely to happen in the future; to compute future probabilities, machine learning, and statistical methods are used [1][2]. Prescriptive analytics focuses on making intelligent decisions toward process optimization based on predictive analytics results [1][2][3].

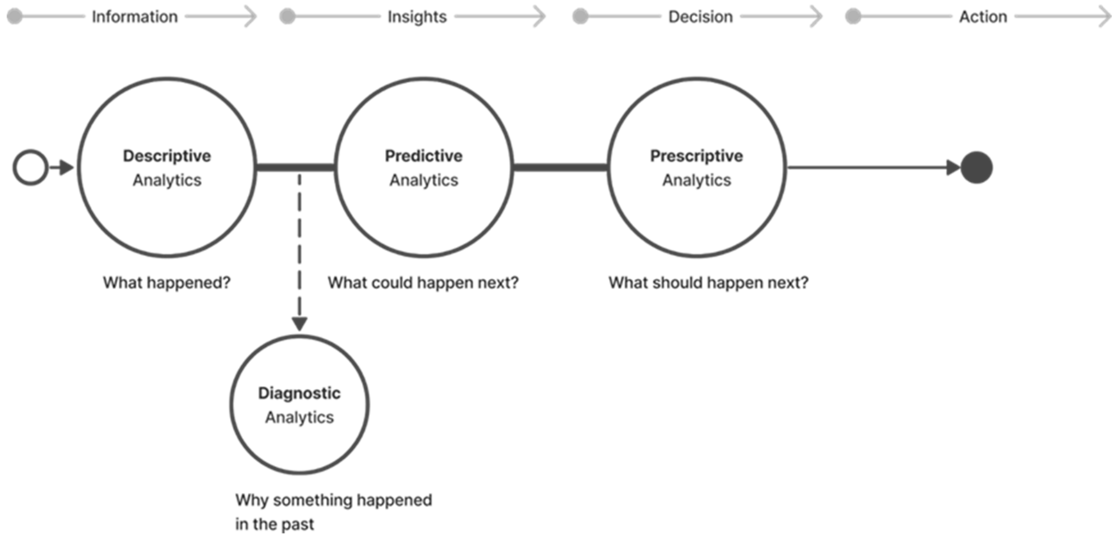


Figure 1 : Data analytics types. (adapted from [1])

3.2. Literature Review Methodology

Our review adopts the systematic method and the approach proposed by Tranfield et al. [4]. Systematic reviews and meta-analysis originally appeared in medical science research [5]. Efforts were made to include systematic method in other disciplines such as business research [4][7]. Although their guidelines were initially proposed for the management discipline, they are suitable for systematic reviews of various sectors [7].

An initial search was conducted in August 2022, followed by a complementary search in February 2023. The online citation databases considered in this review were Scopus, Google Scholar, Emerald. Our research focuses on articles and conference publications of the last decade; therefore, the results were narrowed to the period from 2011 to 2021. To ensure the results' quality, the publishing companies considered are the following: Elsevier ScienceDirect, MDPI, IOP, IEEE, Emer, Springer and IWA.

To derive articles related to the data analysis lifecycle in water treatment, the main search term was "wastewater analytics". A summary of the queries and the number of the initial results are presented in Table 1. Due to the limited relative results, we also tried other keywords combined with the basic term "wastewater" such as "monitoring", "smart", "machine learning", "online monitoring" and topic-specific terms such as "membrane fouling" and "digital twins".

Table 1. Literature Review Fundamental Queries and Results

Database	Query	#Results
Scopus	TITLE-ABS-KEY ("wastewater analytics" OR "wastewater" AND "analytics") AND PUBYEAR > 2011 AND (LIMIT-TO (PUBSTAGE , "final")) AND (LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "cp")) AND (EXCLUDE (PUBYEAR , 2023) OR EXCLUDE (PUBYEAR , 2022)) AND (LIMIT-TO (LANGUAGE , "English")) AND (LIMIT-TO (EXACTKEYWORD , "Wastewater Treatment") OR LIMIT-TO (EXACTKEYWORD , "Wastewater") OR LIMIT-TO (EXACTKEYWORD , "Waste Water"))	103
Google Scholar	"wastewater analytics"	20
Emerald	(content-type:article) AND (wastewater analytics)	40

Search results were refined in three main iterations to elicit each article's relatedness to the subject matter. First, if the title or the abstract is pertinent then the article passes directly to the next phase. Second, the full corpus of the remaining articles was leafed through to exclude irrelevant ones. Finally, all the selected articles were read extensively, shaping the final set for the review. During the iterations, all duplicate articles were removed and the following reasons for exclusion were considered:

- Irrelevant subject
- Review articles
- Not quantitative
- Absence of data analytics techniques

The final set of articles consists of 21 items; their distribution throughout the years and the publishers is shown in Table 2 and Table 3, respectively.

Table 2. Distribution of the final papers throughout the years

Year	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Papers	1	0	1	2	1	1	4	3	2	2	3

Table 3. Number of papers per publisher

Publisher	**Papers**
Elsevier	12
INFORMS	1
IWA	1
MDPI	3
Springer	1
Taylor & Francis	2

3.3. Application Areas

The application areas of data analytics in water treatment as elicited from literature are briefly described in this section. Water quality is a significant concern throughout most of the articles and the water treatment process itself, thus it was not considered as an application area. Water quality refers to recommended standards for the quality of the final effluent or the inflow of a water treatment plant and is determined by measuring water quality indicators against parametric standard values and regulatory requirements. Water quality indicators include physical, chemical (inorganic and organic) and microbiological characteristic parameters [8][9].

3.3.1. Instrumentation and Software Sensoring

A variety of measurement and analysis instruments are utilized in water treatment processes; turbidity, pH analyzers and water quality sensors, to name a few. Along with the hardware development, advancements in big data analytics and software sensors promote online monitoring and lessen the dependency on hardware sensors [11][12].

A soft sensor ("software sensor") is a software that estimates a hardware-like signal. It is utilized to indirectly measure variables that are difficult to measure due to cost or technical limitations. Soft sensors use process data as input to a model that predicts the target variable values. The process data can typically be obtained relatively easily and are composed of signals from hardware sensors and actuators. The prediction model can be classified as data-driven, knowledge-based, or hybrid [13][14][15].

3.3.2. Optical monitoring

Optical monitoring of water is typically performed via examining the water samples manually under a microscope [16][17]. Advances in optical monitoring devices, IoT and data analysis allow modern approaches to suggest real-time (online) optical monitoring for rapid in-situ transmittance measurements. Online optical monitoring combined with image analysis methods and possibly predictive models could provide useful insights to control and optimize the water treatment process [17][18].

3.3.3. Process monitor & control (fault detection, diagnosis, and prognosis)

Advanced monitoring systems enhance the control of the water treatment process by detecting malfunctions, and sensor faults and identifying abnormal process operations or conditions [12][19].

3.3.4. Outlier detection

Outlier detection concerns the identification of observations that fall outside of an expected distribution or pattern; such abnormal observations are called outliers or anomalies [20].

3.3.5. Performance evaluation

The performance of a treatment plant is evaluated via the degree of reduction of BOD, COD and SS, which constitute organic pollution [21]. The performance efficiency of treatment plant depends not only on proper design and construction but also on good operation and maintenance [22][23].

3.4. Analysis of existing approaches

In this section, we further discuss the methods that have been used in literature.

3.4.1. Water Quality Prediction

Han et al. [24] proposed the Flexible Structure Radial Basis Function Neural Network (FS-RBFNN) to develop a water quality prediction model for wastewater treatment systems. The self-organized architecture of the RBFNN is assumed from neuron activity and mutual information (MI). The model was tested in forecasting the Biochemical Oxygen Demand (BOD) of effluent in a wastewater treatment process through a simulation; its accuracy was better compared to other self-organizing algorithm models, namely Adaptive Merging and Growing Algorithm (AMGA), sensitivity-based Radial basis function (SenV-RBF), minimal resource-allocating network (MRAN) and generalized growing and pruning RBF (GGAP-RBF).

Zare Abyaneh [25] evaluated Multivariate Linear Regression (MLR) and Artificial Neural Network (ANN) models in BOD and Chemical Oxygen Demand (COD) prediction of a Wastewater Treatment Plant (WWTP) using minimum input parameters. The data consisted of laboratory measurements throughout 7 years. The ANN model outperformed the MLR and both models appear to predict BOD more accurately. The study suggests that the type of parameters is more significant than their number.

Guo et al. [26] developed and compared ANN and Support Vector Machine (SVM) to predict Total Nitrogen (T-N) concentrations of effluent for integrated food waste and wastewater treatment processes. Meteorological and water quality data were recorded daily during a ten-month period; the latter were produced by laboratory and in situ measurements while the samples were collected from various spots of the wastewater treatment plant. Although both models were sufficient in predicting concentrations, SVM showed higher accuracy. Additionally, a sensitivity analysis was applied and considered the ANN as a better choice through the prism of cause-and-effect relationship between T-N and input parameters.

Tomperi et al. [17] examined the dependencies between process variables and optical measurements, aiming to use optical monitoring to predict common water quality parameters in a WWTP. In-situ optical monitoring was performed by a high- resolution charge-coupled device camera in a real WWTP for over one year. Image analysis variables combined with laboratory process data formed the data set of the experiment. The input variables were selected through five variable selection methods and the prediction model for each water quality parameter was developed using MLR. The study highlights the advantages of online optical monitoring.

Tomperi et al. [27] also demonstrated the prediction of five effluent water quality parameters using MLR and solely optical monitoring variables as inputs. The wastewater samples for the automatic optical monitoring were collected from an industrial activated sludge process of a pulp and paper industry for 13 months. All models were sufficient, resulting that optical monitoring in combination with predicting models could be a powerful tool in process control.

Nadiri et al. [28] introduced a Committee Fuzzy Logic (CFL) and a supervised CFL model that synthesizes three Fuzzy Logic (FL) models to simulate wastewater treatment plant operations and predict effluent quality. In the CFL model the predictions emanate from linear combinations of the FL models, whereas in the Supervised Committee Fuzzy Logic (SCFL) the FL model outputs were combined as inputs to an ANN that constitutes the final prediction. Historical data composed the training and evaluation dataset. As a result, the CFL model performed better than the individual FL models and the SCFL further improved the predictions' accuracy.

Table 4 below consolidates the key findings discussed in this section.

Table 4: Data Analytics Methods for Water Quality Prediction in Reviewed Literature

Paper	Data source	Analytics methods	Input	Output	Evaluation
[24]	Historical	FS-RBFNN	COD, SS, pH, oil, NH_3–N	BOD	MSE RMSE CPU time
[25]	Laboratory	ANN-MLP MLR	TSS, TS, pH, Temperature	BOD, COD	r RMSE bias values
[26]	Laboratory In situ	ANN SVM	month, volumetric inflow flow rate, TSS, COD, T-N, T-P, temperature, pH	T-N	R^2 NSE relative efficiency criteria (drel)

[17]	In situ optical monitoring Laboratory	MLR	different variables per model	SS BOD COD T-N T-P (Separate models)	RMSE R^2
[27]	In situ samples	MLR	optical monitoring variables (amount of filaments, fractal dimension, form factor, roundness, aspect ratio, equivalent diameter, mean area of objects, number of small objects)	(BOD, COD, SS, N, P) of effluent	R^2 RMSE coefficients of regression
[28]	Historical	SCFL ANN CFL TSFL MFL LFL	(BOD, COD, TSS, Temperature, pH) of influent	(BOD, COD, TSS) of effluent	MAPE RMSE R^2

3.4.2. Anomaly, Outlier and Fault Detection

Lepot et al. [29] conducted outlier detection and identification of most representative spectrum in repetitive spectral recording of wastewater samples, using Principal Component Analysis (PCA) and Data Depth Theory (DDT). Ultraviolet-visible (UV-Vis) spectrophotometers collected data from a WWTP in France for four days and a Moving Bed Biofilm Reactor (MBBR) at a WWTP in Switzerland for ten weeks. The results were promising, revealing weaknesses and strengthens of the selected methods.

Chow et al. [30] proposed a real-time anomaly detection system for early warning, using an online UV-Vis spectrophotometer, k-means and visual exploration. The device was placed at the inlet of a WWTP and was collecting data for eighteen months. A developed portal was responsible for the data integrations, visualizations and data analytics procedures. The system was evaluated in an operational environment.

The two papers are summarized in the following table (Table 5).

Table 5. Data Analytics Methods for Anomaly, Outlier and Fault Detection in Reviewed Literature

Paper	Data source	Analytics methods	Input	Output	Evaluation
[29]	UV/Vis spectrophotometer	Data depth theory PCA	*France*: spectral data and for each sample: TSS, total, dissolved COD *Switzerland*: spectral data, ammonium, nitrite and nitrate concentrations	Outliers in repetitive spectra	Confusion matrix Consistency ratio
[30]	online UV-Vis spectrophotometer	visual exploration k-means correlation analysis	spectral data logs	Abnormal inlet water quality	Correlation coefficients against logs

3.4.3. Process simulation, optimization, and control

An in-depth review of data-driven approaches for the analysis of a WWTP performance is provided in [59], also discussing the importance of data-driven decision-making in plant operation optimizations.

Boujelben et al. [22] investigated the performance of four wastewater treatment plants applying Redundancy analysis (RDA), Analysis of Variance (ANOVA) and Duncan test. During the four-year study, physicochemical and biological properties, flow rates, energy consumption and water quality indicators were collected from laboratory analyses, in-situ measurements and external sources. The multivariate analysis showed that differences in capacity, treatment processes or properties of influence affect the performance of the WWTP, which can be improved by a proposed electrolysis of the output wastewater.

Zadorojniy et al. [31] examined reinforcement learning, multivariate adaptive regression splines and Constrained Markov Decision Process (CDMP) analytical approaches to reduce the costs and improve the efficiency in a WWTP. The CDMP was comparatively faster and of higher quality. They built a simulation model and plant-state estimators using historical data from a real WWTP. A transition probability matrix models the process behavior of the plant. Finally, the CMDP optimization unit provides optimal recommendations. The pilot ran for a year in a real WWTP achieving cost reduction alongside other benefits. The solution remains to be applied in larger plants to ensure its generalization.

Arismendy et al. [32] developed an intelligent system to support decision making in WWTP through COD prediction and visualizations. Time-series decomposition, autoregression and ANN were utilized to succeed the COD forecast. The dataset employed in the study contained almost two and a half years of daily sample measurements of major water quality parameters. A web-based platform was also designed to monitor the predictions, present data visualizations, and allow access to the historical data.

Arismendy et al. [33] also presented a data driven decision making system that prescribes actions to optimize the processes of an industrial WWTP, based on predictive data and expert's knowledge. Long-Short Term Memory (LSTM) ANN and decision tree algorithm are utilized to develop the prediction model and the estimation algorithm respectively, using the forementioned dataset. A genetic optimization algorithm is designed targeting the COD value. Future work is said to include more variables and finalize the proposal.

Asami et al. [23] developed and compared ANN and M5 model tree in simulating and predicting the performance of a WWTP and its effluent water quality. A dataset was created mainly from an online wastewater analyzer's daily measurements of water quality parameters for three years. The evaluation proved ANN to be superior to M5 model tree, although both models appeared reliable and could be used in missing data estimations and environmental decisions.

Xiao et al. [34] suggested a fault diagnosis and prognosis framework that combines auto-associative neural networks and Autoregressive Moving Average (ARMA) model, respectively. A recursive minimization strategy is proposed to handle missing data values. Furthermore, Kernel Density Estimation (KDE) control limit was developed to reduce type-I and type-II errors. The models were built and evaluated with simulated WWTP process data. Shallow and deep ANN were compared to and surpassed PCA and kernel PCA models. The framework detected sensor and process faults successfully; additionally, the ARMA model was capable of multi-step-ahead Squared Prediction Error (SPE) prediction.

Andersson et al. [35] proposed a tool to predict future environmental impacts of different WWTP operations and influent conditions, combining process simulation and influent generation models with Life Cycle Assessment (LCA) models. Process models were used to simulate three real WWTPs and the experiment was examined in different scenarios.

The following table (Table 6) provides a concise summary of the literature findings that were discussed in this section.

Table 6. Data Analytics Methods for Process Simulation, Optimization and Control in Reviewed Literature

Paper	Data source	Analytics methods	Input	Output	Evaluation
[22]	Laboratory In situ external sources	RDA ANOVA Duncan test	electrical conductivity, salinity, chlorides, flow rate, energy, data on rainfall, COD, BOD, energy efficiency, TSS,	n/a	n/a

			TKN, NH_4^+, NO_3^-, TP, removal efficiency, faecal coliforms, faecal streptococci, detection of Salmonella and Vibrio cholera, Pb, Cu, Ni, Zn, Cr, Cd, electrolysis		
[31]	Historical sensor data Simulation Three-step interpolation algorithm	Transition probability matrix Multivariate Adaptive Regression Splines (MARS) Constrained Markov decision process (CDMP)	state variables: {influent flow rate, feedback, effluent total nitrogen concentration, effluent total phosphorus concentration, period, total cost}	action variables: {DO set point, waste-activated sludge pump rate, Internal recycle pump rate}	Comparison of policies after 2 years of pilot
[32]	Historical data from different stages of WWTP	ANN-MLP	Flow, COD influent, SS, MLSS, MLVSS, N, pH, DO, F/M	COD	MAPE MSE
[33]	Historical data from different stages of WWTP	i. LSTM ii. Decision Tree iii. GA	i. {BT_C_MLVSS, D_SS, EQ_N, Clari_DO} more than once on different days ii. {EQ_N OxT_PH_PM} more than once on different days iii. pH setpoints	i. EQ_COD ii. EQ N iii. best pH	MAPE Mean & standard deviation T-student & F-Fisher test Box & whisker plot comparison
[23]	online wastewater analyzer TSS St.M.2540-D	ANN M5 model tree	temperature, turbidity, pH, EC, TDS, TSS, DO, BOD_5, COD of the inlet	BOD_5, COD, TSS of output	R^2 $R^2_{adjusted}$ RMSE Standard error of the estimate
[34]	BSM 1 simulation	SANN DANN KPCA ARMA	process data	SPE	SPE Type-I error Type-II error
[35]	process simulation: BSM2G	LCA models Influent Generation model	influent generator BSM-UWS model (scenarios)	Environmental impacts of future processes	

	influent generation: BSM-UWS	Biochemical process models			

3.4.4. Soft Sensors

A comprehensive study of the role and scope of soft sensoring methods and models has been performed by Haimi et al. [60]. Liu et al. [12] developed a probabilistic self-validating software sensor that handles missing and abnormal values and produces interval predictions. Variational Bayesian Principal Component Analysis (VBPCA) is used to detect faults and to reconstruct the corresponding value at the pre-process stage; contribution plots are then utilized to identify the root of the disturbance. Relevant Vector Machine (RVM) was selected as the prediction model that also considers uncertainties from parameters and noise. The soft-sensor application was trained in simulation data and was evaluated in two simulation case studies.

Blanco-Rodríguez et al. [36] employed an electronic nose to discriminate odors and predict their concentration, composing a qualitative and quantitative analysis. The data was collected via flux chamber and direct sampling from six possible odor sources in an urban WWTP. PCA was used for odor pattern identification and Partial Least Squares (PLS) regression for the prediction. The authors concluded that e-nose constitutes a reliable and economic tool for wastewater treatment monitoring.

In Moreira de Lima & Ugulino de Araújo [37] a deep representative learning soft-sensor modeling approach named MISAEL is presented for industrial plants. MISAEL integrates Mutual Information (MI) based stacked autoencoders (SAE) with LSTM. After every SAE layer that extracts the hidden features, an MI analysis is performed to reduce irrelevant information; LSTM networks are then utilized to produce the final result. Two existing industrial case studies were considered to train, test and validate the approach. MISAEL and ensemble MISAEL, that includes a k-fold cross-validation, outperformed other methods under the same conditions.

The key results are summarized in the table below (Table 7).

Table 7:Data Analytics Methods for Soft Sensors in Reviewed Literature

Paper	Data source	Analytics method	Input	Output	Evaluation
[12]	BSM 1 simulation	VBPCA RVM	{COD, NH_4^+, NH_3 nitrogen, Nitrate and nitrite nitrogen, BOD, Flow rate, Oxygen} {BOD, COD, pH, Suspended solids, Sedimentable solids}	BOD, COD	r RMSE LS-SVM PLS
[36]	direct sampling flux chamber sampling	PLS regression PCA	e-nose and olfactometry outputs	odour concentration	RMSE R^2
[37]	pre-existing datasets: i. Debutanizer column ii. Sulfur Recovery Unit	SAE LSTM	i. {top temperature, top pressure, reflux flow, flow to next process, sixth tray temperature, bottom temperature A, bottom temperature B} ii. {gas flow, air flow, secondary air flow, gas flow in SWS zone, air flow in SWS zone}	i. Butane C4 content in IC5 ii. Concentration of SO2in the tail gas	R^2 RMSE Comparison

3.5. Synthesis of Findings and Propositions

Towards the adoption of data analytics for industrial water treatment by AquaSPICE, we synthesize the challenges derived from the literature review and we outline potential directions for future research.

3.5.1. Offline vs. real-time processing approaches

According to our literature review, real-time processing and analysis of data has not been well exploited; most of the reviewed papers deal with offline analysis of data. However, the wide adoption of sensors in modern industries has led to an increasing demand for real-time analytics. Apart from the technical challenges that exist in developing scalable and efficient sensor-driven information systems [38][39], there is also the need for bespoke and algorithms for data analytics, which can process data with time-varying characteristics and thus can solve problems utilizing large scale streaming data. For example, recursion-based data analysis algorithms can be applied in cases where a problem that depends on solving smaller instances of the same problem.

Proposition 1: The development of real-time, sensor-driven data analytics systems and recursive algorithms can promote the adoption of water analytics in industrial applications.

The computational challenges are even higher when analytics algorithms need to be developed for distributed platforms [38], i.e. platforms with components that are located in different networked computers, communicating and coordinating their actions by passing messages to each other [40]. Water analytics in industrial applications can significantly benefit from distributed computing for processing large amounts of unstructured, semi-structured and structured big data. The AquaSPICE water analytics platform has been designed so that it can cope with both batch and streaming data and has provisions for distributed processing of big data.

Proposition 2: Water analytics in industrial applications can benefit from distributed computing for processing large amounts of data.

3.5.2. Targeted and Actionable Analytics

Expertise in data analytics, it is not enough to be versed in analytics in an industrial domain such as water treatment. One needs to understand how to use analytics to solve the specific water management problems. This requires insight into the data themselves as well as the targeted problems. Typical problems in process industries and other industries with large water processing needs are, for example, control and optimization problems. For example, controlling the water flows can benefit from influent and effluent stream flow predictions while prescriptive analytics can facilitate the optimization of water flows and treatment processes. Water analytics provide the ability to leverage optimization — the primary prescriptive analytics tool — to find solutions to complex problems and make optimal decisions. The AquaSPICE data analytics platform will provide generic descriptive, predictive and prescriptive methods but will also cater for the development and deployment of bespoke methods required to address targeted requirements. Moreover, it will synergistically couple with the optimization component of WaterCPS to enable holistic decision making in complex scenarios requiring both data-driven analytics and model-driven optimizations.

Proposition 3: Water analytics in industrial applications should include prescriptive analytics methods that leverage control and optimization of industrial processes.

3.5.3. Synergies with First Principle Models

The cost, time and skill required to develop application-specific models have been barriers to using first-principle modelling tools in industrial water treatment processes. First-principle models require in depth understating of the underlying physical and chemical phenomena, underpinning processes as well as experimental data or/and statistical methods to estimate model parameters. They are not as quick and easy to build, but they have many advantages. In terms of simulation, first-principle models provide extrapolation in addition to the interpolation provided by data-driven models, but they also can be used for monitoring, control and optimization.

There has been growing interest in blending physical and machine learning models to leverage their respective strengths in many fields including weather forecasting, biological systems, materials chemistry, mechanical failure, battery health, and battery safety [41]. Several possible integration architectures for physics based and machine learning models are outlined in [41]. At a high level, there are two broad categories for health forecasting: (A) serial integration of independent models and (B) hybridized PB and ML models. The former category involves architectures more viable in the near term as they can be realized by integration of existing ML and PB tools without any fundamental changes. The latter category will require the development of new approaches.

In the water treatment domain, existing works have focused on employing a priori knowledge to reveal relationships between subsets of regression parameters that serve to restrict their range as well as on constrained regression, i.e., restricting the original problem in the space of predictor and response variables [42]. Existing works have shown that combining data-driven and theory-driven models results in higher quality surrogate models. The improvements are measured by both physical relevance and model accuracy. We aim to take advantage of such approaches; specifically, we will work towards the development of calibration methods for first-principle methods based on data-driven methods as well as other synergistic approaches to address emerging decision making requirements.

Proposition 4: Water analytics in industrial applications can benefit from theory-driven models and should be considered for setting limits on the response variables; establishing known relationships between response and root-cause variables; and relationships among responses.

Further research is needed towards the direction of combining the ‘learned knowledge’ of machine learning and data mining methods with the ‘engineered knowledge’ elicited from domain experts. To this end, the combined use of machine learning and knowledge engineering can complement each other’s strengths and mitigate their weaknesses, since explicitly represented application knowledge could assist data-driven machine-learning approaches to converge faster on sparse data and to be more robust against noise.

With the advancements in big data technologies, artificial intelligence has become an important element of digital systems, because, among others, they make a profound impact on human decision making [43]. As a result, there is an increasing demand for information embedding artificial intelligence and machine learning algorithms for decision making [43]. In this way, there will be the possibility to develop generic, domain-agnostic, data-driven methodologies and algorithms for performing prescriptive analytics.

Proposition 5: AI can assist the development, application and management of water analytics methods in the industrial water treatment domain.

3.5.4. Knowledge Sharing with Analytics

The adoption of knowledge sharing practices may improve the communication between data scientists and domain experts during data analysis projects by capturing and externalizing communicated or created knowledge. Such knowledge sharing practices can support and improve the Problem and Data understanding phases of the data analytics process by providing a common ground when describing the domain and related problem and data aspects. On the other hand, the data scientists could use the knowledge sharing model to define the predictive functions to solve the data analytical tasks and to present to the domain experts, how such functions influence the domain KPIs. The data-driven methods of the AquaSPICE data analytics platform, as well as first-principle models supported by the AquaSPICE Process Simulation and Modelling (PSM) tool will be linked with the Knowledge Graph already developed and delivered with D4.2 in order to enable knowledge sharing between water treatment experts. [44]

Proposition 6: Knowledge sharing may be leveraged by best practices and know-how sharing and can support the cross-process adoption of water data analytics methods.

4. AquaSPICE Analytics Platform

In the context of the AquaSPICE project, a comprehensive data analytics platform for both real-time water monitoring data and historical data has been designed and developed. The target user group of the platform is water management professionals and data analysts. The platform performs various data analytics tasks, including descriptive statistics and analytics, diagnostic reports, predictive analytics, and anomaly detection; the results are then visualized providing the end user with useful insights and graphs. In addition, analysts can perform more complex analyses and experiment with machine learning models via a graphical user interface that also provides experiment tracking and logs throughout the analytics lifecycle. The analysts can select and deploy the models that will afterward be used to produce the data analytics insights.

4.1. Review of Relevant Technologies

As artificial intelligence and data analysis techniques are increasingly integrated into large-scale projects, managing, and maintaining machine learning experiments can present some challenges. These experiments involve various assets, such as algorithms, datasets used for training, evaluating, and testing the algorithms, hyperparameters, parameters, model metrics, artifacts, and logs. Artifacts can include any output during the experiment process, such as files, models, and checkpoints, while logs may contain details such as date, time, duration, resource metrics, errors, and runtime messages. As the number of experiments and project complexity grows, the need to keep track of all the experiments and their assets more effectively becomes more evident. This need was met by experiment tracking tools, which are important assets that accompany the development and maintenance of AI projects. Experiment tracking tools are advantageous for the development and maintenance of AI projects as they help to manage and organize experiments effectively. Additionally, they promote experiment reproducibility, sharing and comparison [45][46].

An overview of the experiment tracking tools that were examined in the research context are summarized in the following table (Table 8) and includes Kubeflow[1], MLflow[2], Comet ML[3], Neptune[4] and Guild AI[5].

[1] https://www.kubeflow.org/

[2] https://mlflow.org/

[3] https://www.comet.ml/

[4] https://neptune.ai/

[5] https://guild.ai/

Table 8: Experiment Tracking Tools Comparison: Pros and Cons

Tool	Pros	Cons
Kubeflow	Open-source Comprehensive platform WEB UI Complex pipelines orchestration Active community	Complexity Can be used with the Kubernetes container orchestration system only Learning curve Resources Intensive Extensive code & infrastructure changes
MLflow	Open-source Comprehensive suite WEB UI Large community	Resources Intensive Simpler ML pipelines
Comet ML	Collaboration Rich visualization WEB UI	Learning curve Pricing
Neptune	Collaborative platform WEB UI	Pricing
Guild AI	Open-source No code changes Local resources Command line and WEB UI	Smaller community Smaller teams

The monitoring needs differ per user role and project. Considering the required resources cost, the tool's complexity and the project's needs, it was deemed appropriate to develop a custom solution. This solution adopted a custom comprehensive approach that combines Guild AI and a user-friendly graphical interface that additionally allows for data selection, experiment conduction and monitoring as well as model optimization, deployment, and management, all wrapped in a microservice that offers additional features for the data scientists and professionals; these features are presented in more detail in the next chapter. Guild AI was selected due to its simplicity and reliability. It requires minor code changes and configuration, making the platform more independent. The lack of user interfaces was considered advantageous, promoting the development of custom graphical user interfaces that include actions that are relevant but not too complex for the users.

4.2. Architecture

The AquaSPICE data analytics platform has been designed based on the principles of microservices architecture. Microservices architecture is a software architectural style that, as opposed to monolithic architecture, divides an application into smaller, independently deployable services, each focused on a specific scope [47]. These services communicate with each other usually through APIs or messaging systems; their collection structures the comprehensive large-scale software application. Microservices are designed to promote modularity, scalability, rapid development, resilience, and fault tolerance. This approach is particularly suitable for complex enterprise applications, like data analytics, where different components may have varying scalability and lifecycle requirements [47][48]. However, it is crucial to carefully design the architecture as well as to effectively manage the added complexity.

Two microservices have been created for the platform: the Data Analytics service and the Data Analytics Workbench service. Each microservice is framed by a well-defined scope that identifies its expertise. Data Analytics service is responsible for computing, presenting, and visualizing the data analytics results to the Water CPS and digital twin systems. Data Analytics Workbench service's aims to provide a place where analysts can create and execute machine learning experiments, compare models, and select to deploy the model to then be retrieved by the Data Analytics service for predictions and anomaly detection.

Separate databases are kept by the microservices. In the Data Analytics service, analytics metadata are stored in the database, the metadata include pilots, processes, assets, and machine learning models identifiers. Experiment logs and metadata produced during data analytics workbench actions are stored in the service's database and might include model hyperparameters, results, dataset variables and selected model per data analytics process. Whenever a new model is deployed from the workbench, the model is kept in cloud storage and the other service is triggered to retrieve the new model and update the related data analytics results. Communications from the external Water CPS and Digital Twins systems are achieved via API calls; the requests are routed to the suitable microservice which in turn responds with the appropriate result. Each microservice receives real-time data streams from an external broker. A diagram that demonstrates the architecture and the high-level communications between the services and the external Water CPS and digital twin systems is presented in Figure 2.

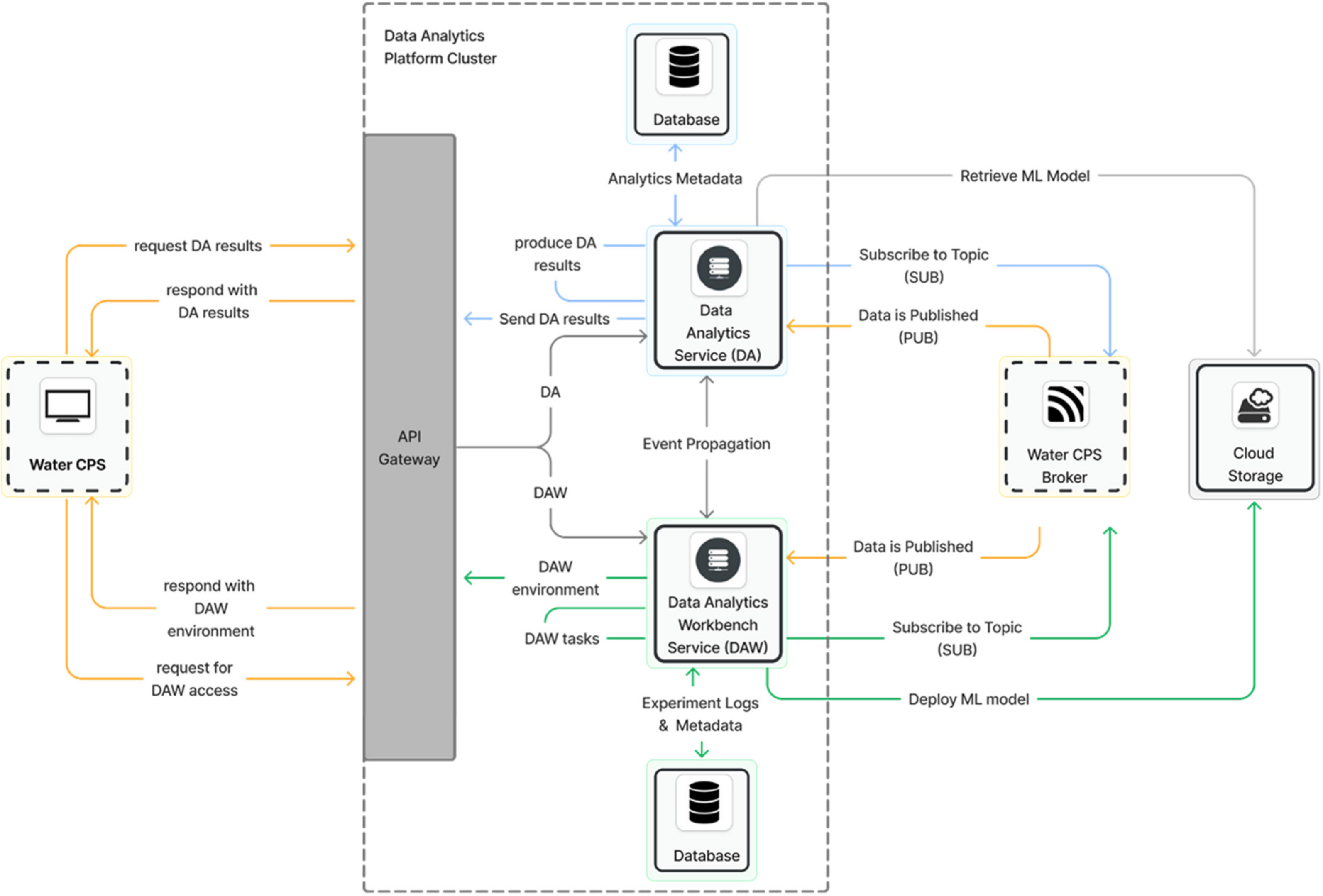


Figure 2: Architecture and High-level Data Flow Diagram

4.3. Development

Reliable state-of-the-art tools were selected for the development and deployment of the platform to create a robust and scalable ecosystem. Each microservice has a specific purpose and utilizes distinct technologies, although there are also shared ones. The technology stack is divided into the following categories:

1. Front-End
2. Back-End
3. Web Infrastructure
4. Cloud Infrastructure
5. Orchestration
6. Development tools

Front-end is the client-side of an application responsible for the information presentation and user interactions; it includes the graphical user interface. For the data analytics service, Streamlit[6] along with CSS[7] and JavaScript programming language were utilized. Streamlit is an open-source Python library designed for the presentation of data science projects. The workbench is developed with Vue.js framework, and it follows Material Design principles. Material Design is an adaptable system of guidelines and components that supports the best practices of user interface design. Additionally, HTML5, CSS and JavaScript were used to adapt the component styles.

[6] https://streamlit.io/

[7] https://www.w3.org/TR/css-2022/

Back-end is the server-side of an application; it is responsible for the business logic, the data processing, database operations and data delivery to the front-end. Both services are developed with Python programming language and FastAPI[8] web framework. PostgreSQL[9] database is selected as the relational database management system while InfluxDB[10] is utilized for the time-series data operations. InfluxDB is a NoSQL database designed for efficiently storing, querying, and visualizing large volumes of time-series data, well suited in IoT and industrial automation scenarios. In order for the data analytics workbench to retrieve real time data streams from the WaterCPS broker, MQTT[11] (Message Queuing Telemetry Transport) protocol is used to create an MQTT client and establish this connection and subscribe to the relevant topic. In both microservices the fundamental back-end libraries, leveraged for the machine learning processes, are Tensorflow[12] and Keras[13] deep learning frameworks, Scikit-learn[14] as a machine learning framework and Plotly[15] for the data visualizations. Tensorflow and Keras are used for deep learning model development, training and ingestion while Scikit-learn provides machine learning algorithms for regression, classification, clustering, and other common ML tasks. Additionally, PyOD[16] is employed for anomaly and outlier detection algorithms. Specifically for the experiment tracking and model optimization in the data analytics workbench, GuildAI is applied.

Uvicorn[17] and Nginx[18] were combined to build a scalable web infrastructure. Uvicorn is an Asynchronous Server Gateway Interface (ASGI) responsible for serving the FastAPI-developed application, while Nginx serves as a reverse proxy managing security and load balancing.

[8] https://fastapi.tiangolo.com/
[9] https://www.postgresql.org/
[10] https://www.influxdata.com/home/
[11] https://mqtt.org/
[12] https://www.tensorflow.org/
[13] https://www.tensorflow.org/guide/keras
[14] https://scikit-learn.org/
[15] https://plotly.com/
[16] https://pyod.readthedocs.io/en/latest/
[17] https://www.uvicorn.org/
[18] https://www.nginx.com/

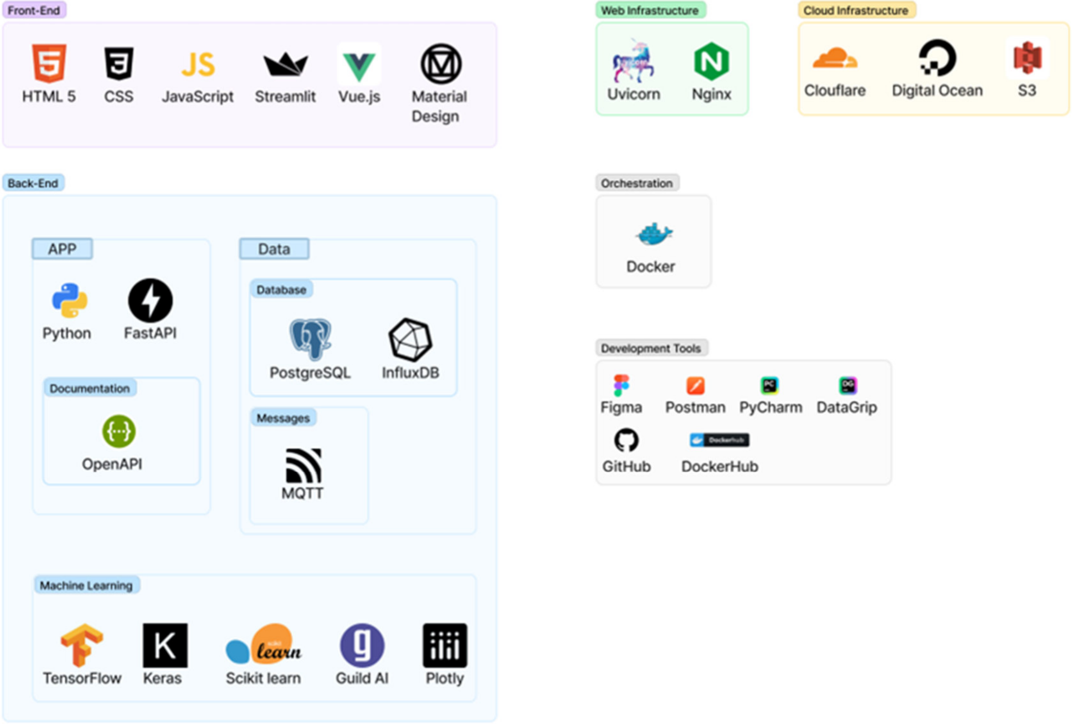


Figure 3: The Technology Stack of the Developed Platform

Concerning the recent state of the cloud infrastructure of the platform, Cloudflare[19] is used as the Content Delivery Network (CDN) and Domain Name System (DNS), Digital Ocean as the cloud provider, and Amazon S3 as cloud storage for objects like ML models. This implies that the platform is hosted on Digital Ocean[20] servers, Cloudflare caches and distributes the static content of the platform, and Cloudflare handles the DNS queries made by a user's device resolving the DNS.

For containerization and packaging of the platform and its dependencies, docker is leveraged. Docker[21] is widely used in microservices architecture promoting the maintainability and scalability of the independently deployable services.

During the development phase, the more significant tools that supported and promoted the results are Figma for the designs, PyCharm IDE[22] for Python as well as DataGrip IDE[23] for databases, Postman[24] for the API testing, GitHub[25] for version control, and DockerHub[26] as the container image registry.

[19] https://www.cloudflare.com/

[20] https://www.digitalocean.com/

[21] https://www.docker.com/

[22] https://www.jetbrains.com/pycharm/

[23] https://www.jetbrains.com/datagrip/

[24] https://www.postman.com/

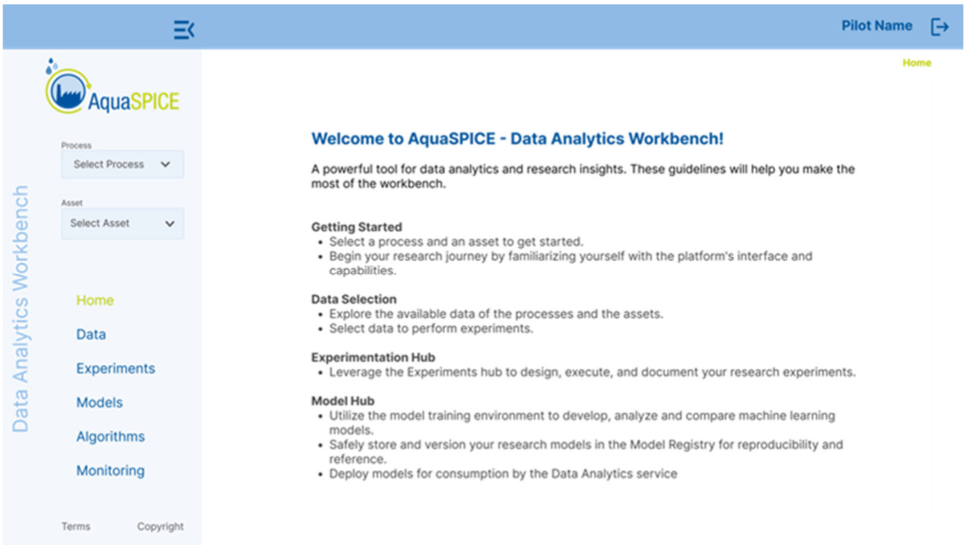
[25] https://github.com/

[26] https://hub.docker.com/

The complete ecosystem of the technologies that are utilized across the platform is summarized in Figure 3.

4.4. Implementation

The main scope of the data analytics workbench is to empower analysts and data professionals to conduct and track experiments using machine and deep learning methods in a simplified yet inclusive and user-friendly environment. Additional functionalities provide insights and information.



Welcome to AquaSPICE - Data Analytics Workbench!

A powerful tool for data analytics and research insights. These guidelines will help you make the most of the workbench.

Getting Started
- Select a process and an asset to get started.
- Begin your research journey by familiarizing yourself with the platform's interface and capabilities.

Data Selection
- Explore the available data of the processes and the assets.
- Select data to perform experiments.

Experimentation Hub
- Leverage the Experiments hub to design, execute, and document your research experiments.

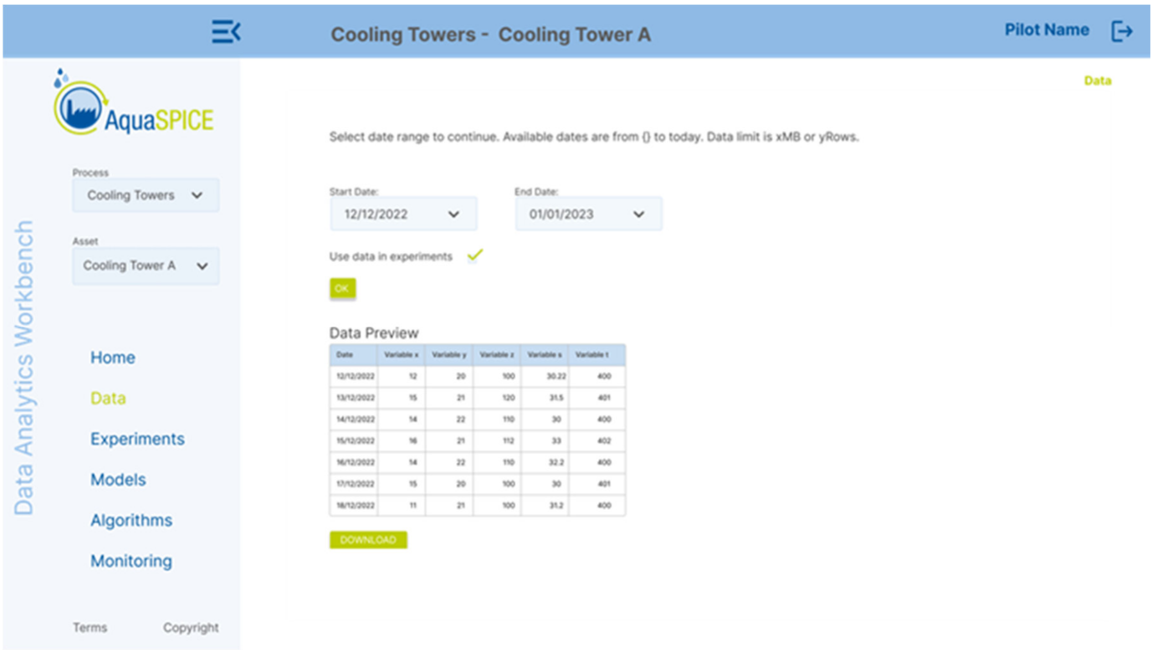
Model Hub
- Utilize the model training environment to develop, analyze and compare machine learning models.
- Safely store and version your research models in the Model Registry for reproducibility and reference.
- Deploy models for consumption by the Data Analytics service

Figure 4: Data Analytics Workbench - Home Page

On the home page core guidelines inform the user of the basic functionalities of the workbench. (Figure 4) From the side menu, the user can select the process and a specific asset to continue to other tasks such as data selection, experiments, and models. The main menu items are always accessible from the side menu. Each menu item corresponds to a specific page; the menu items consist of the following options:

1. Home
2. Data
3. Experiments
4. Models
5. Algorithms
6. Monitoring

On the “Data” page (Figure 5), users can select, and preview existing datasets. Those datasets consist of historical data that are derived from the selected asset for the selected date range and can be used in the experiments. Alternatively, the user can download the data.



Select date range to continue. Available dates are from {} to today. Data limit is xMB or yRows.

Start Date: 12/12/2022

End Date: 01/01/2023

Use data in experiments ✓

OK

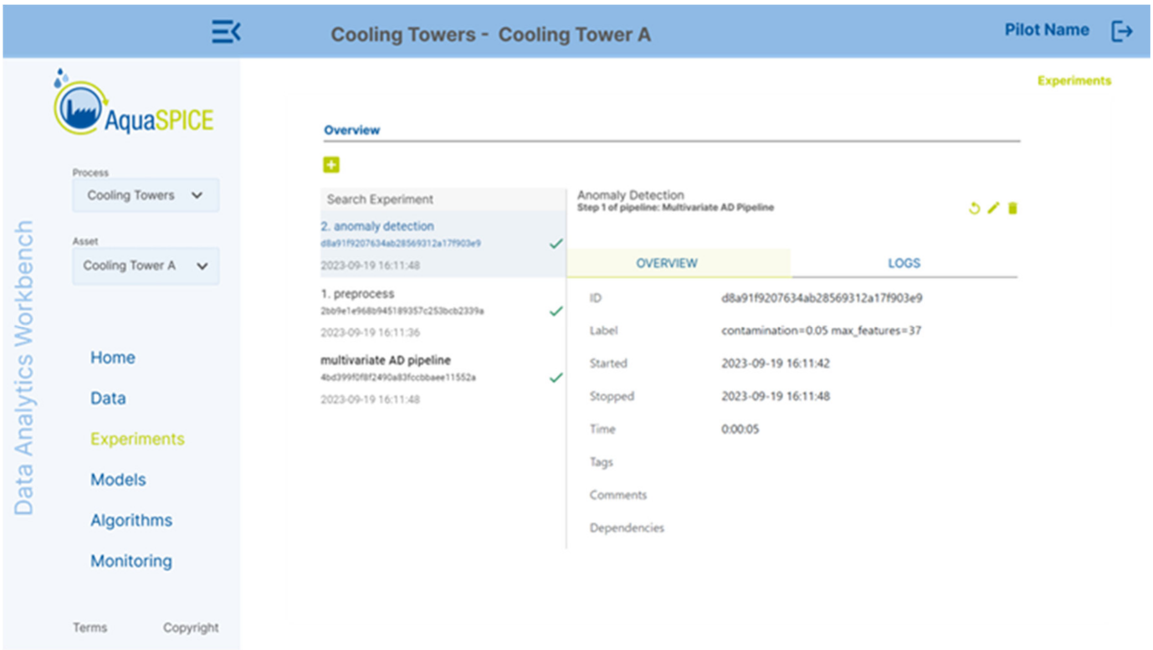
Data Preview

Date	Variable x	Variable y	Variable z	Variable s	Variable t
12/12/2022	12	20	100	30.22	400
13/12/2022	15	21	120	31.5	401
14/12/2022	14	22	110	30	400
15/12/2022	16	21	112	33	402
16/12/2022	14	22	110	32.2	400
17/12/2022	15	20	100	30	401
18/12/2022	11	21	100	31.2	400

DOWNLOAD

Figure 5: Data Analytics Workbench - Data Page

The experiments page provides a list of past experiments (Figure 6) and allows users to conduct and track new experiments (Figure 7) using the selected data, algorithms, and parameters. Popular machine learning frameworks and libraries like TensorFlow, Scikit-learn, and PyOD are supported. Experiment parameters, model parameters and hyperparameters might differ per algorithm and analysis. Experiment parameters include the data analysis type (e.g., anomaly detection, forecasting, prediction), the target variable, and whether the analysis is multivariate or univariate. Additionally, there are two types of parameters in machine learning models: model parameters and hyperparameters. Model parameters are computed during the learning process and might include weights and biases. On the other hand, model hyperparameters are external to the model and need to be set before the training process, since they configure the model's architecture and behavior. The model’s performance is significantly determined by the hyperparameters, making their optimization a crucial task in ML experiments. [49] Common hyperparameters include learning rate, regularizations, number of layers and neurons in an artificial neural network, and contamination in anomaly detection algorithms. During the experiment conduction, metadata is stored including models, hyperparameters, metrics and dataset details. The user can compare and re-run past experiments as well as get informed about the experiment's status, logs, and results. Once an experiment is completed successfully, the produced model becomes available on the model page discussed below.



2. anomaly detection
d8a91f9207634ab28569312a17f903e9 ✓
2023-09-19 16:11:48

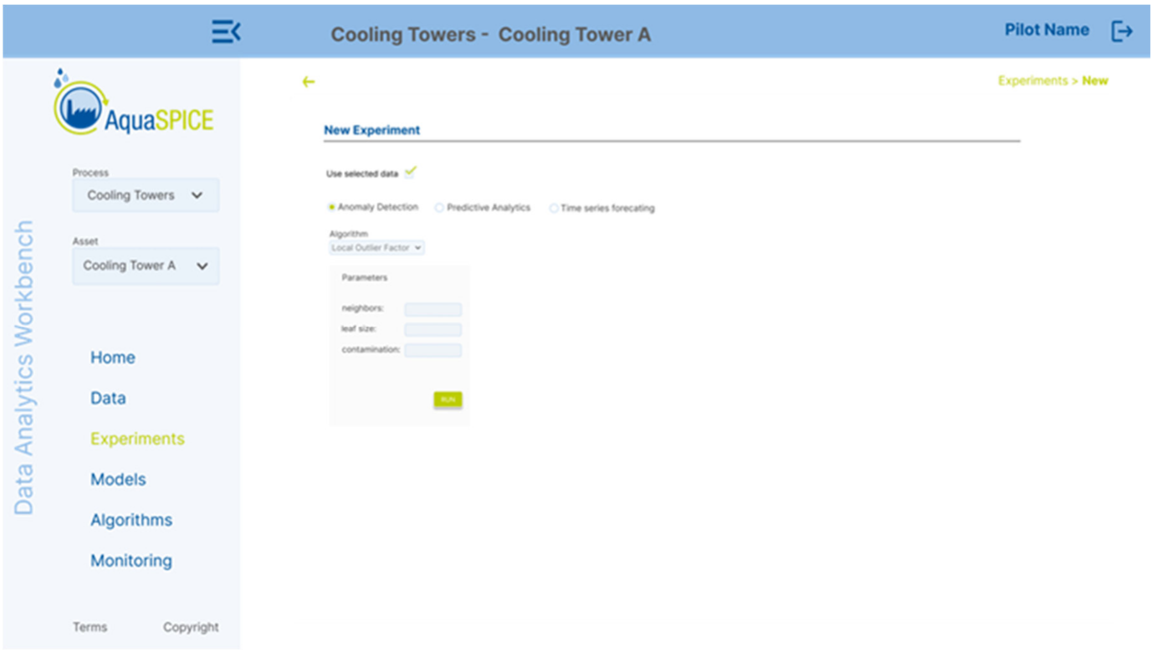
1. preprocess
2bb9e1e968b945189357c253bcb2339a ✓
2023-09-19 16:11:36

multivariate AD pipeline
4bd399f0f8f2490a83fccbbaee11552a ✓
2023-09-19 16:11:48

Anomaly Detection
Step 1 of pipeline: Multivariate AD Pipeline

ID	d8a91f9207634ab28569312a17f903e9
Label	contamination=0.05 max_features=37
Started	2023-09-19 16:11:42
Stopped	2023-09-19 16:11:48
Time	0:00:05
Tags	
Comments	
Dependencies	

Figure 6: Data Analytics Workbench - Experiments Page: Logs



Use selected data ✓

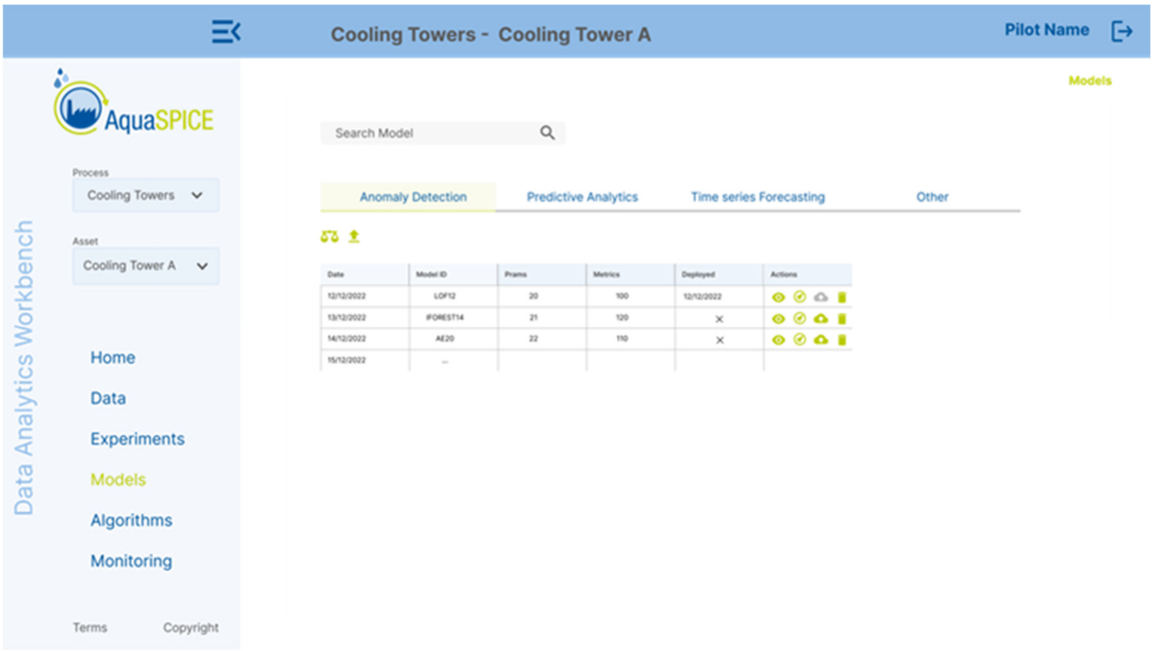
● Anomaly Detection ○ Predictive Analytics ○ Time series forecating

Algorithm
Local Outlier Factor

Parameters
neighbors:
leaf size:
contamination:

Figure 7: Data Analytics Workbench - Experiments Page: New Experiment

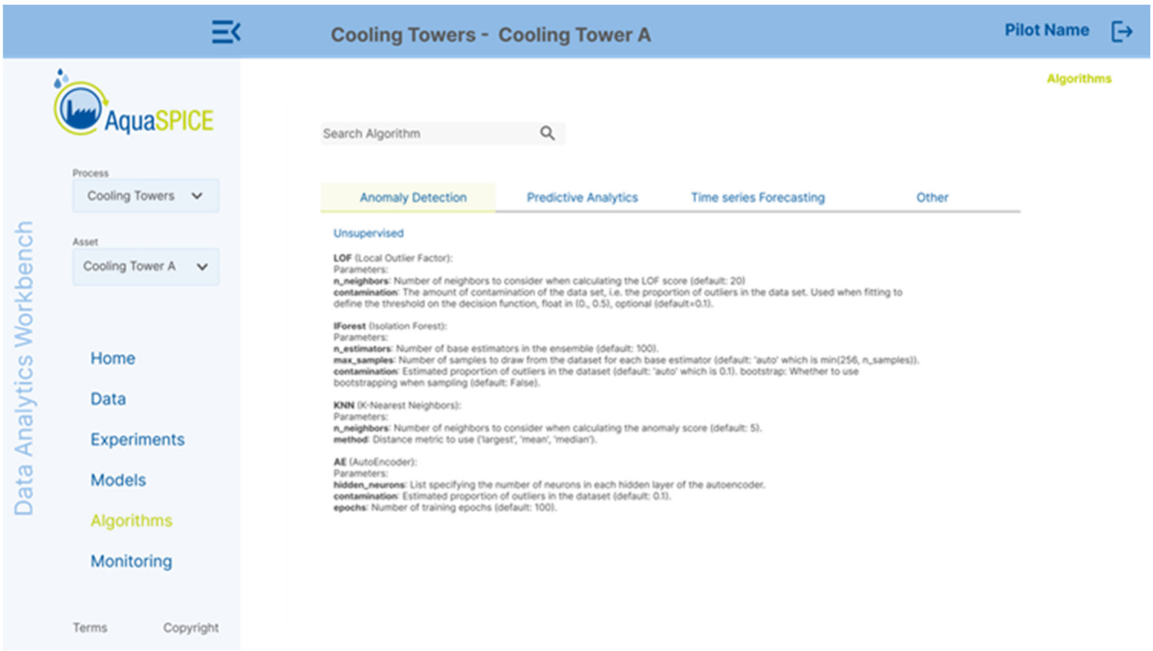
On the model page (Figure 8), insights of the models produced by the ML experiments are presented. The user can select a model from the list to view more details, optimize the model's hyperparameters, compare the model to others, delete a model and select the one to be deployed and afterwards consumed by the Data Analytics Service.



Date	Model ID	Prams	Metrics	Deployed	Actions
12/12/2022	LOF12	20	100	12/12/2022	
13/12/2022	IFOREST14	21	120	×	
14/12/2022	AE20	22	110	×	
15/12/2022	–				

Figure 8: Data Analytics Workbench - Model Page: Model Listing

Users can be informed about the algorithms that are available on the platform, their parameters and hyperparameters on the algorithms page under the relevant data analytics type tab (Figure 9).



Unsupervised

LOF (Local Outlier Factor):
Parameters:
n_neighbors: Number of neighbors to consider when calculating the LOF score (default: 20)
contamination: The amount of contamination of the data set, i.e. the proportion of outliers in the data set. Used when fitting to define the threshold on the decision function, float in (0., 0.5), optional (default=0.1).

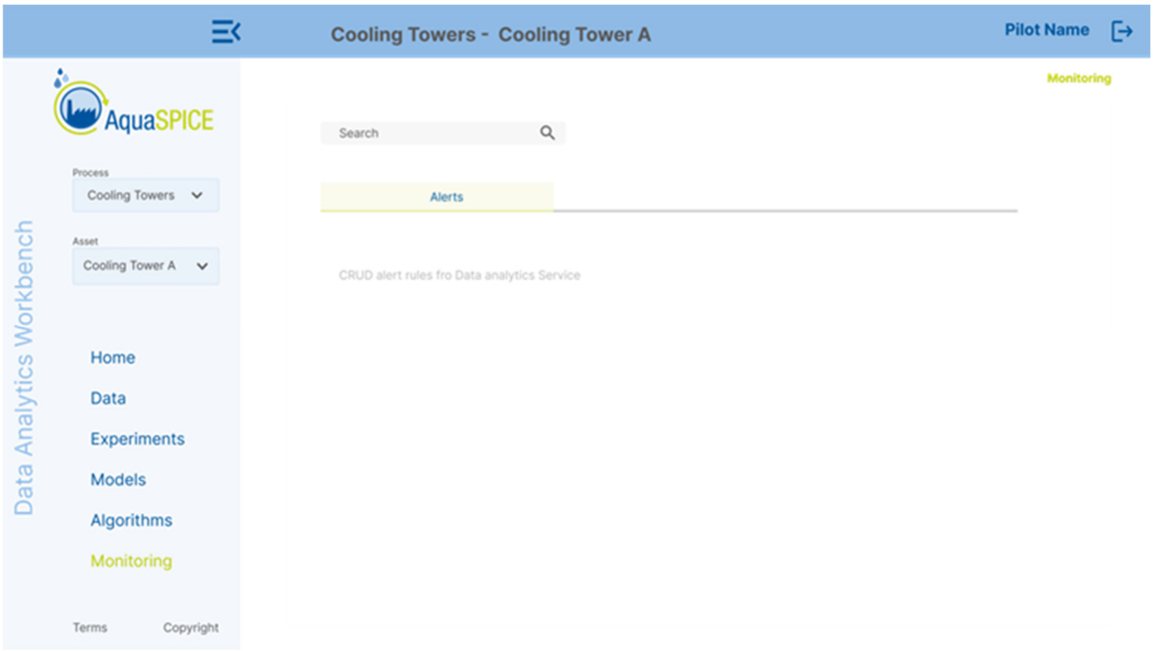
IForest (Isolation Forest):
Parameters:
n_estimators: Number of base estimators in the ensemble (default: 100).
max_samples: Number of samples to draw from the dataset for each base estimator (default: 'auto' which is min(256, n_samples)).
contamination: Estimated proportion of outliers in the dataset (default: 'auto' which is 0.1). bootstrap: Whether to use bootstrapping when sampling (default: False).

KNN (K-Nearest Neighbors):
Parameters:
n_neighbors: Number of neighbors to consider when calculating the anomaly score (default: 5).
method: Distance metric to use ('largest', 'mean', 'median').

AE (AutoEncoder):
Parameters:
hidden_neurons: List specifying the number of neurons in each hidden layer of the autoencoder.
contamination: Estimated proportion of outliers in the dataset (default: 0.1).
epochs: Number of training epochs (default: 100).

Figure 9: Data Analytics Workbench - Algorithms Page

The monitoring page (Figure 10) provides an interface where users can manage alerts and/or actions to be fired when anomalies or outliers are detected in the data.



CRUD alert rules fro Data analytics Service

Figure 10: Data Analytics Workbench - Monitoring Page

5. AI and Analytics Methods, Algorithms & Models

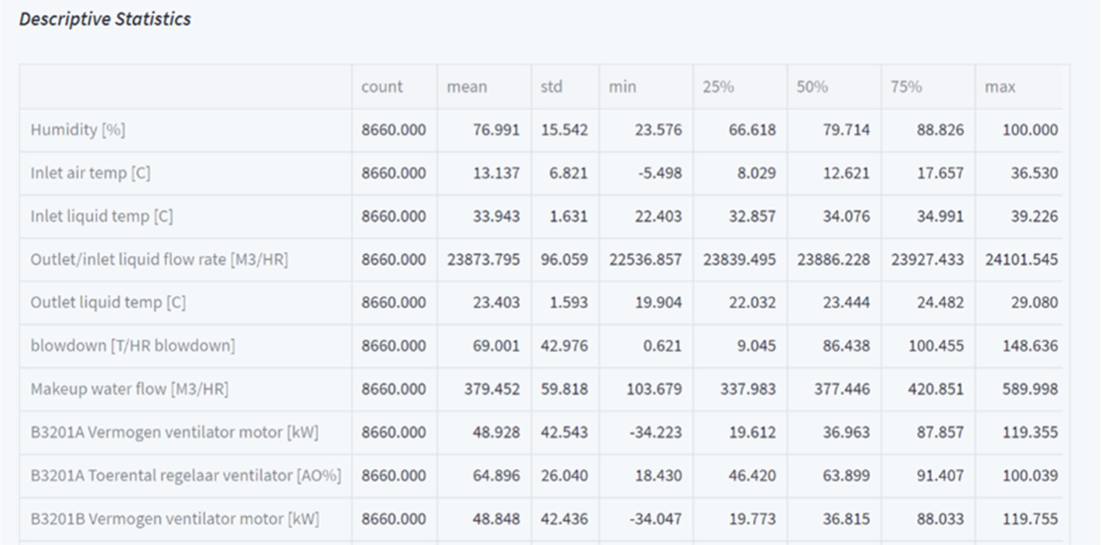
This section describes the AI and Analytics methods, algorithms and models utilized and developed as part of the data analyses.

5.1. Descriptive Analytics

Descriptive statistics are usually the first step in a data analytics process. At first, a summary of descriptive statistics is computed and presented in a table form (Figure 11) to provide an overview of the selected data. The table includes the following metrics:

1. count: The number of non-empty values.
2. mean: The average value.
3. std: The standard deviation (std) measures the amount of variation or dispersion in the dataset.
4. min: The minimum value.
5. 25%: The 25% percentile.
6. 50%: The 50% percentile.
7. 75%: The 75% percentile.
8. max: The maximum value.

Percentile in 5,6 and 7 means how many of the values are less than the given percentile. Regarding std, a low value indicates that the data are close to the mean and a higher value suggests the data are spread out over a larger range of values.



Descriptive Statistics

	count	mean	std	min	25%	50%	75%	max
Humidity [%]	8660.000	76.991	15.542	23.576	66.618	79.714	88.826	100.000
Inlet air temp [C]	8660.000	13.137	6.821	-5.498	8.029	12.621	17.657	36.530
Inlet liquid temp [C]	8660.000	33.943	1.631	22.403	32.857	34.076	34.991	39.226
Outlet/inlet liquid flow rate [M3/HR]	8660.000	23873.795	96.059	22536.857	23839.495	23886.228	23927.433	24101.545
Outlet liquid temp [C]	8660.000	23.403	1.593	19.904	22.032	23.444	24.482	29.080
blowdown [T/HR blowdown]	8660.000	69.001	42.976	0.621	9.045	86.438	100.455	148.636
Makeup water flow [M3/HR]	8660.000	379.452	59.818	103.679	337.983	377.446	420.851	589.998
B3201A Vermogen ventilator motor [kW]	8660.000	48.928	42.543	-34.223	19.612	36.963	87.857	119.355
B3201A Toerental regelaar ventilator [AO%]	8660.000	64.896	26.040	18.430	46.420	63.899	91.407	100.039
B3201B Vermogen ventilator motor [kW]	8660.000	48.848	42.436	-34.047	19.773	36.815	88.033	119.755

Figure 11: Descriptive Statistics Table Sample

Subsequently, interactive descriptive plots allow end users to explore the data and gain insights. User interactions include zoom in/out on specific data points and hovering over plot areas to see more details. Line plots are useful for visualizing data that change over time. In a line plot, the information is displayed as a series of data points connected by line segments. Two-line plots are featured, one that includes all the variables of the selected asset (Figure 12) and another that presents only the selected variable (Figure 13). For the multivariate line plot's needs the data points are scaled to fit in and the user can select which variables to draw inside the plot, allowing for visual comparison. The line plot that contains only one variable, presents the real data values and provides a precise view of the data points over time.

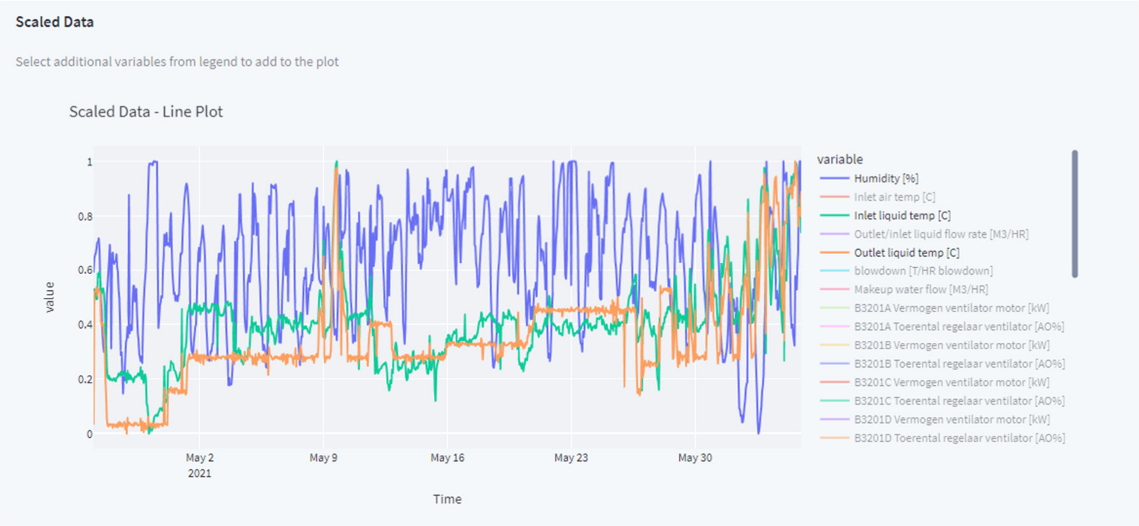


Figure 12: Multivariate Line Plot Sample

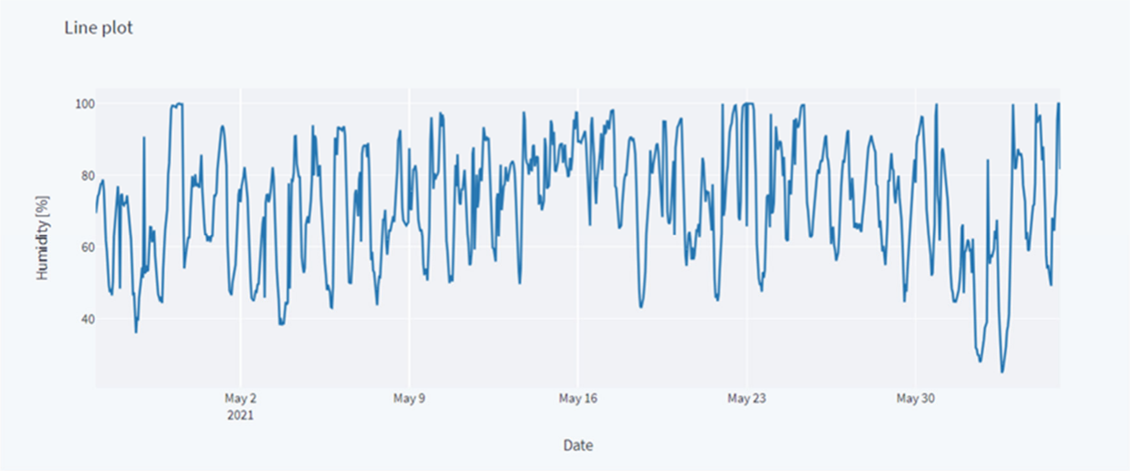


Figure 13: Sample Univariate Line Plot for Humidity

For the selected variable and date range, violin plots are produced (Figure 14). The violin plot introduced by Hintze and Nelson in 1998 features a boxplot together with a plot of the density trace. As in the usual boxplot, the whiskers are extended to the farthest points within 1.5 inter-quartile ranges from the 25th and 75th percentiles [50]. A violin plot is a method of plotting numeric data, similar to box plots, except that they also show the probability density of the data at different values, usually smoothed by a kernel density estimator.

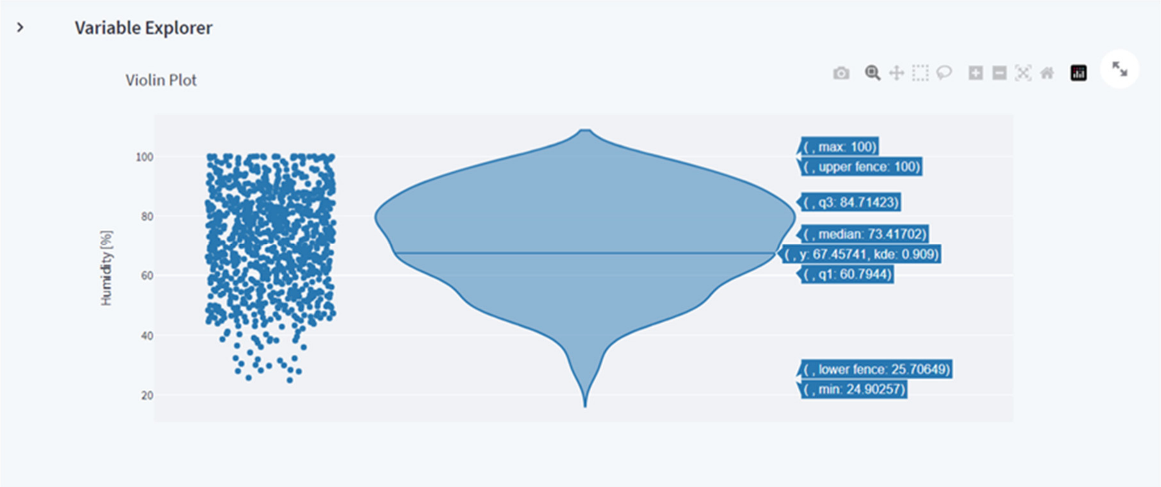


Figure 14: Sample Violin Plot of Humidity

Additionally, for a specific variable and time range, time series decomposition was performed (Figure 15). Time series decomposition involves thinking of a series as a combination of level, trend, seasonality, and noise components described as follows:

1. Observed values (level): real data points
2. Trend: long-term tendency of data
3. Seasonality: regular predictable fluctuations in data
4. Residuals (noise): what remains after removing trend and seasonality from the data [51]

Decomposition serves as a useful abstract model for conceptualizing time series. It aids better understanding of patterns and latent behaviors and can lead to more comprehensive problem-solving and decision-making.

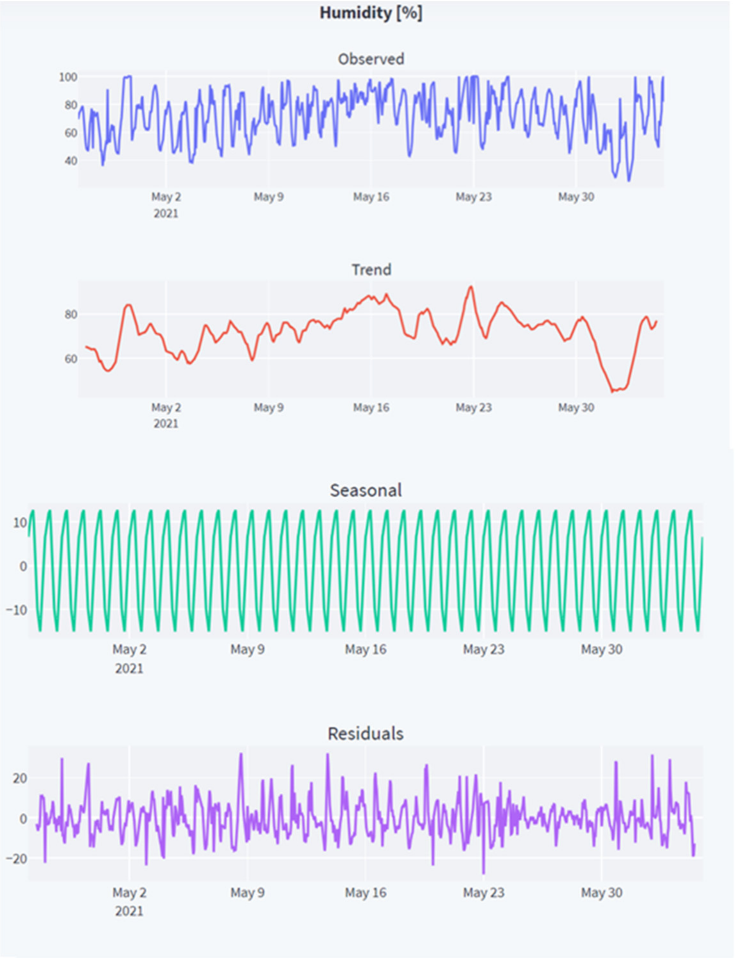


Figure 15: Time Series Decomposition of Humidity Samples

5.2. Predictive Analytics

Predictive analytics encloses techniques for forecasting future values from historical data. In deep learning, neural network models and particularly Long Short-Term Memory (LSTM) networks, are widely utilized to analyze sequential and time series data. LSTM units [52] are a special kind of gated Recurrent Neural Network that, unlike traditional RNNs, overcomes the vanishing and exploding gradients problem. This problem states that gradients during training may vanish or explode exponentially over many time steps. LSTMs excel in learning long-term dependencies which makes them suitable for time series data; this is achieved through a sophisticated system of gates between LSTM cells that allow the model to selectively remember or forget information [53]. LSTM modules consist of a memory cell and four gates: input, input modulation, forget, and output (Figure 16). The memory cell allows for learning long-term dependencies. Concerning the gates, the input gate decides which new data is important to be added to the cell state and typically involves a sigmoid function. As a part of the input gate, an input modulation gate processes the input data. This process involves a normalization of the data values within a specific range to enhance their interpretability by the LSTM cell. To achieve this transformation, a non-linear function, such as the tanh (hyperbolic tangent) function, is applied to the input data. The forget gate is responsible for deciding what data need to be discarded from the cell state and also involves a sigmoid function. Finally, the output gate computes the next hidden state from the cell state [53][54].

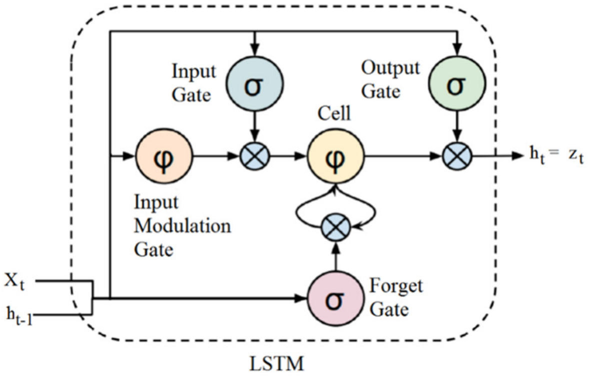


Figure 16: LSTM Cell (adapted from [54])

Deep learning models using LSTM networks were developed to forecast future values of a target variable considering the past values of the whole sequence. Prior to starting the model training, several preparatory steps are essential. These steps include handling missing values in the dataset and scaling the values to be interpretable by the neural network.

The predictive problem considers {y} previous values of all the relevant variables and predicts the next {x} values of a selected target variable. To create a dataset from which the model can learn and solve the forementioned problem, the dataset needs to be converted into a supervised learning format. This conversion in the context of time series forecasting, entails that the dataset is restructured so that each row contains both the {y} previous inputs and the {x} outputs to be forecasted. This is achieved by shifting the observations to align the current time (t) and future times (t+1, t+n) as forecast targets, with past observations (t-1, t-n) serving as input features. {y} and {x} are considered as dynamic variables.

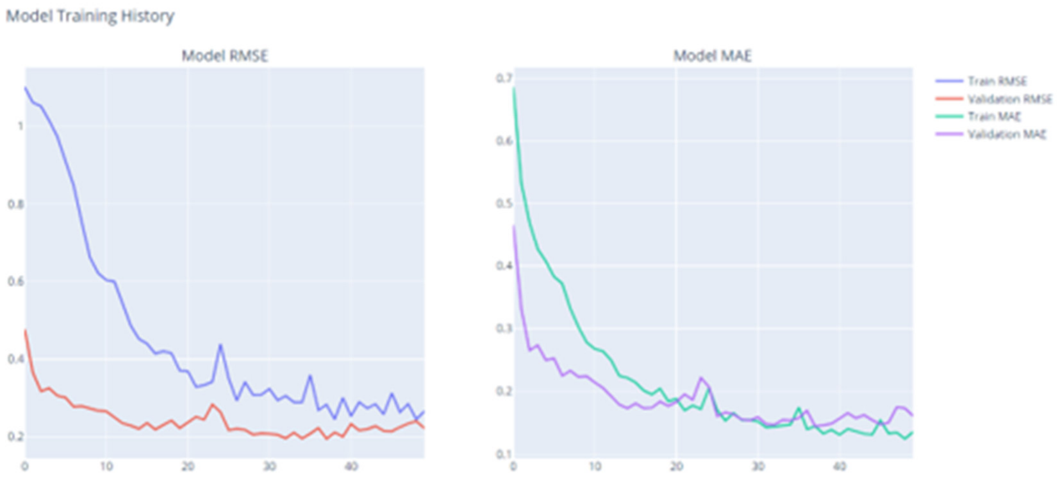


Figure 17: Model Training RMSE and MAE

After the appropriate data preprocessing and transformations are applied, the model is trained and evaluated so that a basis of hyperparameters is identified (Figure 17). Since the base model is determined, hyperparameter optimization process starts to find the optimal number of epochs, neurons per layer, dropout values and regularizers. With the success of the hyperparameter optimisation, the best and more suitable model is selected, evaluated and finalized (trained again) using the whole dataset. Finally, the model is utilized for predictions and its results are visualized in a line plot that contains the real data points, the previously predicted and the forecasted future values (Figure 18).

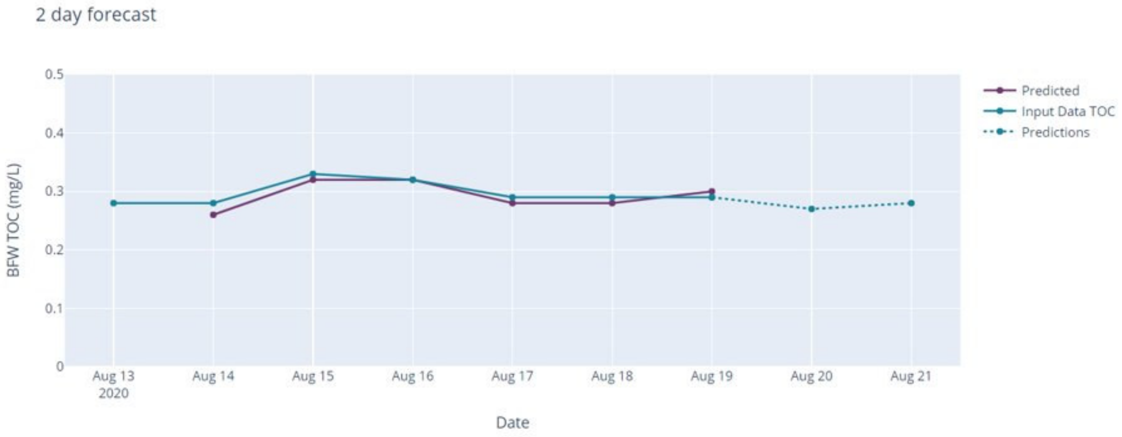


Figure 18: Line Plot of Sample Predicted vs Actual Values

Regarding the detection of anomalies in data sequences, multivariate anomaly detection is performed using unsupervised algorithms such as isolation forest and local outlier factor. Considering the unsupervised approach and the lack of labeled data for evaluation, Isolation Forest was selected due to its faster execution and effectiveness in detecting anomalies in high-dimensional datasets [55][56][57]. Furthermore, the algorithm points out anomalies directly instead of profiling the normal patterns, making it valuable in environments where "normal" is challenging to pinpoint or changes over time. Isolation forest was thought to be a robust choice for such a real-world application that needs to detect anomalies in a timely and efficient way [56][57][58].

Isolation Forest, initially proposed in 2008, builds a structure of decision trees (iTrees) for the dataset to isolate every data value and identifies anomalies as those values that have short average path lengths on the iTrees [58]. The algorithm is based on the premise that the anomalies are easier to isolate and therefore require shorter paths [55][58]. A representation of an Isolation Forest and its iTrees is depicted in Figure 19, where light blue nodes represent the normal observations, blue nodes represent rare normal values and red nodes present detected anomalous observation.

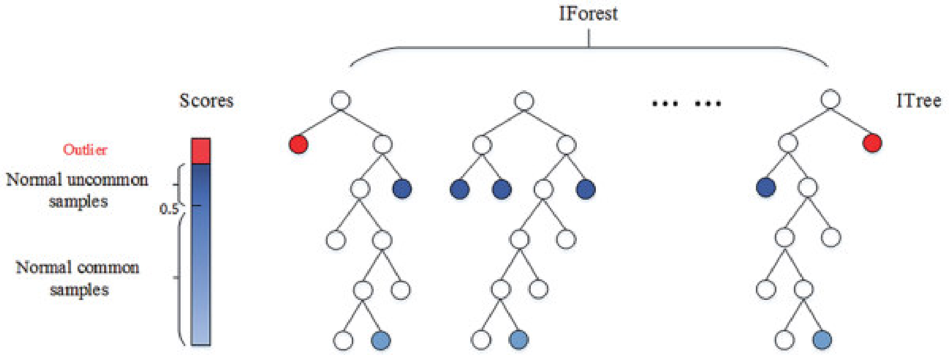


Figure 19: Isolation Forest: Trees and Scores [56][55]

The results of the Isolation Forest for the multivariate analysis are visualized using a bar plot that highlights the five most contributing features (Figure 20). On the other hand, for the univariate outlier detection the results are represented through a line plot that includes the actual data points of the examined variable with the detected outliers marked in red (Figure 21). In both cases the related observations are displayed in a table format.

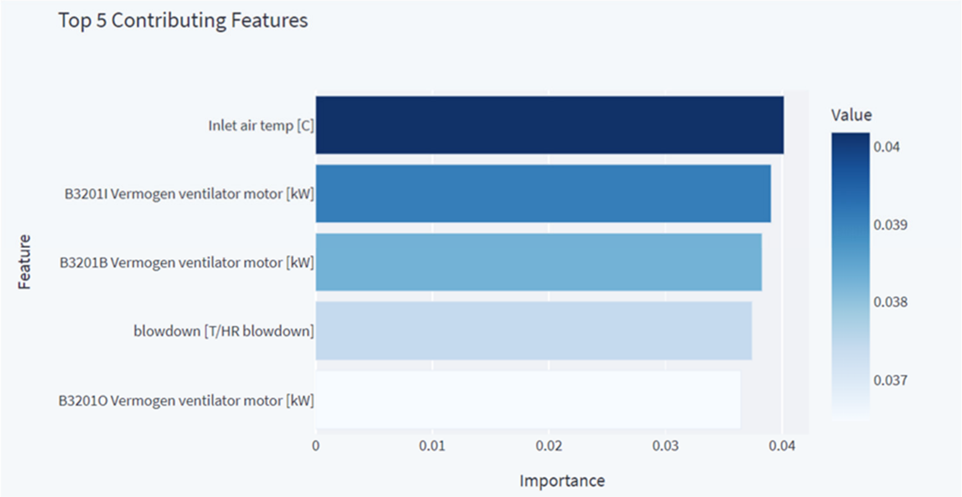


Figure 20: Multivariate Anomaly Detection using Isolation Forest in Cooling Tower Data: Top 5 Contributing Features

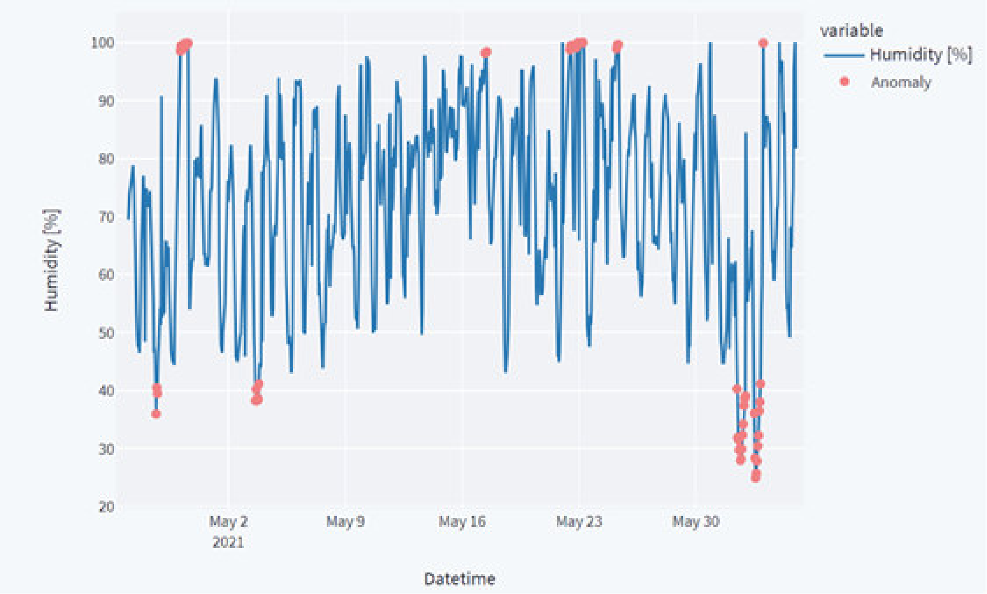


Figure 21: Univariate Anomaly Detection: Line Plot of Humidity and Detected Outliers

6. Conclusions and Next Steps

Based on the broad requirements elicited by use case partners, Task 4.5 designed and developed a data analytics platform for both real-time water monitoring data and historical data. The target user group of the platform is water management professionals (engineers) and data analysts. This platform provides a suite of data analytics functionalities. It includes basic tasks like descriptive statistics and diagnostics, as well as more advanced features like predictive analytics and anomaly detection. The results are visualized to offer clear insights and graphs. Furthermore, it allows analysts to conduct complex analyses and experiment with machine learning models through a user-friendly graphical interface, which includes features for tracking experiments and logging. Analysts can also choose and deploy models that are later used to generate data analytics insights.

In this deliverable, in addition to the platform, a generic, initial set of analytics methods and models are reported, corresponding to the generic requirements elicited from use case partners. These methods will be specialized, adapted and further extended to cover specialised requirements. Further specialization to represent the use cases applications fully and accurately is currently undertaken and will be delivered and reported in the final version of D4.8.

Limitations and risks that may hamper the uptake of the platform by use case partners include the limited availability of labelled data sets that can be used for training of supervised machine learning methods. For example, although all use case partners have requested the provision of anomaly detection methods, no training data sets with labelled anomalies exist. To this end, unsupervised methods for anomaly detection have been used that do not require labelled datasets; however, their accuracy requires further investigation and analysis before use cases can utilise them operationally.

Benefits of our platform include the microservice-based architecture, capability to cope with both historical and real-time data coming from the RTM platform of AquaSPICE, tight integration with the WaterCPS (including look & feel), the provision of generic methods and models as default offering to users but also the capability to support bespoke/custom models.

In addition to the development of bespoke and specific models and methods for the AquaSPICE use cases, our future work will include the development of a soft-sensor method as well as a method for synergistically coupling data-driven and first-principle models that will enable calibration and support improved accuracy and performance.

7. References

[1] Sivarajah, U., Kamal, M.M., Irani, Z. and Weerakkody, V., (2017). Critical analysis of Big Data challenges and analytical methods. Journal of business research, 70, pp.263-286.

[2] Vassakis, K., Petrakis, E., & Kopanakis, I. (2017). Big Data Analytics: Applications, Prospects and Challenges. Lecture Notes on Data Engineering and Communications Technologies, 3–20. doi:10.1007/978-3-319-67925-9_1.

[3] Nishchol Mishra et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (3), 2012, 4434- 4438.

[4] Tranfield, D., Denyer, D. and Smart, P., (2003). Towards a methodology for developing evidence‐informed management knowledge by means of systematic review. British journal of management, 14(3), pp.207-222.

[5] Gopalakrishnan, S. and Ganeshkumar, P., (2013). Systematic reviews and meta-analysis: understanding the best evidence in primary healthcare. Journal of family medicine and primary care, 2(1), p.9.

[6] Snyder, H., Witell, L., Gustafsson, A., Fombelle, P. and Kristensson, P., (2016). Identifying categories of service innovation: A review and synthesis of the literature. Journal of Business Research, 69(7), pp.2401-2408.

[7] Snyder H. (2019). Literature review as a research methodology: An overview and guidelines, Journal of Business Research, Volume 104, p.333-339, ISSN 0148-2963.

[8] Adams, V.D., (2017). Water and wastewater examination manual. Routledge.

[9] ©Energy and Water Utilities Regulatory Authority (EWURA), (2020). Water and Wastewater Quality Monitoring Guidelines- 2nd Edition.

[10] Ching, P.M., So, R.H. and Morck, T., (2021). Advances in soft sensors for wastewater treatment plants: A systematic review. Journal of Water Process Engineering, 44, p.102367.

[11] Stamatelatou, K. and Tsagarakis, K.P., (2015). Sewage treatment plants: economic evaluation of innovative technologies for energy efficiency. IWA Publishing.

[12] Liu, Y., Chen, J., Sun, Z., Li, Y. and Huang, D., (2014). A probabilistic self-validating soft-sensor with application to wastewater treatment. Computers & chemical engineering, 71, pp.263-280.

[13] Shang, C., Yang, F., Huang, D. and Lyu, W., (2014). Data-driven soft sensor development based on deep learning technique. Journal of Process Control, 24(3), pp.223-233.

[14] Liu, J., Jang, S.S. and Wong, D.S.H., (2016). Developing a soft sensor with online variable selection for industrial multi-mode processes. In Computer Aided Chemical Engineering (Vol. 38, pp. 398-403). Elsevier.

[15] Brunner, V., Siegl, M., Geier, D. and Becker, T., (2021). Challenges in the development of soft sensors for bioprocesses: A critical review. Frontiers in Bioengineering and Biotechnology, p.730.

[16] Mesquita, D.P., Amaral, A.L. and Ferreira, E.C., (2013). Activated sludge characterization through microscopy: A review on quantitative image analysis and chemometric techniques. Analytica Chimica Acta, 802, pp.14-28.

[17] Tomperi, J., Koivuranta, E., Kuokkanen, A., & Leiviskä, K. (2016). Modelling effluent quality based on a real-time optical monitoring of the wastewater treatment process. Environmental Technology, 38(1), 1–13. doi:10.1080/09593330.2016.1181674

[18] Koivuranta, E., Stoor, T., Hattuniemi, J. and Niinimäki, J., (2015). On-line optical monitoring of activated sludge floc morphology. Journal of Water Process Engineering, 5, pp.28-34.

[19] Guillot M., Azouzi R., Cote MC. (1994) Process monitoring and control. In: Dagli C.H. (eds) Artificial Neural Networks for Intelligent Manufacturing. Intelligent Manufacturing Series. Springer, Dordrecht.

[20] Zimek, A. and Filzmoser, P., (2018). There and back again: Outlier detection between statistical reasoning and data mining algorithms. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(6), p.e1280.

[21] Kumar, K. & Pitta, Sundara & Babu M, John. (2010). Performance evaluation of waste water treatment plant. International Journal of Engineering Science and Technology. 2.

[22] Boujelben, I., Samet, Y., Messaoud, M., Makhlouf, M.B. and Maalej, S., (2017). Descriptive and multivariate analyses of four Tunisian wastewater treatment plants: a comparison between different treatment processes and their efficiency improvement. Journal of environmental management, 187, pp.63-70.

[23] Asami, H., Golabi, M. and Albaji, M., (2021). Simulation of the biochemical and chemical oxygen demand and total suspended solids in wastewater treatment plants: Data-mining approach. Journal of Cleaner Production, 296, p.126533.

[24] Han, H.G., Chen, Q.L. and Qiao, J.F., (2011). An efficient self-organizing RBF neural network for water quality prediction. Neural networks, 24(7), pp.717-725.

[25] Zare Abyaneh, H., (2014). Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters. Journal of Environmental Health Science and Engineering, 12(1), pp.1-8.

[26] Guo, H., Jeong, K., Lim, J., Jo, J., Kim, Y.M., Park, J.P., Kim, J.H. and Cho, K.H., (2015). Prediction of effluent concentration in a wastewater treatment plant using machine learning models. Journal of Environmental Sciences, 32, pp.90-101.

[27] Tomperi, J., Koivuranta, E. and Leiviskä, K., 2017. Predicting the effluent quality of an industrial wastewater treatment plant by way of optical monitoring. Journal of water process engineering, 16, pp.283-289.

[28] Nadiri, A.A., Shokri, S., Tsai, F.T.C. and Moghaddam, A.A., (2018). Prediction of effluent quality parameters of a wastewater treatment plant using a supervised committee fuzzy logic model. Journal of cleaner production, 180, pp.539-549.

[29] Lepot, M., Aubin, J.B., Clemens, F.H. and Mašić, A., (2017). Outlier detection in UV/Vis spectrophotometric data. Urban Water Journal, 14(9), pp.908-921.

[30] Chow, C.W., Liu, J., Li, J., Swain, N., Reid, K. and Saint, C.P., (2018). Development of smart data analytics tools to support wastewater treatment plant operation. Chemometrics and Intelligent Laboratory Systems, 177, pp.140-150.

[31] Zadorojniy, A., Wasserkrug, S., Zeltyn, S. and Lipets, V., (2019). Unleashing analytics to reduce costs and improve quality in wastewater treatment. INFORMS Journal on Applied Analytics, 49(4), pp.262-268.

[32] Arismendy, L., Cárdenas, C., Gómez, D., Maturana, A., Mejía, R. and Quintero M, C.G., (2020). Intelligent system for the predictive analysis of an industrial wastewater treatment process. Sustainability, 12(16), p.6348.

[33] Arismendy, L., Cárdenas, C., Gómez, D., Maturana, A., Mejía, R. and Quintero M, C.G., (2021). A Prescriptive Intelligent System for an Industrial Wastewater Treatment Process: Analyzing pH as a First Approach. Sustainability, 13(8), p.4311.

[34] Xiao, H., Huang, D., Pan, Y., Liu, Y. and Song, K., (2017). Fault diagnosis and prognosis of wastewater processes with incomplete data by the auto-associative neural networks and ARMA model. Chemometrics and Intelligent Laboratory Systems, 161, pp.96-107.

[35] Andersson, S., Rahmberg, M., Nilsson, Å., Grundestam, C., Saagi, R. and Lindblom, E., (2020). Evaluation of environmental impacts for future influent scenarios using a model-based approach. Water Science and Technology, 81(8), pp.1615-1622.

[36] Blanco-Rodríguez, A., Camara, V.F., Campo, F., Becherán, L., Durán, A., Vieira, V.D., de Melo, H. and Garcia-Ramirez, A.R., 2018. Development of an electronic nose to characterize odours emitted from different stages in a wastewater treatment plant. Water research, 134, pp.92-100.

[37] Moreira de Lima, J.M. and Ugulino de Araújo, F.M., (2021). Industrial semi-supervised dynamic soft-sensor modeling approach based on deep relevant representation learning. Sensors, 21(10), p.3430.

[38] Biegler, L. T., Yang, X., & Fischer, G. A. G. (2015). Advances in sensitivity-based nonlinear model predictive control and dynamic real-time optimization. Journal of Process Control, 30, 104-116.

[39] Yaqoob, I., Hashem, I. A. T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N. B., & Vasilakos, A. V. (2016). Big data: From beginning to future. International Journal of Information Management, 36(6), 1231-1247.

[40] Tanenbaum, A. S., & Van Steen, M. (2007). Distributed systems: principles and paradigms. Prentice-Hall.

[41] Aykol, M., Gopal, C. B., Anapolsky, A., Herring, P. K., van Vlijmen, B., Berliner, M. D., ... & Storey, B. D. (2021). Perspective—combining physics and machine learning to predict battery lifetime. Journal of The Electrochemical Society, 168(3), 030525.

[42] Cozad, A., Sahinidis, N. V., & Miller, D. C. (2015). A combined first-principle and data-driven approach to model building. Computers & Chemical Engineering, 73, 116-127.

[43] Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of Big Data–evolution, challenges and research agenda. International Journal of Information Management, 48, 63-71.

[44] Sarnovsky, M., Bednar, P., & Smatana, M. (2019). Cross-sectorial semantic model for support of data analytics in process industries. Processes, 7(5), 281.

[45] Redyuk, S. (2019). Automated Documentation of End-to-End Experiments in Data Science. 2019 IEEE 35th International Conference on Data Engineering (ICDE), 2076-2080.

[46] Molner Domenech, A., & Guillén, A. (2020). ml-experiment: A Python framework for reproducible data science. Journal of Physics: Conference Series, 1603(1), 012025.

[47] Goniwada, S.R. (2022). Microservices Architecture and Design. In: Cloud Native Architecture and Design. Apress, Berkeley, CA.

[48] Li, Z., Seco, D., & Rodriguez, A., (2019). Microservice-Oriented Platform for Internet of Big Data Analytics: A Proof of Concept. Sensors (Basel, Switzerland), 19.

[49] Li Yang, Abdallah Shami, (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice, Neurocomputing, 415, p. 295-316, ISSN 0925-2312.

[50] Hintze JL, Nelson RD (1998) Violin Plots: A Box Plot-Density Trace Synergism. The American Statistician 52:181-184.

[51] Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). STL: A seasonal-trend decomposition. J. Off. Stat, 6(1), 3-73.

[52] S. Hochreiter and J. Schmidhuber (1997). Long short-term memory, Neural Computation. MIT Press.

[53] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Chapter 10. Deep Learning. pp 367 MIT Press.

[54] Donahue, J., Hendricks, L. A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., & Darrell, T. (2017). Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(4), 677–691. doi:10.1109/tpami.2016.2599174

[55] Jung, C., Lee, Y., Lee, J., & Kim, S. (2020). Performance Evaluation of the Multiple Quantile Regression Model for Estimating Spatial Soil Moisture after Filtering Soil Moisture Outliers. Remote Sensing, 12(10), 1678. doi:10.3390/rs12101678

[56] Chen, W.-R., Yun, Y.-H., Wen, M., Lu, H.-M., Zhang, Z.-M., & Liang, Y.-Z. (2016). Representative subset selection and outlier detection via isolation forest. Analytical Methods, 8(39), 7225–7231. doi:10.1039/c6ay01574c

[57] Togbe, M.U. et al. (2020). Anomaly Detection for Data Streams Based on Isolation Forest Using Scikit-Multiflow. In: Gervasi, O., et al. Computational Science and Its Applications – ICCSA 2020. ICCSA 2020. Lecture Notes in Computer Science, vol 12252. Springer, Cham. https://doi.org/10.1007/978-3-030-58811-3_2

[58] Liu, F. T., Ting, K. M. & Zhou, Z.-H. (2008). Isolation forest. *2008 Eighth IEEE International Conference on Data Mining* (p./pp. 413–422)

[59] Newhart, K.B., Holloway, R.W., Hering, A.S. and Cath, T.Y., (2019). Data-driven performance analyses of wastewater treatment plants: A review. Water research, 157, pp.498-513.

[60] Haimi, H., Mulas, M., Corona, F., & Vahala, R. (2013). Data-derived soft-sensors for biological wastewater treatment plants: An overview. Environmental Modelling & Software, 47, 88–107. doi:10.1016/j.envsoft.2013.05.009